证券研究报告

兴业证券 INDUSTRIAL SECURITIES

分析师：
徐寅
xuyinsh@xyzq.com.cn
S0190514070004

研究助理：
郑兆磊
zhengzhaolei@xyzq.com.cn

当线性模型遇见机器学习

2019 年 09 月 17 日

报告关键点

多因子选股体系中不可规避的一个问题是因子的选择，该领域的学术以及商业研究相对较少。作为机器学习系列二研究，我们通过引入机器学习中的特征选择方法，运用 Filter&Wrapper 构建因子选择体系，并构建线性和非线性融合的动态选股因子，选股能力斐然。因子 IC、ICIR、T 值分别为 0.106、1.50、16.06，且因子在各组别中的单调性和稳定性要优于基准的表现。

相关报告

团队成员：

投资要点

- 多因子选股体系中不可规避的一个问题是因子选择，该领域的学术以及商业研究相对较少。作为机器学习系列二研究，我们通过引入机器学习中的特征选择方法，运用 Filter&Wrapper 构建因子选择体系，进一步构建线性和非线性融合的动态选股因子，该因子选股能力斐然。
- 针对于选择的因子通过线性回归的方式构建动态选股因子 DLS：因子 IC、ICIR、T 值分别为 0.106、1.50、16.06，而基准（选择过去 5 年 ICIR 最优因子等权合成）则分别为 0.061、1.16、12.48；分位数测试来看，DLS 因子多空年化收益、夏普率分别达到 43.0%、4.99，同期基准达到 24.3%，3.48，且 DLS 因子在各组别中的单调性和稳定性要优于基准的表现。
- 为了更好的利用非线性选股信息，我们将机器学习系列一构建的集成学习因子引入到线性回归中（相当于在 Filter&Wrapper 选择的因子基础上添加集成学习因子），构建线性和非线性叠加的 A-DLS 因子。A-DLS 因子 IC 达到 0.13、ICIR 为 2.07，多空组合年化收益 69%、夏普率 6.55，整体表现优于集成学习因子。
- 基于 A-DLS 因子构建主动量化以及增强选股策略，各策略表现优异。其中主动量化策略年化超额收益 26.9%、夏普率 4.61；针对于沪深 300 和中证 500 采用线性优化的方式构建指数增强策略，以针对于中证 500 构建的全市场选股增强策略为例：策略超额收益达到 18.4%、风险收益比达到 4.14、最大回撤 3.7%。

风险提示：本报告模型及结论全部基于对历史数据的分析，当市场环境变化时，存在模型失效风险。

目 录

报告正文

1、引言

1.1、效用递减、进退两难

1952 年 Markowitz 建立了以均值方差模型为基础的现代资产组合管理理论(MPT)，该理论确立了金融学收益风险均衡的分析范式，标志着现代金融学的诞生。Black、Sholes 和 Morton 于 1973 年建立了期权定价模型(OPM)，为衍生品定价问题确立了分析范式；Ross 在 1976 年建立了无套利定价理论（APT），构成量化选股的理论基础;进一步 Fama 与其同事 French 在 1992 年提出了 Fama-French 三因子模型。至此，现代量化投资的理论基础的构建大致完成。

在量化理念从混沌初开到如今枝繁叶茂的漫长征途中，出现过因子选股、统计套利、趋势交易等著名的投资模型，也出现过 LTCM、文艺复兴、AQR 等量化巨头公司……但当下不少国内的量化投资从业者却正在面临着越来越激烈的竞争和不断衰退的阿尔法（超额收益）。以量化选股领域为例：1、曾经选股能力强的因子的有效性在不断降低，且波动愈发剧烈；2、以往的数据源、研究视角/研究方法进一步挖掘的边际效用不断递减，投入产出比在持续收窄；3、随着模型和数据同质化的加剧，踩踏风险也变得越来越高。该如何有效的应对这些问题是每个量化从业者面临的难题。

实际上面对上述挑战，已经有业界人士给出了一些值得尝试的方案：为了应对日渐波动的市场和不断切换的风格，可以考虑将动态因子选择机制常态化；为了应对日益丰富的数据来源和越来越大的数据量，可以采用机器学习/人工智能的方法去提升处理数据、探索规律的效率；为了解决传统的静态模型的弊端（1、对非线性规律把握难度较大；2、因子和权重设置相对固定，很难适应国内市场快速切换的现实情况；3、静态模型的搭建需要大量的历史数据做回测以大概率地保证因子的长期有效性），我们可以构建动态选股模型……

正如所有新事物的诞生、成长都需要一个过程，这些新的方法也面临着诸多挑战。质疑者认为：动态因子选择长期来看不够稳定，同时也增加了交易的成本；机器学习方法过于复杂、难以解释、黑箱属性较重；动态模型主要依靠数据驱动，模型的逻辑性和样本外表现都有待于进一步商榷……但无论怎样，正是这些质疑的声音构成了我们前进的动力，也正是这些质疑让量化选股体系变得更加完善。我们有理由相信这些新的方法经过不断的打磨和雕琢，终将成为未来量化选股模型的中流砥柱。届时我们亦不必进退两难，正应了那句话：前途是光明的、道路是坎坷的、尝试是值得褒奖的。

兴业证券金融工程团队将机器学习应用于量化选股体系的研究正是在这样的背景下产生的。我们并不会简单地将每一个机器学习算法都做一些尝试（实际上在 2015 年-2016 年我们有撰写过 5 篇机器学习的深度报告，详细介绍了各种算法在选股领域的应用），而是更加看重基于某个具体角度的深入挖掘，注重研究的延展性和可落地性。继上一篇《基于集成学习算法的量化选股模型研究》在 2019 年 6 月发布之后，我们推出了机器学习系列的第二篇深度报告《当线性遇上非线

性》。本系列的第一篇报告以改进版的 Adaboost 算法为核心，将非线性信息利用、因子动态选择、因子权重确定等一系列问题有效地融合于同一个分类算法之中，取得了非常好的实践效果；而本文则将目光聚焦于更为传统、却也更受主流投资者喜爱的线性多因子体系，并主要在以下两个方面有所突破::

1、借鉴机器学习中 Filter&Wrapper 的特征选择方法，构建了一套系统化的、有效的动态因子选择机制；

2、引入我们之前开发的集成学习因子 E-NELS，将其视作因子对股票收益非线性预测能力的代表，而后将该因子纳入我们的线性模型框架，构建线性、非线性彼此融合的新一代多因子框架。

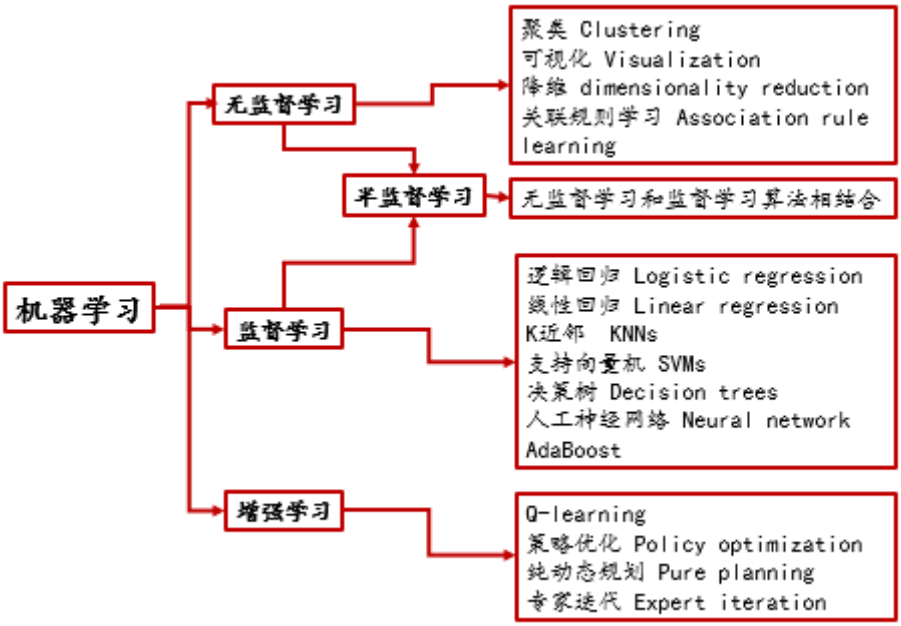
在正式介绍模型之前，我们先将本文的研究测试时段做一个说明：整个研究样本的时段为 2005 年 1 月 4 日-2019 年 7 月 31 日；如不加特殊说明，我们一般采取 60 个月的滚动窗口的实证方式，因此因子和策略的样本外研究自 2009 年 12 月 31 日开始。

2、机器学习之因子选择综述

2.1、机器学习简介

在《基于集成学习算法的量化选股模型研究》的深度报告中，我们详细的对机器学习的大致分类、算法、集成学习的方法等做了详细介绍，并对人工智能/机器学习在投资领域的应用做了简单描述。整体来看机器学习在量化领域的应用越来越广。

图表 1、机器学习算法分类



资料来源：兴业证券经济与金融研究院整理

图表 2、人工智能在投资领域的部分应用案例

时间	事件
2007 年	2007 年，纽约对冲基金 Rebellion Research 推出人工智能产品。
2008 年	首个智能投顾平台诞生：Wealthfront
2011 年	基于 Twitter 用户情绪变化的基金 Derwent Capital Markets 成立。
2013 年	2013 年，Kensho 公司成立并推出数据分析工具 Warren，公司于 2015 年被高盛以 1500 万美元收购。
2015 年	JP Morgan 的对冲基金业务 Highbridge Capital Management 开始与 Sentient 合作创建 AI 策略。Sentient 在资金方面获得 1.43 亿美元的支持，完全通过 AI 进行操作；在算法上，它使用了进化学习，生成了数万亿种专业程序
2017 年 6 月	华夏基金与微软公司在亚太地区设立的微软亚洲研究院举办战略合作发布会，宣布双方将就人工智能在金融服务领域的应用展开战略合作研究
2017 年 10 月	PINTEC（品钛）集团与香港富卫集团（FWD）合作在新加坡成立金融科技公司 PIVOT，这是中国智能投顾产业的首次出海．富卫集团为香港盈科拓展旗下的子公司，而盈科拓展的掌门人是李嘉诚之子李泽楷
2017 年 11 月	工商银行基于人工智能技术的智能投顾品牌“AI 投”已上线运行
2018 年 3 月	贝莱德正式推出了以“进化”统一命名的 7 只人工智能 ETF，这 7 只产品覆盖了消费、金融、医疗保健、科技、媒体等领域，是全球首个以人工智能推出的行业 ETF
2019 年 3 月	同花顺旗下人工智能资产管理有限公司同花顺阿尔法一号私募证券投资基金完成私募投资基金备案

资料来源：兴业证券经济与金融研究院整理

人工智能、机器学习对投资的影响应该说是全流程的，从新数据源的解析与挖掘，到模型构建的方方面面，再到具体的交易模块，越来越多的业界投研人员将两者紧密地结合在一起。具体到我们平日常用的线性多因子框架，其中的因子选择环节就是一个展现机器学习方法优势的非常好的例子。

