版权声明

摘要

共同基金每季度公开披露的基金持股中包含重要的股票特征信息。这部分信息由两方面组成，包括基金经理利用信息优势所掌握的公司状况，以及基金经理进行组合配置时的持股意愿。但是，这些反映股票最新情况的信息是隐含的，在解读和应用时存在困难。为了更有效地提取隐含的股票特征，本文引入嵌入方法（Embedding Methods），从基金持股的“上下文”顺序中学习股票嵌入。其中，股票嵌入表示的空间距离是嵌入方法的关键，同股票显性特征、股票收益表现之间可能存在联系。

本文主要使用 Word2Vec 模型提取基金持股中的股票嵌入。围绕嵌入表示后的股票关联，本文首先进行相关性分析，研究关联度指标和股票行业分类、风格特征、收益率相关性等显性特征的关系，从而考察股票嵌入表示的可解释性。基于关联股票之间收益率相关这一现象，本文构造关联动量因子，并通过 Fama-MacBeth 回归对因子背后的定价异象进行实证检验。在上述研究过程中，本文也通过人工构造的方式从基金持股中提取股票关联，用于同嵌入方法对比。

本文的研究表明，嵌入方法能够从基金持股中获得具有可解释性的股票特征，有助于发现相似的股票和相似的行业。股票嵌入表示的关联度指标和股票行业分类、股票流通市值之间存在单调关系；关联度指标和股票未来收益率的相关性具有单调关系；上述现象在使用人工关联度指标时不明显。嵌入表示获得的股票关联动量因子在 2014.7 至 2023.6 期间日度 rank IC 达到 2.41%，行业和市值中性化后的 ICIR 为 0.305；在控制 Barra 风格因子后能够通过 Fama-MacBeth 检验。

关键词：基金持股，嵌入表示，股票关联，定价异象

STOCK EMBEDDINGS AND STOCK CORRELATION FROM MUTUAL FUND HOLDINGS

wins-m (Master of Finance)
Directed by Prof. Yu-Jane Liu

ABSTRACT

Mutual funds' quarterly disclosures of fund holdings contain important stock characteristic information. This information consists of two aspects: the latest company performance that fund managers grasp using their informational advantage, and the stock holding preferences of fund managers when making portfolio allocations. However, these pieces of information reflecting the latest stock conditions are implicit, leading to difficulties in interpretation and application. To more effectively extract implicit stock characteristics, this paper introduces embedding methods, which learn stock embeddings from the contextual order in fund holdings. The spatial distance between stocks is crucial in embedding methods and may be related to explicit stock features and stock performance.

This paper primarily uses the Word2Vec model to extract stock embeddings from fund holdings. Focusing on the correlation of stocks after representation embedding, this paper first conducts correlation analysis to examine the relationship between stock correlation indicators and industry classification, style characteristics, and return co-movements, thereby investigating the interpretability of embedding representations. Based on the phenomenon of return correlations among associated stocks, this paper constructs a correlation momentum factor and empirically tests the pricing anomalies behind the factor through Fama-MacBeth regression. In the aforementioned research process, this paper also extracts stock correlations from fund holdings manually for comparison with embedding methods.

The research in this paper indicates that embedding methods can obtain interpretable stock characteristics from fund holdings, facilitating the identification of similar stocks and industries. There exists a monotonic relationship between correlation indicators of stock embedding representations and stock industry classification, as well as stock market capitalization. The correlation indicators also exhibit a monotonic relationship with future stock returns; however, this phenomenon is less apparent when using manually constructed correlation indicators. The stock correlation momentum factor obtained from embedding representations achieved a daily rank IC of 2.41% during the period from July 2014 to June

2023, with an ICIR of 0.305 after industry and market neutralization; the results withstand Fama-MacBeth testing after controlling for Barra style factors.

KEY WORDS: Fund holdings, Stock embedding, Stock correlation, Anomaly

目录

第一章　引言

1.1　选题背景

中国共同基金数量远超沪深主板的股票数量。基金管理人的投资能力或许有别于个体投资者，具有专业性和信息优势。但是，也存在一脉文献认为，主动基金管理人的投资存在无效性，无法超越市场指数或随机投资。即使市场有效性存在争议，随着基金数量的增长，投资规模的扩大，投资赛道、风格和方法的多样化，共同基金作为国民投资理财工具的重要性正在日益凸显；与此同时，基金持股已经成为影响市场的重要驱动力以及市场公开信息的重要组成部分，成为市场参与者投资决策和风险控制中不可忽视的力量。

但是，投资人对基金持股数据的利用存在不充分性。在传统的资产管理中，即使各家间机构长期存在信息的勾兑和交换，也依然很少有专业资产管理人能够综合考虑所有同行的持股情况。更一般地，特定基金或基金经理只聚焦于某一赛道或主题，在进行投资决策时仅仅关注相同赛道或主题下的其他资产管理人。这固然源于人类主观决策时“有限注意力”等能力局限，同时也表明公众对于价值、风格、赛道投资等有经济逻辑的投资方法有更高的偏好。在方法论惯性和投资人需求的作用下，共同基金的投资逻辑更倾向于关注行业的趋势及公司本身的财务状况，而几乎不可能对外宣称自身的投资逻辑是基于竞争对手的持股行为做出相应的跟随或对赌反应。

量化投资对于市场公开信息的充分利用具有先天优势。上承自资本资产定价模型（Capital Asset Pricing Model, CAPM）和套利定价理论（Arbitrage Pricing Theory, APT），线性多因子定价模型[①]是实证资产定价（empirical asset pricing）和因子投资（factor investing）的主流方法（石川 等, 2020）。在投资实务中，市场信息能够通过因子的形式对最后的投资决策产生贡献。相关的方法中，围绕基金持股数据的因子或异象也被长期关注，如 Wermers et al（2012）。但是，在这种“先因子，后组合”的投资范式中，基金持股数据必须先经由因子构造，而后才能各自独立地参与投资决策；因为因子构造是人为决定的，因子的表达式、所刻画的维度、多个因子的组合方式、因子间信息重合与互补关系，这些问题都带来信息利用上的困境。从不同信息、不同角度刻画的海量异象组成 “因子动物园”（Cochrane, 2011）而深受诟病。“一千个研究员可以构造出一千种因子”，这种人为加工信息的过程势必会丢失基金持股这种结构化数据[②]中大部分原有的信息，并且引发多种因子间相互竞争和因子组合时的维度灾难等问题。

① 通常称“多因子模型”，具有如下形式：$E[R_i^e] = \boldsymbol{\beta}_i'\boldsymbol{\lambda}$.
② 相对于文本、图象和视频这些非结构化数据。

即使主动投资和量化投资都在基金持股数据的利用过程中存在困境和不充分性，过去的研究者和投资者对于基金持股信息的关注和实践依然充分表明了这一信息来源对实际投资过程的贡献。那么，是否有希望提出新的方法，一方面兼顾所有的基金持股信息，另一方面能够绕开人工构造因子的困境，从而产生更优投资效益并且在逻辑上不失可解释性？得益于近十年来自然语言处理领域在非结构化数据分析处理方面的进展，嵌入方法（Embedding Methods）能够高效地从单词排列而成的文本语句中学习单词本身的含义。单词之于文本，即如股票之于基金，嵌入方法或许能在基金持股中提取单个股票的特征（Gabaix et al., 2023）。这种股票特征的向量化表示具有标准化的输出形式、灵活的输出维度、“端到端”的数据处理过程，因而更容易融入到神经网络中，和其他多个来源的信息共同被使用。这一嵌入表示的方案或许可能降低基金持股数据利用过程中的信息损失。以此为启发，本文希望避开人工构造因子的方式，用嵌入表示的方式来解析基金持股数据中的股票特征；进一步，利用特征中隐含的股票关联信息获得稳定的异象因子，从而揭示市场利用基金持股信息过程中的无效性，同时比较人工提取和嵌入表示这两种方案在应用效果上的差异。

1.2 研究意义与创新点

本文立足于基金持股数据，经过嵌入方法获取股票的特征表示，进而刻画股票关联并检验其中的定价信息，同人工构造的股票关联进行比对。上述研究内容对以下多个方面具有意义——

分析基金投资策略和基金持股影响：本文通过揭示基金持股中股票之间的关联关系，能够还原出基金经理捕捉到的股票特征，这将有助于更加深入地理解基金经理的投资偏好、行业配置和风险管理等行为。基金持股中蕴含股票关联，这一发现能够揭示共同基金影响资产市场的机制，验证基金经理在识别股票特征及相互关系时的信息优势，从而更好地回答基金经理投资决策是否有效这一争议性问题。

检验市场有效性：通过揭示股票之间的关联性，检验关联动量因子，可以揭示相互关联的股票之间信息传导的滞后现象。这种信息滞后带来的定价异象能历史上长期产生超额收益，说明市场在分析基金持股信息时具有无效性。除股票关联以外，本文利用基金持股提取的股票向量表示中也包含其他维度的公司特征，或有助于发现更多异象或改进已有因子。

提示嵌入表示在金融中的应用前景：本研究引入嵌入表示的概念，提供了一种全新的分析视角和方法，可以有效地表征基金持股中股票的多重特征以及股票之间复杂的关联关系。这种方法不仅可以帮助投资者发现潜在的资产特征和相互关联，也可以

启发更多的金融分析人员关注嵌入方法，为解决实际投资中的问题提供新的思路和方案。

发现资产关联、控制投资组合风险：投资组合优化时需要关注公司风格暴露和行业特征所带来的风险敞口，高关联度的股票间的股价相关性更高。当前 A 股市场中显式的公司特征存在缺陷，例如中信、申万等行业分类并不足够及时可靠地对上市公司主营业务所处行业进行正确归类，而来源于“端到端”统计学习的股票关联则更加准确与及时。如果基金持股中包含传统风格和行业分类以外的股票关联信息，这将有助于改进风险模型，帮助投资者把握公司股价的联动性，有针对性地进行风险管理。

提示信息披露效率和金融市场监管：本文揭示基金持股中的股票关联信息，有助于揭示基金持仓披露的市场影响。文献表明更高的披露频率将导致更多的窗饰（window dressing）行为（Xin et al., 2024）并且降低被持有公司的公司投资效率（Du et al., 2021），从而带来负面影响；本文则提出基金持股中包含有效的定价信息，提高披露频率具有积极意义。这提示金融监管在制定持仓披露要求时需要权衡两方面效应。

综上所述，研究基金持股中的股票嵌入和股票关联具有多个方面的研究意义和应用价值，涉及资产定价、市场有效性、资产关联刻画、风险控制、信息披露，以及金融市场监管等多个领域。这些方面的研究成果将为证券投资和金融市场的理论与实践提供一定的启示和指导，推动相关领域的进一步发展和创新。

1.3　研究方法与论文结构

1.3.1　研究方法

本研究关注应用层面，主要采用实证方法。首先，进行文献回顾，简要梳理以下三方面内容：基金持股中的定价信息，嵌入表示方法及其在金融投资中的应用现状，股票关联的刻画与应用。在文献的基础上，本文将简单阐释利用基金持股数据刻画股票特征、构建股票关联度量的合理性，进而提出实证部分值得关注的假设。

其次，本文进行 A 股市场共同基金持股数据的整理和清洗，评估样本量与观测频率。在建模时，根据数据的完整性和实际含义，选择合适的嵌入方法进行训练和评估。在训练嵌入模型时，本文对 PCA、Word2Vec 简单神经网络和 BERT 等具有 Attention 机制的复杂神经网络均进行尝试，并最终选择训练复杂程度较低、初步表现最好的 Word2Vec 方法作为后续主要的建模方法。

最后，在模型应用部分，在风格、行业的划分能力与收益相关性的区分能力上对人工模型与嵌入模型的差异进行统计，并且用因子评估的方式对比所提取的股票关联信息的实际定价能力，对研究的假设给出结论。因子评估围绕分组单调性、信息系数（IC）

和多空组合收益展开；最后，本文使用 Fama-MacBeth 回归（Fama et al., 1973, 2020）进行因子检验。

1.3.1 论文结构

本论文按照以下结构进行组织和撰写：第一章为引言，介绍研究背景和研究意义；第二章对相关领域的文献进行回顾，基于当前研究进展对后续研究问题的可行性与理论依据进行简要分析，并在此基础上提出研究假设；第三章进行数据处理与建模，介绍样本选择、数据处理、模型训练的具体细节；第四章进行模型应用，展示模型在捕捉行业关联、风格关联、收益关联等多个任务中的表现，对模型提取的关联动量异象进行检验；第五章总结全文。

第二章　文献回顾、理论分析与研究假设

2.1　文献回顾

2.1.1　基金持股与基金选股能力

有相当一部分文献认为基金具有选股获取超额收益的能力；并且，这一能力能够反映在基金持股中，通过持仓股票的市场表现得到验证。Jiang et al（2014）分析了 1984 年至 2008 年间美国活跃的股票基金，发现去除市场、规模、价值和动量因子后，被基金大幅超配的股票的回报率依然比被低配的股票高出 7%以上。Cremers et al（2009）提出一种新的衡量基金经理主动性的方法，发现主动份额（Active Share）更高的基金具有更优秀和持续更久的绩效。Wermers et al（2012）利用美国市场共同基金的持仓数据，提出广义逆 Alpha（GIA）指标来捕捉基金经理所预测的公司未来盈利情况，并通过该指标的定价能力间接提供了基金经理基本面分析具有价格发现作用的证据。对于中国 A 股市场，李斌 等（2022）检验了 2005-2019 年间中国主动管理基金挖掘异象因子的能力；樊帅 等（2017）发现基金超配个股的行为与公司基本面指标高度相关，并且受基金投资类型和投资风格的影响；刘莎莎 等（2013）考察基金风险调整及随后的基金业绩后发现，基金经理在主动调整风险策略时具有信息资源优势。

“共同基金能够战胜市场（beat the market）”这一命题在实证检验上存在争议（Harvey et al., 2022）。理论推论上，主动资产管理的费后平均表现将不如市场（Sharpe, 1991），因此全体基金的持股行为并不能直接带来超额信息。基金持股的决策机制可能引入与资产特质无关的噪音，例如基金管理人在基金披露时的窗饰（window dressing）行为（Agarwal et al., 2014），即基金经理在持仓披露前临时对基金持股进行调整，以呈现出对市场表现良好的热门股票更高的持仓。此外，共同基金的市场参与也可能加剧定价的无效：Basak et al（2013）通过理论推导得出机构对基准指数成分股的偏好将导致指数成分股价格和价格相关性偏高、波动加剧；Gompers et al（2001）认为大型机构投资者的市场份额扩大能够解释美国市场 1980 到 1996 期间市值风格的切换。投资经理的持仓组合受限于有限的研究注意力，只能同时覆盖少数的行业和主要的股票，体现为重仓股集中（罗军 等, 2024）：普通股票型基金的前十大重仓股占比中位数在 50%附近，重仓持股占比并未随着持仓股票数的上升而大幅下降。

2.1.2　嵌入方法及其在金融中的应用

嵌入方法（Embedding methods）是机器学习中用于特征选择或特征降维的技术。通过将原始特征空间嵌入到一个更低维度的空间中，该技术能够保留最相关或最具信

息量的特征。嵌入方法常用于处理高维数据，可以帮助提高模型的泛化能力，减少过拟合，并加速训练和预测过程。在自然语言处理中，文本嵌入使用数值形式表示文本数据中单词或短语的含义。

最传统的文本嵌入方法是潜在语义分析（LSA）（Dumais et al., 1988），该方法对“文档-词项”矩阵进行奇异值分解（SVD），将文本数据嵌入到一个语义空间中，用于词义理解、文本分类和信息检索等任务。随着深度学习技术的发展，词嵌入（word embeddings）成为自然语言处理领域的热门方法，Word2Vec（Mikolov et al., 2013a, 2013b）、GloVe（Pennington et al., 2014）等方法开始使用局部的非线性信息。例如在 Word2Vec 中，用浅层神经网络通过文档中某段局部窗口内的其他单词来预测目标位置的中心单词（或者用中心单词来预测附近局部窗口内的上下文）而非整个文档。尽管被广泛使用，Word2Vec 方法也存在局限，包括 ① 无法有效利用上下文窗口内单词的先后顺序，以及 ② 无法考虑单词的多重含义，因而在语义提取的精确性上依旧存在提升空间。近年来，上下文嵌入（contextual embeddings）在广泛的自然语言处理任务上取得了突破性性能，如 ELMo 和 BERT（Liu et al., 2020）。相较于 Word2Vec 中单词和嵌入表示的一一映射关系，上下文嵌入为每个单词分配一个基于其上下文的表示，从而捕捉了单词在不同上下文中的用法，并能跨越相距更远的上下文。

嵌入方法在金融中的应用，主要包含资产定价、资产相似特征的刻画、资产价格相关性的风险控制等方面。对于资产定价，嵌入方法能够引入传统非线性指标之外的其他知识，从而加强传统线性指标。例如，Li et al（2019）用基金-股票二分图和 Skip-Gram 算法进行股票嵌入，抽象化地表示股票特性，从而根据不同股票特性对 7 类交易因子生成不同缩放权重，提升了技术因子的选股能力；Chen et al（2019）借助矩阵分解（Matrix Factorization）拆解出基金持股信息中基金经理偏好和股票内在属性的嵌入表示，并综合历史业绩最好的股票计算市场属性，将未来市场属性（LSTM 时序预测值）和股票内在属性的相关性作为有效的定价特征，同 Alpha 101 因子（Kakushadze Z, 2016）共同使用；叶尔乐（2023）通过类似的矩阵分解方法从基金持仓网络中提取股票属性，而后直接将这些属性同的量价因子拼接后输入到多层感知机（MLP），同样取得了量价因子效果的提升。

更早期的研究更多从新闻文本中获得资产的低维嵌入，评估新闻文本带来的套利机会。例如 Chen et al（2022）利用 16 个国际股票市场和 13 种不同语言的新闻文章数据，为新闻引发的收益可预测性提供证据；Ang et al（2022）利用来自维基百科和维基数据的知识图谱关系提取公司嵌入表示，在收益预测和组合管理中取得更优效果；Du et al（2020）在用神经网络同时编码（encode）新闻文本和历史价格时提出，应该首先获得股票的嵌入表示，后续才便于将训练结果应用于价格预测以外的其他用途。

在资产相似特征和相关性刻画的方面，Dolphin et al（2023a）利用股票历史收益序列学习股票嵌入表示，并讨论了投资组合优化、资产配置、风险管理等多种应用场景。相关作者在后续研究中继续利用历史股票收益时间序列数据，并结合财经新闻提取更加客观的嵌入表示，在行业分类的任务上比使用更传统表示方法的几个基准模型取得了显著的性能改进（Dolphin et al., 2023b）。Gabaix et al（2023）提出用基金持股数据同时学习股票和基金的嵌入表示，发现股票间与基金间的替代关系。Takayanagi et al（2022）提出了一种股票嵌入增强模型（SETN），分别用 Transformer 模型和图神经网络提取文本信息和网络信息，并且基于所获得的股票嵌入特征提取半导体、5G 等主题，构建相关的主题基金。

主流的金融学研究者近年来更关注高维面板数据中低维资产定价模型的提取（Chen et al., 2024; Kelly et al., 2023; Cooper et al., 2021; Kozak et al., 2020; Lettau et al., 2020; Kelly et al., 2015），同样涉及嵌入方法的应用。这些研究更多是在结构化数据的处理过程中应用嵌入方法，与本文关注的词嵌入方法联系不强。嵌入方法在金融中的应用不仅局限于股票和基金的特征向量，还包括金融事件的向量表示（Ding et al., 2015）等。

2.1.3　资产关联的刻画与应用

随着嵌入模型的发展，更多非结构化的图数据、文本数据当中潜在的公开市场信息受到关注。上市公司间的关联信息可能来自上下游产业链和供应链（Wu et al., 2023; Xu et al., 2022; Berger et al., 2020）、共同投资者和卖方分析师的网络（Hameed et al., 2015; Antón et al., 2014; Koch et al., 2016）、投资者社交网络关系（Li et al., 2022; Ivković et al., 2007）、投资者情绪（Kumar et al., 2013）和有限注意力的同步转移（Drake et al., 2017）、新闻舆情文本中的共同出现（Xu et al., 2022; Chen et al., 2018），以及公司风格特征（Wahal et al., 2013; Barberis et al., 2003）。上述文献对多个来源的公司特征进行把握，对上市公司间的价格联动、风险传导等现象进行了有效的捕捉和解释。但遗憾的是，本文在写作时并未发现利用嵌入表示发现基金持股数据中股票关联信息的先例，而聚焦于 A 股市场中基金持股信息定价能力的文献也较为匮乏。

对于基金持股，过去的研究者已经对持股网络、基金信息网络以及在这些网络中股票所处位置的结构特征进行了较多讨论。在投资实务中，徐寅（2020）对基金持股因子提供了一个较为完整的梳理。此外，有卖方研究人员利用基金共同持仓刻画股票关联网络，并据此提出关联网络牵引因子（魏建榕, 2021a）。该作者将基金持股能够反映股票关联关系的理由总结为：① 基金管理人认知——“基金持股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”；② 股东协同行为——“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”。类似的人工提取方法对北向资金等其他机构投资者的持仓同样适用（魏建榕, 2021b）。在后文的模型对比部分，本研究将以这一类从基金持股中提取股票关联网络的方式作为对比基准，称作“人工模型”。

2.2 理论分析

本文重点关注基金持股数据中的公开市场信息和这部分信息在实际投资中的定价能力。特别地，本文提出用嵌入方法提取基金持股中的资产关联，并基于关联网络中的股价联动现象发现市场异象。上一部分的文献回顾带来以下启发：① 主动基金具有选股能力，并且能够经由基金持股进行检验；② 基金持股信息中的股票特征类似于文本序列中单词的含义，可以经由嵌入方法学习；③ 嵌入方法表示出的股票向量信息，在行业分类、风格归因、收益率和风险预测上具有应用可能。下边将对以上启发背后的理论基础进行更深入的讨论。

2.2.1 基金持股中蕴含具有定价能力的股票特征

为什么基金持股信息中蕴含中上市公司股票的特征？主动基金经理能够通过各式分析框架，或从公司基本面出发，或结合行业和赛道格局，捕获公司特征，并通过分配资金的方式对各自看好的公司特征进行持仓：公司特征贯穿人类基金经理选择股票的思维过程。例如，巴菲特的投资业绩可以通过市场（Beta）和质量（Quality）因子解释（Frazzini et al., 2018）。机构投资者对资产的需求可以被看作是对分散的资产特征的需求，因此基金持股中隐含股票特征信息。

作为市场参与的理性主体，机构投资者对于各维度特征完全一致的股票将体现出完全一致的需求量（尽管实际投资中只可能穷举有限的最为重要的维度，而对于股票代码、股票名称等行为偏误带来的不那么重要的特征只能在理想情况下达成一致）。而股票非系统性风险（特质波动性）的存在又促使机构投资者有动机分散化地持仓，即假设两只股票的特征完全符合基金经理期望持有的特征分布，如果不考虑市场摩擦，这一资产持有人将等权重地持有这两只资产，而非持有任意其一。上述简单情况的推广形式将促使机构投资者在有限注意力等约束下尽可能分散化地持有具有相似特征的资产，体现为基金持股中类似特征的股票具有类似的占比。因此，基金持股中的持仓金额相近的股票，在基金经理视角下存在一定关联。

进一步，如果基金经理对股票特征的挖掘足够有效，能够领先于广为人知的定价因子（股票特征），那么机构持仓中潜在的公司风格聚类将包含现有定价因子不能够完全捕捉的定价能力，这也从另一个方面揭示出主动基金管理人的专业性和信息优势，是基金管理业务产生费后收益的基础。

在另一方面，即使抛开“主动基金具有价格发现能力”这一假设，从市场供需的角度出发，投资机构作为资产市场需求方，对不同资产相似的持股偏好也依然会引发相关的价格变动。因此，即使基金持股并不能捕捉任何资产本身的特征，也依然反映资产在“基金投资者需求”这一维度的关联，即基金持股本身就构成一类影响价格的公司特征。综上所述，基金管理人认知和股东持股需求这两个因素共同决定了用机构持仓数据来度量资产特征这种做法的合理性。

2.2.2　基金持股中的资产特征能被人工或嵌入的方法获取

对于投资者持仓信息中的公司特征，过去的投资从业人员已经提出过多种人工构造的形式。例如被共同持有的基金数量，股东中重要机构投资者持股比例，基金主动持股份额及季度变化等等。上述方式从“点对点”的“基金-股票”关系出发，人工设计一些具有直观意义的特征序列，再通过多种经验性的调整手段配合线性或非线性的组合方式来提纯与未来收益序列具有更高相关性的定价因子。上述方法通过因子检验，证明人工方法确实是提取基金持股中有效信息的滤波器。因此下述分析重点关注嵌入方法之所以能获得股票特征的理论依据。

