时间序列数据逐步聚类降维法

郑坦然
（暨南大学产业经济研究院，广东 广州 邮编：510632）

摘　要　为了达到对复杂的多维度时间序列数据进行简化和提取有效信息的目的，本文介绍了一种新的对多维度的时间序列数据进行聚类及降维的方法，处理数据科学及其相关领域的时间序列数据的分类、剔除多重共线性及降维问题。该方法首先通过借鉴 K 临近算法（K-Nearest Neighbor）和贪心算法（Greedy Algorithm）的思想，基于数据横截面有效性为启发式函数（Heuristic Function），按启发式函数大小为顺序逐步对时间序列数据进行聚类；得到已聚类分组的数据后，在每一组内同样通过数据有效性为顺序进行逐步普通最小二乘法（Ordinary Least Squares, OLS）回归分析，逐步回归得到的残差作为新数据，根据新数据回测的有效性进行加权平均，在每个组别内合称为一个新的数据，最终达到降维的效果。通过利用金融量化投资领域的多因子模型（Multi-Factor Model）和过去十年的 A 股数据进行回测及分析，证明采用逐步聚类降维法能够有效的处理高维度、高共线性时间序列因子数据，帮助后续应用模型更好的剔除时间序列数据间的多重共线性关系，同时最大程度上保留甚至提高时间序列数据的有效性。

关键词　无监督聚类；降维；时间序列分析

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 with multicollinearity. 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.

Key Words: Supervised clustering; dimension reduction; time-series data analysis

1 引言

时间序列数据是数据科学领域最重要的数据类别之一，广泛应用于经济金融、机器学习、天气预报、地震预测、医学诊断以及绝大多数涉及到时间数据测量的领域。随着在各个领域可使用数据成几何倍数的增多，很多问题面临着需要处理数量极大、维度极高的时间序列数据。因此 Richard E. Bellman 提出了“维数灾难”（Curse of Dimensionality）[1]，即当数据维数提高时，数据空间的体积提高太快，因此在可用的数据中会出现数据样本稀疏、距离计算困难的情况，进而为获得在统计学上正确且可靠的结果带来了巨大的难度。另外，高维度的数据容易出现多重共线性问题（Multicollinearity）[2]，干扰平时寻求问题时寻找到的特征。因此，给数据降维成为了现在数据科学领域的重要问题。

降维主要是指在数量和维度庞大的原始数据中，在服从一定限定条件的情况下降低随机变量个数，得到一组两两相关性相对较低的主变量的过程。降维进一步又可以细分为变量选择和特征提取两个大类别。而本文介绍的逐步分类降维法，利用的是变量选择的思想，即从原有的数据中找出主要的含有信息最多的数据。

在以往的研究中，时间序列的分类及降维方法主要有基于时间序列距离度量（Dynamic Time Warping）或基于主成分分析（PCA）的衍生出的聚合分类或降维法。这些方法被广泛的应用在信号分析及处理、语音识别和推荐系统等领域，但在处理一些信噪比较低、平稳性较差的数据时，仍存在分类精确度较低、降维导致数据原始信息受损等情况。而且这些方法需要较多的正确样本，否则分类和降维结果也会受到较大影响。

本文介绍的逐步聚类降维算法，主要利用 K 最近邻算法和贪心算法的思想，对所有原始数据进行逐步的无监督聚类分析。逐步聚类的顺序是根据不同数据的特征和应用场景的不同，选用不同的启发式函数对数据含有信息的完整性、独立性和有效性进行量化，得分越高的数据顺序越靠前。在得到聚类标签后，算法将根据标签在类别内对数据进行进一步的启发式函数计算、逐步普通最小二乘法回归分析和加权求和，在每个类别内合成唯一的一组数据，最终达到降维的效果。

相比于传统的时间序列聚类与降维方法，本方法不需要大量的有标签数据，只需要事先对分类的阈值参数进行估计和通过数据应用场景选择合适的启发式函数。算法直接简单，鲁棒性强，同时能尽可能保留原始数据的有效信息。在噪音较大的金融领域也能保持很高的分类准确度。本文将本算法应用在金融工程领域的量化多因子模型中，对大量的未分类股票因子信号进行分类，并在类别内降维。从回测中可以证实，相比于传统的分类降维方法，将使用该方法处理后的因子用于多因子选股模型，模型的选股效果更好，策略表现更优秀。

