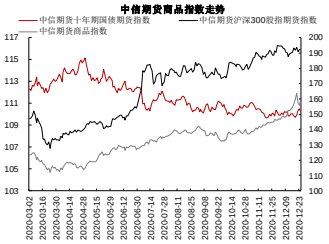
中信期货 CITIC Futures

中信期货研究|量价策略专题报告

2020-12-26

股票因子策略专题报告

投资咨询业务资格：
证监许可【2012】669 号



报告要点

本文回顾了因子和因子策略研究的历史，梳理了 Barra 和 WorldQuant 量价因子体系，并检验了因子表现。本文也对基于两大因子体系的多因子模型进行了检验。

摘要：

综述：对因子和因子研究做了回顾，介绍了两大有代表性的因子体系。

因子评价和策略的构建方法：从两个角度来构建因子，一是 Barra 的大类风格因子体系， 二是基于 WorldQuant Alpha101 和其他类似研究的算法挖掘/机器学习因子。对单因子进行 RankIC 和 RankIC_IR 排序，对多因子模型进行回归统计检验。

最近一年的单因子表现：Barra 风格因子体系中累积收益率范围、历史残差波动率、月换手率的表现最好，RankIC 均值最高（绝对值）为 0.033，RankIC_IR 均值最高（绝对值）为 0.2。算法挖掘/机器学习因子体系中，趋势和反转类中均有表现较好的因子，RankIC 均值最高（绝对值）为 0.02，RankIC_IR 均值最高（绝对值）为 0.257。

中信一级行业的单因子表现：按照中信一级行业分类方法，共分 30 个行业，在每个细分一级行业中，计算并展示 Barra 风格因子体系以及算法挖掘/机器学习因子体系中的单因子表现。结果表明，因子的选股能力受到行业的较大影响，不同因子在不同行业中的表现有明显差异。

多因子模型检验：使用最近一年的日频数据，分别验证基于 Barra 风格因子体系和算法挖掘/机器学习因子体系的多因子模型的因子收益率联合起来是否在统计上显著不为 0。接着，同时考虑 Barra 量价因子以及算法挖掘/机器学习因子，进行模型检验。Barra 多因子模型 R 方 0.37，算法因子模型 R 方 0.49，联合模型 R 方 0.59。

量价策略团队

研究员：
张革
021-60812988
zhangge@citicsf.com
从业资格号 F3004355
投资咨询号 Z0010982

目 录

图目录

一、关于因子研究的综述

（一） 研究简述

自 1964 年 William Sharpe 等提出资本资产定价模型（The capital asset pricing model，CAPM）和 1976 年 Stephen Ross 发明套利定价理论（The arbitrage pricing theory，APT）以来，无论是学术界还是工业界，以此为起点，多因子投资研究和实证定价实践得到了广泛的关注和轰轰烈烈的发展。多因子模型可以记为：

$$E[R_i] = \sum_{k=1}^{K} \beta_{ik} * r_k$$

其中β_{ik}是资产 i 在因子 k 上的暴露，r_k是因子收益率。

在这之后，Fama-French 发表了著名的三因子模型，该模型作为 CAPM 模型的补充，提出股票的市场的 beta 值不能解释不同股票回报率的差异，而上市公司的市值、账面市值比、市盈率可以解释股票回报率的差异。Fama-French 利用对全市场股票进行分组来构建因子的投资子组合，然后利用因子的子投组的多空来构建因子，这种开创性的因子构架方法后来也成为众人竞相模仿的对象。

随着计算机技术的进步，开源语言平台迅猛发展，计算机算力也逐年提升，多因子模型的可行性也随之越来越高。这种可行性主要包含两个角度：一是有易用开源的脚本语言来辅助因子的计算和模型回测与优化，这样的语言要能得到广泛使用，且大多数研究人员有能力学习；二是计算机能够提供相应的算力支持，无论是在硬件还是软件上，计算因子、模型优化和投资组合优化（权重优化）在正常情况下运算量都非常大，如果设计的策略无法在有限时间内给出最优权重，交易上根本无法执行，策略本身设计得再好也没有应用价值。

由于投资者对 alpha 的疯狂追求，近年来越来越多的机器学习算法开始参与到传统统计下的多因子模型中，这些算法主要用来帮助因子生成、因子筛选以及组合优化。机器学习因子大多出于对数据的粗暴挖掘，很多学者认为这些因子背后难有坚固的理论支撑，从海量因子中找出的有效因子可能只是运气成分的体现，出现的显著性也经不起时间考验，所谓的 alpha 会很快消失。出于这个原因，一些学者也设计了排除这种运气成分的统计学思路(Harvey et al. (2016))。本着开放的心态，本文认为机器学习因子可能有一定的解释效力，这种效力或许来自于大量机器学习研究人员得出的相似权重并按之交易导致的市场价格结果，但是这类因子必须得到严谨的处理和统计学论证，否则未来的投资结果可能很难复现回测期内的优异表现。

（二） Barra 和 WorldQuant 因子体系简述

Barra 和 WorldQuant 的因子体系是两种比较有代表性的因子体系。Barra 基于市场研究，所有的因子都代表标的股票在某一个维度的暴露，在业界得到了众多机构的广泛使用。WorldQuant 提供的 Alpha 101 体系中，因子更多的来自于计算机算法挖掘

