



机器学习驱动的基本面量化投资研究

李斌、邵新月和李玥阳
《中国工业经济》，2019

解读人：李斌
2019年11月11日

第一页，说明展示的文献名称、作者、期刊（工作论文写“Working Paper”）、年份等；下方是解读者和解读日期

目录

1. 引言
2. 模型设计
3. 实证结果
4. 结论

2019/11/11

李斌@武大金融

2

目录的结构：

1. 引言：介绍背景、研究动机、研究问题、研究内容概述、创新点和贡献

1. 引言

➤研究背景

- 1. 资产保值和增值是每个家庭和个人都会面临的问题。
- 2. 资产管理是金融服务实体经济的重要手段之一
- 3. 2017年,《新一代人工智能发展规划》,人工智能上升为国家战略
- 4. 智能量化投资是我国金融业高质量发展的重要组成部分。
- 问题: 如何发挥以机器学习为代表的人工智能技术优势,推动我国资产管理水平的提升?

2019/11/11

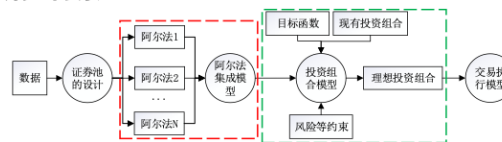
李斌@武大金融

3

背景从现实出发, 研究通常从现实问题出发。条目化列出来, 不要大段拷贝问题, 字体大小最小为20号字体, 建议正文采用24号字体 (强迫自己总结, 利用简洁的文字表达)。

1. 引言

➤研究背景



其中，红色框内为资产定价模块；绿色框内为投资组合模块。

- **基本面量化投资**融合了量化投资（算法驱动）与基本面投资（人为驱动），从**异象因子**出发集成能够提供超额收益的阿尔法信号并构建投资组合。
- 部分投资者表现出“算法厌恶”，根源在于对量化投资的机理和决策机制缺乏深入的了解，亟待开展对智能量化投资方面的研究。

2019/11/11

李斌@武大金融

4

逐步从现实问题步入研究问题。从大到小，研究问题来源于现实问题，同时逐渐抽象化。

1. 引言

➤研究动机

1. 金融研究提出数以百计的异象因子集合，但后续样本外检验发现大部分因子难以持续地提供超额收益，且因子间往往具有较强相关性
2. 传统的组合排序和FM回归并未综合考虑各因子及因子间的交互作用；因子维度变大时，非线性因素使预测的复杂度急剧增加，亟待新方法介入
3. 前美国金融学会会长Cochrane (2011)：在处理如此众多的因子时，将不得不使用“不同的研究工具”(“I suspect we will have to use different methods.”)

2019/11/11

李斌@武大金融

5

研究动机从研究问题所面临的挑战出发，突出为什么要去研究某一个具体的点？通过动机，你能否说服自己去解决所面临的研究挑战？说服不了自己，认真想和读。

1. 引言

➤研究动机

- 机器学习和深度学习能够自动地寻找数据中的**复杂结构和模式**，从而提升预测能力：
 - 众多备选的预测函数形式；
 - 专门被设计用于逼近复杂的非线性关系；
 - 参数正则化和模型选择等技术有效降低过拟合风险。

2019/11/11

李斌@武大金融

6

研究动机2，用新方法解决老问题通常需要说明新方法相对老方法的优势。

研究问题

- **研究问题一**：将机器学习算法与异象因子结合构建的股票收益预测模型能否通过有效识别数据间的非线性关系获得**更好的预测效果**？
- **研究问题二**：若机器学习算法的运用能够提升预测绩效，究竟**哪些因子**能够更好地预测股票未来收益？

2019/1/11

李斌@武大金融

7

在说明研究的动机之后，提出研究问题。一篇文章有1-2个研究问题。后面的实证/理论都是为了解决研究问题而服务，要围绕着研究问题展开。本文有2个研究问题，所以后面分了两章。如果只有1个，分一章也可以。

