金融工程

证券研究报告

2020 年 10 月 21 日

海外文献推荐 第 154 期

作者

吴先兴 分析师

SAC 执业证书编号：S1110516120001

wuxianxing@tfzq.com

相关报告

1 《金融工程：金融工程-市场情绪一览 2020-10-20》 2020-10-20

2 《金融工程：金融工程-市场情绪一览 2020-10-19》 2020-10-19

3 《金融工程：金融工程-FOF 组合推荐周报：上周双鑫 ETF 组合超额收益达 1.25%》 2020-10-19

异象策略的相关性结构

本文从大量异象中筛选出显著的异象，并对其进行聚类，合并为 28 个异象聚类组合。在五因子模型（How 等，2020）下，仍然有三分之一的聚类组合呈现显著的收益。通过最佳优先搜索算法，从现有因子与聚类组合中寻找到 9 个因子，能够解释全部的聚类组合以及显著的异象。其中，预期增长因子（EG）和与应计项目相关的聚类组合是提高模型定价能力的重要因素。

风险提示：本报告基于相关文献，不构成投资建议。

内容目录

图表目录

异象策略的相关性结构

文献来源： Paul Geertsema, Helen Lu, The correlation structure of anomaly strategies, Journal of Banking and Finance, 2020, vol. 119

推荐原因：本文从大量异象中筛选出显著的异象，并对其进行聚类，合并为 28 个异象聚类组合。在五因子模型（How 等，2020）下，仍然有三分之一的聚类组合收益显著。通过最佳优先搜索算法，从现有因子与聚类组合中寻找到 9 个因子，能够解释全部的聚类组合以及显著的异象。其中，预期增长因子（EG）和与应计项目相关的聚类组合是提高定价能力的重要因素。

1. 简介

许多研究通过提出新的资产定价模型来解释逐渐增多的异象。实际上，这些新提出的风险因子最初就以异象的形式表现。例如，Fama & French（1993）三因子模型中的 SMB 因子，便是基于 Banz（1981）中记录的规模异象。本文利用学术界所发现的大量异象，提出了一种基于异象之间的相关性将异象聚类的新方法，来识别实际收益中没有被当前资产定价模型所包含的维度。

本文从现有文献中构建了包含 215 种异象策略的数据库。我们的异象策略是用美股构建的多空组合，以分位数分组并以市值加权构建组合。根据 Hou 等（2015），我们使用 5%显著性水平来挑选异象进行进一步研究。通过这种方法，可以得到 80 个在均值上显著的异象。我们使用聚类分析将这些筛选后的异象划分为 28 个等权的聚类组合。

我们使用基于相关性的相异性指标，运用层次聚类（hierarchical agglomerative clustering）的聚类方法。在本文中，我们希望将显著相关的异象划分为同一类，其中相关性超过 0.4 定义为显著相关。我们发现 28 个集合能够在给定阈值下将异象的错分类水平降至最低，聚类组合的平均相关性为 0.03 而组间相关性最高为 0.57。

聚类组合可以用以识别当前基准模型中遗漏的维度，从而构建更加准确、精确的业绩评价基准。即使使用最严格的模型——五因子模型（Hou 等，2020）仍然能够发现 28 个聚类组合中有 10 个在 5%水平下显著。由于当前的基准模型不能解释这些聚类组合，我们从 41 个待选因子中寻找其他因子以解释收益。通过最佳优先搜索（Best-First Search）算法，我们发现了 9 个因子，能够解释所有 28 个聚类组合及 80 个显著的异象。这 9 个因子分别为预期成长（expected growth）因子（Hou 等,2020），应计聚类组合（Accruals，如 Sloan,1996），SMB 市值因子（Fama & French,1993），发行与收益率溢价聚类组合（如 Basu,1977，Danial&Titman,2006），市场因子，短期反转因子，季节效应聚类组合，资本增长（CapexGrowth）聚类组合（如 Xin,2008），EPS 持续性（epsconsistency）聚类组合（长期盈利增长异象，Alwathainani, 2009)）。

2. 数据

本文使用了在 NYSE, Amex 或者 Nasdaq 交易的普通股构建异象策略，得到每月收益。样本期从 1963 年 7 月开始，到 2019 年 12 月结束。

