华泰证券 HUATAI SECURITIES

数据模式探索：无监督学习案例

华泰人工智能系列之三十三

林晓明/陈烨/李子钰/何康/王晨宇

执业证书编号：S0570516010001

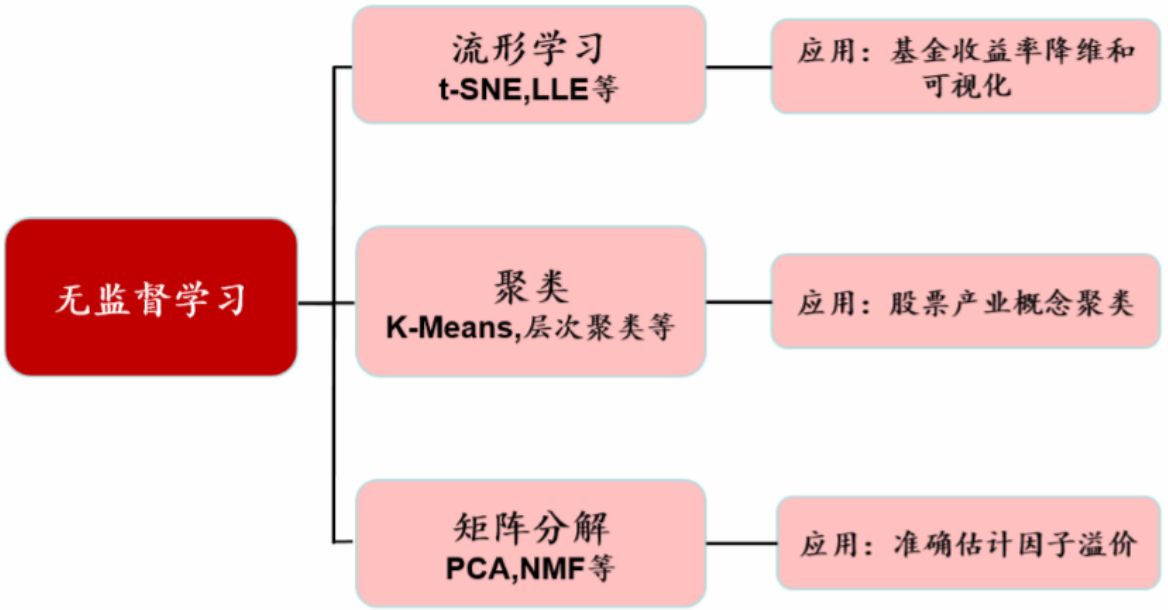
2020年7月

无监督学习

- 无监督学习适用于两种情况：

  1）标签难以获取　2）问题关注的是数据本身内部的结构，不需要标签的参与

图表1：无监督学习及其应用案例



资料来源：华泰证券研究所

1. 流形学习：通过非线性降维的手段将复杂的高维数据映射到低维，对于可视化数据内部结构很有帮助。
2. 聚类：通过给定样本相似度来挖掘样本之间的内在联系。
3. 矩阵分解：将矩阵拆解为数个矩阵的乘积从而提取矩阵内部隐含的信息，被用于数据降维、推荐算法中。

