

基金持股中的股票嵌入与股票关联

*STOCK EMBEDDINGS AND STOCK CORRELATION
FROM MUTUAL FUND HOLDINGS*

wins-m 2024.5

目录

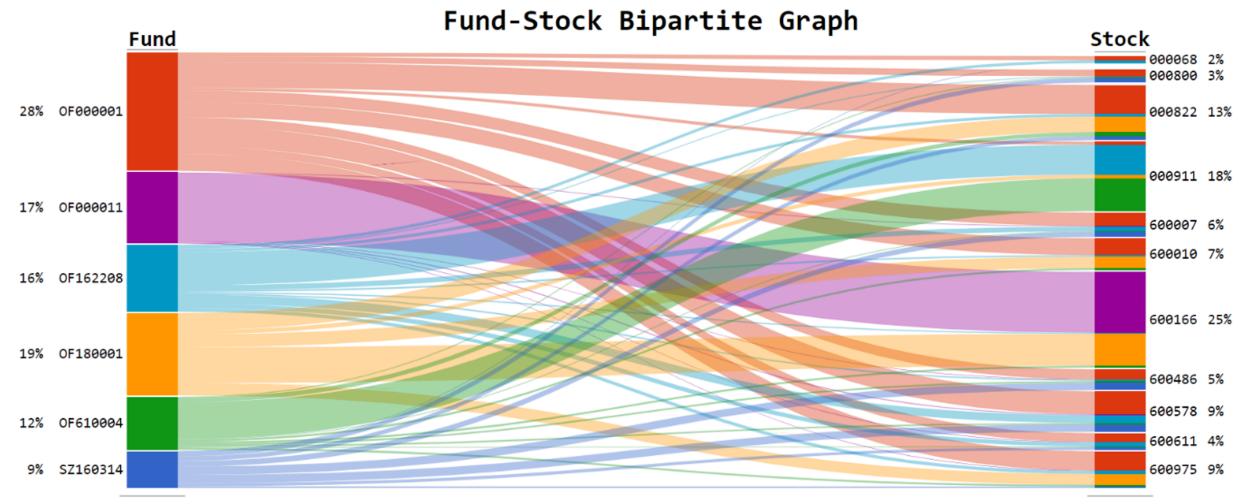
1. 问题背景
2. 研究假设
3. 研究方法
4. 研究结论
5. 研究启发

1. 问题背景

- 基金持股：公开市场信息，但不是结构化的股票特征
 - 问题一：基金持股能否体现主动基金的选股能力？若有，如何提取？

华夏成长混合 2023年4季度股票投资明细							来源: 天天基金 截止至: 2023-12-31
序号	股票代码	股票名称	相关资讯	占净值比例	持股数(万股)	持仓市值(万元)	
1	002025	航天电器	股吧 行情	4.16%	209.92	10,080.49	
2	600862	中航高科	股吧 行情	3.48%	380.43	8,426.47	
3	300034	钢研高纳	股吧 行情	3.39%	403.16	8,204.29	
4	300395	菲利华	股吧 行情	3.27%	216.80	7,926.38	
5	600519	贵州茅台	股吧 行情	2.90%	4.07	7,024.82	
6	000100	TCL科技	股吧 行情	2.74%	1,544.74	6,642.40	
7	600941	中国移动	股吧 行情	2.55%	62.11	6,178.70	
8	002475	立讯精密	股吧 行情	2.47%	173.91	5,991.20	
9	603986	兆易创新	股吧 行情	2.19%	57.35	5,298.94	
10	002371	北方华创	股吧 行情	2.16%	21.23	5,216.94	

[显示全部持仓明细>>](#)



- 股票嵌入：学习股票特征向量
 - 问题二：基金持股中的公司特征能否被嵌入模型有效提取？
 - 问题三：对比已有文献中的人工提取，嵌入方法是否更有效？

Asset Embeddings

Xavier Gabaix Ralph S.J. Koijen Robert J. Richmond Motohiro Yogo*

September 14, 2023. Preliminary and incomplete

Abstract

Firm characteristics are ubiquitously used in economics. These characteristics are often based on readily-available information such as accounting data, but those reflect only a part of investors' information set. We show that useful information about firm characteristics is embedded in investors' holdings data and, via market clearing, in prices, returns, and trading data. Based on insights from the recent artificial intelligence (AI) and machine learning (ML) literature, in which unstructured data (e.g., words or speech) are represented as continuous vectors in a potentially high-dimensional space, we propose to learn asset embeddings from investors' holdings data. Indeed, just as documents arrange words that can be used to uncover word structures via embeddings, investors organize assets in portfolios that can be used to uncover firm characteristics that investors deem important via asset embeddings. This broad theme provides a natural bridge to connect recent advances in the fields of AI and ML to finance and economics. Specifically, we show how *language* models, including transformer models that feature prominently in large language models such as BERT and GPT, can handle *numerical* information, and in particular holdings data to estimate asset embeddings. We provide initial evidence on the value added of asset embeddings through a series of applications in the context of firm valuations, return comovement, and uncovering asset substitution patterns. As a by-product, the models generate investor embeddings, which can be used to measure investor similarity. We propose a programmatic list of potential applications of asset and investor embeddings to finance and economics more generally.