为什么嵌入学习能够捕捉基金持股中的资产特征？特别地，本文重点使用的 Word2Vec 方法为什么能从基金持股金额排序的资产排列中获得资产的特征向量？类比于自然语言处理中的词向量，投资者持仓数据中的公司股票可以被视作“词语”， 而投资者的持仓情况则相当于“上下文”；Word2Vec 通过学习词语在上下文中的语义关系，将每个词语映射为向量表示；因此，Word2Vec 可以将不同公司的持仓关系转化为反映基金持股信息的向量表示，用于捕捉公司特征和公司间的潜在关联。

从原理上，Word2Vec 使用的连续词袋（CBOW）或跳字（Skip-gram）模型学习上下文中隐含的语义关系（见本文 3.2.2 部分）；而在投资者持仓数据中，如果两个公司经常同时出现在同一组投资者的持仓中，又或者在持仓金额排序中位置相邻，又或者总是和同一组股票具有相似的持仓金额，这些现象都意味着它们之间存在某种关联或相似性；通过 Word2Vec，这种资产和资产的关联关系可以被有效地转化为向量空间中的位置关系，使得相似的公司在向量空间中靠近、不相似的公司远离。正是因为上文提到的投资者对股票特征多个维度的分解和考虑，包括行业趋势、市场情绪和公司基本面等，股票嵌入表示相似就意味着公司特征相似。Word2Vec 嵌入方法能够从投资者持仓中获得有效的公司特征向量，学习隐含语义的关系来捕捉投资者偏好，这和机构投资者寻找相互替代的股票时的思维逻辑一致。

2.2.3 所提取的资产特征具有可解释性与应用价值

嵌入模型提取的股票特征向量如果有效，它和公司风格暴露、公司市场表现等“显性”特征的关系是什么？在横截面上，嵌入模型提取的股票特征向量可以被视为隐含特征，它们捕捉了投资者持仓数据中的潜在关联和相似性；与此同时，公司的风格暴露（如行业、市值、估值等）和市场表现（如收益率、波动性等）属于显性特征，是投资者通常关注和考量的公司属性；基金决策时充分考虑显性特征，导致嵌入向量与显性特征关联，可以被显性的公司特征解释；并且，由于显性特征来源于信噪比较低的财务数据和市场数据，嵌入向量所呈现的公司特质可能更纯净、更本质。①

在时间序列上，资产特征的定价能力更多表现为若干风险因子的收益率在历史时期出现波动起伏，而基金持股在时间维度上的变化过程可以体现主动基金对于风格、行业轮动的择时能力。被基金最优先选择的风格和行业具有更优秀的收益，因此在定价时按基金持股来选择股票，或能领先于风格因子收益的变动，产生时序预测能力。

无论是截面选股还是风格择时，基金持股中的嵌入表示都具有信息含量。且不论模型学习到的向量表示本身如何被解释，作为优化过程中的关键损失函数，嵌入向量空间距离所反映的股票相似度或关联程度是体现上述信息含量最直观的指标。如果嵌入模型捕捉到了股票之间的替代关系，那么最终的股票特征将能体现基金持股中隐含的资产特质，具有更高相似度的股票在未来收益和风险上也应更为相似。

2.3 研究假设

在以上理论分析的基础上，本文后续实证部分将重点关注以下研究假设——

假设一：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。

假设二：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。

假设三：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。

假设四：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。

假设五：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。

① 类似的例子包括对历史收益率进行 PCA 分解所得到的统计因子在历史数据上能够比 Barra 等主流风格因子解释更多的收益率方差。

第三章　数据处理与建模

3.1　样本选择和数据处理

3.1.1　样本筛选的规则

本文使用的数据包含中国 A 股市场的基金持股数据，财务指标以及日频的交易行情。市场行情和风格因子来源于 Wind 和 iFinD，基金持股来自 Tushare 和巨潮资讯。

研究的样本为全市场沪深 A 股和披露持仓情况的共同基金。对每个报告期的基金和股票进行如下筛选：① 只考虑重点关注 A 股市场的基金管理人，去除持有 A 股总金额占基金净值比例不足 20%的基金；② 鉴于市值较小的股票中存在“壳污染”以及投资者构成的异质性，去除期末市值最小的 20%的股票，此外，也去除期末停牌、ST、*ST、上市不满三个月的股票；③ 考虑到机构投资者的影响力以及被持有股票的重要程度，去除每个披露期被少于 K 个基金持有的股票，并去除持有上述范围的股票不足 K 只的基金（K 是弱敏感性参数，默认取 K=10）。本文认为，即使是债券型基金，如果持有 A 股的资金占比超过 20% 并且持有股票数量超过 K，也将提供有价值的信息，予以保留。这是因为，基金经理决定持仓权重时需要立足于股票的风险收益特质，即使债券型、混合型和股票型三种基金配置股票的目的不同。

3.1.2　样本时期和观测频率

本文关注的样本范围是 2013 年半年报到 2022 年年报，基金持股数据的频率为半年度，共 20 个报告期。[①]本文在每次报告期后均会用最新的基金持股状况重新提取股票特征，更新股票关联指标。

除了每半年度采样的基金持股数据外，实证部分用到的公司股价、公司上市状态、公司行业分类、公司市值等风格指标，均使用“日频”数据，在每个交易日后取收盘时能够获得的最新数值。在投资组合评估时，也采取“日频”的方式调整组合持仓、评估组合表现。

3.1.3　样本数量

经过本文 3.1.1 部分的样本筛选，后续研究时各个报告期的样本数量如下图 3.1 所示。图左轴表示样本基金数，用浅灰色条形图表示，全市场存在的基金数用深灰色条形图表示，样本筛选后保留的基金数用实线表示。在前后各个历史时期，我国市场的基金数量经历了较大的增长，从原来的不足 1000 只先后经历 2015-2017 和 2019-2022 两个

① 基金持仓的一季报和三季报一般只披露前十大重仓股，因此本研究降采样为半年度，忽略一季报和三季报的持仓披露；类似的研究中也存在每季度进行持仓补全的做法，信息更加及时。

快速增长的阶段，达到 6000 只以上，并且直到样本期末依然呈现较高的增速；深灰色所表示的经过样本筛选后的样本基金数则表现出更加均匀的增长，且未出现两个快速增长时期之间的间隔，这说明样本筛选或可有效去除市场快速扩张过程中较为次要的基金。

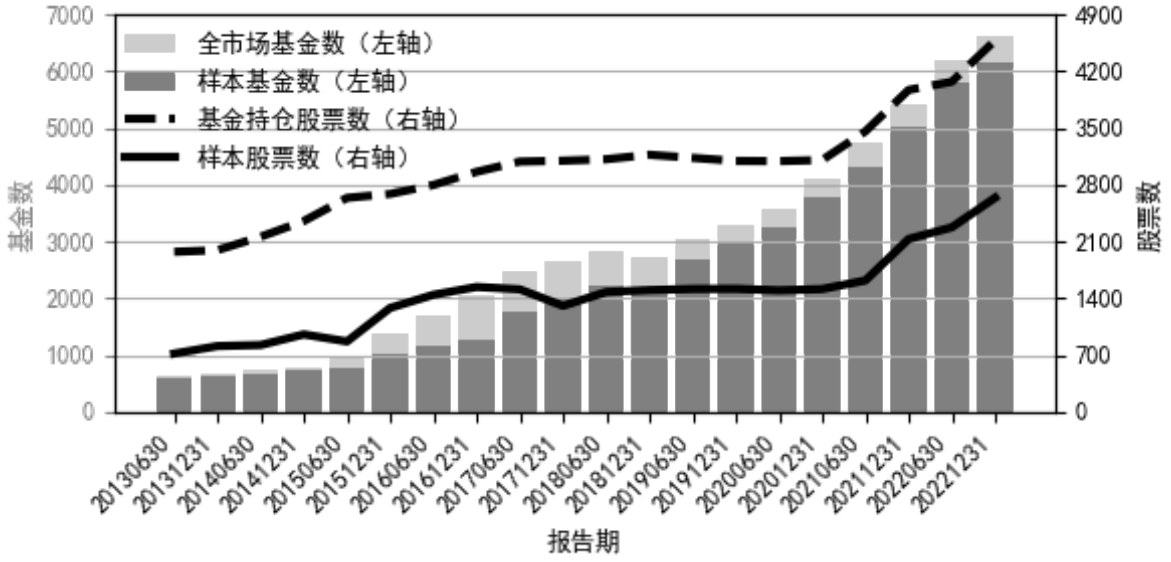


图 3.1　各报告期样本基金、股票数量

图 3.1 中右轴表示股票数量，虚线为全市场中至少被一只基金持股的股票，实线为样本筛选后的股票数量。样本股票数量为基金持股股票数量的一半左右，这和“主动基金管理人具有有限注意力”以及“持仓集中于少数大市值的重仓股”这些直觉相符。样本期初的股票数量在 700 只左右，经历 2016 年的一波增长后长期维持在 1400 附近，并且从 2021 年开始伴随着基金数量的井喷同步呈现增长态势，最新报告期达到 2800 只。综观样本基金数和样本股票数，可以认为基金数量和股票数量的增长趋势基本一致，基金数量在 2019 年后的增长领先于样本股票数的增长。

3.2　模型建立

3.2.1　人工模型

本研究利用基金持股数据，首先建立人工模型，直接刻画股票之间的关联度。在每个报告期末[①]，取任意样本基金 i 所持股票 s，用基金 i 对股票 s 的持仓市值 H_s^i 与股票 s 最近 20 日成交额均值 AMT_s 的比值来衡量基金对个股的影响力强弱，定义影响力 I_s^i 为（式 3.1）：

① 每个报告期单独训练模型，故本节公式中省略“报告期”下标。

$$I_s^i = \frac{H_s^i}{AMT_s} \tag{3.1}$$

取同一基金 i 对任意两只持仓股票 a, b 影响力的最小值作为该基金中两只股票的关联强度 $J_{a,b}^i$（式 3.2）：

$$J_{a,b}^i = min\left(I_a^i, I_b^i\right) \tag{3.2}$$

将任意两只样本股票 a, b 在所有基金中的关联强度加总，得到两只股票的关联度 $K_{a,b}$（式 3.3）：

$$K_{a,b} = \sum_i J_{a,b}^i \tag{3.3}$$

$K_{a,b}$ 越高，表明股票 a 和 b 之间的关联度越大。实际得到的关联度指标在截面上呈现对数化分布，这和股票市值在截面上的对数分布相符。最终我们将股票间的关联度指标取对数，并用 min-max 规范化方法标准化到实数域 $[0, 1]$ 之间（式 3.4）。下文将这一指标称作“人工关联度”。

$$K^* = \frac{ln\, K - min\, ln\, K}{max\, ln\, K - min\, ln\, K} \tag{3.4}$$

上述人工模型的构造方式直接借鉴于相关文献在构造“关联网络牵引因子”时的做法（魏建榕, 2021a）。该方法虽然具有较高的复杂度 $O(N^3)$，但能够较充分地利用基金持股数据：股票被持有的金额更高，或股票被更多的基金持有，或股票在多个基金中共现次数更多，则股票之间的关联度更大。由于额外引入了“近 20 日成交额”这一额外信息，这一指标对股票的流动性水平相对不敏感，成交越活跃的股票，每半年度披露的基金持股金额大小对股票价格的影响能力理应越弱，所反映的公司特征的可信度也应更低。

3.2.2　嵌入模型

本文的研究重点在于用嵌入模型刻画基金持股中的股票关联特征。对标人工模型，本文倾向于只利用更少和更纯净的基金持股信息来尝试捕捉效果更好的公司关联。具体地，本文只关注每个样本基金持股中股票按持仓金额排序后的信息，忽略更细粒度的实际持仓金额。在每个报告期，基金 i 的持仓首先被处理为语句 Λ_i（式 3.5）：

$$\Lambda_i = \left[s_{i1}, s_{i2}, \ldots, s_{im_i}\right] \tag{3.5}$$

s_{ir} 表示基金 i 持仓中的股票，按金额大小排序后依次排在第 $r = 1,2,\ldots,m_i$ 位（m_i 为基金 i 所持股票总数）。每个训练窗口期内，所有样本基金持股分别作为一个独立的语句输入到 Word2Vec 模型中。20 个报告期内所有独立句子长度呈现对数分布的特征（图 3.2），由于样本筛选的规则，在持仓股票数量为 10 处被截断。对数变换前，基金持股数量的均值为 50.5，中位值为 53，最小值和最大值分别为 10 和 1552；在 Word2Vec 中，由于直接关注上下文窗口期内的单词，因此单个句子的长度过长并不会影响模型的训练。

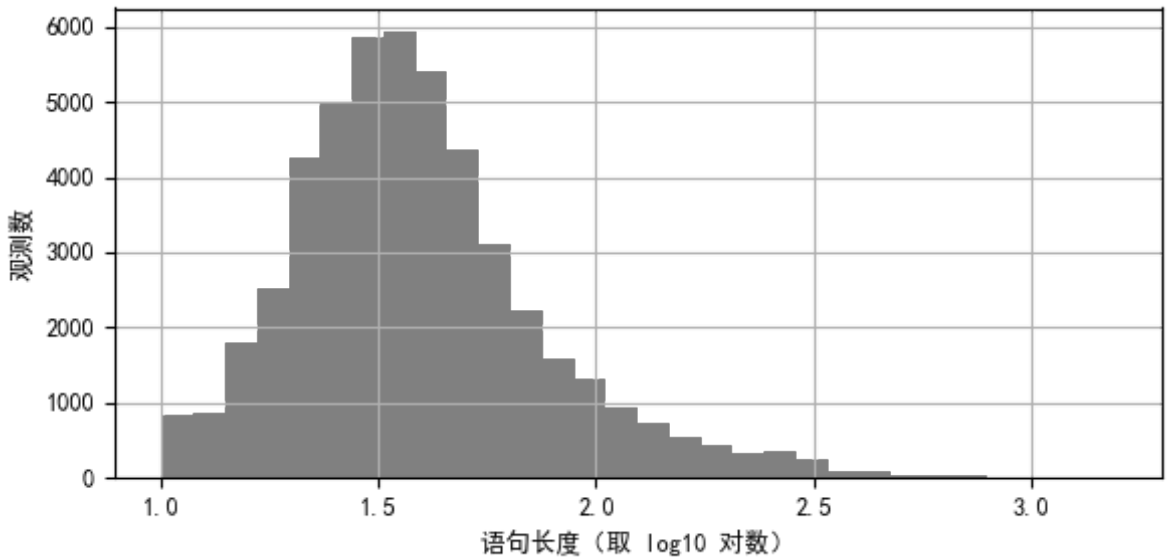


图 3.2　Word2Vec 模型输入的语句长度分布

每个基金经理根据对资产的需求从高到低排列所持股票的金额，尽管少数头部重仓股可能占有绝大部分持仓，但是在局部上下文窗口内（K 个单词，K 取 5 或 10），相邻的极少数股票之间持仓金额的差异可以忽略不计。模型训练中的学习任务是：① 给定在持仓排序 Λ_i 中相邻的 K 个股票，预测居于中心位置的股票代码（CBOW 版本）；或者 ② 给定中心位置的股票，同时预测在持仓排序 Λ_i 中相邻 K 个其他股票的代码（skip-gram 版本）。不同于已有的文献（Gabaix et al., 2023），本文采用 skip-gram 版本的 Word2Vec 算法，原因如下：① skip-gram 通过最大化预测上下文单词的概率来学习词向量，每个目标单词都对应多个预测任务，因此能够处理罕见词，这符合基金持股中股票出现频率的对数分布，即大部分股票出现次数较少；② skip-gram 相比 CBOW 对数据集规模的要求更高，但是相比于自然语言中单词所具有的极为丰富的含义，股票“含义”的稀疏性在直觉上更低，因此对训练集规模的要求并不如自然语言处理中苛刻。综合以上原因，本文选择 skip-gram 版本的 Word2Vec 模型。

具体而言，给定训练样本中的每对排在位次 t 和 $t+j$ 的股票 s_t, s_{t+j}（下文对持仓排序 Λ_i 中表示基金的下标 i 进行省略），希望最大化（式 3.6）：

$$\prod_{r=1}^{m} \prod_{-K\le j\le K, j\neq 0} Pr\left(s_{r+j}|s_r\right) \tag{3.6}$$

其中 $\Pr\left(s_{r+j}|s_r\right)$ 为“中心词”股票 s_t 附近出现关联股票 s_{t+j} 的条件概率，用 softmax 函数定义，即（式 3.7）：

$$Pr\left(s_{r+j}|s_r\right) = \frac{exp\left(v_{s_{r+j}}^T v_{s_r}\right)}{\sum_{s'\in S} exp\left(v_{s'}^T v_{s_r}\right)} \tag{3.7}$$

其中 v_{s_r} 为排在位置 r 的股票 s_r 所对应的词向量（单位向量）。向量余弦距离越小的股票，在同一个窗口期内出现的概率越高。向量维度是一个超参数，本文默认取 30 维。[①]最终，skip-gram 算法优化的损失函数为所有训练样本负对数似然之和，即（式 3.8）：

$$loss = -\sum_{i} \sum_{r=1}^{m_i} \sum_{-K\le j\le K, j\neq 0} ln\, Pr\left(s_{i,r+j}|s_{i,r}\right) \tag{3.8}$$

使用基于随机梯度下降（Stochastic Gradient Descent, SGD）的优化算法迭代股票所对应的词向量 v_s，最小化损失函数（式 3.8），得到词向量模型。[②]取模型中的网络节点，即可以获得股票向量 v_s，以及基于向量余弦距离计算的股票之间的相似度（式 3.9）：

$$sim_{a,b} = \frac{v_a \cdot v_b}{|v_a| \cdot |v_b|} = v_a \cdot v_b \tag{3.9}$$

每半年度进行滚动训练，每次训练时回看过去 1.5 年（3 个报告期）的基金持股披露，由此获得随时间变化的 20 组股票向量以及股票向量之间的相似度，作为每半年度更新的股票关联指标（下文中称为“W2V 相似度”）。

3.2.3　其他模型

本文在研究过程中也考虑了其他更加复杂的语言模型，例如 BERT 模型。[③]根据 Gabaix et al（2023）的工作论文，这些引入上下文嵌入的模型有助于解决 Word2Vec 的

① 考虑到 Barra CNE5 风格因子维度是 10，股票收益率 PCA 分解的前 20 个主成分能够解释绝大部分方差，这一超参数取值是相对合理的。
② 更多 Word2Vec 实现细节见 https://radimrehurek.com/gensim/models/word2vec.html 说明文档。
③ 参考 https://github.com/codertimo/BERT-pytorch，训练时损失函数只使用“masked language model”。

两大局限：① 采取自注意力机制，因此能够捕捉上下文距离更远的资产间的关系[①]；② 模型能够输出多重隐藏状态，而非“股票-向量表示”间的一一映射关系，这有利于捕捉同一股票不同维度的特征[②]。

但是，对于 BERT 等大型模型的训练，半年度披露的基金持股数据可能达不到样本量的要求。本文尝试用 10 年共 20 期基金持股数据，在实验环境的硬件限制下训练了网络尽可能简单的模型：4 层编码器（encoder），2 个注意力头（attention head），词向量维度 30，对于 3813 种样本股票需要学习 277505 个参数。在训练过程中模型较早陷入瓶颈。从嵌入表示在行业聚类和因子检验（同本文 4.2 和 4.3 部分）的表现来看，模型的输出近似于随机输出，说明实验的参数选择下 BERT 模型并不能有效地提取基金持股数据的信息。

① 例如，擅长投资消费行业的基金除了倾向于超配大市值的贵州茅台（600519.SH）、格力电器（000651.SZ）外，也会在小市值股票中配置更多食品饮料、家电、纺织服装行业的优秀股票，如今世缘（603369.SH），三全食品（002216.SZ）等，但是在聚焦于有限窗口期的 Word2Vec 中，因为市值差异，这两部分股票在“语句”中的上下文距离太远，无法被有效表达。

② 例如，在专注于消费行业的基金超配贵州茅台等股票是考虑其在消费赛道内的特质，而专注于其他赛道的基金之所以配置贵州茅台则更多出于跟踪基准指数和大市值 beta 的考虑，即考虑的更多是股票的风格特征。

第四章　模型应用

4.1　刻画资产关联度

4.1.1　人工模型

本文 3.2.1 部分建立了人工关联度指标，该指标共 20 期，每期观测数为共现股票配对数量。报告期 20221231、20181231 和 20141231 的关联度指标的分布如下图 4.1 所示。据图，对数变换后的关联度指标呈现左偏分布。这表明在关联度低于均值或中位值时，数值跨度范围更大，股票间的差异性更明显。此外，随着时间推移，基金持股数量增加，左偏分布的特性随时间变化有所减弱，说明历年来基金持股中体现出的关联度更分散。

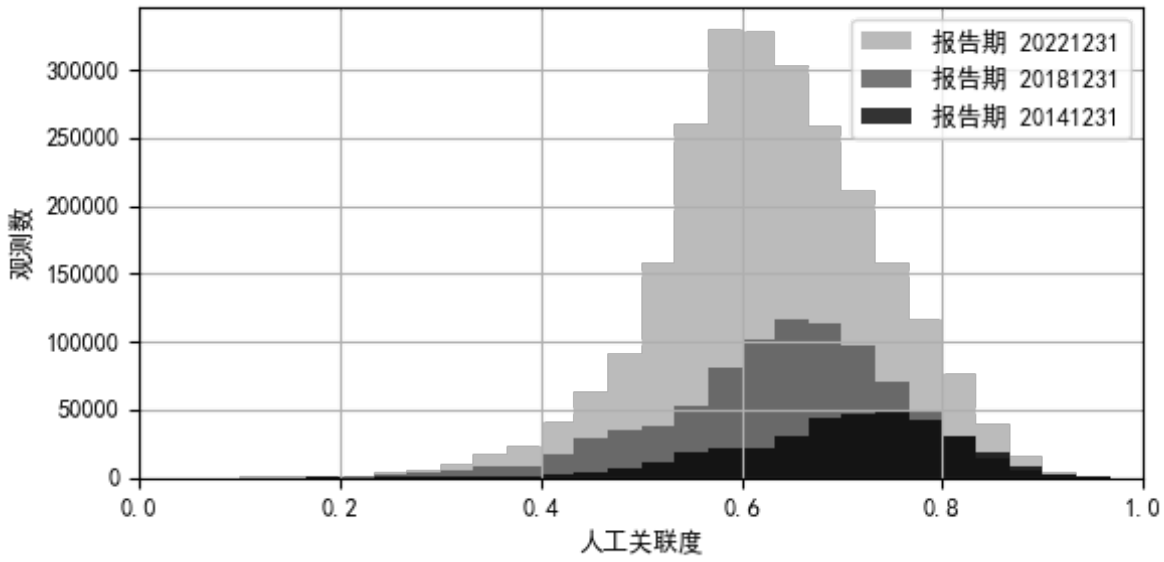


图 4.1　人工关联度指标分布直方图

表 4.1 呈现了各个报告期人工关联度指标的若干统计量。从均值、中位值等统计量的变化规律来看，人工关联度在历年间呈现逐渐下降的趋势，这可能和样本股票数量持续增加有关。关联度指标的标准差维持在 0.12 附近，在历年间变化稳定。从 25%分位数和 75%四分位数来看，人工关联度的分布集中在 0.5~0.8 区间内，样本的数据分布较集中。

表 4.1　人工关联度描述性统计

报告期	观测数	均值	标准差	最小值	25%分位数	中位数	75%分位数	最大值
20130630	213446	0.72	0.12	0.00	0.65	0.73	0.80	1.00

续表 4.1　人工关联度描述性统计

报告期	观测数	均值	标准差	最小值	25%分位数	中位数	75%分位数	最大值
20131231	259778	0.71	0.12	0.00	0.63	0.73	0.80	1.00
20140630	257392	0.69	0.12	0.00	0.61	0.71	0.78	1.00
20141231	366726	0.70	0.11	0.00	0.63	0.71	0.78	1.00
20150630	284659	0.69	0.12	0.00	0.63	0.71	0.78	1.00
20151231	574240	0.68	0.13	0.00	0.61	0.70	0.77	1.00
20160630	762979	0.66	0.13	0.00	0.59	0.68	0.75	1.00
20161231	814492	0.69	0.11	0.00	0.62	0.70	0.77	1.00
20170630	855689	0.65	0.12	0.00	0.58	0.66	0.74	1.00
20171231	625813	0.66	0.12	0.00	0.60	0.68	0.74	1.00
20180630	829651	0.64	0.12	0.00	0.56	0.65	0.72	1.00
20181231	882485	0.64	0.11	0.00	0.58	0.65	0.72	1.00
20190630	774279	0.62	0.12	0.00	0.55	0.63	0.71	1.00
20191231	783090	0.62	0.12	0.00	0.55	0.63	0.70	1.00
20200630	798273	0.62	0.13	0.00	0.54	0.63	0.70	1.00
20201231	772548	0.64	0.11	0.00	0.57	0.64	0.71	1.00
20210630	811328	0.61	0.13	0.00	0.53	0.61	0.70	1.00
20211231	1468973	0.60	0.13	0.00	0.51	0.60	0.69	1.00
20220630	1654530	0.60	0.12	0.00	0.52	0.61	0.69	1.00
20221231	2520885	0.63	0.11	0.00	0.56	0.63	0.70	1.00
全历史	16311256	0.64	0.12	0.00	0.56	0.64	0.73	1.00

