2021年07月09日

新闻舆情数据选股能力初探

——开源量化评论（26）

魏建榕（分析师）
weijianrong@kysec.cn
证书编号：S0790519120001

傅开波（分析师）
fukaibo@kysec.cn
证书编号：S0790520090003

● **因子拥挤导致Alpha衰减，另类数据前景广阔**

伴随着量化投资规模的蓬勃壮大，传统策略的同质化日趋严重，因子拥挤（Factors Crowding）的困境逐渐浮现，最终导致Alpha空间日渐缩窄。另类数据因其孕育的独特Alpha信息，为量化策略的收益提供了新的广袤空间。通联数据（Datayes）作为国内领先的数据智能金融科技公司，旗下的另类数据库种类丰富，其中新闻舆情数据作为其中另类数据的模块之一，蕴含丰富的股票情感信息。通联的新闻舆情数据，主要是对上市公司的新闻进行情感分数打分（SentimentScore），分数值越高通常意味着新闻对该个股有正面情绪。

● **通联新闻舆情数据的基本特征：样本总体略正偏，在财报季时新闻频次较多**

我们对经初筛后共487万条新闻舆情数据进行描述性统计，总体来看，日均出现新闻舆情数据的个股在1600只左右。因新闻舆情数据来源广泛，媒介较多，较能满足总体上的客观公正，我们对所有样本的新闻舆情分数（SentimentScore）进行频数统计，总体来看新闻舆情分数略偏正向；此外在每年的财报季时，新闻频次较多，尤其是4月叠加上市公司年报和一季报时，新闻频次达到了一年最高。

● **新闻舆情均值的变化量在中证500选股域上的绩效表现优异**

我们根据通联新闻舆情数据，计算过去 N 天舆情分数平均值的变化量，记为因子(简记为ΔMS)：ΔMS因子的多空收益比在全样本区间内表现良好，尤其在中证500 选股域上表现优异：在回看天数=20 下，多空收益波动比为 2.2，多头相对中证 500 的年化收益率为 4.6%。

对$\Delta MS_{N=20}$因子进行三种不同换仓频率下的绩效测试：整体表现：双周频>月频>周频。在双周频上该因子的多空对冲年化收益率为 12.00%，，因子的年化 ICIR 为-2.3；月频下，该因子的多空对冲年化收益率 11.92%，因子年化 ICIR 为-2.00。$\Delta MS_{N=20}$因子与过去 20 日涨跌幅的相关性有一定正相关性，相关性接近 0.1，与其余常见因子的相关性较弱，对该因子剔除常见 10 个因子后，剥离得到后的因子在中证 500 选股域上的表现依然优异：多空收益波动比达 2.64，年化 ICIR-2.27，多头相对中证 500 年化收益率 4.86%。

● **风险提示：** 模型测试基于历史数据，市场未来可能发生变化。

金融工程研究团队

魏建榕（首席分析师）
证书编号：S0790519120001

张　翔（分析师）
证书编号：S0790520110001

傅开波（分析师）
证书编号：S0790520090003

高　鹏（分析师）
证书编号：S0790520090002

苏俊豪（研究员）
证书编号：S0790120020012

胡亮勇（研究员）
证书编号：S0790120030040

王志豪（研究员）
证书编号：S0790120070080

盛少成（研究员）
证书编号：S0790121070009

苏　良（研究员）
证书编号：S0790121070008

相关研究报告

《开源量化评论（21）-从茅指数动力学研判抱团现状》-2021.3.30
《开源量化评论（22）-年报展望全扫描：基金经理如何看后市》-2021.4.6
《开源量化评论（23）-“金股+”组合的量化方案》-2021.4.26
《开源量化评论（24）-上市公司招聘数据的选股能力》-2021.5.12
《开源量化评论(25)-业绩超预期 Plus 组合的构建》-2021.7.7》

目　录

图表目录

1、 因子拥挤导致Alpha衰减，另类数据前景广阔

在传统的量化投资领域，模型处理的信息通常来自财务数据和量价数据。然而伴随着量化投资规模的蓬勃壮大，传统策略的同质化日趋严重，因子拥挤（Factors Crowding）的困境逐渐浮现，最终导致Alpha空间日渐缩窄。另类数据因其孕育的独特Alpha信息，为量化策略的收益提供了新的广袤空间。在另类数据的研究领域，开源证券金融工程团队进行了一系列的研究，包括《上市公司招聘数据选股能力初探》、《高频股东数据的隐含信息量》、《北上资金高频数据的CTA潜力》、《从托管机构细窥北向资金选股能力》、《上市公司招聘数据的选股能力》等。