 **开源证券**

金融工程专题

2021年10月02日

从基金持仓行为到股票关联网络

——金融工程专题

金融工程研究团队

魏建榕（首席分析师） 证书编号: S0790519120001	王志豪（联系人） weijianrong@kysec.cn 证书编号: S0790519120001
张 翔（分析师） 证书编号: S0790520110001	魏建榕（分析师） 证书编号: S0790519120001
傅开波（分析师） 证书编号: S0790520090003	王志豪（联系人） wangzihao@kysec.cn 证书编号: S0790120070080
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	

● 基金共同持仓行为是股票关联关系的重要来源

学术界对于股票关联网络的研究由来已久，用于构建关联网络的市场信息主要集中于涨跌幅、成交额、换手率等维度。本报尝试从基金持仓维度出发，探索基金共同持仓股票间的关联网络及应用。本文分别从“基金管理人认知”和“股东协同行为”两个角度理解“基金共同持仓行为是股票关联关系的重要来源”。

基金管理人认知的角度：基金持仓反映管理人在个股层面用脚投票，基金共同持仓两只股票，反映两只股票对管理人而言具有某一方面共性；

股东协同行为的角度：被基金共同持有的股票，其股东成分有交集，从而导致其市场表现存在一定程度关联。

● 关联度指标统计分析：同行业股票间关联度更高

我们按照股票市值划分，将基金持仓股票按照市值由小到大分为5组，测试各组内股票间关联度指标均值，可以看到，大市值股票之间的关联度指标与小市值无显著区别，关联度指标在市值上无暴露。按照两只股票是否属于同一行业的标准划分，可以看到，同行业股票间的关联度指标均值高于不同行业股票间的关联度指标均值。

● 关联网络牵引因子表现稳健

构建股票关联网络之后，我们尝试利用关联网络刻画股票涨跌之间的牵引关系。换言之，股票a的关联股票的涨跌幅有锚定效应，若当月其关联股票普遍上涨，会提高市场对于股票a的涨幅预期。若本月其自身涨幅不高，则预期在次月出现补涨行情。基于这一观点，我们尝试构建关联网络牵引因子 Traction20d。

我们按照月末调仓，双边千三的费率，测算 Traction20d 因子在 20130701 至 20210830 的表现，因子 RankIC 为 3.84%，RankICIR 达到 2.60。5 分组净值曲线分化程度较高，多头组年化收益达到 17.1%，多头换手率为 69.5%，收益波动比

相关研究报告

*xgabaix@fas.harvard.edu, ralph.koijen@chicagobooth.edu, rrichmon@stern.nyu.edu, myogo@princeton.edu. For comments and suggestions, we thank John Campbell, Tengyuan Liang, Tarun Ramadorai, and seminar participants at the Chicago Booth machine learning working group, the NY Fed, the London Quant Group 2022 Autumn Seminar, and the NBER SI 2022 Forecasting & Econometric Methods conference.

嵌入方法（Word2Vec）原理

- “基金持仓-股票” v.s. “文本语句-单词”
 - A 经常在 B 附近出现：A 和 B 相似/相互替代
 - | “宫殿里有国王和王后”
 - A 和 B 都经常在 C 附近出现：A 和 B 相似/相互替代
 - | “国王站在台上” “王后坐在台上”
- 上下文关系相邻 v.s. 投资者需求、股东结构相似

- 股票关联：股票向量余弦距离 & 嵌入模型损失函数的关键
 - 问题四：提取获得的股票关联中蕴含哪些信息？受哪些公司特征影响？与股价收益率的关系？

创新模型 - W2V相似度 (论文 3.2.2 小节)

- 数据：基金 i 持仓 $\Lambda_i = [s_{i1}, s_{i2}, \dots, s_{im_i}]$
 - $s_{ir}, r = 1, \dots, m_i$ 为基金 i 持股中的股票代码，按 **持仓金额大小排序** 后依次排在第 $r = 1, 2, \dots, m_i$ 位
- 学习：股票 s 的嵌入表示 v_s (单位向量，维数取30)
- 损失函数： $\text{loss} = - \sum_{i \in I} \sum_{r=1}^{m_i} \sum_{-K \leq j \leq K, j \neq 0} \ln \Pr(s_{i,r+j} | s_{i,r})$

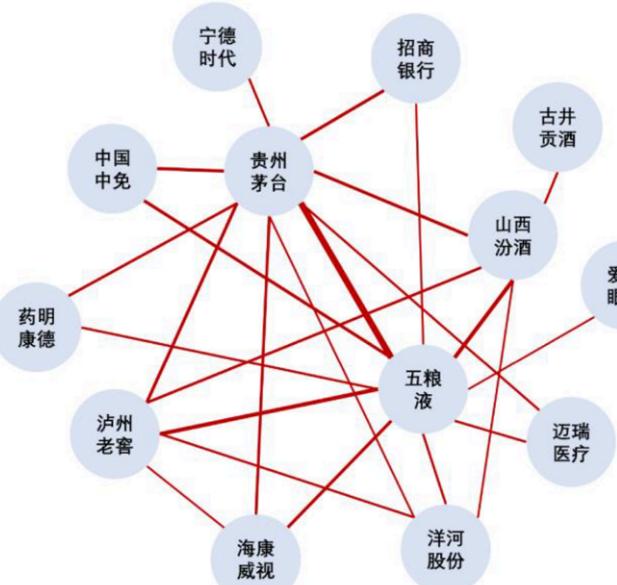
$$\Pr(s_{r+j} | s_r) = \frac{\exp(v_{s_{r+j}}^T v_{s_r})}{\sum_{s' \in S} \exp(v_{s'}^T v_{s_r})}$$

- $\Pr(s_{r+j} | s_r)$ 为 **股票 S_t 附近出现 S_{t+j} 的条件概率**，核心是余弦相似度 $v_{s'}^T v_{s_r}$

$$\text{sim}_{a,b} = v_a \cdot v_b$$

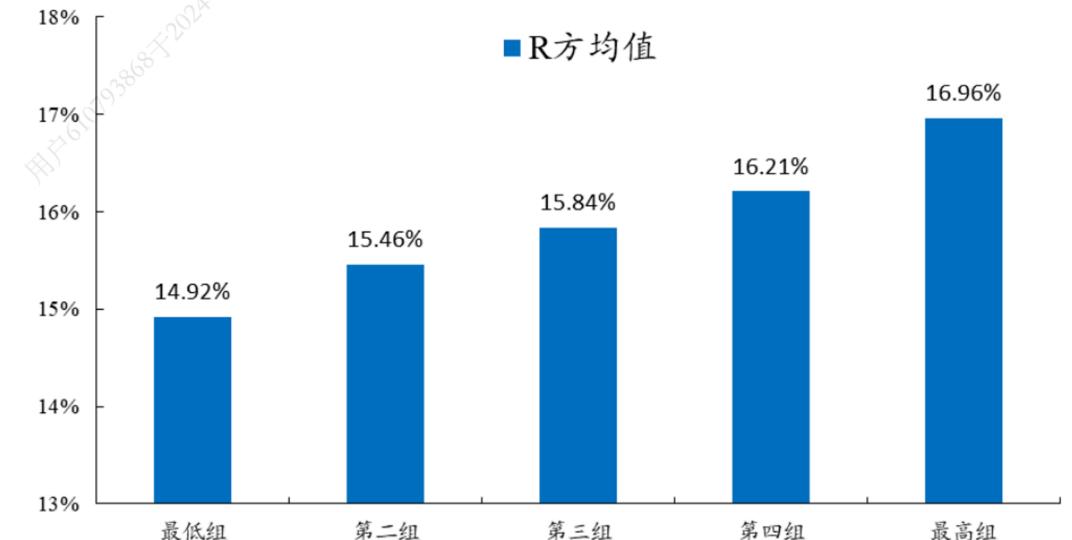
- 关联动量：(1) 相关性分析的统计结果；(2) 模型评估的评判标准
 - 问题五：信息在公开披露后是否被市场价格充分反映？