表 4.2 至表 4.4 分别呈现了报告期为 20221231 时贵州茅台（600519.SH）、宁德时代（300750.SZ）和科大讯飞（002230.SZ）最相似的前 10 只股票。这些过去几年热度较高的股票在人工关联度指标下的相似股票确实符合人们的认知，其最相似的前 10 大股票中不乏相同行业的高替代性资产。此外，值得关注的是表 4.2 中与贵州茅台关联的医药股，表 4.4 中与科大讯飞关联的属于“消费者服务”行业的两个酒店股，均体现了跨行业的公司关联。

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

续表 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）
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	1955.5

表 4.3　宁德时代（300750.SZ）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	300750.SZ	宁德时代	-	电力设备及新能源	7800.4
1	300014.SZ	亿纬锂能	0.9410	电力设备及新能源	1614.3
2	600519.SH	贵州茅台	0.9347	食品饮料	21694.5
3	002812.SZ	恩捷股份	0.9312	基础化工	974.2
4	002594.SZ	比亚迪	0.9299	汽车	2993.1
5	300274.SZ	阳光电源	0.9228	电力设备及新能源	1262.1
6	002475.SZ	立讯精密	0.9222	电子	2250.7
7	603799.SH	华友钴业	0.9219	有色金属	878.9
8	603259.SH	药明康德	0.9179	医药	2072.5
9	002459.SZ	晶澳科技	0.9154	电力设备及新能源	1408.2
10	002049.SZ	紫光国微	0.9121	电子	1120.0

表 4.4　科大讯飞（002230.SZ）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	002230.SZ	科大讯飞	-	计算机	695.6
1	002410.SZ	广联达	0.8620	计算机	593.9
2	002180.SZ	纳思达	0.8613	电子	626.0
3	603345.SH	安井食品	0.8589	食品饮料	471.7
4	300408.SZ	三环集团	0.8580	电子	568.3
5	300451.SZ	创业慧康	0.8578	计算机	102.6

续表 4.4　科大讯飞（002230.SZ）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）
6	000063.SZ	中兴通讯	0.8577	通信	1029.2
7	600862.SH	中航高科	0.8562	国防军工	310.5
8	600754.SH	锦江酒店	0.8554	消费者服务	533.3
9	300207.SZ	欣旺达	0.8550	电子	362.6
10	600258.SH	首旅酒店	0.8548	消费者服务	265.5

再进一步查看与爱尔眼科（300015.SZ）和锦江酒店（600754.SH）关联度最高的股票（见下表 4.5 和表 4.6），则几乎全部是同行业或产业链上下游的公司，不再能找到贵州茅台（表 4.2）和科大讯飞（表 4.4）。这或许表明，主动基金在重仓配置热门股票时，更多考虑与基准指数对齐；而在持有其他非热门的中小市值股票时，更容易同时持有同行业同赛道的类似股票。

表 4.5　爱尔眼科（300015.SZ）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	300015.SZ	爱尔眼科	-	医药	1803.0
1	603259.SH	药明康德	0.9826	医药	2072.5
2	603456.SH	九洲药业	0.9793	医药	352.7
3	300759.SZ	康龙化成	0.9790	医药	640.5
4	300347.SZ	泰格医药	0.9788	医药	592.9
5	600085.SH	同仁堂	0.9709	医药	612.8
6	688050.SH	爱博医疗	0.9695	医药	160.7
7	300760.SZ	迈瑞医疗	0.9694	医药	3830.9
8	000963.SZ	华东医药	0.9655	医药	818.9
9	605369.SH	拱东医疗	0.9598	医药	29.8
10	300357.SZ	我武生物	0.9588	医药	259.7

表 4.6　锦江酒店（600754.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	600754.SH	锦江酒店	-	消费者服务	533.3
1	600258.SH	首旅酒店	0.9667	消费者服务	265.5
2	600600.SH	青岛啤酒	0.9474	食品饮料	752.8
3	000596.SZ	古井贡酒	0.9361	食品饮料	1090.6
4	603345.SH	安井食品	0.9353	食品饮料	471.7

续表 4.6　锦江酒店（600754.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）
5	603885.SH	吉祥航空	0.9314	交通运输	318.1
6	600690.SH	海尔智家	0.9269	家电	1543.1
7	603899.SH	晨光股份	0.9269	轻工制造	507.9
8	002180.SZ	纳思达	0.9263	电子	626.0
9	603868.SH	飞科电器	0.9243	家电	293.3
10	688389.SH	普门科技	0.9235	医药	75.3

4.1.2　嵌入模型

在本文的 3.2.2 部分建立了 W2V 相似度指标，该指标共 18 期，最早从报告期 20140630 开始，每半年滚动训练，单次训练回看期为过去一年半（3 个报告期）。[①]这种滚动训练方式一方面能为 Word2Vec 模型训练提供足够多的“语句”样本，另一方面更加关注最新的基金持股以捕捉股票特征的最新变化。

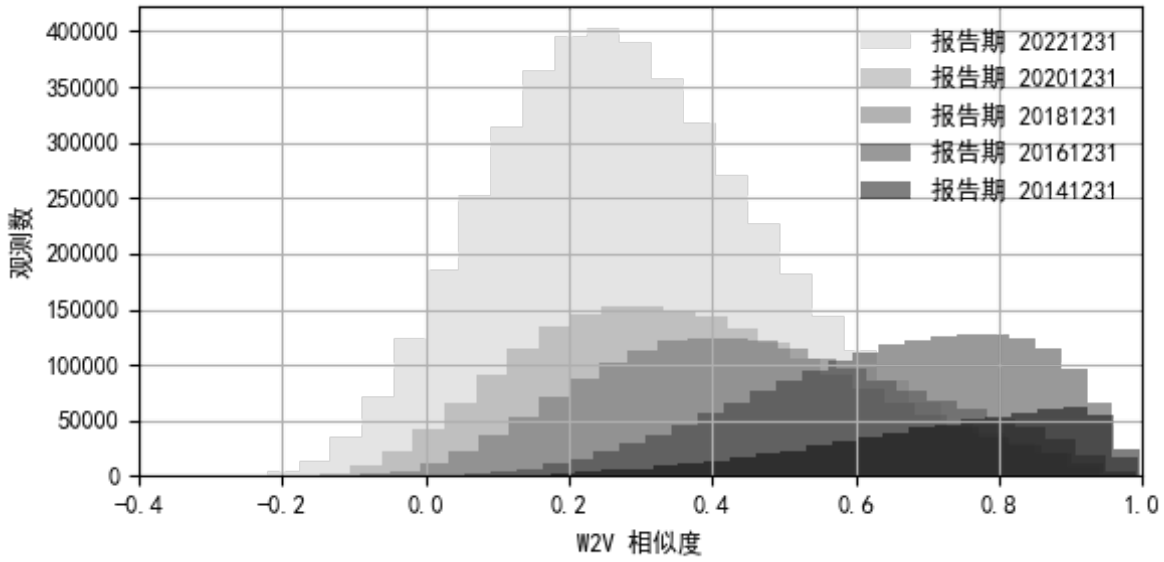


图 4.2　W2V 相似度指标分布直方图

关联度指标在不同报告期内数值分布的直方图如上图 4.2 所示。从该直方图中可以看出，W2V 相似度指标的分布在历史时期内逐步演化。这一变化可以类比成均匀分布的任意两点之间的空间距离从一维空间（线段）向二维空间（凸多边形）的推广，即关

① 由于所有的样本股票均能产生一个向量嵌入，每期观测数是每期样本股票数的二次幂（任意不同股票间均可计算关联度），明显超出人工关联中股票的配对数量。即便如此，得益于矩阵乘法的高效性，该方法计算余弦距离的时间复杂度是 $O(1)$，又因为 Word2Vec 模型训练速度快，因此该方法提取股票相似度指标时的计算效率远超 4.1.1 部分的人工关联度指标。

联度背后的特征维数逐年增长：在最早的报告期 20141231，配对观测数与人工关联度的关系接近线性，可能是因为当时的基金持股不够多样化，持股数量少，Word2Vec 几乎在一个次序给定的股票排序中学习两两距离；而随着时间变化，一维的线性关系逐步变化为二次函数，最初极端的左偏分布在 2018 年之后得到缓解。这表明，随着样本基金数量的增加和配对股票数量的增长，Word2Vec 提取的股票嵌入中的信息含量能随时间增长，最初几个训练期样本量不足的情况得以改善。此外，W2V 相似度指标存在少量负值：由于负值作为向量的余弦距离存在实际含义，因此后续研究中保留了这一分布，不再做 min-max 规范化。

表 4.7 呈现了各个报告期 W2V 相似度指标的若干统计量。这些统计量随着相似度指标的分布变化发生了稳定趋势的位移。从均值、中位值等统计量的变化规律来看，W2V 相似度在历年间呈现明显的下降的趋势，这可能和基金发行数量增加带来的持股偏好多样性有关。W2V 相似度指标的标准差随着时间先上升后下降，在 2019-2020 期间达到峰值；从下四分位数和上四分位数来看，样本的数据分布较集中，尤其是 2020 年后，这在上文中 W2V 相似度直方图当中已经有所体现。

表 4.7　W2V 相似度描述性统计

报告期	观测数	均值	标准差	最小值	25%分位数	中位数	75%分位数	最大值
20131231	407253	0.79	0.17	-0.07	0.70	0.84	0.93	1.00
20140630	536130	0.73	0.19	-0.08	0.61	0.77	0.88	1.00
20141231	705078	0.72	0.18	-0.08	0.60	0.76	0.87	1.00
20150630	833986	0.75	0.18	-0.09	0.64	0.78	0.89	1.00
20151231	1209790	0.72	0.20	-0.06	0.61	0.76	0.88	1.00
20160630	1504245	0.70	0.19	-0.07	0.58	0.73	0.85	1.00
20161231	1811656	0.65	0.19	-0.09	0.52	0.67	0.80	0.99
20170630	1983036	0.58	0.20	-0.15	0.43	0.59	0.74	1.00
20171231	1961190	0.53	0.21	-0.20	0.37	0.53	0.69	1.00
20180630	1867278	0.47	0.22	-0.22	0.31	0.47	0.63	0.99
20181231	1760626	0.46	0.22	-0.26	0.30	0.45	0.62	0.99
20190630	1904176	0.45	0.22	-0.25	0.28	0.44	0.61	0.99
20191231	1949325	0.44	0.23	-0.25	0.26	0.42	0.60	0.99
20200630	2021055	0.40	0.23	-0.29	0.23	0.39	0.57	0.99
20201231	2023066	0.37	0.22	-0.32	0.21	0.35	0.52	0.99
20210630	2063496	0.36	0.21	-0.28	0.21	0.34	0.49	0.99
20211231	2717946	0.33	0.20	-0.33	0.19	0.31	0.46	0.99
20220630	3347578	0.31	0.20	-0.37	0.17	0.30	0.44	0.99

续表 4.7　W2V 相似度描述性统计

报告期	观测数	均值	标准差	最小值	25%分位数	中位数	75%分位数	最大值
20221231	4465566	0.30	0.20	-0.36	0.15	0.28	0.42	0.99
20230630	4111278	0.34	0.21	-0.34	0.19	0.32	0.47	0.99
全历史	39183754	0.45	0.25	-0.37	0.25	0.43	0.64	1.00

和 4.1.1 部分对应，本文呈现报告期为 20221231 时贵州茅台（600519.SH）、宁德时代（300750.SZ）和科大讯飞（002230.SZ）W2V 相似度最高的前 10 只股票（见下表 4.8 至表 4.10）。这些过去几年热度较高的股票在人工关联度指标下的相似股票依然比较符合人们的认知，同 4.1.1 部分人工关联度的结果也高度相似。略有不同的是，所提取的高相似度股票中属于相同或相似行业的比重似乎略有增加，表明 Word2Vec 嵌入方法在捕捉行业关联的方面或许更加有效，这一结论会在后续章节中进一步讨论。对于流通市值，从呈现的三个热门股票来看，市值越大的股票，与之最相似的股票的市值也相应更大，这一结论也可以在后续讨论相似度和股票风格特征时进行进一步探究。

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	0.8379	食品饮料	1955.5

表 4.9　宁德时代（300750.SZ）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	300750.SZ	宁德时代	-	电力设备及新能源	7800.4
1	002594.SZ	比亚迪	0.9420	汽车	2993.1
2	300014.SZ	亿纬锂能	0.9346	电力设备及新能源	1614.3
3	601012.SH	隆基绿能	0.9284	电力设备及新能源	3202.5

续表 4.9　宁德时代（300750.SZ）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）
4	600438.SH	通威股份	0.8765	电力设备及新能源	1736.8
5	300059.SZ	东方财富	0.8738	非银行金融	2151.2
6	002812.SZ	恩捷股份	0.8719	基础化工	974.2
7	002459.SZ	晶澳科技	0.8718	电力设备及新能源	1408.2
8	601888.SH	中国中免	0.8661	消费者服务	4217.9
9	600519.SH	贵州茅台	0.8650	食品饮料	21694.5
10	002475.SZ	立讯精密	0.8641	电子	2250.7

表 4.10　科大讯飞（002230.SZ）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	002230.SZ	科大讯飞	-	计算机	695.6
1	600703.SH	三安光电	0.7948	电子	699.9
2	600745.SH	闻泰科技	0.7919	电子	653.5
3	002008.SZ	大族激光	0.7657	机械	251.1
4	000938.SZ	紫光股份	0.7622	计算机	558.0
5	600570.SH	恒生电子	0.7568	计算机	768.7
6	300033.SZ	同花顺	0.7518	计算机	267.7
7	000776.SZ	广发证券	0.7375	非银行金融	916.9
8	600741.SH	华域汽车	0.7189	汽车	546.4
9	300142.SZ	沃森生物	0.7181	医药	626.2
10	002236.SZ	大华股份	0.7154	电子	221.6

综合 4.1 小节的内容，无论是人工构造的股票关联指标或是嵌入模型提取的股票相似度特征，都能取得较理想的效果。一方面，人工关联度指标和 W2V 相似度指标在每个报告期内的数值分布较为理想，既符合接近正太分布的偏态分布，又在截面上具有相当的分化；另一方面，若干热门股关联度（或相似度）最高的股票，往往具有类似的行业分类和市值特征。由于相似度指标的提取只依赖基金持股数据，这已经初步验证了假设一，即通过人工构造或嵌入表示的方法，基金持股数据中的公司特征能够被有效识别和利用。

4.2　风格、行业与收益表现

在这一部分，本文希望对假设二、假设三和假设四进行更充分的探讨。假设二认为基金持股中获得的股票关联信息对股票的风格、行业特征以及股票价格联动具有区分能力。假设三和假设四则提出，嵌入方法比人工方法在风格-行业划分和股价关联捕捉这两个方面的应用中均更加有效。在检验上述假设时，本文更关注通过人工或嵌入模型提取的股票关联性指标和公司风格、行业的显性特征以及公司股价之间的相关性，因此采用分组对比均值的方式进行直观刻画。具体地，根据显性特征的取值进行分组，比较分组内股票关联性指标统计量在分组之间的差异，从而识别出风格、行业特征对股票关联性指标的影响（4.2.1 和 4.2.2 小节）；或者，根据股票两两之间的关联性指标对股票两两配对进行分组，比较不同分组下的股票未来收益率的相关性，从而识别出关联性指标对于股票收益关联的刻画能力（4.2.3 小节）。

4.2.1　风格特征与股票关联

在前述模型中，经由建模我们获得了股票两两之间的关联度指标。将样本中的基金持股股票按风格排序在截面上进行分组，观察各分组内部股票关联度指标之间是否具有差异，从而检验显性的风格特征和从基金持股中提取的隐形关联特征之间的相关性。

市值因子始终是各类定价模型以及在中国市场最为基础和有效的因子。我们根据流通市值将每一个报告期的全市场股票等分成五组，计算组内股票间的平均人工关联度（或 W2V 相似度）。结果见下图 4.3（或图 4.4）。

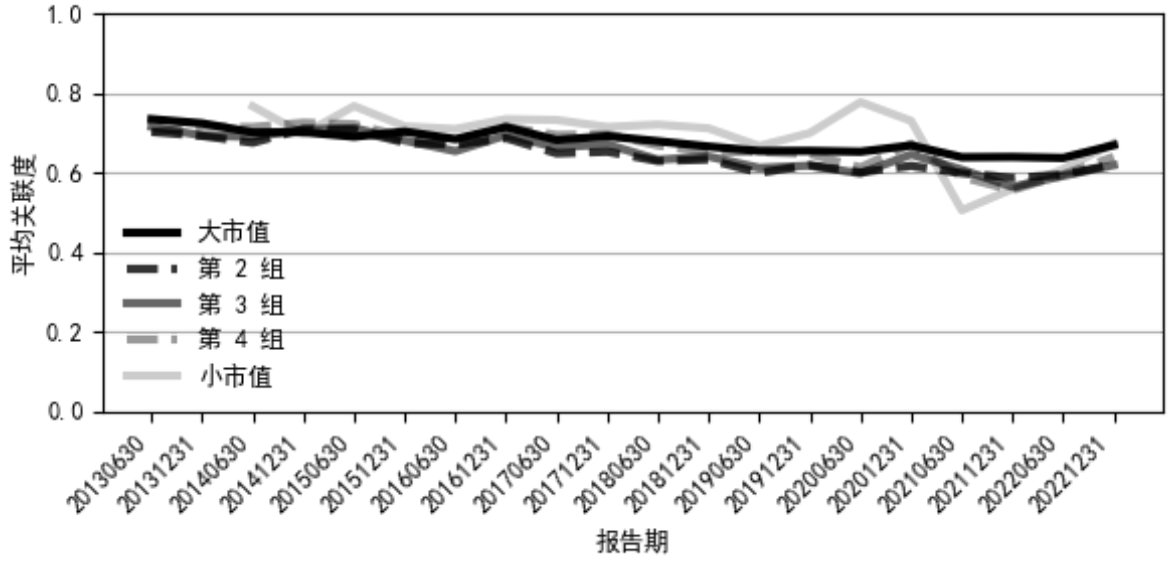


图 4.3　不同流通市值分组下的人工关联度指标历史均值

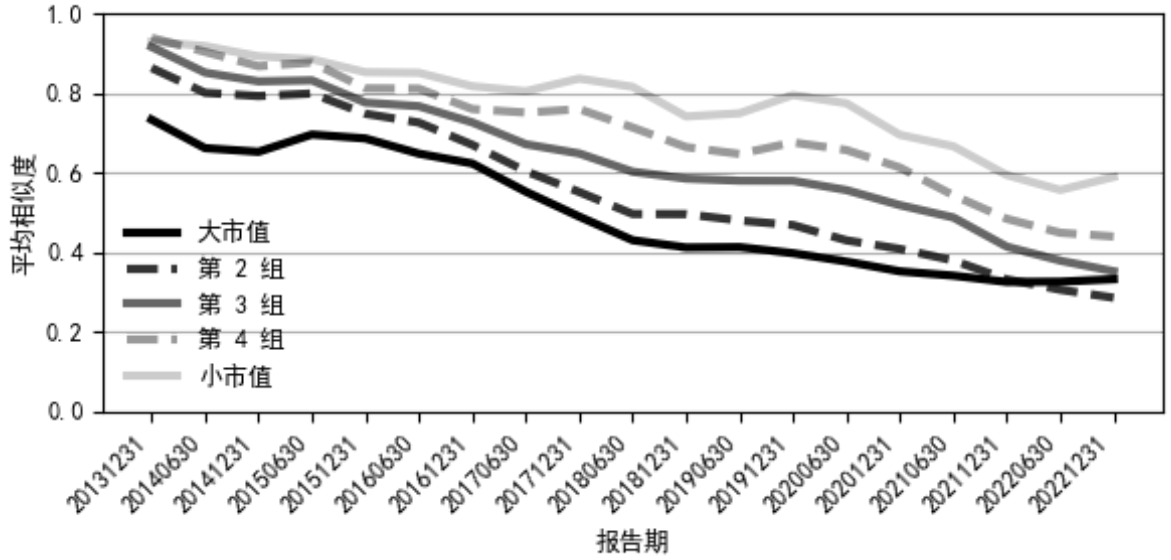


图 4.4　不同流通市值分组下的 W2V 相似度指标历史均值

图 4.3 表明，人工关联度指标在不同流通市值分组之间几乎没有差距；而图 4.4 则表明，W2V 相似度指标能够被流通市值分组加以区分，呈现出“市值越大，股票间的相似度越低”的特点。这一相似度对于市值的单调性在历史各个报告期均存在。这一现象背后可能的解释是，主动基金对于大市值股票有更高的需求，因此更多样化地持有大市值股票，而对于少部分能够被基金重点关注的小市值股票，则可能需要达到更加严苛的入选标准，而因此具有更高的相似性。由于股票之间 W2V 相似度的数值分布在各个历史报告期发生切换，构造两样本均值差异的显著性检验是不合理的；但由于各分组内的股票配对数量足够多，“大市值”组相对于“小市值”组在 W2V 关联度均值上的差异理应显著。与之相反的是，在图 4.3 中，不同分组间平均相似度的单调性非常微弱，且在历史上存在多次反转。对比图 4.3 和图 4.4，可以认为假设三在市值这一显性特征上成立，即嵌入方法比人工方法具有优势。

采用类似的方法，本文也探索了其他重要的风格特征和股票关联指标间的关系（见附录 A）。遗憾的是，无论是人工关联性指标还是 W2V 相似度指标，在除市值以外的诸多风格上均未体现出明显的单调性。这或许表明，基金持股数据中隐藏的股票关联信息独立于常见的风格指标，是常见风格因子的有效补充。这一情况也有利于发掘股票关联当中的“异象”，这在本文 4.3 部分将有所体现。

4.2.2　行业分类与股票关联

类似风格分组，我们也分别计算同行业和不同行业的组内股票相似度均值（图 4.5 和图 4.6）。各报告期使用的行业分类依据“中信一级行业”。从两图中均可看出，各报告期内同行业的平均关联度（相似度）高于不同行业，其中 W2V 相似度指标（图 4.6）

在同行业和不同行业间的均值差异更为明显。[①]即使能注意到差异，同行业和不同行业之间的平均关联度（相似度）差异并不明显。因此，行业并非区分股票关联的决定性因素：基金持股中的关联信息所捕获的同行业内股票间的差异通常可能远超不同行业所带来的差异。

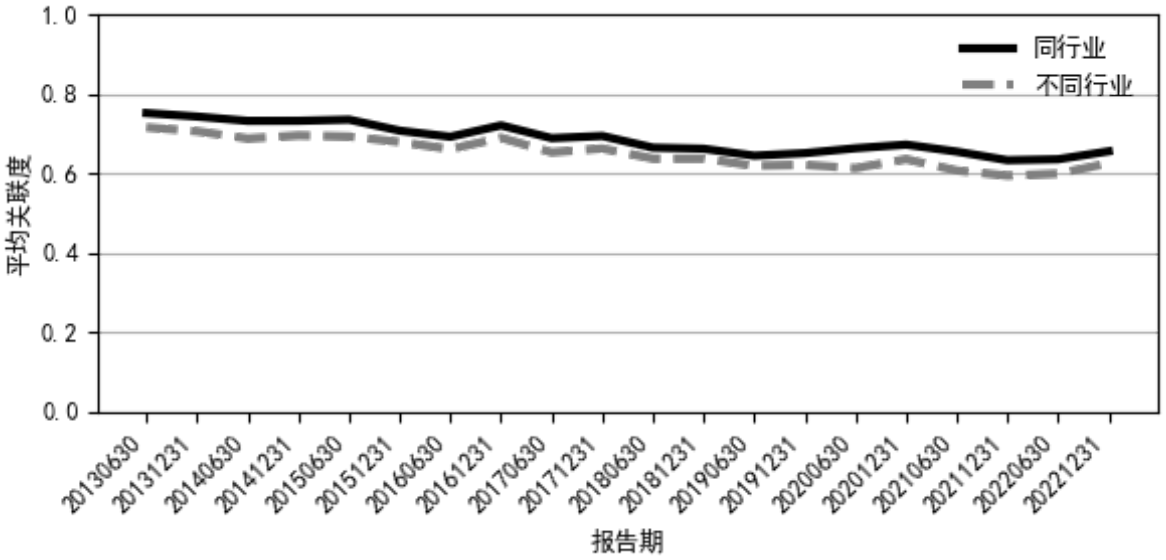


图 4.5　同行业及不同行业之间的人工关联度指标历史均值

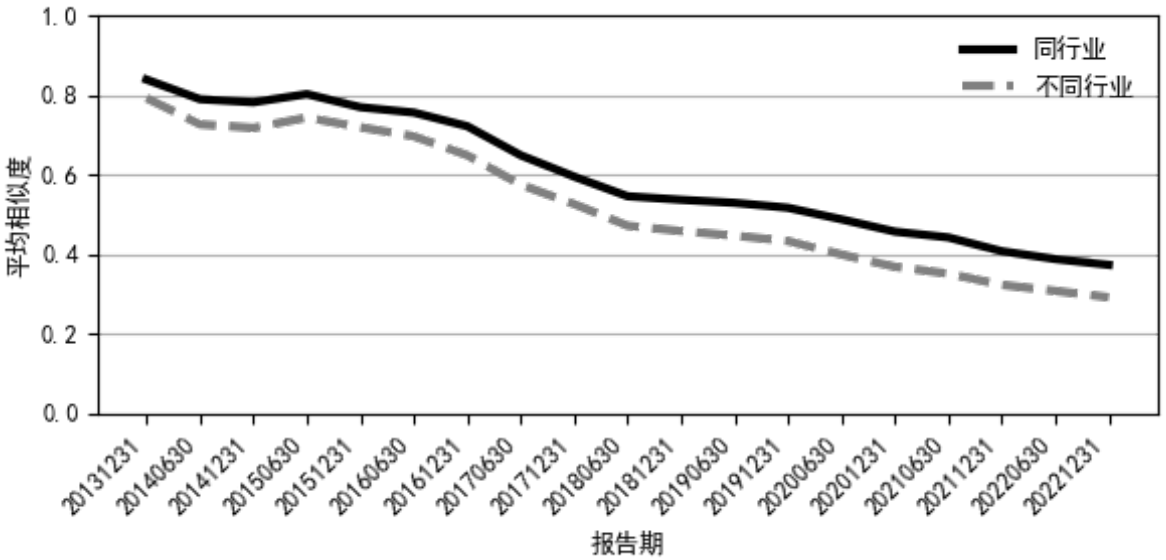


图 4.6　同行业及不同行业之间的 W2V 相似度指标历史均值

Word2Vec 模型输出的嵌入向量作为股票特征的低维表示，其本身也可能反应股票的行业属性。在低维嵌入向量的基础上用 TSNE（t-Distributed Stochastic Neighbor

