

多因子投资模式下的因子挖掘

——基于舆情的因子研究与策略

牛晓健,金圣皓

(复旦大学 国际金融系,上海 200433)

摘要:以东方财富个股股吧评论内容作为原始语料,结合自然语言处理技术构建显性舆情因子后发现,舆情因子与股票收益率具有较强的相关性,因子测试结果表明,使用舆情因子构建投资策略具有可行性。经将舆情因子加入多因子模型中形成指数增强策略,回测结果表明,结合舆情因子的多因子投资策略可以显著获得超过基准的超额收益。利用技术指标作为隐性舆情的替代,探讨隐性情绪指标在 Barra 模型中的利用以期获得更完善的情绪策略研究,结果表明结合隐性策略的 Barra 模型可以获得较显性策略且更稳健的回测效果。

关键词:舆情因子;多因子模型;Barra 风险模型;指数增强策略

中图分类号:F830.91 **文献标志码:**A **文章编号:**1008-5092(2021)05-0037-13

一、绪论

基于市场认知的著名投资家索罗斯指出,市场具有一定程度的反身性,即基本面情况可以影响认知,认知也可以影响基本面。基于以上两方面的因素,本文从逻辑出发挖掘的市场舆情因子为出发点,探讨其与股票市场之间的联系并利用 Barra 模型构建股票的多因子策略。本文试图验证构建的显性舆情因子是否可以取得超额收益、是否具有选股及预测能力。此外,本文探究了舆情因子与其他因子结合的可能性,从而构建有效的指数增强策略。

随着信息技术的不断发展,关于外部信息对于股票收益率的研究有很多,包括投资者情绪与股票指数的相关性的实证检验,以及将外部信息量化为指标,在行业层面上实现策略应用。以上量化手段侧重使用简单的计量方法,如网页搜索量、发帖数量、标注好的看涨或看跌情绪等,本文则试图在基于相关实证检验的基础之上将外部信息根据其语义内容量化为舆情指标,并讨论基于 Barra 模型的择股以及择时策略,因此从另一方面讲也是对 Barra 因子体系的扩充。

二、文献综述

因子模型在理论上已有了较为成熟的发展基础。从 Markowitz 的证券组合理论(1952)^[1]到 Sharpe(1964)^[2]导出的资本资产定价模型(The Capital Asset Pricing Model, CAPM): $\bar{r}_i = r_f + \beta_{im}(\bar{r}_m - r_i)$, Ross(1976)^[3]、套利定价理论(Arbitrage Pricing Theory, APT)的出现将原本的单因素理论发展到多因素理论。

21 世纪末, Fama 和 French(1992)^[4]提出了三因子模型,引入了市场回报、规模效应和市盈率。Fama 和 French(2015)^[5]又在三因素模型中增加了盈利能力和投资水平两个指标。随着信息技术水平的不断发展,多因子模型中加入了技术指标。Asness(1997)^[6]提出了价值策略与动量策略在不同股票上的有效性。业界也不断地推出新的模型新的因子,例如 Barra 公司提出的 USE 风险模型。随着机器学习、深度学习等技术的不断发展,因子的研究中逐渐引入了决策树、XGBoost、神经网络等多种机器学习算法。黄卿(2017)^[7]在 Fama、French 三因子模型的基础上引入支持

收稿日期:2021-05-13

基金项目:国家自然科学基金面上项目(71873039;71573051)。

作者简介:牛晓健(1971—),男,新疆阿克苏人,教授,博士,研究方向:量化投资。

向量机(SVM),在回测期间内取得了年化回报 22.4% 的成绩。

构建因子过程中使用到了分词技术,其中最简单的为机械分词方法,其根据已有字典,按照正向最大匹配法、逆向最大匹配法和双向最大匹配法[邹海山等,2000]^[8]等方法进行匹配。

在舆情的采集方面,可以利用搜索引擎中某些特定词汇的出现频率作为舆情的近似表达,孟雪井等(2016)^[9]在构建投资者情绪指数时就使用了百度搜索指数。孟勇等(2019)^[10]则在研究舆情与市值效应的关系时爬取了新浪财经专栏文章并与传统的可代表舆情信息的其他因子进行了对比分析。

在舆情的度量方面,可通过情感字典和机器学习实现。情感词典可以根据不同的使用场景构建独有的情感词典,游家兴等(2012)^[11]用新闻基调、关注度水平和曝光度三个维度衡量新闻媒体情绪,得出媒体报道偏离客观程度越大、资产价格错误定价的可能性越大的结论。机器学习法则利用语料库作为训练集,根据人工文本标注结果将模型应用到测试集中查看分类效果。

在舆情与收益率的关联性方面,新西兰学者 Frijns 等(2017)^[12]发现投资者情绪解释了股票波动非基本成分中的方差与协方差等情况。国内学者俞庆进、张兵(2012)^[13]通过百度指数分析投资者有限关注与收益率的关系,发现投资者的关注能够影响股票的交易,带来正向价格压力。

现存的“舆情对收益率是否存在影响”相关的文献中大多利用指标模拟舆情得出股票收益率与舆情的关系,单一的搜索或交易指标不能体现完备的舆情信息,使用这种方法有局限性。同时,现存文献较注重舆情对收益率的影响,尽管少数的文献提供了投资策略的想法,但大多数使用的是较为传统的策略。本文通过对中文财经网站上网友的评论进行情感分析,并使用 Barra 多因子模型作为传统多因子模型的替代,更好地融入了风险与收益两个视角构建基于舆情的投资策略。