我们知道多因子体系的基石是大量的选股因子。一般来说，选股因子可以分成以下几个大类，包括：价值、成长、质量、动量/反转、分析师情绪、另类（技术、规模、流动性、风险）等。构建每类因子所需的信息不同，因子选股的表现也是千差万别，而且因子之间（同类或不同类）也存在着或多或少的相关性。因此，如何从已有的因子库中有效地进行因子选择，就成为了每个模型都必须解决的问题。在传统的线性模型中，普遍的做法是根据每个因子的中长期表现，选择具有一定选股能力且符合经济逻辑的因子。这种方法的好处是简单易行，且能够与管理人的主观判断相符，但其缺陷也是显而易见的：筛选过程过于主观、不可复制，容易利用到未来数据，在因子数量较多时效率极低且缺乏一个系统性地应对市场变化的机制。为了解决上述问题，我们将目光转向机器学习的王国，尝试从特征选择和特征提取的角度寻找解决因子动态选择问题的方法。那么，到底什么是特征选择和特征提取呢？

2.2、特征选择与特征提取

在机器学习中，随着数据维度的上升，提供可靠分析所需的数据量将成倍增长。贝尔曼将这种现象称为“维度的诅咒[1]”。

当数据集的维数持续增加时，数据集中有意义的数据将会越来越稀疏，这将增加证明模型结果具有统计学意义的难度。而大数据集中所谓的“大p，小n“问题（其中p是特征数量，n是样本数量）往往使模型过度拟合，从而将小波动误认为

是重要的数据差异而导致分类错误。此外过多的特征也会使得数据集的噪声增加，数据集中的噪声定义为“测量方差的误差”，可能来自测量误差或数据本身的方差[2]。机器学习算法很容易受到嘈杂数据的影响，另外从计算量上来看，随着维度的增加计算成本也会以指数方式提升，因此应该尽可能的减少噪音以避免不必要的复杂性，从而提高算法的效率[3]。

要克服以上问题就必须要找到一种方法来减少备选的特征数量。解决高维数据集问题的一种流行方法是从原始数据集中挑选出有效的变量，删除无效的变量。或者在尽可能保留信息的条件下，寻找一种映射方法将高维数据投影到低维空间上，这两种技术分别称为：特征选择（feature selection）和特征提取（feature extraction）。

➢ 1、特征提取

特征提取通过组合原始变量创建新变量，从而减少所选特征的维数。特征提取算法有两大类：线性和非线性。

线性特征提取假设数据位于较低维度的线性子空间，从而可以直接通过矩阵分解将数据投影在子空间上以实现降维。常见的方法有 PCA-主成分分析[18]，ICA-独立成分分析[19]及 MDS[20]。非线性特征则通过不同方式进行降维，常见的方法有两大类：1、针对特征之间的非线性关系，可以使用提升函数将特征映射到更高维空间。在更高的空间上，特征之间的关系可以被视为线性的，从而我们能使用线性降维的方法，将高维数据映射回较低维度的空间以实现降维。常见的方法有 Kernel PCA[21]；2、另一种方法通称为流形学习 Manifold Learning[22-25]，其思想是若高维数据存在流形结构，则我们能通过非线性方法将高维数据映射到低维空间，同时尽量保有高维数据的本质。

常见的特征提取方法总结如下：

图表 3、特征提取的方法对比

特征提取类别	优点	缺点	例子
线性	有效特征识别能力高	无法处理非线性问题 数据可解释度低	PCA ICA
非线性	有效特征识别能力高 能处理非线性问题	数据可解释度低	Kernel PCA Manifold Learning

资料来源：兴业证券经济与金融研究院整理

➢ 2、特征选择

与特征提取方法不同，特征选择技术不会改变原始数据的表现形式[4]，该技术是通过删除不相关或多余的特征从而达到降维，实现方法分为三类[5-6]：

1) 过滤式特征选择，直接从数据中提取特征（与后续学习过程无关）；
2) 包裹式特征选择，把最终将要使用的模型性能作为特征子集的评价准则；
3) 嵌入式特征选择，将特征选择过程与学习器训练过程融为一体，两者在同一个过程中完成。

✧2.1 过滤式（Filter）

过滤式方法使用模型效能以外的度量来确定该特征是否有用。该方法通过描

述性指标对特征进行排序筛选，而不是以所使用的模型（如包装方法中的模型）的拟合效果来选择特征子集。过滤方法的优点是计算时间非常短、不存在过拟合问题。然而，该方法忽略了特征之间的任何交互或关联。常见的三种不同过滤方法是方差分析[7]、皮尔逊相关分析[8]和信息增益分析[9]。

1) **方差分析**(ANOVA, Analysis of variance) 的思路为：按照不同的特征类别将特征划分为不同的总体，接着检验不同总体之间均值是否相同。如果相同，那么这个特征就不能很好地解释因变量的变化。方差分析检验方法如下，计算每个特征的 F 统计量，接着按每个特征 F 值的大小进行排序，去除 F 值小的特征；
2) **皮尔逊相关分析**使用-1 到 1 之间的数字来度量两个特征之间的相似性。接近 1 或-1 的值表示这两个特征具有很高的相关性。要使用该方法进行特征选择，可以查看所有特征两两相关性的热点图（heatmap），在相关性高的“特征对”中保留与因变量（预测变量）具有最高相关性的特征。高相关与低相关的临界值取决于每个数据集中相关系数的范围。高相关性的范围一般是 0.7 以上；
3) **信息增益分析基本方法**如下：对于一个特征，计算模型有它和没它的时候信息量各是多少，两者的差值就是这个特征给模型带来的信息量（即增益）。通过对信息增益排名即可挑选出效果较好的变量。

✧2.2 包裹式（Wrapper）

包裹式特征选择使用特定的特征子集计算模型效能，该方法通过不断迭代不同的特征子集直到找到最佳解。其缺点是计算时间长，在样本量不够大的情况下模型容易产生过拟合。常见的包裹式特征选择方法有向后选择[10]、向前选择[10]、模拟退火算法[11]及遗传算法[12-13]。

1) **向后选择**从使用所有的数据集开始，该方法需要为每个特征与模型计算 t 检验或 f 检验的 p 值。然后，从模型中删除最不重要的特征（依据 p 值）。重复上述过程，直到模型中不重要特征被删除完毕为止；
2) **向前选择**从零个特征开始，对于每个单独的特征，该方法同样需要计算 t 检验或 f 检验的 p 值，然后选择 p 值最低的特征并将其添加到模型中。接下来，在保有第一个特征的条件下运行添加第二个特征的模型，并选择 p 值最低的第二个特征。以此类推，直到所有具有显著 p 值的特征都被添加到模型中；
3) **遗传算法**首先随机产生一批特征子集，并用评价函数给这些特征子集评分，然后通过交叉、突变等操作繁殖出下一代的特征子集，其中评分越高的特征子集被选中参加繁殖的概率越高。这样经过 N 代的繁殖和优胜劣汰后，种群中就可能产生评价函数值最高的特征子集；
4) **模拟退火算法**目标是要解决局部最优解困境。在随机产生特征子集后，该算法以一定的概率接受一个比当前模型效能要差的模型效能（这个概率随

着时间的推移逐渐降低),这样做可以提高模型跳出局部最优解的可能性,从而达到全局最优解。

✧2.3 嵌入式（Embedded）

嵌入式方法将特征选择作为模型创建过程的一部分。该方法通常为前面两种选择选择方法的折衷。其中 Lasso、岭回归[14]及决策树[15-16]是较为常见的嵌入式特征选择方法。

当希望能在最终模型中保留所有特征，但又不希望模型过于关注任何一个系数时，**岭回归**可以通过对模型的系数（权重）施加惩罚来做到这一点。具体操作是通过在回归的成本函数中添加一个惩罚项(L2 正则项)来对系数过大进行惩罚。所有变量的权重共用一个参数 lambda（λ）来对其进行惩罚，lambda 是一个介于 0 和无穷大之间的值。lambda 越高，系数收缩的越多（惩罚越大）。当 lambda 等于 0 时,结果将是一个不带惩罚项的普通最小二乘模型;与岭回归非常相似,**Lasso** 回归是另一种惩罚模型系数（即变量权重）的方法(惩罚项为 L1 正则项)。与岭回归的区别是，Lasso 趋向于使一部分变量的权重变为 0，这将使得模型的特征数量减少，从而降低复杂性，这就是为什么 Lasso 在某些时候更受欢迎；第三种常用的嵌入式特征选择方法是**决策树**，它可以是回归树，也可以是分类树，具体取决于因变量是连续的还是离散的。在建立树模型时,函数内置了几种特征选择方法,在每次拆分时，用于创建树的函数会尝试对所有特征进行所有可能的组合。简单地说，它选择最能预测树中每个节点的最优特征。而在预测因变量时，最重要的特征在树的根（开始）附近进行拆分，而较不相关的特征是在树的叶（结束）附近进行拆分。生成树之后，可以选择“修剪”一些不向模型提供任何附加信息的节点。这可以防止过拟合，通常通过测试集的交叉验证来实现。

常见的特征选择方法总结如下：

图表 4、特征选择的方法对比

特征选择类别	优点	缺点	例子
过滤式	**单变量**		
	运算快 可扩展性高	忽略特征间相关性 忽略模型效果	信息增益
	多变量		
	考虑特征间相关性	运算速度较慢 忽略模型效果	方差分析 皮尔逊相关
包裹式	**确定性**		
	运算较快（相比随机性） 考虑模型效果	过拟合风险 容易掉入局部最优	向前选择 向后选择
	随机性		
	考虑模型效果 不易掉入局部最优	极高的过拟合风险	模拟退火算法 遗传算法
嵌入式	考虑模型效果	未考虑分类器间相关性	Lasso 回归 决策树

资料来源：兴业证券经济与金融研究院整理

➢ 3、特征选择 Vs 特征提取比较

特征提取和特征选择都属于数据降维的方法。两者主要的不同在于特征提取是在原有特征基础之上去创造出一些新的特征出来，而特征选择则只是在原有特征上进行筛选。因此在数据的解释层面上特征选择能较好的保有原始数据的特征。

图表 5、特征选择 Vs 特征提取方法对比

方法	优点	缺点
特征选择	能保有原始数据的特征 解释度较强	较低的有效特征判别力
特征提取	较高的有效特征判别力 较小的过拟合风险（无监督学习）	较低的数据解释度 提取过程好使较长

资料来源：兴业证券经济与金融研究院整理

通过对特征选择和特征提取的整体分析，我们认为特征选择更加适合选股领域的分析，解释性和接受程度也更高。在特征选择里面，我们首先通过 Filter 限制变量的个数，进一步通过 Wrapper 确定最终的选股变量。

3、基于 Filter&Wrapper 的动态线性选股模型构建

3.1、数据准备&整体流程

为了避免行业市值的影响，我们对于每个因子都会做行业市值中性化，处理方法如下：以中信一级行业为标准，以每个行业内所有股票的流通市值中位数为界来进行大小票的划分，中位数以上者为该行业的大盘股，以下者视为该行业小盘股。而后分别在每个行业市值股票池内进行横截面因子的分位数变换标准化。除了因子层面的标准化之外，对于收益率我们同样通过分位数变换标准化的方式进行处理，以保证可比性。所以后续回归中的收益率均是标准化后的结果。

我们在因子体系构建的时候先用 Filter 的方法通过考虑因子表现和相关性，将因子数目控制在一定范围内；然后用 Wrapper 方法（我们这里采用逐步倒向线性回归的方式筛选，后面会有详细介绍）进一步优化因子数目。

在运用 Filter 以及 Wrapper 选择每期（比如月度选择）的有效因子后，我们运用线性回归的方式来构建最终的线性多因子模型，而这也使得我们保持了从特征选择到预测模型所使用的方法的统一性（Wrapper 使用的就是线性回归）。我们将通过上述一整套流程得到的线性复合因子称作动态线性因子（Dynamic Linear Signal，DLS）。

