长江证券 CHANGJIANG SECURITIES

机器学习

# 机器学习白皮书系列之二：无监督学习的方法介绍及金融领域应用实例

2017-11-27

金融工程 | 专题报告

**分析师** 覃川桃
☎ （8621）61118766
✉ qinct@cjsc.com.cn
执业证书编号：S0490513030001

**联系人** 陈洁敏
☎ （8621）61118706
✉ chenjm5@cjsc.com.cn

**相关研究**

《FOF 系列之首批公募 FOF 产品深入剖析》2017-11-25

《基于 HSAR 算法的阻力位和横盘突破时点识别》2017-10-24

《事件选股方法中的因子暴露与纯化事件收益》2017-10-15

## 报告要点

### ■ 无监督学习方法简介

本篇报告将进行无监督学习方法的介绍。无监督学习方法包括分布估计、因子分析、主成分分析、聚类分析、关联规则和 Google PageRank 算法等，本文主要就常用方法分成两类：聚类和降维进行介绍。

### ■ 降维方法的应用

实践中，将降维思想运用得炉火纯青的是 Barra 风险模型。个股和个券都有几十、上百个指标可以辅助分析其收益风险特征，通过降维的方式，Barra 提取出若干具有代表性的风险因子，找出了资产背后共同驱动因素，使用这些风险因子即可方便的进行绩效归因、组合风险控制等。降维的具体方法包括因子分析和主成分分析等。本文通过因子分析和主成分分析两种方法，结合常见的股票基本面、财务数据、技术指标等，构建选股策略。与基准相比，策略都能获取一定的超额收益，说明了通过降维提取主要特征能够起到一定的提纯和增强作用。

### ■ 聚类方法的应用

聚类分析方法基于相似性概念将数据集再划分，形成较小的组，追求组别间差异尽量大而组内的差异尽量小。根据样本数据特征和预期达到的效果，聚类可选择的方式非常多。本文详细介绍了 K-Means 聚类分析的原理，并且对于几种常见的聚类分析算法：沃德层次聚类、综合层次聚类算法、聚集聚类算法、基于密度的聚类算法、AP 聚类算法、谱聚类算法、小批量法等也一一进行简介。在具体应用上，聚类分析可以用做选股前的预处理，通过重要特征将个股分类之后在每个类别中分别进行选股，效果会优于在全样本内选股。此外，聚类分析的可视化也是重要的应用方式之一，通过热图或最小生成树的方式可以直观的描述资产间的相关性，帮助实现投资组合的风险分散。

### ■ 无监督学习方法的总结

无监督学习相较于上篇的监督学习算法更偏向于数据分析和特征提取，在机器学习中属于算法比较简单基础的类型，因此很多时候容易被忽略，但是不得不强调监督学习及我们系列的下篇将会介绍的深度学习算法如若想要达到较好的效果都离不开对于原始数据分析和处理工作，提升算法的复杂度对于效果的边际提升效应会受到使用的数据本身的局限。

**风险提示：**
1. 模型在使用中存在建模风险；
2. 本文举例均是基于历史数据不保证其未来收益。

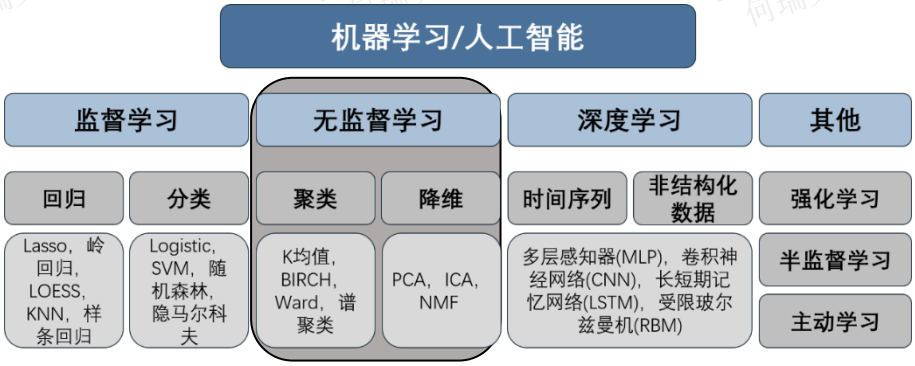
# 目录



# 图表目录

在机器学习白皮书系列的第一篇报告中，我们对于机器学习算法做了梳理，将机器学习/人工智能算法分为了监督学习、无监督学习、深度学习及其他，第一篇报告主要介绍了监督学习的算法及应用实例，通过将样本数据截取出部分作为训练期，在训练期中明确输入指标（X）及对应的标签（Y），在对应输出标签的“监督”下来选择合适的参数是监督学习的主要特征。监督学习的学习目标可分为两类：$P(X|Y)$和$E(X|Y)$，也就是回归问题和分类问题。

图 1：机器学习/人工智能方法介绍




资料来源：JP Morgan，长江证券研究所

回顾第一篇内容，回归中涉及到了惩罚回归模型和非参数回归模型。惩罚回归模型中金融领域使用得较多的有 Lasso 回归、岭回归和弹性网络回归；具有代表性的非参数回归模型则有：K 最近邻、LOESS 及卡尔曼滤波器。同时，也用到两个实例来说明了惩罚回归模型在拟合中的优势，以及卡尔曼滤波器使用时对于趋势判断、状态分辨的灵敏性。

分类算法包括逻辑回归、支持向量机（SVM）、决策树、随机森林以及隐马尔可夫模型。前面四种模型我们给出具体的择时和选股上的实例，使用决策树进行指数周度择时效果较为突出。隐马尔科夫模型我们则是验证其对于国内 A 股市场的状态划分是否有效，但是其月度市场状态划分效果不明显。

本篇报告将进行无监督学习方法的介绍，区别与监督学习，无监督学习是指在不区分输入指标和对应标签，通过直接输入全部样本的情况下学习数据集的分布特征，无监督学习方法包括分布估计、因子分析、主成分分析、聚类分析、关联规则和 Google PageRank 算法等。

# 无监督学习方法的简介

无监督学习的模型众多，本文主要就两类常用方法：聚类和降维进行介绍。图 1 中分别列举了聚类和降维两种类别下对应的部分模型。我们将选取文中实例中使用到的模型进行详细介绍和推导。

降维方法顾名思义就是在众多变量或指标中提取具有代表性的特征，主要包括因子分析、主成分分析、独立成分分析等。以主成分分析为例，主成分分析方法旨在识别数据的主要驱动因素或确定最具代表性的因子组合。例如，收益率曲线变动可以通过收益的平行移动、曲线的陡度变化和曲线的凸度来描述。在多资产组合中，主成分分析可以识别出如动量、价值、波动性、流动性等主要驱动因素。

聚类分析方法基于相似性概念将数据集再划分成较小的组，在金融领域，可以应用于识别波动率、利率等的高、低的状态，而准确的状态识别对不同资产及不同风险溢价的配置具有重要意义。