三、Barra 模型理论体系

美国学者 Barr Rosenberg 在 1975 年提出了 Barra USE1 模型,该模型首次提出了多因子风险模型的概念。

1. 基本表述

Barra 模型基本表述:

$$r_i = \sum_{k=1}^K X_{ik} f_k + u_i \quad (1)$$

其中, r_i 表示股票回报率, X_{ik} 表示因子 k 在股票 i 上的因子值, f_k 为因子回报率, u_i 表示的是特质回报率。

Barra 模型将因子分为国家(市场)因子、行业因子和风格因子三大类,如图 1 所示。

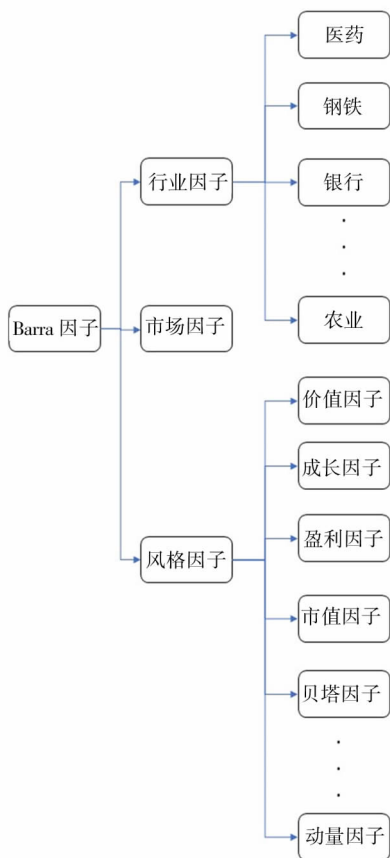


图 1 Barra 因子

N 只股票组成的投资组合收益率为:

$$R_p = \sum_{i=1}^N w_i r_i \quad (2)$$

风险因子 k 在组合上的风险暴露程度为:

$$X_k^p = \sum_{i=1}^N w_i X_{ik} \quad (3)$$

组合收益率为:

$$R_p = \sum_{k=1}^K X_k^p f_k + \sum_{i=1}^N w_i u_i \quad (4)$$

假设因子收益与特质收益率无关,则投资组合整体风险为:

$$\text{var}(R_p) = \sum_{m,n} X_k^p F_{mn} X_k^p + \sum_{i=1}^N w_i^2 \text{var}(u_i) \quad (5)$$

其中 F_{mn} 为因子 m 与因子 n 的协方差矩阵。

综上所述,完整的多因子模型的采用因素之间的协方差度量风险,而后对选股结果以及股票

配置仓位进行优化:

$$\max_w w^T \mu$$
$$s. t. w^T l = 1, w^T var(R_p) w \leq \sigma^2 \quad (6)$$

其中 w 为资产池内各资产的权重比例, l 为全 1 向量, μ 是各资产的期望回报率, σ 表示量化层面上的风险。

2. 因子标准化

因子暴露标准化方法包括:风格归一化和 Z 评分归一化。风格归一化对每个因子减去市值加权均值,再除以其标准差:

$$X_{ik} = \frac{X_{ik}^{raw} - \mu_k}{\sigma_k}$$
$$\mu_k = \sum w_i X_{ik}^{raw} \quad (7)$$

等权组合指数下可使用 Z 值标准化:

$$X_{ik} = \frac{X_{ik}^{raw} - \mu_k}{\sigma_k}$$
$$\mu_k = \frac{1}{N} \sum X_{ik}^{raw} \quad (8)$$

四、单因子研究

随着信息技术的不断发展,投资者们使用线上平台交流对个股和市场的看法时也将影响其他人的投资决策,本文正是利用舆情因子、构建多因子模型。

1. 分词

汉语分词分为两个类别:以字典为基础的词语划分和以机器学习或深度学习算法为基础的词语分词。本文使用的 Jieba 是 Python 中用来为中文分词的库, Jieba 基于概率模型,根据预设好的字典生成输入段落(或句子)的有向无环图(Directed Acyclic Graph, DAG)。DAG 构成了字典(dict), key 为一个词的起始位置, data 为一个词可能的终止位置所构成的集合。常见的精确模式和全模式的流程示例如图 2。

2. 因子构建

本文使用基于情感字典的方式构建个股舆情因子,思路为:

- (1)对语料分词,找到语料种的情感词、带有否定意义的词语和能反映程度的词汇。
- (2)判断程度词或否定词后是否出现情感词,出现该种情况则需要将以上三种词视为一个部分。
- (3)如果有程度副词则该部分分值为情感词汇分数乘以程度副词的程度值。

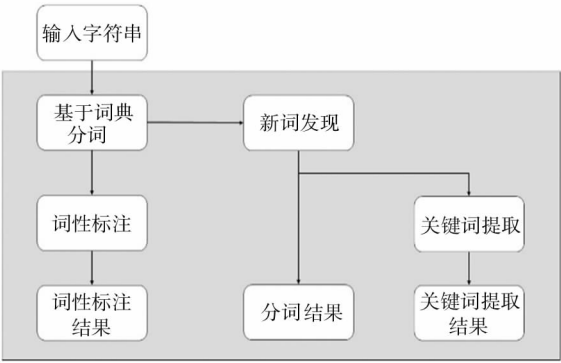


图 2 Jieba 分词流程图

- (4)将所有组的得分加总,若该语料文本否定词的个数为奇数则将文本情感极性归为负值,反之为正。
- (5)将得到的分数除以该段文字的总长度,从而获取因子值。

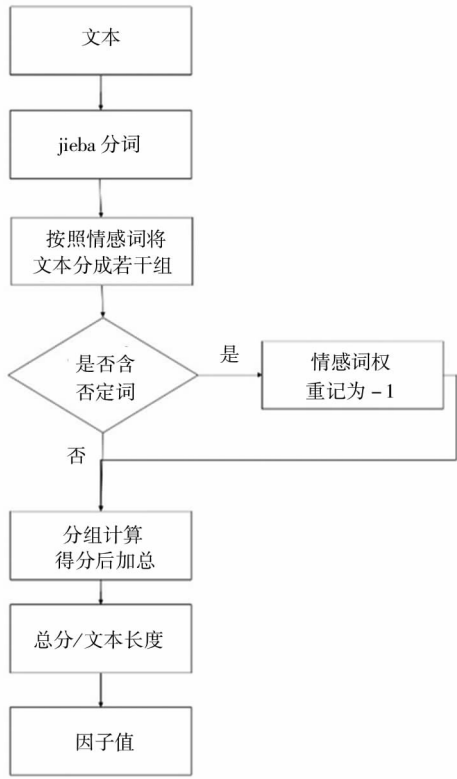


图 3 因子构建流程图

本文中情感词汇分数来自 BosonNLP 情感字典。程度词主要来自知网的情感分析用词语集,部分来自本人总结的金融领域常用程度词(见图 1),并赋予 0~2 范围大小的权重。停用词来自于 Github 上开源的 snownlp 的停用词库。