3.2、动态因子选择

兴业证券金融工程团队所构建的因子库共计包含 165 个量化选股指标，进一步细分为价值、成长、质量、分析师情绪、动量反转、另类这六大类指标，其中另类进一步分为：规模、风险、流动性、技术这四个子类（部分因子定义参见附录）。

图表 6、兴业量化选股因子分类数量统计

	价值	成长	质量	分析师情绪	动量反转	另类
因子数	19	33	39	13	9	52

资料来源：Wind，兴业证券经济与金融研究院整理

以上六大类 165 个因子是我们初始因子池。我们的目标是在每个月月底通过 Filter&Wrapper 方法选择有效的因子，接下来我们详细介绍两者的实现方法。Filter 和 Wrapper 中方法众多，我们这里以 Filter 中的皮尔逊相关系数筛选法以及 Wrapper 中逐步倒向线性回归法为例阐述因子筛选的流程。

- **Filter 筛选**

1、选择每类中表现优秀的因子：

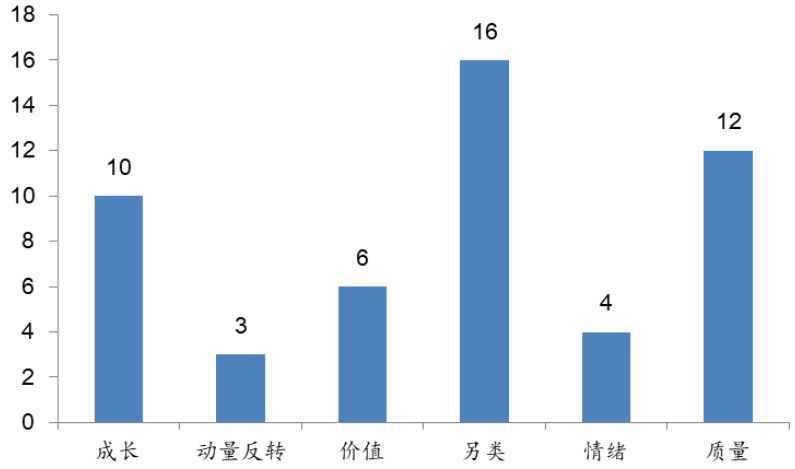
滚动计算过去 5 年每个因子的 ICIR 表现，从六大类别中分别选择表现最优秀的 30%的因子；

2、剔除相关性高的因子：

在筛选时刻计算过去 5 年因子 IC 时间序列的两两相关性，删除相关性高于 0.75 的一对因子中的某一个。在删除的同时需要满足两个准则：i)、一旦确定高相关性因子对之后，要删除相应时点两个因子中过去 5 年 ICIR 表现差的因子；ii)、在删除的同时需保证每大类里面至少保留一个因子（以避免风格有偏）；

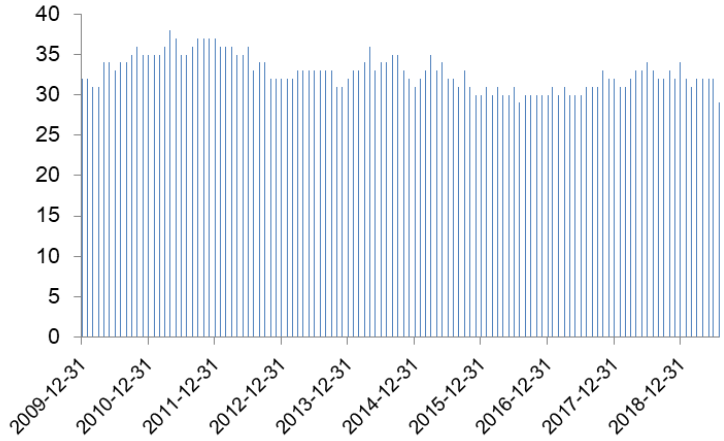
从结果来看，经过第一步筛选之后我们的因子总量从 165 降至 51 个，具体参见图表-8（这里需要注意，由于表现好坏的筛选是一个固定的阈值，所以每期的结果一致）。而经过相关性筛选后因子总数稳定在 32 左右（具体参见图表-9）；而每类选择的数目变动趋势参见图表-10。

图表 8、经过 Filter 中表现筛选后保留的因子数目对比



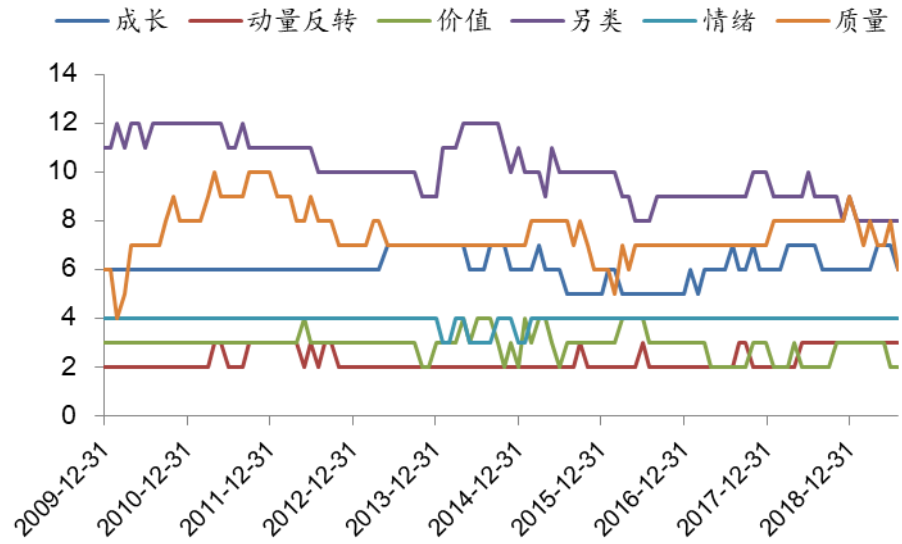
资料来源：Wind，兴业证券经济与金融研究院整理

图表 9、经过 Filter 中相关性筛选后保留的因子数目变动趋势



资料来源：Wind，兴业证券经济与金融研究院整理

图表 10、经过 Filter 步骤之高相关性筛选后每类因子数目变动趋势



资料来源：Wind，兴业证券经济与金融研究院整理

- **Wrapper 筛选**

在完成 Filter 步骤后，我们进一步通过 Wrapper 筛选剩余的因子。Wrapper 有多种实现方式，这里我们采用的是逐步倒向线性回归的方式（与最后利用线性回归构建预测模型保持一致）来挑选因子。逐步倒向回归需要设定因子表现好坏的标准，我们这里以回归系数的 P 值作为参考，阈值设定为 0.05，具体流程和注意事项如下：

1、为保证算法的稳定性，我们用过去 3 年 36 个月月底的横截面数据来构建训练样本。自变量为当期选中的因子过去三年的月度数据，Y 为过去 3 年标准化的月度收益；

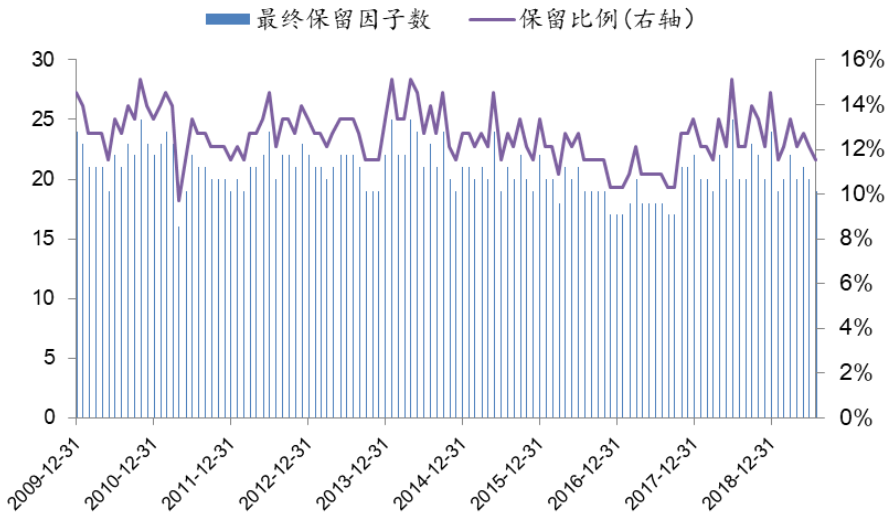
2、 期初回归时，将截面选中因子全部放进去进行回归，如果所有因子 P 值都能达到要求，则程序终止，否则将表现最差因子剔除，同时以剩余因子为自变量再回归，重复该步骤，直至所有因子 P 值均满足要求；

3、 这里需要注意的是，在剔除因子过程中，P 值并不是唯一的参考标准，我们需要保证每大类中至少有一个因子存在。这也就意味着，如果在某次倒向回归中，某个类仅仅剩余一个因子，同时 P 值不达预期，此时即便该因子是表现最差的也不能删除，而这次回归只能追求次优解。这一限制的目的与前面 Filter 步骤中删除高相关性因子时所需要注意的地方是一致的，目的都是为了保证所选因子在大类风格上分布的均衡性。

在完成 Wrapper 筛选后，我们每期保留的因子数大幅度下降，从 Filter 筛选后平均 32 个左右降低至 18 个的水准。观察每类因子的数目变动趋势，我们发现成长、动量反转变动最小，数目也最少，而质量和另类因子的数量变动幅度最大。

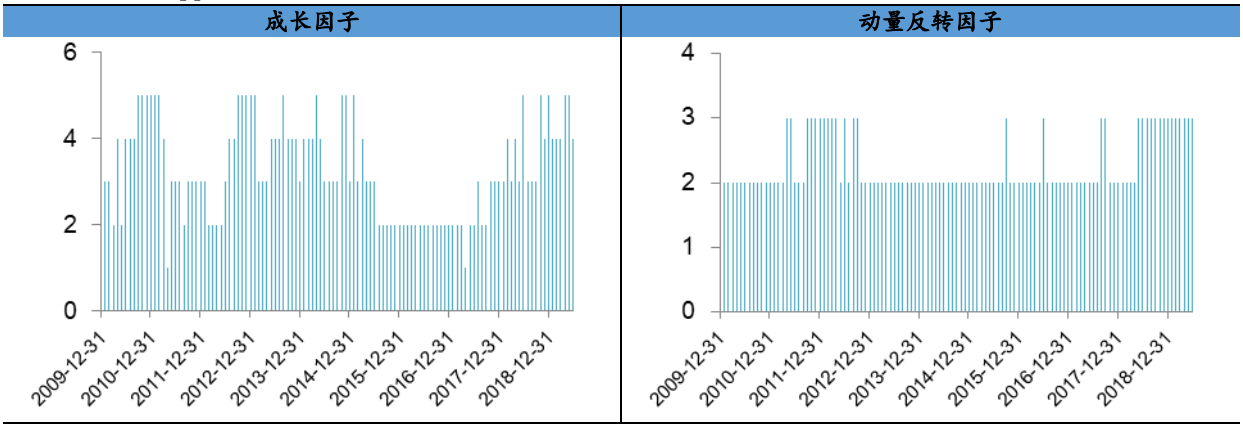
最终我们带着经过 Filter&Wrapper 筛选后的因子进入到下一个环节，用线性回归的方式构建多因子收益率预测模型。完整的 Filter&Wrapper 筛选流程参见图表-15。