## 无监督学习方法的原理

降维和聚类的方法多种多样，理论模型部分我们将不做过多介绍，主要就我们给出应用实例时使用到的模型做详细说明。

### 降维方法：因子分析和主成分分析法

#### 因子分析

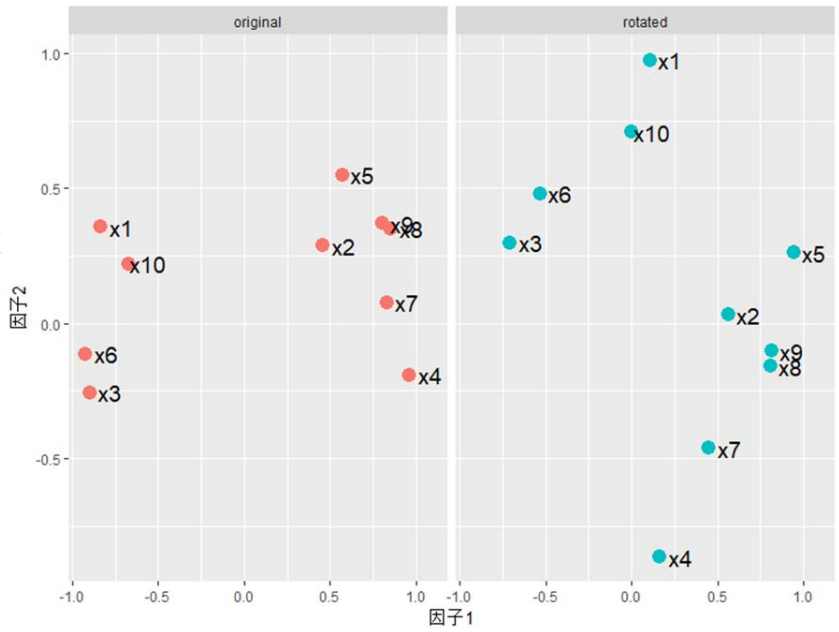
因子分析是一种常用的统计学变量降维和特征重建方法，模型为：

$$x_j = \lambda_{j1} f_1 + \lambda_{j2} f_2 + \cdots + \lambda_{jm} f_m + e_j, \quad \mathrm{j} = 1, \cdots, \mathrm{p}$$

其中$f_k$，$k = 1, \cdots, m$为 $m$ 个潜在公共因子，$\Lambda = (\lambda_{jk})$ 为 $p$ 个随机变量在这些公共因子上的载荷矩阵，$e_j$ 为随机误差项。一般地，该模型假设公共因子个数 $m$ 小于原始变量维数 p，也就是实现变量降维；其次，公共因子之间不相关，在正态分布下等价于独立，这有助于风险的分解；同时还会假设随机误差项与公共因子不相关。

因子分析得以被广泛使用，一方面是因为它能降维整合信息，也可视为去噪；另一方面，因子模型可通过旋转变换调整公共因子和载荷矩阵，进而通过各变量在单个因子上的载荷大小实现对公共因子的合理解释，这在实证分析中尤为重要。举一个简单的例子，用 2 维公共因子张成 10 个变量，对应的$10 \times 2$ 维因子载荷矩阵散点分布图如图 2 所示。

图 2：因子旋转与可解释性




资料来源：长江证券研究所

- （如左图所示）原始因子 1 可通过 $X_3, X_4, X_6, X_7$ 意义综合解释，因为这些变量在因子 1 上的载荷远大于在因子 2 上的载荷；
- （如左图所示）原始因子 2 无法得到合理的解释，因为没有变量在该因子的载荷能远大于在另一个因子的载荷；
- （如右图所示）通过旋转使得变量在两个因子的载荷差异尽量大，此时得到的因子 1 可结合 $X_2, X_8, X_9$ 来解释，因子 2 可结合 $X_1, X_4, X_{10}$ 来解释。

## 主成分分析（PCA）

主成分分析与因子分析类似，通过使用较低维度变量提取相关性较强变量中的信息，是处理变量共线性和充分提取数据信息的有效手段之一。PCA 方法最著名的应用是在人脸识别中特征提取及数据维度的降低。假设仅输入 100*100 维的人脸图像，提取它的灰度值作为原始特征，则这个原始特征将达到 10000 维，这给后面分类器的处理将带来极大的难度。著名的人脸识别 Eigenfaces 算法就是采用 PCA 算法，用一个低维子空间描述人脸图像，同时保存了识别所需要的信息。关于 Eigenfaces 算法的具体过程可以参考 Matthew Turk 和 Alex Pentland 的论文《Eigenfaces for Recognition》。

本质上来看 PCA 是 KL 变换处理离散情况的算法，是 KL 变换的一种应用形式。因此下面先介绍 KL 变换。

离散 KL 变换是一种常用的特征提取方法，用于实现最小均方误差下的最优正交变换，

对于向量$\mathbf{X}$，假设其可以用确定的完备正交归一向量系数$u_i$展开，则有：

$$x = \sum_{i=1}^{\infty} y_i \cdot u_i$$

其中正交基$u_i$满足，$u_i \cdot u_j = \begin{cases} 1, i = j \\ 0, i \neq j \end{cases}$

对上式进行变形可以得到：

$$y_i = y_i \cdot u_i^T \cdot u_i = u_i^T \cdot \left( \sum_{i=1}^{\infty} y_i \cdot u_i. \right) = u_i^T \cdot x$$

假设其中 q 维包含了向量 x 主要信息，我们通过 q 个有限项来尽可能的估计向量 x，公式如下：

$$\hat{x} = \sum_{i=1}^{q} y_i \cdot u_i$$

该估计的均方误差（MSE）为：

$$\varepsilon = E[(x - \hat{x})^T (x - \hat{x})]$$

$$= E\left[ \left( \sum_{i=q+1}^{\infty} y_i \cdot u_i^T \right) \left( \sum_{i=q+1}^{\infty} y_i \cdot u_i \right) \right]$$

$$= E\left[ \sum_{i=q+1}^{\infty} y_i^2 \right]$$

$$= E\left[ \sum_{i=q+1}^{\infty} \left( u_i^T \cdot x \cdot x^T \cdot u_i \right) \right]$$

$$= \sum_{i=q+1}^{\infty} [u_i^T \cdot E(xx^T) \cdot u_i]$$

其中$\Sigma = E(xx^T)$是向量 x 的二阶矩阵，在 PCA 中一般是协方差矩阵。接下来通过拉格朗日乘子法最小化均方误差$\varepsilon$：

$$g(u_i) = \sum_{i=q+1}^{\infty} [u_i^T \cdot \Sigma \cdot u_i] - \sum_{i=q+1}^{\infty} \lambda_i (u_i^T \cdot u_i - 1)$$

对$u_i, i = q+1, \cdots, \infty$求导数，得到$\Sigma \cdot u_i = \lambda_i \cdot u_i$。

即当取向量$\Sigma$的 q 个最大特征值所对应的特征向量为基向量来展开 x 时，其用于估计 x 的均方误差最小，为$\varepsilon = \sum_{i=q+1}^{\infty} \lambda_i$。

以上变换形式适用于 PCA，在实际应用 PCA 时，对于给定随机变量$x_1, \ldots, x_p$，其主成分方向及主成分的定义为：

1) 第一主成分方向$a_1 = (a_{11}, \cdots, a_{1p})'$使得第一主成分$Z_1 = a_{11} \cdot x_1 + \cdots + a_{1p} \cdot x_p$的方差最大化；其中$a_1' a_1 = 1$

2) 第$k, k = 2, \ldots, p$主成分方向$a_k = (a_{k1}, \cdots, a_{kp})'$使得第$k$主成分$Z_k = a_{k1} \cdot x_1 + \cdots + a_{kp} \cdot x_p$的方差最大化；其中$a_k' a_k = 1$，且$Z_k$与$Z_1, \ldots Z_{k-1}$不相关

结合 KL 变换，可以将主成分方向解释为：

- 对于未标准化随机变量，第$k$个主成分方向为协方差矩阵第$k$个特征值对应的特征向量，第$k$个主成分的方差为该特征向量对应的特征值；
- 对于标准化过后的随机变量，第$k$个主成分方向为相关系数矩阵第$k$个特征值对应的特征向量，第$k$个主成分的方差为该特征向量对应的特征值。

因此，实际使用过程中，数据标准化是主成分分析的重要过程，在变量方差近似相同时，我们可直接使用原始数据进行主成分分析，但是若数据数量级差异较大则会对结果产生一定影响。