和检验筛选，很多因子可能本身很难解释其具体含义，但统计研究、回测和实盘交易表明这些因子能够带来超额收益。本文主要聚焦于 Barra 和 WorldQuant 因子体系中的量价因子，对这些量价因子和由其形成的策略表现进行跟踪，并尝试提出对因子和策略的创新性看法或思路。

二、因子评价和策略构建方法

（一） 因子构建方法

本文从两个角度来构建因子：

一是 Barra 的大类风格因子体系，在 Barra 中量价类的风格因子比较清晰的反映了权益资产市场表现的某一个维度，背后的含义也非常明显；

图表 1： Barra 大类风格因子体系（量价类）

风格因子 Style Factor	描述符 Descriptor	说明	因子定义
贝塔 Beta	Historical Beta	历史贝塔	股票收益率对沪深 300 收益率的时间序列回归，取回归系数
流动性 Liquidity	Monthly Share Turnover	月换手率	对最近一个月的股票换手率求和，然后取对数
	Quarterly Share Turnover	季换手率	对最近一个季度的股票换手率求和，然后取对数
	Annual Share Turnover	年换手率	对最近一年的股票换手率求和，然后取对数
	Annualized Traded Value Ratio	年化交易额比率	对最近一年的日换手率进行指数加权求和
长期反转 Long-Term Reversal	Long-term Relative Strength	长期相对强度	计算非滞后的长期相对强度：对股票对数收益率进行加权求和，然后计算滞后交易日的时间窗口内的非滞后值的等权平均值，最后取相反数
	Long-term Historical Alpha	长期历史Alpha	计算非滞后的alpha：取CAPM回归截距项，然后计算滞后交易日的时间窗口内的非滞后值的等权平均值，最后取相反数
规模 Size	Log of Market Capitalization	市值规模	流通市值的自然对数
中市值 Mid Capitalization	Cube of Size Exposure	中等市值	取规模因子的立方，然后对规模因子正交，最后进行去极值和标准化处理
动量 Momentum	Relative Strength 12-month	年相对强度	计算非滞后的相对强度：对股票的对数收益率进行指数加权求和，然后计算滞后交易日的时间窗口内的非滞后相对强度的等权平均值
	Historical Alpha	历史Alpha	在计算贝塔所进行的时间序列回归中取回归截距项，然后计算滞后交易日的时间窗口内的非滞后值的等权平均值
残差波动率 Residual Volatility	Historical Sigma	历史残差波动率	在计算贝塔所进行的时间序列回归中，取回归残差收益率的波动率
	Daily Standard Deviation	日收益率标准差	最近一年日收益率的波动率
	Cumulative Range	累积收益率范围	最近12个月累积对数收益率的最大值减去最小值

资料来源：Barra 中信期货研究部

二是基于 WorldQaunt Alpha101 和其他类似研究的算法挖掘/机器学习因子。如前文所说，这些因子本身可能很难讲出具体的含义，所产生的效果也容易受到 data mining 的影响，在后期这类因子可能会需要严谨的处理和统计学论证。本系列参照早期的多重检验方法或 Harvey et al. (2016)的办法进行验证。

图表 2： WorldQuant 算法挖掘/机器学习因子体系（部分）

因子	说明	因子公式化定义
Alpha#6	趋势类	(-1 * correlation(open, volume, 10))
Alpha#9	反转类	((0 < ts_min(delta(close, 1), 5)) ? delta(close, 1) : ((ts_max(delta(close, 1), 5) < 0) ? delta(close, 1) : (-1 * delta(close, 1))))
Alpha#21	趋势类	((((sum(close, 8) / 8) + stddev(close, 8)) < (sum(close, 2) / 2)) ? (-1 * 1) : (((sum(close,2) / 2) < ((sum(close, 8) / 8) - stddev(close, 8))) ? 1 : (((1 < (volume / adv20)) || ((volume /adv20) == 1)) ? 1 : (-1 * 1))))
Alpha#23	反转类	(((sum(high, 20) / 20) < high) ? (-1 * delta(high, 2)) : 0)
Alpha#28	反转类	scale(((correlation(adv20, low, 5) + ((high + low) / 2)) - close))
Alpha#32	趋势类	(scale(((sum(close, 7) / 7) - close)) + (20 * scale(correlation(vwap, delay(close, 5), 230))))
Alpha#43	反转类	(ts_rank((volume / adv20), 20) * ts_rank((-1 * delta(close, 7)), 8))
Alpha#46	反转类	((0.25 < (((delay(close, 20) - delay(close, 10)) / 10) - ((delay(close, 10) - close) / 10))) ? (-1 * 1) : (((((delay(close, 20) - delay(close, 10)) / 10) - ((delay(close, 10) - close) / 10)) < 0) ? 1 : ((-1 * 1) * (close - delay(close, 1)))))
Alpha#49	反转类	(((((delay(close, 20) - delay(close, 10)) / 10) - ((delay(close, 10) - close) / 10)) < (-1 * 0.1)) ? 1 : ((-1 * 1) * (close - delay(close, 1))))
Alpha#51	反转类	(((((delay(close, 20) - delay(close, 10)) / 10) - ((delay(close, 10) - close) / 10)) < (-1 * 0.05)) ? 1 : ((-1 * 1) * (close - delay(close, 1))))
Alpha#53	反转类	(-1 * delta((((close - low) - (high - close)) / (close - low)), 9))
Alpha#54	反转类	((-1 * ((low - close) * (open^5))) / ((low - high) * (close^5)))

