Sequential Clustering and Dimension Reduction Algorithm of Time Series Data

Tanran Zheng
Institute of Industrial Economics
Jinan University
Guangzhou, China
zhengtr@gmail.com

***Abstract*—This paper presents a new clustering and dimension reduction algorithm that processes high dimensional time series data. This issue has been essential for various fields, especially in data science, medicine, machine learning, weather or earthquake forecast, and finance, that useful information need to be effectively extracted from time-series data of varieties of sources. The proposed algorithm is inspired by classic data science algorithm, such as the K-Nearest Neighbor algorithm and the greedy algorithm. It first applies clustering to the high dimensional time-series data in a specific sequential order, which is determined by a heuristic function deduced by effectiveness of the data; after the completion of the clustering, it sequentially orthonormalizes the data in each group by implementing Ordinary Least Squares (OLS); finally, it calculates the weighted average of the new orthonormalized data in each group to reduce the dimension of the whole data sample. This article also theoretically and empirically examines this algorithm using real-world financial time-series data and Multi-Factor investment model. The results indicate that the proposed algorithm can effectively cluster and reduce dimension of time-series data, therefore significantly improve the model's performance.**

Keywords—Unsupervised clustering; dimension reduction; time-series data analysis

1 Introduction

Time series data is one of the most important data types in the discipline of data science, and it is also widely applied to economics, finance, machine learning, weather forecasting, earthquake forecasting, medical diagnostics, and other fields. As the amount of available data in various fields grows exponentially, problems arise and require dealing with time-series data of great size and dimensionality. For this reason, curse of dimensionality [1] was proposed by Richard E. Bellman. It refers to the phenomena that when the dimensionality of data increases, the space volume of data increases too rapidly such that the sampling data becomes sparse and hard to compute the distance. As a result, it is likely to be prohibitive to obtain reliable statistically significant results. Also, high dimensionality often brings in the issue of multicollinearity [2], misleading the result of feature extraction when solving problems. Therefore, dimension reduction has become a critical topic in the field of data science.

Dimension reduction refers to the process of extracting principle components with relatively low pairwise correlation from the massive and high dimensional raw data, in order to reduce the number of random variables under certain constraints. It can be further divided into two main categories: random variable selection and feature extraction. The sequential clustering and dimension reduction algorithm introduced in this paper will utilize the former, random variable selection, which could also be interpreted as identifying the most informative data from the raw data.

In previous researches, the methods of clustering and dimension reduction for time-series data are mainly based on and derived from Dynamic Time Warping or Principle Components Analysis. These methods are widely used in signal analysis and processing, voice recognition, recommendation system, etc. However, it still may result in issues like low precision for clustering or impairment of original information, when processing noisy or non-stationary data. Additionally, the propriety of the outcomes of such methods is also prone to insufficient correct data samples.

The sequential clustering and dimension reduction algorithm that is introduced in this paper, will mainly employ the ideas of k-nearest neighbor and greedy algorithm to apply unsupervised sequential clustering analysis to the time-series data. The algorithm is conducted in a sequential order, in other words, data will be processed by the algorithm one-by-one. The order of choosing the data will make use of different heuristic functions, based on varying features of the data and scenario, to evaluate the importance and effectiveness of the data. Higher score implies higher priority. At each time step, the proposed algorithm will first cluster the raw data based on cross-sectional pairwise correlation. Then the algorithm will conduct pairwise ordinary least square regression analysis within each cluster, and use the weighted sum of the resulting residual of the regression as a new set of data. In such way, the dimension of the data is reduced to the number of clusters. Each part of the algorithm, clustering, and dimension reduction, can be also implemented by itself as separate algorithm.

Comparing to conventional clustering and dimension reduction methods, the algorithm being introduced does not rely on large amount of labeled data, and is able to take the importance of the raw data into account. It only requires estimating on the thresholds of clustering and assigning the proper heuristic functions for given application scenarios. The algorithm is simple and robust, but at the same time preserves the valid information contained within the data as much as possible. In this paper, the algorithm will be applied to the quantitative multi-factor model to perform sequential clustering and dimension reduction for a large amount of unlabeled stock factors. It can be proved in this paper through backtesting that, the financial data after processed by the proposed algorithm leads to a better result in the multi-factor model, comparing that of conventional clustering and dimension reduction methods.

2 Methodology

The main procedure for sequential clustering and dimension reduction algorithm is as followed:

I. Data pre-processing;

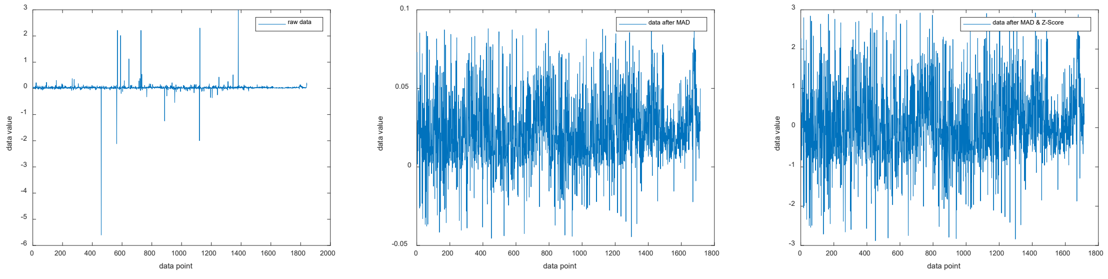
II. Sequential clustering;

III. Sequential dimension reduction.

2.1 Data pre-processing

Raw data needs to be pre-processed before applying the proposed algorithm. Quality of data is one of the determining factors for the prediction precision and generalization capacity of a model. The main purpose of data preprocessing is to ensure accuracy, integrity, consistency, timeliness, reliability, and interpretability of the data, such that the model gets a set of normalized, clean and continuous data as input. To better the performance of the proposed algorithm, the following data pre-processing procedure is conducted.