异象生成方面，作者使用分位数分组并以市值加权构建组合，生成月收益。215 个预测排序变量主要分为企业特征、股票变量或者宏观经济因子载荷。

下图表 A 展示了 215 种策略的概要信息，表 B 给出了过滤后的 80 个异象数据的汇总统计数据，表 C 中的集群投资组合则是通过计算同一集群中的等权平均收益来构建。如预期的那样，经过过滤的异象非常显著，平均月度收益的均值为 0.49%。相关性方面，平均而言，异象之间只有弱正相关。但大多数异象与至少一种异象具有强相关关系。为了处理强相关异象，我们可以将它们分组在同一个聚类组合中。

图 1：异象统计

Table 1
Cross-sectional means of anomaly time-series statistics.
The tables below report the means of important time-series statistics across anomalies. Panel A uses the initial dataset that contains all anomalies. Panel B uses the filtered dataset which include mean significant anomalies (oriented to produce positive mean returns) with *t*-statistics above 1.96. Panel C contains cluster portfolios which consolidate filtered anomalies in Panel B into equal-weighted portfolios by cluster. "Count of observations" is the number of monthly return observations in an anomaly time-series. "Mean monthly return" is the average monthly return of an anomaly time-series in percentage points. "Return volatility" is the times-series standard deviation of monthly anomaly returns. "Mean correlation" is the average of all correlation coefficients for one anomaly with other anomalies in the dataset used in each panel. "Min correlation" and "Max correlation" are defined similarly. The column headings "Average", "SD", "Min" and "Max" indicate cross-sectional statistics taken across all time-series contained in each dataset.

Panel A: Initial dataset (215 anomalies)

	Cross-sectional summary statistics				
Time-series statistic	*N*	Average	SD	Min	Max
Count of observations	215	649.88	66.18	378.00	678.00
Mean monthly return	215	0.01	0.36	−0.78	1.12
Return volatility	215	4.36	1.27	2.17	7.94
Mean correlation	215	0.03	0.06	−0.11	0.15
Min correlation	215	−0.56	0.21	−0.96	−0.18
Max correlation	215	0.72	0.20	0.21	1.00

Panel B: Filtered dataset (80 anomalies)

	Cross-sectional summary statistics				
Time-series statistic	*N*	Average	SD	Min	Max
Count of observations	80	666.59	29.08	570.00	678.00
Mean monthly return	80	0.49	0.16	0.22	1.12
Return volatility	80	4.01	1.12	2.41	7.31
Mean correlation	80	0.06	0.05	−0.07	0.14
Min correlation	80	−0.36	0.12	−0.63	−0.09
Max correlation	80	0.66	0.20	0.18	0.97

Panel C: Clusters dataset (28 cluster portfolios)

	Cross-sectional summary statistics				
Time-series statistic	*N*	Average	SD	Min	Max
Count of observations	28	672.11	22.46	570.00	678.00
Mean monthly return	28	0.45	0.13	0.22	0.78
Return volatility	28	3.27	0.82	2.24	5.48
Mean correlation	28	0.03	0.04	−0.07	0.10
Min correlation	28	−0.30	0.12	−0.49	−0.08
Max correlation	28	0.38	0.12	0.13	0.57

资料来源：Journal of Banking and Finance，天风证券研究所

3. 聚类分析

聚类分析是无监督机器学习的一个分支，用于根据某种相似性概念将实体划分为不同的组。本文使用了聚类分析中的层次聚类。在开始时，每个异象被分配到它自己的集群。在这一点上，有多少异象就有多少集群——在过滤异象的情况下有 80 个。然后，在每次迭代时，根据指定的条件将两个差异最小的聚类合并到一个新的聚类中。每次迭代都减少一个簇的数量，这样在聚类过程结束时，所有异象都被分配到同一个簇中。

我们选择了 N=80 个异象，异象集合为$X = \{x_1, \dots, x_N\}$，相应的收益率序列为$\{r_1, \dots, r_N\}$。

则异象x_i和x_j之间相关性可定义为：

$$\rho_{i,j} := \frac{\sum_{t\in\mathcal{T}}(r_{i,t}-\bar{r}_i)(r_{j,t}-\bar{r}_j)}{\sqrt{\sum_{t\in\mathcal{T}}(r_{i,t}-\bar{r}_i)^2}\sqrt{\sum(r_{j,t}-\bar{r}_j)^2}} \quad (1)$$

其中，$|\mathcal{T}|$表示集合$\mathcal{T}$中元素的个数，$\bar{r}_i := \frac{1}{|\mathcal{T}|}\sum_{t\in\mathcal{T}} r_{i,t}$