本报告将对新闻舆情数据的选股能力进行初步探索。我们基于通联数据提供的新闻舆情数据，进行定量分析，以期探寻其中隐含的Alpha信息。通联数据（Datayes）作为国内领先的数据智能金融科技公司，旗下的另类数据库种类丰富，其中新闻舆情数据作为其中另类数据的模块之一，蕴含丰富的股票情感信息。

2、 新闻舆情数据的基本特征

2.1、 新闻舆情数据的样例

我们使用通联数据中“getNewsRelatedScoreV2”（获取新闻情感信息）和“新闻关联标签行业表”（getNewsTagInd）这两张表，进行新闻舆情原始数据集的构建。

通联的新闻舆情数据，主要是对上市公司的新闻进行情感分数打分（SentimentScore），分数值越高通常意味着新闻对该个股有正面情绪。通联新闻舆情数据的数据实例和表字段含义如图1和表1所示。

图1：通联新闻舆情数据样例[1]

secShortName	degreeProp1St	degreeProp2St	relatedCompanyDegree	relatedCompanyScore	sentiment	sentimentScore	effectiveTime	newsGenre	isEconom	isPolicy	industryName1st	industryName2nd	newsTitle
国泰君安	0.71	0.26	1	0.61	0	0.1	2020/8/19 11:59	普通新闻	0	0	stock	非银金融	因战略安排…
宜宾纸业	1	0	1	0.5	1	0.65	2021/3/1 9:38	普通新闻	0	0	stock	轻工制造	造纸板块直线拉升…
特变电工	0.07	0.92	2	0.95	0	0.11	2020/12/24 17:16	普通新闻	0	0	stock	电气设备	特变电工：拟4亿元投资…
新研股份	0.09	0.9	2	0.95	1	0.57	2021/3/30 10:13	普通新闻	0	0	stock	国防军工	新研股份早盘大涨5.64% 量
雄韬股份	0.44	0.55	2	0.77	0	0.01	2020/8/12 9:47	普通新闻	0	0	stock	银行	投资者提问：请问定增是否
南钢股份	0	1	2	1	-1	-0.4	2020/6/4 8:59	普通新闻	0	0	stock	钢铁	南钢股份去杠杆负债率降30
大为股份	0	1	2	1	0	0	2020/8/1 6:13	公告新闻	0	0	stock	汽车	深圳市特尔佳科技股份有限
永艺股份	0.96	0.01	1	0.48	0	0.1	2020/10/13 9:31	普通新闻	0	0	stock	非银金融	A股10月13日房企股开盘：南
新天然气	0	0.99	2	0.99	0	0.03	2021/3/24 1:50	公告新闻	0	0	stock	采掘	新疆鑫泰天然气股份有限公
中科信息	0.68	0.3	1	0.64	0	0	2020/7/10 14:21	普通新闻	0	0	stock	银行	投资者提问：董秘您好：请

资料来源：通联数据、开源证券研究所

表1：通联新闻舆情数据的字段含义

参数名	参数类型	参数说明
newsID	Int64	新闻ID
ticker	String	证券交易代码
secShortName	String	证券简称
degreeProp1St	Double	关联等级为1的置信度

[1] 由于部分新闻数据描述的是行业或宏观经济，因此在对数据进行初筛时，保留属于上市公司的新闻。此外部分上市公司定期财务报告的信息也频频出现在各家新闻媒体里，因此在清洗原始数据集时，已对数据进行初筛，删除包括属于月度数据、是定期报告、属于图片新闻的记录。

参数名	参数类型	参数说明
degreeProp2St	Double	关联等级为2的置信度
relatedCompanyDegree	Int16	0-不关联，1-弱关联，2-强关联
relatedCompanyScore	Double	关联程度
sentiment	Int16	情感分类0中性，-1负 面，1正面
sentimentScore	Double	情感打分
effectiveTime	Date	新闻有效发布时间
newsGenre	String	普通新闻、价格动态、公告新闻三类
isEconom	Int16	是否包含基本面信息
isPolicy	Int16	是否是国家发布的政策
industryName1st	String	行业新闻（industry）、公司新闻（stock）、 宏观新闻（marco）、债券新闻（bond）、市场新闻（market）、其他新闻（other）
industryName2nd	String	如果一级类别是“行业新闻”，则细分到具体行业， 包括申万一级27个行业（不包含“综合”）
newsTitle	String	新闻标题