主成分分析常与因子分析混淆起来，但它们之间存在本质不同。区别如下：

1) 主成分分析旨在提取原始变量内部最大可能的变动方向，得到方差最大的原始变量线性组合；因子分析则用因子的线性组合近似原始变量；

2) 因子分析有明确的模型形式表达，而主成分分析不能形成有效的模型表达；

3) 因子分析中的估计的因子会随因子数目发生变化，而主成分分析得到的主成分并不会随主成分数目选取变化；

4) 因子模型具有旋转不变性，因此可通过旋转得到合理解释；主成分旋转会改变其原始定义，且很难得到明确的解释意义。

## 聚类方法：K-Means 聚类分析

聚类分析将对象集合分组或分割成子集或集群，使集群内部有较高的相似性，而集群之间相似度较低。聚类分析也用于形成描述性统计以确定数据是否包含不同的子组，观察每一子组数据的特征，集中对特定的聚簇集合作进一步地分析。聚类方法众多，包括 K-Means、沃德（Ward）层次聚类、综合层次聚类算法、聚集聚类算法、基于密度的聚类算法（DBSCAN）、AP 聚类算法、谱聚类算法、小批量法等。在此就 K-Means 聚类算法做详细说明。

K-Means 算法是一种典型的硬聚类算法，通常以欧式距离作为相似度测度，将距离相对靠近的对象组成簇，目标是得到紧凑且尽量独立的簇。

假设我们聚类的对象是空间中点集合$\mathrm{D}=\{x_i|\mathrm{i}=1,2,3\cdots,\mathrm{n}\}$,在聚类初始给定聚类中心点为$\mathrm{C}=\{K_j|j=1,2,3,\cdots,k\}$，并且存在$x_i\in S(K_j)$，即每个数据点均会找到对应的距离较近的聚类中心。采用误差平方和作为聚类准则函数：

$$\mathrm{Cost}=\sum_{i=1}^{k}\Sigma_{x\in S(K_i)}\parallel x-\mu_i\parallel^2$$

其中$\mu_i$是第 i 个聚类$S(K_i)$所包含的对象点的均值，即聚类中心。聚类目标是最小化损失函数 Cost，具体实现步骤如下：

1) 选择 k 个初始中心点$S_0(K_i)$（随机生成或者通过算法来筛选）；

2) 对于集合 D 中每一个数据点$x_i$分别与中心点比较计算 Cost，记录 Cost 最小时每个数据点对应的中心点$K_i$，中心点和距离其相对较近的数据点组成簇；

3) 对一个簇，计算簇中所有点的均值并将均值作为新的聚类中心然后重复上面步骤，直到每个聚类稳定不再发生变化。

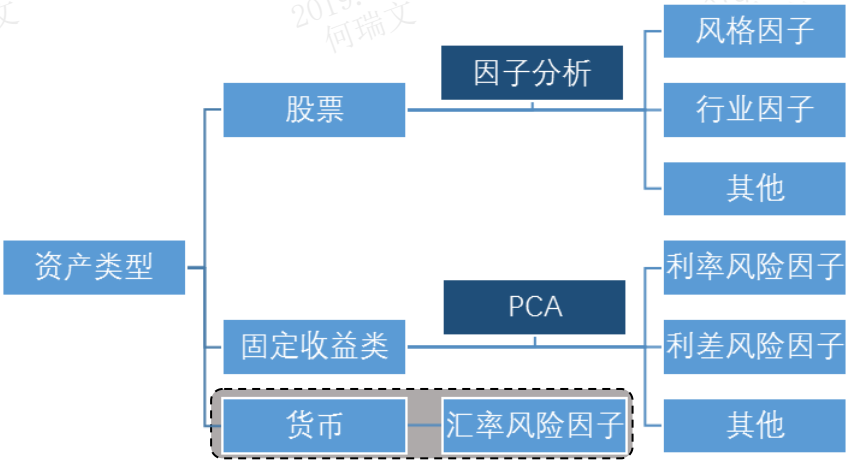
K-Means 之所以成为最常用的聚类算法之一在于其在处理大数据集的高效率，算法快速，逻辑也十分清晰。K-Means 算法的主要局限性在于：1)、需要事先给定类别数 k 和初始中心点，并且这两个参数对于最后分类结果影响较大；2)、对于样本点集合有形态上的要求，在数据呈圆形或高维球形时才能有效划分。

# 无监督学习方法在金融上的应用

降维的思想和因子分析方法在金融上最经典的应用莫过于 Barra 多因素模型，在 BIM301 中详细介绍了模型的构建原理，Barra 的做法是将不同地区不同资产类型综合到一个大框架之下，从底层到上层一步步进行降维和整合。Barra 的传统模型是将资产分为权益类（一般指股票）、固定收益类和货币类（在跨国家或地区时使用）。从市场的微观角度有一系列指标可以刻画这几类资产的走势，例如股票的价量指标、流通市值、PE、PB 等，固定收益类有利率期限结构、债券评级、个券违约风险等，这些指标之间本身存在相关性，因此首先是从众多微观指标到资产层面上的风险因子的降维。在大类资产的维度，又可以将不同国家或地区同类资产的因子进行整合，将其影响因素分成两部分：全

球因子（这一部分解释了不同地区相同类型资产的共同影响因素）和区域因子（仅仅影响当地资产走势）。之后 Barra 在资产类别中添加了商品、对冲基金，并且针对这些资产单独建立模型，不断完善大类资产配置框架。不过无论整体框架多么繁杂，其分析逻辑和应用算法非常清晰明了。

图 3：Barra 多因素模型及可用到的降维方法




资料来源：MSCI，长江证券研究所

聚类应用范围也非常广泛，如果运用在选股上最常见的做法是利用若干维度例如股票的估值、ROE、流通市值、行业等对股票进行聚类，将股票类别与股票收益率特征对应起来。在进行全球资产配置时，由于配置的资产标的丰富，也将聚类方法用于对资产的相关性矩阵进行分层聚类，以用于投资组合风险的分散化。

# 降维方法

在这一部分用因子分析、主成分分析方法构建简单的选股策略，主要目的是抛砖引玉，提供这两种方法在金融上应用的思路。

## 主成分分析法应用实例

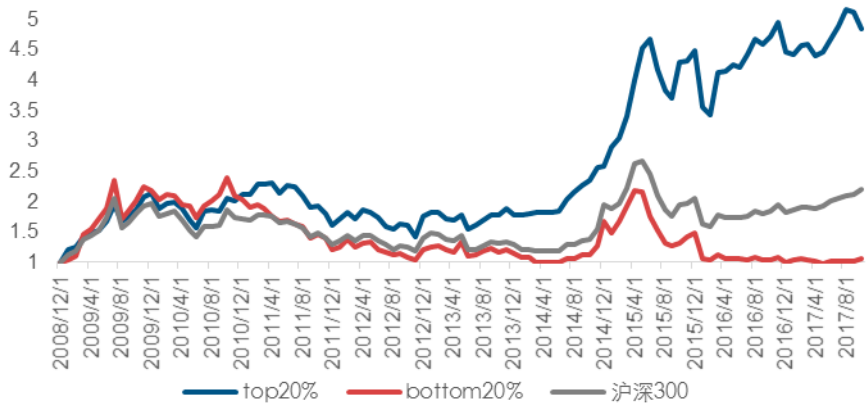
主成分分析是非常广为人知并且操作起来也比较方便的一种降维方法。最常见的应用案例是在进行多因子分析时的准备阶段，我们的因子池里面有接近八十多个子类因子，通过 PCA 可以将众多的因子进行特征抽取，最后将股票收益率归因于具有代表性的 8~9 个特征之上，然后再结合监督学习例如 SVM 等进行多因子选股。

