专题报告

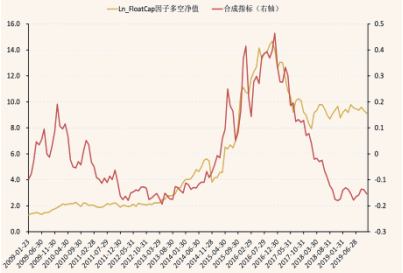
因子拥挤度指标及其择时作用

2020 年 02 月 02 日

机器学习视角下的考察

美国市场对于因子拥挤度指标的重视源于 2009 年动量因子（Momentum Factor）的大幅回撤，研究者认为因子拥挤度可能是影响因子寿命的重要原因。在国外研究的基础上，我们构建了估值价差、配对相关性、因子波动率、因子长期反转等 8 个因子拥挤度指标，并分别用这些指标对单因子收益方向和多因子组合权重进行了择时。在单因子择时方面，我们使用了 XGBoost 和 LSTM 两种机器学习算法，但是并没有取得明显优于纯做多方式的结果。我们又使用合成指标对多因子模型的权重进行调整，最后根据拥挤度指标加权后的多因子模型小幅战胜了因子等权组合的模型。

市值因子多空净值和拥挤度指标



资料来源：Wind 资讯、招商证券定量研究团队整理

- 正像建筑师在设计公共建筑主体时，兼顾商业价值和美观之余，需要着重考虑所需容纳人口的拥挤度，投资者在研究因子策略的时候，需要关注现有的因子拥挤度和资金容量上限。
- 过多的资金追逐同一资产可能会引发尾部风险，因而国外十分重视对于因子拥挤度研究。国外的研究认为因子拥挤度指标本身并非一个因子收益的负向指标，因为必须有资金流入才能推动因子有优秀的收益表现。只有在某个时期有过多的资金聚集在某个因子上的时候，才会使得因子过于拥挤。
- 在国外研究的基础上，我们试图探索因子拥挤度指标是否能在国内市场上对因子进行有效择时。我们构建了 4 类描述因子拥挤度的相对值指标，分别是估值价差、配对相关性、因子波动率、因子长期反转（共 8 个指标）。在 A 股市场，因子拥挤度指标对跟因子的多空收益相关性并不单调。
- 随后，我们分别利用了两种机器学习方法（XGBoost 和 LSTM）基于因子拥挤度指标对单因子未来一周的收益方向进行择时。调用 2009 年以来的数据作为训练和测试的样本，对每个因子分别建立预测模型，在测试集中评估模型的准确性。由于国内的因子多空收益的 Alpha 属性非常显著，利用模型对单因子收益方向进行预测的胜率并不高于每期都做多的简单策略的准确性。
- 我们随后用主成分分析法（PCA）降维合成因子拥挤度单一指标，该指标与因子多空净值走势呈现正相关性，我们以该指标加权构建多因子组合，加权后的组合多空净值小幅战胜等权的多因子组合。
- 总体而言，A 股市场的因子拥挤度有一定的尾部风险警示作用，但是持续用于因子的择时效果并不突出。我们认为一个可能的原因是，A 股市场的投资者结构中个人投资者居多，个人投资者在交易的时候很难形成同一方向的合力，因而在某些时段从合成指标上看因子拥挤度较高，但是即使在最高处，可能也远远没有达到这些常用因子的资金容量上限，不足以使因子发生尾部风险事件。因而在 A 股市场，因子拥挤度指标的指示作用并不显著。

相关报告

《利用 XGBoost 预测规模因子收益方向》

《利用 LSTM 算法估计基金因子暴露度》

任瞳
86-755-83081468
rentong@cmschina.com.cn
S1090519080004

崔浩瀚
86-21-68407276
cuihaohan@cmschina.com.cn
S1090519070004

正文目录

图表目录

因子拥挤度研究背景介绍

近十年来，量化投资在国内市场取得了长足的发展，尤为突出的是因子模型被众多机构投资者所接受和使用。因子模型相对稳健的表现和较大资金容量的优点很能符合机构量化投资的需要。随着国内外投资者对因子模型研究的深入，一些常见、有效且符合经济学逻辑的因子被不少投资机构有针对性地投资。大量的资金有方向性的流入到某个因子中，可能会导致单因子上聚集的资金过多，投资研究者可能会担忧尾部风险事件的发生，“因子拥挤度”的概念随之逐渐被投资者所关注。

建筑师在设计公共场所建筑主体时往往会从多种角度综合考量，力求在商业价值、美学和安全性等众多因素中找到最好的平衡。考量安全性时，人口的拥挤度是重要考察依据，公共场所的人口过度拥挤可能导致灾难性后果。投资者对因子的投资与建筑师设计建筑主体有相似之处，因子拥挤度往往也可以影响到一个因子的生命周期。资金的追逐是让一个因子表现优秀的必要条件，但是过度投资某个因子也会导致尾部风险事件的发生。

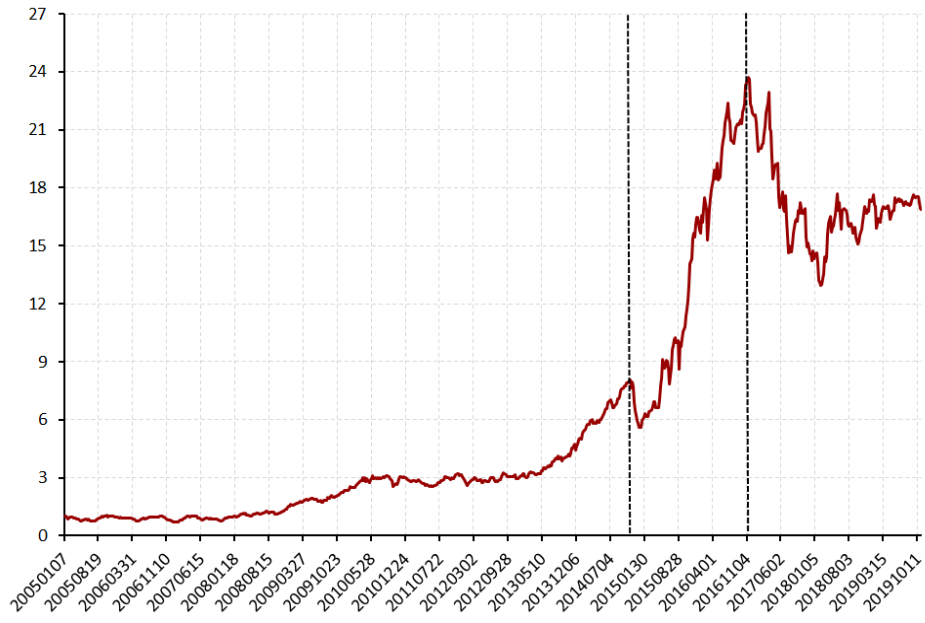
美国市场中，动量因子（Momentum Factor）一直以来是被投资者关注的重要因子，然而动量因子在 2009 年出现了大幅回撤，曾让国外投资者措手不及。投资者和学术界的研究人员对因子失效有多维度的解释，其中一个重要的解释是单因子拥挤度过高导致因子出现大幅回撤，认为因子拥挤度指标是影响因子生命周期的重要因素。

图 1 美国市场动量因子多空净值走势

资料来源：MSCI、招商证券定量研究团队整理

与美股动量因子相似，A 股市场上的市值因子在 2016 年之前曾经是不少机构量化部门重仓的重要因子，但是在 2016 年底和 2017 年却遭遇较大回撤，使量化投资者措手不及，如今市值因子依旧处于长期的震荡之中，风采不再。

图 2 A 股流通市值因子多空净值表现



资料来源：Wind 资讯、招商证券定量研究团队整理

从直觉上来说，因子拥挤度指标似乎应该是一个与因子收益呈负相关的指标，但是在国外，对美股因子的实际测算结果却并非如此。国外的研究认为，因子拥挤度指标本身并非一个因子收益的负向指标，因为必须有资金的稳定流入才能推动因子维持优秀的收益表现。只有在某个时期有过多的资金聚集在某个因子上的时候，才会使得因子过于拥挤。也就是说，在因子拥挤度指标没有达到非常高值的时候，因子拥挤度是可以推动因子取得优秀表现的。为了探究在 A 股市场是否也有类似关系，本报告将着手研究因子拥挤度指标对 A 股市场常用因子的影响。