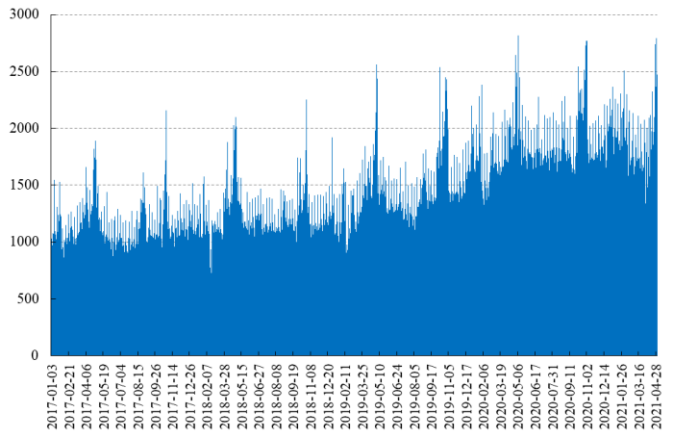
资料来源：通联数据、开源证券研究所

2.2、 新闻舆情分数总体略偏正向，样本分布具有月度效应

我们对经初筛后共487万条新闻舆情数据进行描述性统计，总体来看，日均出现新闻舆情数据的个股在1600只左右（图2）。

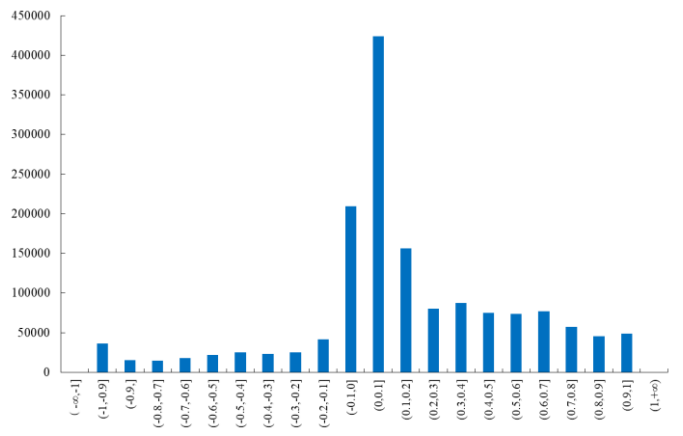
因新闻舆情数据来源广泛，媒介较多，较能满足总体上的客观公正，我们对所有样本的新闻舆情分数（SentimentScore）进行频数统计（图3），**总体来看新闻舆情分数略偏正向**。