我们给出案例是将上面过程简化，选择几个常用指标提取个股在这些指标上的第一主成分，根据提取出来的主成分值进行打分排序。策略的构建过程如下：

- 样本时间：2009 年 1 月到 2017 年 10 月；
- 使用的指标：21 日换手率、252 日换手率、beta 系数、PB、PE_TTM、净利润增速季度同比、反转、波动率、流通市值、上月涨跌幅及本月涨跌幅；
- 组合资产：沪深 300 指数成分股；

- 调仓频率：月度；
- 构建方法：对第一主成分排序打分（排序方向调整到与当期涨跌幅度方向一致），根据排序分成 5 组，选取前 20%（记为 top 20%）和后 20%（记为 bottom20%）分别构建等权组合。

图 4：主成分分析组合收益




资料来源：Wind，长江证券研究所

主成分分析策略的收益率略高于下面的因子分析策略，年化收益率为 19.54%，第一组和最后一组的收益率也有明显差异，说明了选取的指标对于股票收益率的解释一定程度上可以由提取出的第一主成分来代表。不过整体收益率仍然比多因子选股收益率低，也表明了第一主成分中丧失了部分重要信息，这时候就需要在降维和提取的指标解释度之间进行权衡。

表 1：主成分分析组合分年表现

| 年份 | top20% | | | bottom20% | | | 沪深300 | | |
|---|---|---|---|---|---|---|---|---|---|
| | 收益率 | 夏普比率 | 最大回撤 | 收益率 | 夏普比率 | 最大回撤 | 收益率 | 夏普比率 | 最大回撤 |
| 2009 | 112.58% | 3.71 | 14.32% | 118.50% | 2.43 | 26.53% | 96.71% | 2.55 | 24.22% |
| 2010 | -0.61% | -0.02 | 26.33% | -7.51% | -0.27 | 20.43% | -12.51% | -0.44 | 28.32% |
| 2011 | -23.88% | -1.16 | 30.49% | -39.87% | -2.20 | 39.87% | -25.01% | -1.59 | 27.59% |
| 2012 | 9.32% | 0.28 | 23.00% | -1.38% | -0.05 | 24.39% | 7.55% | 0.29 | 18.77% |
| 2013 | 1.60% | 0.08 | 15.94% | -4.14% | -0.15 | 16.90% | -7.65% | -0.35 | 18.38% |
| 2014 | 44.98% | 3.59 | 0.43% | 44.89% | 1.26 | 13.54% | 51.66% | 1.78 | 7.88% |
| 2015 | 73.03% | 2.28 | 20.57% | -10.52% | -0.26 | 41.58% | 5.58% | 0.16 | 33.83% |
| 2016 | -0.35% | -0.01 | 23.71% | -32.58% | -1.04 | 32.58% | -11.28% | -0.41 | 22.88% |
| 2017 | 10.14% | 0.77 | 6.13% | 6.51% | 0.63 | 8.29% | 25.76% | 4.25 | 0.47% |
| 平均 | 19.54% | 0.72 | 38.34% | 0.64% | 0.02 | 59.50% | 9.36% | 0.34 | 42.71% |

资料来源：Wind，长江证券研究所

## 因子分析的应用实例

因子分析用于选股可以基于目前股票市场常用的一些指标如基本面、财务指标、技术指标等，这些指标本身对于股票的收益率具有一定的解释能力，利用因子分析的方法提取它们共同部分可以进行有效降维。这种选股方式有点类似于多因子选股，我们在第一篇监督学习报告中有用支持向量机（SVM）方法进行多因子选股，中间实际上就存在组合因子数据解释下期收益率的过程。利用因子分析方法基于的假设是某些指标的共有信息对股票收益率有更好的解释效果。

为了验证这种方法是否有效，本文尝试使用 63 日换手率、PB、ROE_TTM、净利润增速季度同比、动量、反转和月涨跌幅提取因子构建投资组合。策略的构建过程如下：

- 样本时间：2009 年 1 月到 2017 年 10 月；
- 股票池：沪深 300 指数成分股；
- 调仓频率：月度；
- 构建方法：首先根据上面指标利用因子分析得到两个公共因子；然后计算个股在公共因子上的得分并进行排序（排序方向调整到与当期涨跌幅度方向一致）；根据排序分成 5 组，选取前 20%（记为 top 20%）和后 20%（记为 bottom20%）两部分，分别构建等权组合进行比较。

表 2：因子载荷矩阵

| 日期 | 换手率 | | PB | | ROE_TTM | | 净利润增速同比 | | 动量 | | 反转 | | 月涨跌幅 | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 因子1 | 因子2 | 因子1 | 因子2 | 因子1 | 因子2 | 因子1 | 因子2 | 因子1 | 因子2 | 因子1 | 因子2 | 因子1 | 因子2 |
| 2009/01 | -0.107 | 0.038 | 0.133 | 0.989 | 0.051 | -0.655 | 0.142 | -0.035 | -0.108 | 0.176 | 0.993 | 0.093 | 0.926 | 0.064 |
| 2010/01 | -0.184 | 0.235 | -0.045 | 0.346 | 0.082 | 0.104 | -0.106 | 0.028 | -0.026 | 0.827 | 0.997 | -0.021 | 0.948 | 0.055 |
| 2011/01 | 0.227 | 0.358 | 0.008 | 0.737 | 0.109 | 0.400 | 0.035 | 0.111 | -0.012 | 0.872 | 0.989 | 0.131 | 0.955 | 0.124 |
| 2012/01 | -0.171 | -0.019 | -0.154 | 0.909 | 0.179 | 0.554 | 0.049 | -0.031 | -0.074 | 0.606 | 0.972 | 0.005 | 0.997 | 0.018 |
| 2013/01 | 0.259 | 0.056 | 0.143 | 0.337 | -0.031 | 0.949 | 0.020 | 0.252 | -0.28 | 0.353 | 0.998 | -0.004 | 0.998 | -0.001 |
| 2014/01 | -0.03 | 0.623 | 0.18 | 0.511 | 0.205 | 0.070 | 0.019 | -0.122 | 0.164 | 0.465 | 0.888 | 0.092 | 0.99 | -0.122 |
| 2015/01 | 0.174 | 0.385 | -0.289 | 0.106 | 0.022 | -0.498 | 0.116 | -0.386 | 0.261 | 0.55 | 0.947 | 0.241 | 0.968 | 0.239 |
| 2016/01 | 0.099 | 0.38 | 0.176 | 0.905 | 0.042 | 0.300 | 0.092 | 0.03 | -0.041 | 0.397 | 0.99 | 0.118 | 0.926 | -0.036 |
| 2017/01 | -0.158 | 0.021 | -0.226 | -0.291 | -0.059 | -0.290 | -0.049 | 0.996 | -0.003 | 0.023 | 0.958 | 0.143 | 0.992 | 0.107 |

资料来源：Wind，长江证券研究所

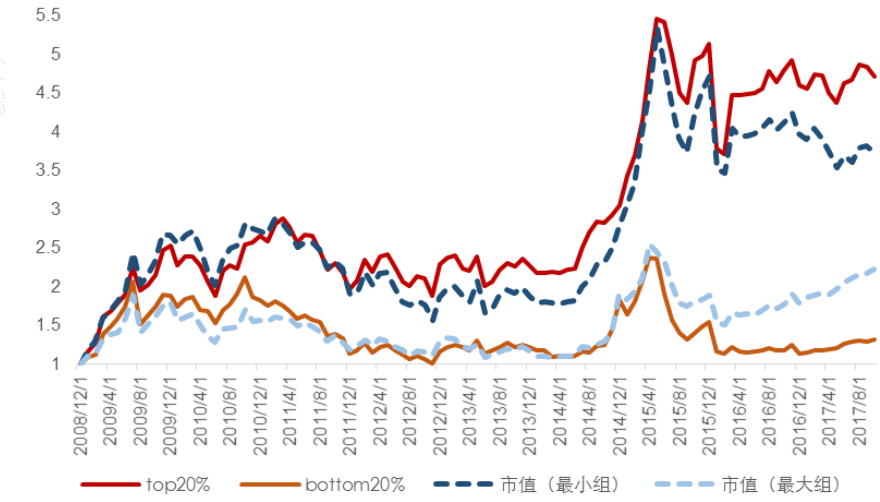
我们截取了每年年初所选变量在两个公共因子上的载荷，由于不同变量对于收益率的解释力度不一样，可以看到在公共因子上载荷大小有差距。并且提取的两个因子相互独立，可以较大程度对变量里的有效信息进行提纯。