图表 11、经过 Filter&Wrapper 筛选后剩余的因子数目及变动趋势



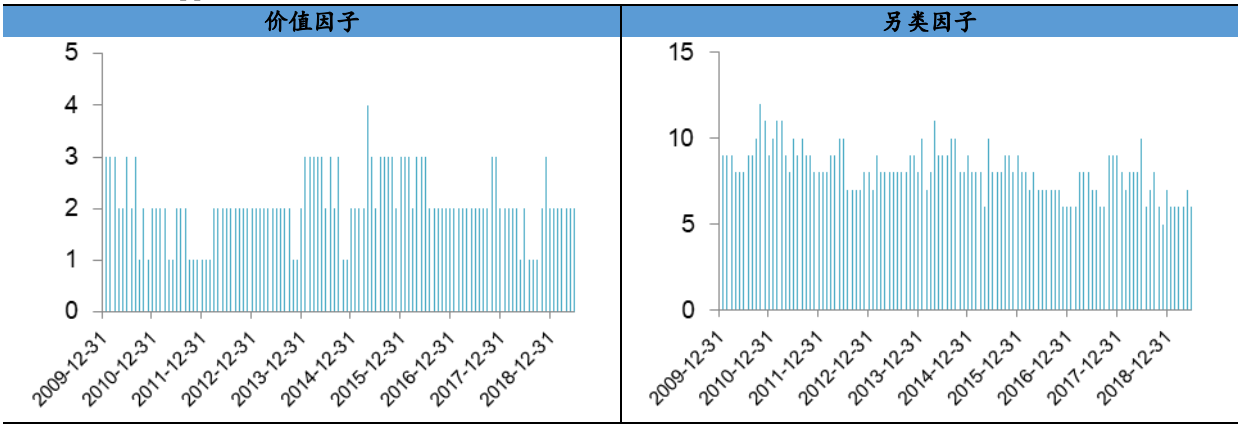
资料来源：Wind，兴业证券经济与金融研究院整理

图表 12、Wrapper 筛选后成长及动量反转因子数变动



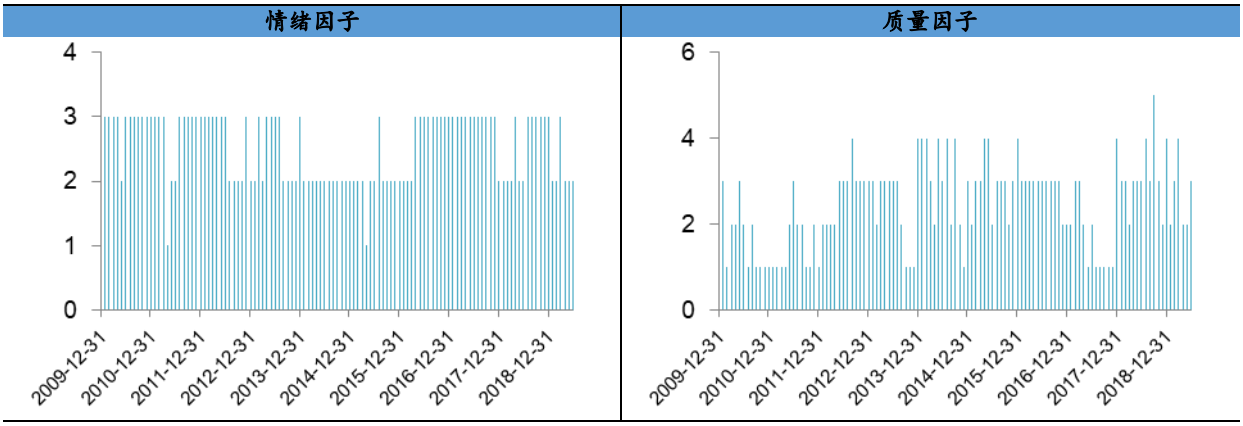
资料来源：Wind，兴业证券经济与金融研究院整理

图表 13、Wrapper 筛选后价值及另类因子数变动



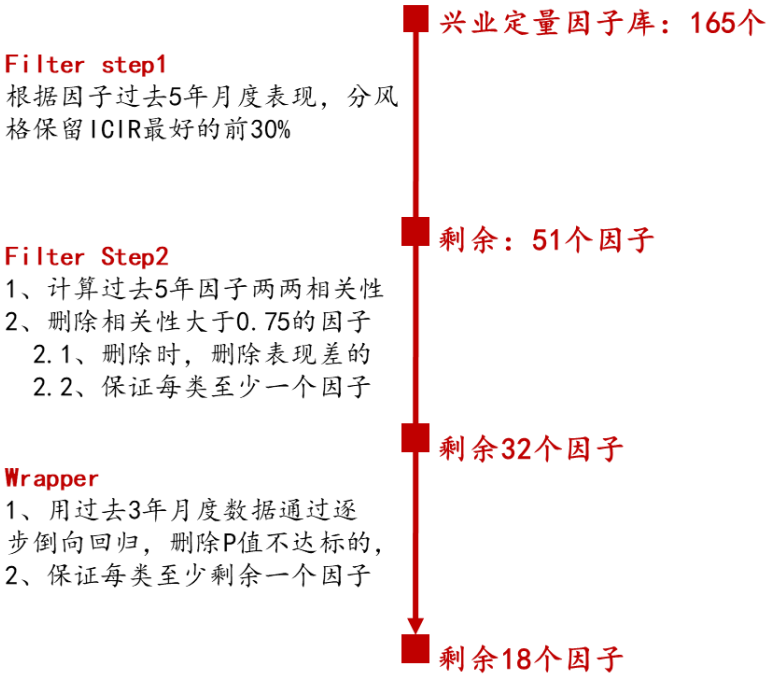
资料来源：Wind，兴业证券经济与金融研究院整理

图表 14、Wrapper 筛选后情绪及质量因子数变动



资料来源：Wind，兴业证券经济与金融研究院整理

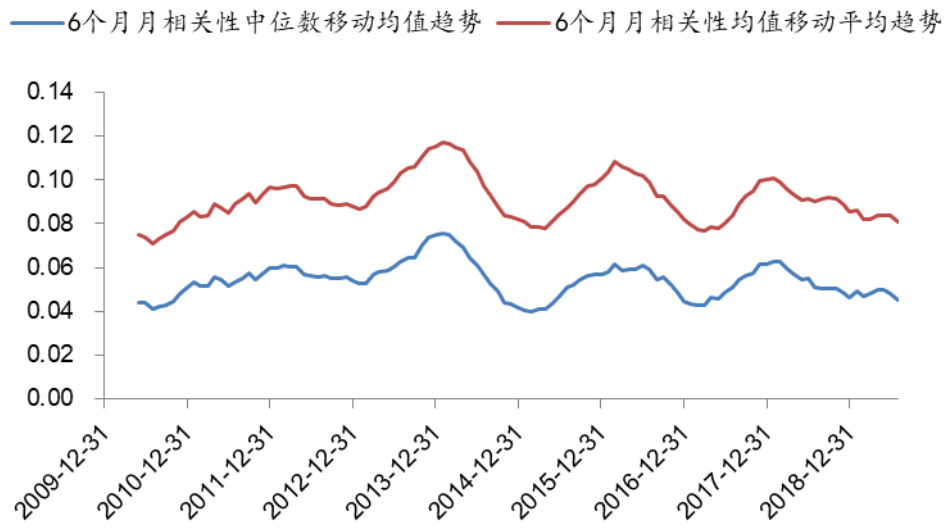
图表 15、因子动态筛选流程图



资料来源：Wind，兴业证券经济与金融研究院整理

我们研究 Filter&Wrapper 流程截面筛选的因子相关性，从结果来看，每期因子的相关性非常低，均值稳定在 0.045 左右。因子相关性较低一方面提升最终合成因子的稳定性同时规避了多重共线性等问题，为后续的研究打下基础。

图表 16、Filter&Wrapper 流程筛选的因子相关性



资料来源：Wind，兴业证券经济与金融研究院整理

3.3、DLS 因子表现分析

经过动态因子选择环节，我们确定了每期选中的因子，进一步通过线性回归的

方式将因子合成。同时为了保证算法的稳定性，与前面一致，我们用选中的因子过去 3 年 36 个月月底的横截面数据来构建回归样本，并通过最小二乘回归确定回归系数，进而得到最终的复合因子 DLS(Dynaimc Linear Signal)

1） DLS 全市场表现分析

IC 以及分位数测试结果显示该因子的选股能力非常强：IC 均值达到 10.6%，ICIR 达到 1.50；而多空年化收益率达到 43.0%、夏普率高达 4.99。

图表 17、DLS 因子 IC 测试

	平均值	标准差	最小值	最大值	IC_IR	t 统计量
DLS	0.106	0.07	-0.08	0.29	1.50	16.06

资料来源：Wind，兴业证券经济与金融研究院整理

图表 18、DLS 因子分位数组合测试

	年化收益	Sharpe	换手	最大回撤	年超额收益	跟踪误差	信息比	胜率
top	21.8%	0.72	1.16	34.4%	14.5%	4.3%	3.40	87.1%
1	16.3%	0.54	1.57	36.2%	9.4%	3.7%	2.50	78.4%
2	15.1%	0.50	1.66	36.6%	8.3%	2.5%	3.32	81.0%
3	12.8%	0.43	1.70	40.0%	6.2%	2.7%	2.28	72.4%
4	9.0%	0.30	1.71	47.0%	2.6%	2.3%	1.16	63.8%
5	5.7%	0.19	1.72	51.6%	-0.3%	2.2%	-0.14	49.1%
6	4.2%	0.14	1.69	56.5%	-1.7%	2.2%	-0.80	35.3%
7	0.3%	0.01	1.66	64.5%	-5.3%	3.1%	-1.69	25.0%
8	-3.6%	-0.11	1.56	69.5%	-9.0%	3.2%	-2.78	17.2%
bottom	-16.2%	-0.51	1.14	83.1%	-20.7%	5.2%	-3.99	8.6%
市场	6.1%	0.20		53.5%				
L_S	43.0%	4.99		4.6%				

资料来源：Wind，兴业证券经济与金融研究院整理

2) DLS 与基准模型的对比分析

前文详细描述了 DLS 因子的构建方式，这里我们尝试通过较为简单的方法为 DLS 因子构建一个基准模型，并对两者进行对比分析。基准模型的构建方法如下：在某个时刻，计算过去 5 年所有 165 个因子表现，从每个类里面选择最为有效(ICIR)的 3 个因子(这一步亦是 Filter 的第一步；而每类选择 3 个也保证了最终因子总量和 DLS 所选择的因子数量基本一致)等权合成，称之为 Benchmark Model (BM)。

从同时期的 IC 以及分位数组合测试结果来看:DLS 因子的 IC 均值达到 0.106，ICIR 以及 T 值分别为 1.50、16.05，远高于 BM 相应指标的水准。从 IC 的移动平均趋势也可以看出 DLS 的有效性要远高于 BM 因子；同样，DLS 的多空年化收益率以及夏普率分别为 43.0%，4.99，也优于基准因子的表现。

为了进一步考察 DLS 的表现，我们还构建了一个在已知全样本因子表现下的复合因子 Future_Sig：回测 2009 年 12 月 31 日-2019 年 7 月底所有因子的月度 ICIR，然后每类中选择月均 ICIR 最高的 3 个因子并等权合成。该因子无论是 IC 亦或是多空表现都非常优异。但需要注意的是，Future_Sig 是站在 2019 年 7 月 31 日，观察并选择在整个历史样本区间上表现最优的因子，而这实际上这是严重的窥探未来数据的做法。当然通过参考 Future_Sig 因子构建方法以及对比 Future_Sig 与 DLS 因子表现我们可得知：1、DLS 因子构建方法没有利用任何未来信息（均是站在当前时点回望过去一段时间的表现）；2、DLS 因子与 Future_Sig