2 逐步聚类降维算法

逐步聚类降维算法主要分为以下几个步骤：

I. 数据预处理；

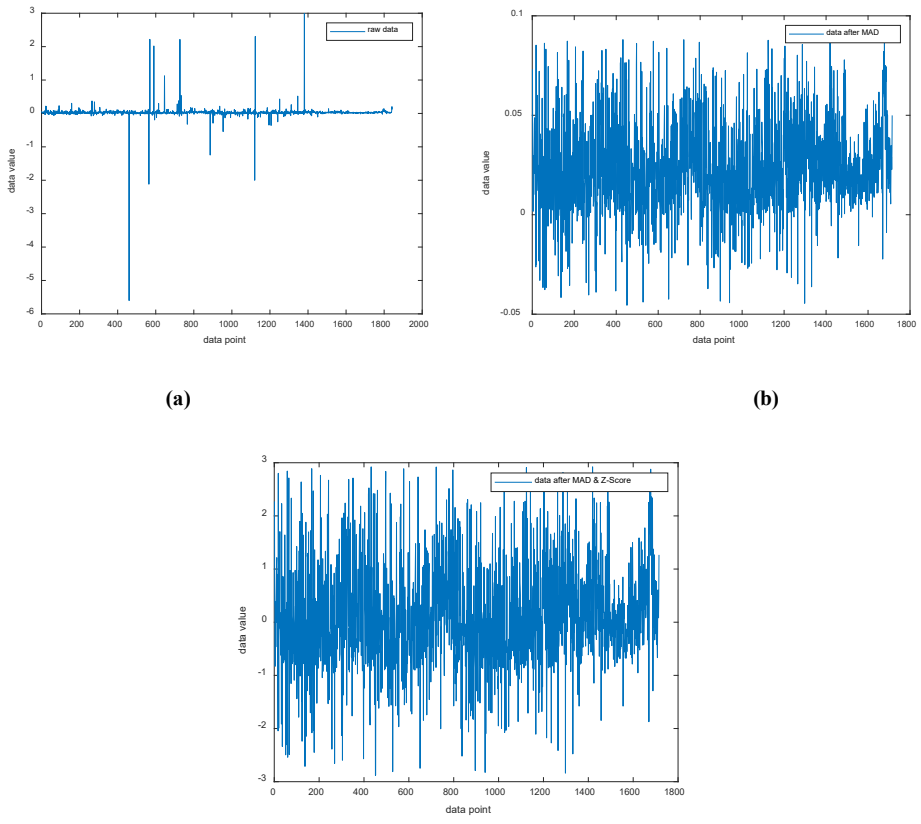
II. 逐步聚类；

III. 逐步降维；

2.1 数据预处理

在进行逐步聚类降维之前，首先要对数据进行预处理。数据的质量，直接决定了模型的预测和泛化能力的好坏。数据预处理的目的主要是确保数据的准确性、完整性、一致性、时效性、可信性和解释性得到保障，让模型的输入是标准的、干净的、连续的数据。

数据预处理主要有两个步骤：去极值和标准化。在本算法中去极值采用的是绝对中位差算法，标准化则用 Z-Score 完成。数据经过处理后，高维度的数据也将化为同一量纲，不同数据间可以有比较性。数据预处理的效果对比见图（1）。



(a) **(b)**

(c)

图 1. 数据预处理效果对比

（a）原始数据；（b）MAD 算法处理后的数据；（c）MAD 及 Z-Score 算法处理后的数据

Figure 1. Results demonstration of data preprocess

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

2.1.1 去极值

绝对中位差（Median Absolute Deviation, MAD）是对单变量数值型数据的样本偏差的一种鲁棒性测量。同时也可以表示由样本的 MAD 估计得出的总体参数。对于单变量数据集$X_1, X_2, \ldots, X_n$，MAD 定义为数据点到中位数的绝对偏差的中位数：

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

其中$\tilde{X}$为X_i的中位数。通过公式（1），先计算出数据与它们的中位数之间的残差（偏差），MAD 就是这些偏差的绝对值的中位数。

绝对中位差是一种统计离差的鲁棒统计量，比标准差更能适应数据集中的异常值。相比之下，标准差使用的是数据到均值的距离平方，所以大的偏差权重更大，异常值对结果也会产生重要影响。而对于 MAD，少量的异常值不会影响最终的结果，因此 MAD 法是一个比样本方差或者标准差更鲁棒的度量，因此在本算法选择 MAD 法对数据极值进行处理。

在将 MAD 当作标准差估计的一种一致估计量之前，需要将 MAD 转换为使用标准差σ：

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

其中κ是一个取决于数据分布类型的比例常数[3]。假设数据服从标准正态分布，则可以利用标准正态分布的逆累计分布函数（inverse of the cumulative distribution function），可以算得κ的取值。其中需要确保±MAD包含标准正态累积分布函数的 50%（从 1/4 到 3/4 的范围值）[4]。

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

所以，

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

而

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

将$\frac{3}{4}$的取值带入得到：

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

最后得到

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

得到κ后，将距离样本数据的中位数$\tilde{X}$超过$3\cdot\kappa\cdot\mathrm{MAD}$（即 3 倍标准差）外的数据定义为异常值，并将其数值修正为$\kappa\cdot\mathrm{MAD}$:

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

2.1.2 标准化

数据中不同特征的量纲可能不一致，数值间的差别可能很大，不进行处理可能会影响到数据分析的结果，因此，需要对数据按照一定比例进行缩放，使之落在一个特定的区域，转化为统一量纲，便于进行综合分析。

因为在对数据的异常值采用 MAD 法进行处理后，可以确保数据样本的均值μ和标准差σ不受极端值影响，因此接下来可以采用 Z-Score 算法对每组数据进行横截面的标准化处理。选择使用 Z-Score 算法是因为降维前需要对大量的不同的数据源进行比较，而 Z-Score 标准化能够将不同量级的数据转化为统一量度的 Z-Score 分值，这样数据之间才有可比性。同时由于数据的均值和标准差是可知的，Z-Score 分值也是可以计算得到的。

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

得到每组数据的 Z-Score 分值后，就可以用 Z-Score 代替原始数据取值，进而对对各组数据进行进一步的分析[5]。

2.2 逐步聚类

逐步聚类算法是以 K 最近邻（K-Nearest Neighbor, KNN）算法的思想为基础，在其中加入了根据启发式函数为顺序、逐步加入数据样本进行聚类分析的聚类算法。K 最近邻算法是一种常用于分类的算法，是有成熟理论支撑的、较为简单的经典机器学习算法之一，最早于 1967 年由 Cover T 和 Hart P 提出。但 K 最近邻算法缺点也十分明显，它对当前待分类样本的分类，需要大量已知分类的样本的支持[6]，而并不是所有数据都在事前有已知的分类样本。当数据维度过高，或者数据是通过机器学习或深度学习挖掘出来的时候，数据本身的意义并不是十分的明确，K 最近邻算法就不能发挥出它的优势；又例如在金融量化投资领域的因子，会出现因子间的相关性随时间的变化而不断变化的情况，因此之前的分类样本就不能稳定地处理因子间的共线性问题，因此本文中的算法在对 K 最近邻算法进行改进，加入贪心算法（Greedy algorithm）[7]中启发式函数（heuristic function）排序的方法，对数据进行分类。

逐步聚类算法的基本思路是：对待分类样本过去一段时间内的数据进行横截面的相关性分析，得到数据样本的平均 Spearman 相关系数；之后，用相关系数作为数据间的距离，依次为依据在每个横截面上对所有数据样本进行聚类。待分类的数据样本将被逐个地判断是否属于已知类别，在特征空间中找到与之相关