随后，将相关性转化为相异性度量，定义为：

$$d(x_i, x_j) := (1-\rho_{i,j})/2 \quad (2)$$

其中，$d(x_i, x_j)$的范围为$[0,1]$，$d=0$对应完全正相关$(\rho=1)$，$d=1$对应完全负相关$(\rho=-1)$。通过指定异象间的相异性，可以将强相关异象聚在一组，而将弱相关异象划分在不同的组。

我们使用平均关联方法（average linkage method）衡量两个类别之间的关系。将 X 的两个不重叠子集C_a和C_b（$C_a, C_b \subset X$，$C_a \cap C_b = \varnothing$)之间的平均联系 L(·)定义为

$$L(C_a, C_b) = \frac{\sum_{x\in C_a}\sum_{y\in C_b} d(x,y)}{|C_a||C_b|} \quad (3)$$

注意，在特殊情况下$C_a = x_a$和$C_b = x_b$平均连杆$L(C_a, C_b)$ 等于 $d(x_a, x_b)$

迭代 k = 1 时，初始集群划分包括 N 个集群，记为$C_1, \ldots, C_N$，每个集群包含一个(唯一的)元素。有：

$$P_1 = \{C_1, \ldots, C_N\} = \{\{x_1\}, \ldots, \{x_N\}\} \quad (4)$$

迭代 k > 1 处的P_k被递归定义为

$$P_k = \left(P_{k-1} / \{C_a^{k-1}, C_b^{k-1}\}\right) \cup \{C_a^{k-1} \cup C_b^{k-1}\} \quad (5)$$

其中C_a^{k-1}和 C_b^{k-1}这两个簇是P_{k-1}中满足要求的最不相似的集群，而

$$L(C_a^{k-1}, C_b^{k-1}) = \inf\{L(C_x, C_y) \,|\, C_x, C_y \in P_{k-1} \text{ and } C_x \neq C_y\} \quad (6)$$

因此，在每一次迭代中，两个最不相似的集群C_a^{k-1}和C_b^{k-1}在前一个集群分区P_{k-1}.中被识别出来。通过删除C_a^{k-1}和C_b^{k-1}，加上由C_a^{k-1}和C_b^{k-1}并集形成的新集群，形成新的集群分区P_k。这些集群都是在 X 的元素的基础上形成的集合，每个 X 的元素都出现在给定集群分区中的一个且仅一个集群中。集群分区的序列$S = P_1, \ldots, P_N$被称为集群结构。上述描述也可以表示为一种算法;具体细节如图 2 所示：

图 2：聚类算法

Algorithm 1 Hierarchical Agglomerative Clustering (HAC)

Require: $X, L(\cdot)$ ▷ Set of entities X and linkage method $L(\cdot)$
Ensure: $S = P_1, ..., P_N$ ▷ Cluster structure is a sequence of cluster partitions

procedure HAC($X, L(\cdot)$)
  $P_1 \leftarrow \{\{x_1\}, ..., \{x_N\}\}$ ▷ Initial cluster partition
  for k = 2 to N **do**
    $P_k \leftarrow P_{k-1}$ ▷ Copy the previous partition
    Find C_a^{k-1} and $C_b^{k-1} \in P_{k-1}$ such that $L(C_a^{k-1}, C_b^{k-1})$ is a minimum
    $C' \leftarrow C_a^{k-1} \cup C_b^{k-1}$ ▷ Create a new cluster C'
    Remove C_a^{k-1} and C_b^{k-1} from P_k
    Add C' to P_k
  end for
  $S = P_1, ..., P_N$ ▷ Cluster structure S contains N cluster partitions
  return S
end procedure

资料来源：Journal of Banking and Finance，天风证券研究所

生成集群组合需要确定适当的集群数量。为了让组合能够实现某种程度的独立性，集群的最佳数量应该能最大限度地减少对异象的误分类。考虑到一些因子具有中度甚至实质性的相关性，文章取相关性阈值为 0.4。在发生误分类时，如果成对相关性低于阈值却将一对不同的异象分配给同一聚类，称为假阳性；而若成对相关性高于阈值却将一对不同的异象分配给不同的聚类，则称为假阴性。因此，将 X 上的实体在给定相关性阈值上的对群集P_k的误分类计数定义为：

$$M_{k,\rho^*} = \text{Count of False Positive Pairs} + \text{Count of False Negative Pairs} = \left(\sum_{C\in P_k}\sum_{x_i,x_j\in C; i\neq j} 1_{(\rho_{i,j}<\rho^*)}\right) + \left(\sum_{(x_i,x_i)\in X'} 1_{(\rho_{i,j}>\rho^*)}\right) \quad (7)$$