华泰证券
HUATAI SECURITIES

02

流形学习

流形学习

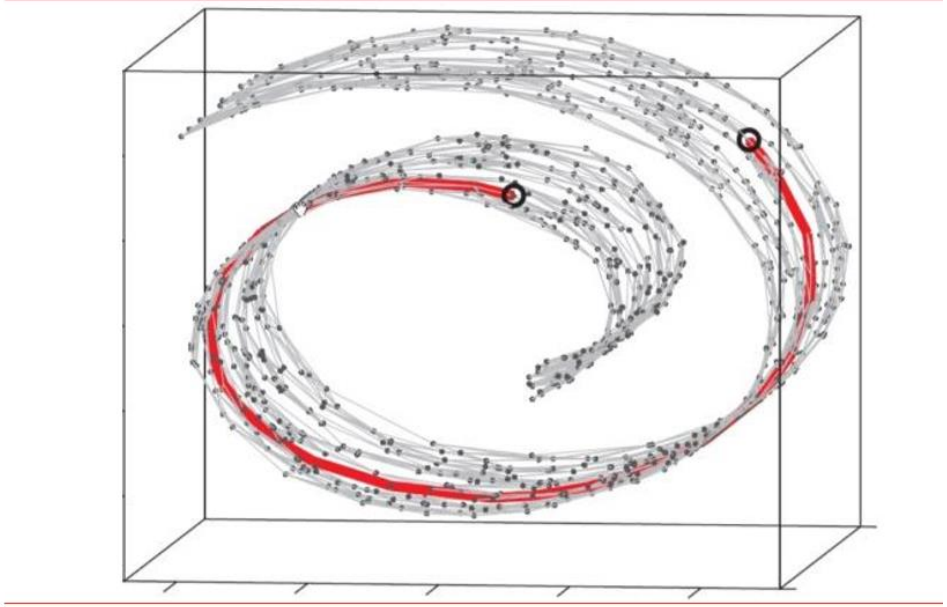
- 给瑞士卷曲面建模时，我们可以采用两种方法：

  （1）用三维空间刻画瑞士卷，并用三维坐标 P(x, y, z)来表示曲面上的点

  （2）将瑞士卷可以在二维平面展开，得到一个维度更低的流形空间

- 三维空间中大多数点并不在瑞士卷曲面上，说明使用三维空间刻画瑞士卷存在冗余。高维空间中的冗余可能会造成两个后果：1）维度灾难 2）测量误差

图表2：三维空间中的瑞士卷



资料来源：华泰证券研究所

- 在流形空间(把瑞士卷展开)上两个点的距离(红色的线)很远，但是用三维空间的欧氏距离来计算它们的距离则要近得多。

- 流形空间上点之间距离可以用欧氏距离测量，不代表低维流形所展开的高维空间中也可以使用欧氏距离测量，只有在流型空间中使用欧氏距离才有意义。

流形学习常用模型

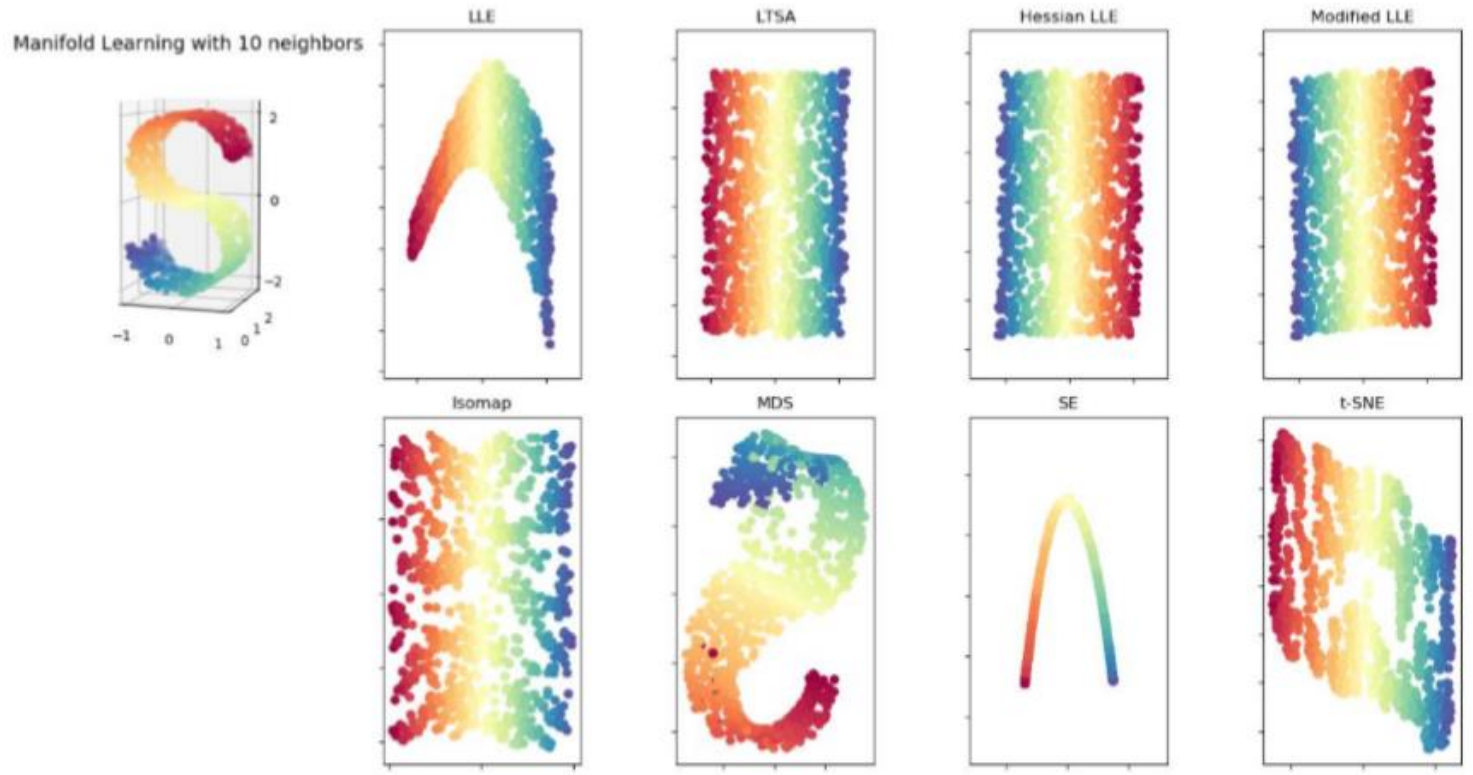
流形学习常用来数据降维并可视化。常用的模型如下：

- LLE(Locally Linear Embedding)：局部线性嵌入模型，目标为保持邻域内样本之间的线性关系。
- LTSA(Local Tangent Space Alignment)：局部切空间对齐模型，其基本思想是将流形的局部几何先用切坐标表示，那么流形中的每一个点处的切空间可以和欧式空间中的一个开子集建立同构，也就是切映射。
- Hessian LLE：相比于LLE，其用其已有邻域点的低维坐标线性表示新增样本点，来得到新增点的低维嵌入，使得算法更加简便。
- Modified LLE：相比于LLE，其利用多重权重向量解决LLE正则化的问题。
- Isomap(Isometric Mapping)：等距特征映射模型，其引进了邻域图，即样本只与其相邻的样本连接，使得较远的点可通过最小路径算出距离，在此基础上进行降维保距。
- MDS(Multidimensional Scaling)：多维尺度分析模型，其思路是保持新空间与原空间的相对位置关系，先用原空间的距离矩阵D，求得新空间的内积矩阵B，再由内积矩阵B求得新空间的表示方法Z。
- SE(Spectral Clustering)：谱嵌入模型，利用相似矩阵的谱(特征值)来对数据降维。
- t-SNE(t-distributed Stochastic Neighbor Embedding)：通过仿射(affinitie)变换将数据点映射到概率分布上，降维目标是保持近似的概率分布。

流形学习案例一：S型三维数据降维

- 图表3的案例来自sklearn，案例使用流形学习将左侧的S型的三维数据降维到二维平面。 可见，各算法的降维结果中都保持了原始数据的单调颜色变化，但由于各算法的原理不同，所得的二维流形也有一定差异。

图表3： S 型三维数据降维图

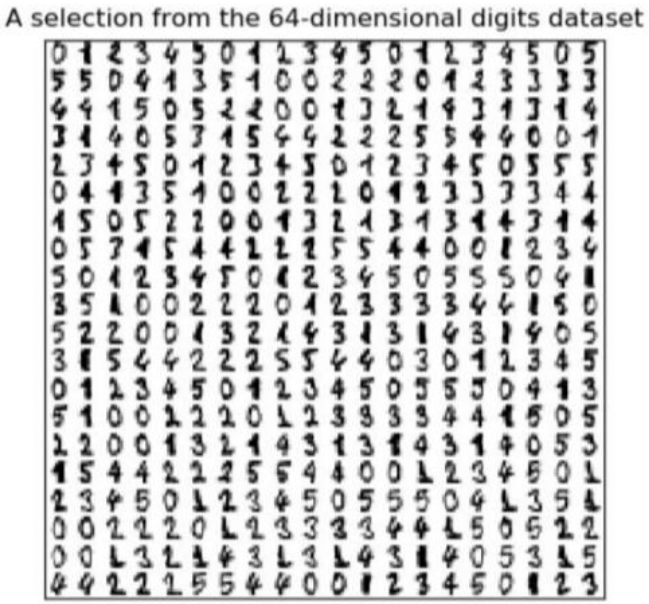


资料来源：sklearn，华泰证券研究所

流形学习案例二：手写体数字降维

- 数据来源于 sklearn 手写数字数据集，每张数字图片有 8*8 个像素，将数据集图片按行展开，每张图可转换为 1*64 的手写数字向量。
- 该案例的目的是通过各种流形学习算法将 64 维的手写数字向量降维到二维平面，并观察数字 0-5 的图片在降维后的分布情况。