图7：关联网络预览图（局部）



资料来源：开源证券研究所（连线粗细表示关联强弱）

图10：更高的关联度指标意味着股票涨跌间更高的 R 方



数据来源：Wind、开源证券研究所

1. 问题背景

- 基金持股：公开市场信息，但不是结构化的股票特征
 - 问题一：基金持股能否体现主动基金的选股能力？若有，如何提取？
- 股票嵌入：学习股票特征向量
 - 问题二：基金持股中的公司特征能否被嵌入模型有效提取？
 - 问题三：对比已有文献中的人工提取，嵌入方法是否更有效？
- 股票关联：股票向量余弦距离 & 嵌入模型损失函数的关键
 - 问题四：提取获得的股票关联中蕴含哪些信息？
- 关联动量：(1) 相关性分析的统计结果；(2) 模型评估的评判标准
 - 问题五：信息在公开披露后是否被市场价格充分反映？

2. 研究假设

1. 依赖基金持股信息，通过人工或嵌入表示的方法能够获得具有实际意义的公司数值特征。
2. 上述公司数值特征所刻画的公司关联度，能够反映股票的风格暴露和行业分类特征，并且与未来的价格表现相关。
3. 相比人工构造，嵌入方法在提取基金持股中股票的显性特征时更加高效，与风格暴露和行业分类之间的相关性更高。
4. 相比人工构造，嵌入方法在提取基金持股中股票的隐性特征时更加高效，与股票未来收益之间的相关性更高。
5. 通过基金持股信息获得的关联股票的未来收益之间存在关联动量，在控制公司风格特征后能够获得超额收益。

3. 研究方法

1. 模型建立：基金持股 (fund, stock, amount) --> 股票关联 (stock1, stock2, corr)

- (1) 人工构造 (2) 嵌入模型

2. 相关性分析：关联度指标统计性质

- 关联度指标分布
- 股票关联 vs. 风格特征 (*)
- 股票关联 vs. 行业分类 (**)
- 股票关联 vs. 收益相关性 (**)

3. 模型应用：关联股票收益相关 (统计规律) --> 关联动量 (定价异象)

- 因子评估
- Fama-MacBeth检验

模型

基准模型 - 人工关联度 (论文 3.2.1 小节)

- 基金对个股的影响力: $I_s^i = H_s^i / AMT_s$
 - H_s^i : 基金 i 对股票 s 的持仓市值
 - AMT_s : 股票 s 最近 20 日成交额均值
- 基金 i 同时持仓股票 a 和 b 带来的关联强度: $J_{a,b}^i = \min(I_a^i, I_b^i)$
- 股票 a, b 的关联度: $K_{a,b} = \sum_{i \in \text{Inst}} J_{a,b}^i$
 - 规范化: $K_{a,b}^* = \frac{\ln K - \min \ln K}{\max \ln K - \min \ln K}$

创新模型 - W2V相似度 (论文 3.2.2 小节)

- 数据: 基金 i 持仓 $\Lambda_i = [s_{i1}, s_{i2}, \dots, s_{im_i}]$
 - $s_{ir}, r = 1, \dots, m_i$ 为基金 i 持股中的股票代码, 按 **持仓金额大小排序** 后依次排在第 $r = 1, 2, \dots, m_i$ 位
- 学习: 股票 s 的嵌入表示 v_s (单位向量, 维数取30)
- 损失函数: $\text{loss} = - \sum_{i \in I} \sum_{r=1}^{m_i} \sum_{-K \leq j \leq K, j \neq 0} \ln \Pr(s_{i,r+j} | s_{i,r})$

$$\Pr(s_{r+j} | s_r) = \frac{\exp(v_{s_{r+j}}^T v_{s_r})}{\sum_{s' \in S} \exp(v_{s'}^T v_{s_r})}$$

- $\Pr(s_{r+j} | s_r)$ 为 股票 S_t 附近出现 S_{t+j} 的条件概率, 核心是余弦相似度 $v_{s'}^T v_{s_r}$

$$\text{sim}_{a,b} = v_a \cdot v_b$$

4. 研究结论

1. 嵌入模型能有效利用基金持股数据提取公司数值特征，对股票关联进行定量刻画
2. 关联度：受行业分类、市值特征影响
 - 同行业，低于不同行业
 - 市值越高，平均关联度越低
 - 嵌入模型：分异更明显
3. 关联度：与未来股价相关
 - 嵌入模型：关联度更高，下季度收益相关性更高
 - 人工模型：无分异
4. 关联动量异象：基金持股隐含股票特征，且未被市价充分反映
 - 因子检验：持续有效
 - 嵌入模型：更显著
 - Fama MacBeth回归：不能被定价模型解释