上式中，$X' = \{(x_i, x_j): x_i, x_j \in X \text{ and } i \neq j \text{ and } (\forall_{C\in P_k}: i \in C \Longrightarrow j \notin C\}$，是分区$P_k$中所有不在同一簇内的异象对的集合。条件 x 为真时，指示函数$1_{(x)} = 1$，否则为 0。

下图绘制了假阳性数、假阴性数和误分类总数相对于聚类数在相关阈值为 0.4 时的图。结果表明，集群组合个数为 28 时，在 3160 对相关中，错误分类的计数最小，为 236。

图 3：聚类数量

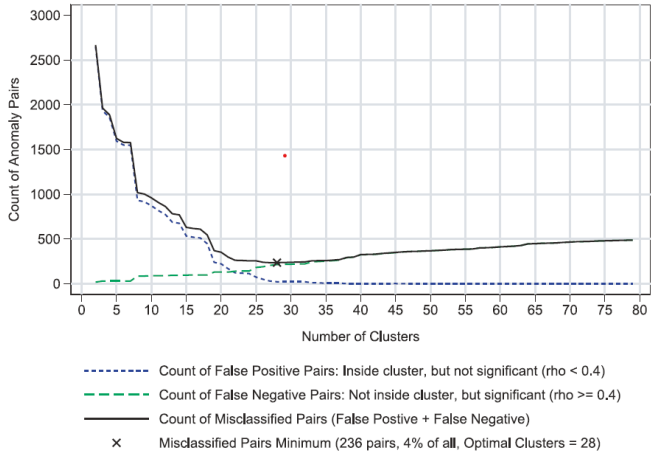

Fig. 1. Number of clusters that minimises misclassification when the significance threshold for pairwise correlation is $\rho > 0.4$. The figure above plots the count of false positive pairs, false negative pairs and misclassified pairs (y-axis) against the number of clusters (x-axis). A false positive pair is a pair of anomalies that are allocated to the same cluster but have a pairwise correlation below 0.4 (the significance threshold for *clusters*). A false negative pair is a pair of anomalies that are allocated to different clusters but have a pairwise correlation above the significance threshold. A *misclassified* pair is either a false positive pair or a false negative pair. The optimal number of clusters minimises the number of misclassified pairs, marked with an "X" in the figure above. The analysis is based on the filtered anomalies described in Panel B of Table 1. The cluster partitions are constructed using a "one-minus" dissimilarity measure and an average linkage method, as described in Section 3.1.

资料来源：Journal of Banking and Finance，天风证券研究所

4. 收益的维度

4.1. 聚类组合的 alpha 与无法解释的方差

在本节中，本文进行了标准时间序列的 alpha 检验，以衡量当前的基准模型是否可以解释异象和聚类组合。下图列出了六种不同基准模型和测试中使用的三个数据集，在显著性水平不断增加时，具有显著 alpha 的异象或集群组合的个数。其中，表 A 选择所有异象作为测试样本，表 B 与 C 的测试样本分别为 80 个过滤异象和 28 个集群组合。

测试结果表明，对于不同的样本，HXZ4 和 HMXZ5 模型的表现均为最佳，它们是消除显著 alpha 的两个最佳模型。此外，在表 C 的检验中，即使是表现最好的 HMXZ5 模型，在显著性水平 5%情况下，也有 10 个（36%）集群组合结果显著，这意味着可以向当前的基准模型添加更多维度。

在给定的显著性阈值下，集群的显著性 alpha 值的百分比要高于异象策略。这表明，当异象本身强相关时，所解释异象的比列可能会夸大基准模型的能力。从这个意义上说，集群投资组合可能提供了一个更均衡的基准。

图 4：显著的 Alpha 数量

Table 5
Count of significant alphas.
The tables below report the count of significant mean returns ("mean") and significant benchmark alphas under the CAPM, FF3, HZX4, FF5, FF6, and HMXZ5 models for three datasets. Refer to Table 1 for a description of the three datasets. The column headers of each table indicate statistical significance: "not sig" means not significant at a 10% level; "$p<10\%$", "$p<5\%$" and "$p<1\%$" indicates p-values below 10%, 5% or 1% respectively; "$|t|>3$" indicates an absolute t-statistic above 3, or equivalently, a p-value below 0.27%. Standard errors are Newey-West HAC adjusted with a lag of 6 months.