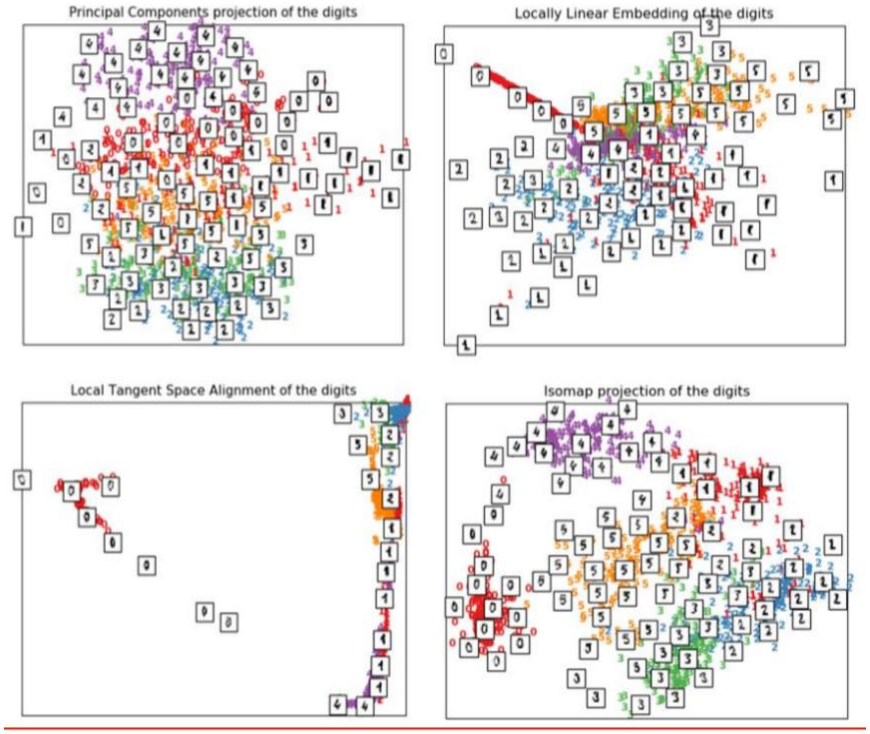
图表4：手写数字数据集



资料来源：sklearn，华泰证券研究所

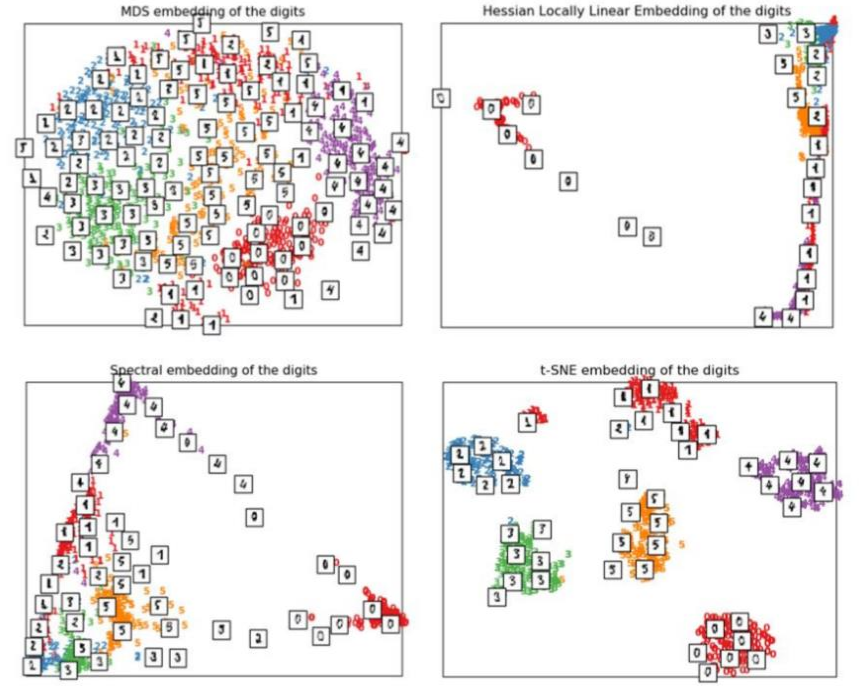
流形学习案例二：手写体数字降维

图表5：手写数字降维图 1



资料来源：sklearn，华泰证券研究所

图表6：手写数字降维图 2

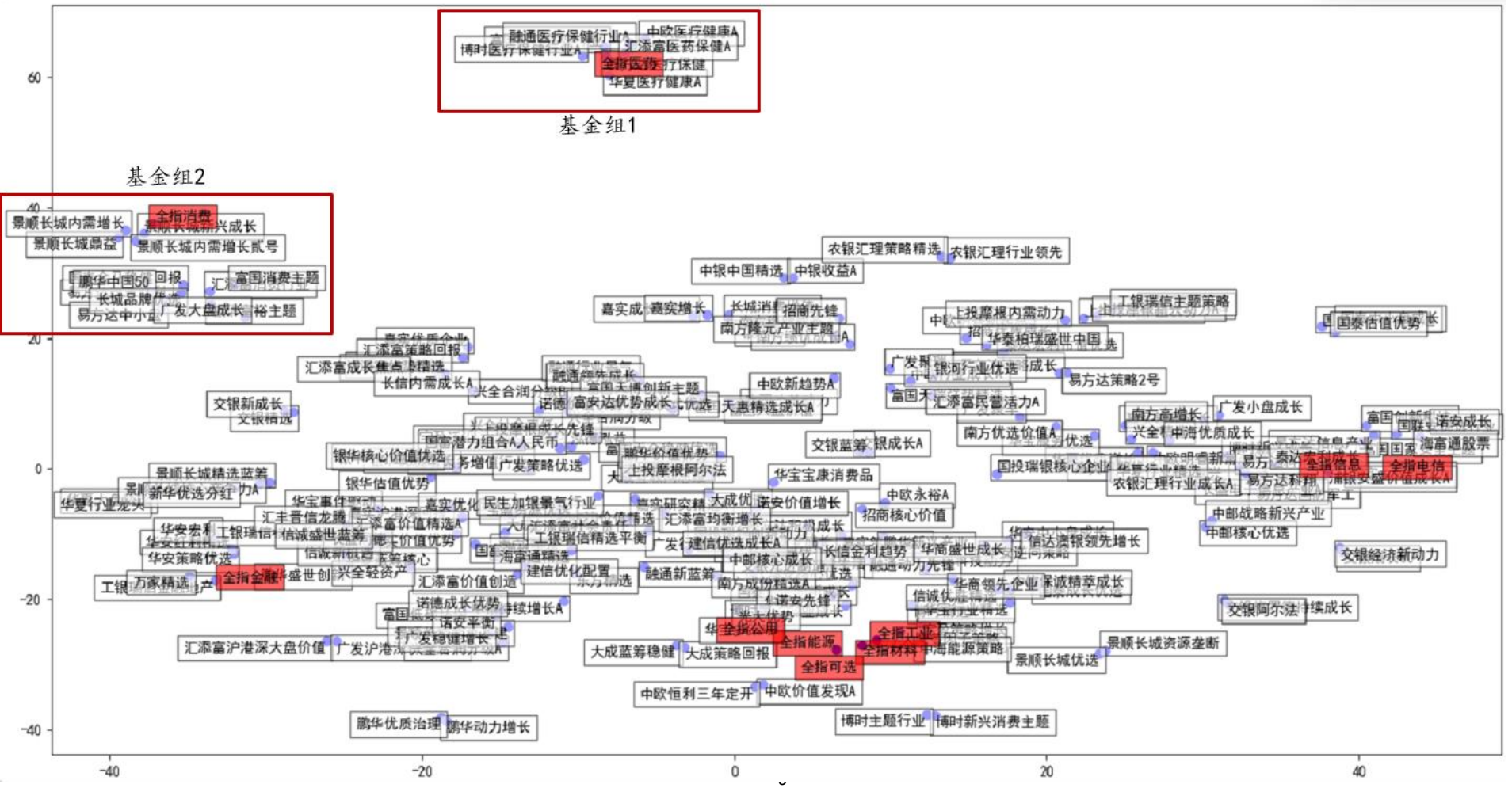


资料来源：sklearn，华泰证券研究所

- 图表 5 和图表 6 展示了各算法的降维效果，可知，t-SNE 的降维效果较好，各数字对应的样本点自然地聚成了 6 个簇。

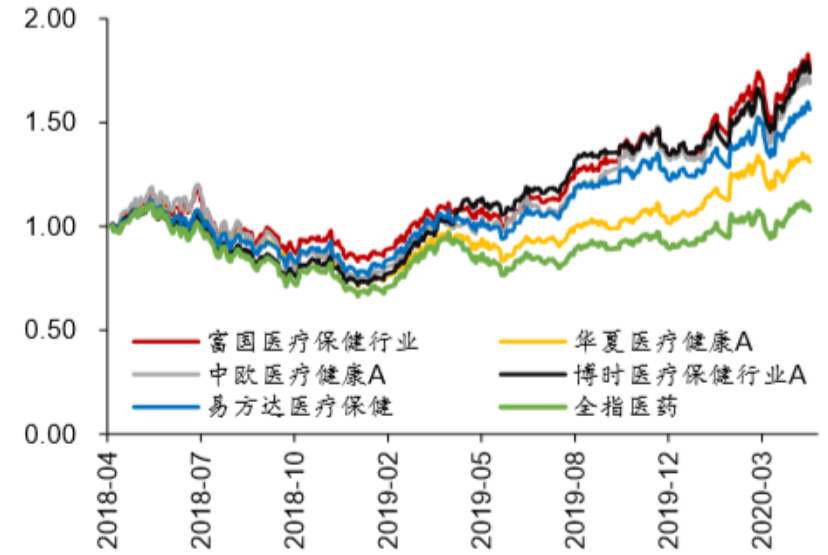
流形学习案例三： t-SNE 进行基金收益率降维和可视化

- 使用 t-SNE 进行基金收益率降维和可视化，收益率相近的基金出现了聚集现象。



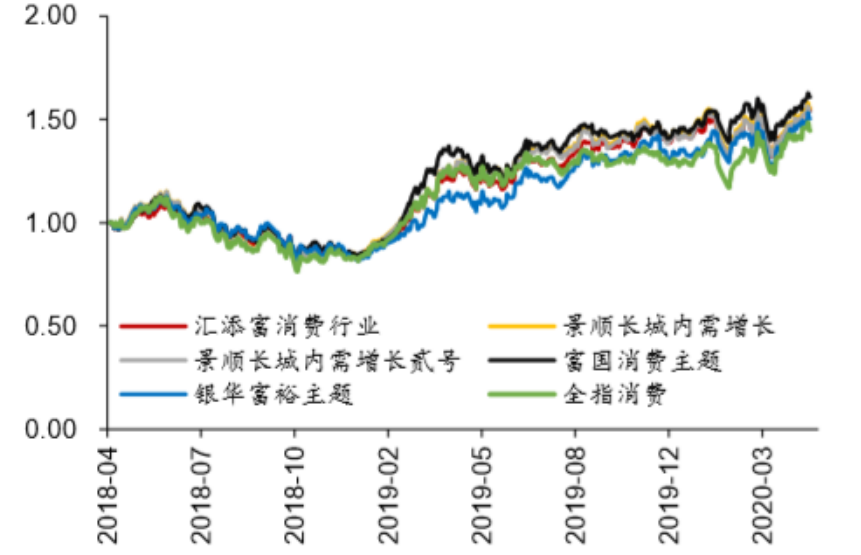
流形学习案例三： t-SNE 进行基金收益率降维和可视化

图表8： 偏股混合型基金组 1 净值



资料来源：Wind，华泰证券研究所