资料来源：WorldQaunt 中信期货研究部

Barra 的大类风格因子体系中大致有 7 类量价因子，分别是贝塔、流动性、长期反转、规模、中市值和动量，所有因子均清晰的反映了一个维度的特征，计算方法也比较便于代码实现。

基于 WorldQaunt Alpha101 和其他类似研究的算法挖掘/机器学习因子无法清晰的看出是反应了个股哪一方向的特征，本文将其大致归为趋势和反转两类，趋势类因子的因子收益率为正，投组应超配这些因子；反转类因子的因子收益率为负，投组应尽量降低在这些因子上的暴露。基于因子意义的明晰度和计算机算力的限制，本文选取了部分计算相对方便且没有明显机器学习过拟合挖掘情况（例如因子在计量单位上就没有意义）的因子作为算法挖掘因子的代表。

图表 3： WorldQuant 算法挖掘/机器学习因子体系使用公式一览

rank(x)	cross-sectional rank
delay(x, d)	value of x d days ago
correlation(x, y, d)	time-serial correlation of x and y for the past d days
covariance(x, y, d)	time-serial covariance of x and y for the past d days
scale(x, a)	rescaled x such that sum(abs(x)) = a (the default is a = 1)
delta(x, d)	today's value of x minus the value of x d days ago
signedpower(x, a)	x^a
decay_linear(x, d)	weighted moving average over the past d days with linearly decaying weights d, d – 1, …, 1 (rescaled to sum up to 1)
indneutralize(x, g)	x cross-sectionally neutralized against groups g (subindustries, industries, sectors, etc.), i.e., x is cross-sectionally demeaned within each group g
ts_{O}(x, d)	operator O applied across the time-series for the past d days, non-integer number of days d is converted to floor(d)
ts_min(x, d)	time-series min over the past d days
ts_max(x, d)	time-series max over the past d days
ts_argmax(x, d)	which day ts_max(x, d) occurred on
ts_argmin(x, d)	which day ts_min(x, d) occurred on
ts_rank(x, d)	time-series rank in the past d days
min(x, d)	ts_min(x, d)
max(x, d)	ts_max(x, d)
sum(x, d)	time-series sum over the past d days
product(x, d)	time-series product over the past d days
stddev(x, d)	moving time-series standard deviation over the past d days

资料来源：WorldQaunt 中信期货研究部

（二） 单因子评价方法

依照现有的因子库，可以直接对相应的单因子进行评价，进而形成相应的策略。这样的策略只关注一个因子对权益收益率的影响，主要关注这些因子 RankIC 值和 RankIC_IR 值。RankIC 代表因子的选股能力，其绝对值越大越好；RankIC_IR 代表因子稳定获得超额收益的选股能力，其绝对值越大越佳。这一过程可以简要描述为：

- 计算图表 1 和图表 2 中所有细分因子的值；
- 在每一个截面上，分别计算属于同一个风格大类（例如动量）的细分因子与个股下期收益率计算 Spearman 秩相关系数，即得到该细分因子的 RankIC 值，在时序上取 RankIC 均值；
- 利用因子的 RankIC 值计算 RankIC_IR 值，RankIC_IR 是某一时间维度的 RankIC 均值除以标准差，时间维度与策略的调仓周期相匹配；
- 计算风格因子的 RankIC 和 RankIC_IR：计算风格大类中所有细分因子的 RankIC 和 RankIC_IR 的算术平均（如需）。

（三） 多因子模型检验和策略构建方法

依照现有的因子库，可以构建多因子策略。回归检验的过程总的来说可以简要概括为：

- 挑选和计算因子，在时间序列上计算个股在一篮子因子上的暴露；

- 对因子进行标准化和正交化处理，这里采用 Schmidt 正交化法，以排除因子之间的相互影响；
- 通过截面回归，找到个股（超额）收益率均值和因子暴露在截面上的关系；
- 计算每个个股的定价错误，联合检验是否在统计上为零。

无论投资者用什么因子（基本面因子、宏观经济因子、技术面因子），不管在确定截面关系时采用何种回归方式，对多因子模型的最终评判都转化成判断因子收益率联合起来是否在统计上显著不为零。现在一般使用 Fama-MacBeth 两步回归法。