Panel A: All anomalies (215 anomalies)

Benchmark	Not Sig	$p<10\%$	$p<5\%$	$p<1\%$	$\|t\|>3$	Total
MEAN	114	101	80	47	35	215
CAPM	90	125	106	86	70	215
FF3	73	142	125	91	77	215
HXZ4	144	71	52	39	28	215
FF5	101	114	101	60	50	215
FF6	132	83	68	49	34	215
HMXZ5	161	54	40	19	12	215

Panel B: Filtered dataset (80 anomalies)

Benchmark	Not Sig	$p<10\%$	$p<5\%$	$p<1\%$	$\|t\|>3$	Total
MEAN	0	80	80	47	35	80
CAPM	2	78	72	63	55	80
FF3	11	69	64	56	49	80
HXZ4	37	43	34	28	23	80
FF5	27	53	49	37	34	80
FF6	36	44	40	31	22	80
HMXZ5	54	26	24	12	8	80

Panel C: Cluster portfolio dataset (28 cluster portfolios)

Benchmark	Not Sig	$p<10\%$	$p<5\%$	$p<1\%$	$\|t\|>3$	Total
MEAN	0	28	28	23	19	28
CAPM	0	28	28	23	20	28
FF3	3	25	25	22	18	28
HXZ4	10	18	16	13	9	28
FF5	5	23	21	16	13	28
FF6	11	17	17	14	11	28
HMXZ5	16	12	10	6	4	28

资料来源：Journal of Banking and Finance，天风证券研究所

4.2. 降维搜索

我们识别了备选因子中能够最好地解释所有 28 个集群投资组合的子集。候选因子一共 41 个，包括 28 个聚类组合本身加上 13 个共同因子(MKTRF、SMB、HML、RMW、CMA、UMD、ROE、EG、LTREV、STREV、QMJ、BAB 和 HMLdevil)。本文考虑两个目标：第一个目标是减少 alpha 显著的集群投资组合数量，而第二个目标是减少集群投资组合平均的无法解释的方差$(1-R^2)$。

本文选择了一种贪婪（最佳优先）搜索算法，该策略在每个步骤中都选择能最大限度地减少目标的额外因子。其中，目标可设为显著集群投资组合的数量或平均无法解释的方差，具体视情况而定。在模型中添加额外的因子总是会减少无法解释的方差，直至最后为零。然而，通常情况下，不能通过向当前模型中添加额外的因子来进一步减少显著的集群投资组合的数量。当发生这种情况时，切换到第二阶段搜索策略，该策略将最强的集群投资组合添加到模型中（以 alpha 的 t 统计量衡量）。

最终，减少 alpha 的工作确定了组成模型的 9 个因子，它们可以解释 28 个集群组合。这些因子包括 4 个共同因子和 5 个集群组合；按顺序排列，它们分别是，1）HMXZ5 模型的 EG（预期增长）因子，2）Accruals 集群投资组合，3）FF3 模型中的 SMB 规模因子，4）IssuanceAndYield 集群投资组合，5）MKTRF（市场超额收益因子），6）STREV（短期反转因子），7）seasonality 集群投资组合，8）CapexGrowth 集群投资组合和 9）epsconsistency 集群投资组合。我们验证了，选择的 9 个因子可以将筛选的 80 个异象减少到了不重要的

程度，这表明本文所阐述的 28 个聚类组合是对 80 个筛选的原始异象的更简介的表达。

需要强调的是，还有其他候选因子集合也能够解释所有的集群组合，我们的搜索结果只证明了上界的存在，可能还存在更简单的模型。

文章还考虑了从预先选定的基准模型的因子开始进行降维搜索。这种分析能够突出在描述异象回报横截面方面能够提供增量的集群投资组合。其中，EG 因子和 Accruals 集群投资组合影响力最大，它们在一起解释了 28 个集群投资组合中的 18 个(未报告的结果)——比任何一个基准模型都多。在所有情况下，STREV 因子或 STReversal 集群组合(包含 STREV 因子)也会被选择。最后，季节性和 epsconsistency 集群组合也经常选入。这些集群投资组合是高度独立的，将其包含到模型中似乎是解释其的唯一方法。