图表9： 偏股混合型基金组 2 净值



资料来源：Wind，华泰证券研究所

- 在图7的基金组 1 中选取 5 只基金并加入全指医药指数，可得到图表 8 中的净值曲线，基金净值以及指数的走势相似。

- 在图7的基金组 2 中选取 5 只基金并加入全指消费指数，可得到图表 9 中的净值曲线，基金净值以及指数的走势相似。

华泰证券
HUATAI SECURITIES

03

聚类

聚类常用算法

常用的聚类算法如下：

- K-Means：一种迭代求解的聚类分析算法，原理是把每个对象分配给距离它最近的聚类中心，直到达到平衡。

- AP聚类(Affinity Propagation)：基于图论的聚类算法，通过迭代过程不断更新每一个点的吸引度和归属度值，直到产生m个高质量的Exemplar(类似于质心)，同时将其余的数据点分配到相应的聚类中。

- 谱聚类：基于图论的聚类方法，通过对样本数据的拉普拉斯矩阵的特征向量进行聚类，可以理解为将高维空间的数据映射到低维，然后在低维空间用其它聚类算法(如K-Means)进行聚类。

- 层次聚类：分为凝聚(自底向上)和分裂(自顶向下)两种方法，常用的方法是凝聚法，通过某种相似性测度计算节点之间的相似性，并按相似度由高到低排序，逐步重新连接个节点。