表 1 BosonNLP 情感字典概览

消极情感词	分数	中性情感词	分数	积极情感词	分数
最尼玛	-6.704	冒昧	0.328	超有爱	4.755
扰民	-6.498	董事会	0.328	Happy	4.792
fuck. . .	-6.330	教育部	0.328	变废为宝	4.820
RNM	-6.219	厚望	0.329	心想事成	4.825
wcnmlgb	-5.967	造诣	0.329	阖家幸福	5.151
Fxxk	-5.872	运势	0.329	特邀	5.262
MLP	-5.872	储蓄	0.329	薄礼	5.285
吃哑巴亏	-5.771	静脉曲张	0.329	敬赠	5.426
IAQI	-5.771	消费	0.330	永葆	5.426
MLGBD	-5.694	有增无减	0.330	携起手来	5.615

3. 因子检验方法

本文在因子测试时对规模 and 行业因子做了剔除,假设初始因子暴露由三个因素组成:行业因子、市值因子,和去除行业、市值影响的因子暴露,即:

$$\begin{bmatrix} \beta_{iu} \\ \vdots \\ \beta_{in} \end{bmatrix} = \begin{bmatrix} I_{ilu} \cdots I_{ilv} & M_{ilm} \\ \vdots & \vdots \\ I_{inu} \cdots I_{inv} & M_{imn} \end{bmatrix} \begin{bmatrix} f_{ii} \\ \vdots \\ f_{in} \end{bmatrix} + \begin{bmatrix} u_{ii} \\ \vdots \\ u_{in} \end{bmatrix} \quad (9)$$

其中: β 代表股票 i 在所测试因子上的原始因子值; I_{ilu} 表示股票 i 的行业值,本文的行业标准选用了申万行业清单,而 I 为哑变量,其取值取决于股票 i 是否属于某一行业,是即为 1,否即为 0; M 代表股票 i 的市值因子暴露; u 则是真正去除行业 and 市值影响后的因子暴露,后文的单因子检验也以 u 为基础的。

判断因子是否合适所采用的方法通常是 IC 测度法。 IC 表示的是当期回报率与上期因子的相关系数,数学表述为:

$$IC_{t,i} = Corr(X_{t-1,i}, r_t) \quad (10)$$

X 表示因子暴露(应为去除行业 and 市值影响的暴露), r 表示收益率。下文将使用 Spearman 的计算方式计算上期因子值与当期回报率的相关系数 IC 。

本文选用了 IC 平均值以及 IR 。 IC 的绝对值越大,说明当期因子能更好地与下期的回报率产生正向或负向的相关性,该因子就越有效。选用 IR 指标可评价 IC 绝对值序列在尽可能大的情况下的稳定性:

$$IR = abs \left(\frac{\frac{1}{N} \sum IC_i}{\frac{1}{N-1} \sum \left(IC_i - \frac{1}{N} \sum IC_i \right)^2} \right) \quad (11)$$

综上所述,好的因子是 IC and IR 都较大的因子,这些因子可以提供 alpha 收益,而 IC and IR 较小的因子则会影响收益,造成较大的波动。

4. 显性舆情因子研究

本文选取的股票池为上证 50 成分股,回测时间区间为 2019 年 1 月 1 日至 2020 年 10 月 1 日。本文希望将量化策略应用到真正的白马股上,因此文章最将回测标的锁定在上证 50 和沪深 300 等高质量公司上。

(1) 数据预处理

本文爬取的数据是东方财富股吧评论的正文部分,发现数据存在以下几方面问题:

- ①“名不副实”:股友在一只股票的股吧下谈论另一只股票。
- ②评论中存在系统推送的该只股票当日的融资融券信息、行情信息、广告等。
- ③评论重复。

表 2 为浦发银行(600000. SH)于 2019 年 1 月 10 日的评论节选,其中无用信息较多,因此在计算舆情因子前有必要对内容进行预处理。

问题 1 的解决办法:利用金融数据库(Choice 终端或 Wind 终端)将 A 股所有上市公司的代码 and 简称导出,根据习惯将一部分简称再简化,如“贵州茅台”可转化成“茅台”,得到字段名为“代码”“简称”和“其他简称”的股票清单,剔除不匹配的评论。

问题 2 的解决办法:挑选上述推送信息的关键词,剔除含有关键词的评论内容。本文删除掉的关键词包括:“净值播报”“融资融券信息”“数据显示”“法律责任”“图片来源”“图表”“本公司董事会及全体董事”。

表 2 浦发银行(600000.SH)评论节选(2019-01-10)

存在问题	评论内容
无	拉升浦发,功德无量
无	估值修复,就在 2019
问题 1	赶紧买入 000401,有肉吃
问题 2	华宝银行 ETF(代码:512800)净值播报…
问题 2	数据显示,截至 01 月 09 日 XXXX 股票沪深股通共计持有 3.74 亿股,比上一交易日增加 XXX.XX 万股。