性最高的一个样本，如样本大于事前预定的阈值，则被分为同一类，反之归为新的一类。具体算法如表（1）所示。

表 1. 逐步聚类算法
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 labeled data sample $<h_0,h_1,h_2,\ldots,h_n>$, in which at least one of samples h_0 is a new group
Get the correlation coefficient matrix of n data samples for the past alpha period, let $r_{(i,j)}$ be correlation coefficient of data i and data j
for t from 1 to T **do**
    for i from 1 to n **do**
        Find data sample h_j that has the highest correlation coefficient with x_i in H
        if $r_{(i,j)} \geq M$ **do**
            Let x_i and h_j be a new group
        else do
            Let x_i be a new group
        end if
    end for
end for

2.2.1 逐步聚类顺序的选择——启发式函数

本算法在确认逐步聚类顺序时根据数据类型的不同，采用不同的启发式函数(heuristic function)，对数据的重要性或者有效性进行打分，从打分最高的数据开始分类。例如在量化多因子模型中，因子数据用于预测下一期的股票收益，因此回测时可以采用因子数据的信息比率作为启发式函数来决定逐步聚类的顺序。信息比率（Information Ratio, 以下简称 *IR*）是衡量因子数据预测准确度及稳定性的指标，启发式函数计算方法如下[8]：

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

其中 *IC* 为信息系数（Correlation Coefficient），*n* 为计算 *IC* 均值的周期，σ_{IC}为 *IC* 的标准差。而 $IC(f,t)$表示所选股票的因子值与股票下期收益率的截面相关系数，$IC(f,t)$的计算方法是计算全部股票在调仓周期期初排名和调仓周期期末收益排名的线性相关度，即斯皮尔曼等级相关系数（Spearman's rank correlation coefficient）：

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

其中IC_t为因子在 *t* 期的 *IC* 值，f_{t+1}为因子对 *t*+1 期股票收益率的预测值（或向量），ret_{t+1}为 *t*+1 期股票实际的收益率（或向量）。

当某一因子数据 *IR* 的值越高时，说明该因子在过去 *n* 期的预测能力越高、预测胜率越稳定，因此在金融多因子模型中，数据的有效性最强。所以在多因子模型中使用逐步聚类降维法时，因子数据 *IR* 越高、越先进行聚类。而如果要将逐步聚类降维法应用到其它领域和作用的数据时，需要对数据进行回测打分，重要性或有效性越高的因子，在聚类顺序中优先级越高。

2.2.2 聚类阈值的选取

在聚类的过程中，本算法需要预先设定阈值 M 来判断待分类数据和最近邻数据是否为同一类别。对于不同的数据和不同的应用场景，阈值 M 可以有不同的选择标准。对于多因子模型来说，由于相关性的取值为$r \in [-1,1]$， 因子相关性大于 0.4-0.6 可以被认为是高相关性因子。在之后的回测结果中，阈值 M 取值为 0.6。

2.2.3 分类依据

随着新的数据的逐步加入，算法需要判断出新数据是否应加入已有类别或独立分为一个新类别。这里借鉴了 K 最近邻聚类算法的思想[9]，通过将数据间的相关系数作为距离，算出待分类数据（current data）与全部已分类数据的距离，找出所有已分类数据中距离待分类数据最近的那一组数据，若其相关系数小于阈值，则将待分类数据分到与之距离最近的数据中的这一组别中；否则将新数据分至新的类别。由于选择的是最近的一组数据，因此这相当于一个改进的 1 最近邻算法（即 k=1 的 K 最近邻算法）[10]。