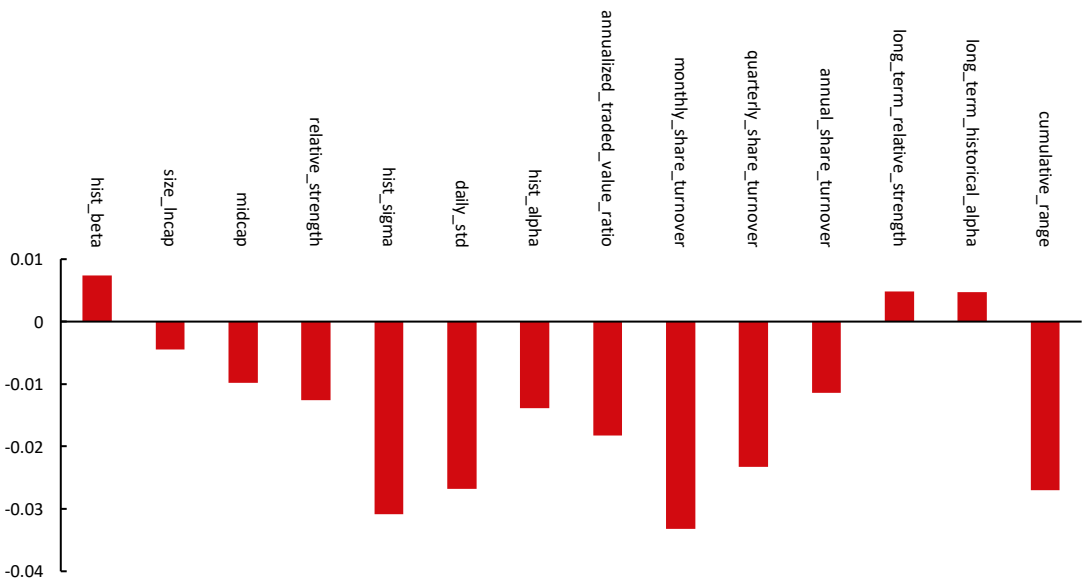
进行回归检验后，可以清晰地看到哪些因子是显著的，哪些因子不是，通过对因子收益率的时间序列进行统计分析，最终判定该因子能否在长期稳定的贡献超额收益，进而构建相应的组合多空策略，即通过控制组合在风险因子上的暴露，增加组合在收益因子上的暴露，降低组合在负收益因子上的暴露来达到实现超额收益的目的。

三、最近一年的单因子表现（全行业）

（一） Barra 风格因子表现

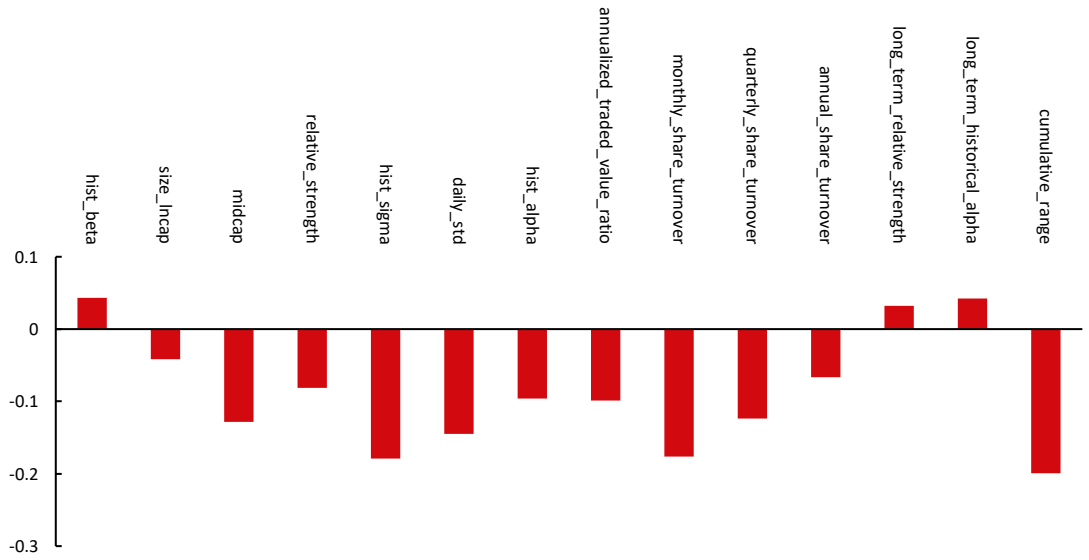
使用近一年的数据回测，按因子 RankIC，Barra 风格因子里面选股能力前 3 位的是月换手率，历史残差波动率，累积收益率范围，策略应选择在这些因子上载荷尽量低的股票；按因子 RankIC_IR，Barra 风格因子里面选股能力前 3 位的是累积收益率范围，历史残差波动率，月换手率，策略应选择在这些因子上载荷尽量低的股票。