图2：日均有新闻舆情的个股约1600只



数据来源：通联数据、开源证券研究所

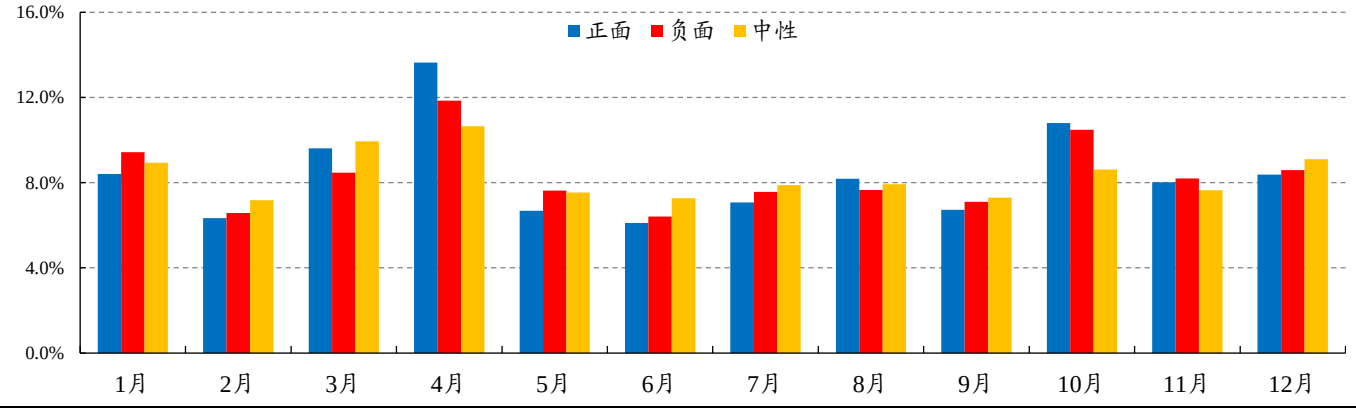
图3：新闻舆情分数的分布总体偏正



数据来源：通联数据、开源证券研究所

此外，我们对新闻舆情（Sentiment）的样本分布进行月度效应的统计（图4），在每年的财报季时，新闻频次较多，尤其是4月叠加上市公司年报和一季报时，新闻频次达到了一年最高。

图4：新闻舆情呈现一定的月度效应：新闻舆情样本数在每年财报季偏多，尤其是每年4月

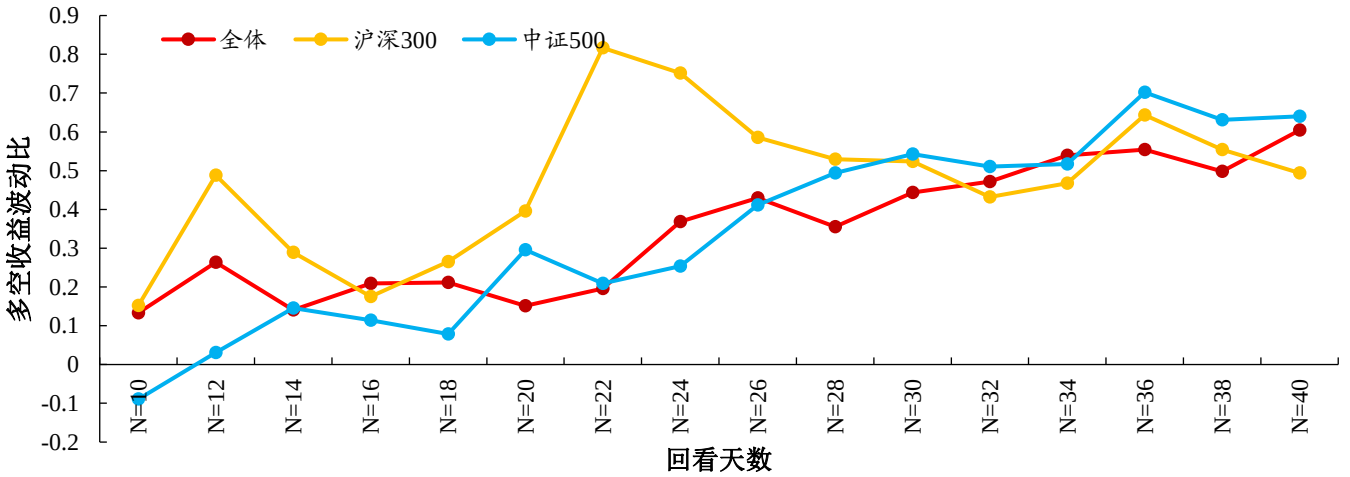


数据来源：通联数据、开源证券研究所

3、 ΔMS_N因子在中证500选股域上表现优异

舆情值蕴含市场对于该股的情绪，我们用过去N天舆情分数的平均值（mean sentiment，简记为$MS_{t-N,t}$）作为选股因子进行单因子测试。从效果上来看，不管是全样本、沪深300还是中证500选股域上看，回测结果均不尽如人意（图5）。