4.1 嵌入模型能有效利用基金持股数据提取公司数值特征，对股票关联进行定量刻画

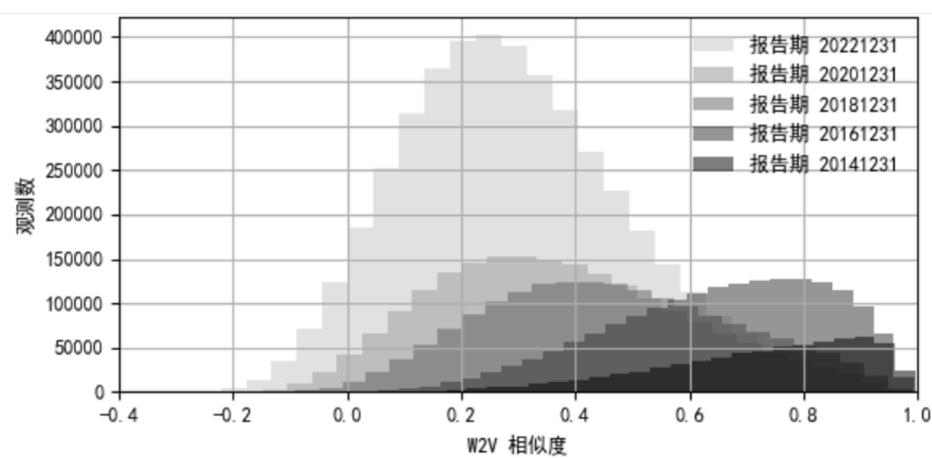
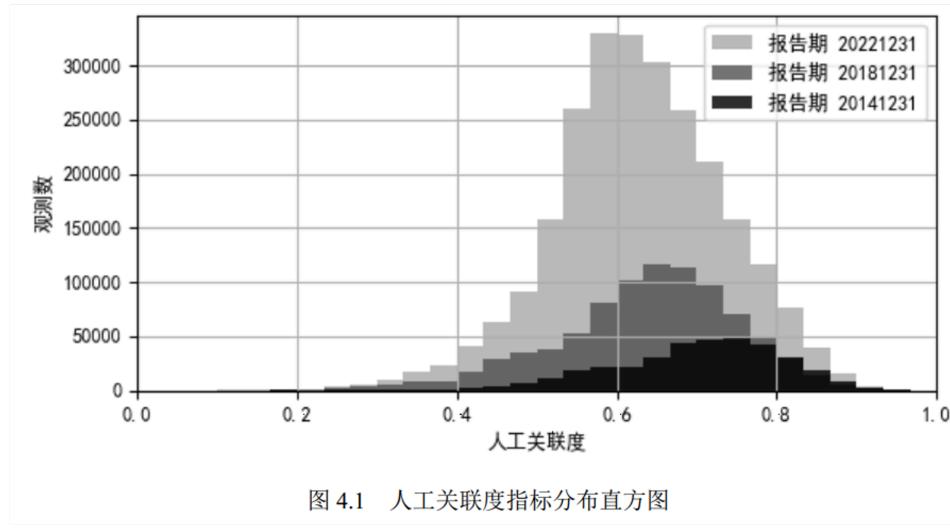


表 4.2 贵州茅台 (600519.SH) 人工关联度最高的股票

排名	股票代码	股票简称	人工关联度	所属行业	流通市值 (亿)
0	600519.SH	贵州茅台	-	食品饮料	21694.5
1	000568.SZ	泸州老窖	0.9582	食品饮料	3284.6
2	000858.SZ	五粮液	0.9511	食品饮料	7013.4
3	600809.SH	山西汾酒	0.9475	食品饮料	3470.4
4	300015.SZ	爱尔眼科	0.9416	医药	1803.0
5	000596.SZ	古井贡酒	0.9388	食品饮料	1090.6
6	603259.SH	药明康德	0.9384	医药	2072.5
7	300750.SZ	宁德时代	0.9347	电力设备及新能源	7800.4
8	300760.SZ	迈瑞医疗	0.9346	医药	3830.9
9	601888.SH	中国中免	0.9342	消费者服务	4217.9
10	600887.SH	伊利股份	0.9277	食品饮料	1055.5

表 4.8 贵州茅台 (600519.SH) W2V 相似度最高的股票

排名	股票代码	股票简称	W2V 相似度	所属行业	流通市值 (亿)
0	600519.SH	贵州茅台	-	食品饮料	21694.5
1	000568.SZ	泸州老窖	0.9381	食品饮料	3284.6
2	000858.SZ	五粮液	0.9217	食品饮料	7013.4
3	601888.SH	中国中免	0.9160	消费者服务	4217.9
4	300059.SZ	东方财富	0.8994	非银行金融	2151.2
5	600036.SH	招商银行	0.8920	银行	7686.3
6	600809.SH	山西汾酒	0.8856	食品饮料	3470.4
7	300750.SZ	宁德时代	0.8650	电力设备及新能源	7800.4
8	000596.SZ	古井贡酒	0.8515	食品饮料	1090.6
9	002304.SZ	洋河股份	0.8383	食品饮料	2411.8
10	600887.SH	伊利股份	0.8370	食品饮料	1055.5