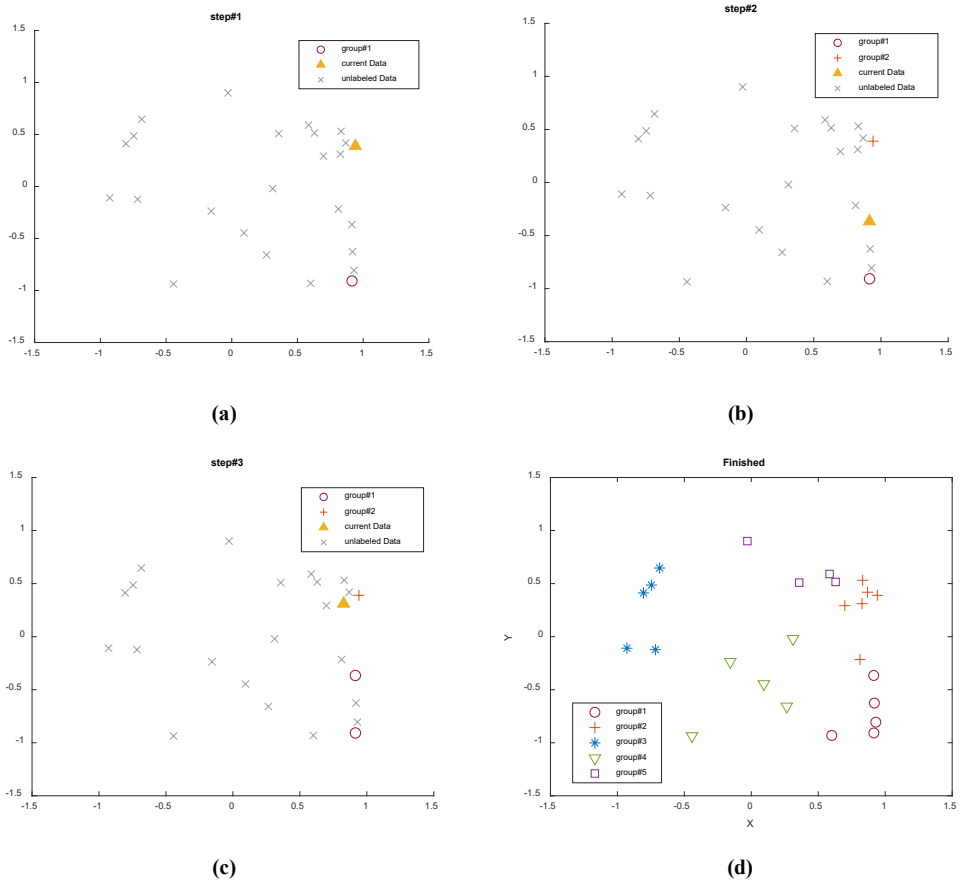


图 2. 逐步聚类算法示意图

（a）第一步（默认排序最高的数据为第一组）；（b）第二步；（c）第三步；（d）最终分类结果

Figure 2. Demonstration of sequential clustering algorithm

(a) Step#1; (b) Step#2; (c) Step#3; (d) Final result

2.3 逐步降维

为了解决数据维度过高导致样本稀疏等维数灾难问题，以及数据多重共线性问题，本算法将时间序列数据进行聚类之后，还需要以此为依据，对数据进行降维。由于数据已经根据其之间的相关性进行聚类分类，类别总数及每个类别内的数据数量都减少了很多。因此该算法利用聚类后的特性，将属于同一个类别内的因子进行正交及加权，在最大程度保留各数据所包含的信息的基础上，降低数据维度。而降维后的数据维度，即为聚类的类别数量。

通过对逐步聚类算法中的分类阈值进行调整，可以调整聚类的类别数量，从而调整降维后的数据维度。当维度降低到一定水平后，传统的正交方法（例如普通最小二乘法）便能有效的排除各维度数据间的线性相关性，达到最优化模型输入数据的效果。

降维的算法与聚类算法相似，在每一个类别中，首先对各时间序列数据的重要性或者有效性，即公式（10）的启发式函数，进行打分，从打分最高的数据开始，逐步将打分更低的数据作为自变量、将打分高的数据作为因变量，做线性回归并求得残差，并将残差作为新的时间序列数据。每做一次回归并得到一个新的时间序列数据，再对新数据的有效性进行打分。若新的时间序列数据有效性大于一定阈值，则将新数据保留并一同纳入因变量中；否则舍弃该数据。该方法旨在判断自变量数据中的信息有多少能被因变量数据解释。舍弃有效性不足的数据是降维算法中非常重要的一部分，因为如果回归后的残差的 *IR* 不足以到达阈值，说明新加入的数据中大部分信息已经被已有数据解释，剩下的信息大部分为噪音，所以应该被排除。

当依次对组内所有数据进行回归过后，保留下来的新的时间序列数据，即残差，将根据其各自的打分进行加权求和。加权求和后的新数据即为代表该分类的数据。

表 2. 逐步降维算法
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 $X_{i,j}$, $j \in \{1,2,\cdots,n-1\}$
Sort $IR_{i,j}$ within group i that $IR_{i,j} \geq IR_{i,j+1}$, $j \in \{1,2,\cdots,n-1\}$
for i from 1 to k **do**
    Add first data within group i into the set of dependent variables, $DV=\{X_{i,1}\}$
    for j from 2 to n_i **do**
        Perform Ordinary Least Squares: $DV = \beta_{i,1}X_{i,j+1} + \varepsilon_{i,j+1}$
        Calculate IR^{ε}_{j} of $\varepsilon_{i,j+1}$
        if $IR^{\varepsilon}_{j} \geq M$ **do**
            $DV = DV \cup \varepsilon_{i,j}$
        else do
            Abandon $\varepsilon_{i,j}$
        end if
    end for
    Calculate weighted average of DV_j with a weight of IR^{ε}_{j}
end for

2.3.1 类别内逐步回归分析

在分组之后首先要对组内的因子进行逐步回归分析。以多因子模型为例，假设聚类后第i组内共有n个因子，每个因子记为$X_{i,j}, j = \{1,2,\dots,n\}$，首先先求得每个因子的启发式函数（公式（10））， 即信息比率$IR_{i,j}$，并根据$IR_{i,j}$从高到低的顺序对因子进行线性回归分析。

线性回归有许多模型，逐步回归算法中利用普通最小二乘法（Ordinary Least Squares）的模型，对因子回归并取残差[11]。普通最小二乘法是一种直接简便的数学优化方法，该方法对线性方程组进行回归分析，通过最小化误差的平方和寻找数据的最佳函数匹配。本算法选取普通最小二乘法是因为其鲁棒性强[12]，避免过拟（overfitting）。逐步回归中利用最小二乘法这个特性，寻找自变量与因变量的线性关系，即最小二乘法的拟合函数[13]，然后将回归的残差作为自变量数据排除掉与因变量数据的线性关系后残留的信息数据。对于有 n 组样本的数据而言，普通最小二乘法的公式为[12]：

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

其中y_i为因变量，x_{in}为自变量，β为回归系数，ε_i为回归残差。

由于逐步回归旨在最大程度上保留因子所包含的信息，因此选择 IR 最大的因子作为第一个因变量，再将其它因子逐个加入模型中。因此假设对于因子$X_{i,j}, j = \{1,2,\dots,n\}$，$j$ 是各个因子对应的信息比率在所有 n 个因子信息比率数量中递减排序的序数，即：

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

逐步回归算法中作为自变量的因子是逐个加入的，因此根据公式（12），降维算法的第一步变形为：

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

回归的残差$\varepsilon_{i,2}$即为$X_{i,2}$排除掉与$X_{i,1}$的线性关系后的新数据，此时再对$\varepsilon_{i,2}$计算信息比率（公式（10）），记为$\widehat{IR}_{i,2}$，若信息比率超过阈值，说明$X_{i,2}$含有较多$X_{i,1}$因子中未包含的信息，则保留该新数据，记为$\tilde{X}_{i,2}$，在之后的回归中，$\tilde{X}_{i,2}$将作为因变量加入到回归模型中；反之，若新数据信息比率未超过阈值，说明$X_{i,2}$与$X_{i,1}$相关性较强，$X_{i,2}$中的信息大部分能被$X_{i,1}$因子中的信息所解释，因此新数据则作为残差被忽略，不在之后的数据加权中加入。