1. 引言

➤现有研究

1. 机器学习在资产定价中的应用

- Feng, Giglio, and Xiu(2017)基于LASSO方法衡量因子对资产定价的贡献，发现盈利性和投资因子更具有统计上显著的解释力
- 后续也有用Adaptive LASSO和Adaptive Group LASSO进行因子的筛选
- 李斌等（2017）分别采用支持向量机、神经网络、Adaboost算法预测股价涨跌方向，发现机器学习算法具有更高的准确率
- Gu et al. (2018)检验了常见的机器学习算法在美国市场上的预测能力，发现机器学习模型可以有效地超越线性模型

2019/11/11

李斌@武大金融

8

现有研究是如何解决研究问题，主要阅读文章的文献综述部分。尽量总结。

1. 引言

➤现有研究

2. 异象因子研究

- Green, Hand和Zhang (2017)采用Fama-MacBeth回归检验了1980-2014年美国股票市场的94项异象因子，发现12个因子对股票月度收益具有显著的预测效果。
- Light, Maslov, and Rytchkov (2017)采用“偏最小二乘法”(PLS)来检验公司特征对期望收益的预测能力；
- Hou et al. (2019)检验了447个文献中的异象因子，发现大部分并不能预测股票未来收益；
- Jiang等 (2019)分别采用Fama-MacBeth回归、PCA、PLS和FC方法整合A股市场中的75个异象因子，证明上述线性方法能够从因子中提取出有助于预测的信息。

2019/11/11

李斌@武大金融

9

1. 引言

➤研究内容

1. 搜集中国A股市场**96个异象因子数据**，采用**12种线性&非线性机器学习算法**构建异象因子-超额收益预测模型，并构建投资组合，以检验模型在A股市场的表现
2. 系统性地对比各种机器学习算法的绩效，证明机器学习算法能够显著超越传统线性回归的绩效；且**非线性算法**的预测绩效明显超过线性机器学习算法；深度学习算法**DFN和LSTM**绩效提升最为明显
3. 分析异象因子的重要性发现**交易摩擦类因子**在A股市场具有更强的预测能力

2019/11/11

李斌@武大金融

10

1. 所读文献是如何解决研究问题？概述研究内容，不要超过1页。用大字，强迫自己总结。
2. 解决研究问题所得最重要的结论是什么？概述，1-2个条目结课，用24号以上的大字。

1. 引言

➤创新点

1. **因子集合的创新**。本文构建的96个异象因子集合是**目前中国A股市场相关研究中最大的因子集合**，基于全面的数据集上的分析预测能够获取更有效的信息。
2. **预测模型的创新**。将基本面量化投资与机器学习算法结合构建预测模型的设计在中国股票市场的研究中相对缺乏，本文使用12种线性/非线性机器学习算法构建多因子预测模型，是目前**中国股票市场最全面的利用机器学习算法进行的异象因子研究**。