Two main steps for data pre-processing are fixing outliers and standardizing. In the introduced algorithm, median absolute deviation (MAD) will be applied to fix the outliers, and Z-Score is used for standardizing. Upon the completion of data pre-processing, all data will be converted to the same dimension and made calculation and comparison between these data possible. See Figure 1 for the results demonstration of data pre-process. Two main steps for data pre-processing are fixing outliers and standardizing. In the introduced algorithm, median absolute deviation (MAD) will be applied to fix the outliers, and Z-Score is used for standardizing. Upon the completion of data pre-processing, all data will be converted to the same dimension and made calculation and comparison between these data possible. See Figure 1 for the results demonstration of data pre-process.



(a) **(b)** **(c)**

Figure 1. Results demonstration of data preprocess

(a) Raw data; (b) Data after MAD; (c) Data after MAD & Z-Score

2.1.1 Fix outliers

Median absolute deviation is a robust measurement for the sample variance of single-variable numerical data. The sample MAD also can be interpreted as the MAD of the population. For single-variable datasets $X_1, X_2, \ldots, X_n$, MAD is defined as the median of absolute standard deviation from data points to data's median,

$$\mathrm{MAD} = \mathrm{median}\left(\left|X_i - \tilde{X}\right|\right) \tag{1}$$

where $\tilde{X}$ is the median of X_i. The residuals (deviation) of the data points from the data's median can be calculated from Equation (1), then MAD is the median of the absolute values of such residuals.

MAD is a robust measurement of statistical dispersion, and it can handle the extreme outliers within the datasets more robustly than standard deviation. To compare, standard deviation is calculated using the squares of the distance from data point to data's mean, so the higher deviations will be given higher weights and have a bigger impact on the result. For MAD, a small number of outliers will not influence the final result, thus a more robust measure of scale than sample variance or standard deviation. Therefore, in the introduced algorithm, MAD is used to deal with outliers.

Before MAD is applied as a consistent estimator for standard deviation, it needs to be converted through,

$$\sigma = \kappa \cdot \mathrm{MAD} \tag{2}$$

where κ is a proportionality constant[3] depending on data distribution. Assuming the data follows normal distribution, then the value of κ could be computed from the inverse of the cumulative distribution function of the normal distribution. It is also important to make sure that ±MAD covers 50% of the normal cumulative distribution function (for the range from ¼ to ¾) [4].

$$\frac{1}{2} = \mathrm{P}(|X - \mu| < \mathrm{MAD}) = \mathrm{P}\left(\left|\frac{X-\mu}{\sigma}\right| < \frac{\mathrm{MAD}}{\sigma}\right) = P\left(|Z| < \frac{\mathrm{MAD}}{\sigma}\right) \tag{3}$$

So,

$$\Phi\left(\frac{\mathrm{MAD}}{\sigma}\right) - \Phi\left(-\frac{\mathrm{MAD}}{\sigma}\right) = \frac{1}{2} \tag{4}$$

and

$$\Phi\left(-\frac{\mathrm{MAD}}{\sigma}\right) = 1 - \Phi\left(\frac{\mathrm{MAD}}{\sigma}\right) \tag{5}$$

Substitute $\frac{3}{4}$ into the equation

$$\frac{\mathrm{MAD}}{\sigma} = \Phi^{-1}\left(\frac{3}{4}\right) = 0.67449 \tag{6}$$

to get

$$\kappa = \frac{1}{\Phi^{-1}\left(\frac{3}{4}\right)} = 1.4826 \tag{7}$$

Upon getting the constant k, the data points with distance over 3 · κ · MAD (3 standard deviations) from the sample data's median are labeled as outliers, and will in turn be fixed as κ · MAD.

$$\mathrm{X}_i = \begin{cases} \mathrm{X}_i, & \tilde{X} - 3\cdot\kappa\cdot\mathrm{MAD} \le \mathrm{X}_i \le \tilde{X} + 3\cdot\kappa\cdot\mathrm{MAD} \\ \tilde{X} - 3\cdot\kappa\cdot\mathrm{MAD}, & \mathrm{X}_i \le \tilde{X} - 3\cdot\kappa\cdot\mathrm{MAD} \\ \tilde{X} + 3\cdot\kappa\cdot\mathrm{MAD}, & \mathrm{X}_i \ge \tilde{X} + 3\cdot\kappa\cdot\mathrm{MAD} \end{cases} \tag{8}$$

2.1.2 Standardizing

Different input data might be on different dimensions and the difference between the magnitudes of data might be large. If this issue is not addressed, it may negatively influence the analysis result. So, the data needs to be proportionally scaled to a particular range and converted to the same dimension, preserving the original ranking within each data, thus is better for later analysis.

Since fixing the outliers of the data with MAD will preserve the sample mean μ and standard deviation σ, Z-Score could be applied in the next step for cross-sectional standardizing of every dataset. Before the procedure of proposed algorithm, a large number of various data needs to be compared. And by converting the data values of different dimensions to Z-Score values, which are of consistent scale, can make sure that the data is available for comparison. Z-Score values could be calculated as the data's mean and standard deviation are possible to obtain.

$$\mathrm{Z-Score} = \frac{\mathrm{X}_i - \mu}{\sigma} \tag{9}$$

Upon getting the Z-score values of all datasets, it is possible to further analyze the data based on those instead of raw data values [5].

2.2 Sequential Clustering

Sequential clustering algorithm is inspired by k-nearest neighbor (K-NN) algorithm and greedy algorithm, adding more features that include a heuristic function to give a particular order to the data when performing pairwise clustering. K-nearest neighbor is a frequently used clustering algorithm and also a relatively straightforward classic machine learning algorithm with solid theoretical support. It was introduced by Cover T and Hart P as early as 1967. However, K-Nearest Neighbor also has noticeable limitations. As one of the supervised learning algorithms, it requires a large number of hand-labeled data as training data sample [6], but not all datasets have such samples. Especially when dealing with high-dimensional data or data obtained by data mining, the data may not be able to clearly explain itself, thus K-NN algorithm would not be utilized to the full extent. One example is that, for the factors of quantitative investment in financial industry, the correlations between factors may vary from time to time. In such case the labeled training data sample could not be safely relied on to deal with multicollinearity issues. Moreover, financial data has high dimension and low signal-to-noise ratio, placing more obstacles for clustering. Therefore, the sequential clustering algorithm introduced in this paper is an unsupervised approach that can deal with the issues stating above for data clustering [7].