问题 3 的解决办法:利用程序去除当日该只股票评论的重复内容。

理后,得到了因子的原始数据(未标准化、未中性化),表 3 为其中 20 只股票的面板数据描述性统计。

(2)舆情因子描述统计

按照前文因子构建方法对预处理后的文本处

表 3 舆情因子描述性统计(简化版)

	count	mean	std	min	25%	50%	75%
浦发银行	635	0.07	0.09	-0.37	0.01	0.06	0.12
上海机场	607	0.05	0.08	-0.32	0.00	0.04	0.08
民生银行	636	0.07	0.09	-0.13	0.02	0.06	0.11
中国石化	637	0.06	0.07	-0.30	0.02	0.05	0.09
中信证券	626	0.06	0.07	-0.25	0.01	0.04	0.09
三一重工	622	0.06	0.08	-0.17	0.02	0.05	0.09
招商银行	625	0.06	0.11	-0.57	0.00	0.05	0.11
保利地产	618	0.06	0.14	-0.34	-0.01	0.04	0.08
中国联通	629	0.04	0.10	-0.36	-0.01	0.03	0.07
上汽集团	625	0.05	0.09	-0.31	0.00	0.04	0.10
复星医药	593	0.05	0.09	-0.41	0.00	0.04	0.08
恒瑞医药	633	0.07	0.10	-0.55	0.02	0.06	0.12
万华化学	625	0.07	0.07	-0.13	0.02	0.07	0.11
贵州茅台	625	0.08	0.06	-0.10	0.05	0.07	0.11
山东黄金	631	0.05	0.06	-0.20	0.01	0.04	0.07
海螺水泥	628	0.07	0.09	-0.20	0.02	0.06	0.10
用友网络	601	0.04	0.12	-0.26	-0.03	0.02	0.08
海尔智家	631	0.06	0.09	-0.55	0.02	0.06	0.10
三安光电	622	0.04	0.06	-0.10	0.00	0.03	0.06
闻泰科技	600	0.04	0.07	-0.32	0.00	0.03	0.07

为了验证因子构建方法的有效性,本文选用了股票涨跌幅和指数涨跌幅两个维度进行验证:

①股票涨跌幅

本文将 50 只股票在回测区间的涨跌幅与上证 50 进行对比,并按涨跌幅大小进行分组,部分分组情况见表 4。因子中位数为标的股票在回测区间因子序列的中位数,用于反映股民对标的股票的乐观程度。基准倍数是指在回测区间内股票的涨跌幅是基准(上证 50)的 X 倍。

按涨跌幅将 50 只股票分组,分组依据见表

5,分组后各组的因子情况如表 6,涨幅较大的股票的因子平均值明显高于涨幅较低的组(4、5)组,从这个角度看舆情因子值与股票涨跌幅呈正比关系,且结果较为准确。

②指数涨跌幅

本小节将从指数每日涨跌幅的角度入手,检验因子构建的有效性。根据上证 50 指数在回测区间内的涨跌幅降序排列,得到分组如表 7。为了使各组内天数近似呈正态分布,分组标准如表 8,分组结果如表 9。

表 4 股票涨跌幅情况及分组情况(简化版)

简称	因子中位数	区间涨幅/%	基准倍数	组别
中信建投	0.063	481.464	12	1
闻泰科技	0.040	453.629	11	1
隆基股份	0.062	437.584	11	1
红塔证券	0.048	431.154	11	1
中国中免	0.043	274.157	7	2
三一重工	0.051	210.034	5	2
用友网络	0.049	207.475	5	2
贵州茅台	0.036	190.331	5	2
...	...	...	...	...
上证 50	上证 50 指数	40.963	1	
...	...	...	...	...
交通银行	0.050	-12.565	0	5
中国石化	0.056	-13.047	0	5
邮储银行	0.031	-14.563	0	5
中国铁建	0.040	-20.277	0	5
上汽集团	0.049	-20.665	-1	5
中国石油	0.026	-39.248	-1	5

表 5 股票分组标准

组别	分组标准
1	高于基准 10 倍
2	高于基准 5~10 倍
3	高于基准 1~5 倍
4	正收益但低于基准
5	负收益

表 6 分组因子值(股票)

组别	股票数	因子平均值
1	4	0.053
2	4	0.045
3	19	0.047
4	13	0.035
5	10	0.038

表 7 指数涨跌幅及分组情况(简化版)

时间	因子中位数	当日上证 50 涨跌幅/%	组别
2020 年 07 月 06 日	0.039	6.805	1
2019 年 02 月 25 日	0.044	6.276	1
2019 年 03 月 31 日	0.088	3.823	1
2019 年 05 月 12 日	0.103	3.519	1
2019 年 06 月 20 日	0.046	3.514	1
2019 年 04 月 16 日	0.042	3.354	1
2020 年 07 月 20 日	0.037	3.202	1
2020 年 03 月 24 日	0.031	3.116	1
...	...	...	...
2019 年 03 月 10 日	0.125	-3.732	9
2020 年 03 月 16 日	0.011	-3.743	9
2020 年 07 月 26 日	0.086	-3.831	9
2020 年 07 月 16 日	0.013	-4.579	9
2019 年 05 月 06 日	0.019	-4.761	9
2020 年 02 月 03 日	0.032	-7.004	9

表 8 指数涨跌幅分组标准

组别	指数涨跌幅 x 范围
1	$x \geq 2.5$
2	$2 \leq x < 2.5$
3	$1.5 \leq x < 2$
4	$1 \leq x < 1.5$
5	$0 \leq x < 1$
6	$-0.5 \leq x < 0$
7	$-1.5 \leq x < -0.5$
8	$-3 \leq x < -1.5$
9	$x < -3$

表 9 分组因子值(指数)

组别	分组因子平均值	天数
1	0.056	14
2	0.058	15
3	0.051	15
4	0.050	32
5	0.049	164
6	0.041	102
7	0.043	69
8	0.045	39
9	0.049	8