此外，本文还考虑第二个目标，即哪些因子最适合减少无法解释的方差。减少方差的搜索和减少 alpha 的搜索会生成稍微不同的因子排名，结果如下图中的表 B 所示。通过减少方差搜索选择的前九个因子是来自 HXZ4 模型的 ROE（获利能力）因子，其次是 IssuanceAndYield， AssetGrowth， Accruals， MarginGrowth， termbeta， coskewness, STReversal 和 pchquickratio 集群投资组合。

总而言之，一个相对简洁的九因子模型包含了全部 28 个集群组合以及所有 80 个显著的异象。当增加因子扩充当前的基准模型时，几乎总是首先选择 EG（预期增长）因子和 Accruals 集群投资组合。

图 5：降维

Table 8
Dimensionality reduction.
The tables below report dimensionality reduction search results. The sample runs from July 1972 to December 2019, or the maximum period during which all cluster portfolios and common factors have non-missing returns. The pool of candidate factors comprises the 28 cluster portfolios and 13 common factors (MKTRF, SMB, HML, RMW, CMA, UMD, ROE, EG, LTREV, STREV, QMJ, BAB, and HMLdevil). We initially search without pre-selecting a model (column heading "Start with nothing") and then search after pre-selecting models (models shown in columns headings). The factors in pre-selected models are in bold font. The search process iteratively augments the current model with an additional factor that reduces the number of significant cluster portfolio alphas ($|t_\alpha| > 1.96$) by most (Panel A) or that reduces the average unexplained variance ($\overline{1-R^2}$) by most (Panel B). The search for alpha reduction terminates when all cluster portfolio alphas are subsumed; it consists of two stages. The first-stage search iteratively adds the factor that reduces the count of significant cluster portfolio alphas by most. It switches to the second stage when no additional factor can further reduce this count (indicated by a horizontal line). The second stage search proceeds by adding the most statistically significant of the surviving cluster portfolios to the current model. Panel B lists only the first 15 selected factors of to conserve space.

Panel A: α-reduction search

	Start with nothing				Start with HMXZ5				Start with FF3				Start with FF5				Start with FF6			
N	Added Factor	#Sig	$\lvert\alpha\rvert$	$\overline{1-R^2}$	Added Factor	#Sig	$\lvert\alpha\rvert$	$\overline{1-R^2}$	Added Factor	#Sig	$\lvert\alpha\rvert$	$\overline{1-R^2}$	Added Factor	#Sig	$\lvert\alpha\rvert$	$\overline{1-R^2}$	Added Factor	#Sig	$\lvert\alpha\rvert$	$\overline{1-R^2}$
1	EG	14	35	94	**MKT**	28	47	97	**MKTRF**	28	47	97	**MKTRF**	28	47	97	**MKTRF**	28	47	97
2	Accruals	10	28	87	**ME**	28	47	93	**SMB**	28	47	93	**SMB**	28	47	93	**SMB**	28	47	93
3	SMB	5	21	84	**I2A**	23	42	87	**HML**	24	44	85	**HML**	24	44	85	**HML**	24	44	85
4	IssuanceAndYield	3	18	77	**ROE**	17	36	78	EG	9	22	82	**RMW**	21	40	79	**RMW**	21	40	79
5	MKTRF	2	16	76	**EG**	11	24	76	Accruals	4	18	75	**CMA**	19	38	77	**CMA**	19	38	77
6	STREV	1	14	73	AssetGrowth	6	21	72	IssuanceAndYield	2	15	72	EG	7	21	75	**UMD**	17	32	71
7	seasonality	2	12	69	Research	5	18	69	STReversal	2	14	68	Accruals	2	16	70	EG	7	21	70
8	CapexGrowth	1	9	64	STREV	4	17	66	epsconsistency	2	13	64	STReversal	3	14	66	Accruals	2	15	65
9	epsconsistency	0	9	60	termbeta	3	16	63	CapexGrowth	1	9	60	MarginGrowth	2	12	61	STReversal	3	13	62
10					Accruals	4	13	58	seasonality	0	8	57	epsconsistency	3	11	57	MarginGrowth	2	12	58
11					ExternalFinance	3	11	55					ExternalFinance	1	9	53	epsconsistency	3	11	55
12					EarningsMomentum	2	10	51					seasonality	0	8	50	ExternalFinance	1	9	51
13					epsconsistency	1	9	49									seasonality	0	8	48
14					seasonality	1	7	46												
15					earnincq	0	7	43												