2019/11/11

李斌@武大金融

11

总结文章的创新点。文章都会讲，提取出来即可。

1. 引言

➤ 研究意义

1. 丰富了经济学和管理学研究的工具箱。
2. 丰富了量化投资的理论和实践研究。
3. 本文丰富了中国市场中股票截面收益影响因素的研究。

2019/11/11

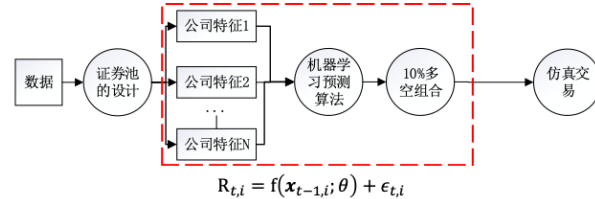
李斌@武大金融

12

这里是研究意义或者贡献，主要是从理论上来讲。同样要大字。

2. 研究设计

►模型的总体设计



$$R_{t,i} = f(\mathbf{x}_{t-1,i}; \theta) + \epsilon_{t,i}$$

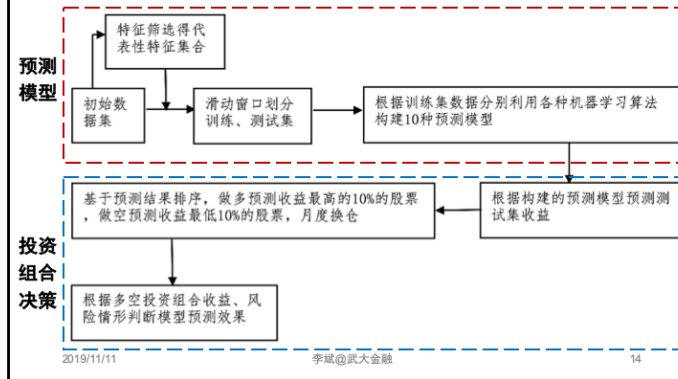
- $f(\cdot)$ 定义一个参数为 θ 的函数，在本文中为丰富的机器学习和深度学习方法中的函数形式
- $R_{t,i}$ 为股票 i 第 t 期的超额收益
- $\mathbf{x}_{t-1,i} = (x_{t-1,i,1}, x_{t-1,i,2}, \dots, x_{t-1,i,N})$ 为公司 i 第 $t-1$ 期的 N 个异象因子
- $\epsilon_{t,i}$ 为误差项。

第一部分“引言”讲完后，基本上小同行专家都已经了解本文做什么了。

第二部分讲“研究设计”，主要告诉听众，本文解决研究问题的主要途径，通过这一套研究设计检验的结果在第3部分实证结果中。比如通过实证验证的话，说明研究假设、实证研究方法和实证数据等等。本文是讲整套的研究设计。这里需要更加详细化，需要用学术的语言讲。不要随意发挥，很容易被专家看出来。多想想为什么：为什么推导出本文的假设？是否可能还有其他的推导路径？解决研究问题为什么需要采用本文的研究方法？是否还能采用其他更好或更合适的研究方法？为什么采用本文的实证数据？X和y变量是否会有更好的代理变量？等等。

2. 研究设计

模型的总体设计



继续总体设计。

2. 研究设计 – 数据

➤时间区间：1997年1月-2018年10月A股市场月频数据

➤因子集合

- 借鉴Green, Hand and Zhang (2017)，选取了96个公司特征代理异象因子，分为交易摩擦因子、动量因子、价值因子、成长因子、盈利因子、财务流动因子共六大类。

➤数据处理

- 对于季度财务数据均进行月度填充，数据来源于CSMAR。
- 缺失值处理：①若股票在第t月收益数据存在缺失，则剔除该股票在月份t上的所有数据；②若某只股票的因子值缺失，则以0填充。
- 剔除掉ST、金融股及上市一年内数据后，共381062条有效样本

1月 ^t	2月 ^t	3月 ^t	4月 ^t	5月 ^t	6月 ^t	7月 ^t	8月 ^t	9月 ^t	10月 ^t	11月 ^t	12月 ^t
t-1年三季度报表填充 ^a				t-1年年度报表填充 ^a				t年半年度报表填充 ^a		t年三季度报表填充 ^a	

2019/1/11

字海星AI入图

15

说明数据的选择：1. 数据的区间、频率、市场等；2. 选取数据的维度；3. 关键的数据处理步骤，本文作了填充，需要特别说明；4. 最后所得数据的统计（样本数等）。同样，每一步多想想为什么？常见的选择原因：1. 最长可用的数据区间、全股票集合等；2. 参考别人的做法（文献）。

2. 研究设计 – 数据

➤ 单因子检验中显著的因子集合（22项）

异常因子	size	std_dvol	LM	lagretn	rd_mve	size_ia	volumed
年化收益率	21.26%	18.18%	16.71%	15.17%	13.93%	12.13%	11.65%
t-statistics	3.8701	8.7984	7.7101	3.8841	5.4255	2.911	5.9414
std_turn	illq	mom36	momchg	chfeps	depr	aeavol	retnmax
11.62%	9.21%	9.07%	9.01%	8.44%	8.09%	6.67%	5.99%
6.5113	4.2123	2.563	2.8739	4.7377	2.1071	3.6105	4.5582
SP	vol	idvol	skew	pchsaleinv	SglNVg	CRG	
5.23%	5.00%	4.74%	4.54%	3.70%	3.63%	2.91%	
2.0128	2.3778	3.2926	3.774	2.8627	2.924	2.1851	