图5：仅用过去N天舆情分数的平均值作为选股因子（$MS_{t-N,t}$），选股效果不尽如人意



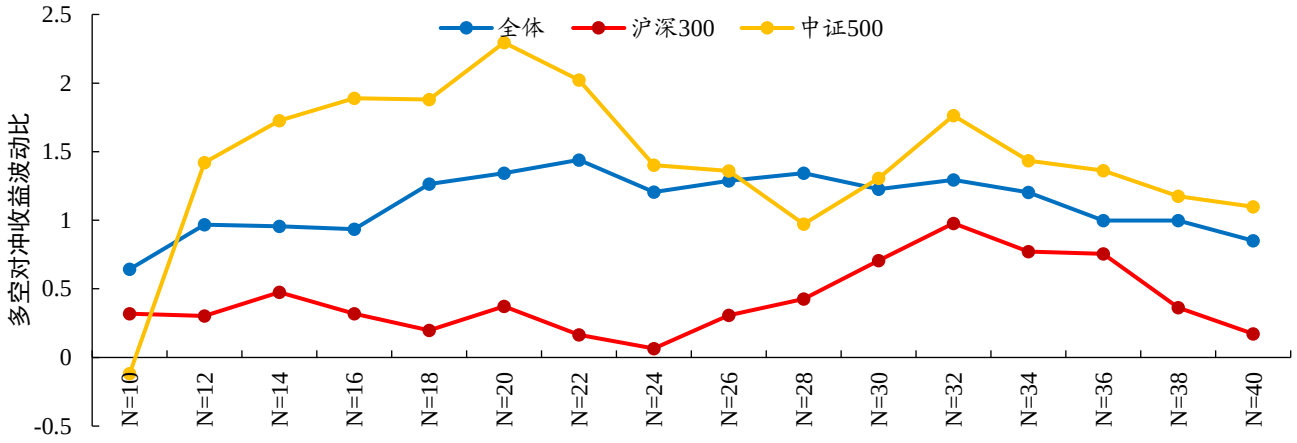
数据来源：通联数据、开源证券研究所（手续费设置为双边千三）

为此，我们对因子$MS_{t-N,t}$上做一定的变化，记过去N天舆情分数平均值的变化量记为我们新因子ΔMS_N：

$$\Delta MS_N = MS_{t-N,t} - MS_{t-2N-1,t-N-1}$$

我们对ΔMS_N进行单因子测试，图6为在不同回看天数N下，三个不同样本域下的选股效果。从图6可以看到：该因子的多空收益比在全样本区间内表现良好，尤其在中证500选股域上表现优异。

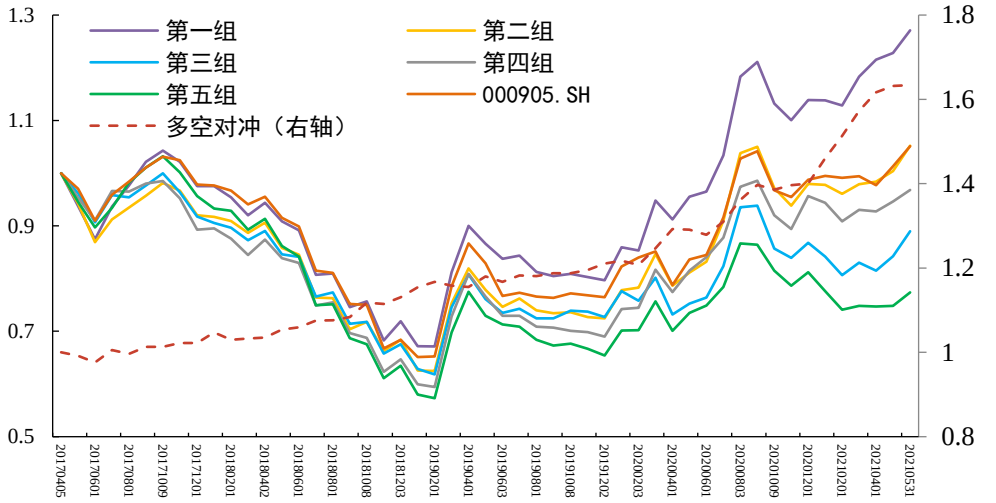
图6：新因子ΔMS_N在中证500选股域上表现出色



数据来源：通联数据、开源证券研究所（手续费设置为双边千三）

为了更进一步展现ΔMS_N因子在中证500上优秀的选股效果，我们取回看天数N=20进行中证500选股域上的测试。图7为按月调仓下，$\Delta MS_{N=20}$因子在中证500上五分组和多空对冲净值的走势。从中可以看出，该因子在多空以及多头组上的表现优异：多空收益波动比为2.2，多头相对中证500的年化收益率为4.6%。