整体来看,指数涨幅较大的 1~5 组(涨幅大于 0),其因子值也较大,另一方面,6~9 组(涨幅小于 0)因子值较小,因此从指数来看,因子的构建较为合适。

从个体来看,第 2 组因子值较高于第 1 组,第 9 组因子值也略高,上述情况出现在涨跌幅度较大的时间里,原因可能在于舆情因子在市场波动剧烈时效果要稍逊于波动较小的时段。在涨幅小于 0 的组别中(6~9 组)因子值出现了倒挂,说明舆情因子在识别差股票能力上弱于识别好股票。

(3) 舆情因子 IC 与 IR

由于计算 IC 的过程中存在时间差,即因子暴露和预期收益率存在时间间隔,本文选用了月频的收益率与舆情因子计算的 IC、IR 值。回测范围内的月 IC 值如图 4 所示。

在回测范围内(2019 年 1 月至 2020 年 9 月)月频显性舆情因子的 IC 均值为 0.038,标准差为 0.154, IC 序列均值绝对值大于 0.03,因子较好。从图形来看,舆情因子存在一定波动,尤其发生在 2020 年,主要原因可能为当年股票市场的巨大波动。IC 均值大都大于 0.1,说明因子值与收益率依然保持着较好的相关性。因此,本文构建的舆情因子是比较有效的。

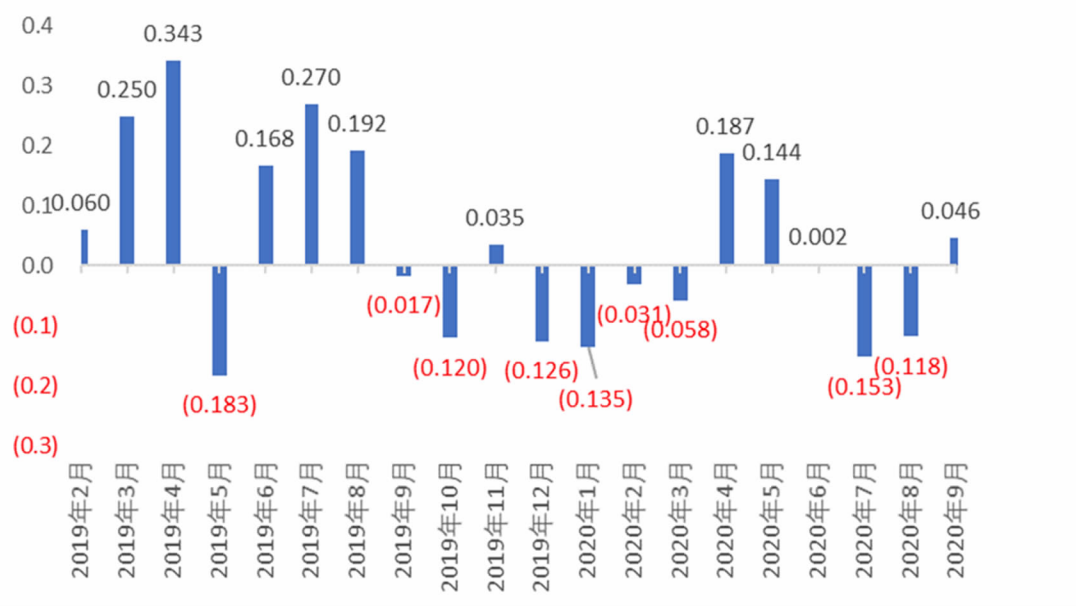


图 4 舆情因子月 IC 值

五、结合舆情因子的多因子研究

下文使用如上所述的收益—风险模型构建结合舆情因子的多因子指数增强组合,从收益的角度完善 Barra 模型。

1. 模型概述

本文采用的收益模型形式为传统的预期收益模型

w^T \mu (12)

式中:w^T 为股票权重向量,而 \mu 为预期收益率。

预期收益率 \mu 选用的是 alpha 因子的预期收益率:

\mu = IC^T X (13)

式中:IC^T 表示因子的 IC 值序列组成的向量,而 X 为因子暴露。IC 序列与因子暴露序列的向量积便可以很好地表示预期收益。

本文采用 Barra 模型具体表述为:已知有股票 S_1, S_2, \dots, S_n, 它们的收益率作为随机变量分别为 R_1, R_2, \dots, R_n。

E(R_i) = \mu_i (14)

多因子的分解如下:

R_1 = \beta_{11} F_1 + \dots + \beta_{1k} F_k + \epsilon_1
R_2 = \beta_{21} F_1 + \dots + \beta_{2k} F_k + \epsilon_2
...
R_n = \beta_{n1} F_1 + \dots + \beta_{nk} F_k + \epsilon_n (15)

这里每个因子 F_i 也是随机变量,而且具有一个协方差矩阵 \Lambda。因此股票收益 R_n 构成的协方

差矩阵应该满足

\Omega = B \Lambda B^T + S (16)

其中矩阵 B 是由所有的系数 \beta_{ij} 组成, S 是个对角矩阵,

S = \begin{bmatrix} \sigma_1^2 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \sigma_n^2 \end{bmatrix} (17)

给定一组权重 \omega_i, 作为投资组合的收益是 \mu^T \cdot \omega, 作为投资组合的方差也就是本文的风险模型为:

\omega^T \Omega \omega = \omega^T (B \Lambda B^T + S) \omega (18)

Barra 模型假定因子收益率与特质收益率不相关。将投资组合的收益与风险合二为一, 本文使用效用函数为:

\max_{\omega} \mu^T \omega - \frac{1}{2} \lambda \omega^T \Omega \omega \quad (\lambda > 0) (19)