图表 4： Barra 风格因子 RankIC 均值



资料来源：同花顺 中信期货研究部

图表 5： Barra 风格因子 RankIC_IR

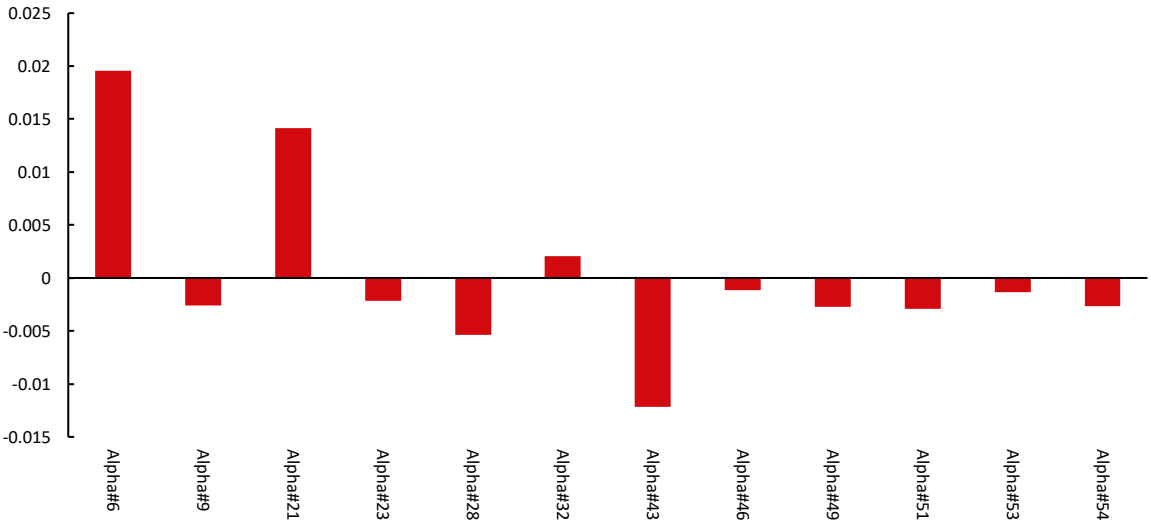


资料来源：同花顺 中信期货研究部

（二） 算法挖掘/机器学习因子表现

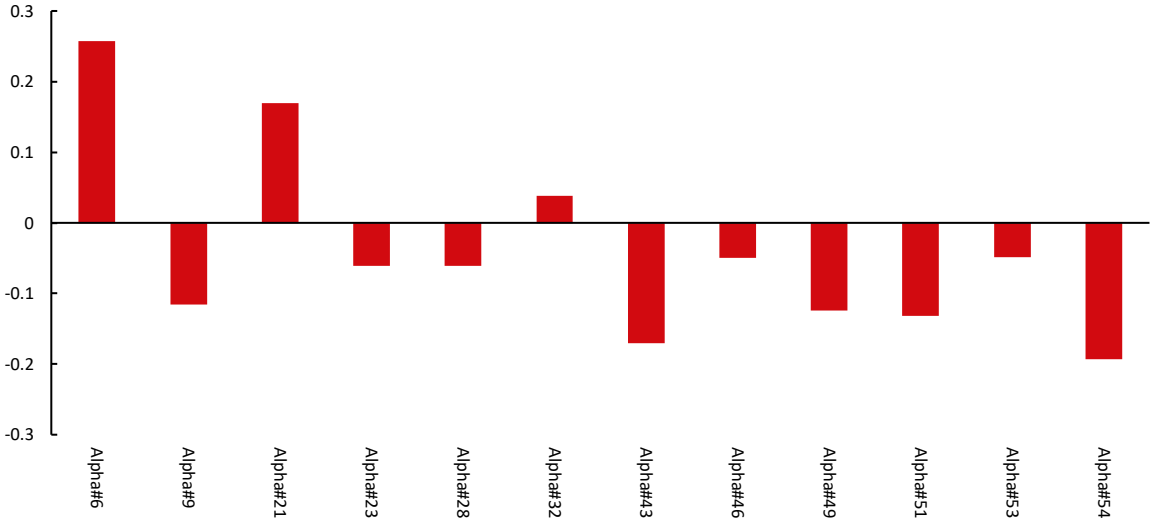
使用近一年的数据回测，按因子 RankIC 绝对值大小，算法挖掘/机器学习因子里面选股能力前 3 位的是 Alpha#6,Alpha#21,Alpha#43,策略应选择在 Alpha#6,Alpha#21 上载荷尽量高，同时在 Alpha#43 上载荷比较低的股票；按因子 RankIC_IR，算法挖掘/机器学习因子里面选股能力前 3 位的是 Alpha#6，Alpha#54，Alpha#43，策略应选择在 Alpha#6 上载荷尽量高，同时在 Alpha#54，Alpha#43 上载荷比较低的股票。

图表 6： 算法挖掘/机器学习因子 RankIC 均值



资料来源：同花顺 中信期货研究部

图表 7： 算法挖掘/机器学习因子 RankIC_IR



资料来源：同花顺 中信期货研究部

四、中信一级行业的单因子表现

（一） Barra 风格因子表现

使用中信一级行业分类，本文首先计算基于 Barra 风格因子体系的单因子 RankIC 和 RankIC_IR 值。可以发现，不同行业的最佳因子，无论是按 RankIC 还是按 RankIC_IR，均差异明显。这可能是由于不同行业间上市公司本身特质差别较大，以此导致的上市公司权益回报率对因子的敏感性的天差地别。

图表 8： Barra 风格因子表现（中信一级行业）

指数代码	成分股数	行业分类	最佳因子（按RankIC均值）	RankIC均值	最佳因子（按RankIC_IR）	RankIC_IR
CI005001	47	石油石化	size_lncap	0.081	size_lncap	0.359
CI005002	36	煤炭	cumulative_range	0.041	cumulative_range	0.214
CI005003	111	有色金属	hist_beta	0.068	size_lncap	0.300
CI005004	170	电力及公用事业	hist_sigma	0.086	cumulative_range	0.387
CI005005	52	钢铁	size_lncap	0.050	midcap	0.274
CI005006	360	基础化工	hist_sigma	0.057	cumulative_range	0.341
CI005007	132	建筑	hist_sigma	0.104	hist_sigma	0.485
CI005008	81	建材	cumulative_range	0.061	cumulative_range	0.331
CI005009	128	轻工制造	hist_sigma	0.066	hist_sigma	0.397
CI005010	405	机械	quarterly_share_turnover	0.047	cumulative_range	0.287
CI005011	257	电力设备及新能源	quarterly_share_turnover	0.051	long_term_historical_alpha	0.270