4.2 关联度：受行业分类、市值特征影响

- 同行业，高于不同行业；市值越高，平均关联度越低；嵌入模型：分异更明显

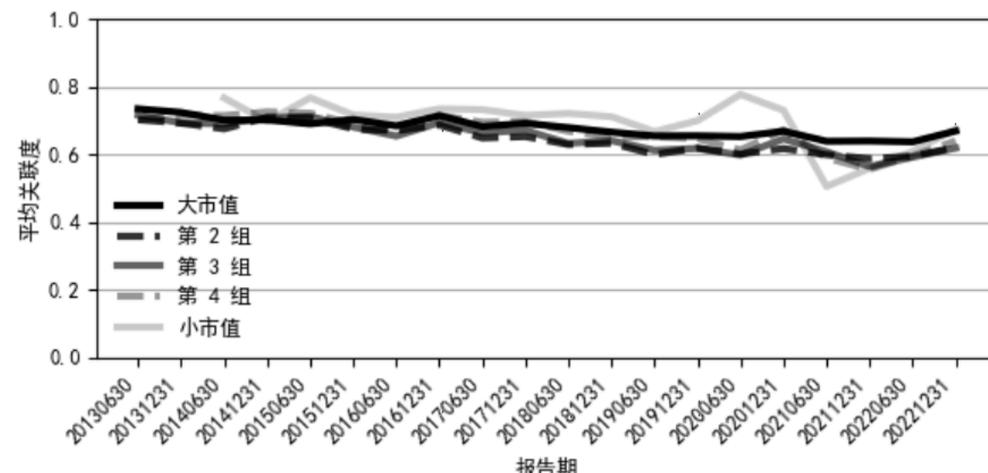


图 4.3 不同流通市值分组下的人工关联度指标历史均值

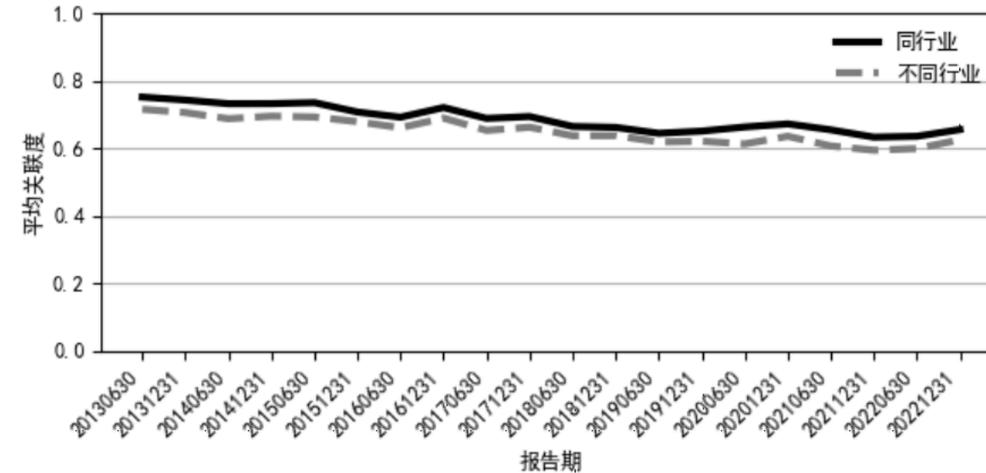
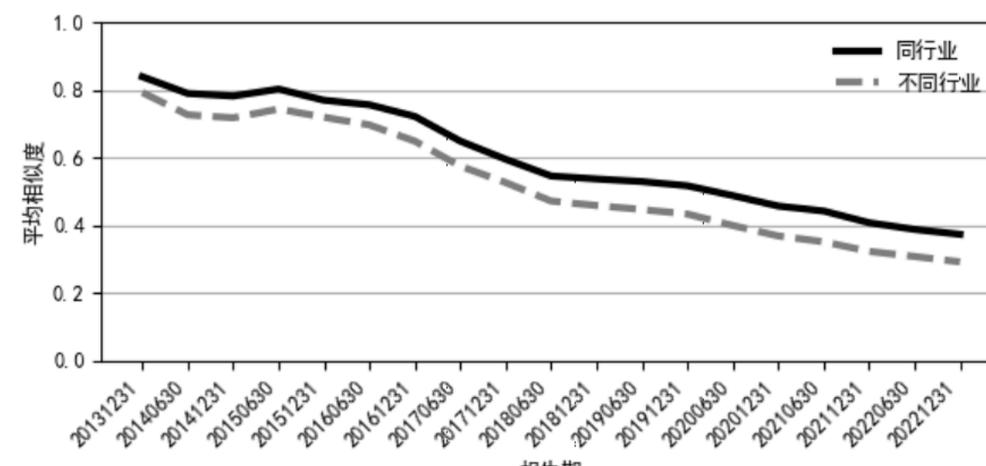
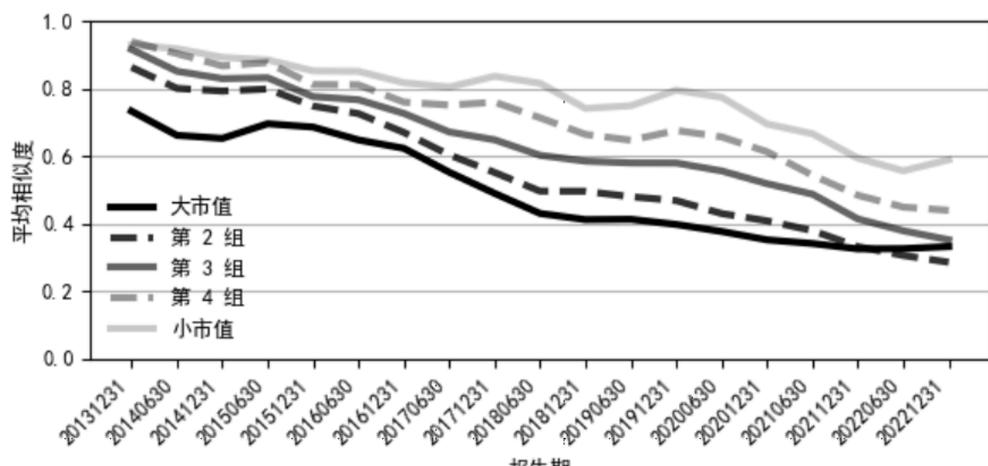