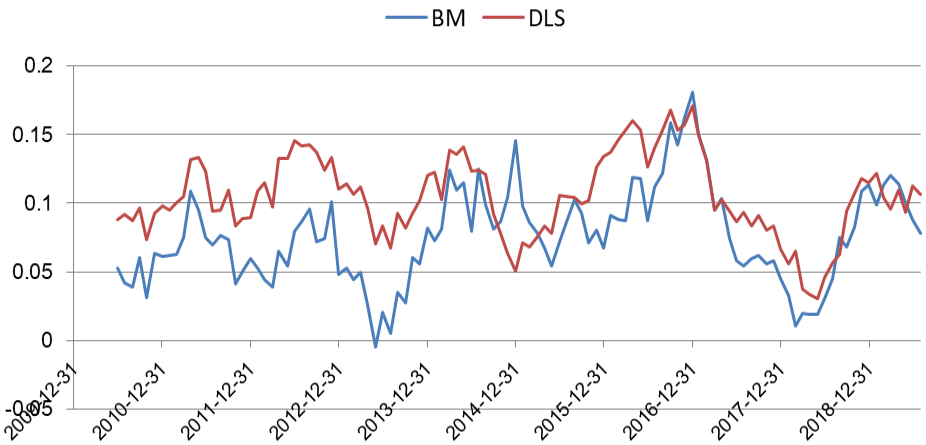
的表现非常接近，也在一定程度上说明了上述动态因子选择方法的有效性。

图表 19、BM、DLS、Future_Sig 因子对比分析

	平均值	标准差	IC_IR	t 统计量	多空年化收益	多空夏普率	最大回撤
BM	0.061	0.053	1.16	12.48	24.3%	3.48	4.9%
DLS	0.106	0.070	1.50	16.05	43.0%	4.99	4.6%
Future_Sig	0.118	0.064	1.86	20.03	47.2%	5.16	2.5%

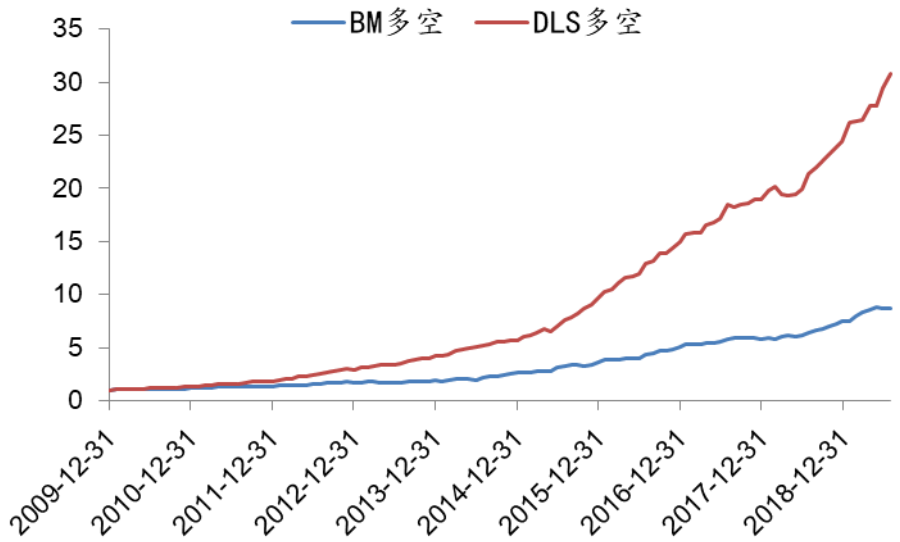
资料来源：Wind，兴业证券经济与金融研究院整理

图表 20、6 个月 DLS、BM 因子的 IC 移动平均趋势



资料来源：Wind，兴业证券经济与金融研究院整理

图表 21、DLS、BM 因子的多空组合净值曲线



资料来源：Wind，兴业证券经济与金融研究院整理

3） DLS 在宽基指数成分股内的表现分析

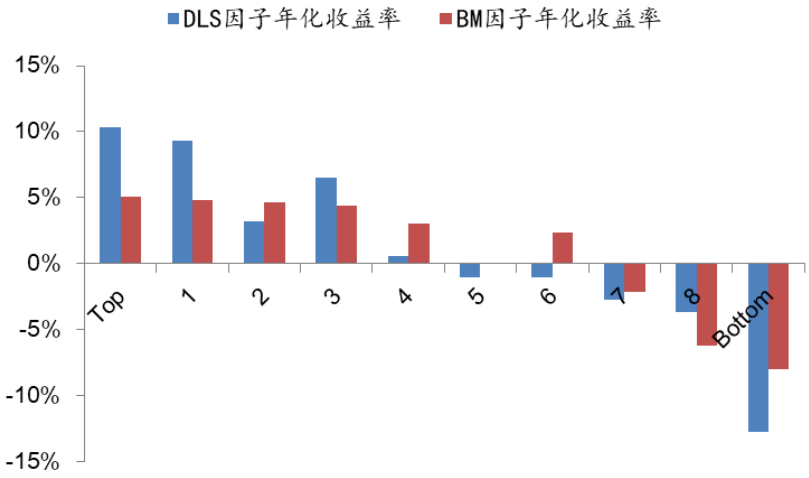
我们将股票池缩小至主要宽基指数范围内（沪深 300/中证 500），从测试结果来看，DLS 因子的有效性和稳定性规律依然不变，以在中证 500 测试为例：DLS 因子 IC、ICIR、T 值分别达到 0.88、0.97、10.45，多空组合年化收益率 31.3%，夏普率 3.02，大幅度优于基准的表现，且各组别的单调性要优于基准因子的表现。

图表 22、DLS、BM 因子在主流宽基指数范围内的有效性测试对比分析

		平均值	标准差	IC_IR	t 统计量	多空年化收益	多空夏普率
沪深 300	BM	0.049	0.09	0.51	5.52	13.1%	1.10
	DLS	0.071	0.11	0.63	6.76	24.6%	1.86
中证 500	BM	0.067	0.10	0.68	7.30	19.5%	2.21
	DLS	0.088	0.09	0.97	10.45	31.3%	3.02

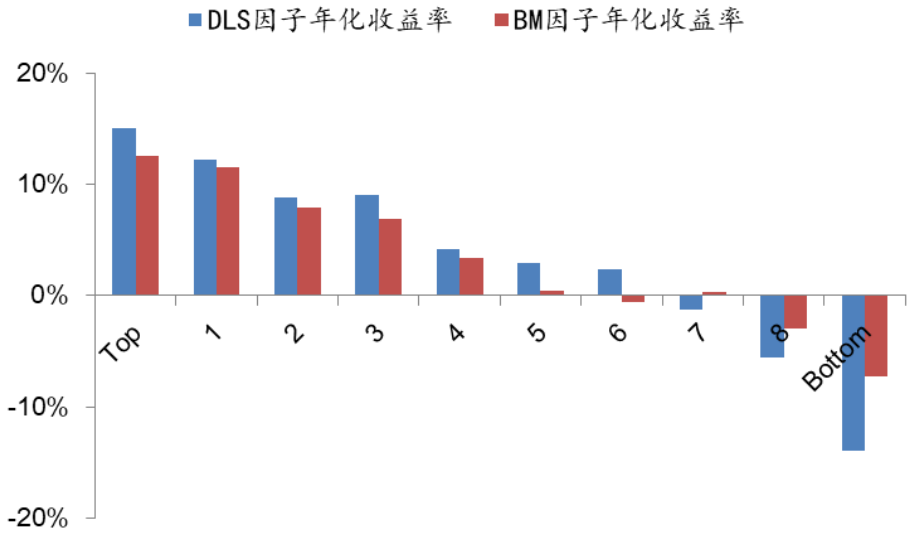
资料来源：Wind，兴业证券经济与金融研究院整理

图表 23、DLS、BM 因子在沪深 300 指数内分位数测试之年化收益



资料来源：Wind，兴业证券经济与金融研究院整理

图表 24、DLS、BM 因子在中证 500 指数内分位数测试之年化收益



资料来源：Wind，兴业证券经济与金融研究院整理

3.4、A-DLS 因子表现分析

整个 DLS 因子的构建逻辑（包括 Filter&Wrapper 的因子筛选流程）都是假设因子和收益之间的关系是线性的，因此那些具有非线性选股能力的因子将会被全部剔除。为了能够更好地将非线性预测能力与上述模型融为一体，我们引入了在《基于集成学习算法的量化选股模型研究》中所构建的集成学习因子（基于改进

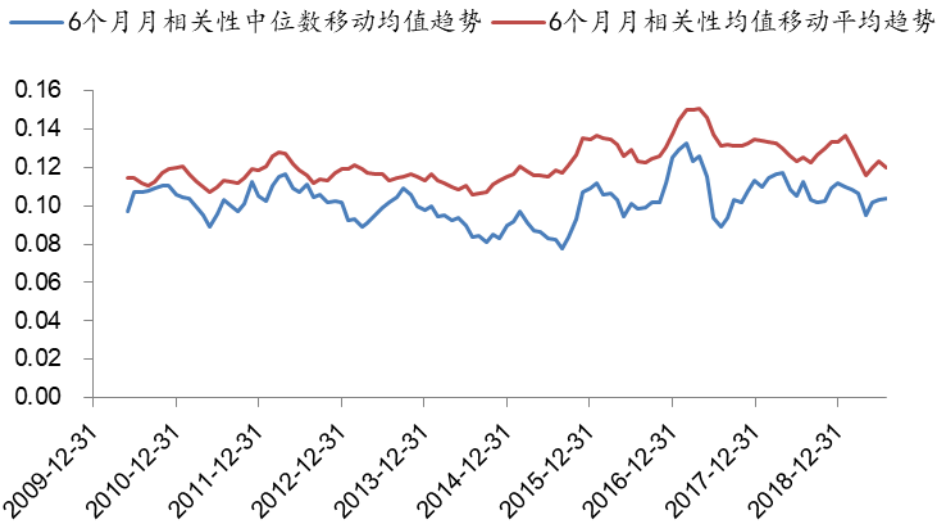
版的 Adaboost 算法)。如何将 DLS 与集成学习因子结合起来呢？这里大致有两种思路：

1. 将 DLS 与集成学习因子分别当成两个因子，通过等权、经验比例、相关性调整 IC（附录有详细介绍）加权等方式将两个因子合成；
2. 将集成学习因子作为一个体现非线性选股能力的单因子，与 Filter&Wrapper 选择的因子一起进行线性回归，并得到最终的复合因子。

我们对上述两种方式都进行了实证分析，这里重点呈现第二种方式（第一种方法的结果请参见附录 1），并把利用该方法生成的因子称为 A-DLS (Adaboost &Dynamic Linear Signal)。注意在构建 A-DLS 因子时，所有方法和细节处理均保持不变，只是在每期回归时，多增加了一个集成学习因子。

同时我们证实了集成学习因子与前面 Filter&Wrapper 筛选的因子相关性非常低，均值稳定在 0.1 左右。低相关性进一步支持将集成学习因子纳入我们的回归模型当中。