① W2V 相似度取值范围略微更大（见表 4.7），但对比均值时可忽略不计。

Embedding）进一步降维到二维平面。[①]下图 4.7 和图 4.8 呈现了样本最新的两个报告期中股票数量前十的行业的聚集情况。从图中容易观察出，相同行业的股票在二维空间中存在聚集现象，例如“食品饮料”和“农林牧渔”，相对于其他行业较为独立；一些行业之间的空间关系较为靠近，例如“基础化工”“电力设备及新能源”和“汽车”；同一行业也可能分散成多个聚落，例如“通信”和“传媒”。综合来看，基金持股嵌入表示中的股票空间关系既包含了行业分类信息，也能兼顾其他维度的股票差异，。

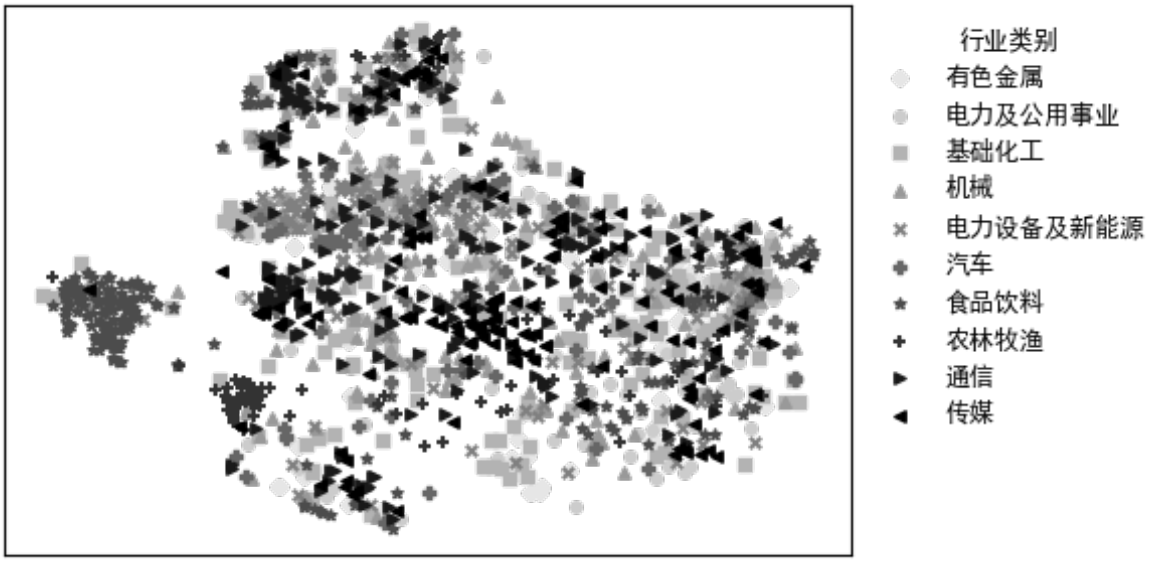


图 4.7　W2V 嵌入 TSNE 降维后聚类结果 - 数量前 10 的行业（2022 中报）

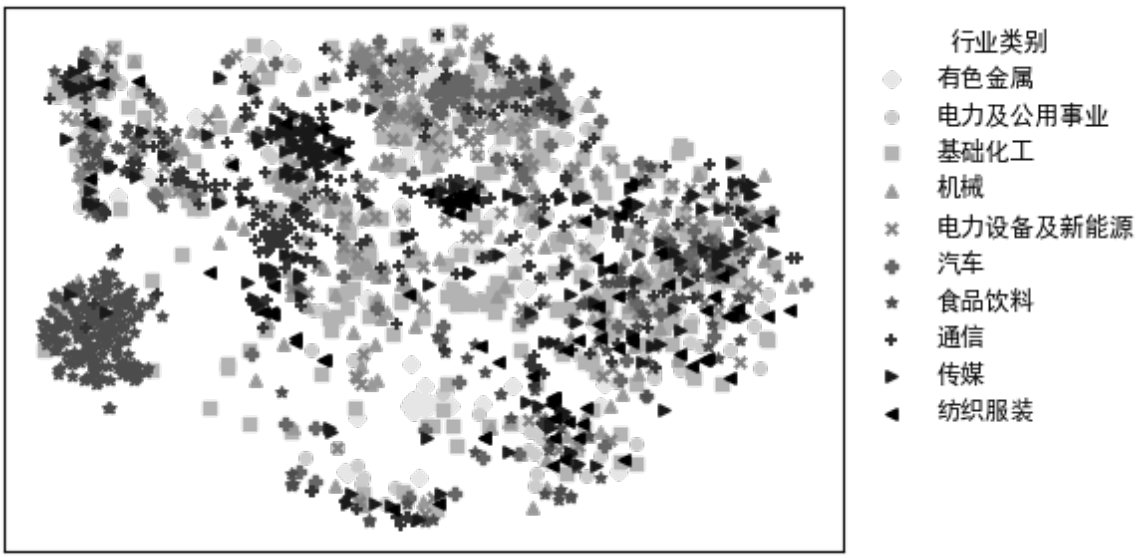


图 4.8　W2V 嵌入 TSNE 降维后聚类结果 - 数量前 10 的行业（2022 年报）

① 不同于 PCA 线性降维，TSNE 通过优化目标函数，使得高维空间中距离相近的数据点在二维（或三维）空间中仍然保持相近的关系，从而呈现数据中的群集和关联关系。本文写作时发现 TSNE 的可视化效果比 PCA 更好。

4.2.3　股票关联与未来收益

更高的关联度是否意味着收益率之间具有更高的同步性？根据上季度末尾的基金持股计算股票关联度，根据关联度高低将股票分组，统计基金持股披露后下一季度内、不同关联度分组中、股票两两之间日收益序列的相关性（图 4.9 至图 4.12）。

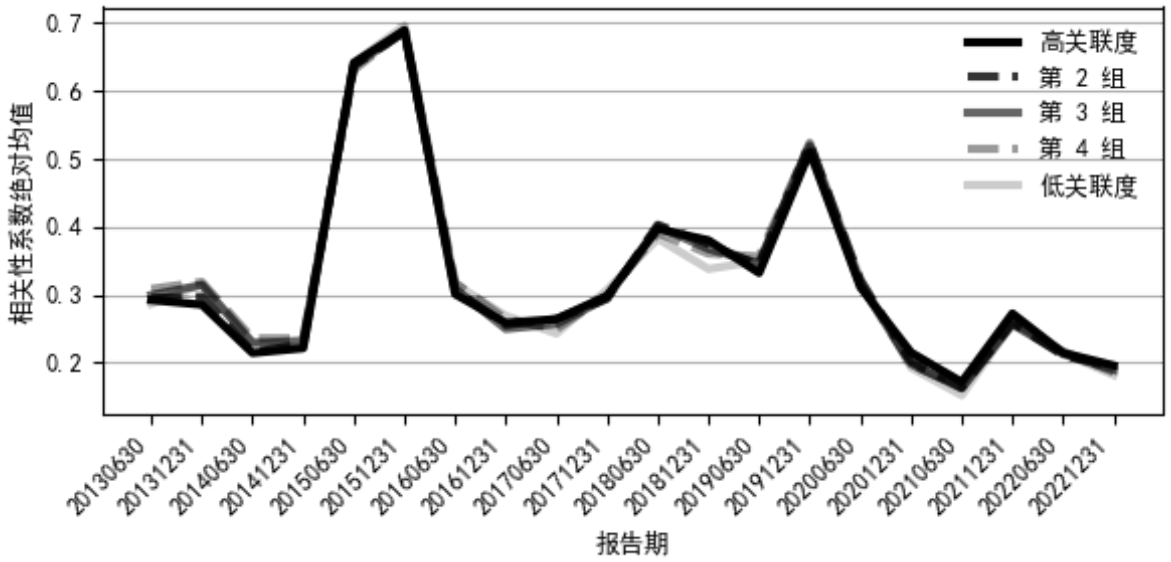


图 4.9　不同人工关联度分组内股票收益率相关性系数绝对值均值（每期抽样 10 万对）

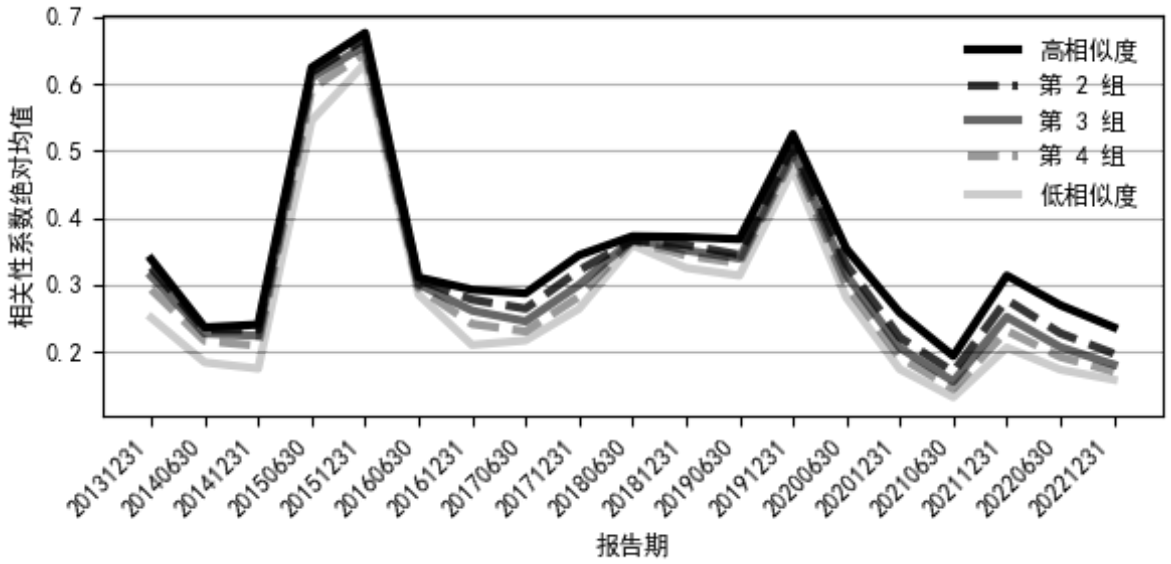


图 4.10　不同 W2V 相似度分组内股票收益率相关性系数绝对值均值（每期抽样 10 万对）

首先，从图 4.9 和图 4.10 中相关性系数绝对均值在不同报期间的大幅波动中可以看出，股票收益率的相关性在历史不同时期呈现明显的差异，这一变化来自全市场在时序上的系统性变化，而同人工关联度（或 W2V 相似度）分组的关系不大。时间因素

带来的股票收益相关性波动远远超过关联性指标带来的差异。因此，需要比较去除均值后各分组内的相关性差异（图 4.11 和图 4.12）。

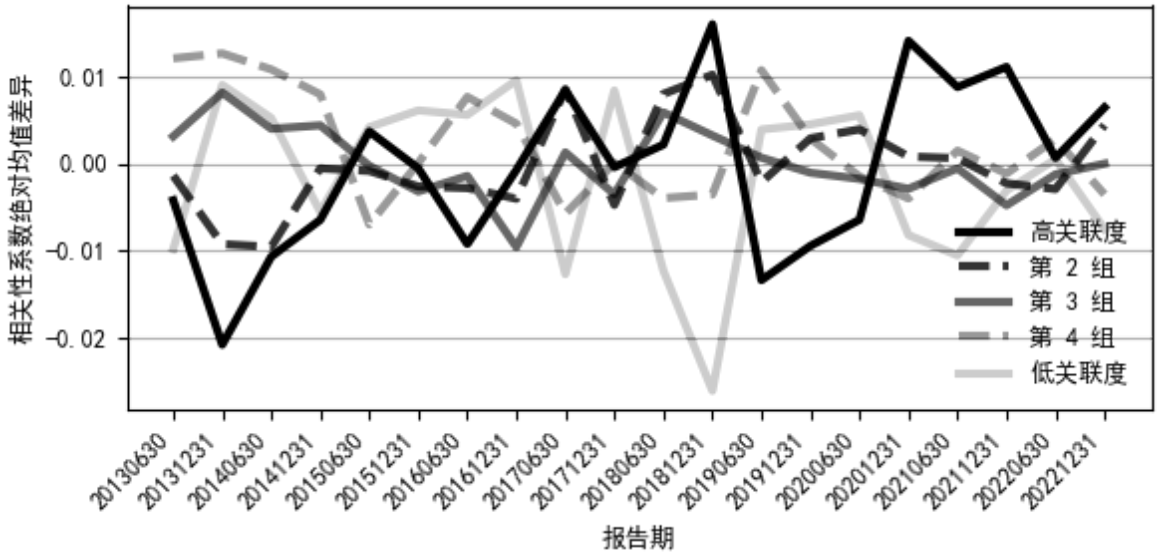


图 4.11　不同人工关联度分组内股票收益率相关性系数绝对值均值（相对于均值）

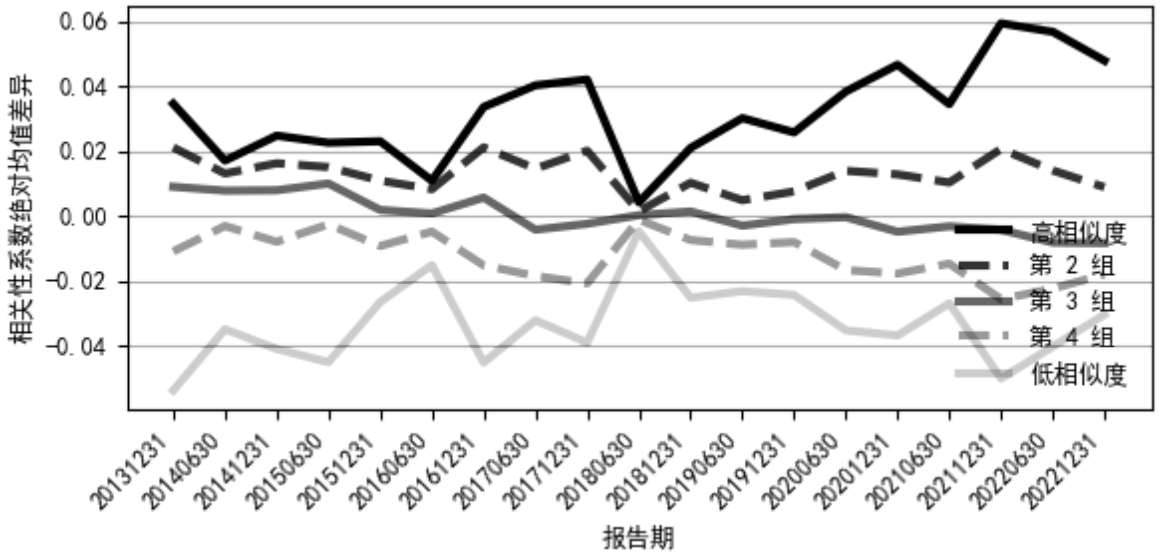


图 4.12　不同 W2V 相似度分组内股票收益率相关性系数绝对值均值（相对于均值）

人工关联度的分组并不能够区分资产未来收益的相关性（图 4.10），但是 W2V 相似度则在股价收益率相关性上具有一定的区分能力（图 4.12）：体现为关联度更高的股票在基金持股披露后一个季度内股价日收益率的相关性也更高。但是，收益相关性同关联性指标间的单调关系并不太稳定，尤其是报告期 20180630 之后的一个季度。因此，对于股票关联指标和股价变动的关系，需用下文 4.3 部分更进一步的方法来识别。

4.3　具有定价能力的关联动量

4.3.1　因子构造

从基金持股中提取获得的资产关联信息，如果仅仅是对资产风格特征、行业分类、收益表现等更显性的股票特征的重复，而不具有额外信息，那么对基金持股数据应用嵌入方法也将失去实际的应用价值。嵌入表示的资产关联特征中蕴含的公开市场信息，不仅仅应该和其他资产特征之间相互联系，还应该反映出那些不够直观、用其他来源的数据无法捕捉到的股票特征。本文 2.2 部分进行理论分析时，提出基金持股信息具有定价能力的两大机制：基金管理人认知和股东持股需求。无论是基金经理挖掘出的未能反映在市场价格中的股票特质，还是买方需求视角下、股票之间的替代或互补关系，这些信息可能都未能引起其他市场参与者足够的重视和发掘，导致基金披露持仓之后的半年到一年内依然存在关联股票之间滞后的关联动量现象。

具体而言，在基金持股得到的股票关联网络中，如果信息传导足够迅速，那么驱动某些公司股价变动的新息将会迅速波及相邻股票；基金持股中特有的股票关联是基金经理利用私人信息优势所捕获的，就全市场而言信息传导并不及时，那么相似股票背后共同的价格驱动因素在关联网络的中的传导将存在时间上的滞后。基于上述分析，可以猜想，关联股票的价格变动对中心股票具有滞后的“牵引”作用：如果关联股票普遍上涨，相对而言中心股票的价格在过去涨幅不足，那么随着关联股票间相似的特征被市场挖掘和传播以及市场参与者对作为替代品的中心股票需求上升，关联股票的历史价格和中心股票的未来价格之间应该存在滞后的动量效应。由此，本文构建如下形式的关联动量指标：

（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（式 4.1），减去截面中位值，刻画股票 i 在过去 20 日的超额收益（式 4.2）：

$$R_{i,t}^{20} = \frac{Open_{i,t}}{Open_{i,t-20}} - 1 \tag{4.1}$$

$$AR_{i,t}^{20} = R_{i,t}^{20} - Median_i\left(R_{i,t}^{20}\right) \tag{4.2}$$

（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）：

$$ER_{i,t}^{20} = \frac{\sum_{j\neq i} AR_{i,t}^{20} \cdot \lambda_{i,j}}{\sum_{j\neq i} \left|\lambda_{i,j}\right|} \tag{4.3}$$

其中 $\lambda_{i,j}$ 为最新基金持股数据中股票 i 和 j 的人工关联度 $K_{i,j}^{*}$（式 3.4）或 W2V 相似度 $\mathrm{sim}_{i,j}$（式 3.9）。

（3）每期截面上，预期超额收益对实际的过去 20 日超额收益回归（式 4.4）[①]，取残差 $\varepsilon_{i,t}$ 即得到关联动量牵引因子原始值:

$$ER_{i,t}^{20} = \beta_1 \cdot AR_{i,t}^{20} + \varepsilon_{i,t} \tag{4.4}$$

上述式 4.4 的回归能够捕捉预期超额收益未能反应在实际已实现收益中的部分，即关联动量的强度。关联动量取值越高，未来期望收益越高。接下来，分别使用人工关联度和 W2V 相似度作为股票关联指标，构造两组因子，对因子的截面选股能力进行检验。

4.3.2 因子评估

首先，基于人工关联度指标计算“人工关联度牵引因子”，因子表现见图 4.13 和图 4.14，其中图 4.13 展示因子截面分层能力[②]，图 4.14 展示了人工关联度牵引因子的信息系数（IC，Information Coefficient）月度均值和累计值[③]。

由图 4.13 可知，基于人工关联度指标构造的牵引因子具有一定的分层能力，表现为累计相对收益随着分组因子值位次增加而依次上升；但在 2016-2020 期间因子对前 80%的股票分层能力一般，表现为“多头组”至“第 4 组”的分化不明显；5 组曲线的形状分布相对 $y = 0$ 轴不对称，表明因子的分层能力主要由空头组贡献。

图 4.14 中，人工关联度牵引因子因子 rank IC 均值为 1.85%，ICIR 为 0.1786，具有一定的预测能力。累积 rank IC 在 2020 年下半年有明显回撤，说明因子在历史上存在失效情况，这在图 4.13 中也有所体现；2017-2018 年期间因子表现一般，第一和第三季度 rank IC 月度均值（左轴）相对更高、第二和第四季度更低，可能是这一时期基金持股风格切换较为频繁所致。这表明对基金持股数据降采样到半年度后或存在失效风险。

接下来，基于 W2V 相似度指标计算“W2V 相似度牵引因子”，因子表现展示在图 4.15 和图 4.16 中。对比图 4.13，图 4.15 中 W2V 相似度牵引因子在历年间各分组走势的趋势更加均匀，单调性优秀，表明因子的分层能力更加稳定；折线图分布相对 $y = 0$ 轴更为对称，说明因子分层能力在截面分布上同样更为连续。

图 4.16 反映 W2V 相似度牵引因子的 rank IC 情况。用嵌入模型改进后的关联动量牵引因子，rank IC 均值达到 2.41%，比人工关联度因子提高 30%；ICIR 达到 0.1845，也高于图 4.14 的结果。观察累计 rank IC 曲线，不存在明显的回撤，并且历年上升斜率基本不变，表明该因子的截面预测能力相当稳定。

① 回归无截距项，否则存在共线性问题（式 4.3）。

② 根据因子值截面排名，将样本股票分为 5 组，各组内部等权持有股票，计算各分组的相对收益率（基准为全部样本股票该日期的平均收益率），最终绘制各分组的累计超额收益率。

③ $rank\ IC$ 为每日因子值和次日收益率在截面上的 Spearman 相关性系数，反映因子对未来收益率的预测能力；图右下方注记中的“均值”为 rank IC 均值，“ICIR”为 rank IC 的信息比率（Information Ratio），计算方式为 rank IC 均值除以 rank IC 标准差 $ICIR = \text{Mean}(IC)/\text{Std}(IC)$，反映因子预测能力在时间上的稳定性。