因子拥挤度指标构建

构建因子拥挤度指标，衡量因子当前的拥挤情况，统计指标可以分为绝对值指标和相对值指标。绝对值指标是直接对市场上投资者的持仓数据进行分析，了解具体有多少资金投入到单因子上，这种方式逻辑直接，但是在实现上难度高。相对值指标则是统计得到的能代表某个因子的一组股票的相对估值、配对相关性等指标，用这些指标来间接判断是否出现了因子投资过于拥挤的情况。

表 1：因子拥挤度指标类型

指标类型	指标名称
绝对值指标	投资者持仓数据
相对值指标	估值价差
	配对相关性
	因子波动率
	因子长期反转

资料来源：招商证券定量研究团队整理

公募基金持仓数据

所谓基金持仓是指该基金重点投资股票、债券、货币等的资产数量和比例。投资者可以从基金持仓了解该基金的风险程度。常规地，我们获取基金持仓数据的依据是公募基金在每季度末公布的季度公告中的数据。但是由于季度报告的公布存在比较长的滞后期（参见表 2），且公布的持仓数据是期末的持仓数据的快照，是一个静态数据，用报告数据计算因子拥挤度的方法在实际中并不是一个理想的方法。虽然目前也有些方法通过研究基金复权单位净值增长率和因子收益率序列来反推基金产品在各因子上的暴露，然而就目前而言，并没有十分尽如人意的解决方案。在国外也同样面临类似问题，比如美股的 13F Filling 每季度公布一次，且截止日为报告期结束后的 45 天。

表 2：证监会规定的公募基金各类公开报告公布的截止日期

季度报告	半年度报告	年度报告
15 个工作日	60 日	90 日
重仓股持仓数据（前 10 只）	全部持仓数据	

资料来源：Wind 资讯、招商证券定量研究团队整理定量研究团队整理

由于上述原因，我们没有采用基金产品公布的季报中的持仓数据来计算因子的拥挤度，而是采用因子模型在应用中产生的交易数据来计算因子的拥挤度，我们将这类指标称为相对值指标。

相对值指标的构建

在说明我们如何构建因子拥挤度指标之前，有必要对我们构建因子拥挤度的方法进行简述。

表 3：大类风格因子构建方式和细分因子表

因子名称	大类风格代码	合成该风格因子所用到的细分因子
价值	Value	BP_LR, EP_Fwd12M, SP_TTM, OCFP_TTM, Sales2EV
市值	FloatCap	Ln_flaotcap
成长	Growth	Gr_Q_Earning, Gr_Q_OpEarning, Gr_Q_Sale
情绪	Sentiment	EPS_FY0_R1M, Rating_R3M, TargetReturn
反转	Momentum	RTN_20D, RTN_60D, RTN_1200D
交易行为	TradingBehavior	VolAvg_20D_240D, RealizedSkewness_240D, VolCV_20D, TurnoverAvg_20D, SpreadBias_120D, IVR, VWAPP_OLS

资料来源：招商证券定量研究团队整理定量研究团队整理

我们根据招商多因子库中的细分类别下的单因子合成了价值、成长、情绪、反转和交易行为五大类风格因子。其中，价值因子由 BP_LR，EP_Fwd12M，SP_TTM 等五个描述子以等权方式合成。成长因子由 Gr_Q_Earning，Gr_Q_OpEarning，Gr_Q_Sale 三个描述子以等权方式合成。情绪因子由 EPS_FY0_R1M，Rating_R3M，TargetReturn 三个描述子以等权方式合成。反转因子由 RTN_20D，RTN_60D，RTN_1200D 三个描述子以等权方式合成。交易行为因子由 VolAvg_20D_240D，RealizedSkewness_240D，VolCV_20D 等七个描述子以等权方式合成。合成各大类风格因子的细分因子如表 3 所示，具体单个因子的定义见附录。以下我们对因子拥挤度指标进行一一解释。

相对指标 1：估值价差

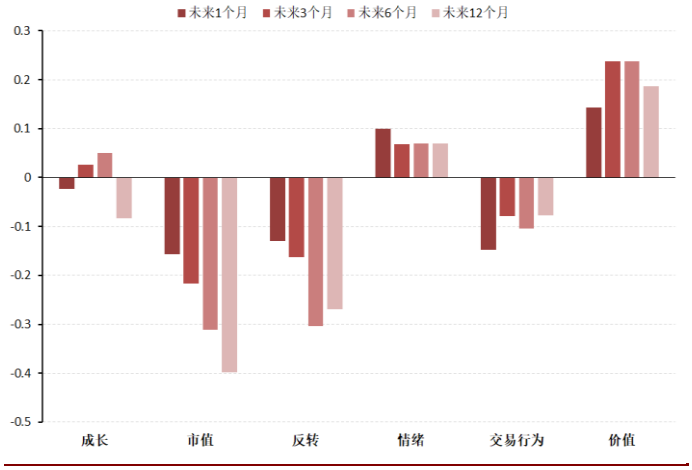
一般认为，若有大量的资金追逐某只股票时，股票的估值倾向于上升，而不被资金所偏好的个股估值倾向于走低。与研究股票估值的思路类似，当资金大量追逐某个因子时，能代表该单因子一组股票（多头组合）的估值倾向于上升，同时，该因子的空头组合的估值倾向于走低。以反转因子为例，我们根据如下步骤进行多头组合与空头组合的划分：

1. 取每月月底为截面观测日，选取 A 股市场上的在该时点上符合计算要求的个股（剔除 ST 股和上市不满 100 个交易日的股票）。
2. 按其当期的因子暴露度大小进行排序，然后依次平均分成 10 组，抽取排名第 1 的一个组合作为多头组合，排名第 10 的组合为空头组合，等权组合。
3. 分别计算多头组合和空头组合的相对估值指标的中位数，相对估值指标我们取最通用的 P/E（市盈率）、P/B（市净率）和 P/S（市销率），为了解决估值指标在收益接近 0 时取值接近无穷等不良情况，我们取估值指标的倒数 E/P，B/P 和 S/P，进行估值的刻画。
4. 构建估值价差指标：

估值价差 = 空头组合的估值指标倒数中位数 - 多头组合的估值指标倒数中位数

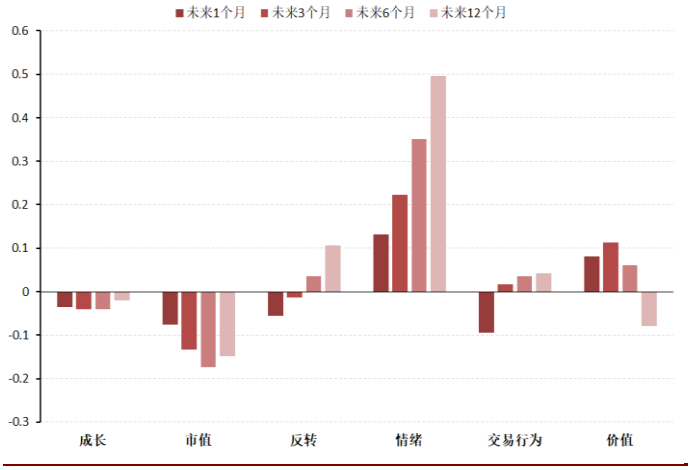
估值价差指标越高，则说明该时点上因子拥挤度越高。我们计算了估值价差指标和因子未来 1 个月、3 个月、6 个月和 12 个月的收益相关性。结果如下：