The basic procedure for sequential clustering algorithm is: conduct cross-sectional correlation analysis on all the unlabeled data for a past period of time to get the average of sample Spearman's correlation coefficient. Then, one at a time, assign unlabeled data to cluster based on the correlation coefficient between the unlabeled data and the cluster: if the correlation coefficient is higher than a preset threshold, the unlabeled data sample will be clustered into the group; otherwise it will be put into a new group.

Table 1. Algorithm of sequential clustering

```
Algorithm: Sequential Clustering
Parameters: $n$ unlabeled data, $T$ period data, threshold $M$
Let $X$ be unlabeled data sample $<x_1,x_2,\ldots,x_n>$
Let $H$ be the set of labeled data sample, and $h_j$ be labeled data
Let $r_{(i,j)}$ be the correlation coefficient of data $i$ and data $j$ for the past
certain periods.
for episode $t$ from 1 to $T$ do
    Choose unlabeled data $x_i$ with highest heuristic value
    if $H \neq \emptyset$ then
        Find labeled data $h_j \in H$ that has the highest
         correlation coefficient
        if $r_{(i,j)} \geq M$ then
            Assign $x_i$ to cluster $j$, add $h_j$ to $H$
            Continue
        end if
    end if
end for
```

2.2.1 Determine the order of clustering

The introduced algorithm employs different heuristic functions according to different data types in order to determine the order of choosing unlabeled data. The importance or effectiveness of the data is scored on, and the clustering starts with the data of highest score. For example, when dealing with multi-factor model in finance, stock's factors are used to predict excess return of stock in the future; therefore in this case the heuristic function should give an estimation of precision and stability of each factor's prediction. Information ratio is a measurement that is commonly used to evaluate the effectiveness of stock's factor. It can be calculated as [8]:

$$heuristic\ function: h(f,t) = IR_t = \frac{\sum_{t-n}^{t} IC(f,t)}{n \cdot \sigma_{IC}} \tag{10}$$

where IC is the correlation coefficient; n is the period for calculating mean of IC; σ_{IC} is the standard deviation of IC. Particularly, $IC(f,t)$ represents the cross-sectional correlation coefficient between the factor and the stock return rate of next term. To determine $IC(f,t)$, the Spearman's rank correlation coefficient between factor at the start of the period and return at the end of the period is computed:

$$IC(f,t) = correlation(f_{t+1}, ret_{t+1}) \tag{11}$$

where $IC(f,t)$ is the IC value at period t; f_{t+1} is the predicted value (or vector) of return based on the factor; $ret_{t+1}$is the real value of return at period $t+1$.

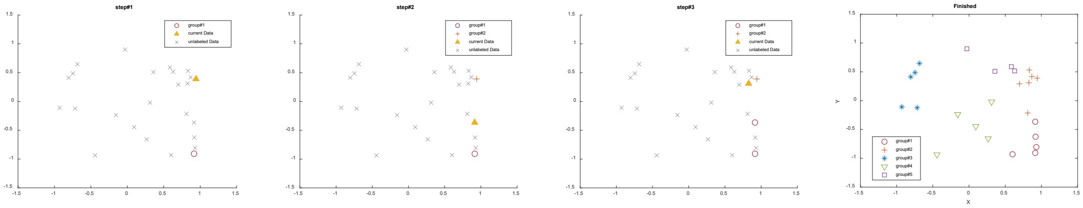
High IR value of a specific factor indicates that the factor has high prediction precision and stability for the past n period terms, which also implies that the data has strong effectiveness in multi-factor models. Thus, when applying sequential clustering and dimension reduction algorithm in multi-factor models, the factors with highest IR value is given the first order. When the proposed algorithm is going to be applied on data of other fields, the heuristic function needs to be chosen such that it can evaluate the importance and effectiveness of the data. Priority in ordering will be assigned to the data with high score calculated by the heuristic function.

2.2.2 Threshold of clustering

During the process of clustering, the algorithm requires a hyperparameter M, a preset threshold, to determine if the unlabeled data sample could be clustered into a certain cluster. For various data and its application scenarios, the threshold M could be chosen by different set of standards. As for the multi-factor model, the valid correlation is $r \in [-1, 1]$ and the range of 0.4 to 0.6 is consider high correlation between factors. Thus, the threshold M is set to 0.6 for the following back testing results.

2.2.3 Clustering

When a new set of data is being evaluated, the algorithm first needs to determine if it can be clustered into a specific group or put into a new group. Using the correlation coefficient between data as distance metric, an unlabeled data (a query) is classified by assigning the label which is nearest to that query point if the distance is smaller than the preset threshold; otherwise, the unlabeled data will be assigned a new label. This procedure is inspired by the methodology of K Nearest Neighbor algorithm [9]. Since the group with the minimum distance is chosen, the process can be considered similar as 1 Nearest Neighbor algorithm (KNN when *k*=1) [10].



(a) **(b)** **(c)** **(d)**

Figure 2. Demonstration of sequential clustering algorithm
(a) Step#1; (b) Step#2; (c) Step#3; (d) Final result

2.3 Sequential Dimension Reduction

To deal with the issues of sparse sample and multicollinearity caused by high dimensionality of the data, the introduced algorithm also features dimension reduction after the procedure of clustering based on the same concept. Since the data has already been clustered, the total number of clusters, comparing to that of the original data, has been significantly reduced. Exploiting such feature, the algorithm will then apply orthogonal and weighted summation within each cluster to perform dimension reduction on the basis of preserving the information contained within the data to the greatest possible extent. The total number of groups being clustered is also considered as the dimension of the new data after reduction.

The number of groups of clustering, which also is the dimension of the resulting new data, can be managed by adjusting the preset threshold in sequential clustering procedure. When the dimensionality is reduced to a certain degree, conventional method of orthogonality (e.g. least squares method) could effectively eliminate the multicollinearity issue across the data while preserving the information of the data. The performance of the model will be better using the processed new data than using the raw data.