图 5：因子分析选股策略净值


资料来源：Wind，长江证券研究所

从净值曲线上来看，因子分析选股第一组和最后一组的差异比较明显，相比于沪深 300 也有一定的超额收益，但是整体收益率比起我们使用 SVM 进行多因子选股时第一组的收益率低不少。

图 6：因子分析选股与单因子选股




资料来源：Wind，长江证券研究所

在选取的几个指标中，利用单指标进行选股时流通市值的表现较好，因此将因子分析选股方法与市值指标分组选股进行比较，效果好于单指标选股。

表 3：因子分析组合分年表现

| 年份 | top20% | | | bottom20% | | | 沪深300 | | |
|---|---|---|---|---|---|---|---|---|---|
| | 收益率 | 夏普比率 | 最大回撤 | 收益率 | 夏普比率 | 最大回撤 | 收益率 | 夏普比率 | 最大回撤 |
| 2009 | 153.02% | 4.40 | 13.65% | 88.86% | 2.11 | 26.49% | 96.71% | 2.55 | 24.22% |
| 2010 | 5.24% | 0.17 | 25.58% | -3.33% | -0.12 | 19.30% | -12.51% | -0.44 | 28.32% |
| 2011 | -25.89% | -1.26 | 31.48% | -37.59% | -2.02 | 37.59% | -25.01% | -1.59 | 27.59% |
| 2012 | 15.79% | 0.47 | 22.17% | 2.76% | 0.11 | 20.42% | 7.55% | 0.29 | 18.77% |
| 2013 | -0.39% | -0.02 | 16.47% | 4.40% | 0.21 | 13.66% | -7.65% | -0.35 | 18.38% |
| 2014 | 33.84% | 2.17 | 4.43% | 49.69% | 1.56 | 10.48% | 51.66% | 1.78 | 7.88% |
| 2015 | 68.73% | 2.23 | 19.98% | -15.84% | -0.39 | 44.32% | 5.58% | 0.16 | 33.83% |
| 2016 | -10.56% | -0.29 | 27.77% | -25.93% | -0.90 | 26.01% | -11.28% | -0.41 | 22.88% |
| 2017 | 3.00% | 0.25 | 7.73% | 19.45% | 3.56 | 0.60% | 25.76% | 4.25 | 0.47% |
| 平均 | 19.18% | 0.67 | 34.67% | 3.22% | 0.11 | 52.20% | 9.36% | 0.34 | 42.71% |

资料来源：Wind，长江证券研究所

分年来看，在市场整体表现较为强势时，因子分析选股策略能够获取一定的超额收益，但是在市场风格发生较大转变，例如 17 年，与多因子选股一样策略整体收益率比不上指数。不过值得注意的是在 14 年和 17 年排序靠后的 20%反而跑赢了大多数时候较为占优的第一组，也表明在实际分析的过程中，由于是按照月份来进行调仓，可以根据每个月的实际情况来选择适合当前市场的组合，进行灵活的操作。

综合上面结果可以看到依赖于因子分析方法构建的选股策略在获取有效的超额收益上不具备太大优势，与贴标签的监督学习方法不同，因子分析为代表的无监督学习还是偏重于线性关系的挖掘，在降维时并未考虑提取因子对目标变量的解释能力，所以有较大可能出现提取信息对于解释目标变量低效的问题。所以在进行进一步改进时，我们可考虑在因子分析模型中加入公共因子对目标变量的解释模型，它将对有效因子的提取过程进行一定限制，增强得到的因子对目标变量的解释力。

# 聚类方法

聚类的方法非常多，需要根据分析的数据特征以及期望达到的效果来选择合适的方法。这一部分将会对常见的聚类方法进行简介，介绍的所有聚类方法以及评价指标都能够在 Python 的 sklearn.cluster 模块中找到对应的函数来实现，每种方法对应的参数也均是依据对应函数的要求。

## 聚类方法的比较和评价

聚类是观察式的学习方式，其分析的结果可以提供多个可能的解，最终解的选择需要研究者的主观判断和后续的分析。聚类方法可以划分为层次聚类、谱聚类等。下面介绍一下几种典型的聚类方法：

沃德（Ward）层次聚类是与 K-Means 类似的分层聚类方法，所不同的是它使用树状图来进行样本点的聚类，并且适用于样本量较大聚类类别数较多的情况。

综合层次聚类算法（BIRCH）是为比较庞大的数据库设计的分层聚类方法，基于特征树进行聚类，主要参数有四个，重要参数用于调整树结构。它可以逐步聚类实时流数据，因此在许多情况下，只需要单遍扫描数据集就能进行聚类。

集聚聚类算法（Agglomerative Clustering）是另一种层次聚类方法，开始时将每个点作为聚类中心，然后将点合并成一个聚类类别。和 Ward 层次聚类中分析每个聚类内的离差平方和所不同的是，该方法分析所有聚类的观测值之间的离差均值。

近邻传播算法（Affinity Propagation）即 AP 算法基于数据点间的信息传递来形成聚类。将每个点都看作是潜在的聚类中心点，然后将两两数据点之间连线构成一个网络，通过不断迭代更新网络中各条边的消息来达到最终收敛确定聚类。该算法可以有效的发现大量的子类。用到的参数有阻尼系数和参考度（perference），通过调整参考度可以间接控制聚类数量。

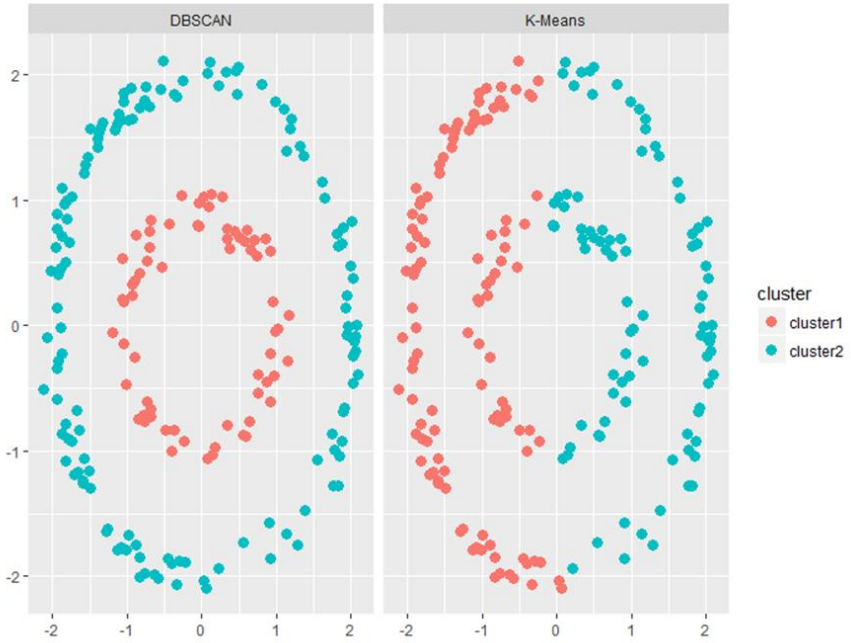
谱聚类算法（Spectral Clustering）和 AP 算法相似，它也在点之间传递信息，但不识别每个聚类的中心点，该算法需要指定聚类的数目，通过计算样本之间的关联矩阵，映射到低维空间中，然后运用 K-Means 来确定聚类，这种方法在聚类的数量较少的情况下效果较好。