图 3 估值价差 EP 与因子未来多空收益相关系数走势



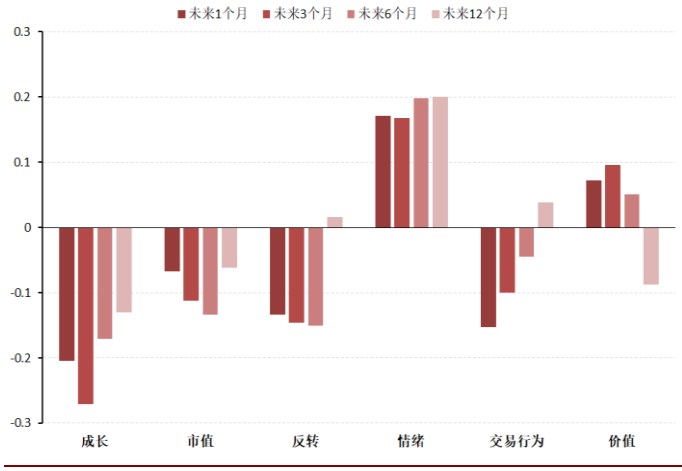
资料来源：Wind 资讯、招商证券定量研究团队整理

图 4 估值价差 BP 与因子未来多空收益相关系数走势



资料来源：Wind 资讯、招商证券定量研究团队整理

图 5 估值价差 SP 与因子未来多空收益相关系数走势



资料来源：Wind 资讯、招商证券定量研究团队整理

A 股市场上，估值价差指标与因子未来一年的收益走势相关性并不是单调的，比如情绪因子未来收益与估值价差呈现持续正相关，而市值因子却呈现持续的负相关。

相对指标 2：配对相关性

当大量资金追逐某个因子的时候，投资者可能会同向买入在该因子上暴露度较大的股票，同时同向卖出在该因子上暴露度低的股票，因此多头组合内部的股票可能会同涨同跌的态势，空头组合内部的股票也有类似效应。

根据这一逻辑我们计算了各因子多头组合与空头组合过去 3 个月收益与组内平均收益的相关系数。构成配对相关性指标：

1. 取每月月底为截面观测日，选取 A 股市场上的在该时点上符合计算要求的个股（剔除 ST 股和上市不满 100 个交易日的股票）。

2. 按其当期的因子暴露度大小进行排序，然后依次平均分成 10 组，抽取排名第 1 的一个组合作为多头组合，排名第 10 的组合为空头组合，等权组合。

3. 在截面观测日回看 3 个月，将多（空）头组合内的成分股日收益等权，计算多（空）头组合的日收益。分别计算组内成分股和多（空）头组合的日收益相关系数，并取均值，计算多（空）头配对相关性。

$$\text{多头配对相关性} = \text{Mean}(\text{Corr}(\text{个股 } i_{3\text{个月}}, \text{多头组合均收益}_{3\text{个月}}))$$

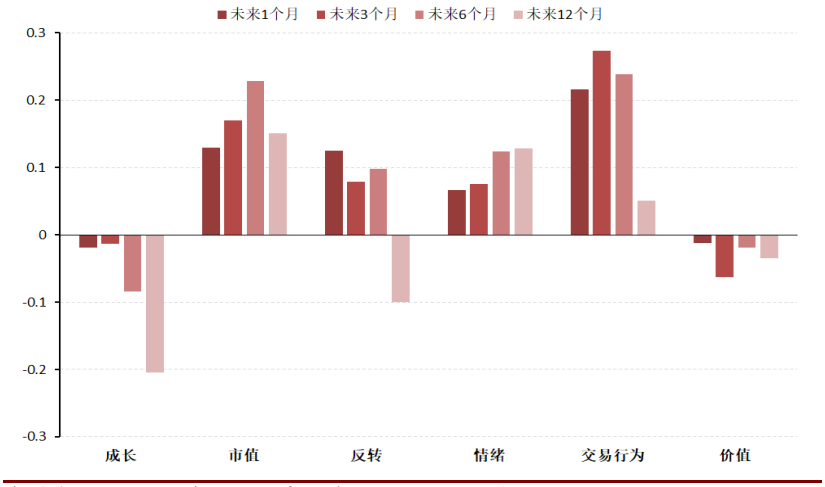
$$\text{空头配对相关性} = \text{Mean}(\text{Corr}(\text{个股 } i_{3\text{个月}}, \text{空头组合均收益}_{3\text{个月}}))$$

4. 最终的配对相关性指标为多头配对相关性与空头配对相关性之和。

$$\text{配对相关性指标} = \text{多头配对相关性} + \text{空头配对相关性}$$

从经济学逻辑出发，配对相关性指标越大，则因子越拥挤。我们同样给出了配对相关性指标与因子未来 1 个月、3 个月、6 个月和 12 个月的相关系数，结果如下：

图 6 配对相关性与因子未来多空收益相关系数走势



资料来源：Wind 资讯、招商证券定量研究团队整理

大部分因子的未来一年收益和配对相关性指标呈现正相关。

相对指标 3：因子收益波动率

因子收益波动率指标的思路是，当有较多资金聚集在某资产上时，资产价格的波动可能会变大。根据这一逻辑构建因子收益的波动率指标。

1. 取每月月底为截面观测日，选取 A 股市场上的在该时点上符合计算要求的个股（剔除 ST 股和上市不满 100 个交易日的股票）。

2. 按其当期的因子暴露度大小进行排序，然后依次平均分成 10 组，抽取排名第 1 的一个组合作为多头组合，排名第 10 的组合为空头组合，等权组合。

3. 在截面观测日回看 3 个月，将多（空）头组合内的成分股日收益等权，计算多（空）头组合的日收益，并计算多（空）头组合日收益序列的标准差。将历史时点上的 A

股市场上符合要求的股票等权组合，构建市场组合，用一样的方法计算市场组合的收益。

4. 构建因子波动率指标公式如下：

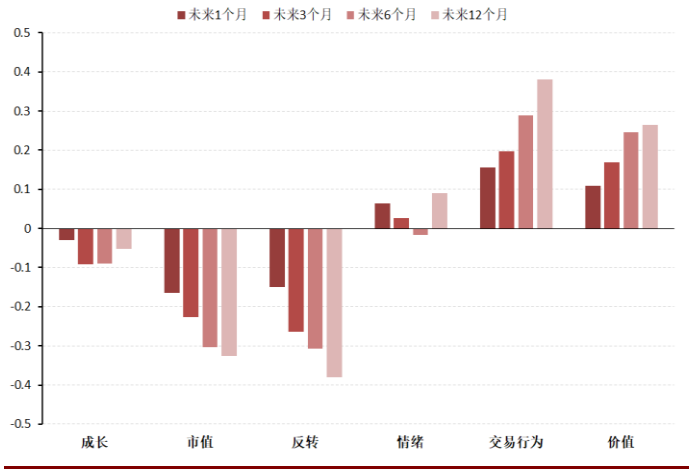
$$多头组合因子波动 = \frac{Std(多头组合)}{Std(市场组合)}$$

$$多空组合因子波动 = \frac{Std(多头组合)}{Std(空头组合)}$$

多头组合因子波动指标和多空组合因子波动指标的区别在于分母。多头组合的因子波动用多头组合收益标准差除以市场组合的收益标准差，来剔除市场整体波动对多头组合波动的影响。而多空组合因子波动则是反映多头组合的波动异于空头组合波动的程度。

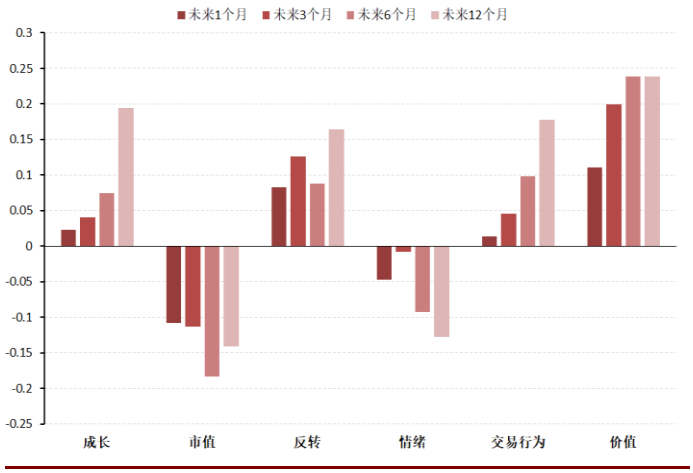
同样地，我们给出了该类指标对因子未来一年各时间段收益的相关系数，示列如下：

图 7 多头组合因子波动与因子未来多空收益相关系数



资料来源：Wind 资讯、招商证券定量研究团队整理

图 8 多空组合因子波动与因子未来多空收益相关系数



资料来源：Wind 资讯、招商证券定量研究团队整理

相对指标 4：因子长期反转

当大量资金涌向某个因子时，能代表该因子的一组股票的价格会上涨，股票资产收益的优秀表现又会进一步吸引资金进入，形成正向反馈，最终导致该因子拥挤。因而因子的长期反转可能也是衡量因子拥挤度的一个重要指标。我们分别设置了 3 年和 4 年的回溯期，去计算某个因子多空组合在回溯期里收益，滚动观测，形成因子长期反转指标。