CI005012	86	国防军工	monthly_share_turnover	0.071	hist_sigma	0.384
CI005013	172	汽车	cumulative_range	0.036	cumulative_range	0.312
CI005014	112	商贸零售	cumulative_range	0.092	cumulative_range	0.631
CI005015	50	消费者服务	monthly_share_turnover	0.098	monthly_share_turnover	0.425
CI005016	75	家电	hist_sigma	0.059	cumulative_range	0.275
CI005017	89	纺织服装	hist_sigma	0.099	hist_sigma	0.758
CI005018	352	医药	daily_std	0.090	hist_sigma	0.401
CI005019	112	食品饮料	hist_sigma	0.083	hist_sigma	0.558
CI005020	88	农林牧渔	hist_alpha	0.068	hist_alpha	0.331
CI005021	37	银行	long_term_relative_strength	0.103	size_lncap	0.261
CI005022	69	非银行金融	hist_sigma	0.119	hist_sigma	0.512
CI005023	127	房地产	monthly_share_turnover	0.094	monthly_share_turnover	0.652
CI005024	116	交通运输	monthly_share_turnover	0.066	hist_alpha	0.406
CI005025	287	电子	quarterly_share_turnover	0.033	quarterly_share_turnover	0.197
CI005026	120	通信	hist_alpha	0.059	hist_alpha	0.441
CI005027	261	计算机	hist_sigma	0.055	hist_sigma	0.393
CI005028	149	传媒	monthly_share_turnover	0.074	cumulative_range	0.374
CI005029	57	综合	cumulative_range	0.062	cumulative_range	0.472
CI005030	17	综合金融	monthly_share_turnover	0.067	monthly_share_turnover	0.199

资料来源：同花顺　中信期货研究部

（二） 算法挖掘/机器学习体系因子表现

使用中信一级行业分类，本文接着计算基于算法挖掘/机器学习因子体系的单因子 RankIC 和 RankIC_IR 值。可以发现，不同行业的最佳因子，无论是按 RankIC 还是按 RankIC_IR,均差异明显。这亦可能是由于不同行业间上市公司本身特质差别较大，将单一因子均匀的应用于全市场选股会具有较大的风险。

图表 9： 算法挖掘/机器学习体系因子表现（中信一级行业）

指数代码	成分股数	行业分类	最佳因子（按RankIC均值）	RankIC均值	最佳因子（按RankIC_IR）	RankIC_IR
CI005001	47	石油石化	alpha028	0.062	alpha028	0.378
CI005002	36	煤炭	alpha043	0.071	alpha043	0.328
CI005003	111	有色金属	alpha053	0.049	alpha053	0.422
CI005004	170	电力及公用事业	alpha021	0.051	alpha021	0.371
CI005005	52	钢铁	alpha043	0.083	alpha043	0.503
CI005006	360	基础化工	alpha101	0.039	alpha101	0.387
CI005007	132	建筑	alpha101	0.060	alpha101	0.446
CI005008	81	建材	alpha021	0.064	alpha021	0.403
CI005009	128	轻工制造	alpha032	0.035	alpha046	0.309

CI005010	405	机械	alpha101	0.025	alpha021	0.216
CI005011	257	电力设备及新能源	alpha043	0.043	alpha043	0.403
CI005012	86	国防军工	alpha043	0.042	alpha043	0.340
CI005013	172	汽车	alpha101	0.052	alpha101	0.388
CI005014	112	商贸零售	alpha032	0.083	alpha032	0.671
CI005015	50	消费者服务	alpha006	0.070	alpha046	0.400
CI005016	75	家电	alpha043	0.039	alpha043	0.291
CI005017	89	纺织服装	alpha032	0.076	alpha032	0.608
CI005018	352	医药	alpha032	0.041	alpha021	0.433
CI005019	112	食品饮料	alpha032	0.043	alpha032	0.267
CI005020	88	农林牧渔	alpha021	0.067	alpha021	0.478
CI005021	37	银行	alpha032	0.089	alpha032	0.305
CI005022	69	非银行金融	alpha021	0.077	alpha021	0.439
CI005023	127	房地产	alpha101	0.072	alpha101	0.616
CI005024	116	交通运输	alpha023	0.048	alpha023	0.350
CI005025	287	电子	alpha021	0.020	alpha021	0.199
CI005026	120	通信	alpha028	0.024	alpha053	0.234
CI005027	261	计算机	alpha101	0.028	alpha051	0.230
CI005028	149	传媒	alpha101	0.029	alpha101	0.214
CI005029	57	综合	alpha028	0.057	alpha021	0.429
CI005030	17	综合金融	alpha021	0.067	alpha021	0.478

资料来源：同花顺　中信期货研究部

五、多因子模型检验

（一） 因子池和数据选择

本文首先分别验证基于 Barra 风格因子体系和算法挖掘/机器学习因子体系的多因子模型的因子收益率联合起来是否在统计上显著不为 0。首先选取 Barra 风格因子池中的所有因子，进行多因子模型检验，即 Fama-MacBeth 两步回归。接着选取算法挖掘/机器学习因子池中的所有因子进行统计检验，同样采用两步回归法。最后，我们同时考虑 Barra 量价因子以及算法挖掘/机器学习因子，进行模型检验。