考虑到本文要做的是指数增强组合, 给出如下约束条件:

l^T \omega = 0, \omega^T B = 0, \omega + h \geq 0 (20)

l^T 为全 1 向量, l^T \omega = 0 代表所计算的组合仅考虑在原本的指数基础上实现超额收益。 \omega^T B = 0 的含义是该零权重组合在矩阵 B 内的因子暴露是 0。 h 向量表示的是原指数组合内的股票权重, \omega + h \geq 0 则表示该投资组合内股票不能做空。

本文在收益模型中将会使用 alpha 因子, 即 IC 绝对值较大的因子, 在风险模型中则希望剔除 IC 绝对值较小的因子。

2. 因子特征分析

多因子模型的关键在于因子的选择,本文选用了 wind 金融终端作为底层数据库。但 wind 中

部分因子分组出现重合,为避免分类出现错误,现将这 19 个因子全部归入到行情衍生数据库中,修改分组后的因子(月频)总览如表 10。

表 10 因子总览

	因子 <i>IC</i>	因子个数
行情衍生数据		236
正 <i>IC</i>	0.028	167
中国 A 股能量、量价与压力支撑技术指标	0.040	34
中国 A 股日行情	0.036	7
中国 A 股强弱与趋向技术指标	0.035	26
中国 A 股摆动与反趋向技术指标	0.027	25
中国 A 股资金流向数据	0.022	64
中国 A 股成交量技术指标	0.011	4
中国 A 股 Level2 指标	0.011	7
负 <i>IC</i>	-0.024	69
中国 A 股摆动与反趋向技术指标	-0.008	2
中国 A 股强弱与趋向技术指标	-0.016	11
中国 A 股成交量技术指标	-0.020	12
中国 A 股日行情	-0.020	2
中国 A 股能量、量价与压力支撑技术指标	-0.027	7
中国 A 股资金流向数据	-0.028	30
中国 A 股 Level2 指标	-0.029	5
A 股量化因子库		116
正 <i>IC</i>	0.024	76
趋势风险因子	0.026	2
一致预期因子	0.026	28
收益类技术因子	0.024	28
成交量换手率技术因子	0.022	18
负 <i>IC</i>	-0.016	40
趋势风险因子	-0.008	4
收益类技术因子	-0.015	10
一致预期因子	-0.017	15
成交量换手率技术因子	-0.018	11
财务数据		470
正 <i>IC</i>	0.023	214
中国 A 股现金流量表	0.030	45
中国 A 股资产负债表	0.023	50
中国 A 股 TTM 指标历史数据	0.023	25
中国 A 股利润表	0.021	26
中国 A 股财务衍生指标表	0.019	47
中国 A 股公布重要财务指标	0.017	21
负 <i>IC</i>	-0.024	256
中国 A 股公布重要财务指标	-0.014	4
中国 A 股资产负债表	-0.020	80
中国 A 股现金流量表	-0.023	28
中国 A 股利润表	-0.025	40
中国 A 股财务衍生指标表	-0.027	55
中国 A 股 TTM 指标历史数据	-0.027	49
总计		822

3. 多因子的选择

(1)收益因子

选用相关系数均值绝对值较大、波动较小的因子作为收益因子,相关系数均值绝对值较小、波

动较大的因子作为风险因子。本文选用了 822 个因子中 *IR* 最大的 10 个因子作为收益因子,如表 11。

表 11 收益因子概览

序号	因子名称	<i>IC</i> 均值	<i>IR</i>	一级分类	二级分类
1	S_WEST_EPS_FYGROWTH	0.106	0.806	A 股量化因子库	一致预期因子
2	VOLUME_DIFF_INSTITUTE	0.089	0.771	行情衍生数据	中国 A 股资金流向数据
3	ASFI_S_FA_TANGASSETTOINTDEBT	-0.080	0.771	财务数据	中国 A 股财务衍生指标表
4	ASFI_S_FA_EQUITYTOINTERESTDEBT	-0.072	0.705	财务数据	中国 A 股财务衍生指标表
5	S_TECH_AR	0.064	0.683	行情衍生数据	中国 A 股能量、量价与压力支撑技术指标
6	S_VAL_PE_TTM	0.089	0.670	行情衍生数据	A 股日行情
7	BR_ADJ	0.063	0.665	行情衍生数据	中国 A 股能量、量价与压力支撑技术指标
8	BIAS36_ADJ	0.063	0.631	行情衍生数据	中国 A 股摆动与反趋向技术指标
9	ASCF_CASH_RECP_SG_AND_RS	-0.072	0.616	财务数据	中国 A 股现金流量表
10	ASBS_DEFERRED_TAX_LIAB	-0.059	0.613	财务数据	中国 A 股资产负债表

表 11 相关指标含义如下:

1. “S_WEST_EPS_FYGROWTH”: EPS 的增长

2. $S_WEST_EPS_FYGROWTH = \frac{EPS_{FY2}^{一致预测} - EPS_{FY1}^{一致预测}}{ABS(EPS_{FY1}^{一致预测})}$ (21)

3. “VOLUME_DIFF_INSTITUTE”: 机构量差(含主动被动),即为(特大单买量 - 特大单卖量)

4. “ASFI_S_FA_TANGASSETTOINTDEBT”: 有形资产/带息债务

5. “ASFI_S_FA_EQUITYTOINTEREST-DEBT”: 归属于母公司的股东权益/带息债务

6. “S_TECH_AR”为人气指标

7. $AR = \frac{N \text{ 日内(最高价 - 开盘价)之和}}{N \text{ 日内(开盘价 - 最低价)之和}} \times 100$ (22)

8. “S_VAL_PE_TTM”为滚动市盈率

9. “BR_ADJ”为意愿指标

10. $BR = \frac{N \text{ 日内(最高价 - 前收盘价)之和}}{N \text{ 日内(前收盘价 - 最低价)之和}} \times 100$ (23)

11. “BIAS36_ADJ”

12. $BIAS = \frac{\text{收盘价} - N \text{ 日平均价格}}{N \text{ 日平均价格}} \times 100\%$ (24)