- DBSCAN：基于密度的聚类算法，其将簇定义为密度相连的点的最大集合，能够把具有足够高密度的区域划分为簇，并可在噪声的空间数据库中发现任意形状的聚类。

聚类算法对比

图表10：聚类算法对比

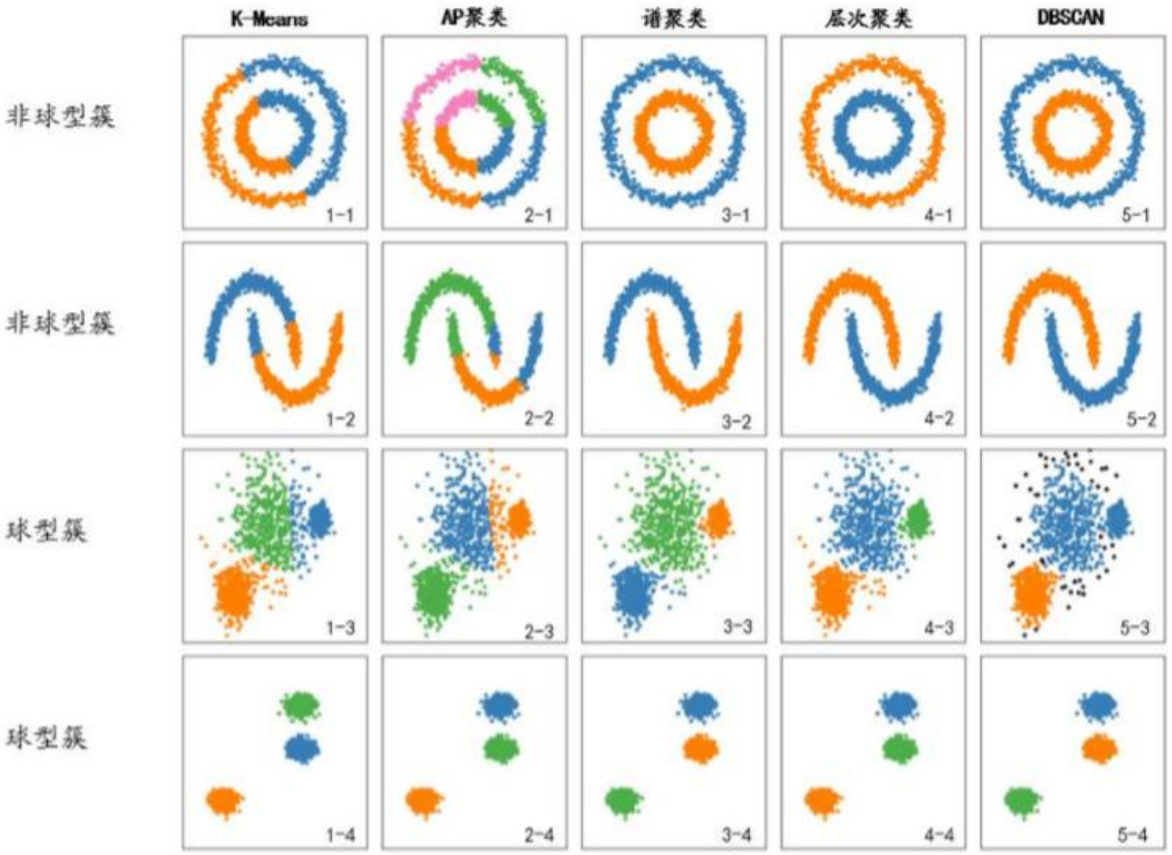
模型名称	是否需要指定聚类数目	模型输入	优点	缺点
K-Means	是	样本的特征取值或样本间的相似度	简单快速、适用于大量数据	初始值敏感、异常值敏感、容易局部最优、不能处理非球型簇(如图表 11 的 1-1 与 1-2 所示，图形连接处被断开)、不支持过多的簇
AP 聚类	否	样本的特征取值或样本间的相似度	支持大量的簇	时间复杂度高、数据量不可拓展
谱聚类	是	样本的特征取值或样本间的相似度	簇形状不敏感(如图表 11 的 3-1 与 3-2 所示，能够识别出链状簇)、适用于稀疏数据、处理高维数据复杂度低	不支持过多的簇、对相似度定义和参数敏感
层次聚类	是	样本的特征取值或样本间的距离	支持大量的簇、簇形状不敏感(如图表 11 的 4-1 与 4-2 所示，能够识别出链状簇)、适用于大量数据、能够显示出聚类层次、全局最优	时间复杂度高
DBSCAN	否	样本的特征取值或样本间的距离	适用于大量数据、簇形状不敏感、对噪音不敏感(如图表 11 的 5-3 所示，黑色数据点(噪音)不被分类)	空间复杂度高、需自定义变量多且参数敏感

资料来源：sklearn，华泰证券研究所

聚类算法对比

- 图11展示了不同聚类算法对于不同结构簇的聚类效果。可以看到谱聚类、层次聚类和DBSCAN对非球型簇有较好的聚类能力，即对簇形状不敏感。

图表11：球形簇和非球形簇的聚类结果



资料来源：sklearn，华泰证券研究所

聚类算法案例：基于股票产业概念的聚类

- 我们将对股票按照所属产业概念进行聚类，以观察 A 股概念的分布情况，股票的概念数据来自于 Wind。聚类算法的模型常用的输入是样本间的相似度或距离，我们使用股票概念的余弦相似度来衡量股票的相似度：

$$Similarity(A,B)=\frac{|A\cap B|}{\sqrt{|A|*|B|}}$$

- 其中 A 为股票 1 所属概念集合，B 为股票 2 所属概念集合。例如：
（1）股票 1 所属概念集合为：{干细胞;肺炎概念;创新药;生物疫苗;国产化创新;大消费}
（2）股票 2 所属概念集合为：{大消费;国产化创新;流感;肺炎概念;生物疫苗;血液制品;创新药}

- 则它们的余弦相似度为：

$$Sim(\text{股票 1},\text{股票 2})$$
$$=\frac{length\{\text{肺炎概念;创新药;生物疫苗;国产化创新;大消费}\}}{\sqrt{length\{\text{干细胞;肺炎概念;创新药;生物疫苗;国产化创新;大消费}\}*length\{\text{大消费;国产化创新;流感;肺炎概念;生物疫苗;血液制品;创新药}\}}}=\frac{5}{\sqrt{6*7}}$$
$$\approx 0.77$$

聚类算法案例：基于股票产业概念的聚类

- 计算两两股票之间的相似度就可得到相似度矩阵：

$$\begin{bmatrix} Sim_{1,1} & Sim_{1,2} \cdots & Sim_{1,n} \\ \vdots & \ddots & \vdots \\ Sim_{n,1} & \cdots & Sim_{n,n} \end{bmatrix}$$

- 其中，$Sim_{i,j}$ 表示股票 i 与股票 j 之间的相似度，$Sim_{i,i} = 0$。$Sim_{i,j}$ 越大，则股票概念越相似，股票在高维空间中越靠近。由于股票相似度和距离呈现负相关，可取$(1-Simi,j)$作为距离矩阵中的元素，得到距离矩阵：

$$\begin{bmatrix} 1-Sim_{1,1} & 1-Sim_{1,2} \cdots & 1-Sim_{1,n} \\ \vdots & \ddots & \vdots \\ 1-Sim_{n,1} & \cdots & 1-Sim_{n,n} \end{bmatrix}$$

聚类算法案例：基于股票产业概念的聚类

- 在无法获取真实标签时，聚类常用的评价指标有以下三个：
  1. Silhouette Coefficient(轮廓系数)：该指标反映了不同簇类之间的分离度，值域为[-1，1]，值越大说明簇与簇之间距离越明显。
  2. Calinski-Harabasz Index(方差比准则)：该指标是类间离差矩阵的迹与类内离差矩阵的迹的比值。数值越大，组间协方差很大，组与组之间界限明显。
  3. Davies-Bouldin Index(分类适确性指标)：该指标度量每个簇类最大相似度的均值。值域大于0，值越小表示聚类效果越好。

图表12：沪深300成分股聚类评价指标

	轮廓系数	方差比准则	分类适确性指标
K-means	0.21	45.98	1.57
层次聚类	0.24	44.46	1.51
谱聚类	0.17	34.83	1.50

资料来源：Wind，华泰证券研究所

图表13：中证500成分股聚类评价指标

	轮廓系数	方差比准则	分类适确性指标
K-means	0.17	41.95	1.99
层次聚类	0.15	39.38	1.87
谱聚类	0.10	24.72	2.30