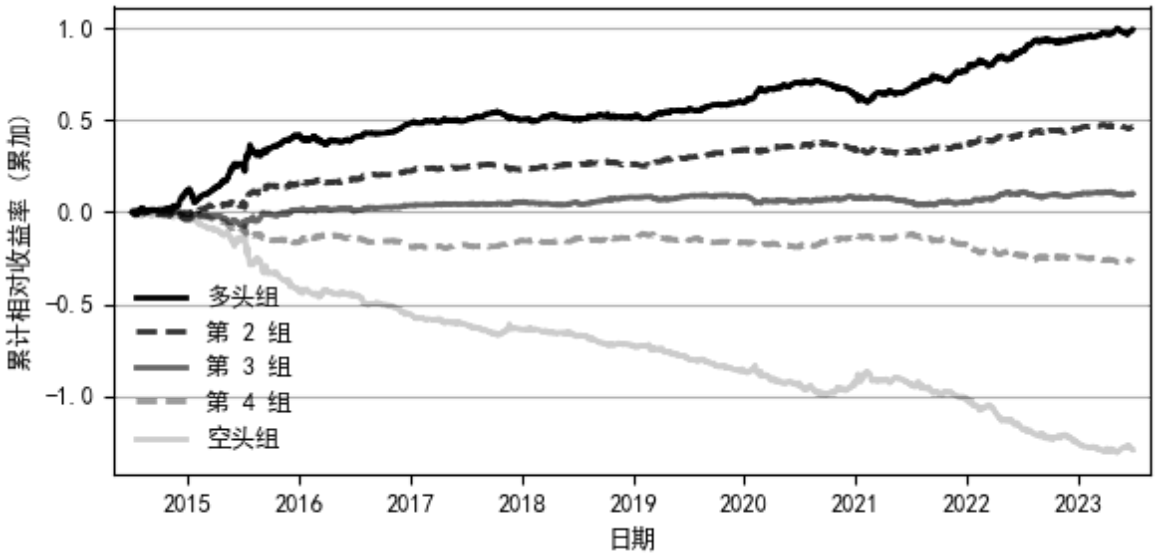


图 4.13　人工关联度牵引因子 5 分组收益分化情况

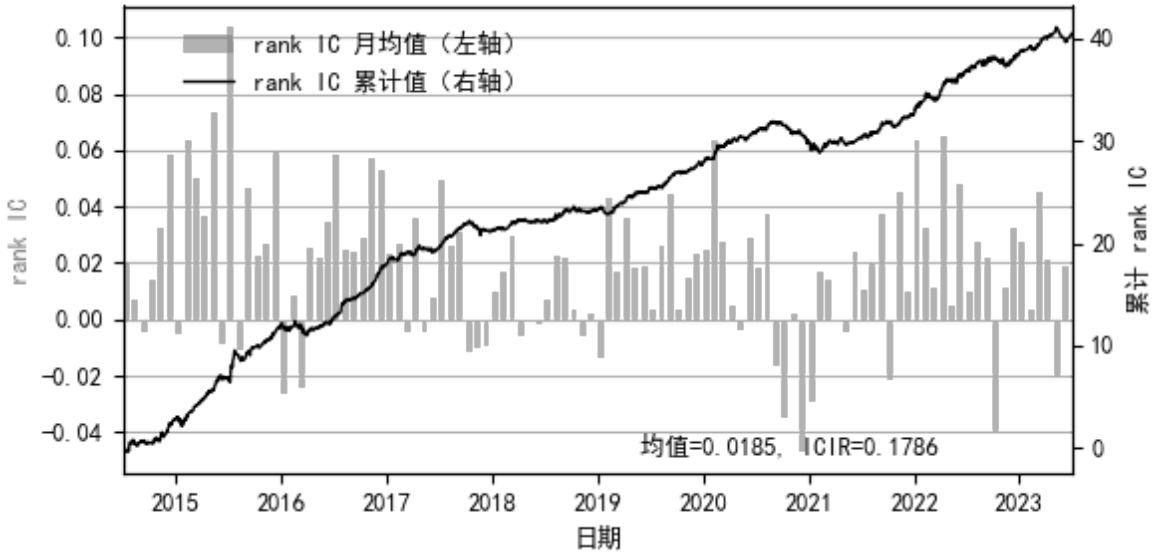


图 4.14　人工关联度牵引因子 rank IC 和累计 rank IC

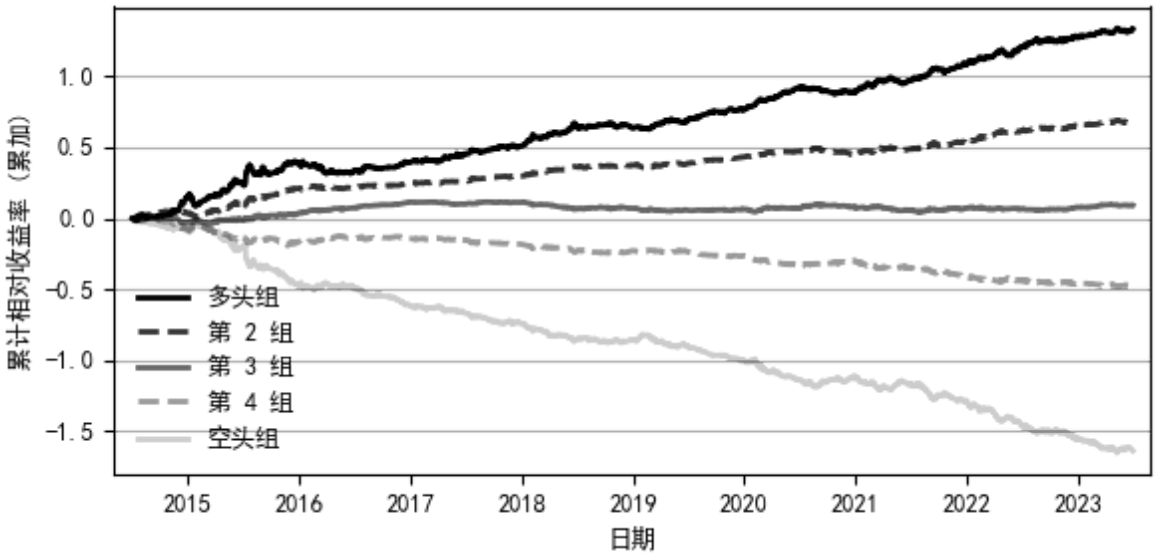


图 4.15　W2V 相似度牵引因子 5 分组收益分化情况

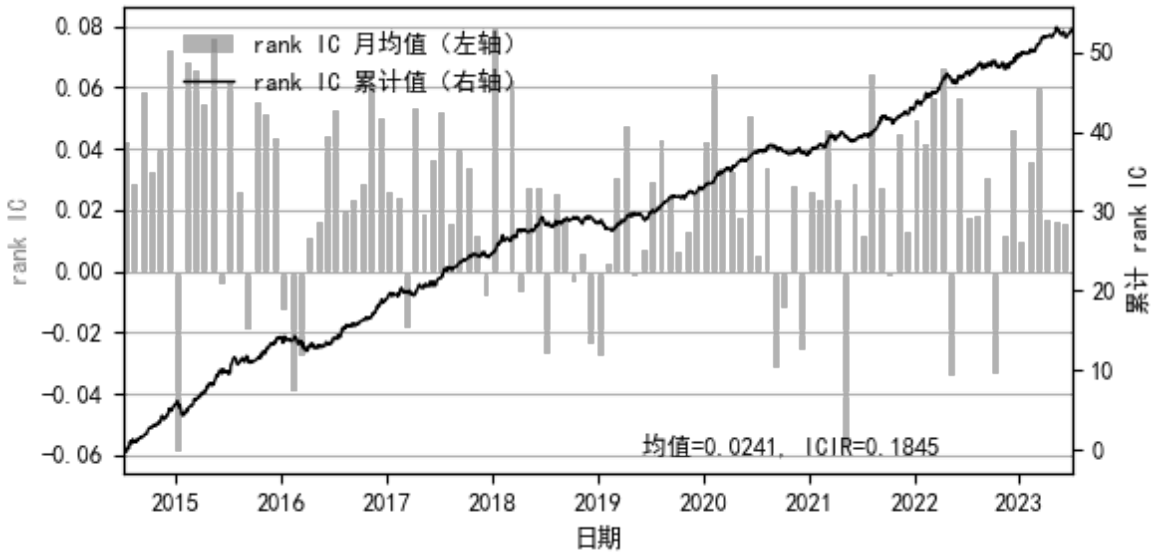


图 4.16　W2V 相似度牵引因子 rank IC 和累计 rank IC

最后，对因子进行回测，分别测试“多头组”“空头组”“多 - 空”组历史表现，并与全市场基准指数（中证全指）对比。①

下表 4.11 和图 4.17 展示了人工关联度牵引因子的回测表现。②因子“多头组”年化收益率为 15.5%，夏普率 0.51；“多 - 空”组年化收益率为 12.5%，夏普率 2.07，最大回撤为 11.6%，具有良好的对冲效果；多头组日均换手率达到 38%，日度的交易成本

① 根据因子值对每个截面的股票排序，等权持有因子值最高的 20%的样本股票，构成“多头组”；等权持有因子值最低的 20% ，构成“空头组”；买入“多头组”、卖出“空头组”构成“多 - 空”组。
② 组合每半年度的表现见附录 B 中表格。

较高。观察图 4.17，“多 - 空”表现在 2021 年初有所回撤。以上回测结果表明，通过人工构造的方式刻画基金持股中的股票关联，已经能在关联资产间观测到关联动量。

表 4.11　人工关联度牵引因子全历史回测结果（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	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%

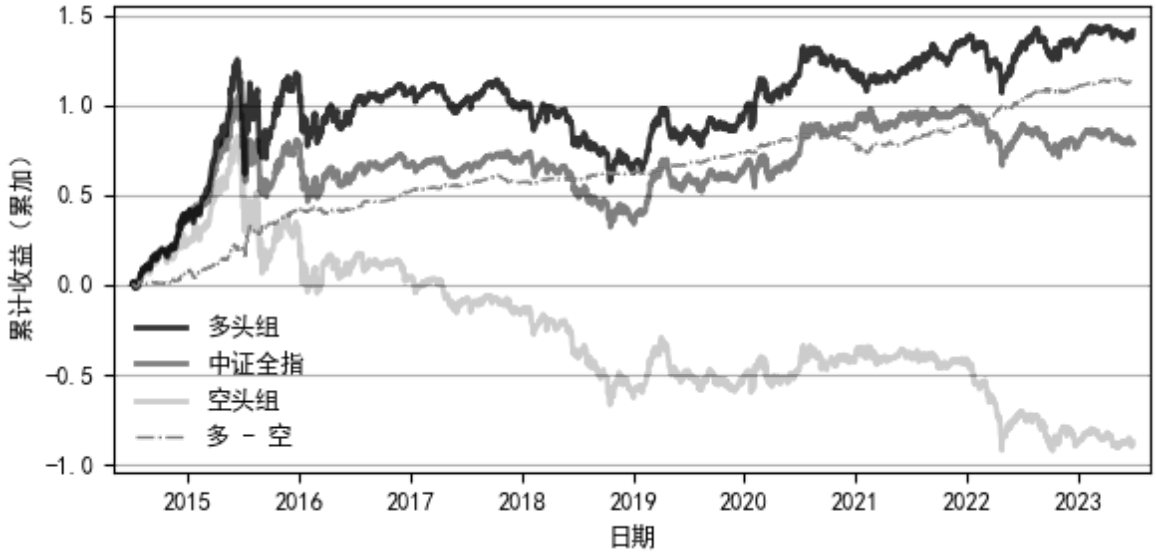


图 4.17　人工相似度牵引因子 5 分组多头/多空策略回测表现

下表 4.12 和图 4.18 展示了 W2V 相似度牵引因子的回测表现。因子“多头组”年化收益率为 17.9%，夏普率 0.60；“多 - 空”组年化收益率 16.3%，夏普率 2.35，最大回撤为 7.9%，各项指标均好于人工关联度因子；策略日均换手率略低于人工关联度因子，但仍然不可忽视。图 4.18 所示的“多 - 空”分组收益更平滑，回撤较小。嵌入表示提取的股票相似度中的关联动量牵引作用更明显，信息提取比人工模型更有效。

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%

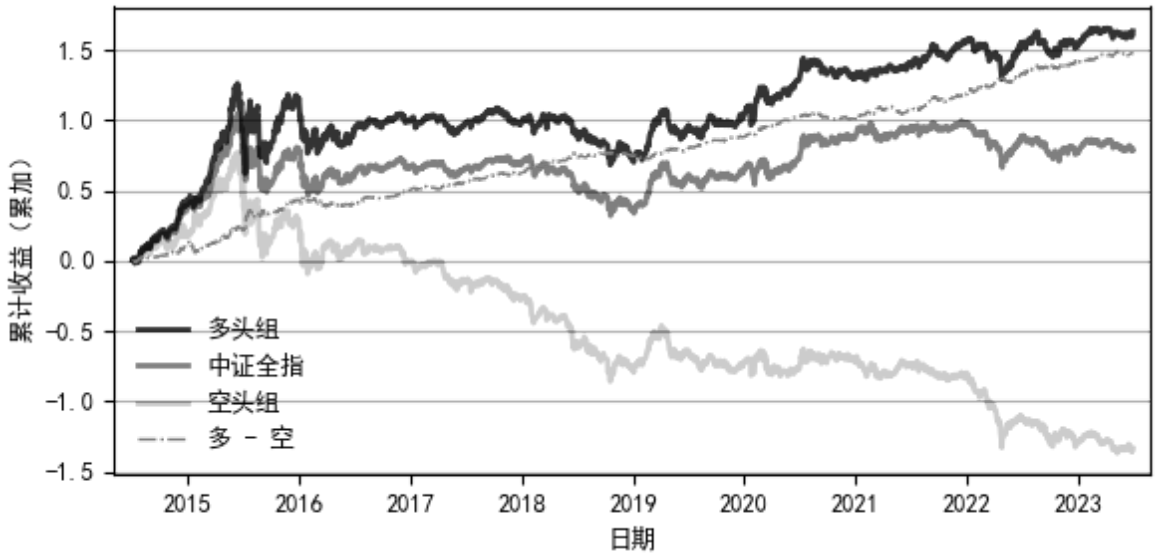


图 4.18　W2V 相似度牵引因子 5 分组多头/多空策略回测表现

前文 4.2 小节的描述性统计表明，关联度指标中含有一定的行业和市值信息，具有选择行业、选择风格的能力。作为对照，对因子原始值依次进行行业、市值中性化。[①] 比较中性化前后的 IC 指标（表 4.13）和多空策略回测表现（表 4.14），简要结论如下：（1）行业中性化后，Rank IC 下降，多头、多空策略的超额收益下降，但是 ICIR 和夏普率明显提升；（2）市值中性化后，Rank IC 和多头、多空策略的收益能力进一步下降，ICIR 和夏普率略有提升。以上差异表明，本文提取的人工关联度牵引因子和 W2V 相似度牵引因子具有行业择时能力和风格（市值）择时能力，基金持股中所提取的关联动量牵引因子能够通过有效的风格暴露来提升收益表现。此外，去除行业暴露后，ICIR 大幅提升，说明因子截面选股能力的波动更多来自于行业暴露的波动，因子在各个行业内的选股能力相对更稳定。

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

① 行业中性化方式为因子值行业内中心化（原始值减行业均值）；市值中性化方式为回归（行业中心化后的因子值对截距项和对数流通市值回归，取残差）。中性化后的详细表现见附录 C 部分。

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%
多 - 空	13.9%	3.47	5.3%	3.7%	60.0%	-
空头组	-14.2%	-0.47	-22.8%	76.0%	51.4%	37.1%

4.3.3　因子检验

以上因子评估表明——关联度牵引因子本身具有截面定价能力，存在“关联动量”异象。但经过行业、市值中性化后因子的 IC 和多空净值表现有所下降，说明关联动量的定价能力部分来源于风格暴露。基金持股中相似股票间的关联动量携带的 alpha 信息和常见的公司特征之间可能存在重合。下面用 Fama-MacBeth 截面回归（Fama et al., 1973, 2020）进行异象检验。

检验的设定为：每个截面，取样本股票下期相对中证全指的超额收益率（单位：%）作为被解释变量，用异象变量和多因子模型中的变量作为解释变量，每个截面进行一次回归。每次回归产生异象因子的估计系数，即异象因子的因子收益率。检验异象收益率序列的均值和 *t* 值。[①]如果异象收益率显著不为 0，则认为异象因子能够获得多因子模型无法解释的超额收益。

为更贴近 A 股市场近年来的实际状况，本文选用 Barra CNE5 风险模型中的 10 大风格因子的因子暴露。[②]回归的变量说明见表 4.15；所有因子均进行去极值和标准化处理。[③]回归的时间范围为 2015 年 1 月至 2023 年 6 月的所有交易日，共 2064 期。

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。

① 由于因子收益（回归估计系数）在时序上潜在的异方差和自相关性，计算 t 值时经过 Newey-West 调整。
② 也称作因子载荷。除市场因子 beta 来源于 CAPM 模型的时序回归外，其余因子为股票特征指标的原始值。详细因子定义见 https://www.msci.com/www/research-report/the-barra-china-equity-model/014459336 附件。
③ 去极值方法为绝对中位差（median absolute deviation）法；标准化方法用 Z-Score 法，公式为：ZScore = (变量 − 变量均值)/变量标准差。

对以上因子计算截面因子暴露的相关性系数，并在时序上取均值（图 4.19）。首先，4 个待检验的异象因子之间相关性较高，尤其是行业中性化前后的相同因子高度相关，说明因子反映同质化信息，符合逻辑；其次，异象因子和风格因子之间相关性不高，除动量（相关性约为 0.2）外，同其余风格因子的平均相关性系数均在 0.1 以下；对于风格因子，因为构造逻辑来源于相似的公司特征，存在较高的相关性系数，例如“流动性”和“残余波动率”，“非线性市值”和“市值”，因此在后续回归时需要注意共线性问题。①

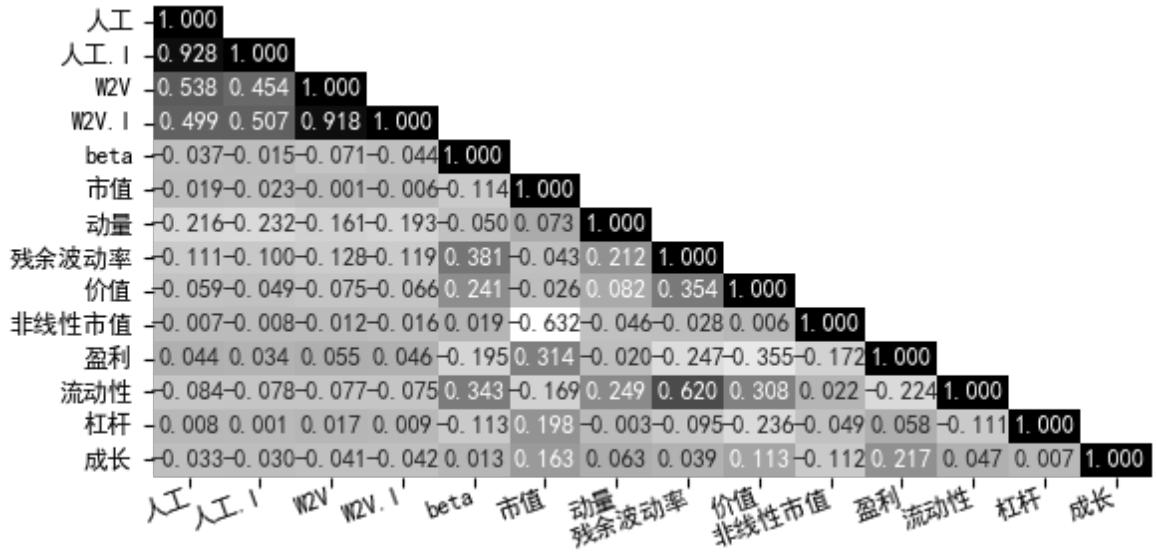


图 4.19　因子相关性分析

本文根据 VIF 统计情况（附录表 D.1）未发现截面回归中存在明显的共线性。此外，在回归前统计风格因子本身的定价能力（附录表 D.2 和表 D.3）。结果显示，风格因子本身具有定价能力，历年间的 rank IC 和 ICIR 相对稳定。

分别进行五组 Fama-MacBeth 回归，其中前四组分别包含四类关联度牵引因子，第五组只包含风格因子；五组回归的样本范围保持一致，以便对比。回归结果如下文表 4.17 所示。根据回归 (1) 至 (4) 组，异象因子在控制其他风格因子后均具有显著的因子收益②；从因子收益率规模来看，第 (3) 组回归中的“W2V”变量，即 W2V 相似度牵引因子的收益均值最高——因子载荷每提升 1 个标准差，控制其他风格因子后的预期收益提高 0.0226%；因子收益规模和本文 4.3.2 部分因子评估指标的单调性一致。

① 具体做法：每个截面进行回归前进行 VIF 检验，若不通过（VIF 最大值超过 5 或 10），即当天的估计系数不精确，则在计算时序均值和 t 检验时对该截面予以排除。历次回归时的 VIF 分布情况见附录表 D.1，实际情况表明，共线性问题并不严重（VIF 截面最大值均小于 10，VIF 最大值 75%分位数为 2.45）。此外，如果风格因子之间存在共线性，可根据风格因子评估表现进行变量筛选，或对高相关的风格因子根据历史IC均值进行加权合成。

② 检验单个因子时一般 t 值大于 2（小于-2）即可在 5%的显著性水平上拒绝原假设；评估多组因子时，由于多重假设检验问题，文献建议 t 值绝对值需要达到 3.0（Harvey et al., 2016）甚至 3.4 （Chordia et al., 2020）。

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	0.0224	0.0242	0.0228	0.0235
	[2.79]	[2.76]	[3.00]	[2.81]	[2.88]
市值	-0.0225	-0.0223	-0.0207	-0.0217	-0.0232
	[-3.60]	[-3.54]	[-3.32]	[-3.44]	[-3.68]
动量	-0.0351	-0.0354	-0.0342	-0.0340	-0.0408
	[-4.40]	[-4.44]	[-4.15]	[-4.13]	[-5.25]
残余波动率	0.0463	0.0463	0.0460	0.0458	0.0453
	[6.89]	[6.83]	[6.99]	[6.90]	[6.73]
价值	0.0005	0.0004	0.0019	0.0009	0.0006
	[0.08]	[0.06]	[0.30]	[0.14]	[0.09]
非线性市值	0.0084	0.0085	0.0094	0.0091	0.0089
	[2.62]	[2.65]	[2.91]	[2.82]	[2.77]
盈利	0.0050	0.0050	0.0059	0.0052	0.0053
	[1.21]	[1.19]	[1.44]	[1.26]	[1.29]
流动性	-0.0989	-0.0977	-0.0981	-0.0981	-0.0970
	[-9.35]	[-9.87]	[-9.87]	[-9.75]	[-9.74]
杠杆	-0.0045	-0.0047	-0.0051	-0.0049	-0.0047
	[-1.23]	[-1.27]	[-1.42]	[-1.34]	[-1.27]
成长	0.0244	0.0244	0.0246	0.0247	0.0244
	[7.37]	[7.35]	[7.54]	[7.50]	[7.39]
日均观测数	1399.4	1399.4	1399.4	1399.4	1399.4
日均 Adj. R2	0.1249	0.1241	0.1271	0.1252	0.1224

注：中括号内为 *t* 值

表 4.17 中，因子定价能力的强弱还体现在对多因子模型解释能力的提升方面。对比回归 (1) 至 (5) 组的日均 Adj. R2（Adjusted R-squared，即各期截面回归调整后拟合优度），发现加入异象因子后多因子模型对未来收益率的解释力度有所提升，并且提升幅度和所估计的因子收益率正相关。此外，风格因子本身的收益率高度相似，并且显著性和系数规模与人们的认知一致，即 A 股市场最重要的因子包括市值、动量、残余波动率、成长、流动性等。为了各组回归之间容易比较，表 4.17 的五组 Fama-MacBeth 回归中保持各截面的股票样本范围一致（受限于“人工”因子的覆盖度）；附录表 D.4 则放宽样本范围限制，但是在因子检验方面的结论依然一致，即关联动量异象显著。

因子检验部分的结果表明，人工关联度牵引因子、W2V 相似度牵引因子在控制主要的风格因子后对未来收益具有显著的解释能力。在常见的股票特征风格以外，从基金持仓中的股票关联信息能够带来关联动量异象。

第五章　结论与展望

5.1　论文主要结论

本文用嵌入方法（Word2Vec），从 2013 年至 2023 年 A 股市场共同基金持股数据中提取股票的特征向量，并取向量余弦距离作为股票关联度，从而创新性地提取了基金持股中的股票关联信息。本文依次考察了（1）嵌入模型获得的关联度指标同股票风格特征、行业分类的关系，（2）关联度指标同未来资产收益相关性的关系，以及（3）关联股票之间的关联动量异象。在上述各项考察中，本文将嵌入表示的关联度指标同人工构造的关联指标进行对比。本文的主要结论包括：

第一，嵌入模型能有效利用基金持股数据提取公司数值特征，进而对股票关联进行定量刻画。利用基金持仓能够发现相似的股票，这些股票通常具有相似的行业和市值；嵌入表示得到的股票特征向量还对行业聚类有提示作用；基于 Word2Vec 的词嵌入模型对基金持股数据的提取较为有效，基于 BERT 的上下文嵌入则可能存在样本不足的问题。

第二，由基金持股提取的股票关联度受到公司行业分类和市值特征的影响。同行业内公司的平均关联度高于不同行业；公司市值越高，相互之间的平均关联度越低；嵌入模型提取的股票关联度受行业和市值分类的影响更大。基金持股中潜藏了公司风格、行业特征，嵌入模型在提取这些特征时更加高效。

第三，由基金持股提取的股票关联度和公司未来股价表现相关。公司嵌入表示的相似度越高，未来收益的相关性也越高；人工模型提取的关联度则无此现象。嵌入模型刻画的公司特征对公司未来股价的影响力更强，在提取影响公司未来股价的隐性公司特征时比人工模型更有效。

第四，由基金持仓提取的关联股票之间存在关联动量异象。关联股票的历史收益对中心股票的未来收益具有预测能力；关联股票过去的收益相对中心股票越高，中心股票未来的收益表现相应越高；据此构造的关联动量牵引因子，在 2014 年 7 月至 2023 年 6 月的回测区间内具有稳定的截面选股能力；该因子的预测能力部分来源于对行业和风格的暴露，中性化后因子 IC 下降，ICIR 上升；在控制行业和主要风格因子后，关联动量异象具有显著的因子收益，能够通过 Fama-MacBeth 检验；嵌入模型提取的因子比人工模型在各项因子评估、多空回测和因子检验指标上都更加显著。基金持仓中隐藏的公司关联信息能够被嵌入方法有效地提取，并且在刻画关联股票间的关联动量时，嵌入模型的表现好于人工模型。

5.2　研究不足与未来展望

本文尝试用嵌入方法提取基金持股中的股票特征，并围绕特征间的相似度考察了股票关联指标的应用场景。由于实验环境和时间限制，本文的研究还有以下不足，值得在未来继续探究：

第一，利用基金持仓数据不够充分和及时。忽略了第一和第三季度对前十大重仓股的披露信息，不够充分；嵌入表示和股票关联的更新时间固定为每年六月底和十二月底，与基金持股实际披露时间有 2 至 3 个月的间隔，不够及时。后续研究和实践中若希望更有效地从基金持股中提取公司特征，可考虑定期还原最新的基金持仓。

第二，对嵌入向量的解读和应用聚焦于股票关联，对嵌入向量本身的实际内涵缺乏探讨，包括行业、风格暴露，对收益和波动的解释能力等。此外，资产特征向量可以经由神经网络同其他公司特征连接（concatenate），这在现有文献中已有案例，值得未来研究时考虑。

第三，输入嵌入模型的基金持股信息预处理过于简单，只保留了持股金额排序后的位序，未能充分利用持股金额、持股金额相比上一季度的变化量、变化方向和加速度等；同一批训练样本中也未能体现最新报告期和历史报告期差异、不同规模的基金所具有的市场影响力差异等。训练数据过于简单或许也是 Word2Vec 模型能够取得最佳效果的原因。后续使用更复杂的嵌入模型时可考虑提取出更复杂的特征作为模型的共同输入。

参考文献

樊帅, 张千玉, 姜国华, 2017. 基金经理利用基本面信息选股吗?——来自基金持仓方面的证据[J]. 投资研究, 36(8): 65-81.