Regarding the order of choosing unprocessed data, the algorithm for dimension reduction shares the same concept with clustering algorithm. The order is determined by the importance and the effectiveness of each time-series data, which is calculated by the heuristic function in Equation (10). Within each cluster, the heuristic function is scored on. Within each cluster, starting from the data with highest score and successively taking the data with lower score as independent variable and that of higher score as dependent variable, a linear regression model is fitted and the residuals are computed to form a new time-series data. For every new set of such time-series data of residuals obtained from the regression model, a score of effectiveness of this data will be evaluated again, by using the same heuristic function. If this score is greater than another preset threshold, then the new data will be kept and combined into the dependent variable; otherwise, the data will be discarded. The preset threshold should be a score that is high enough so that the query data can be defined as effective or important. The main purpose of the algorithm is to assess how much information contained in the dependent variable can be explained by the independent variables. Discarding data with low effectiveness is an important component of dimension reduction algorithm. If the heuristic value (IR for multi-factors model) for the residuals of regression is lower than the threshold, it indicates that the information contained in the newly added data has already been explained by existing data while the rest part is mainly noise, thus should be discarded.

After sequentially applying linear regression to all the data, the new time-series data remained, which are essentially the residuals, will be assigned weighted sums using their new heuristic scores as weights. These resulting data will be considered to represent its cluster.

Table 2. Algorithm of sequential dimension reduction

```
Algorithm: Sequential Dimension Reduction
Parameters: $k$ number of groups of sample data, $n_i$ number of data within group $i$,
            threshold $M$
Let $X_{i,j}$ be the data sample, $IR_{i,j}$ be the IR of $X_{i,j}$, $j∈\{1,2,…,n-1\}$
Sort $IR_{i,j}$ within group $i$ that $IR_{i,j} \geq IR_{i,j+1}$, $j∈\{1,2,…,n-1\}$
for $i$ from 1 to $k$ do
     Add first data within group $i$ into the set of
     dependent variables, $DV=\{X_{i,1}\}$
     Choose unprocessed data $X_{i,j}$
     Perform Ordinary Least Squares:
          $DV = \beta_{i,1}X_{i,j+1} + \varepsilon_{i,j+1}$
     Calculate $IR^{\varepsilon}_{j}$ of $\varepsilon_{i,j+1}$
     if $IR^{\varepsilon}_{j} \geq M$ do
         $DV = DV \cup \varepsilon_{i,j}$
     else do
         Discard $\varepsilon_{i,j}$
     end if
     Calculate weighted average of $DV_j$ with a weight of $IR^{\varepsilon}_{j}$
end for
```

2.3.1 Sequential Regression within Clusters

After clustering, the first step is to conduct sequential regression analysis on the factors with in each cluster. Taking the multi-factor model as example, and assuming that there are n factors in group i with the factors denoted as $X_{i,j}, j = \{1,2,\dots,n\}$, the heuristic function, or the information ratio $IR_{i,j}$ for each factor is calculated by Equation(10). In descending order of $IR_{i,j}$, a linear regression model is then fitted to each factor.

Linear regression may be fitted with different methods, and for the sequential regression analysis in this paper, ordinary least squares (OLS) approach will be used to build the regression models for the factors [11]. Ordinary least squares is a straightforward mathematical regression method, which performs regression analysis through minimizing the sum of the squares of the residuals and finding the best matched function. It is employed in the algorithm introduced in this paper for its robustness [12] to avoid overfitting issues. In the procedure of sequential regression analysis, the relationship between dependent and independent variables, which also is the fitted function of ordinary least squares [13], is calculated. Then the residuals will be kept as the information contained within independent variables which has not been explained by the regression model. The formula of OLS for n-variable sample is [12]:

$$y_i = \beta_1 x_{i,1} + \beta_2 x_{i,2} + \cdots + \beta_n x_{i,n} + \varepsilon_i \tag{12}$$

where y_i is the dependent variable, $x_{i,n}$ is the independent variable, β is the regression coefficient, ε_i is the residual.

Since the purpose of the sequential regression analysis is to preserve the information contained within the data to the greatest possible extent, therefore, the proposed algorithm chooses the order of data to perform regression analysis based on the rank of their respective information ratio. Suppose for factors $X_{i,j}, j = \{1,2,\ldots,n\}$, j is the descending rank of the factor, then

$$IR_j \geq IR_{j+1}, j \in \{1,2,\ldots,n-1\} \tag{13}$$

Since the regression analysis is performed in a sequential fashion, according to Equation (12), the first step then becomes

$$X_{i,1} = \beta_{i,1} X_{i,2} + \varepsilon_{i,2} \tag{14}$$

The residual $\varepsilon_{i,2}$ in Equation 14 will be the remaining part of $X_{i,2}$, after eliminating the linear dependency with $X_{i,1}$. To determine if $\varepsilon_{i,2}$ contains enough useful information to be kept as a new data, the information ratio of $\varepsilon_{i,2}$ ($\widehat{IR}_{i,2}$) will be calculated again with Equation (10). An $\widehat{IR}_{i,2}$ larger than the preset threshold indicates $X_{i,2}$ has enough information which is not contained in $X_{i,1}$. Therefore, in this case, $\varepsilon_{i,2}$ will be kept as $\tilde{X}_{i,2}$. On the contrary, an $\widehat{IR}_{i,2}$ smaller than the preset threshold indicates $X_{i,2}$ does not have enough information which is not contained in $X_{i,1}$. In this case, both $X_{i,2}$ and $\varepsilon_{i,2}$ will be discarded.

If the residual is kept, for example, as $\tilde{X}_{i,2}$, it will be added to the regression model as dependent variable for the sequencing steps. The formula of the sequencing steps can be represented as:

$$X_{i,1} + \tilde{X}_{i,j} = \beta_{i,j+1} X_{i,j+1} + \varepsilon_{i,j+1}, j \in \{2,3,\ldots,n-1\} \Leftrightarrow \widehat{IR}_{i,j} \geq M \tag{15}$$

2.3.2 Weighted Sum within each Cluster

After all the data in a given cluster have been processed by regression analysis, the remaining data will be combined using weighted sum. The weights are the new heuristic value (information ratio $\widehat{IR}$ in this case). The new combined data X' will be the only data for this cluster.

$$X'_i = X_{i,1} \cdot IR_{i,1} + \tilde{X}_{i,j} \cdot \widehat{IR}_{i,j} \Leftrightarrow \widehat{IR}_{i,j} \geq M \tag{16}$$

Therefore, the number of dimensions of the resulting set of the data after sequential dimension reduction algorithm will be at most the number of the clusters, which is determined by the previous clustering's result.