Panel B: $1-R^2$-reductionsearch

	Start with nothing				Start with HMXZ5				Start with FF3				Start with FF5				Start with FF6			
N	Added Factor	#Sig	$\lvert\alpha\rvert$	$\overline{1-R^2}$	Added Factor	#Sig	$\lvert\alpha\rvert$	$\overline{1-R^2}$	Added Factor	#Sig	$\lvert\alpha\rvert$	$\overline{1-R^2}$	Added Factor	#Sig	$\lvert\alpha\rvert$	$\overline{1-R^2}$	Added Factor	#Sig	$\lvert\alpha\rvert$	$\overline{1-R^2}$
1	ROE	17	40	88	MKT	28	47	97	MKTRF	28	47	97	MKTRF	28	47	97	MKTRF	28	47	97
2	IssuanceAndYield	18	38	79	ME	28	47	93	SMB	28	47	93	SMB	28	47	93	SMB	28	47	93
3	AssetGrowth	13	28	73	I2A	23	42	87	HML	24	44	85	HML	24	44	85	HML	24	44	85
4	Accruals	9	24	67	ROE	17	36	78	ROE	17	37	77	RMW	21	40	79	RMW	21	40	79
5	MarginGrowth	10	25	63	EG	11	24	76	Accruals	13	26	71	CMA	19	38	77	CMA	19	38	77
6	termbeta	10	24	59	Momentum	11	24	70	AssetGrowth	10	22	65	Momentum	17	31	70	UMD	17	32	71
7	coskewness	9	23	55	FinAssets	10	22	65	MarginGrowth	10	23	62	FinAssets	14	27	65	FinAssets	14	27	66
8	STReversal	9	18	51	coskewness	9	21	61	coskewness	10	22	58	Accruals	9	20	60	Accruals	10	20	61
9	pchquickratio	9	18	47	ExternalFinance	10	21	57	termbeta	10	22	54	ExternalFinance	9	19	56	ExternalFinance	9	19	57
10	earnincq	9	17	44	pctdisaccruals	7	19	54	earnincq	11	22	50	coskewness	9	18	53	coskewness	9	18	53
11	ExternalFinance	7	15	41	InventoryEfficiency	7	18	50	pchquickratio	10	22	47	fscore	9	17	49	fscore	9	17	50
12	pctdisaccruals	7	14	37	dFF5w6RMWdbeta	7	18	46	pctdisaccruals	9	20	44	pctdisaccruals	9	16	46	earnincq	8	16	46
13	seasonality	7	13	34	pchdeprn	7	16	43	ExternalFinance	8	18	40	earnincq	8	15	42	pctdisaccruals	8	15	43
14	pchdeprn	7	13	31	pchquickratio	7	16	40	STReversal	8	14	37	pchquickratio	8	15	39	pchquickratio	8	15	40
15	epsconsistency	6	12	28	earnincq	7	16	37	seasonality	7	13	34	pchdeprn	8	15	36	pchdeprn	8	15	37

资料来源：Journal of Banking and Finance，天风证券研究所

5. 资产定价

在基准模型种增加新的因子可以显著提高模型的定价能力。与以减少方差的搜索相比，以减少 alpha 为目的的搜索更加有效。

参照 Barillas 和 Shanken（2017）的研究，我们使用基准因子的月度最大夏普比平方（MS 比率）来比较基准模型的定价能力。下图结果显示，在全样本中，基准模型里 HMXZ5 模型的 MS 比最高，为 0.37。而搜索增强模型（SG9）为 0.51，远远强于 HMXZ5。在利用数据重抽样法进行 100,000 模拟后，上述结论依旧不变。

图 6：最大平方夏普比

Table 10
Maximum squared Sharpe ratios.
These tables report the maximum monthly squared Sharpe ratio (MS ratio) for different models. The sample runs from July 1972 to December 2019, or the maximum period during which all cluster portfolios and common factors have non-missing returns. Panel A considers benchmark models. Panel B considers nine-factor models aimed at alpha-reduction (as indicated by "[α]" at the end of a model name). "SG9 [α]" refers to the α-reduction search-generated nine-factor model without any pre-selected factors. Each of "HMXZ5 + 4 [α]", "FF5 + 4 [α]" and "FF6 + 3 [α]" refers to an α-reduction search-augmented nine-factor model that pre-selects benchmark factors. Panel C considers nine-factor models aimed at unexplained variance-reduction (as indicated by "[$1-R^2$]" at the end of a model name). All models in Panel B and Panel C are constrained to have nine factors to ensure fair comparisons of MS ratios. The α-reduction and $1-R^2$-reduction search procedures are described in Table 8. The Mean MS ratio and Median MS ratio are estimated through bootstrapped simulations. Specifically, to simulate one time-series of a benchmark factor we randomly draw 570 observations (with replacement) from the full sample of its monthly returns. In one simulation run, the same seed is used to generate pseudo time-series for all benchmark factors to maintain the cross-sectional correlation structure. One simulation run produces one MS ratio. This procedure is repeated 100,000 times to simulate a sample of MS ratios.