13. “ASCF_CASH_RECP_SG_AND_RS”: 销售商品、提供劳务收到的现金

14. “ASBS_DEFERRED_TAX_LIAB”: 递延所得税负债

(2) 风险因子

风险因子选择中剔除银行类和保险类特有指标,共选用了 5 个风险因子,如表 12。

表 12 风险因子概览

序号	因子名称	<i>IC</i> 均值	<i>IR</i>	一级分类	二级分类
1	ASFI_S_FA_YOY_TR	-0.0004	0.0023	财务数据	中国 A 股财务衍生指标表
2	MONEYFLOW_PCT_VALUE_L	0.0004	0.0025	行情衍生数据	中国 A 股资金流向数据
3	TTMHis_S_FA_EQUITY_MRQ	-0.0005	0.0031	财务数据	中国 A 股 TTM 指标历史数据
4	ASBS_CONTRACT_LIABILITIES	0.0006	0.0040	财务数据	中国 A 股资产负债表
5	ASFI_S_FA_EQUITYTODEBT	0.0006	0.0041	财务数据	中国 A 股财务衍生指标表

表 13 相关指标含义如下:

“ASFI_S_FA_YOY_TR”:营业总收入同比增长率(%)。

“MONEYFLOW_PCT_VALUE_L”:大单资金流向占比(%)。

“TTMHis_S_FA_EQUITY_MRQ”:滚动的归属于母公司的股东权益。

“ASBS_CONTRACT_LIABILITIES”:合同负债。

“ASFI_S_FA_EQUITYTODEBT”:归属于母公司的股东权益/负债合计。

4. 模型结果

本次多因子模型中假定手续费为双边万分之八,初始金额为一千万人民币,设立基准为上证综指,同时每日以开盘价格买入当天持仓更新的股票。股票选择的是上证 50 成分股,回测周期为 2019 年 1 月 1 日至 2020 年 9 月 30 日,调仓周期为两周。

本文在上证 50 股票池的基础上也设定了几条筛选规则:剔除选股日上市前两个月的股票;调仓日时无法买涨停的股票,也无法卖跌停的股票;删除调仓日因退市或停止交易无法交易的股票。

表 13 多因子回测数据(不含舆情因子)

回测收益	多头年化	基准年化	多头夏普	最大回撤	超额年化	超额夏普	超额最大回撤
35.98%	20.64%	17.26%	0.65	-33.36%	3.38%	0.15	-27.30%



图 5 多因子回测图(不含舆情因子)

图 5 中蓝色线为策略结果,黑色线为基准(上证综指)情况,绿色线为超额收益。第一个计算期末结束时,为了使用计算日当日前推两周的滚动 IC 值,本模型不开仓。从策略结果看,上文所用模型的年化收益为 20.64%,超过基准但表

现平平。

以下将把舆情因子投入到多因子模型中使用以便获得更好的收益,现模型共有 11 个收益因子,5 个风险因子,其回测效果如表 14 和图 6。

表 14 多因子回测数据(含舆情因子)

回测收益	多头年化	基准年化	多头夏普	最大回撤	超额年化	超额夏普	超额最大回撤
57.67%	33.09%	17.26%	1.02	-34.97%	15.83%	0.63	-29.38%

将舆情因子加入收益模型后,年化收益达到了 33.09%,比未加入舆情因子的模型高 12.46%。综上所述,多因子模型的策略是比较成功的,同时本文构建的因子也是相对有效的。

六、舆情的其他表示及策略

投资者舆情或情绪按照表征特点可以分为显性和隐性两类。本文通过利用情感字典分析股吧



图 6 因子回测图(含舆情因子)

评论得到的情绪指标也属于显性情绪。隐性情绪是通过交易反映的投资者情绪,隐性情绪是根据技术指标及市场情况构建而成的,也称为客观指

标。本章使用了能表征日频情绪的技术因子并将其应用到 Barra 模型中,如表 15。

表 15 隐性情绪指标概述

序号	因子名称	含义
1	S_VAL_PE_TTM	滚动市盈率
2	S_DQ_FREETURNOVER	换手率(自由流通股本)
3	PSY_ADJ	PSY 心理指标(12、6 天)
4	VOSC	VOSC 成交量震荡(12、26 天)
5	VROC_12D	VROC 量变动速率(12 天)

市盈率体现出来投资者对股票的预期。

换手率反映交易活跃程度,同时也可评判投资风格。

PSY 心理指标刻画投资人的心理波动,通常用 N 日内上涨天数的频率来计算,本文选用了 M 日简单平均的方式作滤波,参数 N 为 12(天),参数 M 为 6(天)。

VOSC 指标主要用来分析成交量的变化规律。计算方法为(短期平均成交量 - 长期平均成

交易)/短期平均成交量,其中短/长期平均成交量的时间参数分别为 12(天)和 26(天)。

VROC 也称成交量变动比率,其通过比较当日成交量与 N 日前成交量的关系分析趋势并判断交易信号。参数 N 通常取 12(天)。

模型首先需将因子分为收益因子和风险因子两类,市盈率和自由流通股换手率归类为收益因子; PSY 指标和 VOSC、VROC 的 IC 均值绝对值较低, IR 较低,为风险因子。

表 16 隐性情绪指标因子 IC 及分类情况

序号	因子名称	IC 均值	IR	一级分类	二级分类
1	S_VAL_PE_TTM	0.09	0.67	行情衍生数据	A 股日行情
2	S_DQ_FREETURNOVER	0.03	0.22	行情衍生数据	A 股日行情
3	PSY_ADJ	0.02	0.19	行情衍生数据	A 股能量、量价与压力支撑技术指标
4	VOSC	-0.01	0.12	行情衍生数据	A 股成交量技术指标
5	VROC_12D	-0.01	0.11	行情衍生数据	A 股成交量技术指标