图 4.5 同行业及不同行业之间的人工关联度指标历史均值



4.3 关联度：与未来股价相关

- 嵌入模型：关联度更高，下季度收益相关性更高；人工模型：无分异

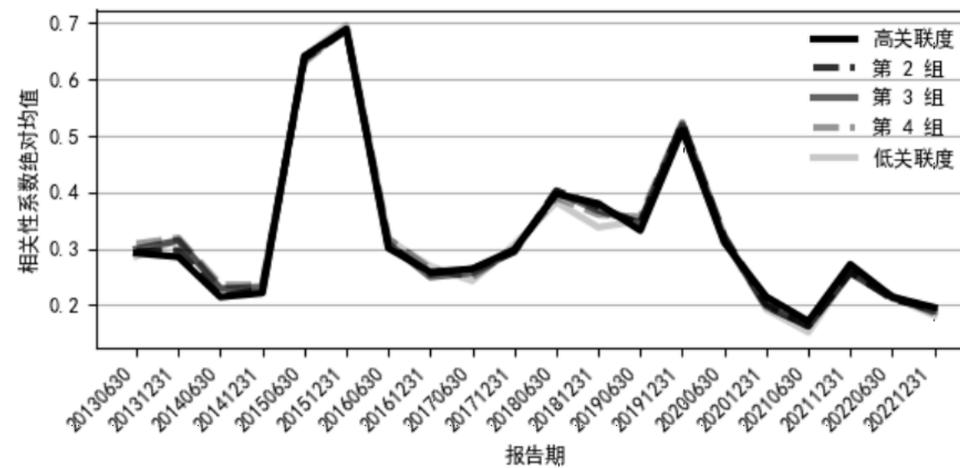


图 4.9 不同人工关联度分组内股票收益率相关性系数绝对值均值（每期抽样 10 万对）

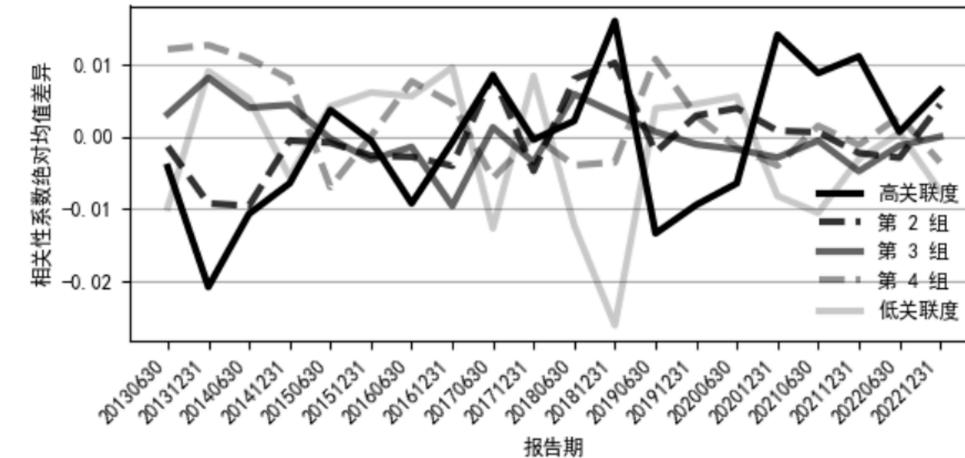
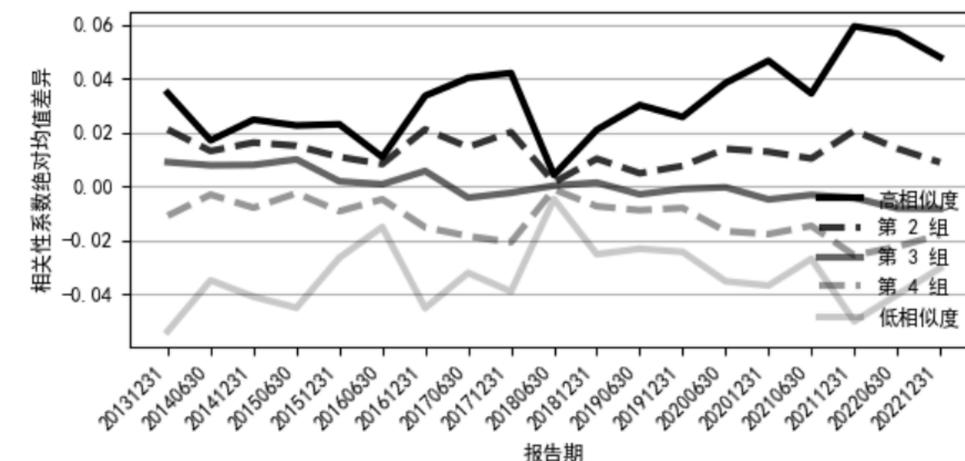
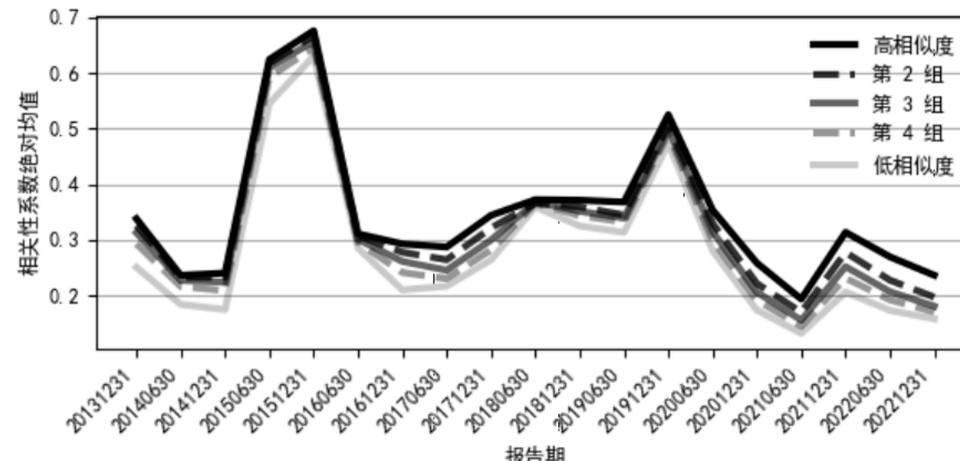


图 4.11 不同人工关联度分组内股票收益率相关性系数绝对值均值（相对于均值）



4.4 关联动量异象：基金持股隐含股票特征，且未被市价充分反映

模型应用 - 因子构造 (论文 4.3.1 小节)