本文选取的数据是最近一年的日频数据。

（二） Barra 体系多因子模型检验

按照第二章（三）多因子模型检验的方法，本段对 Barra 多因子模型进行检验，得到下面结果：

图表 10：Barra 体系多因子模型检验结果

Item	Estimate	Std.Error	z-value	Pr(>|z|)	Signif.codes

(Intercept)	-4.0026e-03	2.3093e-03	-1.7332	0.083051	*
hist_beta	1.3000e-02	1.7119e-03	7.5941	3.100e-14	****
size_lncap	7.3257e-04	9.7475e-05	7.5154	5.673e-14	****
midcap	7.2055e-04	5.4087e-05	13.3220	<2.2e-16	****
relative_strength	-5.7030e-02	1.9400e-03	-29.3976	<2.2e-16	****
hist_sigma	4.8735e-01	1.7467e-01	2.7901	0.005269	***
daily_std	-1.0801e+00	1.9942e-01	-5.4163	6.086e-08	****
hist_alpha	1.1305e+01	3.0925e-01	36.5564	<2.2e-16	****
annualized_traded_value_ratio	-8.6202e-04	1.2206e-04	-7.0625	1.635e-12	****
monthly_share_turnover	3.2936e-03	4.9519e-04	6.6511	2.908e-11	****
quarterly_share_turnover	1.0184e-03	4.1490e-04	2.4546	0.014106	**
annual_share_turnover	-1.9242e-03	3.0340e-04	-6.3420	2.268e-10	****
long_term_relative_strength	1.6926e-02	8.7868e-04	19.2626	<2.2e-16	****
long_term_historical_alpha	-7.8898e-01	1.3856e-01	5.6941	1.241e-08	****
cumulative_range	1.2528e-03	5.1558e-04	2.4300	0.015101	**

Signif. codes: 0 '****' 0.001 '***' 0.01 '**' 0.05 '*' 0.1
Multiple R-squared: 0.37281
资料来源：Wind 中信期货研究部

Barra 体系中共有 14 个量价因子，模型 R 方为 0.37，显示模型具有一定的解释能力。绝大多数变量在 0.05 的置信度下显著，变量的显著性较强。

（三） 算法挖掘/机器学习因子体系模型检验

按照第二章（三）多因子模型检验的方法，本段对算法挖掘/机器学习多因子模型进行检验，得到下面结果：

图表 11：算法挖掘/机器学习多因子模型检验结果

Item	Estimate	Std.Error	z-value	Pr(>|z|)	Signif.codes
(Intercept)	1.3477e+00	7.5254e-02	17.9085	<2.2e-16	****
Alpha#6	1.0134e-01	2.9802e-02	3.4005	0.0006726	****
Alpha#9	-4.9648e-04	7.6654e-05	-6.4770	9.357e-11	****
Alpha#21	-5.9746e-01	1.9312e-02	-30.9377	<2.2e-16	****
Alpha#23	-3.1298e-02	2.8719e-03	-10.8982	<2.2e-16	****
Alpha#28	-1.1757e+03	2.7619e+01	-42.5702	<2.2e-16	****
Alpha#32	1.7032e+01	6.2583e-01	27.2142	<2.2e-16	****
Alpha#43	-1.4228e-02	5.2476e-04	-27.1128	<2.2e-16	****
Alpha#46	-5.7723e-03	8.4118e-04	-6.8621	6.784e-12	****
Alpha#49	-4.5868e-03	1.1551e-03	-3.9711	7.156e-05	****
Alpha#51	4.2908e-03	1.1325e-03	3.7888	0.0001514	****
Alpha#53	3.2223e-02	1.0382e-02	3.1039	0.0019100	***
Alpha#54	2.6859e-16	3.1619e-16	0.8495	0.3956253	*

Signif. codes: 0 '****' 0.001 '***' 0.01 '**' 0.05 '*' 1

Multiple R-squared: 0.49187
资料来源：Wind　中信期货研究部

本段选取的算法挖掘/机器学习因子体系中共有 12 个量价因子，模型 R 方为 0.49，显示模型具有一定的解释能力。除 Alpha#53、Alpha#54 外，所有因子均在 0.001 的置信度下显著，变量的显著性非常强。仅从统计检验的结果看，算法挖掘/机器学习量价因子体系比业界广泛使用的 Barra 因子体系解释能力更强，变量也更显著。但是，这可能是受到数据挖掘影响的过拟合结果，海量的信息挖掘找到显著因子的机率非常高。