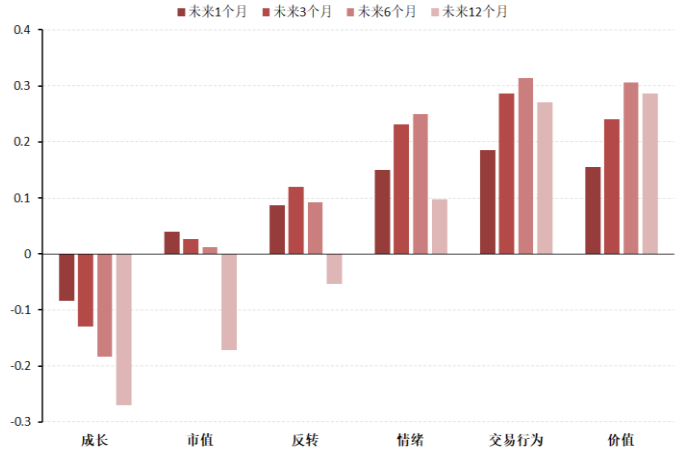
1. 取每月月底为截面观测日，选取 A 股市场上的在该时点上符合计算要求的个股，剔除 ST 股和上市不满 100 个交易日的股票。
2. 按其当期的因子暴露度大小进行排序，然后依次平均分成 10 组，抽取排名第 1 的一个组合作为多头组合，排名第 10 的组合为空头组合，等权组合。继而构建多空组合，计算多空组合的日收益序列。
3. 分别计算多空组合过去 3 年和过去 4 年的累积收益。

$$因子长期反转（3年）= Ret(因子多空组合_{近3年})$$

$$因子长期反转（4年）= Ret(因子多空组合_{近4年})$$

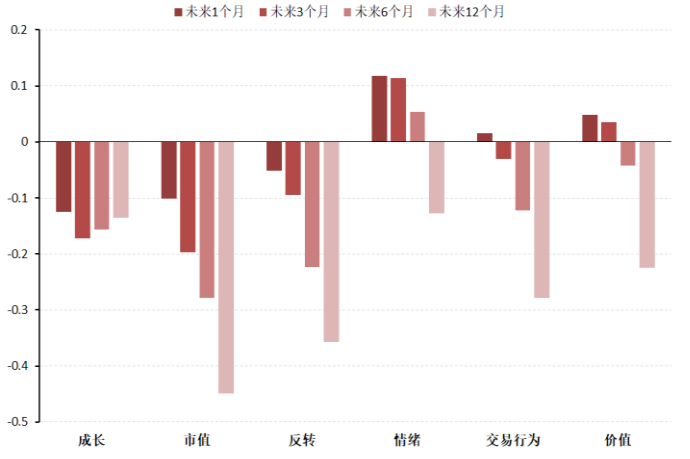
根据我们的研究，A 股市场上，常用因子的 Alpha 属性强于美国市场的因子表现，美国市场因子收益大约是 3 年发生较大回撤，而 A 股市场的这一回撤大约是 48 个月，因而我们分别计算了 3 年和 4 年的长期反转。图 9 和图 10 给出了这两个指标对因子未来 1 个月、3 个月、6 个月和 12 个月的收益相关系数走势。

图 9 因子长期反转（3 年）与因子多空收益相关系数



资料来源：Wind 资讯、招商证券定量研究团队整理

图 10 因子长期反转（4 年）与因子多空收益相关系数



资料来源：Wind 资讯、招商证券定量研究团队整理

仅从数据上来看，在 A 股市场，3 年长期反转指标与因子收益更多地呈现出正相关性，而 4 年长期反转指标则更多地呈现负相关，这和我们之前的推测相符合。

以上，我们构建了 4 类（共 8 个）因子拥挤度指标，这些指标背后都具有一定的经济学逻辑支撑，用这些指标能在一定程度上反映因子的拥挤程度，而且在数据可获得性上要显著优于其他类型的指标。在美国市场，还有一些指标被用来刻画因子拥挤度，比如能代表某个因子的一组股票的做空成本，但是由于在 A 股市场上，做空的成本要显著大于美国市场，做空行为在 A 股市场上并不流行，因而该类型指标在 A 股市场上的应用也有较大的局限性。所以在本篇报告中，没有借鉴美国市场上的这类指标。下文中，我们将用这些指标尝试对单因子和多因子模型进行择时。

机器学习算法对单因子多空收益进行择时

我们将计算得到的 8 个因子拥挤度指标作为特征变量，来预测单因子多空组合在未来一周的收益方向，分别尝试 XGBoost 和 LSTM 两种不同的机器学习算法进行预测。

XGBoost 算法预测单因子收益

XGBoost（Extreme Gradient Boosting）是一种非常有效的机器学习方法，也是时下比较前沿的机器学习算法之一。XGBoost 的一个显著特征是其在很多应用场景下的可扩

展性。我们认为它或许可以应用于对因子下期方向的判断。相较于其他机器学习算法，XGBoost 对于金融数据分析的优势主要有三点：

首先，该算法处理数据非常高效，在单台机器上的运行速度比现有流行解决方案快十倍以上，这对于处理海量的行情数据尤其重要。

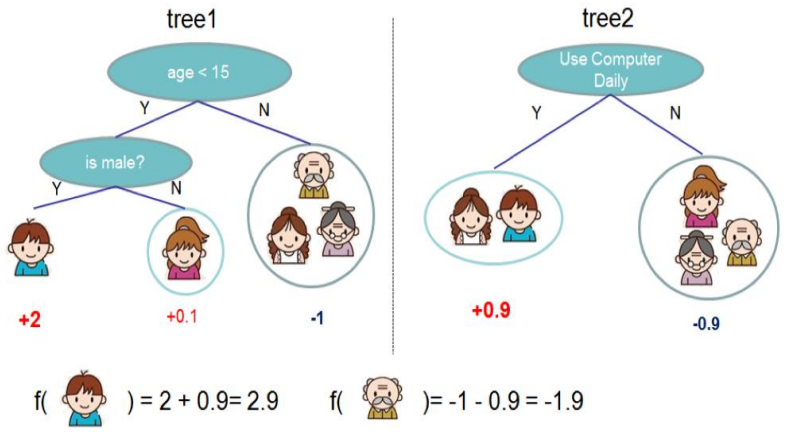
再者，该算法擅长处理稀疏矩阵（稀疏数据），即存在大量缺失值的数据。

第三，XGBoost 可以学习特征之间更高级别的相互关系。因子拥挤度和因子收益之间的关系很可能并不是线性的，而是比线性关系更复杂的存在。因而需要用 XGBoost 对更高级别关系进行探索和学习，往往可以更加吻合特征变量之间真实的、更高级别的关系。

XGBoost 是一个树集成算法（Tree Ensemble Model）。对于一个有n个样本、m个特征的数据集，$\mathcal{D}=\{(x_i,y_i)\}$ $\left(|\mathcal{D}|=n,,x_i\in\mathbb{R}^m,y_i\in\mathbb{R}\right)$，树集成模型会用$K$个可加的函数来预测结果：

$$\hat{y}_i=\emptyset(x_i)=\sum_{k=1}^{K}f_k(x_i)$$

图 11 XGBoost 算法示意图



资料来源：Chen, Tianqi, and Carlos Guestrin. "Xgboost: A scalable tree boosting system." *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. ACM, 2016、招商证券定量研究团队整理

表 4：XGBoost 超参数设置表

参数名称	取值	释义
max_depth	3	树的最大深度。树越深通常模型越复杂，更容易过拟合。
learning_rate	0.1	学习率或收缩因子。学习率和迭代次数 / 弱分类器数目 n_estimators 相关。
n_estimators	100	弱分类器数目。
gamma	0	节点分裂所需的最小损失函数下降值。
min_child_weight	1	叶子结点需要的最小样本权重和。
subsample	1	构造每棵树的所用样本比例（样本采样比例），同 GBM。