图表 25、集成学习因子与 Filter&Wrapper 流程筛选的因子相关性



资料来源：Wind，兴业证券经济与金融研究院整理

从最终合成因子的测试结果来看，A-DLS 因子的表现十分优秀：从 IC 来看，A-DLS 因子 IC 均值、ICIR 以及 T 值分别为 0.141、2.01 和 21.58；从分位数组合测试来看，A-DLS 因子多空组合年化收益率高达 60.5%、夏普率达到 6.55，且 A-DLS 因子各分为组合的单调性、换手率均优于集成学习因子的表现。整体来看 A-DLS 表现略胜一筹。

图表 26、A-DLS、集成学习因子 IC 测试

	平均值	标准差	最小值	最大值	IC_IR	t 统计量
A-DLS	0.141	0.070	-0.06	0.35	2.01	21.58
集成学习因子	0.137	0.066	-0.03	0.32	2.08	22.32

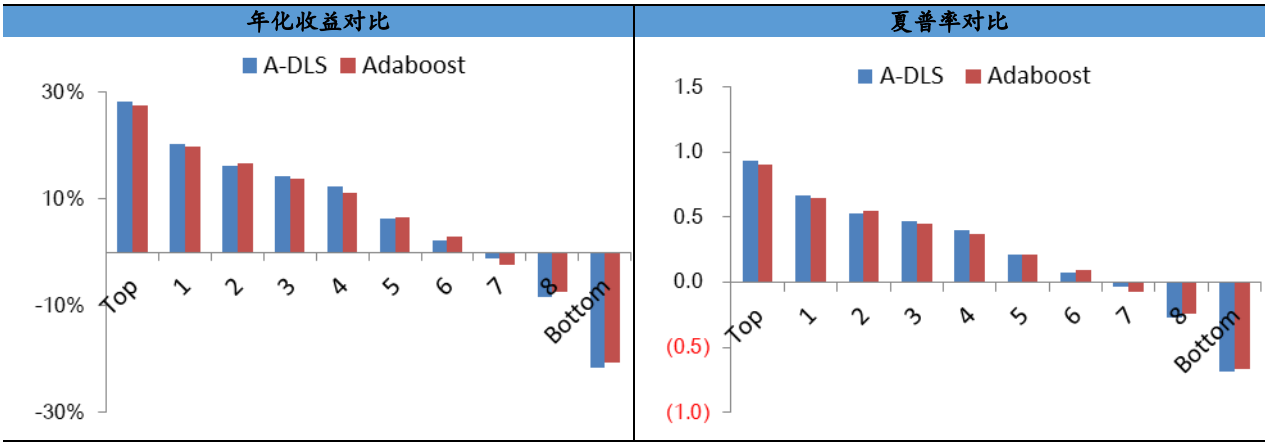
资料来源：Wind，兴业证券经济与金融研究院整理

图表 27、A-DLS、集成学习因子分位数组合测试

	A-DLS 因子表现				集成学习因子表现			
	年化收益	Sharpe	换手率	最大回撤	年化收益	Sharpe	换手率	最大回撤
top	28.3%	0.94	1.28	30.0%	27.6%	0.90	1.42	29.5%
1	20.3%	0.67	1.63	33.1%	19.9%	0.65	1.69	34.6%
2	16.2%	0.53	1.70	36.5%	16.8%	0.55	1.74	36.3%
3	14.2%	0.47	1.73	35.2%	13.8%	0.45	1.77	35.4%
4	12.5%	0.41	1.75	39.8%	11.2%	0.37	1.78	44.3%
5	6.3%	0.21	1.75	51.2%	6.5%	0.21	1.78	51.8%
6	2.3%	0.07	1.74	61.2%	3.1%	0.10	1.77	58.8%
7	-1.0%	-0.03	1.72	63.2%	-2.3%	-0.07	1.76	65.9%
8	-8.4%	-0.27	1.64	76.0%	-7.4%	-0.24	1.70	73.3%
bottom	-21.6%	-0.68	1.29	90.5%	-20.6%	-0.67	1.49	89.8%
市场	6.1%	0.20		53.5%	6.1%	0.20		53.5%
L-S	60.5%	6.55		1.8%	58.0%	6.10		0.6%

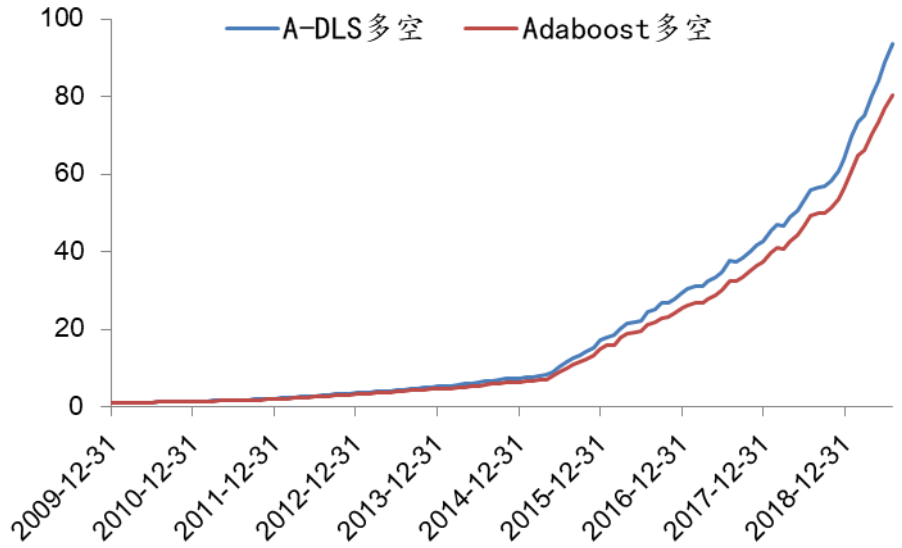
资料来源：Wind，兴业证券经济与金融研究院整理

图表 28、A-DLS VS 集成学习因子分位组合年化收益和夏普率对比



资料来源：Wind，兴业证券经济与金融研究院整理

图表 29、A-DLS、集成学习因子多空表现



资料来源：Wind，兴业证券经济与金融研究院整理

4、基于 A-DLS 因子的选股策略研究

一般而言，投资者会结合自己的需求（如投资风险偏好不同、资金容量要求不同、换手频率不同等）构建不同风格的投资策略。接下来我们基于 A-DLS 因子构建主动量化以及增强选股策略。

4.1、主动量化策略构建

基于 A-DLS 因子我们构建了主动量化选股模型：

1. 每期选择 100 只股票，以中证 500 作为业绩比较基准；
2. 调仓日：若当期持仓的股票下一期没有跌出前 200，仍然继续持有；
3. 其他设定：选股池需删除 ST、同时保证上市天数满 180 天；交易成本双边 0.3%。

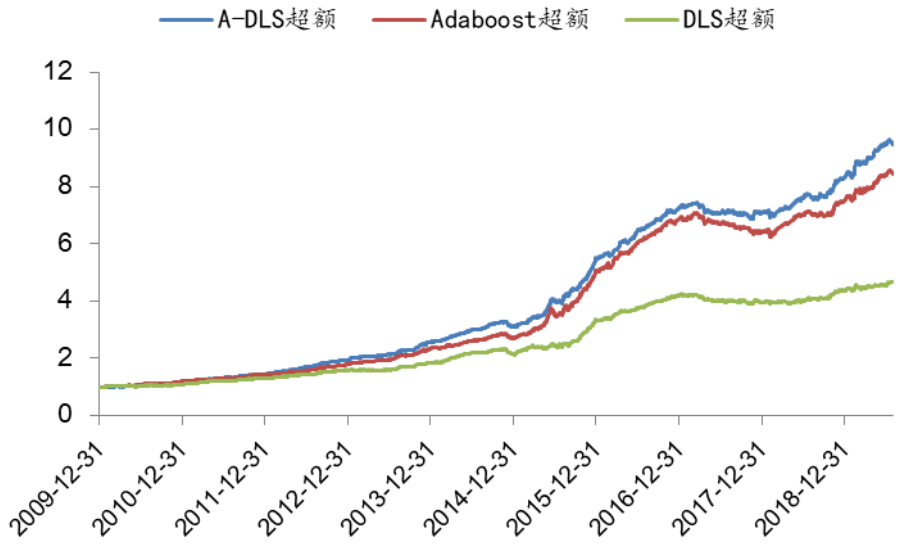
从测试结果来看，我们基于 A-DLS 因子构建的主动量化策略成绩斐然，策略多头超额收益风险比为 4.61，最大回撤为 6.6%。

图表 30、月度主动量化策略表现汇总

	年化收益率	波动率	风险收益比	胜率	回撤
中证 500 表现	0.9%	26.7%	0.03	54.1%	65.2%
A-DLS 策略多头	27.4%	27.7%	0.99	57.4%	45.1%
A-DLS 策略超额	26.9%	5.8%	4.61	62.3%	6.6%

资料来源：Wind，兴业证券经济与金融研究院整理

图表 31、月度主动量化策略超额收益净值曲线



资料来源：Wind，兴业证券经济与金融研究院整理

4.2、指数增强选股策略构建

指数增强策略旨在控制跟踪误差的前提下，尽可能获取超越基准的表现。综合考虑投资者的偏好，针对于沪深 300 以及中证 500 宽基指数，我们构建了 2 个增强模型，不同模型的实现细节参见图表 33。

图表 32、基于月度 A-DLS 因子构建的增强策略

	针对于沪深 300 的增强策略	针对于中证 500 的增强策略
策略一	1、换手率约束单边 0.2 2、全市场选股	1、换手率约束单边 0.4 2、全市场选股
策略二	1、换手率约束单边 0.2 2、每期沪深 300 股票池内股票权重占比>80%	1、换手率约束单边 0.4 2、每期中证 500 股票池内股票权重占比>80%
统一要求	1、 选股池需删除 ST 股、同时保证上市天数满 180 天； 2、 保持行业、市值、Beta 中性； 3、 个股最大权重不超过 10%； 4、 相对基准的权重波动不超过+-0.5%； 5、 交易成本：0.3%	

资料来源：兴业证券经济与金融研究院整理

1. 基于 A-DLS 因子的沪深 300 增强策略

基于沪深 300 的增强策略表现非常优异，以策略二为例：策略相对沪深 300 指数的年化超额收益稳定在 7.5%左右，最大回撤为 2.5%，风险收益比高达 3.14。分年度来看，选股策略在每一年均能稳定的跑赢基准，2019 年以来（截至 7 月 31 日）超额收益稳定在 4.5%。

同时我们发现策略一和策略二表现非常接近，这实际上意味着：沪深 300 具有非常鲜明的市值、行业、Beta 特点，只要这三类风格做到相对中性，那么即便不添加沪深 300 成分股权重占比 80%以上的要求，策略依然可以做到稳定的跟踪沪深 300 的表现。

图表 33、针对于沪深 300 的增强策略表现汇总

	年化收益率	波动率	风险收益比	胜率	回撤
沪深 300 表现	0.7%	23.0%	3.2%	51.1%	46.7%
策略一多头	8.9%	22.8%	0.39	52.5%	39.7%
策略二多头	8.3%	22.8%	0.37	52.8%	40.1%
策略一超额	8.0%	2.9%	2.80	57.3%	4.1%
策略二超额	7.5%	2.4%	3.14	58.6%	2.5%