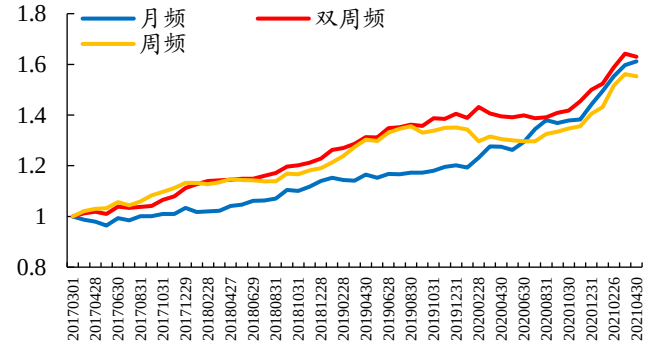
图7：按月调仓下，$\Delta MS_{N=20}$因子在中证500选股域上多空收益波动比2.2



数据来源：通联数据、开源证券研究所（手续费设置为双边千三）

另外，我们对$\Delta MS_{N=20}$因子进行三种不同换仓频率下的绩效测试（手续费设置为双边千三），如图 8 和图 9 所示，该因子在三种换仓频率上的整体表现：双周频>月频>周频。在双周频上该因子的多空对冲年化收益率为 12.00%，多头相对中证 500 年化收益率为 1.72%，因子的年化 ICIR 为-2.3；**月频下，该因子的多空对冲年化收益率 11.92%，多头相对中证 500 的年化收益率为 4.61%，因子年化 ICIR 为-2.00**。

图8：三种调仓频率下的多空净值：双周频>月频>周频



数据来源：通联数据、开源证券研究所（手续费设置为双边千三）

图9：三种调仓频率下的绩效指标

指标\调仓频率	月频	双周频	周频
IC均值	-0.032	-0.024	-0.017
rankIC均值	-0.029	-0.023	-0.017
年化收益率(多空对冲)	11.92%	12.00%	10.74%
收益波动比(多空对冲)	2.166	2.649	2.037
年化收益率（多头）	2.58%	3.43%	0.15%
收益波动比（多头）	0.057	0.027	0.030
年化收益率（空头）	27.77%	12.51%	14.42%
年化收益率（多头相对中证500）	4.61%	1.72%	1.74%
收益波动比（多头相对中证500）	0.609	0.321	0.253
年化ICIR	-2.009	-2.311	-2.394

数据来源：通联数据、开源证券研究所（手续费设置为双边千三）

此外，我们对该因子与常见因子进行相关性分析，从表2可以看到：**$\Delta MS_{N=20}$因子与过去20日涨跌幅的相关性有一定正相关性，相关性接近0.1，与其余因子的相关性较弱**。

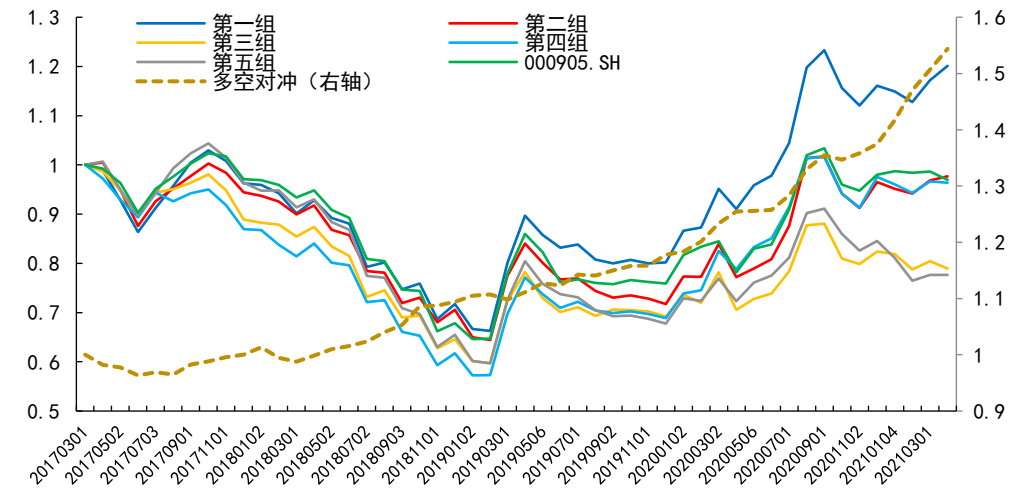
表2：$\Delta MS_{N=20}$与常见因子的相关性：因子与过去 20 日涨跌幅因子相关性较高，与其余因子相关性较弱

因子名称	20日涨跌幅	Beta	价值	杠杆	盈利	成长	流动性	规模	非线性市值	波动
相关性	0.098	-0.005	-0.005	0.003	0.000	-0.004	-0.006	0.005	-0.002	0.005