小批量法（Mini Batch K-Means）为了减少处理时间，使用 K-Means 算法处理小批量数据。与 K-Means 方法不同的是，只对每个新的批次更新聚类的中心点，而不是每个新的点。Mini Batch 通常会比 K-Means 收敛速度更快，但是效果可能略差。

基于密度的聚类算法（DBSCAN）主要目标是区分被低密度区域分离的高密度区域。不同于基于距离的 K-Means 算法，DBSCAN 可以发现任意形状的聚类，并且能够排除奇异点的影响。最为重要的两个参数为 eps 和 min samples，分别用来控制样本分到聚类类别的距离阈值和类别中最小样本点个数。

综合上述关于聚类算法的介绍，我们可以看到不同聚类方法的聚类准则和适用场景有所区别。以 DBSCAN 和 K-Means 这两种聚类方法为例，在分类样本形态呈环形时，DBSCAN 将样本聚成内环和外环两类，解释性较强；K-Means 在 K=2 时将样本切分成左右平面，解释性相对较弱，且聚类结果不稳定（可能以任意直径方向切分）。因此，给定数据集，分析数据集的结构并选取合适的聚类方法将对聚类效果的影响较大。

图 7：K-Means 和 DBSCAN 在双环分布聚类上的比较




资料来源：长江证券研究所

合适的聚类方法的选择也可以从聚类效果出发，有一系列指标可以用于评价聚类效果。评价指标可以分为两类，第一类是当数据点存在可比照的聚类标签时，例如我们对个股收益率聚类，此时已知的个股行业分类就可以作为比照对象，用于评价聚类结果与行业分类是否一致。对于已知标签 $y_i$ 的数据，我们假设 $\hat{y}_i$ 为它对应的聚类标签预测。

第二类是数据点没有贴标签，此时只是单纯分析聚类是不是达到尽可能区分差异性的目的。下表列举了常用的一些聚类效果评价指标。

表 4：聚类效果评价指标

| 指标 | 名称 | 说明 |
|---|---|---|
| CP | 紧密性 | 每一个类内数据点到聚类中心的平均距离 |
| SP | 间隔性 | 各聚类中心两两之间平均距离 |
| DBI | 分类适确性指标 | 对任意两类别的类内距离平均距离之和比上两聚类中心距离求最大值，越小表示类内距离越小，同时类间距离越大 |
| DVI | 邓恩指数 | 任意两个簇元素的最短距离(类间)比上任意簇中的最大距离(类内) |
| RI | 维兰德系数 | 两种聚类结果中任意成对数据同时聚为一类或不同类的比例 |
| ARI | 调整维兰德系数 | RI的修正，利用两种聚类结果的列联表进行评价 |
| AS | 准确率 | 与基准相比分类正确的比例，$AS = \sum_{i=1}^{n} 1(y_i = \hat{y}_i)/n$ |
| PS | 精确率 | $PS = TP/(TP+FP)$，TP：真阳性，FP：假阳性 |
| F1 | F值 | F1=2P*R/P+R；P：精确率；R：召回率 |
| HS | 一致性 | 不同类别中元素无重叠 |
| CS | 完全性 | 给定分类的所有元素须分配到同一个类别 |
| HCV | 一致完全性 | HS和CS的均值 |
| HL^ | 汉明损失 | 分类错误的样本点占比 |
| JS | 杰卡德相似系数 | 预测值与真实值交集与二者并集的比值 |
| MI | 互信息 | $MI = \sum_{x \in X} \sum_{y \in Y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)}$ |
| A_MI | 调整互信息 | 根据最大值调整指标到0~1范围内 |
| Z_MI | 标准化互信息 | 根据均值将MI进行标准化 |
| Avg^ | 平均排名（升序） | 依据所有指标（除CP,SP,DBI,DVI和RI）进行排名，取均值 |

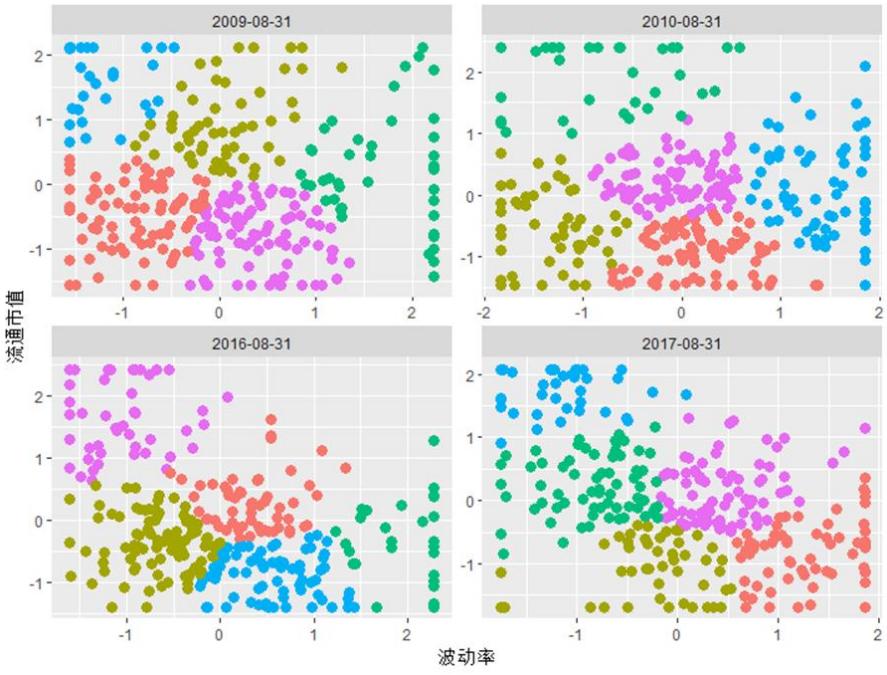
资料来源：Scikit-learn: Machine Learning in Python，长江证券研究所

# 聚类方法的应用实例

## 挖掘样本数据特征

聚类方法与机器学习中监督学习主要区别在于，其是一种探索性的分析方法，即通过模型挖掘出样本数据的特征，进而展示基于样本数据分出的不同聚类群体的差异及聚类群体在时间序列上的变化。图 8 反应了聚类结果在时间序列上的变化，选取的样本数据是沪深 300 指数成分股，根据股票的两个重要特征：流通市值和波动率进行聚类。对比 2009-08-31，2010-08-31，2016-08-31 和 2017-08-31 四个时期聚成五类的效果，上下两组图的对比明显，在 2009 年市值相对较大的组别里面，个股的波动率分布也比较均匀，分出了中、低波动率类别；在 2017 年市值相对最大的一组只对应低波动率特征。以市值大小来划分股票的话，体现了几年下来，不同市值个股对应的波动率水平的变化。