资料来源：Wind，兴业证券经济与金融研究院整理

图表 34、针对于沪深 300 构建的增强策略分年度表现

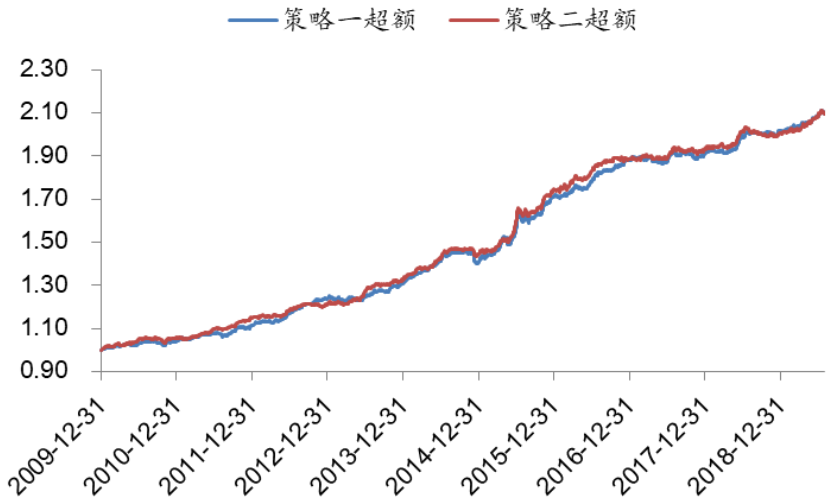
		多头		多头超额	
	基准	策略一	策略二	策略一	策略二
2010 年	-12.5%	-8.6%	-6.6%	4.2%	6.1%
2011 年	-25.0%	-19.9%	-18.8%	6.8%	8.2%
2012 年	7.6%	19.9%	13.5%	11.4%	5.5%
2013 年	-7.6%	-2.2%	1.2%	5.6%	9.2%
2014 年	51.7%	63.6%	65.7%	7.3%	9.0%
2015 年	5.6%	28.9%	27.8%	22.1%	20.9%
2016 年	-11.3%	-2.9%	-4.3%	9.7%	8.1%
2017 年	21.8%	23.8%	24.7%	1.7%	2.5%
2018 年	-25.3%	-21.0%	-22.4%	5.6%	3.8%
2019 年	27.4%	32.2%	33.1%	3.9%	4.5%

资料来源：Wind，兴业证券经济与金融研究院整理

图表 35、针对于沪深 300 构建的增强策略多头净值曲线

资料来源：Wind，兴业证券经济与金融研究院整理

图表 36、针对于沪深 300 构建的增强策略超额净值曲线



资料来源：Wind，兴业证券经济与金融研究院整理

2. 基于 A-DLS 因子的中证 500 增强策略

与沪深 300 的增强策略框架一致，我们构建了中证 500 指数增强策略。从结果来看策略表现突出，以策略一为例：策略年化超额收益率高达 18.4%，收益风险比达到 4.14，回撤为 3.7%。分年度来看，各策略也能稳定的战胜基准，2019 年以来策略一、策略二分别战胜基准 10.4%、6.4%个点。

图表 37、针对于中证 500 的增强策略表现汇总

	年化收益率	波动率	风险收益比	胜率	回撤
中证 500 表现	0.9%	26.7%	3.5%	54.1%	65.2%
策略一多头	19.7%	26.1%	0.76	56.0%	43.8%
策略二多头	14.4%	26.3%	0.55	55.5%	46.3%
策略一超额	18.4%	4.4%	4.14	60.1%	3.7%
策略二超额	13.2%	3.7%	3.60	59.0%	3.0%

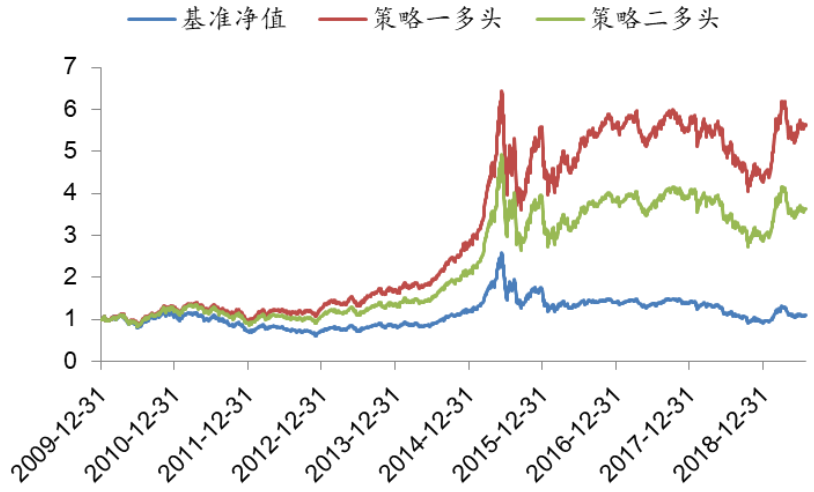
资料来源：Wind，兴业证券经济与金融研究院整理

图表 38、针对于中证 500 构建的增强策略分年度表现

		多头		超额	
	中证 500	策略一	策略二	策略一	策略二
2010 年	10.1%	26.0%	22.9%	14.5%	11.5%
2011 年	-33.8%	-21.6%	-25.7%	17.1%	11.3%
2012 年	0.3%	29.4%	19.2%	28.8%	18.6%
2013 年	16.9%	33.4%	25.4%	14.0%	7.0%
2014 年	39.0%	62.0%	53.7%	16.9%	10.7%
2015 年	43.1%	98.7%	83.9%	39.3%	29.0%
2016 年	-17.8%	1.2%	-1.3%	22.3%	20.1%
2017 年	-0.2%	0.6%	3.9%	0.7%	4.1%
2018 年	-33.3%	-22.7%	-26.8%	15.4%	9.4%
2019 年	17.6%	29.8%	25.0%	10.4%	6.4%

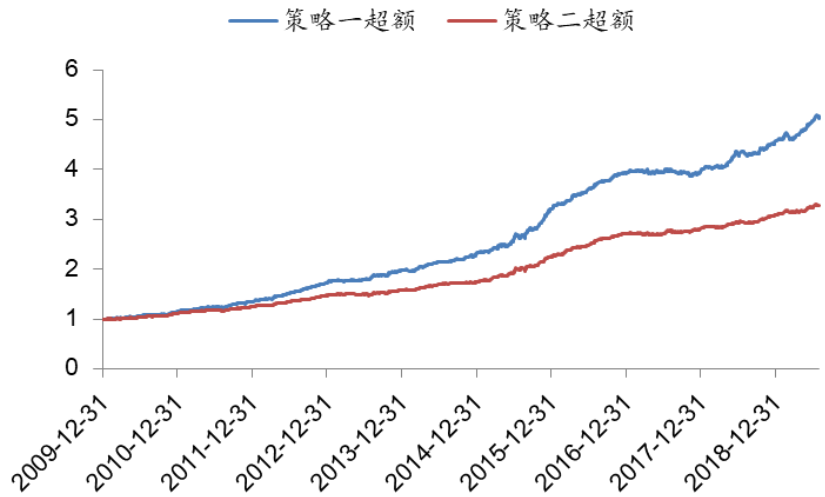
资料来源：Wind，兴业证券经济与金融研究院整理

图表 39、针对于中证 500 构建的增强策略多头净值曲线



资料来源：Wind，兴业证券经济与金融研究院整理

图表 40、针对于中证 500 构建的增强策略超额净值曲线



资料来源：Wind，兴业证券经济与金融研究院整理

5、总结

本文的研究主要聚焦于两个方面展开：1、借鉴机器学习中特征选择的方法，在不窥探未来数据的情况下构建了一套完整的动态因子选择方法，并验证了该方法的有效性；2、尝试将非线性因子与线性模型相融合，进一步构建完整的选股体系。回顾我们的机器学习系列研究，每个系列侧重解决的问题不尽相同：

1、在 2015 年初-2016 年的 5 篇系列里面，我们详细的介绍了各个机器学习模型的优缺点，以及在选股领域的应用尝试。该系列更多的是让大家对机器学习在选股领域有一个初步的认知；

2、在《基于集成学习算法的量化选股模型研究》中，我们以 Adaboost 为基础，详细探讨了机器学习和选股结合的一系列深度问题：机器学习的可解释性、过拟合问题、低换手大容量模型、高换手模型等等，更加具有针对性和落地意义；

3、本篇报告，我们聚焦于机器学习的前序篇章：因子选择问题。通过将特征选择引入进来，层层递进，我们构建了一个完整的因子选择体系。

文章最后还是回到研究的初衷：人工智能在量化投资领域的应用之路也许不会一帆风顺，但我们的努力绝不会停止，我们将坚守卖方研究的初衷，砥砺前行，争取为大家提供更多有价值的成果。

6、附录

附录一、部分中间测试结果

➢ DLS 与集成学习因子合成测试

图表 41、DLS 与集成学习因子按照不同比例或者修正 IC 合成测试

	平均值	标准差	最小值	最大值	IC_IR	t 统计量
DLS_Ada_37	0.142	0.07	-0.05	0.33	2.09	22.36
DLS_Ada_46	0.141	0.07	-0.05	0.33	2.05	21.94
DLS_Ada_Adj_IC	0.141	0.07	-0.03	0.33	2.07	22.17

资料来源：Wind，兴业证券经济与金融研究院整理

备注：以 DLS_Ada_37 为例，其表示两因子合成时，LS 占比 30%；

DLS_Ada_Adj_IC 表示：1、计算两 LS 与 Ada 修正的 IC→$IC_{ada}=(IC_{ada\_raw}-cor*IC_{DLS\_raw})/(1-cor^2)$；$IC_{DLS}=(IC_{DLS\_raw}-cor*IC_{ada\_raw})/(1-cor^2)$;2、然后以修正的 IC 权重化后加和处理得到结果；

图表 42、DLS_Ada_37 因子分位数组合测试结果

	年化收益	Sharpe	换手	最大回撤	年超额收益	跟踪误差	信息比	胜率
top	27.1%	0.89	1.31	32.1%	19.6%	4.3%	4.54	90.5%
1	20.8%	0.69	1.65	31.8%	13.7%	3.2%	4.23	86.2%
2	17.8%	0.58	1.72	36.4%	11.0%	3.2%	3.45	85.3%
3	13.9%	0.46	1.75	37.0%	7.3%	2.6%	2.81	69.8%
4	12.0%	0.39	1.76	40.3%	5.5%	2.4%	2.28	72.4%
5	7.1%	0.23	1.77	52.0%	1.0%	2.3%	0.42	55.2%
6	1.6%	0.05	1.76	58.8%	-4.2%	2.5%	-1.66	27.6%
7	-1.5%	-0.05	1.73	64.7%	-7.1%	2.7%	-2.60	19.8%
8	-7.9%	-0.26	1.67	75.6%	-13.1%	3.9%	-3.35	12.1%
bottom	-21.8%	-0.69	1.36	90.8%	-26.2%	5.9%	-4.44	2.6%
市场	6.1%	0.20		53.5%				
L_S	59.4%	6.39		1.9%				

资料来源：Wind，兴业证券经济与金融研究院整理

图表 43、DLS_Ada_46 因子分位数组合测试结果

	年化收益	Sharpe	换手	最大回撤	年超额收益	跟踪误差	信息比	胜率
top	27.0%	0.89	1.28	32.1%	19.6%	4.3%	4.59	92.2%
1	20.1%	0.67	1.63	33.5%	13.0%	3.4%	3.80	87.9%
2	18.3%	0.60	1.71	34.2%	11.4%	2.8%	4.03	88.8%
3	14.2%	0.47	1.74	37.8%	7.6%	2.8%	2.73	78.4%
4	11.7%	0.38	1.76	40.0%	5.3%	2.2%	2.40	75.9%
5	6.4%	0.21	1.76	50.7%	0.3%	2.1%	0.13	45.7%
6	2.5%	0.08	1.75	59.0%	-3.4%	2.4%	-1.44	36.2%
7	-1.8%	-0.06	1.73	65.3%	-7.3%	2.7%	-2.76	13.8%
8	-7.7%	-0.25	1.66	75.4%	-12.9%	3.9%	-3.35	13.8%
bottom	-21.7%	-0.69	1.32	90.7%	-26.0%	5.9%	-4.43	4.3%
市场	6.1%	0.20		53.5%				
L_S	59.0%	6.44		2.8%				