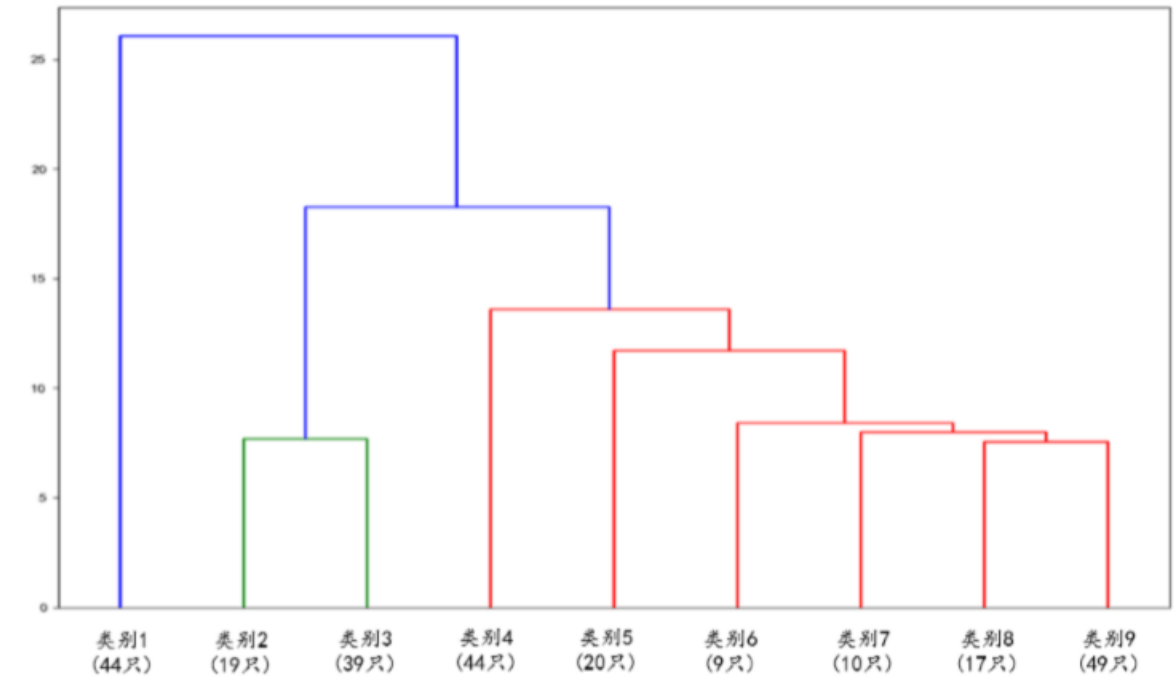
资料来源：Wind，华泰证券研究所

- 从三个评价指标来看，谱聚类表现最差，K-means和层次聚类的表现接近。

聚类算法案例：基于股票产业概念的聚类

- 层次聚类可以通过分层的方式显示各个类别的联系。图 14 展示了沪深 300 成分股的层次聚类效果图。

图表14：沪深 300 成分股层次聚类图

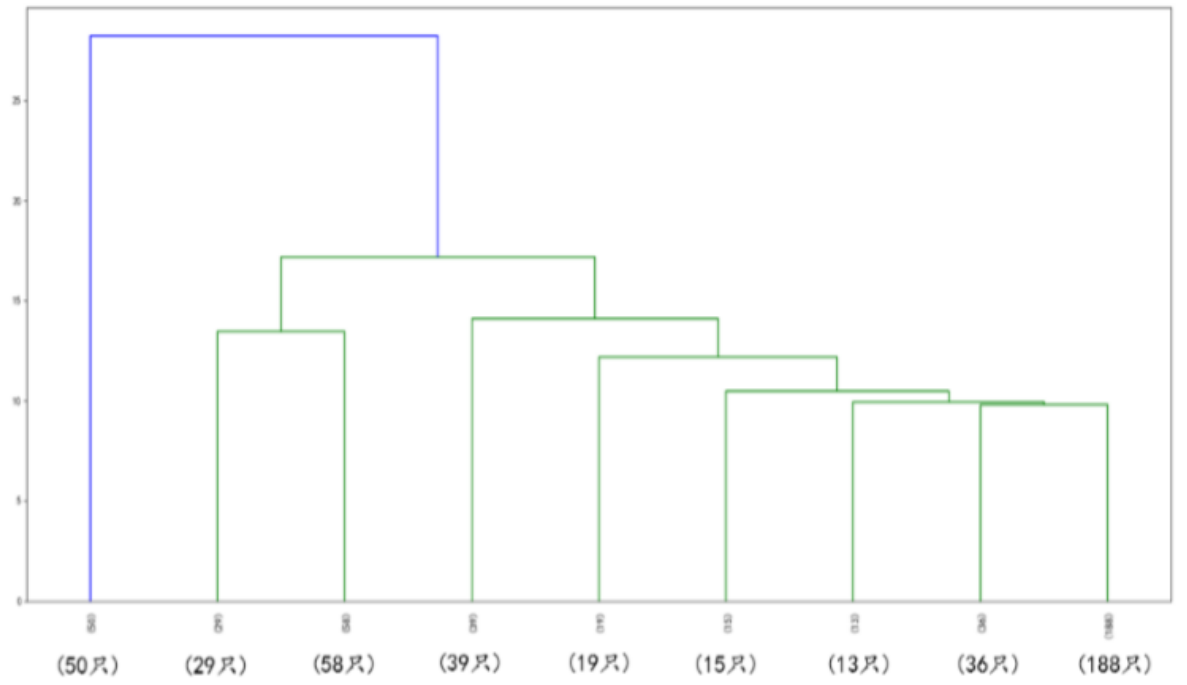


资料来源：Wind，华泰证券研究所

聚类算法案例：基于股票产业概念的聚类

- 层次聚类可以通过分层的方式显示各个类别的联系。图 15 展示了中证 500 成分股的层次聚类效果图。

图表15： 中证 500 成分股层次聚类图



资料来源：Wind，华泰证券研究所

聚类算法案例：基于股票产业概念的聚类

- 统计各聚类簇中概念出现次数，可得到图表16和图表17中各聚类簇的概念词云，概念的字体越大，表示簇中含有该概念的股票越多。
- 可以看出，聚类簇中的概念具有高度相似性，说明层次聚类将具有相似概念的股票聚到了一起。

图表16：沪深300成分股层次聚类簇概念词云

资料来源：Wind，华泰证券研究所

图表17：中证500成分股层次聚类簇概念词云

资料来源：Wind，华泰证券研究所

聚类算法案例：基于股票产业概念的聚类

- 图表 18 展示了沪深300成分股各个聚类中相似度较高的一些股票。

图表18：沪深 300 层次聚类

股票名称	股票概念	所属聚类	相似度
智飞生物	干细胞;肺炎概念;创新药;生物疫苗;国产化创新;大消费	2	0.77
沃森生物	大消费;国产化创新;流感;肺炎概念;生物疫苗;血液制品;创新药		
华东医药	仿制药;医保概念;抗生素;国产化创新;创新药;医疗改革	2	0.72
恒瑞医药	大消费;国产化创新;医疗改革;干细胞;抗癌;医保概念;创新药;仿制药		
深南电路	制造业单项冠军企业;珠三角;消费电子产业;科技龙头;华为概念;深圳;高价股;电路板;手机产业;基站;半导体材料;5G	4	0.71
生益科技	聚酰亚胺;科技龙头;5G;珠三角;华为概念;电路板;基站;消费电子产业		
韦尔股份	华为概念;芯片国产化;集成电路;半导体产业;消费电子产业;浦东新区;科技龙头;5G 应用;TWS 耳机;半导体分立器件;摄像头;虚拟现实	4	0.56
兆易创新	高价股;消费电子产业;集成电路;数字中国;存储器;芯片国产化;半导体产业;华为概念;国产软硬件;科技龙头;手机产业;TWS 耳机;出口型企业		
中航飞机	航空发动机;高端装备制造;十大军工集团;航母;通用航空;大飞机	7	0.83
航发动力	通用航空;十大军工集团;高端装备制造;大飞机;军民融合;航母		
通威股份	新能源;光伏;国产化创新;异质结电池(HIT);CDM 项目	9	0.68
隆基股份	新能源;国产软硬件;国产化创新;中非合作概念;异质结电池(HIT);数字中国;光伏		