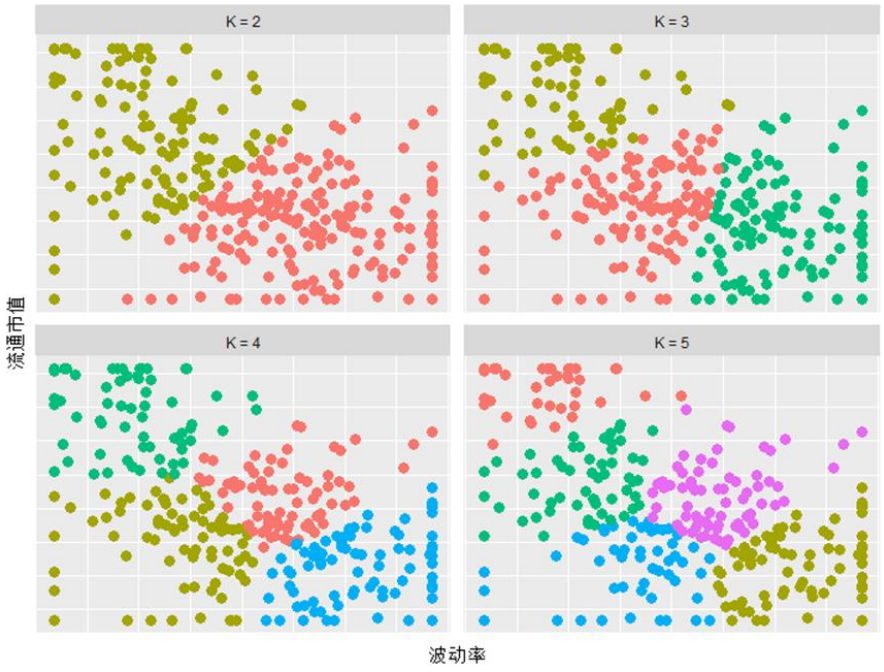
图 8：沪深 300 成分股聚类（时间序列上的变化）




资料来源：Wind，长江证券研究所

某些聚类算法如 DBSCAN 不需要提供聚类类别数，算法会根据数据特征来确定合适的类别数，选定不同的类别数决定了对数据分析时能够刻画出的特征。本文选取 2017 年 9 月的沪深 300 成分股，仍然是基于波动率和流通市值两个指标使用 K-Means 进行聚类，分别设置聚类类别数 K=2,3,4,5，比较聚类结果。

图 9：沪深 300 成分股聚类（不同聚类类别数下的变化）




资料来源：Wind，长江证券研究所

由于流通市值和波动率本身具有较强的相关性，当 K=2 时，基本上等同于根据波动率高低进行划分；当 K=3 时，在市值相对较低的类别里，依据波动率特征划分出了高波动率组和低波动率组；可以看到在一定范围内选择的聚类类别越多时，得到的聚类结果使用到的特征越多，分出来的结果也更细致。

聚类分析可以作为构建策略前的预处理，在金融领域分析很多问题时会用到情景分析的方法，就是根据重要变量预先设置好情景，在不同情景下采取不同的策略，这是一种较主观的分类方式。选准变量和模型，使用聚类方法也可以达到类似效果。

下面就举一个简单例子来说明使用聚类方法之后，在不同类别里面选股的效果。

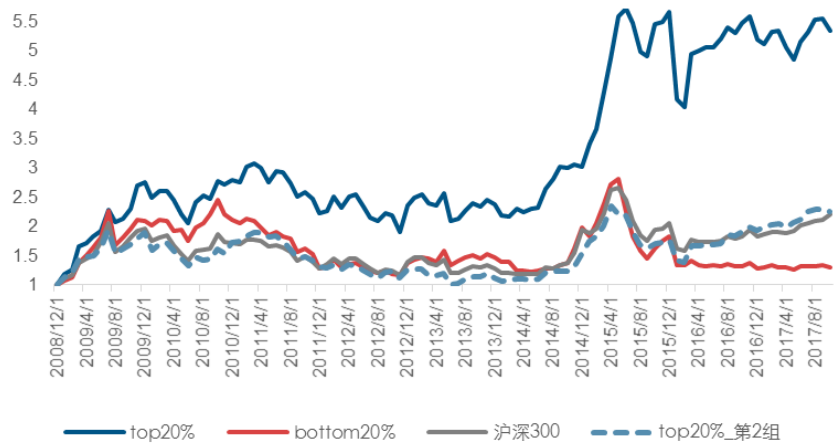
首先选取 252 日换手率和流通市值两个变量，使用 K-Means 算法进行聚类。为了类中心解释的稳定性和每个类别中数据点尽量充足，选择 K=2，将股票分为两类。然后基于前一部分主成分分析提取特征的有效性验证，分别提取两个聚类中个股的第一主成分，并按照主成分的值进行打分构建四种投资组合。提取主成分的原始变量为：21 日换手率、beta 系数、PB、PE_TTM、ROE_TTM、反转、流通市值、上月涨跌幅、本月涨跌幅。策略的构建过程如下：

- 样本时间：2009 年 1 月到 2017 年 10 月；
- 组合资产：当期沪深 300 指数成分股；
- 调仓频率：月度；

- 构建方法：使用 K-Means 将个股聚为两类，提取每个类别中个股的第一主成分值，进行排序打分（排序方向调整到与当期涨跌幅度正相关）；根据排序分成 5 组，选取排序前 20%（记为 top20%）和后 20%（记为 bottom20%）构建组合。

在两个类别里面，市值相对较小的一组选股的效果更好，我们在下图中也将市值相对较大一组排序前 20%的股票组合（top20%_第 2 组）的净值曲线以虚线列出。比较可以看出，通过流通市值和换手率将股票进行分组，不同组别里面用相同的指标选股的效果有差异。

图 10：K-Means 聚类组合收益




资料来源：Wind，长江证券研究所

分年来看，效果和直接使用 PCA 相比略有改善，说明了使用聚类算法做预处理的方式在提高组合收益率上有一定程度的效果。不过在沪深 300 内的选股本身样本池中股票数量就比较少，限制了分类的类别数，也造成效果提升幅度不大。建议在全市场采取聚类方式后再进行选股，不仅会缩小我们的选股范围，我们也可以在分出的不同特征的股票池内找出不同的适用选股指标，做到更精细化。

表 5：聚类分析组合年化收益

| 年份 | top20% | | | bottom20% | | | 沪深300 | | |
|---|---|---|---|---|---|---|---|---|---|
| | 收益率 | 夏普比率 | 最大回撤 | 收益率 | 夏普比率 | 最大回撤 | 收益率 | 夏普比率 | 最大回撤 |
| 2009 | 175.04% | 4.93 | 9.13% | 110.28% | 2.51 | 26.02% | 96.71% | 2.55 | 24.22% |
| 2010 | 1.74% | 0.06 | 25.55% | 0.84% | 0.03 | 17.39% | -12.51% | -0.44 | 28.32% |
| 2011 | -20.18% | -0.94 | 27.18% | -39.15% | -1.93 | 39.67% | -25.01% | -1.59 | 27.59% |
| 2012 | 5.62% | 0.16 | 25.14% | 6.21% | 0.24 | 19.22% | 7.55% | 0.29 | 18.77% |
| 2013 | 0.42% | 0.02 | 18.43% | 6.89% | 0.28 | 14.76% | -7.65% | -0.35 | 18.38% |
| 2014 | 27.40% | 1.43 | 8.17% | 35.83% | 1.13 | 15.76% | 51.66% | 1.78 | 7.88% |
| 2015 | 87.38% | 3.01 | 14.23% | -8.15% | -0.18 | 48.17% | 5.58% | 0.16 | 33.83% |
| 2016 | -8.31% | -0.22 | 28.59% | -30.26% | -1.03 | 30.26% | -11.28% | -0.41 | 22.88% |
| 2017 | 3.51% | 0.26 | 9.16% | 2.01% | 0.24 | 5.11% | 25.76% | 4.25 | 0.47% |
| 平均 | 20.87% | 0.70 | 37.88% | 2.98% | 0.09 | 55.12% | 9.36% | 0.34 | 42.71% |