- ① 取关联网络中所有样本股票过去 20 日复权开盘价涨跌幅，减去截面均值，刻画股票 i 在过去 20 日的超额收益（式 4.2）

$$R_{i,t}^{20} = Open_{i,t}/Open_{i,t-19} - 1$$

$$AR_{i,t}^{20} = R_{i,t}^{20} - \frac{1}{n} \sum_{i=1}^n R_{i,t}^{20}$$

- ② 取股票 i 所有关联股票 j 的超额收益，根据关联度（或相似度）加权求和，获得所有股票过去 20 日来源于关联股票的 预期超额收益（式 4.3）

$$ER_{i,t} = \sum_{j \neq i} AR_{i,t}^{20} \cdot \lambda_{i,j}$$

其中 $\lambda_{i,j}$ 为最新基金持仓数据中股票 i 和 j 的关联度（相似度）

② 取股票 i 所有关联股票 j 的超额收益，根据关联度（或相似度）加权求和，获得所有股票过去 20 日来源于关联股票的预期超额收益（式 4.3）

$$ER_{i,t} = \sum_{j \neq i} AR_{i,t}^{20} \cdot \lambda_{i,j}$$

其中 $\lambda_{i,j}$ 为最新基金持仓数据中股票 i 和 j 的关联度（相似度）

③ 每期截面上，预期超额收益对实际的过去 20 日超额收益，以及其他控制的风格、行业因子载荷进行回归（式 4.4）

$$ER_{i,t}^{20} = \beta_1 \cdot AR_{i,t}^{20} \left(+ \sum_k \beta_k f_{i,t}^{(k)} \right) + \varepsilon_{i,t}$$

取残差 $\varepsilon_{i,t}$ 即得到关联动量牵引因子

- 因子检验：持续有效

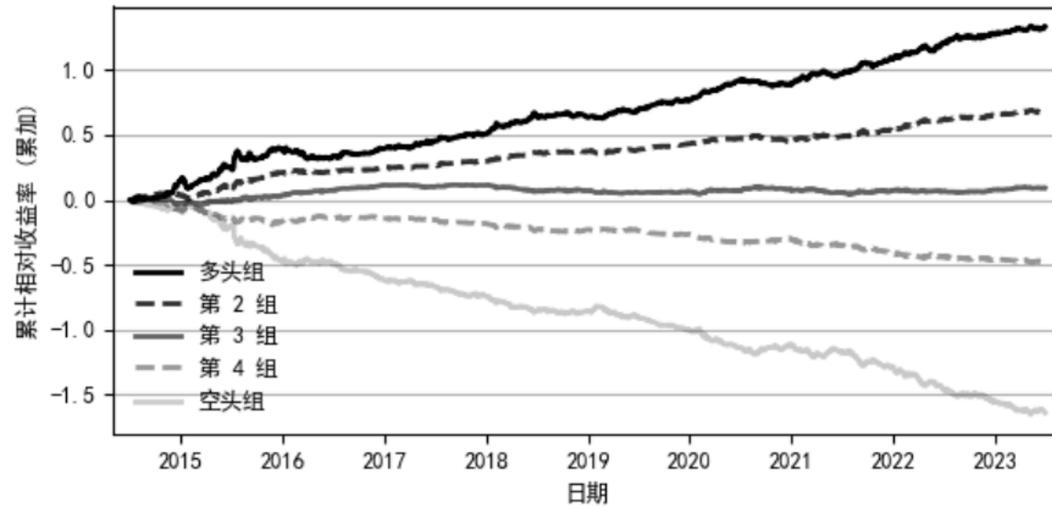


图 4.15 W2V 相似度牵引因子 5 分组收益分化情况



图 4.18 W2V 相似度牵引因子 5 分组多头/多空策略回测表现

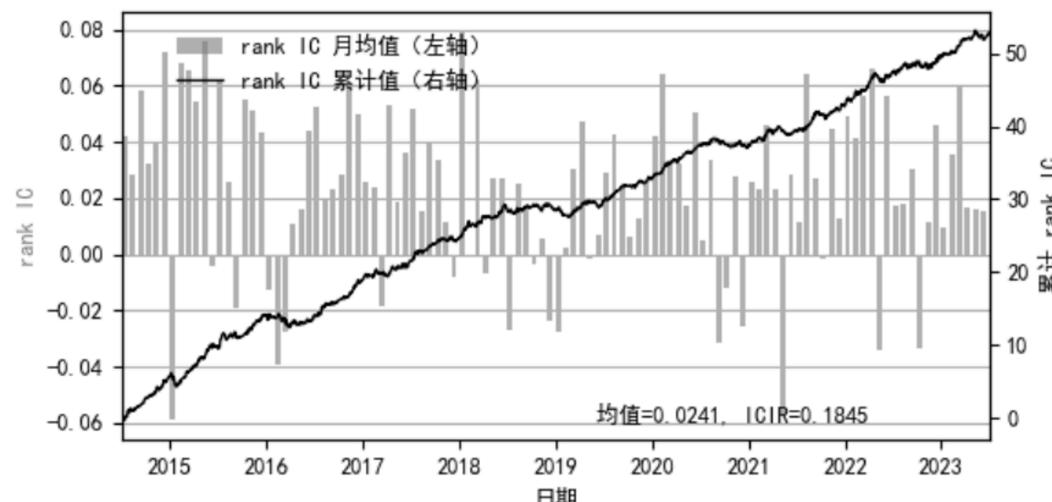


表 4.12 W2V 关相似牵引因子回测结果 (2014.7-2023.6)

分组	年化收益	夏普率	年化超额收益	最大回撤	日度胜率	日均换手率
多头组	17.9%	0.60	9.2%	64.3%	56.3%	36.3%
多 - 空	16.3%	2.35	7.6%	7.9%	56.8%	-
空头组	-14.6%	-0.48	-23.3%	86.3%	51.3%	37.4%

• 嵌入模型：更显著

表 4.13 中性化前后因子评估表现 (2014.7-2023.6)

因子	rank IC 均值	ICIR	日均样本数
人工关联度牵引因子	1.85%	0.1786	1362.6
人工关联度牵引因子（行业中性）	1.63%	0.2252	1362.6
人工关联度牵引因子（行业、市值中性）	1.55%	0.2488	1362.6
W2V 相似度牵引因子	2.41%	0.1845	1706.4
W2V 相似度牵引因子（行业中性）	2.31%	0.2526	1706.4
W2V 相似度牵引因子（行业、市值中性）	2.27%	0.3052	1706.3

表 4.14 中性化前后因子多头/多空策略回测表现 (2014.7-2023.6)