3 Experiment

The sequential clustering and dimension reduction algorithm fits with the multi-factor model in quantitative finance very well, as it can effectively eliminate the multicollinearity issue across the dimensions of data while preserving the information of the data. Its effectiveness can be proved by comparing the model's performance between using the raw time-series data of factors as input, versus using the processed data of factors as input. The results between conventional dimension reduction method PCA and the proposed method are also compared and analyzed.

3.1 Model Selection

Multi-factor model, best known as the generalized form of Fama-French three-factor model [14], is an extension of the single-factor capital asset pricing model [15]. It was designed to explain the factors that affect the average excess returns of the stock market [16].

$$\mathrm{r} = \mathrm{R}_f + \beta_3(\mathrm{R}_m - \mathrm{R}_f) + \mathrm{b}_s \cdot SMB + \beta_v \cdot HML + \varepsilon \tag{17}$$

Based on the three-factor model, multi-factor model adds more factors as input. The model computes a set of scores for each stock based on each of these factors, which are represented as $X_i, i = \{1, 2, \ldots, n\}$. By computing the linear weighted sum of these scores, the model assigns a total score for each of the stock, which is used to predict the future excess return of the corresponding stock. Then the portfolio can be constructed by buying stocks with high score, short selling stocks with low score.

$$\mathrm{r} = \mathrm{R}_f + \beta_1 X_1 + \cdots + \beta_n X_n + \varepsilon \tag{18}$$

Multi-factor model is one of the best candidates to test the effectiveness of sequential clustering and dimension reduction algorithm, because (1) it relies on high-dimensional data for different aspects of each stock, including value, growth, profitability, market capitalization, volatility, technical trend, etc.; (2) it is a linear model, therefore it can be used to test if the issue multicollinearity is addressed properly.

3.2 Backtest Parameters

The sample space of the back test is all the stocks in the Chinese A-shares market that satisfied the following conditions:

a) is member of CSI 500 index;
b) has not been labeled as Special treatment stock (ST stock) in the last three months;
c) has not been placed transaction hold throughout the day of adjusting position;
d) has trading volume larger than 1 million CNY.

Parameters of the back test:

a) One month of position holding period, long top 10% of stocks that has highest score, and short bottom 10% of stocks that has lowest score;
b) Using unweighted sum to calculate the scores of stocks;
c) Transaction fee: 0.3%;
d) Benchmark: CSI 500 index (000905.SH);
e) Backtest period: Jan 1, 2010 – Oct 31, 2018

Table 3. List of factors and calculation methods [14][15][16]

Category	Factor	Factor Abbr.	Computation Method
Size	Natural log of total market capitalization	LNMCAP	Natural log of total market capitalization
Value	Price-to-book ratio	PB	Market capitalization / Book value
	Consensus PB ratio	ESTPB	Rolling consensus price-to-book ratio
	Price-to-earnings ratio	PE	Market capitalization / Total equity
	Consensus PE ratio	ESTPE	Rolling consensus price-to-earnings ratio
Technical	3-month reversal	REVERSE60D	3-month accumulative price changed
	6-month reversal	REVERSE120D	6-month accumulative price changed
Growth	Operating Revenue YoY	OperRevYoY	Operating Revenue(t) / Operating Revenue(t-1)
	Net Profit YoY	NetProfYoY	Net Profit (t) / Net Profit (t-1)
	Quarterly ROE YoY	deltaROE	Quarterly ROE (t) / Quarterly ROE (t-1)
	Quarterly ROA YoY	deltaROA	Quarterly ROA (t) / Quarterly ROA (t-1)
Profitability	Quarterly return on equity	ROE	Quarterly net profit/Average total equity
	Quarterly return on assets	ROA	Quarterly net profit/Average total assets
Liquidity	1-month turnover	TURNOVER20D	Average of turnover last month
	3-month turnover	TURNOVER60D	Average of turnover last 3 months
	6-month turnover	TURNOVER120D	Average of turnover last 6 months
Volatility	1-month average true range	ATR20D	1-month average true range
	3-month average true range	ATR60D	3-month average true range
Dividend	Dividend yield	DividendYield	Rolling 12-month dividend yield

3.3 Backtest Data

In order to test the effectiveness of the sequential clustering and dimension reduction algorithm, this paper uses 19 factors that are commonly used in multi-factor model as the raw data. Each of these factors covers daily data for more than 3600 stocks in China A-Shares market for the past decade. The source of the data comes from public data of China A-Shares market. The calculation methods of the factors are shown in Table 3. The conventional way to categorize these factors is also presented as 'factor category' for references. These categories are based on the underlying financial logic of each factor, and are widely used in finance industry. However, these categories might not be the best representation of the correlation between these factors in a technical way. The differences between the result of conventional categories and that of new clusters will be compared.

3.4 Evaluation

According to the algorithm, the effectiveness and stability of the preprocessing data, which is represented by information ratio (IR), needs to be evaluated first. The result is shown in figure 3. All the factors' information ratio is positive within the backtest period, indicating that these factors have good prediction precision and stability. After getting the information ratio of each factor, sequential clustering algorithm will be performed in a descending order based on the information ratio.