参数名称	取值	释义
colsample_bytree	1	构造每棵树的所用特征比例。
colsample_bylevel	1	树在每层每个分裂的所用特征比例。
reg_lambda	1	L2 正则的惩罚系数。

资料来源：招商证券定量研究团队整理

和其他机器学习的模型一样，XGBoost 也有较多的超参数需要设置，我们给出了我们的模型中超参数的设置方案，总体来说，我们在超参数的设置上尽量控制模型的复杂度，因为我们能使用的样本量有限，若模型的复杂度过高，不利于模型的准确性和泛化能力。

特征变量和标签数据处理

1. 输入的特征变量数据是 2005 年 1 月到 2019 年 11 月的周频数据，共 755 个样本。
2. 标签（被解释变量）为下一周的因子多空组合收益方向，收益为正取 1，否则取 0。
3. 进行了数据归一化处理（XGBoost 算法在进行树分裂的时候，依据的是数据的排序，因而是否对输入的特征变量数据进行归一化基本没有影响。）
4. 按时间序列对样本数据进行排序，取前 80%的数据作为训练集，后 20%的数据作为测试集。

预测结果评估

表 5：XGBoost 模型预测结果评价表

因子名称	模型预测准确性	做多策略准确性
成长	60.54%	63.53%
市值	40.70%	58.75%
反转	64.00%	67.24%
情绪	69.01%	72.81%
交易行为	82.76%	84.08%
价值	46.43%	62.86%

资料来源：Wind 资讯、招商证券定量研究团队整理

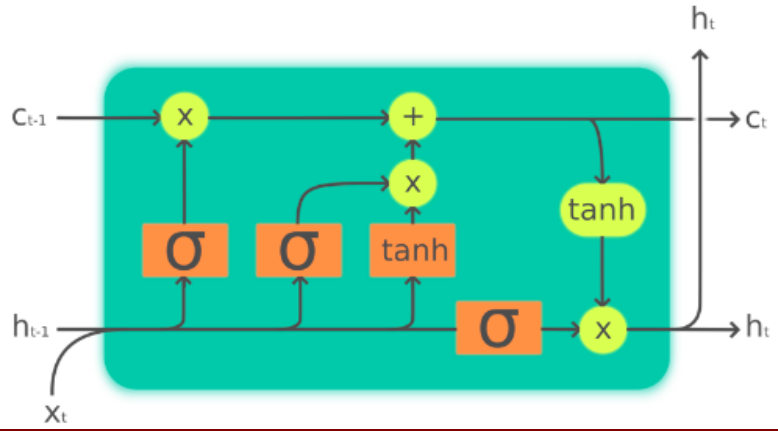
我们考察了 XGBoost 算法在测试集中的准确性，对于成长因子、反转因子、情绪因子和交易行为因子的预测准确性均在 60%以上，但是 XGBoost 的预测效果仍然不太理想。因为若我们不进行择时，每期都坚持做多，则胜率会高于 XGBoost 模型预测的胜率，交易行为因子的胜率甚至高达 84.08%。

LSTM 算法预测单因子收益

LSTM 即长短期记忆网络（Long Short-Term Memory），是一种时间循环神经网络（RNN），循环神经网络适合于处理和预测时间序列问题。LSTM 是在原有的 RNN 的神经网络上进行了神经元结构改造，在每个神经元中引入了三个“门”，分别叫做输入门、遗忘门和输出门，用于数据的筛选和保留。

一般的神经网络不具备像人一样的思维延续性。为了能使机器像人一样连续思考，科学家在原有的前馈神经网络中加入了内部反馈链接，RNN 应运而生，是专门用于处理时间序列问题的神经网络。但是 RNN 在刚出现的时候是难以被训练的，存在“梯度爆炸”（Exploding Gradients）和“梯度消失”（Vanishing Gradient）的问题。这两个问题导致 RNN 在刚出现的时候难以训练。为了解决上述问题，于是又发明了 LSTM，LSTM 较好的处理了“梯度爆炸”和“梯度消失”的问题。

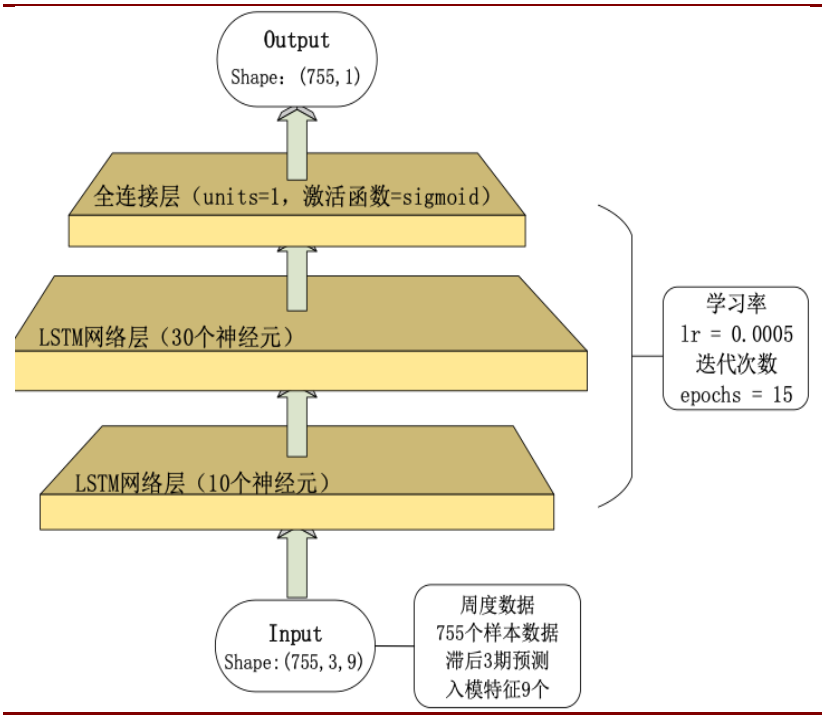
图 12 LSTM 网络神经元示意图



资料来源：维基百科、招商证券金融工程组整理

神经网络结构

图 13 LSTM 网络结构图



资料来源：招商证券定量研究团队整理

我们构建的神经网络的结构为 3 层，第 1 层和第 2 层为 LSTM 层，用于对特征数据和标签进行学习，并在测试集中对标签数据进行预测，第 3 层是全连接层，这层对数据进行整理，通过 sigmoid 函数输出最后我们需要的方向信号数据。特征变量为我们计算的 8 个拥挤度指标和 1 个标签的一阶滞后项，回看周期为 3 期，学习率设置为 0.0005，迭代次数（epochs）设置为 15。

特征变量和标签数据处理

1. 输入的特征变量数据是 2005 年 1 月到 2019 年 11 月的周频数据，共 755 个样本。
2. 标签（被解释变量）为下一周的因子多空组合收益方向，收益为正取 1，否则取 0。
3. 进行了数据归一化处理。
4. 按时间序列对样本数据进行排序，取前 80%的数据作为训练集，后 20%的数据作为测试集。

预测结果评估

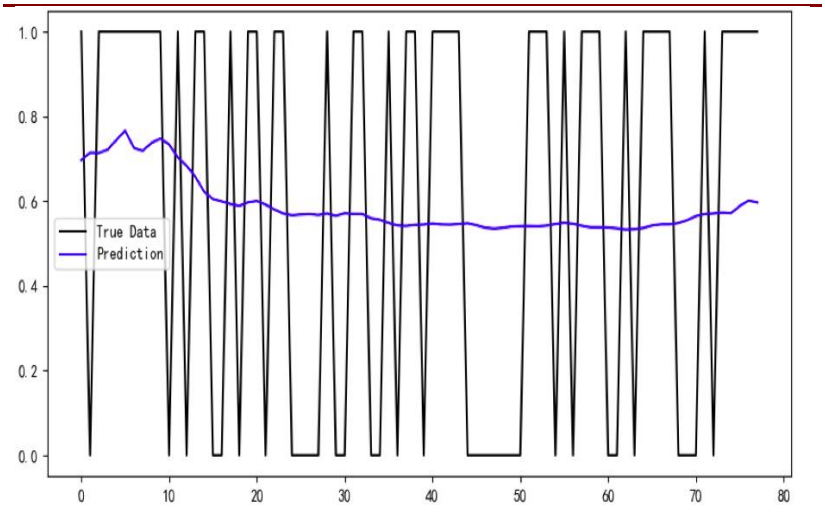
表 6：LSTM 模型预测结果评价表

因子名称	模型预测准确性	做多策略准确性
成长	63.53%	63.53%
市值	58.75%	58.75%
反转	67.24%	67.24%
情绪	72.81%	72.81%
交易行为	84.08%	84.08%
价值	62.86%	62.86%