李斌, 雷印如, 2022. 中国公募基金挖掘了股票市场异象吗?[J]. 金融研究(9): 188-206.

刘莎莎, 刘玉珍, 唐涯, 2013. 信息优势、风险调整与基金业绩[J]. 管理世界(8): 67-76.

罗军, 季燕妮, 2024. 基于主动权益基金的投资注意力研究[R]. 广发证券.

石川, 刘洋溢, 连祥斌, 2020. 因子投资：方法与实践[M]. 电子工业出版社.

魏建榕, 2021a. 从基金持仓行为到股票关联网络[R]. 开源证券.

魏建榕, 2021b. 从北向资金持仓行为到股票关联网络[R]. 开源证券.

徐寅, 2020. 公募基金持仓因子全解析[R]. 兴业证券.

叶尔乐, 2023. 深度学习如何利用公募持仓网络优化选股效果？[R]. 民生证券.

Agarwal V, Gay G D, Ling L, 2014. Window Dressing in Mutual Funds[J]. The Review of Financial Studies, 27(11): 3133-3170.

Ang G, Lim E P, 2022. Learning knowledge-enriched company embeddings for investment management[C]//Proceedings of the Second ACM International Conference on AI in Finance. New York, NY, USA: Association for Computing Machinery: 1-9.

Antón M, Polk C, 2014. Connected Stocks[J]. The Journal of Finance, 69(3): 1099-1127.

Barberis N, Shleifer A, 2003. Style investing[J]. Journal of Financial Economics, 68(2): 161-199.

Basak S, Pavlova A, 2013. Asset Prices and Institutional Investors[J]. American Economic Review, 103(5): 1728-1758.

Berger T, Ramazan G, 2020. Volatility spillover along the supply chains: a network analysis on economic links[J]. Journal of Risk, 22(5).

Chen C, Zhao L, Bian J, et al., 2019. Investment Behaviors Can Tell What Inside: Exploring Stock Intrinsic Properties for Stock Trend Prediction[C]//Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. New York, NY, USA: Association for Computing Machinery: 2376-2384.

Chen K, Luo P, Liu L, et al., 2018. News, search and stock co-movement: Investigating information diffusion in the financial market[J]. Electronic Commerce Research and Applications, 28: 159-171.

Chen L, Pelger M, Zhu J, 2024. Deep Learning in Asset Pricing[J]. Management Science, 70(2): 714-750.

Chen Y, Kelly B T, Xiu D, 2022. Expected Returns and Large Language Models[A]. Rochester, NY.

Chordia T, Goyal A, Saretto A, 2020. Anomalies and false rejections[J]. The Review of Financial Studies, 33(5): 2134-2179.

Cochrane J H, 2011. Presidential Address: Discount Rates[J]. The Journal of Finance, 66(4): 1047-1108.

Cooper I, Ma L, Maio P, et al., 2021. Multifactor Models and Their Consistency with the APT[J]. Review of Asset Pricing Studies, 11(2).

Cremers K J M, Petajisto A, 2009. How Active Is Your Fund Manager? A New Measure That Predicts Performance[J]. The Review of Financial Studies, 22(9): 3329-3365.

Ding X, Zhang Y, Liu T, et al., 2015. Deep learning for event-driven stock prediction[C]//Proceedings of the 24th International Conference on Artificial Intelligence. Buenos Aires, Argentina: AAAI Press: 2327-2333.

Dolphin R, Smyth B, Dong R, 2023a. Stock Embeddings: Representation Learning for Financial Time Series[J]. Engineering Proceedings, 39(1): 30.

Dolphin R, Smyth B, Dong R, 2023b. A Machine Learning Approach to Industry Classification in Financial Markets[C]//Longo L, O'Reilly R. Artificial Intelligence and Cognitive Science. Cham: Springer Nature Switzerland: 81-94.

Drake M S, Jennings J, Roulstone D T, et al., 2017. The Comovement of Investor Attention[J]. Management Science, 63(9): 2847-2867.

Du K, Huang J, Louis H, et al., 2021. Monthly Mutual Fund Portfolio Disclosures and the Efficiency of Portfolio Firms' Investment Decisions[J]. Available at SSRN 3948716.

Du X, Tanaka-Ishii K, 2020. Stock Embeddings Acquired from News Articles and Price History, and an Application to Portfolio Optimization[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Online: Association for Computational Linguistics: 3353-3363.

Dumais S T, Furnas G W, Landauer T K, et al., 1988. Using latent semantic analysis to improve access to textual information[C]//Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. New York, NY, USA: Association for Computing Machinery: 281-285.

Fama E F, French K R, 2020. Comparing cross-section and time-series factor models[J]. The Review of Financial Studies, 33(5): 1891-1926.

Fama E F, MacBeth J D, 1973. Risk, Return, and Equilibrium: Empirical Tests[J]. Journal of Political Economy, 81(3): 607-636.

Frazzini A, Kabiller D, Pedersen L H, 2018. Buffett's Alpha[J]. Financial Analysts Journal, 74(4): 35-55.

Gabaix X, Koijen R S J, Richmond R, et al., 2023. Asset Embeddings[A]. Rochester, NY.

Gompers P A, Metrick A, 2001. Institutional Investors and Equity Prices*[J]. The Quarterly Journal of Economics, 116(1): 229-259.

Hameed A, Morck R, Shen J, et al., 2015. Information, Analysts, and Stock Return Comovement[J]. The Review of Financial Studies, 28(11): 3153-3187.

Harvey C R, Liu Y, 2022. Luck versus Skill in the Cross Section of Mutual Fund Returns: Reexamining the Evidence[J]. The Journal of Finance, 77(3): 1921-1966.

Harvey C R, Liu Y, Zhu H, 2016. … and the cross-section of expected returns[J]. The Review of Financial Studies, 29(1): 5-68.

Ivković Z, Weisbenner S, 2007. Information Diffusion Effects in Individual Investors' Common Stock Purchases: Covet Thy Neighbors' Investment Choices[J]. The Review of Financial Studies, 20(4): 1327-1357.

Jiang H, Verbeek M, Wang Y, 2014. Information Content When Mutual Funds Deviate from Benchmarks[J]. Management Science, 60(8): 2038-2053.

Kakushadze Z, 2016. 101 formulaic alphas[J]. Wilmott, 2016(84): 72-81.

Kelly B, Malamud S, Pedersen L H, 2023. Principal Portfolios[J]. The Journal of Finance, 78(1): 347-387.

Kelly B, Pruitt S, 2015. The three-pass regression filter: A new approach to forecasting using many predictors[J]. Journal of Econometrics, 186(2): 294-316.

Koch A, Ruenzi S, Starks L, 2016. Commonality in Liquidity: A Demand-Side Explanation[J]. The Review of Financial Studies, 29(8): 1943-1974.

Kozak S, Nagel S, Santosh S, 2020. Shrinking the cross-section[J]. Journal of Financial Economics, 135(2): 271-292.

Kumar A, Page J K, Spalt O G, 2013. Investor Sentiment and Return Comovements: Evidence from Stock Splits and Headquarters Changes*[J]. Review of Finance, 17(3): 921-953.

Lettau M, Pelger M, 2020. Factors That Fit the Time Series and Cross-Section of Stock Returns[J]. The Review of Financial Studies, 33(5).

Li L, Li Y, Wang X, et al., 2022. Hedge fund networks, information dissemination, and stock price comovement: Evidence from China[J]. International Review of Financial Analysis, 83: 102224.

Li Z, Yang D, Zhao L, et al., 2019. Individualized Indicator for All: Stock-wise Technical Indicator Optimization with Stock Embedding[C]//Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. Anchorage AK USA: ACM: 894-902.

Liu Q, Kusner M J, Blunsom P, 2020. A Survey on Contextual Embeddings[A]. arXiv.

Mikolov T, Chen K, Corrado G, et al., 2013a. Efficient Estimation of Word Representations in Vector Space[A]. arXiv.

Mikolov T, Sutskever I, Chen K, et al., 2013b. Distributed Representations of Words and Phrases and their Compositionality[C]//Advances in Neural Information Processing Systems: Vol. 26. Curran Associates, Inc.

Pennington J, Socher R, Manning C, 2014. GloVe: Global Vectors for Word Representation[C]//Moschitti A, Pang B, Daelemans W. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). Doha, Qatar: Association for Computational Linguistics: 1532-1543.

Sharpe W F, 1991. The Arithmetic of Active Management[J]. Financial Analysts Journal, 47(1): 7-9.

Takayanagi T, Sakaji H, Izumi K, 2022. SETN: Stock Embedding Enhanced with Textual and Network Information[C]//2022 IEEE International Conference on Big Data (Big Data). 2377-2382.

Wahal S, Yavuz M D, 2013. Style investing, comovement and return predictability[J]. Journal of Financial Economics, 107(1): 136-154.

Wermers R, Yao T, Zhao J, 2012. Forecasting Stock Returns Through an Efficient Aggregation of Mutual Fund Holdings[J]. The Review of Financial Studies, 25(12): 3490-3529.

Wu D, Wang Q, Olson D L, 2023. Industry classification based on supply chain network information using Graph Neural Networks[J]. Applied Soft Computing, 132: 109849.

Xin X, Yeung P E, Zhang Z, 2024. Wrong Kind of Transparency? Mutual Funds' Higher Reporting Frequency, Window Dressing, and Performance[J]. Journal of Accounting Research: 1475-679X.12527.

Xu W, Liu W, Wang L, et al., 2022. HIST: A Graph-based Framework for Stock Trend Forecasting via Mining Concept-Oriented Shared Information[A]. arXiv.