分组	年化收益	夏普率	年化超额收益	最大回撤	日度胜率	日均换手率
人工关联度牵引因子						
多头组	15.5%	0.51	6.8%	63.9%	55.6%	38.0%
多 - 空	12.5%	2.07	3.9%	11.6%	57.6%	-
空头组	-9.6%	-0.32	-18.2%	87.2%	51.8%	37.9%
人工关联度牵引因子（行业中性）						
多头组	12.8%	0.42	4.1%	64.0%	55.2%	37.6%
多 - 空	10.7%	2.44	2.0%	9.2%	58.2%	-
空头组	-8.6%	-0.29	-17.3%	83.4%	52.0%	37.3%
人工关联度牵引因子（行业、市值中性）						
多头组	11.3%	0.38	2.7%	68.4%	55.0%	37.5%
多 - 空	9.4%	2.56	0.7%	5.4%	58.4%	-
空头组	-7.4%	-0.25	-16.0%	74.0%	52.3%	37.5%
W2V 相似度牵引因子						
多头组	17.9%	0.6	9.2%	64.3%	56.3%	36.3%
多 - 空	16.3%	2.35	7.6%	7.9%	56.8%	-
空头组	-14.6%	-0.48	-23.3%	86.3%	51.3%	37.4%
W2V 相似度牵引因子（行业中性）						
多头组	16.1%	0.53	7.4%	65.4%	56.3%	36.2%
多 - 空	15.5%	3.21	6.9%	3.7%	59.5%	-
空头组	-15.0%	-0.5	-23.7%	86.6%	51.4%	37.1%
W2V 相似度牵引因子（行业、市值中性）						
多头组	13.7%	0.46	5.0%	65.8%	55.8%	36.3%

• Fama MacBeth回归：不能被定价模型解释

表 4.15 Fama-MacBeth 截面回归变量说明

变量	简介
被解释变量	
超额收益	相对中证全指的超额收益 (%)；复权收盘价, T+1 买入, T+2 卖出。
异象变量	
人工	人工关联度牵引因子。
人工.I	人工关联度牵引因子，经过行业中心化。
W2V	W2V 相似度牵引因子。
W2V.I	W2V 相似度牵引因子，经过行业中心化。
控制变量	
beta	表征股票相对于市场的波动敏感度。
市值	捕捉大盘股和小盘股之间的收益差异。
动量	描述了过去两年里相对强势的股票与弱势股票之间的差异。
残余波动率	解释了剥离了市场风险后的波动率高低产生的收益率差异。
价值	描述了股票估值高低不同而产生的收益差异，即价值因子。
非线性市值	描述了无法由规模因子解释的但与规模有关的收益差异。
盈利	描述了由盈利收益导致的收益差异。
流动性	解释了由股票相对的交易活跃度不同而产生的收益率差异。
杠杆	描述了高杠杆股票与低杠杆股票之间的收益差异。
成长	描述了对销售或盈利增长预期不同而产生的收益差异。
人工	1.000
人工.I	-0.928 1.000
W2V	-0.538 0.454 1.000
W2V.I	-0.499 0.507 0.918 1.000
beta	-0.037-0.015-0.071-0.044 1.000
市值	-0.019-0.023-0.001-0.006-0.114 1.000
动量	-0.216-0.232-0.161-0.193-0.050 0.073 1.000
残余波动率	-0.111-0.100-0.128-0.119 0.381-0.043 0.212 1.000
价值	-0.059-0.049-0.075-0.066 0.241-0.026 0.082 0.354 1.000
非线性市值	-0.007-0.008-0.012-0.016 0.019-0.632-0.046-0.028 0.006 1.000
盈利	-0.044-0.034 0.055 0.046-0.195 0.314-0.020-0.247-0.355-0.172 1.000
流动性	-0.084-0.078-0.077-0.075 0.343-0.169 0.249 0.620 0.308 0.022-0.224 1.000
杠杆	-0.008-0.001 0.017 0.009-0.113 0.198-0.003-0.095-0.236-0.049 0.058-0.111 1.000
成长	-0.033-0.030-0.041-0.042 0.013 0.163 0.063 0.039 0.113-0.112 0.217 0.047 0.007 1.000

表 4.17 Fama-MacBeth 回归结果 (2015.1-2023.6)

	(1)	(2)	(3)	(4)	(5)
人工	0.0182 [5.42]				
人工.I		0.0158 [5.96]			
W2V			0.0226 [4.57]		
W2V.I				0.0202 [5.53]	
beta	0.0224 [2.79]	0.0224 [2.76]	0.0242 [3.00]	0.0228 [2.81]	0.0235 [2.88]
市值	-0.0225 [-3.60]	-0.0223 [-3.54]	-0.0207 [-3.32]	-0.0217 [-3.44]	-0.0232 [-3.68]
动量	-0.0351 [-4.40]	-0.0354 [-4.44]	-0.0342 [-4.15]	-0.0340 [-4.13]	-0.0408 [-5.25]
残余波动率	0.0463 [6.89]	0.0463 [6.83]	0.0460 [6.99]	0.0458 [6.90]	0.0453 [6.73]
价值	0.0005 [0.08]	0.0004 [0.06]	0.0019 [0.30]	0.0009 [0.14]	0.0006 [0.09]
非线性市值	0.0084 [2.62]	0.0085 [2.65]	0.0094 [2.91]	0.0091 [2.82]	0.0089 [2.77]
盈利	0.0050 [1.21]	0.0050 [1.19]	0.0059 [1.44]	0.0052 [1.26]	0.0053 [1.29]
流动性	-0.0989 [-9.35]	-0.0977 [-9.87]	-0.0981 [-9.87]	-0.0981 [-9.75]	-0.0970 [-9.74]
杠杆	-0.0045 [-1.23]	-0.0047 [-1.27]	-0.0051 [-1.42]	-0.0049 [-1.34]	-0.0047 [-1.27]
成长	0.0244 [7.37]	0.0244 [7.35]	0.0246 [7.54]	0.0247 [7.50]	0.0244 [7.39]

5. 研究启发

- **基金持股** 数据蕴含股票特征信息，未能被市场充分关注。
- **嵌入方法** 能够提取基金持股等非结构化数据中的资产特征，有时优于人工构造。
- 股票特征的相似程度——股票关联，部分地被市值、行业等显性特征解释，额外也包含隐形的价格关联。

谢谢！