资料来源：Wind，华泰证券研究所

聚类算法案例：基于股票产业概念的聚类

- 图表 19 展示了中证500成分股各个聚类中相似度较高的一些股票。

图表19： 中证 500 层次聚类

股票名称	股票概念	所属聚类	相似度
深天马 A	5G 应用;超高清视频;国产化创新;小米产业链;国产软硬件;科技龙头;触摸屏;手机屏幕;液晶显示;华为概念;Micro LED;珠三角;虚拟现实;手机产业;新型显示技术;智能手表;OLED;消费电子产业	2	0.75
维信诺	手机屏幕;新型显示技术;华为概念;智能手表;消费电子产业;出口型企业;手机产业;5G 应用;超高清视频;小米产业链;OLED;液晶显示		
太极股份	华为鲲鹏;金融科技;网络可视化;科技龙头;核高基;智慧城市;国产软硬件;云计算;自主可控;电子政务;工业互联网;5G 应用;操作系统;网络安全;十大军工集团;华为概念	2	0.76
中国软件	知识产权;自主可控;核高基;华为概念;华为鲲鹏;科技龙头;电子政务;消费电子产业;国产软硬件;云计算;十大军工集团;操作系统;5G 应用		
浦东金桥	迪士尼;上海国资改革;上海市国资;上海自贸区;浦东新区	3	0.91
外高桥	上海国资改革;上海自贸区;创投;浦东新区;上海市国资;迪士尼		
航发控制	大飞机;十大军工集团;高端装备制造;航母;航空发动机;通用航空;军民融合	5	0.93
洪都航空	无人机;大飞机;高端装备制造;通用航空;军民融合;航空发动机;航母;十大军工集团		
浙江医药	超级细菌;医保概念;肺炎概念;维生素;抗生素;医疗改革	5	0.67
哈药股份	医保概念;医疗改革;抗生素;肺炎概念;工业大麻;东北振兴		

资料来源：Wind，华泰证券研究所

华泰证券
HUATAI SECURITIES

04

无监督学习应用于因子投资

无监督学习应用于因子投资——PCA算法准确估计因子溢价

- 论文：Asset Pricing with Omitted Factors，来自耶鲁大学的Stefano Giglio和芝加哥大学的Dacheng Xiu。

- 论文主要研究了如何在有遗漏变量的情况下使用PCA进行更精确的因子溢价估计，特别是针对通胀率、GDP等宏观因子，这些因子属于不可交易因子(nontradable factors)。

- 对于不可交易因子，经典的因子溢价估计方法包括Fama-MacBeth回归和因子模拟组合方法(mimicking-portfolio approach)，但这两种方法都会面临以下两个问题：
  1. 估计结果会随着控制变量的变化而改变，例如当选取的控制变量为市场因子和Fama三因子时，会得出不一样的因子溢价。
  2. 遗漏控制变量问题，经典Fama三因子是从经济学角度提取的描述市场共性的因子，然而市场的复杂性可能导致一些潜在的定价因子难以通过人脑构造得出，从而遗漏控制变量。

- 基于以上问题，论文创造性地提出了一种无需观测到全部真实因子便可准确估计因子溢价的方法：Three-Pass Estimator，可以有效解决资产定价模型中遗漏变量和测量误差的问题。

无监督学习应用于因子投资——PCA算法准确估计因子溢价

- Giglio与Xiu提出三个步骤得到γ_g的估计值：

1. 设n为资产数目，T为截面数，R是大小为$n*T$的超额收益矩阵，$\bar{R}$是R去均值后的矩阵。使用PCA算法从$n^{-1}T^{-1}\bar{R}^{-1}\bar{R}$矩阵中提取主成分：

$$\hat{V} = T^{\frac{1}{2}}(\xi_1 : \xi_2 : \ldots : \xi_p)^T$$

其中$(\xi_1 : \xi_2 : \ldots : \xi_p)$为矩阵的前p个主成分，并得到系数$\hat{\beta} = T^{-1}\bar{R}\hat{V}^T$。
为方便计算将$\hat{V}$标准化：$\hat{V}\hat{V}' = I_{\hat{p}}$；

2. 截面回归：用平均收益$\bar{r}$对潜在因子暴露$\hat{\beta}$进行截面回归，得到平均收益和主成分的回归系数，即主成分因子的因子溢价 $\hat{\gamma}$：

$$\hat{\gamma} = \left(\hat{\beta}^T\hat{\beta}\right)^{-1}\hat{\beta}^T\bar{r}$$

3. 时序回归：设G是大小为$d*T$的可观测因子矩阵，d为需要估计因子溢价的可观测因子数目，$\bar{G}$为G去均值后的矩阵，通过时序回归$\bar{G} = \hat{\eta}\hat{V}$可得到$\hat{\eta}$：

$$\hat{\eta} = \bar{G}\hat{V}^T\left(\hat{V}\hat{V}^T\right)^{-1}$$

- 最终，可观测因子的因子溢价为：

$$\hat{\gamma}_g = \hat{\eta}\hat{\gamma} = \bar{G}\hat{V}^T\left(\hat{V}\hat{V}^T\right)^{-1}\left(\hat{\beta}^T\hat{\beta}\right)^{-1}\hat{\beta}^T\bar{r}$$

无监督学习应用于因子投资——PCA算法准确估计因子溢价

- 实证部分中，Giglio与Xiu采用了647个资产1976~2010年的月频数据，资产类型包括美国股票、各类债券和外汇。包含的资产类型越多，其覆盖到的风险类型越丰富，构建的“风险空间”就越完整。

- 待评估的资产定价因子包括：

1. 可交易因子：市场因子(Market)、规模因子(SMB)、价值因子(HML)、盈利因子(RMW)、投资因子(CMA)、动量因子(MOM)、押注β因子(BAB，Frazzini 和Pedersen (2014))、质量因子(QMJ，Asness et al. (2013))

2. 不可交易因子：

工业产值增长的AR(1)新息(IP growth)、流动性因子(Liquidity)

279个宏观变量前三个主成分的VAR(1)新息(Macro PC1-3，Ludvigson 和Ng (2010))

2个中间资本因子(Interm.(He)，He et al. (2017)以及Interm.(Adrian)，Adrian et al. (2014))

4个来自Novy-Marx (2014)的因子(NY temp、Global temp、El Niño和Sunsplots)

2个基于消费的因子(Cons. Growth和Stockholder cons.，Malloy et al. (2009))