2019/1/11

李斌@武大金融

16

数据的描述性统计，不要超过1页。没啥特殊说明，可以不展示。

2. 研究设计 – 研究方法

➤机器学习算法，采用监督学习算法

0. 基准算法：线性回归模型

1. 线性算法：FC、Ridge、Lasso、ElasticNet、PLS

2. 非线性算法：SVM、GBDT、XGBoost、EN-ANN

3. 深度学习算法：DFN、RNN、LSTM

➤训练、测试集划分：滑动窗口法，采用网格搜索和交叉验证选参数

第1组：

199701	199702	...	199711	199712	199801
				1	2

第2组：

199702	199703	...	199712	199801	199802
				1	2

第3组：

199703	199704	...	199801	199802	199803
				1	2

.....

第n组：

201712	201801	...	201810	201811	201812
				1	2

2019/11/11

于第10次输入数据

17

1. 说明在前述研究框架中采用的方法。这里只列出了方法，口述说明选取理由，本文的预期结果（或想验证的假设）有三点：1. 线性的表现>基准; 2. 非线性>线性; 3. 深度学习>非线性。
2. 框架中比较重要的滑动窗口方法。

2. 研究设计 – 研究方法

- 为了探索和发现影响中国市场股票截面收益的重要因子，本文采用如下的方式从机器学习的视角进一步审视中国股票截面收益的影响因素。

单因子	统计学意义上显著的因素（多空组合收益 $t\text{-statistic} > 1.96$ ）
机器学习方法	从全变量中剔除单一因子之后的收益损失

2019/1/11

李斌@武大金融

18

1. 该页说明了第二个研究问题的研究方法，可以放到研究设计中去（论文中是这样写的）。
2. 研究问题2事实上是研究问题1进一步检验的结果，研究问题1的建议在总体设计中已经说明。当然也可以将检验研究问题1和2的方式显式说明，会更好。

3. 实证结果1：机器学习模型在A股市场的实证绩效

2019/1/11

李斌@武大金融

19

“实证结果”是为了通过“研究设计”去解决“研究问题”所给出，通过对实证结果的解读，回答研究问题。注意：论文通常有很多表格，但最重要的表格通常1个研究问题1-2张，其他稳健性等等，都是分支，不是最重要的，文献解读时可以忽略。

回到本文的实证结果的解读从两个角度展开。

1. 机器学习模型在A股市场的实证绩效
2. 在机器学习的视角下，因子的重要性。

因为下一页的表格太大，因此这边用了一个空白页，以显示这是研究问题1的实证结果。大家做的话，可以将这一页和下一页合并。

机器学习算法全变量预测结果对比(12月滑动窗口)						
	多头组合			多空组合		
	mean	FF5- α	夏普比率	mean	FF5- α	夏普比率
OLS	2.47%	0.90%	0.73	2.10%	1.67%	1.52
FC	2.71%	1.01%	0.80	2.42%	1.62%	1.33
Ridge	2.52%	0.95%	0.75	2.17%	1.73%	1.57
Lasso	2.55%	0.96%	0.76	2.27%	1.82%	1.61
Elastic	2.52%	0.95%	0.75	2.17%	1.73%	1.57
PLS	2.52%	0.95%	0.75	2.17%	1.73%	1.57
SVM	2.60%	1.07%	0.79	2.42%	2.01%	1.87
EN-ANN	2.51%	0.94%	0.76	2.22%	1.72%	1.74
Xgboost	2.69%	0.98%	0.81	2.54%	1.92%	1.80
GBDT	2.71%	1.00%	0.82	2.56%	1.92%	1.79
DFN	2.94%	1.34%	0.88	2.88%	2.34%	2.01
RNN	2.52%	1.10%	0.78	2.10%	1.79%	1.97
LSTM	2.86%	1.18%	0.85	2.57%	2.01%	1.96
SIZE	2.45%	0.51%	0.70	1.73%	0.27%	0.68
MKT	0.61%	-0.01%	0.1795			