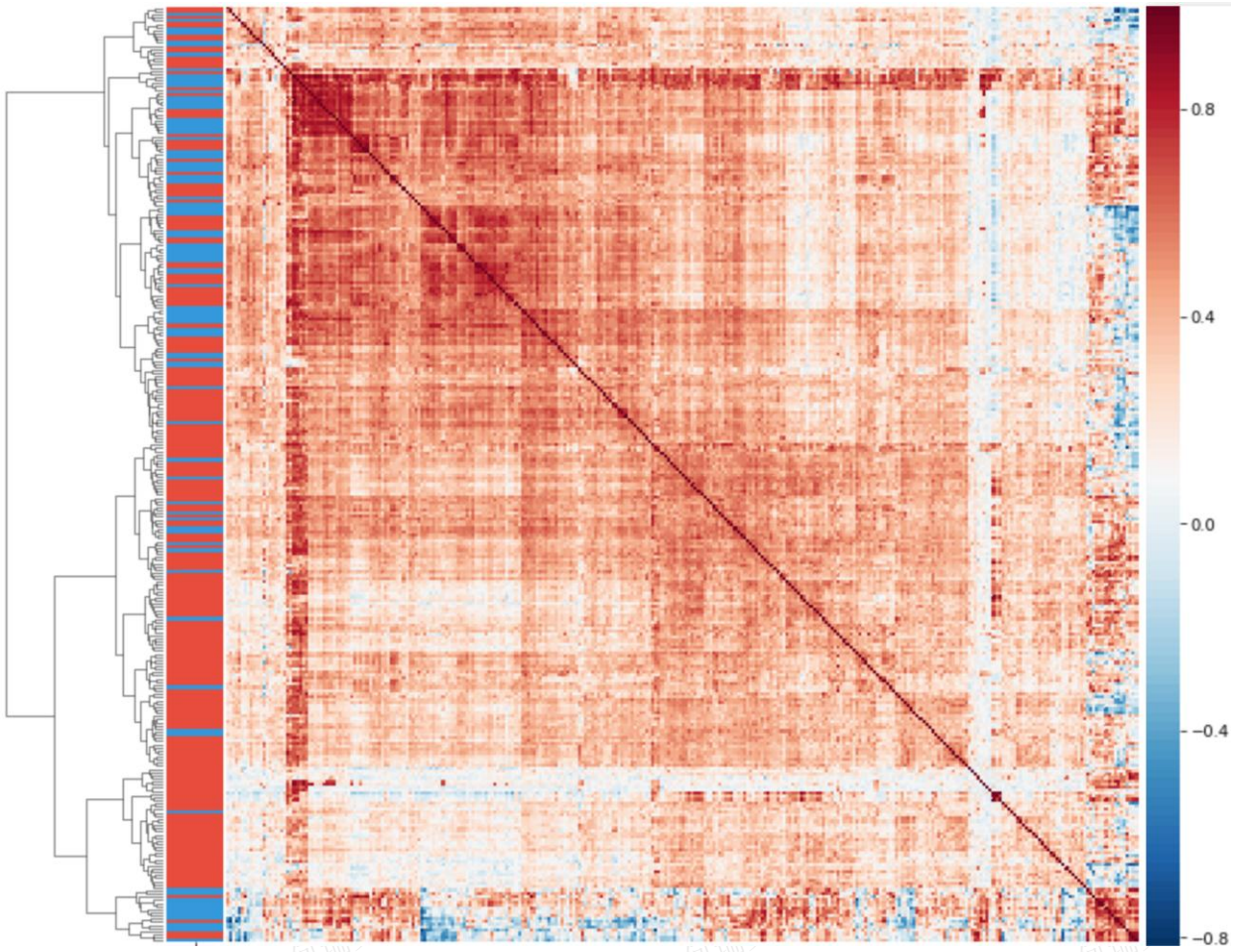
资料来源：Wind，长江证券研究所

除了在选股上的应用，在进行基金研究时也可引入聚类方法，我们在进行基金分类的时候会发现依据基金的名称或者基金的比较基准并不能对基金进行准确分类，这个时候可以利用基金的阶段收益率数据进行聚类分析，能够更合理的根据基金投资组合特征找出相似基金，然后进行类内基金比较筛选。

### 聚类的可视化

可视化是聚类分析的重要组成部分。下面的例子是根据沪深 300 成分股从 2008 年到 2017 年月度收益率的相关性使用 Ward 层次聚类得到的结果，左侧的纵轴代表几年下来个股涨跌情况（红色代表上涨，蓝色代表下跌），右侧的纵轴代表相关性高低的颜色分布。

图 11：Ward 分层聚类的可视化




资料来源：Wind，长江证券研究所

如果我们组合中涉及到多种资产也可以用上面的聚类方法来分析资产相关性，以尽量分散组合风险。下表中列举出根据上面层次聚类方法得到的部分聚类类别中的个股数。

表 6：Ward 分层聚类部分结果展示

| 层数 | 每层中个股数量 | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| L1 | 300 | | | | | | | | | | | | | | |
| L2 | 140 | 160 | | | | | | | | | | | | | |
| L3 | 20 | 120 | 104 | 56 | | | | | | | | | | | |
| L4 | 12 | 8 | 7 | 113 | 44 | 60 | 39 | 17 | | | | | | | |
| L5 | 5 | 7 | 7 | 5 | 37 | 76 | 17 | 27 | 23 | 37 | 8 | 31 | 9 | 8 | 3 |

资料来源：Wind，长江证券研究所

除了上面热力图的形式，当资产数量较少时，可以用最小生成树来可视化聚类。最小生成树是一种将资产间相关性进行可视化表示的方式，各个资产由点表示，通过线连接所有资产（点），资产间的相关性高低与线的长度成反比。最小生成树通过最小化连接线距离之和来选择连接资产之间的线，通过这种方式可以找到风险收益特征比较相近的策略或资产，以辅助我们的组合配置决策。

# 总结

作为白皮书系列的第二篇报告，本篇主要介绍了机器学习方法中的非监督学习方法，分为了降维和聚类两部分来介绍。降维方法主要有因子分析和主成分分析，聚类则有很多不同方法可以选择，本文主要介绍了 K-Means，并对其他常用的聚类方式及评价方法进行了简介。延续前一系列的风格在对原理进行介绍之后，也给出其在金融上的实践案例。不过在构建策略上无监督学习方法并无优势，其主要作用还是在数据分析和特征提取之上，可用作监督学习方法使用前的预处理。

## 投资评级说明

| | | |
|---|---|---|
| 行业评级 | 报告发布日后的 12 个月内行业股票指数的涨跌幅度相对同期沪深 300 指数的涨跌幅为基准，投资建议的评级标准为： | |
| | 看　　好： | 相对表现优于市场 |
| | 中　　性： | 相对表现与市场持平 |
| | 看　　淡： | 相对表现弱于市场 |
| 公司评级 | 报告发布日后的 12 个月内公司的涨跌幅度相对同期沪深 300 指数的涨跌幅为基准，投资建议的评级标准为： | |
| | 买　　入： | 相对大盘涨幅大于 10% |
| | 增　　持： | 相对大盘涨幅在 5%～10%之间 |
| | 中　　性： | 相对大盘涨幅在-5%～5%之间 |
| | 减　　持： | 相对大盘涨幅小于-5% |
| | 无投资评级： | 由于我们无法获取必要的资料，或者公司面临无法预见结果的重大不确定性事件，或者其他原因，致使我们无法给出明确的投资评级。 |

# 联系我们


## 上海

浦东新区世纪大道 1198 号世纪汇广场一座 29 层（200122）

## 武汉

武汉市新华路特 8 号长江证券大厦 11 楼（430015）

## 北京

西城区金融街 33 号通泰大厦 15 层（100032）

## 深圳

深圳市福田区福华一路 6 号免税商务大厦 18 楼（518000）


## 重要声明