因此若选定阈值为 M，后续的回归公式为：

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

2.3.2 类别内打分加权

当第i组内所有的因子都经过逐步回归计算后，便对保留下来的因子$\tilde{X}$进行加权求和，权重为因子对应的新的信息比率，即$\widehat{IR}$。新因子X'可以表示为：

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

新因子X'便作为第i组降维后的因子。因此可以得知，对于一共n个时间序列数据，若数据被分为I组，逐步降维法可以将原本n维的时间序列数据降维至I维，其中$I \leq n$。而 I 的大小取决于前序 n 个时间序列数据聚类的结果。

3 算法效果验证

对时间序列数据的聚类和降维的最终目的都是为了，排除数据间的相关性干扰，更好的提取到所有时间序列数据的信息，因此非常适合应用在在金融领域的多因子模型中。通过将同一组因子（原始时间序列数据）直接带入多因子模型，统计模型在回测时间段内的收益表现，对比将因子经过逐步聚类及降维算法处理过后再带入多因子模型的表现，可以验证该算法是否使多因子模型更好的从高维度、存在多重共线性的时间序列因子中提取有效信息，提升多因子模型的表现。

3.1 验证模型：多因子模型

多因子模型（Multi-Factor model）是指根据法马-佛伦奇三因子模型（Fama-French three-factor model）[14]拓展而来的资本资产定价模型（Capital Asset Pricing Model, CAPM）[15]的改进理论，目的在于解释股票市场的平均回报率受到哪些风险溢价因素的影响[16]。

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

而多因子模型就是在三因子模型的基础上，加入更多的因子。通过对股票不同维度进行评价并打分，形成因子，即$X_i, i = \{1,2,\dots,n\}$。对各因子进行线性加权后，对未来股票的收益进行预测，买入得分高的股票，卖空得分低的股票。

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

多因子模型是测验时间序列逐步聚类降维法的最佳模型之一，因为（1）多因子模型涉及多个维度的数据：涉及股票不同方面的表现，如估值、成长能力、盈利能力，和股票的风险，如市值、波动率、价量因素等；（2）由于基本多因子模型（根据公式（18））涉及时间序列数据的线性加权，因此可以检验算法是否有效的处理数据的多重共线性问题。

3.2 回测参数

回测样本空间为每月末满足以下条件的所有 A 股：

a) 中证 500 指数（000905.SH）成分股；

b) 非 ST 股及 ST 摘帽后三个月以上；

c) 调仓当天收盘非涨跌停且非停牌；

d) 当日成交额大于等于 1 百万人民币。

回测参数为：

a) 每个月月末调仓，选取样本空间中因子加权综合得分最大的前 20%只股票作为多头持仓；

b) 因子加权方式：简单等权相加；

c) 交易费用设为双边 0.3%；

d) 比较基准为中证 500 指数（000905.SH），即做空 IC 中证 500 股指期货（忽略基差及滚仓成本）；

e) 回测时间为 2010 年 1 月 1 日至 2018 年 10 月 31 日

3.3 因子列表

为了测试逐步聚类降维法，原始的时间序列数据选用了多因子模型中常用的 19 个因子，每个因子涵盖了中国 A 股股票的 3600 只股票（含已退市）在过去十年中每天的数据。因子列表及其对应算法如下：

表 3. 因子算法表[14][15][16]
Table 3. List of factors and calculation methods

因子维度	因子	简称	因子说明
规模	总市值对数	LNMCAP	总市值对数
估值	市净率	PB	总市值/股东权益（不含少数股东权益）
	一致预期市净率	ESTPB	一致预期滚动 PB
	市盈率	PE	总市值/归属母公司股东的净利润
	一致预期市盈率	ESTPE	一致预期滚动 PE
技术	三个月反转	REVERSE60D	过去 60 个交易日累计涨跌幅
	六个月反转	REVERSE120D	过去 120 个交易日累计涨跌幅
成长	营收同比增速	OperRevYoY	本季度营业收入/去年同季度营业收入
	净利润同比增速	NetProfYoY	本季度净利润/去年同季度净利润
	单季度净资产收益率同比变化	deltaROE	单季度净资产收益率-去年同期单季度净资产收益率
	单季度总资产收益率同比变化	deltaROA	单季度总资产收益率-去年同期单季度总资产收益率
盈利	单季度净资产收益率	ROE	归属母公司股东净利润/当期平均归属母公司股东权益
	单季度总资产收益率	ROA	净利润(含少数股东损益)/当期总资产
流动性	一个月日均换手	TURNOVER20D	过去 20 个交易日换手率均值
	三个月日均换手	TURNOVER60D	过去 60 个交易日换手率均值
	六个月日均换手	TURNOVER120D	过去 120 个交易日换手率均值
波动	一个月真实波动率	ATR20D	过去 20 个交易日日内真实波幅均值
	三个月真实波动率	ATR60D	过去 60 个交易日日内真实波幅均值
分红	股息率	DividendYield	近 12 月股息率

3.4 效果对比

在对原始因子数据进行预处理后，根据算法，先对因子单独的有效性和稳定性打分，即因子的信息比率（*IR*），如图（3）所示。因子 *IR* 在回测时间段内大部分显著大于 0，也就是说选择的因子对股票下一期收益都有一定的预测能力。得到每个因子的 *IR* 后，之后的逐步聚类算法将根据因子 *IR* 来排序。