1. 解决研究问题1所需要的实证结果，有2个panel，1个是数值，1个是统计检验。
2. 表格比较大，因此分成2张ppt。有必要的话，只截取其中重要的行/列即可，展示结果的含义最重要，细枝末节可以忽略。
3. 严格注意表格中字体的大小，保证可读性。不要小于20号字体，截图的话，将其中的文字与20号字体比较下。
4. 通过不同颜色的框图，将重点突出。

➤ 机器学习算法全样本预测结果对比——NW-t检验(12个月滑动窗口)

OLS与其它算法		DFN与其它算法	
	多空组合		多空组合
FC	1.1523	FC	1.8546
Ridge	1.5746	Ridge	3.3132
Lasso	1.9734	Lasso	2.7663
ElasticNet	1.9544	ElasticNet	2.8118
PLS	1.7723	PLS	2.0576
EN-ANN	1.0578	EN-ANN	4.1133
Xgboost	2.6956	Xgboost	1.3189
GBDT	3.0773	GBDT	1.0130
DFN	3.2764		
RNN	3.7663		
LSTM	3.8879		

- 机器学习算法都较OLS回归投资绩效存在明显提升
- DFN能够显著超越Ridge、Lasso、ElasticNet、PLS和EN-ANN算法获得更高的投资绩效

2019/1/11

李斌@武大金融

21

在讲完之后，一定要用文字总结下该表最主要的观察。注意用大字，提炼总结。

➤ 剔除市值因子后的预测绩效（12个月滑动窗口）

检验模型是否受到市值因子（单因子平均月度收益最高，为1.77%）的驱动而并非多个因子的聚合效果。

	多空组合（全变量）			多空组合（去除市值）		
	mean	FF5- α	夏普比率	mean	FF5 α	夏普比率
OLS	2.10%	1.67%	1.5211	2.12%	1.35%	1.2933
FC	2.42%	1.62%	1.3314	1.41%	0.77%	0.6318
Ridge	2.17%	1.73%	1.5707	2.23%	1.61%	1.4803
Lasso	2.27%	1.82%	1.6134	1.93%	1.58%	1.2059
Elastic	2.17%	1.73%	1.5707	1.92%	1.57%	1.2016
PLS	2.17%	1.73%	1.5707	1.60%	1.19%	0.9233
SVM	2.42%	2.01%	1.8759	2.44%	1.80%	1.4424
ENANN	2.22%	1.72%	1.7479	1.77%	1.38%	1.3008
Xgboost	2.54%	1.92%	1.8027	2.11%	1.58%	1.2777
GBDT	2.56%	1.92%	1.7962	2.03%	1.49%	1.2533
DFN	2.88%	2.34%	2.0132	2.98%	2.07%	1.8516
RNN	2.10%	1.79%	1.9794	2.11%	1.70%	1.4679
LSTM	2.57%	2.01%	1.9670	2.59%	2.17%	1.6036

2019/1/11

李斌@武大金融

22

1. 稳健性检验，时间紧可以不要。
2. Ppt上一定要用1-2句话说明做什么和结果说明了什么。

➤ 集成各类算法预测结果

为了更直观的说明机器学习算法相对于传统线性模型的绩效提升，本文简单加权11种机器学习算法（其中未包括线性组合的FC）构建集成预测模型： $R_{t,i}^{ensemble} = \frac{1}{11} \sum_{j=1}^{11} R_{t,i}^j$

3个月滑动窗口				12个月滑动窗口		
	多空组合	多头组合	空头组合	多空组合	多头组合	空头组合
mean	2.56%	2.60%	0.04%	2.98%	2.89%	-0.09%
FF3- α	2.38%	1.34%	-1.20%	2.67%	1.41%	-1.46%
FF5- α	2.18%	1.20%	-1.21%	2.55%	1.28%	-1.48%
夏普比率	1.4698	0.7715	-0.0651	2.1736	0.8720	-0.1085
24个月滑动窗口				36个月滑动窗口		
	多空组合	多头组合	空头组合	多空组合	多头组合	空头组合
mean	2.56%	2.60%	0.04%	2.98%	2.89%	-0.09%
FF3- α	2.38%	1.34%	-1.20%	2.67%	1.41%	-1.46%
FF5- α	2.18%	1.20%	-1.21%	2.55%	1.28%	-1.48%
夏普比率	1.4698	0.7715	-0.0651	2.1736	0.8720	-0.1085