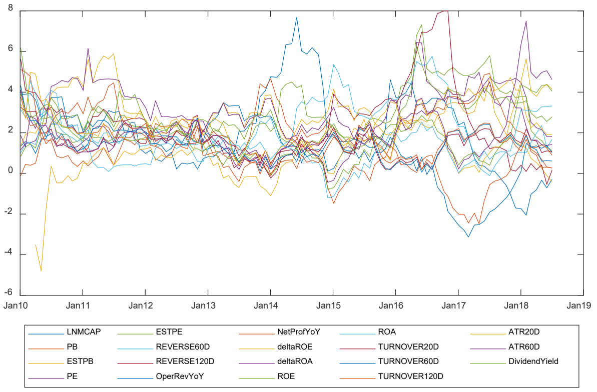


Figure 3. Time-series of raw data's information ratio

Next step, the cross-sectional 12-month rolling average of the Spearman correlation coefficient of each factor will be calculated. The resulting correlation matrix will be used by the sequential clustering algorithm to perform clustering. The result of the correlation matrix and clustering are shown in Table 4. According to the correlation matrix, most of the factors have at least one other high-correlated factor. From the result, the proposed algorithm accurately clusters all the factors based on their pairwise correlations. This result is slightly different from the conventional category, but turns out to be more sensible. For example, the Operating Revenue YoY does not fall into the cluster of other growth factors, as it is the only one that is computed by operating revenue instead of net profit; the other example is the PB and PE, which both are considered as value factors by convention, do not fall in to same cluster by the proposed algorithm. This is because they are computed by book value per-share versus earnings per-share respectively, which have low correlation.

Based on the cluster shown in Figure 4, the sequential dimension reduction algorithm combines and reduces the original 19 factors into 11 new factors. The correlation matrix of the 11 new factors is shown in Table 5. It's obvious that the cross-sectional pairwise correlations between the new factors are significantly lower than that of the original 19 factors. The time-serial average pairwise correlations of the 11 new factors are also substantially lower than that of the 19 original factors.

Table 4. Correlation matrix of raw data and results of Sequential Clustering

Cluster	Catagory	Factors	LNMCAP	PB	ESTPB	PE	ESTPE	REVERSE60D	REVERSE120D	OperRevYoY	NetProfYoY	deltaROE	deltaROA	ROE	ROA	TURNOVER20D	TURNOVER60D	TURNOVER120D	ATR20D	ATR60D	DividendYield
11	Size	LNMCAP	1	-0.01	-0.03	0.04	0	0	-0.03	-0.14	0.2	0.1	0.07	0.05	0.06	0.14	0.16	0.14	-0.02	-0.03	0.09
10	Value	PB	-0.01	1	0.6	0.37	0.29	0.12	0.3	-0.11	-0.04	-0.14	-0.17	-0.26	-0.29	0.3	0.34	0.36	0.38	0.46	0.2
10		ESTPB	-0.03	0.6	1	0.33	0.42	0.1	0.22	-0.11	-0.03	-0.11	-0.15	-0.2	-0.22	0.17	0.18	0.2	0.23	0.27	0.15
5		PE	0.04	0.37	0.33	1	0.58	0.1	0.16	-0.03	0.05	0.09	0.06	0.25	0.2	0.13	0.12	0.12	0.2	0.22	0.47
7		ESTPE	0	0.29	0.42	0.58	1	0.13	0.21	-0.01	-0.01	0	-0.02	0.09	0.04	0.07	0.08	0.08	0.11	0.13	0.23
6	Technical	REVERSE60D	0	0.12	0.1	0.1	0.13	1	0.65	-0.07	-0.11	-0.17	-0.18	-0.09	-0.09	-0.07	-0.09	-0.09	-0.1	-0.08	0.02
6		REVERSE120D	-0.03	0.3	0.22	0.16	0.21	0.65	1	-0.13	-0.19	-0.25	-0.28	-0.19	-0.23	0.05	0.09	0.09	0.09	0.15	0.02
9	Growth	OperRevYoY	-0.14	-0.11	-0.11	-0.03	-0.01	-0.07	-0.13	1	0.25	0.31	0.34	0.12	0.15	-0.01	-0.01	-0.01	-0.05	-0.07	-0.09
1		NetProfYoY	0.2	-0.04	-0.03	0.05	-0.01	-0.11	-0.19	0.25	1	0.65	0.63	0.31	0.29	0.03	0.04	0.03	-0.03	-0.04	-0.04
1		deltaROE	0.1	-0.14	-0.11	0.09	0	-0.17	-0.25	0.31	0.65	1	0.88	0.5	0.45	-0.03	-0.03	-0.04	-0.06	-0.09	-0.03
1		deltaROA	0.07	-0.17	-0.15	0.06	-0.02	-0.18	-0.28	0.34	0.63	0.88	1	0.51	0.52	-0.02	-0.01	-0.02	-0.07	-0.1	-0.05
2	Profitability	ROE	0.05	-0.26	-0.2	0.25	0.09	-0.09	-0.19	0.12	0.31	0.5	0.51	1	0.85	-0.11	-0.14	-0.15	-0.05	-0.09	0.2
2		ROA	0.06	-0.29	-0.22	0.2	0.04	-0.09	-0.23	0.15	0.29	0.45	0.52	0.85	1	-0.14	-0.17	-0.19	-0.05	-0.1	0.24
8	Liquidity	TURNOVER20D	0.14	0.3	0.17	0.13	0.07	-0.07	0.05	-0.01	0.03	-0.03	-0.02	-0.11	-0.14	1	0.91	0.82	0.48	0.49	0.11
8		TURNOVER60D	0.16	0.34	0.18	0.12	0.08	-0.09	0.09	-0.01	0.04	-0.03	-0.01	-0.14	-0.17	0.91	1	0.93	0.51	0.55	0.08
8		TURNOVER120D	0.14	0.36	0.2	0.12	0.08	-0.09	0.09	-0.01	0.03	-0.04	-0.02	-0.15	-0.19	0.82	0.93	1	0.46	0.55	0.06
4	Volatility	ATR20D	-0.02	0.38	0.23	0.2	0.11	-0.1	0.09	-0.05	-0.03	-0.06	-0.07	-0.05	-0.05	0.48	0.51	0.46	1	0.85	0.18
4		ATR60D	-0.03	0.46	0.27	0.22	0.13	-0.08	0.15	-0.07	-0.04	-0.09	-0.1	-0.09	-0.1	0.49	0.55	0.55	0.85	1	0.18
3	Dividend	DividendYield	0.09	0.2	0.15	0.47	0.23	0.02	0.02	-0.09	-0.04	-0.03	-0.05	0.2	0.24	0.11	0.08	0.06	0.18	0.18	1