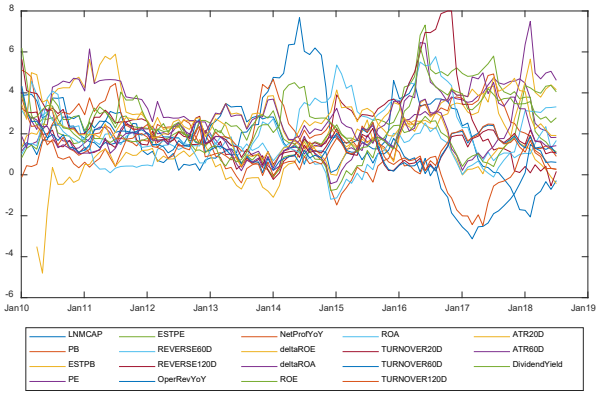


图 3. 原始因子滚动信息比率（IR）时序图
Figure 3. Time-series of raw data's information ratio

下一步，根据因子横截面的相关系数计算出滚动 12 期因子的相关系数矩阵。得到相关系数矩阵后，算法将对因子进行逐步聚类。相关系数矩阵及分类结果标签见下表（4）。从相关系数矩阵可以看出，大部分因子都存在与其相关性较高的其它单个或多个因子。分类结果准确地根据不同因子数据间相关性对因子进行了分类，也可以注意到有的分类并不是完全按照传统的因子维度来分类的，但仔细观察因子计算方法，会发现算法的分类标准更加科学。例如成长维度中的营收同比增速没有与同维度中的另外三个因子分到同一类别，这是因为在成长维度的四个因子中，它是唯一的一个用营业收入去计算的因子，而另外三个的计算方法都用到了企业的净利润；又例如估值维度中的市净率与市盈率并未被分到同一组，是因为虽然两个因子在计算时都采用股票总市值作为分子，但分母分别是相关性并不高的股东权益和归属母公司股东的净利润，因此这两个因子的相关性并不足以让他们分到同一组。

表 4. 原始因子相关系数矩阵及逐步聚类结果

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

分类标签	因子维度	因子	总市值对数	市净率	一致预期市净率	市盈率	一致预期市盈率	三个月反转	六个月反转	营收同比增速	净利润同比增速	单季度净资产收益率同比变化	单季度总资产收益率同比变化	单季度净资产收益率	单季度总资产收益率	一个月日均换手	三个月日均换手	六个月日均换手	一个月真实波动率	三个月真实波动率	股息率
11	规模	总市值对数	1.00	-0.01	-0.03	0.04	0.00	0.00	-0.03	-0.14	0.20	0.10	0.07	0.05	0.06	0.14	0.16	0.14	-0.02	-0.03	0.09
10	估值	市净率	-0.01	1.00	0.60	0.37	0.29	0.12	0.30	-0.11	-0.04	-0.14	-0.17	-0.26	-0.29	0.30	0.34	0.36	0.38	0.46	0.20
10	估值	一致预期市净率	-0.03	0.60	1.00	0.33	0.42	0.10	0.22	-0.11	-0.03	-0.11	-0.15	-0.20	-0.22	0.17	0.18	0.20	0.23	0.27	0.15
5	估值	市盈率	0.04	0.37	0.33	1.00	0.58	0.10	0.16	-0.03	0.05	0.09	0.06	0.25	0.20	0.13	0.12	0.12	0.20	0.22	0.47
7	估值	一致预期市盈率	0.00	0.29	0.42	0.58	1.00	0.13	0.21	-0.01	-0.01	0.00	-0.02	0.09	0.04	0.07	0.08	0.08	0.11	0.13	0.23
6	技术	三个月反转	0.00	0.12	0.10	0.10	0.13	1.00	0.65	-0.07	-0.11	-0.17	-0.18	-0.09	-0.09	-0.07	-0.09	-0.09	-0.10	-0.08	0.02
6	技术	六个月反转	-0.03	0.30	0.22	0.16	0.21	0.65	1.00	-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	成长	营收同比增速	-0.14	-0.11	-0.11	-0.03	-0.01	-0.07	-0.13	1.00	0.25	0.31	0.34	0.12	0.15	-0.01	-0.01	-0.01	-0.05	-0.07	-0.09
1	成长	净利润同比增速	0.20	-0.04	-0.03	0.05	-0.01	-0.11	-0.19	0.25	1.00	0.65	0.63	0.31	0.29	0.03	0.04	0.03	-0.03	-0.04	-0.04
1	成长	单季度净资产收益率同比变化	0.10	-0.14	-0.11	0.09	0.00	-0.17	-0.25	0.31	0.65	1.00	0.88	0.50	0.45	-0.03	-0.03	-0.04	-0.06	-0.09	-0.03
1	成长	单季度总资产收益率同比变化	0.07	-0.17	-0.15	0.06	-0.02	-0.18	-0.28	0.34	0.63	0.88	1.00	0.51	0.52	-0.02	-0.01	-0.02	-0.07	-0.10	-0.05
2	盈利	单季度净资产收益率	0.05	-0.26	-0.20	0.25	0.09	-0.09	-0.19	0.12	0.31	0.50	0.51	1.00	0.85	-0.11	-0.14	-0.15	-0.05	-0.09	0.20
2	盈利	单季度总资产收益率	0.06	-0.29	-0.22	0.20	0.04	-0.09	-0.23	0.15	0.29	0.45	0.52	0.85	1.00	-0.14	-0.17	-0.19	-0.05	-0.10	0.24
8	流动性	一个月日均换手	0.14	0.30	0.17	0.13	0.07	-0.07	0.05	-0.01	0.03	-0.03	-0.02	-0.11	-0.14	1.00	0.91	0.82	0.48	0.49	0.11
8	流动性	三个月日均换手	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	1.00	0.93	0.51	0.55	0.08
8	流动性	六个月日均换手	0.14	0.36	0.20	0.12	0.08	-0.09	0.09	-0.01	0.03	-0.04	-0.02	-0.15	-0.19	0.82	0.93	1.00	0.46	0.55	0.06
4	波动	一个月真实波动率	-0.02	0.38	0.23	0.20	0.11	-0.10	0.09	-0.05	-0.03	-0.06	-0.07	-0.05	-0.05	0.48	0.51	0.46	1.00	0.85	0.18
4	波动	三个月真实波动率	-0.03	0.46	0.27	0.22	0.13	-0.08	0.15	-0.07	-0.04	-0.09	-0.10	-0.09	-0.10	0.49	0.55	0.55	0.85	1.00	0.18
3	分红	股息率	0.09	0.20	0.15	0.47	0.23	0.02	0.02	-0.09	-0.04	-0.03	-0.05	0.20	0.24	0.11	0.08	0.06	0.18	0.18	1.00