2019/1/11

李斌@武大金融

23

稳健性说明。说明做了什么？建议再加个结果说明。

3. 实证结果2：机器学习的视角下因子的重要性。

2019/11/11

李斌@武大金融

24

同样

➤ 被选中次数超过7次(>=50%)的因子集合

序号	因子	因子名称	因子类别	N
1	aeavol	收益公告异常交易量	交易摩擦因子	11
2	turnsd	换手率的波动率	交易摩擦因子	11
3	egr	股东权益变化	成长因子	10
4	LM	标准化的换手率	交易摩擦因子	10
5	retvol	总波动率	交易摩擦因子	9
6	skewness	总偏态	交易摩擦因子	9
7	vold	交易额	交易摩擦因子	9
8	illq	非流动性风险	交易摩擦因子	8
9	CFdebt	现金流负债比	财务流动性因子	7
10	idvol	异质波动率	交易摩擦因子	7
11	lagretn	短期反转	动量因子	7
12	tang	偿债能力/总资产	财务流动性因子	7

观察：交易摩擦因子在重要因子中占比较大。

2019/1/11

李斌@武大金融

25

为了证明研究问题2所给出的实证结果表。

同时，并非是交易摩擦因子本身占总体比例较大导致

因子类别	因子总数	重要因子数	占比
交易摩擦因子	21	10	48%
财务流动性因子	10	2	20%
动量因子	6	1	17%
盈利因子	14	1	7%
成长因子	35	1	3%
价值因子	10	0	0%

交易摩擦因子具有较强预测能力的原因？有待继续探索。

2019/1/11

李斌@武大金融

26

1. 很可能会有人猜测，是因为交易摩擦类因子比例过大所导致。为了排除这一猜测，继续检验。
2. 标题直接说明了验证的假设。
3. 说明后续并未被解释。红色高亮

➤ 集成结果筛选出代表性因子12个月滑动窗口预测结果

	多空组合（全变量）			多空组合（筛选变量）		
	mean	FF5- α	夏普比率	mean	FF5- α	夏普比率
OLS	2.10%	1.67%	1.52	2.51%	1.78%	1.65
FC	2.42%	1.62%	1.33	3.03%	2.14%	1.47
Ridge	2.17%	1.73%	1.57	2.53%	1.83%	1.59
Lasso	2.27%	1.82%	1.61	2.60%	1.89%	1.60
Elastic	2.17%	1.73%	1.57	2.61%	1.93%	1.61
PLS	2.17%	1.73%	1.57	2.63%	1.92%	1.63
SVM	2.42%	2.01%	1.87	2.75%	2.03%	2.26
EN-ANN	2.22%	1.72%	1.74	2.33%	1.88%	1.84
Xgboost	2.54%	1.92%	1.80	2.61%	2.37%	2.01
GBDT	2.56%	1.92%	1.79	2.65%	2.34%	2.04
DFN	2.88%	2.34%	2.01	3.51%	2.99%	2.68
RNN	2.10%	1.79%	1.97	3.23%	2.68%	2.71
LSTM	2.57%	2.01%	1.96	3.39%	2.61%	2.69

结合研究问题1和研究问题2所衍生的稳健性检验。

4. 结论

1. 机器学习算法能够**识别异常因子间的非线性关系**而获得更好的投资收益，即使考虑交易成本和做空限制也能获得显著的超额收益
2. **非线性机器学习算法**的绩效表现总体**优于线性算法**；深度学习算法(DFN和LSTM)获得了最好的投资绩效
3. 因子重要性显示，以收益公告异常交易量为代表的**交易摩擦类因子**对股票收益具有较强的预测能力。
4. 多数方法在筛选后的因子集合的投资绩效能够超越其在全变量集合上的投资绩效。
5. 未来工作主要在交易摩擦类因子的影响渠道。

2019/1/11

李斌@武大金融

28

总结做了什么；和未来要做什么。

谢谢！

李斌@武大金融

2019/11/11

李斌@武大金融

29

标准动作。