附录 A　其他风格特征与股票关联指标

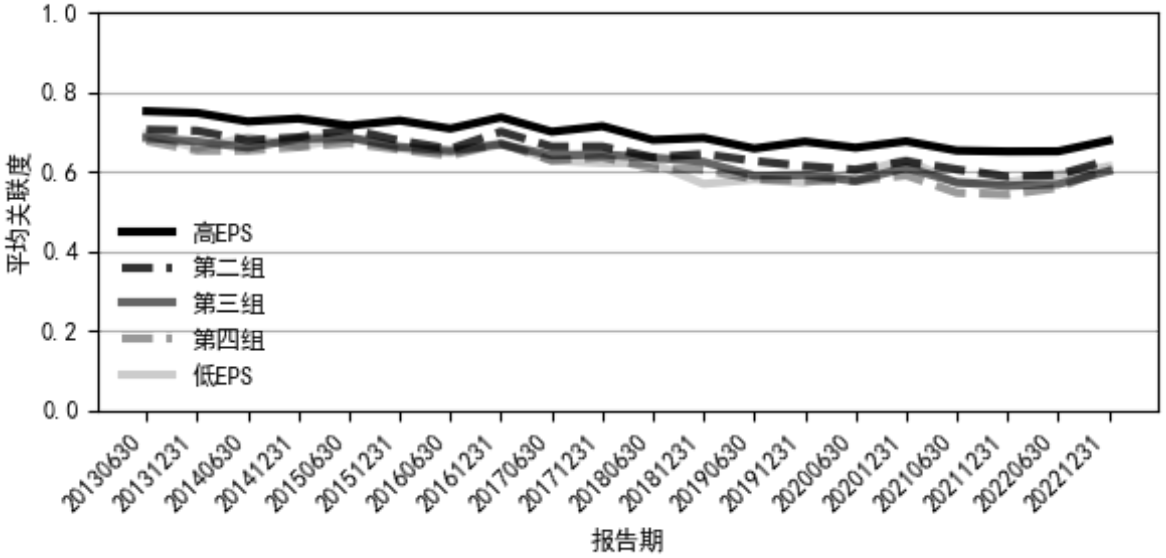


图 A.1　不同每股盈余（EPS）分组下人工关联度指标历史均值

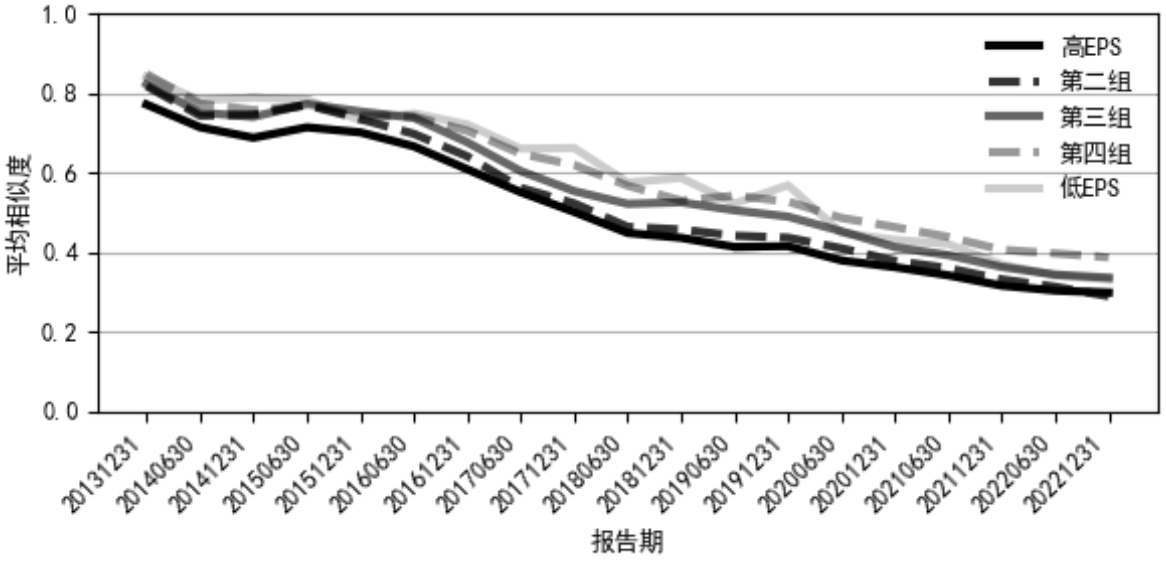


图 A.2　不同每股盈余（EPS）分组下 W2V 相似度指标历史均值

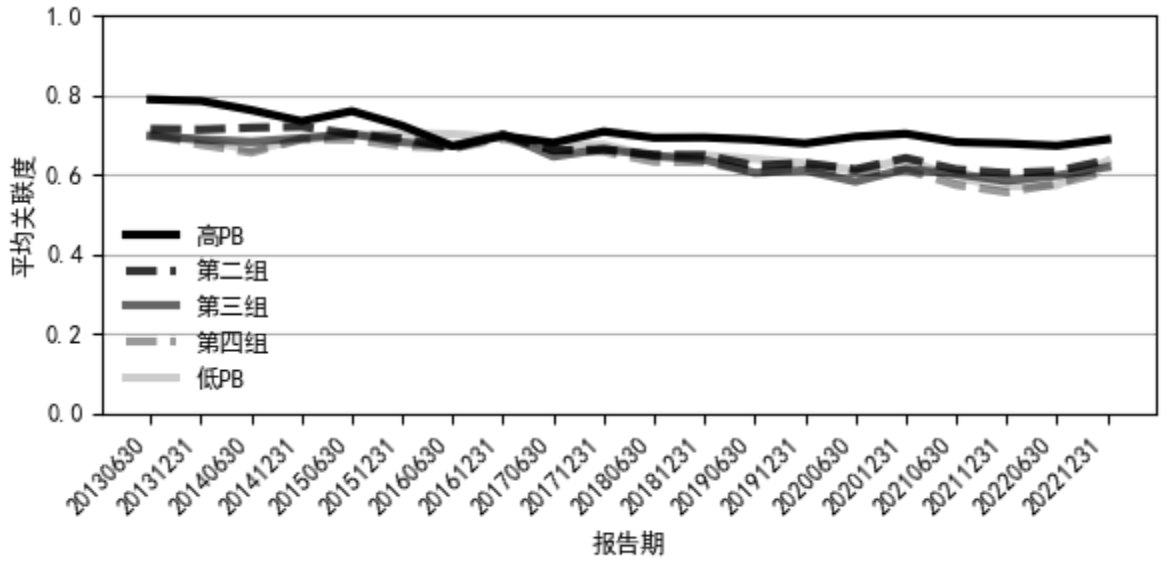


图 A.3　不同市净率（PB）分组下人工关联度指标历史均值

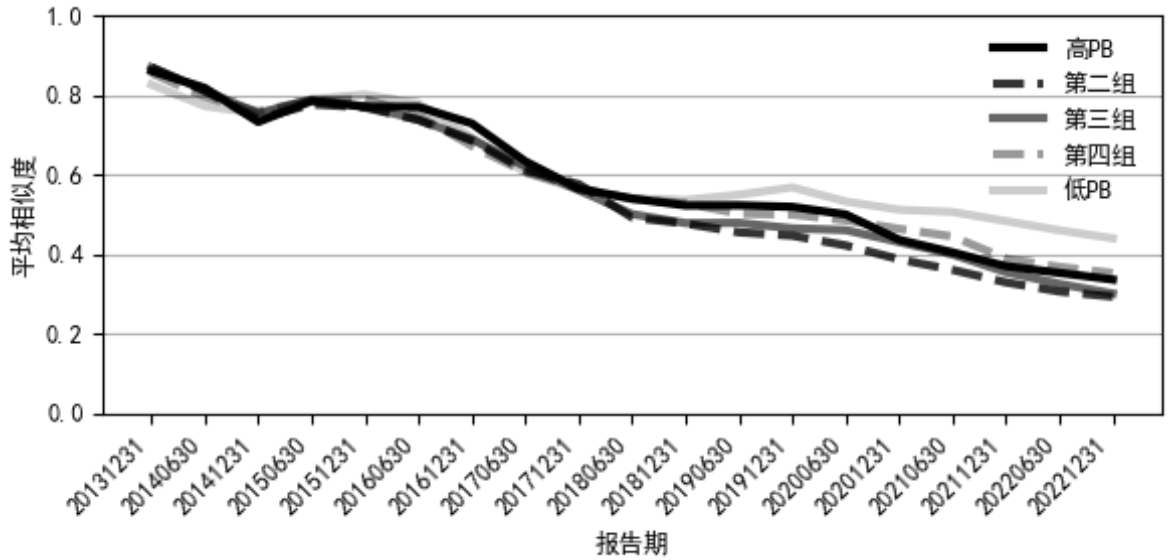


图 A.4　不同市净率（PB）分组下 W2V 相似度指标历史均值

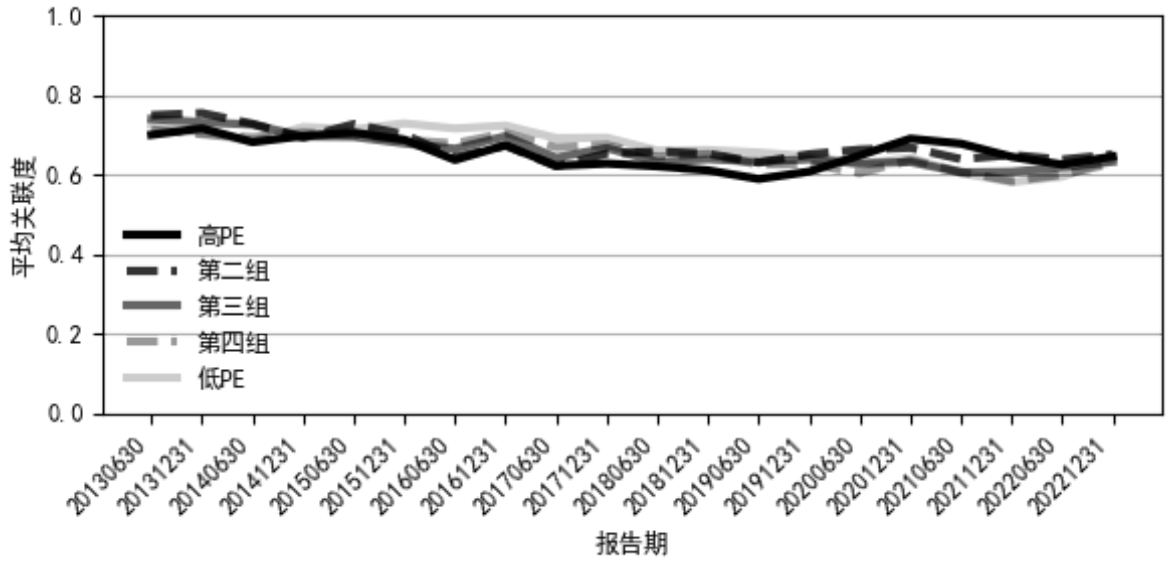


图 A.5　不同市盈率（PE）分组下人工关联度指标历史均值

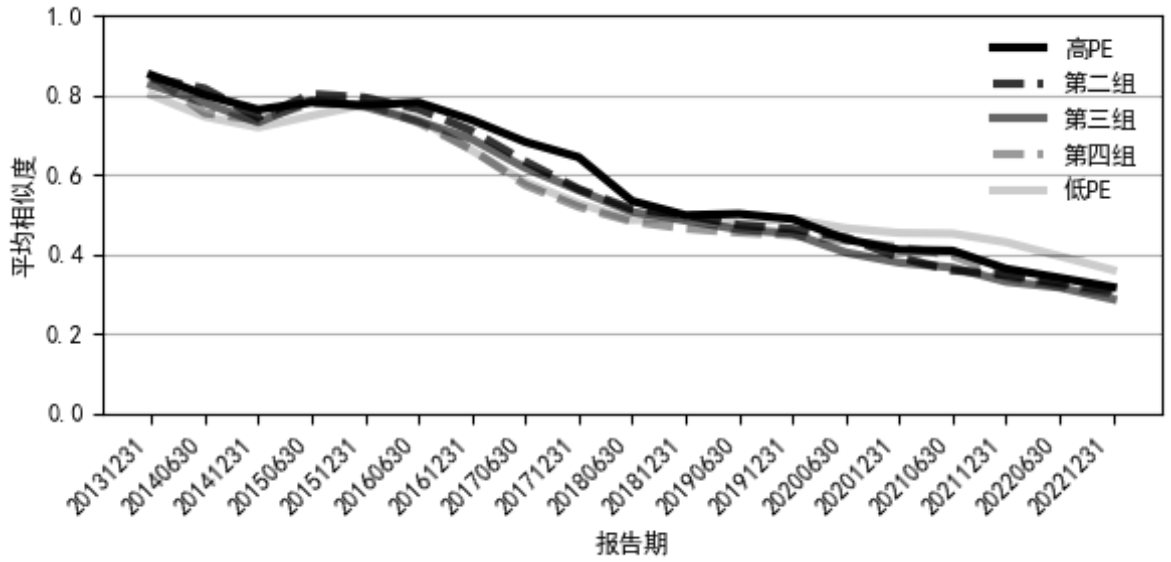


图 A.6　不同市盈率（PE）分组下 W2V 相似度指标历史均值

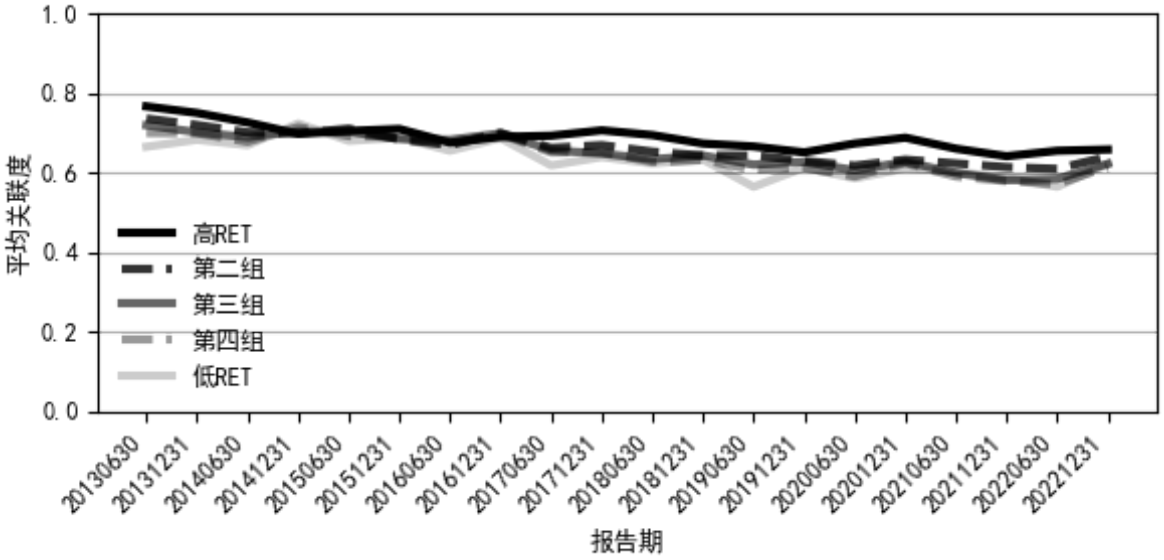


图 A.7　不同历史收益率（RET）分组下人工关联度指标历史均值

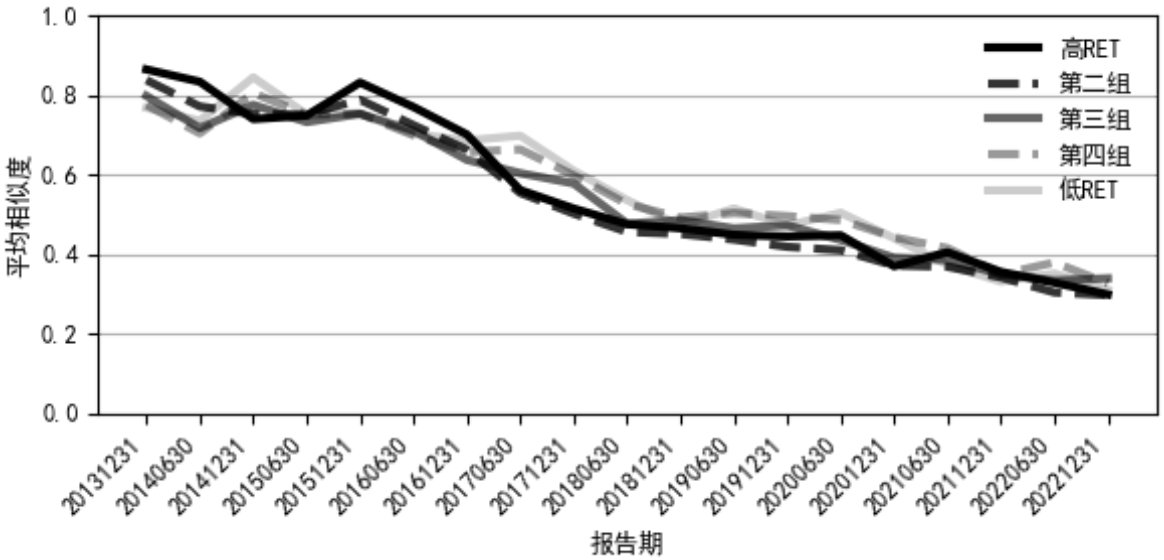


图 A.8　不同历史收益率（RET）分组下 W2V 相似度指标历史均值

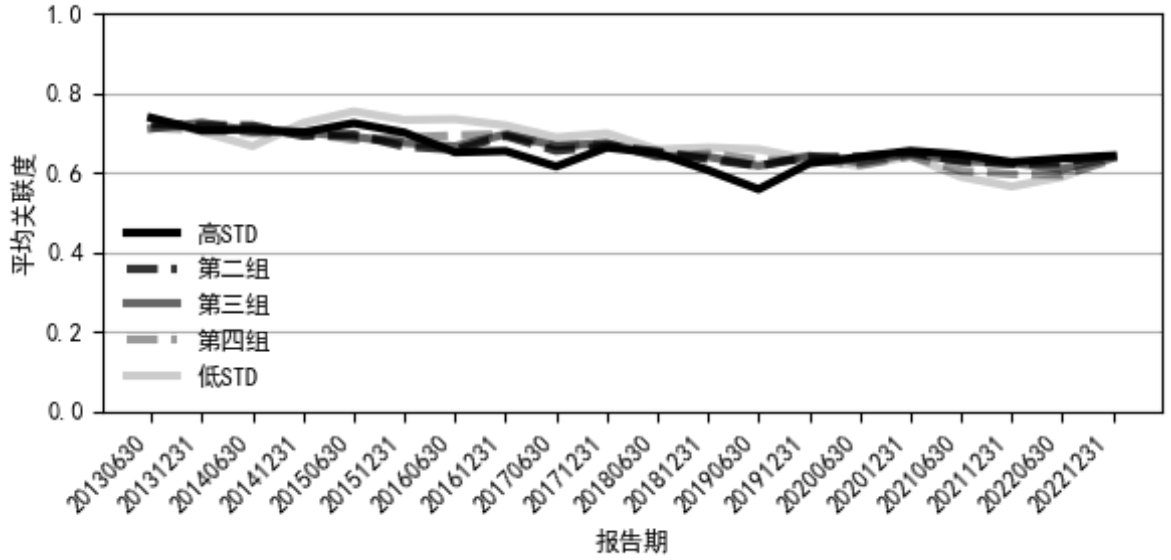


图 A.9　不同历史波动率（STD）分组下人工关联度指标历史均值

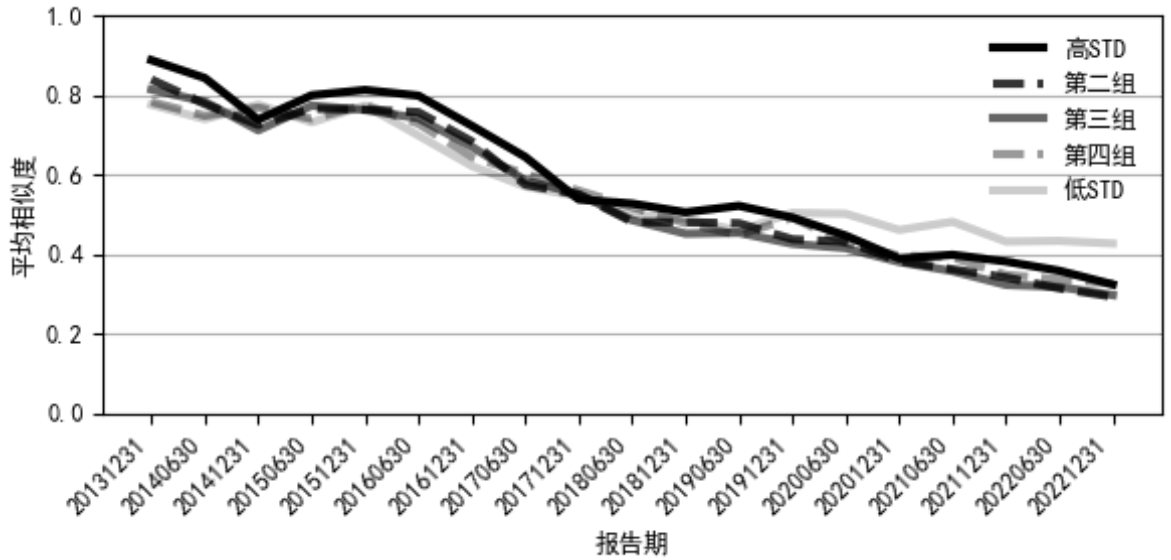
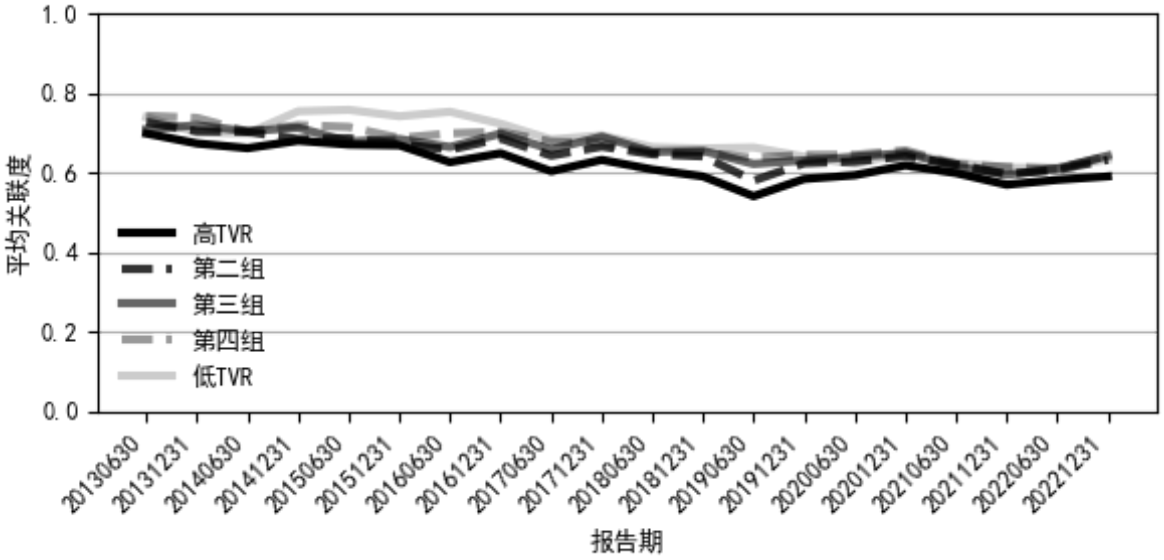


图 A.10　不同历史波动率（STD）分组下 W2V 相似度指标历史均值



A.11　不同换手率（TVR）分组下人工关联度指标历史均值

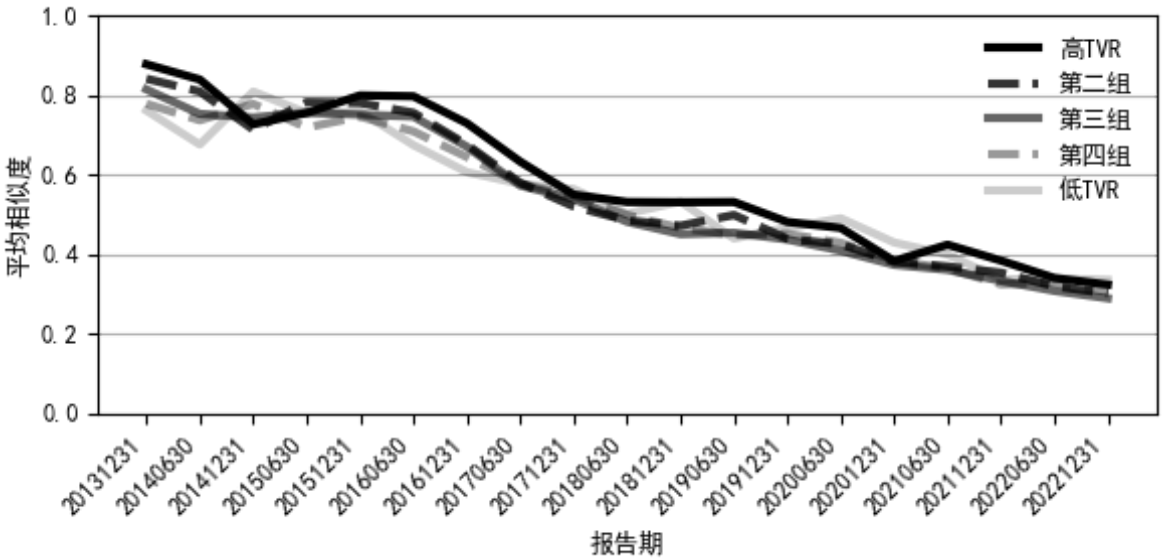


图 A.12　不同换手率（TVR）分组下 W2V 相似度指标历史均值

附录 B　半年度回测表现

表 B.1　人工关联度牵引因子 5 分组回测“多头组”历史表现

半年度	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
2014-H2	73.2%	3.82	1.0%	4.3%	64.8%	35.1%
2015-H1	123.5%	2.97	45.5%	33.1%	64.7%	43.9%
2015-H2	29.1%	0.37	23.5%	63.9%	56.0%	53.0%
2016-H1	-29.2%	-0.64	3.3%	48.2%	51.7%	45.2%
2016-H2	13.8%	0.89	7.5%	18.9%	57.3%	37.5%
2017-H1	-8.9%	-0.60	-8.4%	16.1%	53.8%	30.1%
2017-H2	-6.3%	-0.39	-11.9%	16.4%	53.6%	35.6%
2018-H1	-35.7%	-1.51	1.4%	36.5%	50.4%	35.7%
2018-H2	-33.5%	-1.31	-5.1%	54.5%	50.0%	37.2%
2019-H1	46.2%	1.78	-1.7%	39.4%	54.2%	40.2%
2019-H2	13.8%	0.77	3.7%	20.9%	56.3%	33.7%
2020-H1	47.8%	1.49	31.5%	13.9%	59.0%	37.7%
2020-H2	0.9%	0.04	-30.8%	17.4%	51.6%	42.1%
2021-H1	9.9%	0.59	-1.0%	24.8%	55.1%	37.0%
2021-H2	26.8%	1.60	20.2%	12.1%	60.0%	34.9%
2022-H1	-5.2%	-0.21	16.8%	32.0%	53.8%	34.5%
2022-H2	-5.4%	-0.26	13.4%	17.8%	52.0%	37.0%
2023-H1	20.0%	1.43	16.8%	9.1%	56.8%	32.6%
全历史	15.5%	0.51	6.8%	63.9%	55.6%	38.0%

表 B.2　人工关联度牵引因子 5 分组回测“多 - 空”组历史表现

半年度	年化收益	夏普率	年化超额收益	最大回撤	日度胜率
2014-H2	13.4%	2.38	-58.7%	1.6%	53.6%
2015-H1	25.5%	3.13	-52.5%	4.2%	64.7%
2015-H2	43.8%	3.20	38.2%	6.4%	64.8%
2016-H1	1.8%	0.30	34.4%	3.7%	45.0%
2016-H2	16.5%	5.08	10.3%	0.9%	62.9%
2017-H1	7.2%	2.10	7.7%	1.2%	55.5%
2017-H2	3.0%	0.79	-2.6%	5.2%	57.6%
2018-H1	3.9%	1.10	40.9%	4.5%	58.0%
2018-H2	5.8%	1.60	34.2%	2.5%	58.9%

2019-H1	11.6%	2.87	-36.3%	1.9%	58.5%
2019-H2	9.9%	3.62	-0.1%	1.3%	64.3%
2020-H1	18.1%	3.31	1.8%	1.4%	58.1%
2020-H2	-5.7%	-1.04	-37.3%	7.2%	46.8%
2021-H1	3.3%	0.43	-7.7%	11.6%	55.9%
2021-H2	16.5%	2.92	9.9%	4.7%	60.8%
2022-H1	28.7%	4.31	50.7%	3.3%	61.5%
2022-H2	13.5%	2.18	32.3%	1.8%	54.4%
2023-H1	9.0%	2.18	5.8%	3.5%	55.9%
全历史	12.5%	2.07	3.9%	11.6%	57.6%

表 B.3　W2V 相似度牵引因子 5 分组回测“多头组”历史表现

半年度	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
2014-H2	81.5%	4.26	9.3%	4.9%	64.8%	33.3%
2015-H1	113.4%	2.66	35.4%	34.4%	63.9%	41.6%
2015-H2	29.3%	0.38	23.7%	64.3%	55.2%	53.7%
2016-H1	-39.8%	-0.86	-7.2%	49.7%	51.7%	44.1%
2016-H2	10.7%	0.72	4.5%	25.6%	57.3%	35.5%
2017-H1	-3.0%	-0.21	-2.5%	14.8%	58.0%	26.4%
2017-H2	0.5%	0.04	-5.0%	12.0%	54.4%	34.9%
2018-H1	-12.4%	-0.57	24.7%	18.8%	54.6%	28.7%
2018-H2	-37.9%	-1.48	-9.6%	39.2%	50.8%	38.9%
2019-H1	45.6%	1.76	-2.4%	34.2%	55.9%	40.6%
2019-H2	16.5%	0.94	6.5%	20.1%	57.9%	28.4%
2020-H1	54.0%	1.70	37.7%	13.9%	59.8%	34.5%
2020-H2	4.5%	0.21	-27.1%	14.9%	54.0%	45.4%
2021-H1	17.0%	1.11	6.1%	15.6%	53.4%	37.5%
2021-H2	30.8%	1.84	24.2%	10.8%	62.4%	31.0%
2022-H1	-3.4%	-0.13	18.7%	29.1%	53.8%	29.0%
2022-H2	-4.6%	-0.22	14.2%	18.2%	54.4%	35.4%
2023-H1	21.7%	1.52	18.5%	9.5%	51.7%	33.4%
全历史	17.9%	0.60	9.2%	64.3%	56.3%	36.3%

表 B.4　人工关联度牵引因子 5 分组回测“多 - 空”组历史表现

半年度	年化收益	夏普率	年化超额收益	最大回撤	日度胜率
2014-H2	22.9%	3.90	-49.3%	1.5%	55.2%

2015-H1	19.7%	1.88	-58.3%	7.9%	59.7%
2015-H2	41.9%	3.46	36.3%	5.3%	64.0%
2016-H1	-3.2%	-0.39	29.4%	6.4%	46.7%
2016-H2	16.7%	3.58	10.4%	3.9%	55.6%
2017-H1	11.5%	2.27	12.0%	2.2%	58.8%
2017-H2	11.8%	2.51	6.2%	1.7%	56.8%
2018-H1	25.6%	4.11	62.7%	2.5%	63.0%
2018-H2	-0.5%	-0.11	27.8%	3.1%	49.2%
2019-H1	10.2%	1.96	-37.7%	4.9%	50.8%
2019-H2	15.6%	4.64	5.6%	1.0%	59.5%
2020-H1	30.8%	5.51	14.5%	1.2%	69.2%
2020-H2	-3.9%	-0.64	-35.5%	5.6%	47.6%
2021-H1	12.1%	1.83	1.2%	5.5%	54.2%
2021-H2	21.7%	2.91	15.1%	4.0%	61.6%
2022-H1	31.5%	3.72	53.5%	4.2%	58.1%
2022-H2	15.2%	1.85	34.0%	2.2%	56.8%
2023-H1	14.1%	2.73	10.9%	3.2%	55.1%
全历史	16.3%	2.35	7.6%	7.9%	56.8%

附录 C　行业、市值中性化后因子表现

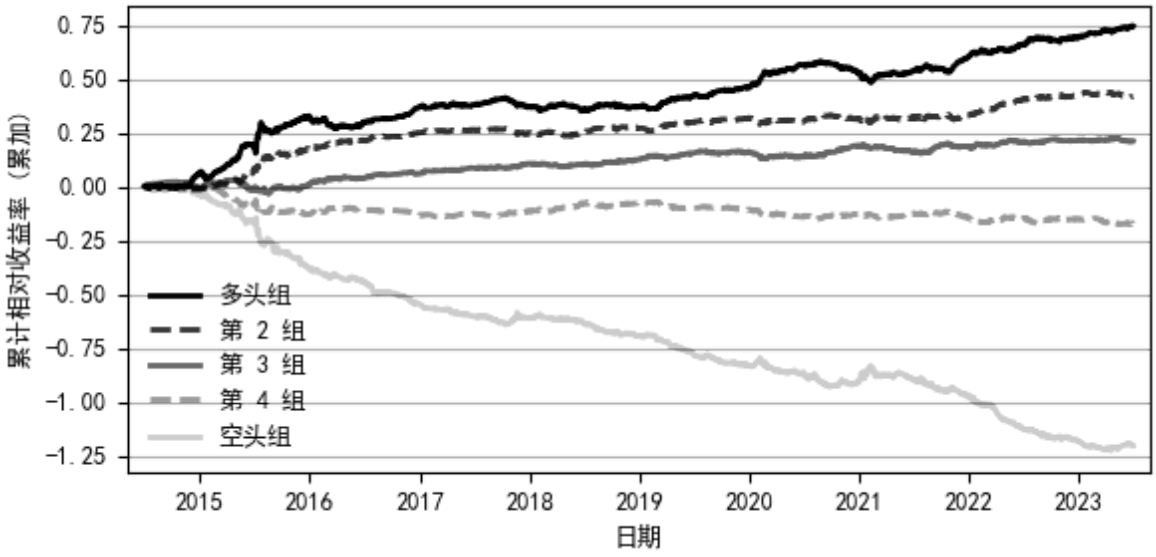


图 C.1　人工关联度牵引因子（行业中性）5 分组收益分化情况

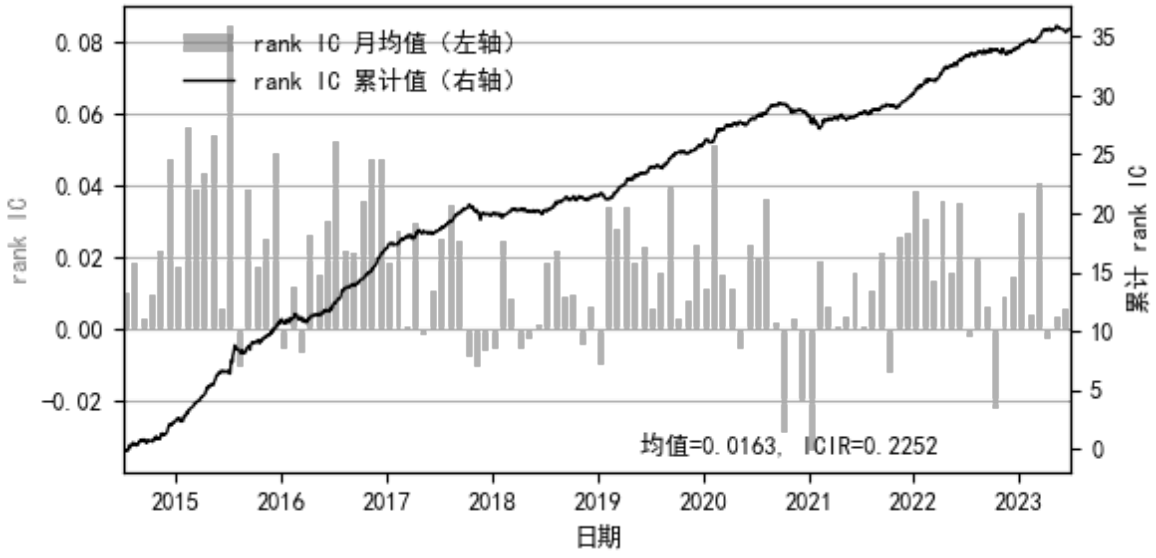


图 C.2　人工关联度牵引因子（行业中性）rank IC 和累计 rank IC

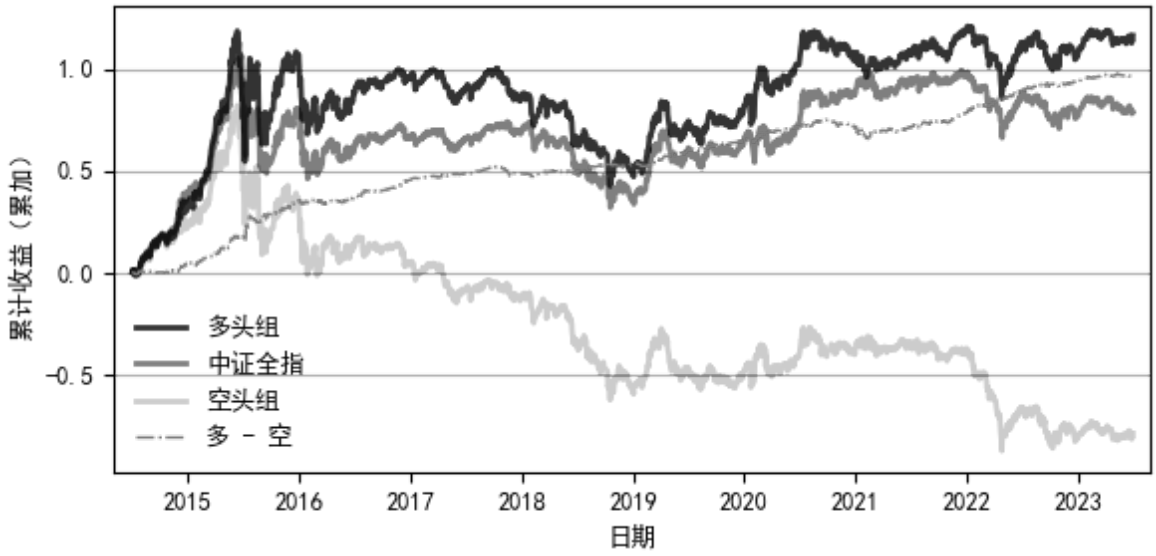


图 C.3　人工关联度牵引因子（行业中性）5 分组多头/多空策略回测表现

表 C.1　人工关联度牵引因子（行业中性）5 分组回测“多头组”历史表现

半年度	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
2014-H2	63.4%	3.29	-8.8%	5.8%	60.0%	34.9%
2015-H1	119.8%	2.89	41.8%	34.2%	66.4%	41.6%
2015-H2	24.6%	0.31	19.1%	64.0%	56.8%	53.3%
2016-H1	-32.5%	-0.70	0.0%	50.6%	53.3%	44.1%
2016-H2	12.6%	0.79	6.3%	20.5%	58.9%	37.4%
2017-H1	-10.0%	-0.65	-9.5%	16.5%	52.9%	30.2%
2017-H2	-8.7%	-0.53	-14.3%	16.1%	52.0%	34.9%
2018-H1	-38.8%	-1.66	-1.8%	37.9%	48.7%	35.3%
2018-H2	-32.8%	-1.28	-4.5%	55.5%	48.4%	37.1%
2019-H1	48.2%	1.85	0.2%	40.8%	53.4%	39.5%
2019-H2	13.7%	0.77	3.7%	21.0%	54.8%	33.8%
2020-H1	46.9%	1.44	30.7%	14.1%	57.3%	38.0%
2020-H2	6.0%	0.27	-25.7%	14.5%	53.2%	41.8%
2021-H1	5.4%	0.31	-5.6%	22.7%	53.4%	36.2%
2021-H2	21.3%	1.36	14.7%	13.4%	59.2%	33.9%
2022-H1	-16.0%	-0.63	6.0%	35.5%	53.0%	34.1%
2022-H2	-12.3%	-0.62	6.5%	21.6%	52.0%	36.8%
2023-H1	21.0%	1.52	17.8%	12.5%	59.3%	32.9%
全历史	12.8%	0.42	4.1%	64.0%	55.2%	37.6%