资料来源：Wind 资讯、招商证券定量研究团队整理

十分凑巧的是，我们发现模型预测准确性简单做多策略的准确性是一致的，为了弄清楚其中的原因，我们查看了模型中间过程的计算结果：

图 14 LSTM 中间过程预测结果



资料来源：Wind 资讯、招商证券定量研究团队整理

上图中，蓝色折表示的是经过第二层 LSTM 层输出的预测值，尽管预测值有变动，但是由于始终大于 0.5，因而经过第三层的 sigmoid 函数之后，输出的是接近 1 的值，也就是说模型始终给出做多的投资建议，这也就是模型预测准确性与多头策略的准确性是一致的原因。

综上，LSTM 算法也未能在因子拥挤度指标中，学习到有利于单因子收益方向判断的增量信息。既然在单因子多空收益预测上，因子拥挤度指标的应用并不是很理想，我们转变思路，看能否在多因子组合上，获得择时方面的增量信息，取得一些超额收益。

利用合成指标对多因子模型进行择时

我们对这 8 个拥挤度指标按不同的因子，进行了相关系数计算和展示。

图 15 成长因子各拥挤度指标两两相关系数

	ValueBP	ValueEP	ValueSP	PairCorr	VolLong	VolLS	Rever3	Rever4
ValueBP	1.000	-0.265	0.056	0.193	-0.123	0.288	-0.233	-0.385
ValueEP	-0.265	1.000	-0.240	-0.320	-0.204	-0.118	0.266	0.231
ValueSP	0.056	-0.240	1.000	0.139	0.048	-0.114	0.168	-0.191
PairCorr	0.193	-0.320	0.139	1.000	0.040	-0.374	0.324	0.129
VolLong	-0.123	-0.204	0.048	0.040	1.000	0.445	0.085	0.270
VolLS	0.288	-0.118	-0.114	-0.374	0.445	1.000	-0.198	0.145
Rever3	-0.233	0.266	0.168	0.324	0.085	-0.198	1.000	0.520
Rever4	-0.385	0.231	-0.191	0.129	0.270	0.145	0.520	1.000

资料来源：Wind 资讯、招商证券定量研究团队整理

图 16 市值因子各拥挤度指标两两相关系数

	ValueBP	ValueEP	ValueSP	PairCorr	VolLong	VolLS	Rever3
ValueBP	1.000	0.315	0.697	-0.248	0.516	0.439	0.480
ValueEP	0.315	1.000	0.566	-0.057	0.169	0.089	0.333
ValueSP	0.697	0.566	1.000	-0.175	0.368	0.333	0.250
PairCorr	-0.248	-0.057	-0.175	1.000	-0.435	-0.637	-0.021
VolLong	0.516	0.169	0.368	-0.435	1.000	0.888	0.130
VolLS	0.439	0.089	0.333	-0.637	0.888	1.000	0.157
Rever3	0.480	0.333	0.250	-0.021	0.130	0.157	1.000
Rever4	0.517	0.277	0.177	-0.145	0.351	0.343	0.833

资料来源：Wind 资讯、招商证券定量研究团队整理

图 17 反转因子各拥挤度指标两两相关系数

	ValueBP	ValueEP	ValueSP	PairCorr	VolLong	VolLS	Rever3
ValueBP	1.000	0.023	0.694	-0.022	-0.056	-0.114	0.423
ValueEP	0.023	1.000	0.248	0.048	0.327	0.074	0.339
ValueSP	0.694	0.248	1.000	-0.268	0.070	0.062	0.203
PairCorr	-0.022	0.048	-0.268	1.000	-0.021	-0.624	0.235
VolLong	-0.056	0.327	0.070	-0.021	1.000	0.239	-0.050
VolLS	-0.114	0.074	0.062	-0.624	0.239	1.000	-0.120
Rever3	0.423	0.339	0.203	0.235	-0.050	-0.120	1.000
Rever4	0.410	0.400	0.252	0.150	0.003	-0.053	0.752

资料来源：Wind 资讯、招商证券定量研究团队整理

图 18 情绪因子各拥挤度指标两两相关系数

	ValueBP	ValueEP	ValueSP	PairCorr	VolLong	VolLS	Rever3
ValueBP	1.000	0.449	0.676	0.142	0.111	0.037	0.332
ValueEP	0.449	1.000	0.484	0.008	0.024	0.136	0.332
ValueSP	0.676	0.484	1.000	0.058	-0.134	0.106	0.284
PairCorr	0.142	0.008	0.058	1.000	0.007	-0.651	0.349
VolLong	0.111	0.024	-0.134	0.007	1.000	0.030	0.147
VolLS	0.037	0.136	0.106	-0.651	0.030	1.000	-0.177
Rever3	0.332	0.332	0.284	0.349	0.147	-0.177	1.000
Rever4	0.212	0.142	0.058	0.149	-0.023	0.015	0.666

资料来源：Wind 资讯、招商证券定量研究团队整理

图 19 交易行为因子各拥挤度指标两两相关系数

	ValueBP	ValueEP	ValueSP	PairCorr	VolLong	VolLS	Rever3
ValueBP	1.000	0.125	0.658	-0.077	0.137	0.227	0.416
ValueEP	0.125	1.000	0.116	-0.084	0.326	0.186	0.311
ValueSP	0.658	0.116	1.000	-0.101	0.017	0.131	0.224
PairCorr	-0.077	-0.084	-0.101	1.000	-0.169	-0.693	0.464
VolLong	0.137	0.326	0.017	-0.169	1.000	0.433	0.327
VolLS	0.227	0.186	0.131	-0.693	0.433	1.000	-0.051
Rever3	0.416	0.311	0.224	0.464	0.327	-0.051	1.000
Rever4	0.467	0.309	0.332	0.155	0.107	0.080	0.755

资料来源：Wind 资讯、招商证券定量研究团队整理

图 20 价值因子各拥挤度指标两两相关系数

	ValueBP	ValueEP	ValueSP	PairCorr	VolLong	VolLS	Rever3
ValueBP	1.000	0.668	0.859	0.213	0.339	-0.026	0.586
ValueEP	0.668	1.000	0.401	0.131	0.566	-0.014	0.753
ValueSP	0.859	0.401	1.000	0.055	0.008	0.054	0.478
PairCorr	0.213	0.131	0.055	1.000	0.440	-0.533	0.353
VolLong	0.339	0.566	0.008	0.440	1.000	-0.072	0.447
VolLS	-0.026	-0.014	0.054	-0.533	-0.072	1.000	-0.252
Rever3	0.586	0.753	0.478	0.353	0.447	-0.252	1.000
Rever4	0.651	0.807	0.566	0.143	0.369	-0.121	0.744

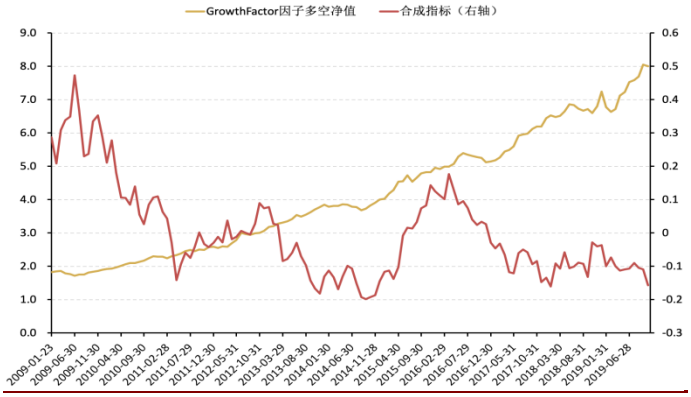
资料来源：Wind 资讯、招商证券定量研究团队整理

对于不同的因子，这些指标的相关系数有差异，但是总体而言，比较显著的特征是同一类型指标的相关系数较高，比如因子长期反转（3 年）和因子长期反转（4 年）指标。而其中一个比较显著的特例是配对相关性指标和多空组合因子波动指标有比较显著的负相关关系。在 A 股市场上，部分拥挤度指标存在自相矛盾的情况。