Table 5. Correlation matrix of new data after sequential clustering and dimension reduction algorithm

New Factors	#1	#2	#3	#4	#5	#6	#7	#8	#9	#10	#11
#1	1.00	0.27	-0.06	-0.09	0.16	-0.09	0.09	0.04	0.22	-0.07	-0.07
#2	0.27	1.00	0.11	-0.07	0.26	-0.07	0.21	-0.01	0.13	-0.08	0.14
#3	-0.06	0.11	1.00	0.13	0.26	-0.01	0.26	0.03	-0.02	0.20	0.09
#4	-0.09	-0.07	0.13	1.00	0.05	0.29	0.03	-0.08	-0.04	0.17	0.01
#5	0.16	0.26	0.26	0.05	1.00	0.01	0.44	0.04	0.14	0.16	0.06
#6	-0.09	-0.07	-0.01	0.29	0.01	1.00	0.05	0.14	-0.11	0.07	0.02
#7	0.09	0.21	0.26	0.03	0.44	0.05	1.00	0.07	0.04	0.12	-0.04
#8	0.04	-0.01	0.03	-0.08	0.04	0.14	0.07	1.00	0.02	-0.05	0.01
#9	0.22	0.13	-0.02	-0.04	0.14	-0.11	0.04	0.02	1.00	-0.04	0.20
#10	-0.07	-0.08	0.20	0.17	0.16	0.07	0.12	-0.05	-0.04	1.00	0.01
#11	-0.07	0.14	0.09	0.01	0.06	0.02	-0.04	0.01	0.20	0.01	1.00

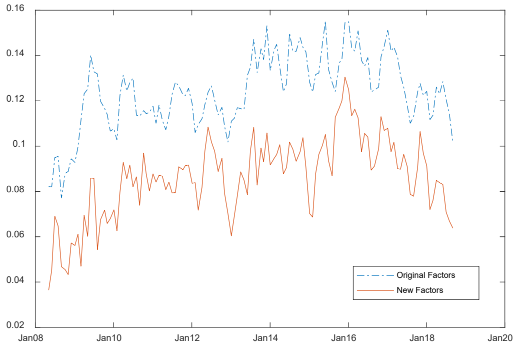


Figure 4. Comparison of average correlation coefficient

The rolling average information ratio of the new 11 factors is also increased comparing to that of the original 19 factors, shown in Figure 5. This proves that the proposed algorithm not only effectively eliminate the multicollinearity issue across the dataset, the useful information of the data is also preserved. Moreover, because a large proportion of the noise and repeated information is discarded during the process, the resulting data have larger information ratio, in other words, more effective when combined together.

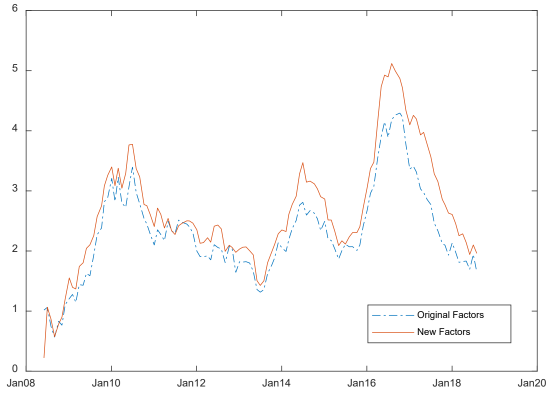


Figure 5. Comparison of average Information Ratio

Lastly, also most importantly, the best way to evaluate the effectiveness of the proposed algorithm is to check if using the new factors improves the performance of multi-factors model.

With the same setup and backtest parameters, the multi-factor model with new factors as input will achieve better performance, comparing to model with unprocessed factors. Moreover, it can also be shown that model with new factors outperforms model with PCA factors, which are processed by one of the most commonly used dimension reduction algorithm nowadays[17][18][19], by a large margin. The comparison of net asset value of these models are shown in Figure 6 and Figure 7.

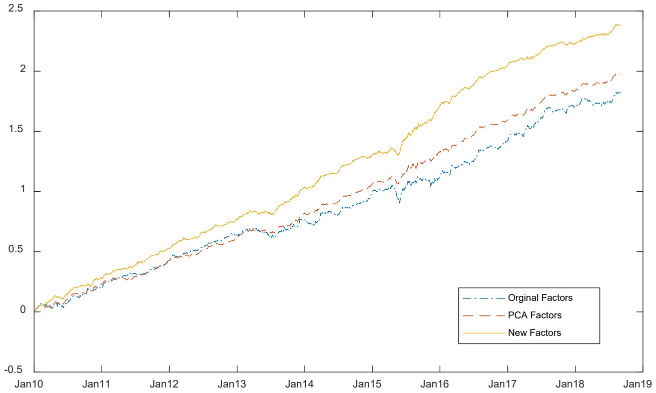


Figure 6. Comparison of net asset value of multi-factor models using different data

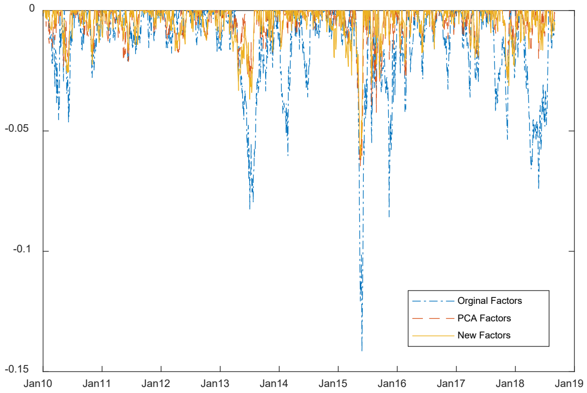


Figure 7. Comparison of drawdown of Multi-Factor models using different data

Table 6. Comparison of back test data

	Annual Excess Return	**Volatility**	**Sharpe Ratio**	**Maximum Drawdown**	**Return / Drawdown**
Original Factors	22%	0.085	2.531	-14%	1.525
PCA Factors	23%	0.061	3.81	-6%	4.07
New Factors	**28%**	**0.061**	**4.58**	**-7%**	**4.157**

The backtest result of new factors, comparing to that of original factors and PCA factors, has a better result, in other words, improves the performance of the multi-factors model. The model using new factors achieves better annually excess return, risk-adjusted return (Sharpe ratio) [20][21], and lower volatility [22][23]. These can be quantified and evaluated by classic portfolio management evaluation methods, shown in Table 6. Especially, the model using new factors has a Sharpe ratio [22][23] of 4.58, which is a substantial improvement comparing to 2.53 (original factors) and 3.81 (PCA factors), indicating a better excess return given same amount of risk exposure.