（四） “Barra”+“算法挖掘/机器学习”因子体系模型检验

按照第二章（三）多因子模型检验的方法，本段同时选择 Barra 风格因子和算法挖掘/机器学习因子，进行多因子模型进行检验，得到下面结果：

图表 12：“Barra”+“算法挖掘/机器学习”多因子模型检验结果

Item	Estimate	Std.Error	z-value	Pr(>|z|)	Signif.codes
(Intercept)	2.0668e-03	5.2469e-04	3.9390	8.183e-05	*****
Alpha#6	6.1873e-04	1.3237e-03	0.4674	0.6401863	*
Alpha#9	-3.1830e-01	1.2206e-01	-2.6077	0.0091152	****
Alpha#21	2.4066e-03	1.3912e-03	1.7299	0.0836559	**
Alpha#23	-2.1301e-02	5.6601e-03	-3.7633	0.0001677	*****
Alpha#28	7.0604e-03	7.5498e-03	0.9352	0.3496968	*
Alpha#32	9.1286e-03	4.8593e-03	1.8786	0.0603015	**
Alpha#43	-1.1987e-02	3.6760e-03	-3.2608	0.0011110	****
Alpha#46	-1.8073e-01	6.8454e-02	-2.6402	0.0082854	****
Alpha#49	4.3530e-02	1.6523e-02	2.6345	0.0084265	****
Alpha#51	3.3947e-02	1.4381e-02	2.3605	0.0182484	***
Alpha#53	2.9652e+00	1.1441e+00	2.5917	0.0095516	****
hist_beta	-3.4607e-03	4.8289e-04	-7.1666	7.687e-13	*****
size_lncap	1.6509e-03	3.3159e-04	4.9787	6.402e-07	*****
midcap	2.5744e-03	7.5267e-05	34.2031	<2.2e-16	*****
relative_strength	1.7693e-03	1.2259e-04	14.4328	<2.2e-16	*****
hist_sigma	7.6929e-05	1.4798e-04	0.5198	0.6031694	*
daily_std	2.1285e-04	7.2032e-05	2.9550	0.0031265	****
hist_alpha	-4.9812e-03	1.7027e-04	-29.2545	<2.2e-16	*****
annualized_traded_value_ratio	7.3484e-05	6.7859e-05	1.0829	0.2788525	*
monthly_share_turnover	-6.9422e-04	1.4027e-04	-4.9492	7.453e-07	*****
quarterly_share_turnover	8.9843e-04	1.0826e-04	8.2990	<2.2e-16	*****
annual_share_turnover	-2.7870e-04	8.2914e-05	-3.3612	0.0007759	*****
long_term_relative_strength	-1.0981e-03	9.0541e-05	-12.1277	<2.2e-16	*****
long_term_historical_alpha	-1.2072e-03	5.5796e-05	-21.6369	<2.2e-16	*****
cumulative_range	-5.3674e-04	6.9236e-05	-7.7524	9.019e-15	*****

Signif. codes: 0 ‘*****’ 0.001 ‘****’ 0.01 ‘***’ 0.05 ‘**’ 0.1 ‘*’ 1
Multiple R-squared: 0.59089
资料来源：Wind 中信期货研究部

同时考虑 Barra 风格因子和算法挖掘/机器学习因子给模型带来了更强的解释能力，本段选取的模型 R 方高达 0.59，显示模型具有较强的解释能力，这些因子从统计上看，能够解释绝大部分收益的来源。

考虑了 Barra 风格因子以后，算法挖掘/机器学习因子的显著性整体上出现了比较显著的下降，这说明算法因子可能确实受到了数据挖掘的影响，导致模型产生了过拟合的现象，在加上 Barra 风格因子后这些算法因子便不再特别显著。这也从侧面印证了本文之前的观点，即算法因子必须得到严谨的处理和统计学论证。本文建议，可以尝试参照早期的多重检验方法或 Harvey et al. (2016)的办法进行验证，以在层出不穷的因子中排除纯靠 data mining 挖掘的、找到真正能够解释股票预期收益截面差异的因子。

免责声明

除非另有说明，中信期货有限公司拥有本报告的版权和/或其他相关知识产权。未经中信期货有限公司事先书面许可，任何单位或个人不得以任何方式复制、转载、引用、刊登、发表、发行、修改、翻译此报告的全部或部分材料、内容。除非另有说明， 本报告中使用的所有商标、服务标记及标记均为中信期货有限公司所有或经合法授权被许可使用的商标、服务标记及标记。未经中信期货有限公司或商标所有权人的书面许可，任何单位或个人不得使用该商标、服务标记及标记。

如果在任何国家或地区管辖范围内，本报告内容或其适用与任何政府机构、监管机构、自律组织或者清算机构的法律、规则或规定内容相抵触，或者中信期货有限公司未被授权在当地提供这种信息或服务，那么本报告的内容并不意图提供给这些地区的个人或组织，任何个人或组织也不得在当地查看或使用本报告。本报告所载的内容并非适用于所有国家或地区或者适用于所有人。

此报告所载的全部内容仅作参考之用。此报告的内容不构成对任何人的投资建议，且中信期货有限公司不会因接收人收到此报告而视其为客户。

尽管本报告中所包含的信息是我们于发布之时从我们认为可靠的渠道获得，但中信期货有限公司对于本报告所载的信息、观点以及数据的准确性、可靠性、时效性以及完整性不作任何明确或隐含的保证。因此任何人不得对本报告所载的信息、观点以及数据的准确性、可靠性、时效性及完整性产生任何依赖，且中信期货有限公司不对因使用此报告及所载材料而造成的损失承担任 何责任。本报告不应取代个人的独立判断。本报告仅反映编写人的不同设想、见解及分析方法。本报告所载的观点并不代表中信期货有限公司或任何其附属或联营公司的立场。

此报告中所指的投资及服务可能不适合阁下。我们建议阁下如有任何疑问应咨询独立投资顾问。此报告不构成任何投资、法律、会计或税务建议，且不担保任何投资及策略适合阁下。此报告并不构成中信期货有限公司给予阁下的任何私人咨询建议。

中信期货有限公司
深圳总部
地址：深圳市福田区中心三路 8 号卓越时代广场（二期）北座 13 层 1301-1305、14 层
邮编：518048
电话：400-990-8826