鉴于部分指标的相关系数绝对值偏高，我们用主成分分析法（PCA）分析每个指标贡献的方差，保留能贡献较大方差的指标，同时利用因子旋转降维、组合合成指标，具体方法如下：

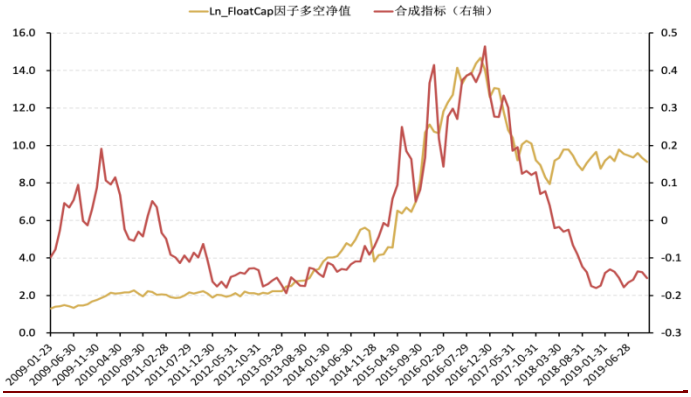
1. 利用因子旋转将 8 个指标降维成 3 个指标(考虑到前 3 个因子的方差贡献显著高于其他因子)；
2. 将降维得到的 3 个指标等权组合，合成一个因子拥挤度指标。

图 21 成长因子多空净值与拥挤度合成指标



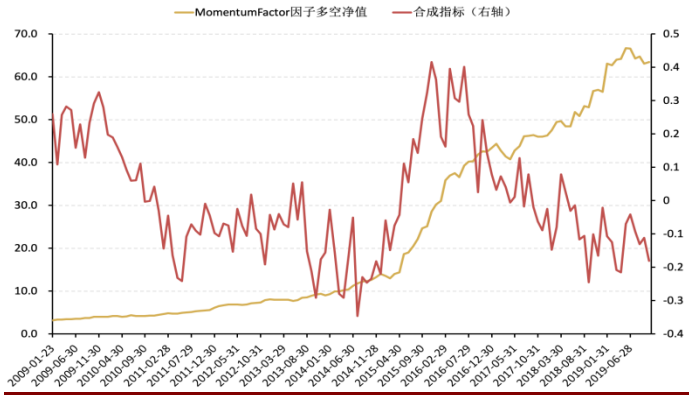
资料来源：Wind 资讯、招商证券定量研究团队整理

图 22 市值因子多空净值与拥挤度合成指标



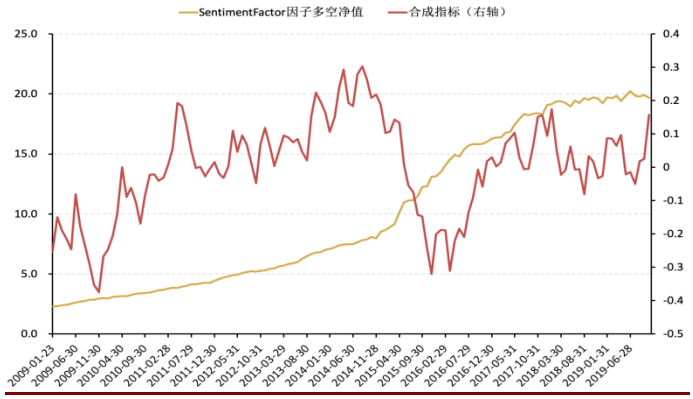
资料来源：Wind 资讯、招商证券定量研究团队整理

图 23 反转因子多空净值与拥挤度合成指标



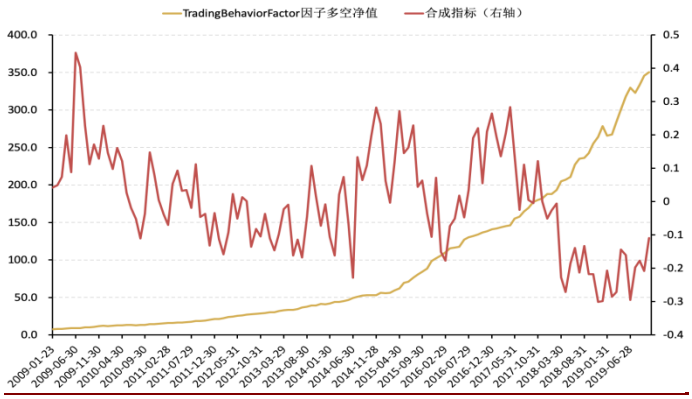
资料来源：Wind 资讯、招商证券定量研究团队整理

图 24 情绪因子多空净值与拥挤度合成指标



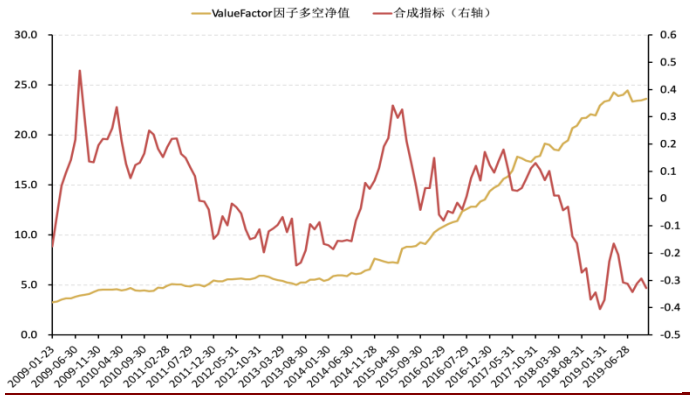
资料来源：Wind 资讯、招商证券定量研究团队整理

图 25 交易行为因子多空净值与拥挤度合成指标



资料来源：Wind 资讯、招商证券定量研究团队整理

图 26 价值因子多空净值与拥挤度合成指标



资料来源：Wind 资讯、招商证券定量研究团队整理

图 21 至图 26 中，红色折线为因子拥挤度合成指标的走势，黄色折线则代表因子多空累积收益。从上图看，由于以量价指标计算的因子（比如情绪因子和交易行为因子）有很强的 Alpha 属性，在过去的观测期中很少出现明显的回撤，因而因子拥挤度指标对于这些因子的风险指示作用并不显著。相对而言，对市值因子的尾部风险指示作用更为明显一些。

由于市值因子在 2009 年至 2016 年底这段时间里有很稳定的正向收益，小市值的股票走势长期好于大市值股票，而且一般股票在市值因子上的暴露度波动不会很大，市值因子在该期间是十分受追捧。因子拥挤度合成指标显示，在 2013 年至 2016 年，市值因子的拥挤度快速上升，在 2016 年底达到顶峰之后，市值因子发生了较大回撤，蓝筹股和大市值股票的收益要好于小市值股票，因子拥挤度指标也几乎在同时发生下降。2018 年以来，市值因子的收益方向始终维持在震荡走势，因子拥挤度指标也处于历史低位。

然而总体而言，在 A 股市场上，因子拥挤度指标的指示作用弱于国外市场。我们猜测一个可能的原因是，A 股市场的投资者结构中个人投资者居多，个人投资者在交易的时候很难形成同一方向的合力，因而在某些时段从合成指标上看因子拥挤度较高，但是即使在最高处，可能也远远没有达到这些常用因子的资金容量上限，不足以使因子发生尾部风险事件。因而在 A 股市场，因子拥挤度指标的指示作用并不显著。