资料来源：Wind，兴业证券经济与金融研究院整理

图表 44、DLS_Ada_Adj_IC 因子分位数组合测试结果

	年化收益	Sharpe	换手	最大回撤	年超额收益	跟踪误差	信息比	胜率
top	27.9%	0.92	1.34	30.3%	20.4%	4.4%	4.62	92.2%
1	20.6%	0.68	1.66	34.5%	13.6%	3.0%	4.60	91.4%
2	17.0%	0.56	1.72	36.2%	10.1%	3.2%	3.20	85.3%
3	14.0%	0.46	1.75	35.2%	7.4%	2.9%	2.56	78.4%
4	11.8%	0.38	1.76	41.0%	5.4%	2.2%	2.42	74.1%
5	7.1%	0.23	1.76	51.8%	1.0%	2.2%	0.44	53.4%
6	2.0%	0.06	1.76	59.8%	-3.8%	2.5%	-1.50	31.0%
7	-1.7%	-0.06	1.73	65.3%	-7.3%	3.0%	-2.44	19.0%

8	-7.5%	-0.25	1.67	74.1%	-12.8%	3.7%	-3.42	16.4%
bottom	-22.0%	-0.70	1.38	91.0%	-26.3%	6.1%	-4.34	2.6%
市场	6.1%	0.20		53.5%				
L_S	60.5%	6.36		0.7%				

资料来源：Wind，兴业证券经济与金融研究院整理

附录二、文献引用

【1】R. E. Bellman,“Dynamic Programming, Princeton University Press,”Princeton, NJ, USA, 1957.

【2】J. Han,“Data Mining: Concepts and Techniques,”Morgan Kaufmann Publishers, San Francisco, Calif, USA, 2005.

【3】D. M. Strong, Y.W. Lee, and R. Y. Wang, “Data quality in context,” Communications of the ACM, vol. 40, no. 5, pp. 103–110, 1997.

【4】Y. Saeys, I. Inza, and P. Larra˜naga, “A review of feature selection techniques in bioinformatics,” Bioinformatics, vol. 23, no. 19, pp. 2507–2517, 2007.

【5】A. L. Blum and P. Langley, “Selection of relevant features and examples in machine learning,” Artificial Intelligence, vol. 97,no. 1-2, pp. 245–271, 1997.

【6】S. Das, Filters,“Wrappers and a boosting-based hybrid for feature selection,”in Proceedings of the 18th International Conference on Machine Learning (ICML’01), pp. 74–81, Morgan Kaufmann Publishers, San Francisco, Calif, USA, 2001.

【7】C. Ding and H. Peng, “Minimum redundancy feature selection from microarray gene expression data,”in Proceedings of the IEEE Bioinformatics Conference (CSB’03), pp. 523–528, IEEE Computer Society,Washington, DC, USA, August 2003.

【8】M. A. Hall, “Correlation-based feature selection for machine learning,” Tech. Rep., 1998.

【9】P.Yang, B. B. Zhou, Z. Zhang, and A. Y. Zomaya,“Amulti-filter enhanced genetic ensemble system for gene selection and sample classification of microarray data,” BMC Bioinformatics, vol. 11, supplement 1, article S5, 2010.

【10】H. Glass and L. Cooper, “Sequential search: a method for solving constrained optimization problems,” Journal of the ACM, vol. 12, no. 1, pp. 71–82, 1965

【11】Van Laarhoven P J M, Aarts E H L. “Simulated annealing: Theory and applications,” Springer, Dordrecht, 1987: 7-15.

【12】T. Jirapech-Umpai and S. Aitken, “Feature selection and classification for microarray data analysis: evolutionary methods for identifying predictive genes,” BMC Bioinformatics, vol. 6, article 148, 2005.

【13】C. H.Ooi and P. Tan, “Genetic algorithms applied tomulti-class prediction for the analysis of gene expression data,” Bioinformatics, vol. 19, no. 1, pp. 37–44, 2003.

【14】S. Ma, X. Song, and J. Huang, “Supervised group Lasso with applications to microarray data analysis,” BMC Bioinformatics, vol. 8, article 60, 2007.

【15】R. D´ıaz-Uriarte and S. Alvarez de Andr´es, “Gene selection and classification of

microarray data using random forest," BMC Bioinformatics, vol. 7, article 3, 2006.

【16】H. Jiang, Y. Deng, H.-S. Chen et al., "Joint analysis of two microarray gene-expression data sets to select lung adenocarcinoma marker genes," BMC Bioinformatics, vol. 5, article 81, 2004.

【17】Saeys Y, Inza I, Larrañaga P., "A review of feature selection techniques in bioinformatics," bioinformatics, 2007, 23(19): 2507-2517.

【18】Jolliffe I. , "Principal component analysis," Springer Berlin Heidelberg, 2011.

【19】Hyvärinen A, Oja E. , "Independent component analysis: algorithms and applications," Neural networks, 2000, 13(4-5): 411-430.

【20】Kruskal J B, Wish M. , "Multidimensional scaling," Sage, 1978.

【21】B. Scholkopf , A. Smola , and K.R. Muller. , "Nonlinear component analysis as a kernel eigenvalue problem," Neural Computation, 10(5): 1299- 1319, 1998

【22】J. B. Tenenbaum , V. de Silva, and J. C. Langford, "A global geometric framework for nonlinear dimensionality reduction," Science, 290, pp. 2319 - 2323, 2000

【23】Sam T. Roweis , and Lawrence K. Saul, "Nonlinear Dimensionality Reduction by Locally Linear Embedding," Science 22 December 2000

【24】Mikhail Belkin , Partha Niyogi , "Laplacian Eigenmaps for Dimensionality Reduction and Data Representation," Computation , 2003

【25】Xiaofei He, Partha Niyogi, "Locality Preserving Projections," Advances in Neural Information Processing Systems 16 (NIPS 2003), Vancouver, Canada, 2003

风险提示：本报告模型及结论全部基于对历史数据的分析，当市场环境变化时，存在模型失效风险。

分析师声明

本人具有中国证券业协会授予的证券投资咨询执业资格并注册为证券分析师，以勤勉的职业态度，独立、客观地出具本报告。本报告清晰准确地反映了本人的研究观点。本人不曾因，不因，也将不会因本报告中的具体推荐意见或观点而直接或间接收到任何形式的补偿。

投资评级说明

投资建议的评级标准	类别	评级	说明
报告中投资建议所涉及的评级分为股票评级和行业评级(另有说明的除外)。评级标准为报告发布日后的12个月内公司股价(或行业指数)相对同期相关证券市场代表性指数的涨跌幅。其中：A股市场以上证综指或深圳成指为基准，香港市场以恒生指数为基准；美国市场以标普500或纳斯达克综合指数为基准。	股票评级	买入	相对同期相关证券市场代表性指数涨幅大于15%
		审慎增持	相对同期相关证券市场代表性指数涨幅在5%～15%之间
		中性	相对同期相关证券市场代表性指数涨幅在-5%～5%之间
		减持	相对同期相关证券市场代表性指数涨幅小于-5%
		无评级	由于我们无法获取必要的资料，或者公司面临无法预见结果的重大不确定性事件，或者其他原因，致使我们无法给出明确的投资评级
	行业评级	推荐	相对表现优于同期相关证券市场代表性指数
		中性	相对表现与同期相关证券市场代表性指数持平
		回避	相对表现弱于同期相关证券市场代表性指数

信息披露

本公司在知晓的范围内履行信息披露义务。客户可登录 www.xyzq.com.cn 内幕交易防控栏内查询静默期安排和关联公司持股情况。

使用本研究报告的风险提示及法律声明

兴业证券股份有限公司经中国证券监督管理委员会批准，已具备证券投资咨询业务资格。

“本公司”）的客户使用，本公司不会因接收人收到本报告而视其为客户。本报告中的信息、意见等均仅供客户参考，不构成所述证券买卖的出价或征价邀请或要约。该等信息、意见并未考虑到获取本报告人员的具体投资目的、财务状况以及特定需求，在任何时候均不构成对任何人的个人推荐。客户应当对本报告中的信息和意见进行独立评估，并应同时考量各自的投资目的、财务状况和特定需求，必要时就法律、商业、财务、税收等方面咨询专家的意见。对依据或者使用本报告所造成的一切后果，本公司及/或其关联人员均不承担任何法律责任。

本报告所载资料的来源被认为是可靠的，但本公司不保证其准确性或完整性，也不保证所包含的信息和建议不会发生任何变更。本公司并不对使用本报告所包含的材料产生的任何直接或间接损失或与此相关的其他任何损失承担任何责任。

本报告所载的资料、意见及推测仅反映本公司于发布本报告当日的判断，本报告所指的证券或投资标的的价格、价值及投资收入可升可跌，过往表现不应作为日后的表现依据；在不同时期，本公司可发出与本报告所载资料、意见及推测不一致的报告；本公司不保证本报告所含信息保持在最新状态。同时，本公司对本报告所含信息可在不发出通知的情形下做出修改，投资者应当自行关注相应的更新或修改。

除非另行说明，本报告中所引用的关于业绩的数据代表过往表现。过往的业绩表现亦不应作为日后回报的预示。我们不承诺也不保证，任何所预示的回报会得以实现。分析中所做的回报预测可能是基于相应的假设。任何假设的变化可能会显著地影响所预测的回报。

本公司的销售人员、交易人员以及其他专业人士可能会依据不同假设和标准、采用不同的分析方法而口头或书面发表与本报告意见及建议不一致的市场评论和/或交易观点。本公司没有将此意见及建议向报告所有接收者进行更新的义务。本公司的资产管理部门、自营部门以及其他投资业务部门可能独立做出与本报告中的意见或建议不一致的投资决策。

本报告并非针对或意图发送予或为任何就发送、发布、可得到或使用此报告而使兴业证券股份有限公司及其关联子公司等违反当地的法律或法规或可致使兴业证券股份有限公司受制于相关法律或法规的任何地区、国家或其他管辖区域的公民或居民，包括但不限于美国及美国公民（1934 年美国《证券交易所》第 15a-6 条例定义为本「主要美国机构投资者」除外）。

特别声明

在法律许可的情况下，兴业证券股份有限公司可能会持有本报告中提及公司所发行的证券头寸并进行交易，也可能为这些公司提供或争取提供投资银行业务服务。因此，投资者应当考虑到兴业证券股份有限公司及/或其相关人员可能存在影响本报告观点客观性的潜在利益冲突。投资者请勿将本报告视为投资或其他决定的唯一信赖依据。

兴业证券研究

上 海	北 京	深 圳
地址：上海浦东新区长柳路36号兴业证券大厦15层	地址：北京西城区锦什坊街35号北楼601-605	地址：深圳市福田区皇岗路5001号深业上城T2座52楼
邮编：200135	邮编：100033	邮编：518035
邮箱：research@xyzq.com.cn	邮箱：research@xyzq.com.cn	邮箱：research@xyzq.com.cn