		Bootstrapped ($N = 100,000$)	
	Full-sample MS ratio	Mean	Median
Panel A: Benchmark models			
CAPM	0.016	0.018	0.016
FF3	0.040	0.046	0.044
HXZ4	0.155	0.165	0.163
FF5	0.101	0.112	0.110
FF6	0.131	0.147	0.144
HMXZ5	0.373	0.389	0.386
Panel B: Nine-factor models (α - reduction search)			
SG9 [α]	0.512	0.540	0.537
HMXZ5 + 4 [α]	0.455	0.485	0.482
FF5 + 4 [α]	0.487	0.516	0.513
FF6 + 3 [α]	0.477	0.505	0.502
Panel C: Nine-factor models ([$1-R^2$] - reduction search)			
SG9 [$1-R^2$]	0.301	0.328	0.325
HMXZ5 + 4 [$1-R^2$]	0.390	0.418	0.415
FF5 + 4 [$1-R^2$]	0.268	0.293	0.290
FF6 + 3 [$1-R^2$]	0.271	0.295	0.292

资料来源：Journal of Banking and Finance，天风证券研究所

6. 总结

本文从 215 个异象策略中筛选出了 80 个显著异象，并利用聚类分析将它们合并为 28 个集群投资组合。在当前的基准模型中，即使是表现最佳的 HXMZ5 模型仍然使 28 个集群组合中的 10 个具有显著性。集群投资组合体现了当前基准模型没有考虑到的预期回报的维度。

本文从 41 个候选因子中寻找可以解释 28 个集群组合和 80 个异象的因子。以降低 alpha 显著性为目标，最佳优先搜索策略找到 9 个因子。分别是 1)EG(预期增长因子), 2)Accruals 集群投资组合，3) SMB 规模因子，4) IssuanceAndYield 集群投资组合，5) MKTRF (市场超额收益因子), 6)STREV(短期反转因子), 7)seasonality 集群投资组合，8)CapexGrowth 集群投资组合和 9) epsconsistency 集群投资组合。

分析师声明

本报告署名分析师在此声明：我们具有中国证券业协会授予的证券投资咨询执业资格或相当的专业胜任能力，本报告所表述的所有观点均准确地反映了我们对标的证券和发行人的个人看法。我们所得报酬的任何部分不曾与，不与，也将不会与本报告中的具体投资建议或观点有直接或间接联系。

一般声明

特别声明

投资评级声明

类别	说明	评级	体系
股票投资评级	自报告日后的 6 个月内，相对同期沪深 300 指数的涨跌幅	买入	预期股价相对收益 20%以上
		增持	预期股价相对收益 10%-20%
		持有	预期股价相对收益-10%-10%
		卖出	预期股价相对收益-10%以下
行业投资评级	自报告日后的 6 个月内，相对同期沪深 300 指数的涨跌幅	强于大市	预期行业指数涨幅 5%以上
		中性	预期行业指数涨幅-5%-5%
		弱于大市	预期行业指数涨幅-5%以下

天风证券研究

北京	武汉	上海	深圳
北京市西城区佟麟阁路 36 号	湖北武汉市武昌区中南路 99 号保利广场 A 座 37 楼	上海市浦东新区兰花路 333 号 333 世纪大厦 20 楼	深圳市福田区益田路 5033 号平安金融中心 71 楼
邮编：100031	邮编：430071	邮编：201204	邮编：518000
邮箱：research@tfzq.com	电话：(8627)-87618889	电话：(8621)-68815388	电话：(86755)-23915663
	传真：(8627)-87618863	传真：(8621)-68812910	传真：(86755)-82571995
	邮箱：research@tfzq.com	邮箱：research@tfzq.com	邮箱：research@tfzq.com