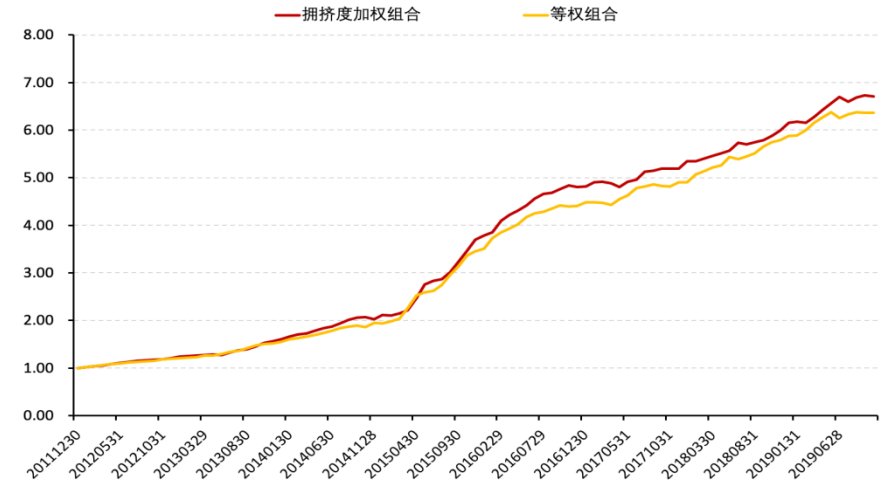
多因子加权组合构建

尽管因子拥挤度能在一定程度上提示因子的尾部风险，但是在大部分情况下是呈现正相关的，而不是逻辑直觉上的负相关关系。因此，我们可以构建以因子拥挤度指标加权的多因子组合，具体做法如下：

1. 滚动观测因子拥挤度指标，计算当前因子拥挤度指标在过去 3 年中的分位数。以分位数作为因子加权权重，月频调仓构建多因子组合。
2. 模型只对基本面类因子（价值、市值、成长）进行择时，对 Alpha 属性较强的量价类因子依然是等权配置，持续做多。

以等权多因子组合的多空净值为对标基准，因子拥挤度加权组合净值走势如图 27 所示：

图 27 拥挤度加权组合与等权组合净值走势



资料来源：Wind 资讯、招商证券定量研究团队整理

图 27 中，黄色折线是因子的等权组合净值走势线，等权组合常被用来当做比较的基准。红色折线是根据因子拥挤度指标进行加权后的多因子组合净值走势。自 2015 年以来，因子的加权组合小幅战胜了等权组合。

总结

由于过多的资金追逐同一资产可能会引发尾部风险，因而国外十分重视对于因子拥挤度研究。国外的研究认为因子拥挤度指标本身并非一个因子收益的负向指标，因为必须有资金流入才能推动因子有优秀的收益表现。只有在某个时期有过多的资金聚集在某个因子上的时候，才会使得因子过于拥挤。

在国外研究的基础上，我们试图探索因子拥挤度指标是否能在国内市场上对因子进行有效择时。我们继而构建了 4 种描述因子拥挤度的相对值指标，分别是估值价差、配对相关性、因子波动率、因子长期反转。在 A 股市场，因子拥挤度指标对跟因子的多空收益相关性并不单调。

我们分别利用了两种机器学习方法（XGBoost 和 LSTM）基于因子拥挤度指标对单因子进行择时，但是由于国内的因子 alpha 属性很显著，在单因子层面并没有获得良好的择时效果。我们随后用主成分分析法降维合成因子拥挤度单一指标，并以该指标加权构建多因子组合，组合净值线小幅战胜等权的多因子组合。

总体而言，A 股市场的因子拥挤度有一定的尾部风险警示作用，但是持续用于因子的择时效果并不显著。我们猜测一个可能的原因是，A 股市场的投资者结构中个人投资者居多，个人投资者在交易的时候很难形成同一方向的合力，因而在某些时段从合成指标上看因子拥挤度较高，但是即使在最高处，可能也远远没有达到这些常用因子的资金容量上限，不足以使因子发生尾部风险事件。因而在 A 股市场，因子拥挤度指标的指示作用并不显著。

附录

表 7：细分因子定义

因子类	因子名称	因子描述
价值	EP_Fwd12M	每股收益_未来 12 个月预测值 / 收盘价
	SP_TTM	营业收入_TTM / 总市值
	OCFP_TTM	经营活动产生的现金流量净额_TTM / 总市值
	BP_LR	股东权益合计(不含少数股东权益)_最新财报 / 总市值
	Sales2EV	营业收入_TTM / (总市值 + 非流动负债合计_最新财报 - 货币资金_最新财
成长	Gr_Q_OpEarning	单季度营业利润同比增长率
	Gr_Q_Earning	单季度净利润同比增长率
	Gr_Q_Sale	单季度营业收入同比增长率
反转	RTN_20D	复权收盘价 / 复权收盘价_20 天前 - 1
	RTN_60D	复权收盘价 / 复权收盘价_60 天前 - 1
	RTN_1200D	复权收盘价 / 复权收盘价_1200 天前 - 1
情绪	EPS_FY0_R1M	一致预期 EPS_FY0 过去 20 天的变化率
	Rating_R3M	分析师综合评级 3 个月的变化率
	TargetReturn	一致预测目标价 / 收盘价 - 1
交易行为	VolAvg_20D_240D	过去 20 天日均成交量 / 过去 240 天日均成交量
	TurnoverAvg_20D	过去 20 天日均换手率
	VolCV_20D	过去 20 天日成交量的标准差 / 过去 20 天日均成交量
	RealizedSkewness_240D	过去 240 天日收益率数据计算的偏度
	SpreadBias_120D	价差偏离度因子
	IVR	1 - Fama-French 三因子回归的调整 R 平方
	VWAPP_OLS	VWAP 收益率与原始收益率回归的残差

资料来源：招商证券定量研究团队整理

分析师承诺

负责本研究报告的每一位证券分析师，在此申明，本报告清晰、准确地反映了分析师本人的研究观点。本人薪酬的任何部分过去不曾与、现在不与，未来也将不会与本报告中的具体推荐或观点直接或间接相关。

任瞳：首席分析师，定量研究团队负责人，管理学硕士，16 年证券研究经验，2010 年、2015 年、2016、2017 年新财富最佳分析师。在量化选股择时、基金研究以及衍生品投资方面均有深入读到的见解。

崔浩瀚：量化分析师，浙江大学经济学硕士，3 年量化策略研究开发经验。研究方向是机器学习在金融领域的应用和多因子选股策略开发。

投资评级定义

公司短期评级

以报告日起 6 个月内，公司股价相对同期市场基准（沪深 300 指数）的表现为标准：

强烈推荐：公司股价涨幅超基准指数 20%以上
审慎推荐：公司股价涨幅超基准指数 5-20%之间
中性：公司股价变动幅度相对基准指数介于±5%之间
回避：公司股价表现弱于基准指数 5%以上

公司长期评级

A：公司长期竞争力高于行业平均水平
B：公司长期竞争力与行业平均水平一致
C：公司长期竞争力低于行业平均水平

行业投资评级

以报告日起 6 个月内，行业指数相对于同期市场基准（沪深 300 指数）的表现为标准：

推荐：行业基本面向好，行业指数将跑赢基准指数
中性：行业基本面稳定，行业指数跟随基准指数
回避：行业基本面向淡，行业指数将跑输基准指数

重要声明

本报告由招商证券股份有限公司（以下简称“本公司”）编制。本公司具有中国证监会许可的证券投资咨询业务资格。本报告基于合法取得的信息，但本公司对这些信息的准确性和完整性不作任何保证。本报告所包含的分析基于各种假设，不同假设可能导致分析结果出现重大不同。报告中的内容和意见仅供参考，并不构成对所述证券买卖的出价，在任何情况下，本报告中的信息或所表述的意见并不构成对任何人的投资建议。除法律或规则规定必须承担的责任外，本公司及其雇员不对使用本报告及其内容所引发的任何直接或间接损失负任何责任。本公司或关联机构可能会持有报告中所提到的公司所发行的证券头寸并进行交易，还可能为这些公司提供或争取提供投资银行业务服务。客户应当考虑到本公司可能存在可能影响本报告客观性的利益冲突。