数据来源：通联数据、开源证券研究所

我们对$\Delta MS_{N=20}$因子进行剔除表2中的10个因子，剥离得到后的因子在中证500选股域上的表现依然优异（图10和表3）：**多空收益波动比达2.64，年化ICIR-2.27，多头相对中证500年化收益率4.86%**。

图10：$\Delta MS_{N=20}$因子在剔除常见因子后，在中证500上的表现依然优异



数据来源：通联数据、开源证券研究所（手续费设置为双边千三）

表3：$\Delta MS_{N=20}$因子剥离掉常见因子后在中证500上的绩效表现：多空收益波动比2.64，年化ICIR为-2.27

指标	第一组	第二组	第三组	第四组	第五组	多空对冲	中证500指数	多头对冲基准
年化收益率	5.89%	0.89%	-3.64%	0.23%	-5.14%	11.60%	1.00%	4.86%
年化波动率	20.39%	19.99%	19.67%	19.93%	19.75%	4.40%	19.04%	6.94%

指标	第一组	第二组	第三组	第四组	第五组	多空对冲	中证500指数	多头对冲基准
收益波动比	0.29	0.04	-0.19	0.01	-0.26	2.64	0.05	0.70
最大回撤	35.62%	35.85%	39.50%	41.11%	42.80%	2.50%	36.88%	4.84%
胜率	50.98%	47.06%	45.10%	45.10%	43.14%	80.39%	47.06%	56.86%
盈亏比	1.29	1.26	1.13	1.33	1.15	1.60	1.27	1.35

数据来源：通联数据、开源证券研究所

4、 风险提示

模型测试基于历史数据，市场未来可能发生变化。

特别声明

《证券期货投资者适当性管理办法》、《证券经营机构投资者适当性管理实施指引（试行）》已于2017年7月1日起正式实施。根据上述规定，开源证券评定此研报的风险等级为R3（中风险），因此通过公共平台推送的研报其适用的投资者类别仅限定为专业投资者及风险承受能力为C3、C4、C5的普通投资者。若您并非专业投资者及风险承受能力为C3、C4、C5的普通投资者，请取消阅读，请勿收藏、接收或使用本研报中的任何信息。
因此受限于访问权限的设置，若给您造成不便，烦请见谅！感谢您给予的理解与配合。

分析师承诺

负责准备本报告以及撰写本报告的所有研究分析师或工作人员在此保证，本研究报告中关于任何发行商或证券所发表的观点均如实反映分析人员的个人观点。负责准备本报告的分析师获取报酬的评判因素包括研究的质量和准确性、客户的反馈、竞争性因素以及开源证券股份有限公司的整体收益。所有研究分析师或工作人员保证他们报酬的任何一部分不曾与，不与，也将不会与本报告中具体的推荐意见或观点有直接或间接的联系。

股票投资评级说明

	评级	说明
证券评级	买入（Buy）	预计相对强于市场表现20%以上；
	增持（outperform）	预计相对强于市场表现5%～20%；
	中性（Neutral）	预计相对市场表现在－5%～＋5%之间波动；
	减持（underperform）	预计相对弱于市场表现5%以下。
行业评级	看好（overweight）	预计行业超越整体市场表现；
	中性（Neutral）	预计行业与整体市场表现基本持平；
	看淡（underperform）	预计行业弱于整体市场表现。
备注：评级标准为以报告日后的6~12个月内，证券相对于市场基准指数的涨跌幅表现，其中A股基准指数为沪深300指数、港股基准指数为恒生指数、新三板基准指数为三板成指（针对协议转让标的）或三板做市指数（针对做市转让标的）、美股基准指数为标普500或纳斯达克综合指数。我们在此提醒您，不同证券研究机构采用不同的评级术语及评级标准。我们采用的是相对评级体系，表示投资的相对比重建议；投资者买入或者卖出证券的决定取决于个人的实际情况，比如当前的持仓结构以及其他需要考虑的因素。投资者应阅读整篇报告，以获取比较完整的观点与信息，不应仅仅依靠投资评级来推断结论。		

分析、估值方法的局限性说明

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。本报告采用的各种估值方法及模型均有其局限性，估值结果不保证所涉及证券能够在该价格交易。

法律声明

开源证券股份有限公司是经中国证监会批准设立的证券经营机构，已具备证券投资咨询业务资格。