根据表（4）中的分类，通过逐步降维法，原本 19 个因子被降维及组合成 11 个。新的 11 因子（根据其对应的分类标签命名）的相关系数矩阵如表（5）。同时也可以计算所有新因子时间序列上的平均相关系数，与逐步聚类降维前原始因子的平均相关系数对比，如图（4）所示，因子数据间的横截面相关系数在逐步聚类降维后显著下降。

表 5. 逐步聚类降维后的新因子相关系数矩阵
Table 5. Correlation matrix of new data after sequential clustering and dimension reduction algorithm

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

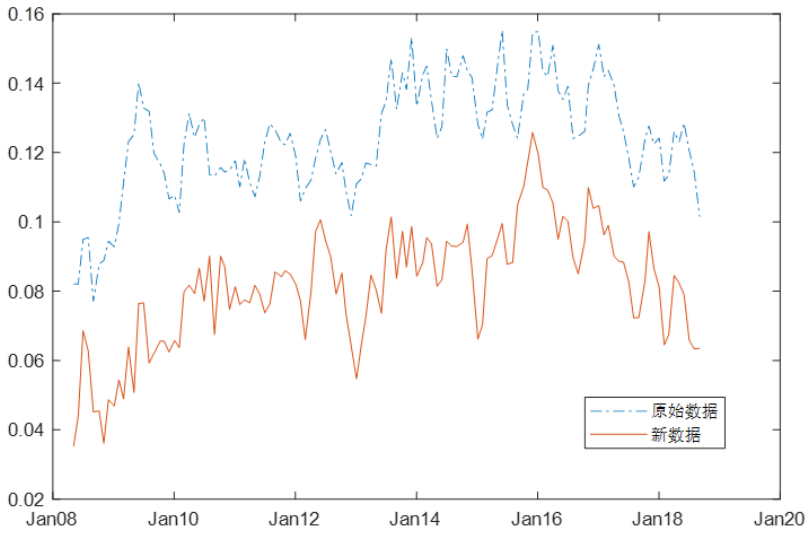


图 4. 平均相关系数时序对比图
Figure 4. Comparison of average correlation coefficient

根据图（3）的因子 *IR* 时序图可以看到，原始因子数据的 *IR* 已经在回测区间内的大多数时间显著大于 0，即因子有效。同样，可以对比原始因子与新因子在时间序列上的平均信息系数（*IR*），见图（5）。可以看到通过逐步聚类降维法，不仅新因子间的相关性减小，因子 *IR* 还能有一定的提升，证明逐步聚类降维法在降维的过程中能排除一部分因子间由于多重共线性导致的重复暴露和噪音，能够有效地保存甚至更高效提取因子数据中的有效信息。

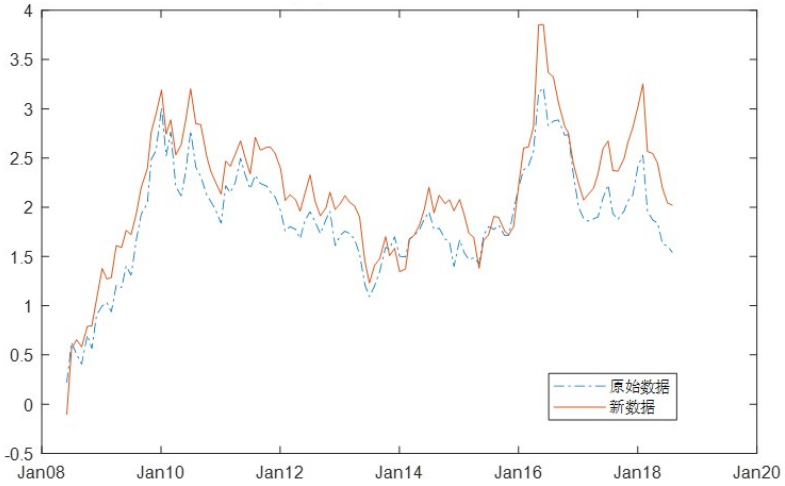


图 5. 平均信息比率（*IR*）时序对比图

Figure 5. Comparison of average Information Ratio

最后，也是最重要的，检验逐步聚类降维法的效果的标准就是：在经过逐步聚类降维法处理后的新因子是否能够有效地提高多因子模型的回测表现。本文通过比较同样框架下的多因子模型、使用同样的回测参数，分别对比用原始因子、主成分分析（PCA）降维后的因子和逐步聚类降维法处理后的新因子的回测结果，可证明逐步聚类降维法处理后的新因子能够更有效地提炼因子中的有效信息，使用新因子作为输入数据的多因子模型表现最好。其中，本文选取的作为结果对比的主成分分析法（PCA）是现在使用最为普遍的降维算法之一[17]。通过用改进后的主成分分析法[17][18][19]与新算法进行对比，可以对逐步聚类降维法在处理高维度因子时的效果进行一个直观的比较。