4 Conclusion

This paper introduces a new algorithm for clustering and dimension reduction, which is designed to handle high dimensional and intercorrelated time-series data. The proposed algorithm sequentially performs clustering and ordinary least squared regression analysis based on a data-related heuristic function. It combines some classic and robust algorithm, such as K-Nearest Neighbor algorithm, Greedy algorithm, and ordinary least squares regression, with a heuristic function that measures the effectiveness or importance of the data, to form a new robust, unsupervised algorithm.

The effectiveness of this algorithm is examined by using real-world financial time-series data and multi-factors investment model. The 19 original factors can be clustered and combined into 11 new factors correctly by the proposed algorithm. The new factors have a lower (nearly zero) average pair-wise correlation coefficient, indicating that the proposed algorithm can eliminate multicollinearity within the dataset; the new factors also have a better information ratio, which is a quantitative measurement of the effectiveness of the factor in finance, indicating that the proposed algorithm can better filter out the noise and repeated information within the data set. Through a rigorous backtest using the multi-factor model, this paper shows that the proposed algorithm can achieve a better output comparing to PCA, which is a commonly used dimension reduction algorithm. Therefore, the conclusion is the sequential clustering and dimension reduction algorithm can efficiently deal with high dimensional, correlated, noisy data, by converting them into a new set of barely correlated, less noisy and lower dimensional data.

In the future, we are looking forward to improving the sequential clustering and dimension reduction algorithm, by further optimizing its runtime complexity and its robustness. Moreover, we will actively engage to apply this algorithm to more industries, such as weather forecast, medical treatment, and data science.

References

[1] Bellman, Richard E. (1957). *Dynamic programming*. Princeton University Press.

[2] Goldberger, Arthur S. (1991). *A Course in Econometrics*. Harvard University Press. pp. 248–250.

[3] Ruppert, David. (2010). *Statistics and Data Analysis for Financial Engineering*, Springer Science & Business Media.

[4] Leys, C.; et al. (2013). Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. *Journal of Experimental Social Psychology*. 49 (4): 764–766.

[5] Glantz, Stanton A; Slinker, Bryan K; Neilands, Torsten B. (2016), *Primer of Applied Regression & Analysis of Variance* (Third ed.), McGraw Hill.

[6] Cover, Thomas M.; Hart, Peter E. (1967). Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*. 13 (1): 21–27.

[7] Black, Paul E. (2005). Greedy algorithm. *Dictionary of Algorithms and Data Structures*. U.S. National Institute of Standards and Technology (NIST).

[8] Bacon, Carl R. (2012). *Practical Risk-Adjusted Performance Measurement.* John Wiley & Sons.

[9] Jaskowiak, Pablo A.; Campello, Ricardo J. G. B. (2011) Comparing Correlation Coefficients as Dissimilarity Measures for Cancer Classification in Gene Expression Data. *Brazilian Symposium on Bioinformatics*, 1–8. CiteSeerX 10.1.1.208.993.

[10] Everitt, Brian S.; Landau, Sabine; Leese, Morven; and Stahl, Daniel. (2011). Miscellaneous Clustering Methods. *Cluster Analysis,* 5th Edition, John Wiley & Sons, Ltd., Chichester, UK.

[11] Goldberger, Arthur S. (1964). Classical Linear Regression. *Econometric Theory*. New York: John Wiley & Sons. pp. 158.

[12] Hayashi, Fumio. (2000). *Econometics*. Princeton University Press. p. 15.

[13] Williams, M. N; Grajales, C. A. G; Kurkiewicz, D. (2013). Assumptions of multiple regression: Correcting two misconceptions. *Practical Assessment, Research & Evaluation*. 18 (11).

[14] Fama, E. F.; French, K. R. (1993). Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics*. 33: 3–56. CiteSeerX 10.1.1.139.5892.

[15] Eugene F. Fama & Kenneth R. French. (2004). The Capital Asset Pricing Model: Theory and Evidence, *Journal of Economic Perspectives*, American Economic Association, vol. 18(3), pages 25-46, Summer.

[16] Harvey, Campbell R., Yan Liu, and Heqing Zhu. (2015) ... And the cross-section of expected returns. *Review of Financial Studies,* hhv059.

[17] Jolliffe I.T. *Principal Component Analysis*, Series: Springer Series in Statistics, 2nd ed., Springer, NY, 2002, XXIX, 487 p. 28 illus.

[18] Abdi. H. & Williams, L.J. (2010). Principal component analysis. *Wiley Interdisciplinary Reviews: Computational Statistics*. 2 (4): 433–459. arXiv:1108.4372.

[19] Pueyo, Laurent. (2016). Detection and Characterization of Exoplanets using Projections on Karhunen Loeve Eigenimages: Forward Modeling. *The Astrophysical Journal*. 824 (2): 117. arXiv:1604.06097.

[20] Sharpe, W. F. (1966). Mutual Fund Performance. *Journal of Business*. 39 (S1): 119–138.

[21] Scholz, Hendrik. (2007). Refinements to the Sharpe ratio: Comparing alternatives for bear markets. *Journal of Asset Management*. 7 (5): 347–357.

[22] Loth, Richard. (2019) 5 Ways To Measure Mutual Fund Risk, *Investopedia*

[23] Wilmott, Paul. (2007). *Paul Wilmott introduces Quantitative Finance* (Second ed.). Wiley. pp. 429–432.