无监督学习应用于因子投资——PCA算法准确估计因子溢价

- 论文中因子溢价估计结果如图20：

图表20：因子溢价估计结果

Table 1: Three-Pass Regression: Empirical Results

Factors	Avg. Ret. γ	Avg. Ret. stderr	two-pass no controls γ	two-pass no controls stderr	two-pass w/ R_m γ	two-pass w/ R_m stderr	two-pass w/ FF3 γ	two-pass w/ FF3 stderr	Mimick.-portf. w/ R_m γ	Mimick.-portf. w/ R_m stderr	Mimick.-portf. w/ FF3 γ	Mimick.-portf. w/ FF3 stderr	three-pass regression γ	three-pass regression stderr	R^2_g	p-value g weak
Market	0.51**	(0.23)	0.56**	(0.23)	0.56**	(0.23)	0.51**	(0.22)	0.51**	(0.22)	0.51**	(0.22)	0.51**	(0.23)	99.57	0.00
SMB	0.25	(0.15)	0.82**	(0.34)	0.07	(0.16)	0.10	(0.16)	0.08**	(0.04)	0.25	(0.16)	0.20	(0.16)	97.24	0.00
HML	0.35**	(0.17)	−0.85**	(0.38)	0.30*	(0.18)	0.35**	(0.17)	−0.13**	(0.06)	0.35**	(0.15)	0.20	(0.15)	83.03	0.00
MOM	0.69***	(0.24)	−2.01**	(0.88)	0.20	(0.26)	0.71***	(0.24)	−0.05	(0.05)	−0.21*	(0.11)	0.49**	(0.23)	89.82	0.00
RMW	0.38***	(0.13)	0.04	(0.16)	−0.00	(0.17)	0.27**	(0.13)	−0.07**	(0.03)	−0.09	(0.06)	0.22*	(0.11)	71.48	0.00
CMA	0.32***	(0.11)	−0.59**	(0.24)	0.34**	(0.14)	0.42***	(0.12)	−0.10**	(0.05)	0.12	(0.08)	0.14	(0.10)	59.03	0.00
BAB	0.94***	(0.22)	−1.59*	(0.85)	1.10***	(0.29)	1.21***	(0.27)	−0.06	(0.05)	0.23**	(0.11)	0.57***	(0.15)	47.43	0.00
QMJ	0.44***	(0.14)	−0.50**	(0.21)	0.01	(0.16)	0.25*	(0.14)	−0.15**	(0.07)	−0.29***	(0.09)	0.06	(0.13)	84.29	0.00
Liquidity			2.26**	(0.90)	3.44***	(1.09)	0.57	(0.68)	0.21*	(0.11)	0.32**	(0.14)	0.37**	(0.16)	12.11	0.00
Interm. (He)			1.01**	(0.45)	0.19	(0.49)	0.43	(0.45)	0.57**	(0.25)	0.78***	(0.27)	0.60**	(0.31)	69.05	0.00
Interm. (Adrian)			1.37***	(0.30)	1.52***	(0.28)	1.58***	(0.27)	0.10*	(0.06)	0.61***	(0.15)	0.72***	(0.16)	51.99	0.00
NY temp.			−319.01	(255.73)	125.89	(152.76)	−277.96**	(124.08)	−2.35	(5.42)	10.71	(10.94)	−0.69	(13.90)	0.76	0.84
Global temp.			−6.65	(4.85)	−5.29	(4.92)	−3.33	(2.07)	−0.01	(0.09)	0.11	(0.17)	0.05	(0.21)	2.21	0.09
El Niño			56.85***	(17.42)	19.23*	(11.08)	−15.34**	(7.11)	0.39	(0.33)	0.94	(0.59)	0.41	(0.82)	1.58	0.43
Sunspots			−409.37	(937.73)	1637.60***	(467.40)	882.89**	(405.40)	−19.30	(19.49)	−4.33	(30.42)	4.01	(35.63)	0.86	0.72
IP growth			−0.36**	(0.14)	−0.27***	(0.07)	−0.14***	(0.05)	−0.00	(0.00)	−0.01	(0.01)	−0.01*	(0.00)	2.25	0.21
Macro PC 1			84.90***	(24.76)	87.26***	(20.95)	39.96***	(13.57)	1.22	(0.75)	2.49*	(1.43)	3.26**	(1.58)	2.34	0.29
Macro PC 2			9.35	(15.93)	9.28	(16.34)	23.91***	(8.97)	−0.91	(0.59)	−2.05**	(1.03)	−0.88	(1.27)	4.05	0.09
Macro PC 3			−5.94	(14.30)	−6.70	(12.11)	−31.24***	(9.74)	−0.99	(0.64)	−0.61	(1.21)	−1.25	(1.51)	6.60	0.01
Cons. growth			0.26*	(0.16)	−0.03	(0.11)	0.07	(0.05)	−0.00	(0.00)	−0.00	(0.01)	0.00	(0.01)	4.07	0.07
Stockholder cons.			6.26***	(2.14)	2.48**	(1.20)	1.08*	(0.58)	0.05	(0.04)	0.03	(0.06)	0.17**	(0.08)	2.50	0.32

Note: For each factor, the table reports the risk premia estimates using different methods, with the restriction that the zero-beta rate is equal to the observed T-bill rate: "Avg. Ret.", the time-series average return of the factor, available when the factor is tradable; three versions of the two-pass cross-sectional regression, using no control factors in the model, using the market, and using the Fama-French three factors, respectively; two versions of the mimicking-portfolio estimator, projecting factors onto the market portfolio and the Fama-French three factors (given that we have more portfolios than observations, it is not feasible to use the mimicking-portfolio approach with all test portfolios); the three-pass estimator we propose in this paper, using $\hat{p} = 7$ latent factors; the R^2 of the projection of g_t onto the latent factors; and the p-value of the test that factor g_t is weak.

无监督学习应用于因子投资——PCA算法准确估计因子溢价

- 图20展示了三个模型下每个因子的风险溢价。该表各列的含义如下：
  1. 第一列展示了可交易因子的时序平均收益来作为对照标准。对于不可交易因子，无法计算其时序平均收益。
  2. 第二到第四列展示了Fama-MacBeth回归估计的因子溢价。分别为采用无控制变量、市场因子作为控制变量和Fama三因子作为控制变量估计的因子溢价。
  3. 第五列和第六列展示了因子模拟组合法估计的因子溢价。分别为采用市场因子作为控制变量和Fama三因子作为控制变量估计的因子溢价。
  4. 第七列展示了论文提出的三步法(Three-Pass Estimator)得到的因子溢价。这里设定PCA主成分为7，即有7个潜在因子。第8列是将可观测因子映射到潜在因子上的R^2。第9列为可观测因子是弱因子的检验p值。

- 图20显示，三步法得到了更符合逻辑的结果，其估计的可交易因子的溢价和因子本身的平均收益(第一列)接近。随着控制变量的变化，Fama-MacBeth回归法和因子模拟组合法所估计的因子溢价会有变化，甚至有些因子溢价的符号与第一列中的结果符号相反，说明这两种方法表现欠佳。

- 结果还显示，一些标准的宏观因子(如Macro PC2等)没有显著的风险溢价，而与市场摩擦相关的因子(如流动性因子和Interm.(He)和Interm.(Adrian))有很显著的风险溢价。

感谢聆听，欢迎您多提意见！

- 个人微信：李子钰-华泰金工
- 无法扫码可手动搜索：**18924616742**
- 加好友麻烦您备注公司名称+个人姓名谢

- 华泰金工团队公众号
- 内含华泰金工所有深度报告，欢迎扫描关注！