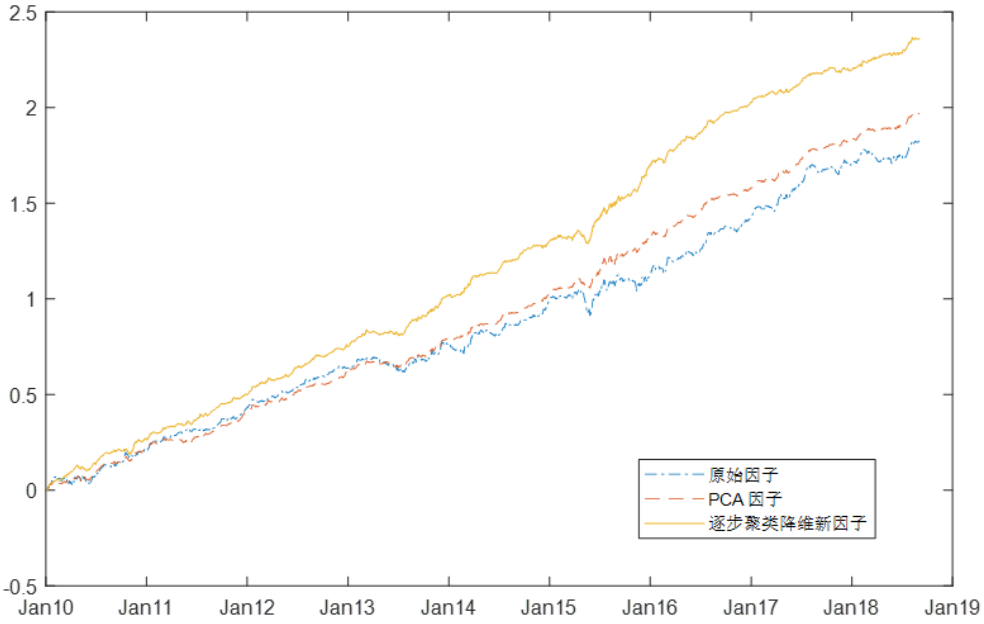


图 6. 多因子模型回测：净值对比图

Figure 6. Comparison of net asset value of Multi-Factor models using different data

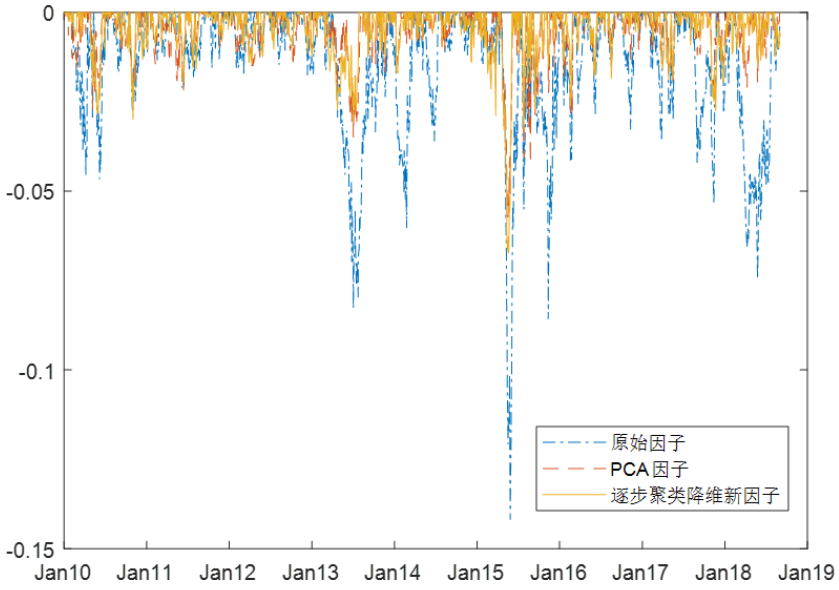


图 7. 多因子模型回测：回撤对比图
Figure 6. Comparison of drawdown of Multi-Factor models using different data

表 6. 多因子模型回测数据对比
Table 6. Comparison of back test data

	年化收益	年化波动率	夏普比率	最大回撤	收益回撤比
原始因子	22%	0.085	2.531	-14%	1.525
PCA 因子	23%	0.061	3.81	-6%	4.07
逐步聚类降维新因子	**28%**	**0.061**	**4.58**	**-7%**	**4.157**

通过对比原始因子、PCA 因子和新因子的回测表现（图（6）、图（7）），可以看到通过逐步聚类降维的新因子对多因子模型的贡献最大，回测效果最好：模型对比比较基准（中证 500 股指期货）的超额收益更高，超额收益净值的稳定性也显著提高。更进一步来说，通过计算模型超额收益的年化收益、夏普比率[20][21]、收益回撤比等参数可以量化模型表现[22][23]。通过表（6）的数据可以看到，除了最大回撤这一项新因子的表现和 PCA 因子表现相当，在其它量化指标上，新因子的效果都是最好的。以经典的衡量金融模型的夏普比率[22][23]为例，使用新因子的多因子模型的夏普比率可以达到 4.58，显著高于使用原始因子和 PCA 因子的 2.53 和 3.81。

4 结论

本文研究了一种处理高维度、高共线性时间序列因子数据聚类降维算法。该算法首先借鉴 K 最近邻聚类算法和贪心算法的思想，采用根据数据有效性为顺序，对数据进行逐步聚类；在得到聚类标签后，同样通过数据有效性为顺序，对数据进行基于普通最小二乘法的逐步组内加权、降维。

通过量化金融领域的多因子模型对算法进行验证，对比原始因子及新因子的相关性，可以证明算法能够有效地剔除时间序列数据间的多重共线性关系；通过对比因子有效性——信息比率，可以证明算法能够更有效地提取时间序列数据的有效信息；通过将逐步聚类降维的新数据带入到传统多因子模型中进行回测比较，证明了经过该算法处理后的时间序列数据能够显著提高多因子模型的回测表现，以经典的衡量金融模型的夏普比率[22][23]为例，使用新因子的多因子模型的夏普比率可以达到 4.58，显著高于使用原始因子和 PCA 因子的夏普比率，证明了该算法的有效性。因此可以得出结论：逐步聚类降维法能够有效地处理高维度、高共线性时间序列因子数据，帮助后续应用模型更好地剔除时间序列数据间的多重共线性关系，同时最大程度上保留甚至提高时间序列数据的有效性。

目前，我们在致力于进一步优化逐步聚类降维法的算法，希望使其在计算更高维度的数据（上百甚至上千个时间序列数据）时，算法速度更快，从而加强算法的实用性。同时也将研究该算法在应用到其它领域（如天气预测、医学医疗等）的模型时能有更好、更稳健的效果。

参考文献

[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.