图 7 和表 17 回测结果表明,只使用隐性情绪

最大回撤为 15.88%,超额最大回撤较小,为

因子的策略超额年化可以达到 10.52%,策略的

-5.81%。

表 17 隐性情绪指标模型回测结果

回测收益	多头年化	基准年化	多头夏普	最大回撤	超额年化	超额夏普	超额最大回撤
48.42%	27.78%	17.26%	1.34	-15.88%	10.52%	1.93	-5.81%



图 7 隐性情绪指标回测结果

表 18 显示了显性舆情因子策略与隐性情绪策略结果对比情况。两种策略均取得了相对基准的超额收益;从最大回撤角度看,隐性情绪策略的回撤更小;受到较大回撤及整体波动的影响,显性情绪策略的超额夏普较小。总体而言,隐性策略更为稳健,显性策略收益略佳。

表 18 舆情因子(显性情绪)与隐性情绪策略结果比较

	隐性情绪策略	显性情绪策略
回测收益	48.42%	57.67%
多头年化	27.78%	33.09%
基准年化	17.26%	
多头夏普	1.34	1.02
最大回撤	-15.88%	-34.97%
超额年化	10.52%	15.83%
超额夏普	1.93	0.63
超额最大回撤	-5.81%	-29.38%

本文认为隐性情绪策略更稳健的原因在于文章所度量的显性情绪为中小投资者所有的,其对上证 50 成分股的影响可能不强于机构投资者。隐性情绪更能体现投资者的真实情绪。因此显性情绪策略应用在大市值股票上的收益可能会稍弱于隐性情绪策略。

七、结论

本文主要研究了基于上证 50 成分股的情绪策略,包括显性情绪策略和隐性情绪策略及两者在 Barra 模型中的实现。在研究显性情绪策略时,本文使用基于语义的舆情因子,并验证了因子及策略的有效性,在研究基于交易指标的隐性策略时也发现基于上证 50 成分股的隐性情绪策略好于显性情绪策略。

在显性舆情因子有效性检验上,本文使用的方法是因子分组数据分析以及 IC 检验法,前者的检验分为股票涨跌幅与指数涨跌幅两部分,通过股票和指数涨跌幅大小分组排名的方式验证涨跌幅大小与舆情因子的相关性;而 IC 检验法则计算舆情因子的月频因子值与月频收益率的相关性序列,通过序列均值和标准差验证其确实有效。

在结合显性舆情因子的 Barra 多因子模型有效性检验上,本文以 Barra 多因子组合作为对照组,将 IC 检验法筛选后的舆情因子作为收益因子加入该多因子组合后,证明了舆情因子可以融入 Barra 多因子模型中以获得更高地收益。

在研究隐性情绪策略时,我们选用了滚动市盈率、自由流通股换手率、成交量及衍生指标等高

频数据。根据隐性情绪策略构建的 Barra 模型具有较显性情绪策略更稳健的策略效果。本文的研究可为相关机构投资提供参考借鉴。

参考文献:

[1] MARKOWITZ H. Portfolio Selection[J]. The Journal of Finance, 1952,7(1):77-91.

[2] SHARPE W F. Capital Asset Prices:A Theory of Market Equilibrium Under Conditions of Risk[J]. The Journal of Finance, 1964,19(3):425-442.

[3] ROSS S A. The arbitrage theory of capital asset pricing[J]. Journal of Economic Theory, 1976,13(3):341-360.

[4] FAMA E F, French K R. The Cross-Section of Expected Stock Returns[J]. The Journal of Finance, 1992,47(2):427-465.

[5] FAMA E F, French K R. A five-factor asset pricing model[J]. Journal of Financial Economics, 2015,116(1):1-22.

[6] ASNESS C S. The Interaction of Value and Momentum Strategies[J]. Financial Analysts Journal, 1997,53(2):29-36.

[7] 黄卿. 支持向量机在中国 A 股市场量化策略应用研究——基于 Fama-Fench 三因子模型[J]. 时代金融,2017(11):172-173.

[8] 邹海山,吴勇,吴月珠,等. 中文搜索引擎中的中文信息处理技术[J]. 计算机应用研究,2000(12):21-24.

[9] 孟雪井,孟祥兰,胡杨洋. 基于文本挖掘和百度指数的投资者情绪指数研究[J]. 宏观经济研究,2016(1):144-153.

[10] 孟勇,常静. 投资者情绪对规模效应的影响[J]. 统计与信息论坛,2019(4):14.

[11] 游家兴,吴静. 沉默的螺旋:媒体情绪与资产误定价[J]. 经济研究, 2012(7):141-152.

[12] FRIJNS B, VERSCHOOR W F C, ZWINKELS R C J. Excess stock return comovements and the role of investor sentiment [J]. Journal of International Financial Markets, Institutions and Money, 2017,49:74-87.

[13] 俞庆进,张兵. 投资者有限关注与股票收益——以百度指数作为关注度的一项实证研究[J]. 金融研究,2012(8):152-165.

Factor Mining Under Multi-factor Investment Mode ——Factor Research and Strategy Based on Public Opinion

NIU Xiaojian, JIN Shenghao

(Department of International Finance, Fudan University, Shanghai 200433, China)

Abstract: It is found that public opinion factors have a good correlation with stock returns. The factor test results show that it is feasible to use public opinion factors to construct investment strategies. And the factor values are positively related with stock returns. This paper adds public opinion factors to the multi-factor model to form an index enhancement strategy. The results show that the multi-factor investment strategy combined with public opinion factors can significantly obtain excess returns that exceed the benchmark. This article also uses technical indicators as a substitute for recessive emotions, and explores the use of recessive emotion indicators in the Barra model to obtain a more complete emotional strategy research. The results show that the Barra model combined with recessive strategies can achieve more robustness than explicit strategies.

Keywords: public opinion factor; multi-factor model; Barra Model; index enhancement strategy

(责任编辑:沈建新)