表 C.2　人工关联度牵引因子（行业中性）5 分组回测“多 - 空”组历史表现

半年度	年化收益	夏普率	年化超额收益	最大回撤	日度胜率
2014-H2	8.9%	2.39	-63.3%	1.3%	52.8%
2015-H1	25.2%	4.79	-52.8%	2.2%	62.2%
2015-H2	35.1%	3.60	29.5%	3.3%	63.2%
2016-H1	3.9%	0.93	36.5%	2.5%	54.2%
2016-H2	16.1%	6.04	9.8%	0.6%	66.9%
2017-H1	6.9%	2.46	7.4%	1.0%	56.3%
2017-H2	-0.2%	-0.06	-5.8%	4.9%	52.0%
2018-H1	1.3%	0.40	38.3%	5.4%	51.3%
2018-H2	6.9%	2.40	35.2%	3.2%	58.9%
2019-H1	14.0%	4.38	-34.0%	1.4%	59.3%
2019-H2	8.5%	3.66	-1.5%	0.8%	61.1%
2020-H1	14.0%	3.42	-2.2%	1.2%	63.2%
2020-H2	0.6%	0.14	-31.1%	3.8%	53.2%
2021-H1	-0.3%	-0.05	-11.2%	9.2%	49.2%
2021-H2	13.6%	3.39	7.0%	3.8%	63.2%
2022-H1	22.5%	5.49	44.5%	1.0%	65.0%
2022-H2	7.8%	2.02	26.5%	1.3%	52.0%
2023-H1	7.9%	3.25	4.7%	1.2%	63.6%
全历史	10.7%	2.44	2.0%	9.2%	58.2%

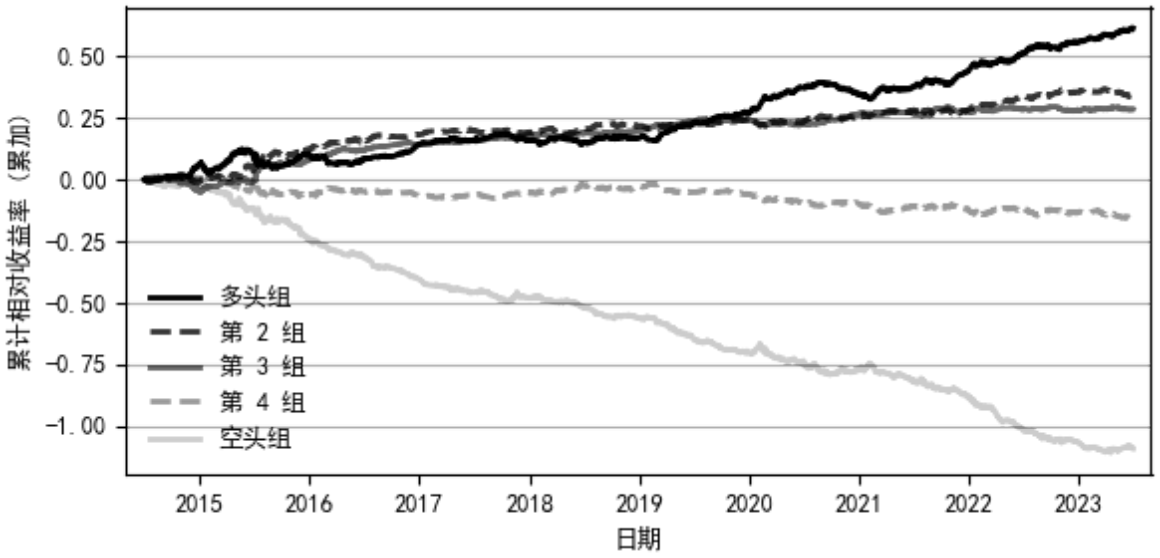


图 C.4　人工关联度牵引因子（行业、市值中性）5 分组收益分化情况

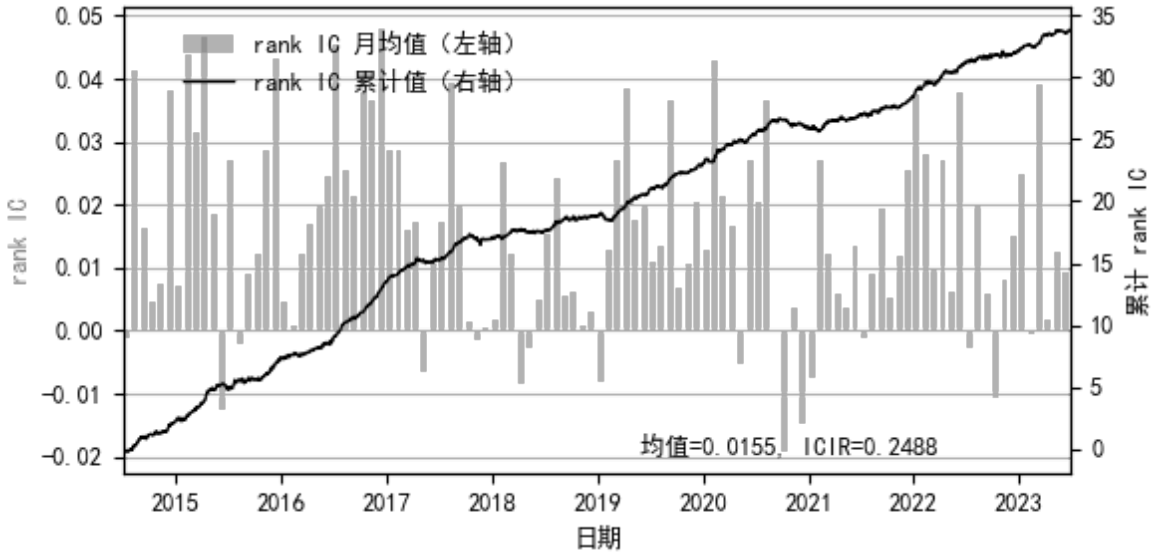


图 C.5　人工关联度牵引因子（行业、市值中性）rank IC 和累计 rank IC

图 C.6　人工关联度牵引因子（行业、市值中性）5 分组多头/多空策略回测表现

表 C.3　人工关联度牵引因子（行业、市值中性）5 分组回测“多头组”历史表现

半年度	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
2014-H2	63.4%	3.22	-8.7%	6.0%	60.0%	34.9%
2015-H1	101.3%	2.49	23.3%	35.6%	63.9%	41.1%
2015-H2	-0.2%	-0.00	-5.8%	68.4%	57.6%	51.9%
2016-H1	-31.4%	-0.69	1.2%	64.0%	54.2%	44.6%
2016-H2	10.4%	0.64	4.1%	20.2%	57.3%	36.9%
2017-H1	-8.3%	-0.54	-7.8%	16.6%	51.3%	30.7%
2017-H2	-7.1%	-0.44	-12.7%	14.9%	52.0%	35.3%
2018-H1	-38.7%	-1.66	-1.6%	36.3%	49.6%	34.9%
2018-H2	-31.2%	-1.21	-2.9%	53.6%	48.4%	37.4%
2019-H1	49.4%	1.92	1.4%	39.9%	52.5%	40.1%
2019-H2	14.7%	0.84	4.7%	19.5%	54.0%	34.0%
2020-H1	46.2%	1.43	30.0%	14.0%	56.4%	38.1%
2020-H2	8.4%	0.37	-23.3%	14.1%	53.2%	43.2%
2021-H1	9.2%	0.57	-1.7%	20.0%	55.9%	35.6%
2021-H2	21.5%	1.37	14.9%	10.9%	60.0%	33.6%
2022-H1	-13.4%	-0.53	8.6%	35.1%	55.6%	33.6%
2022-H2	-9.6%	-0.48	9.2%	20.1%	51.2%	36.6%
2023-H1	21.7%	1.58	18.5%	10.4%	56.8%	32.4%
全历史	11.3%	0.38	2.7%	68.4%	55.0%	37.5%

表 C.4　人工关联度牵引因子（行业、市值中性）5 分组回测“多 - 空”组历史表现

半年度	年化收益	夏普率	年化超额收益	最大回撤	日度胜率
2014-H2	6.8%	2.16	-65.4%	1.8%	55.2%
2015-H1	13.8%	2.44	-64.2%	1.9%	52.1%
2015-H2	13.6%	2.12	8.0%	3.8%	62.4%
2016-H1	5.3%	1.56	37.9%	1.1%	55.0%
2016-H2	13.8%	6.82	7.6%	0.6%	68.5%
2017-H1	7.5%	3.17	8.0%	0.9%	62.2%
2017-H2	2.0%	0.67	-3.6%	3.4%	49.6%
2018-H1	2.9%	1.06	40.0%	3.3%	54.6%
2018-H2	6.0%	2.19	34.3%	1.3%	61.3%
2019-H1	14.3%	4.73	-33.6%	1.4%	58.5%
2019-H2	9.5%	4.34	-0.6%	0.7%	61.1%
2020-H1	14.0%	3.49	-2.2%	1.3%	65.0%
2020-H2	1.1%	0.27	-30.6%	3.4%	50.8%
2021-H1	7.2%	1.84	-3.7%	5.4%	57.6%
2021-H2	12.4%	3.11	5.8%	1.4%	56.8%
2022-H1	21.9%	5.63	43.9%	1.1%	64.1%
2022-H2	8.6%	2.34	27.4%	0.9%	55.2%
2023-H1	8.6%	3.72	5.4%	1.1%	61.0%
全历史	9.4%	2.56	0.7%	5.4%	58.4%

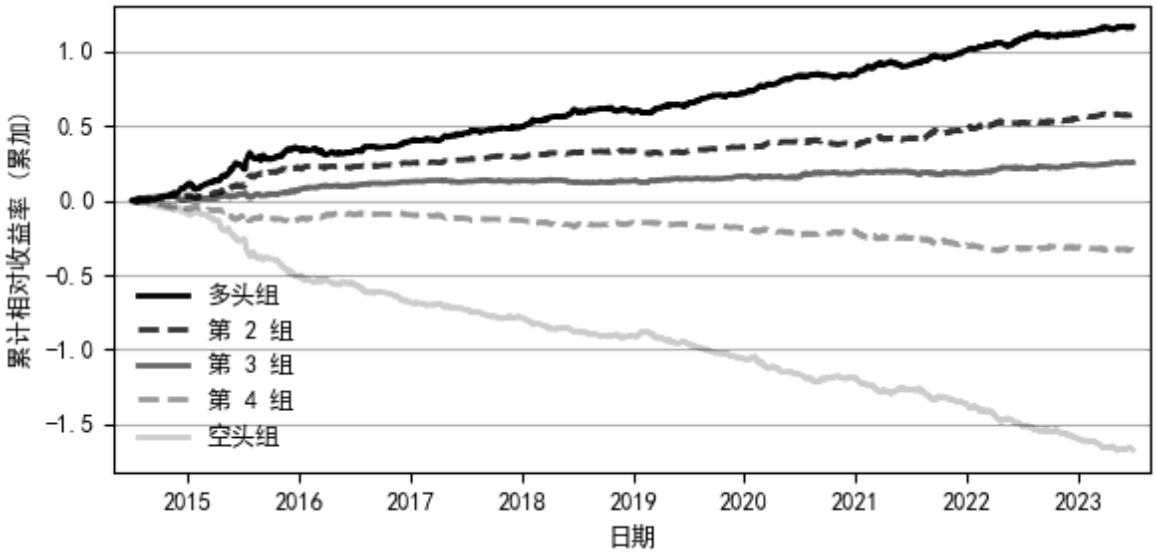


图 C.7　W2V 相似度牵引因子（行业中性）5 分组收益分化情况

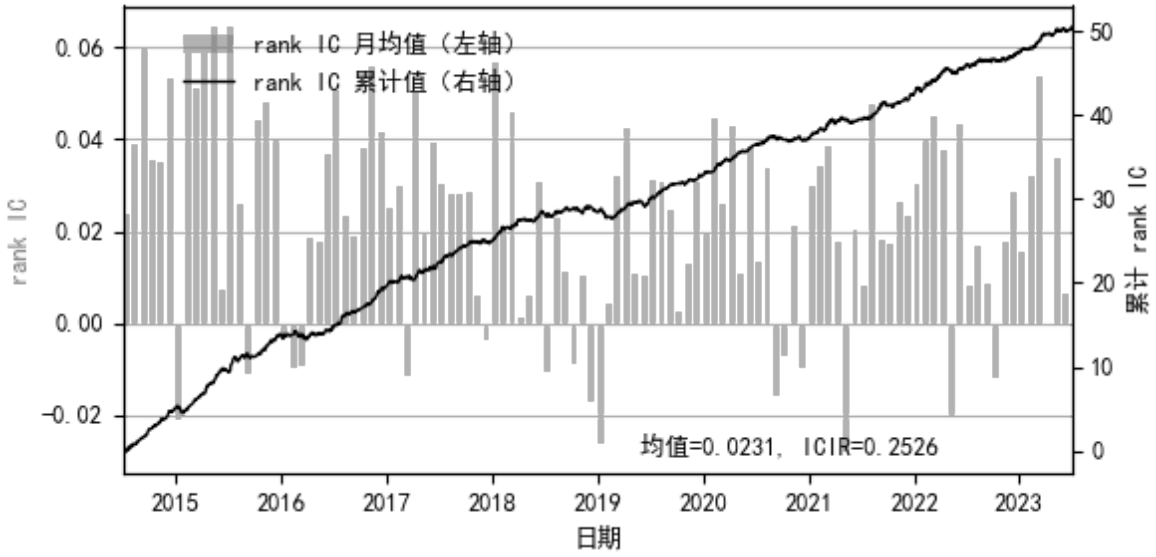


图 C.8　W2V 相似度牵引因子（行业中性）rank IC 和累计 rank IC

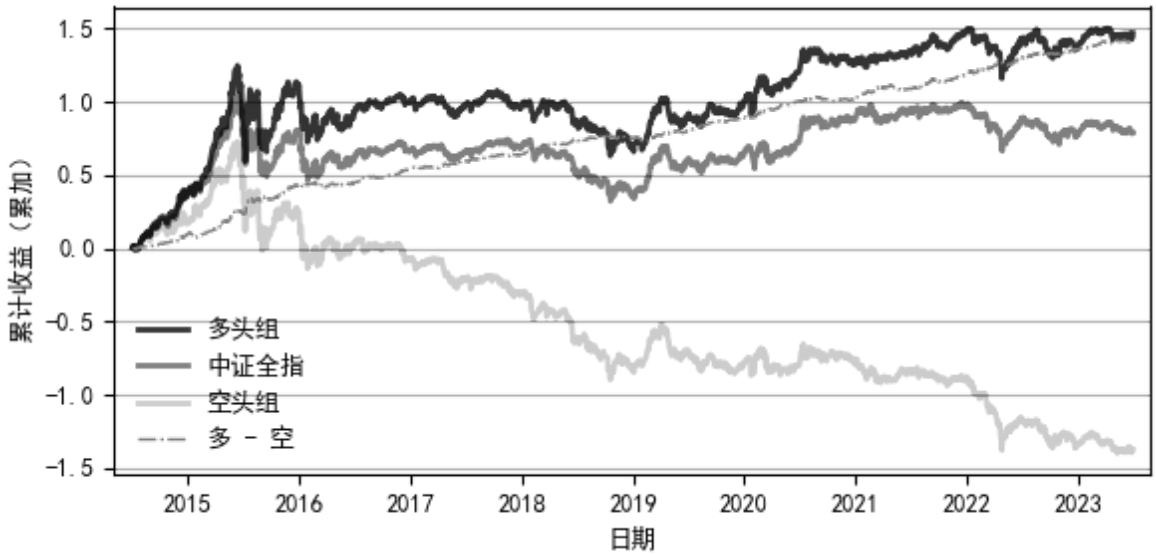


图 C.9　W2V 相似度牵引因子（行业中性）5 分组多头/多空策略回测表现

表 C.5　W2V 相似度牵引因子（行业中性）5 分组回测“多头组”历史表现

半年度	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
2014-H2	72.4%	3.70	0.2%	5.3%	64.0%	34.6%
2015-H1	118.8%	2.81	40.8%	35.4%	65.5%	41.1%
2015-H2	25.1%	0.33	19.5%	65.4%	56.8%	53.2%
2016-H1	-32.4%	-0.70	0.2%	52.1%	50.8%	44.8%
2016-H2	12.2%	0.80	5.9%	21.1%	58.1%	35.9%
2017-H1	-3.7%	-0.25	-3.2%	14.8%	54.6%	26.5%
2017-H2	-2.4%	-0.17	-8.0%	12.0%	54.4%	35.3%
2018-H1	-18.9%	-0.86	18.2%	22.4%	52.9%	29.2%
2018-H2	-36.7%	-1.42	-8.3%	42.8%	52.4%	38.5%
2019-H1	44.8%	1.70	-3.1%	35.1%	55.1%	40.0%
2019-H2	16.8%	0.96	6.7%	19.3%	55.6%	29.3%
2020-H1	45.8%	1.45	29.6%	14.1%	59.8%	35.2%
2020-H2	11.3%	0.53	-20.4%	11.1%	54.0%	44.8%
2021-H1	14.3%	0.92	3.4%	12.7%	53.4%	37.1%
2021-H2	27.7%	1.73	21.1%	9.9%	60.0%	31.5%
2022-H1	-12.6%	-0.48	9.4%	34.0%	56.4%	27.6%
2022-H2	-12.1%	-0.60	6.6%	20.2%	51.2%	34.4%
2023-H1	20.7%	1.44	17.5%	11.6%	57.6%	32.4%
全历史	16.1%	0.53	7.4%	65.4%	56.3%	36.2%

表 C.6　W2V 相似度牵引因子（行业中性）5 分组回测“多 - 空”组历史表现

半年度	年化收益	夏普率	年化超额收益	最大回撤	日度胜率
2014-H2	18.8%	4.99	-53.4%	0.8%	62.4%
2015-H1	28.1%	4.36	-49.9%	3.7%	61.3%
2015-H2	37.4%	4.31	31.8%	3.2%	68.0%
2016-H1	4.0%	0.71	36.6%	3.2%	51.7%
2016-H2	16.9%	4.58	10.6%	1.1%	58.1%
2017-H1	11.3%	2.85	11.8%	1.5%	63.0%
2017-H2	8.2%	2.31	2.6%	1.4%	52.8%
2018-H1	20.9%	5.22	58.0%	1.5%	62.2%
2018-H2	2.2%	0.58	30.6%	2.3%	53.2%
2019-H1	10.4%	2.26	-37.5%	3.7%	52.5%
2019-H2	16.1%	6.24	6.1%	0.5%	66.7%
2020-H1	23.0%	5.26	6.8%	1.3%	65.8%
2020-H2	3.6%	0.87	-28.0%	3.1%	51.6%
2021-H1	13.7%	2.76	2.8%	3.7%	56.8%
2021-H2	18.3%	3.59	11.7%	2.6%	66.4%
2022-H1	24.2%	4.16	46.2%	2.2%	60.7%
2022-H2	10.8%	2.19	29.5%	1.3%	57.6%
2023-H1	12.7%	3.76	9.5%	1.4%	61.0%
全历史	15.5%	3.21	6.9%	3.7%	59.5%

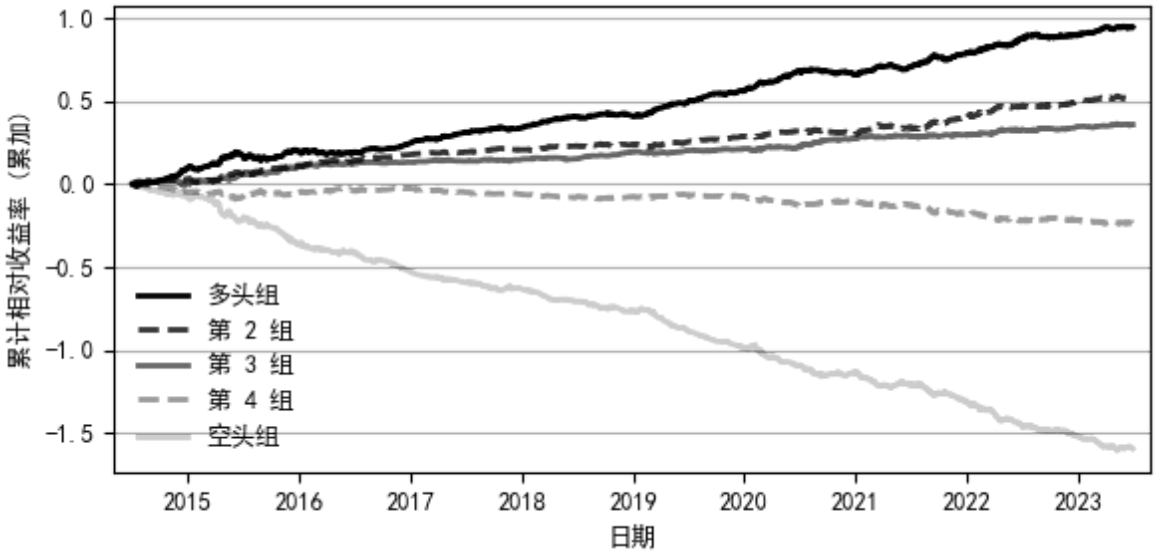


图 C.10　W2V 相似度牵引因子（行业、市值中性）5 分组收益分化情况

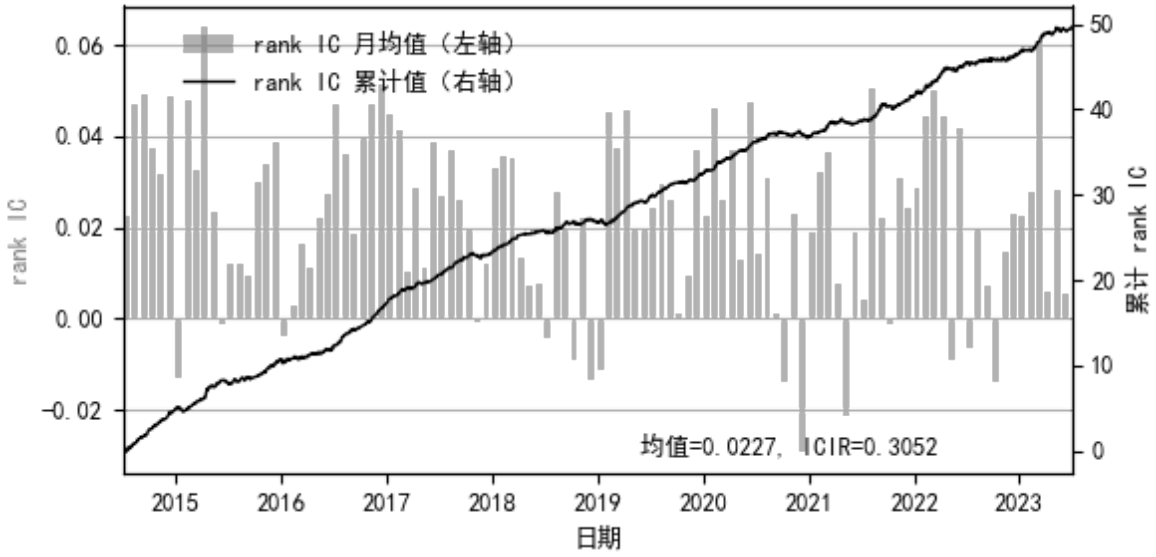


图 C.11　W2V 相似度牵引因子（行业、市值中性）rank IC 和累计 rank IC

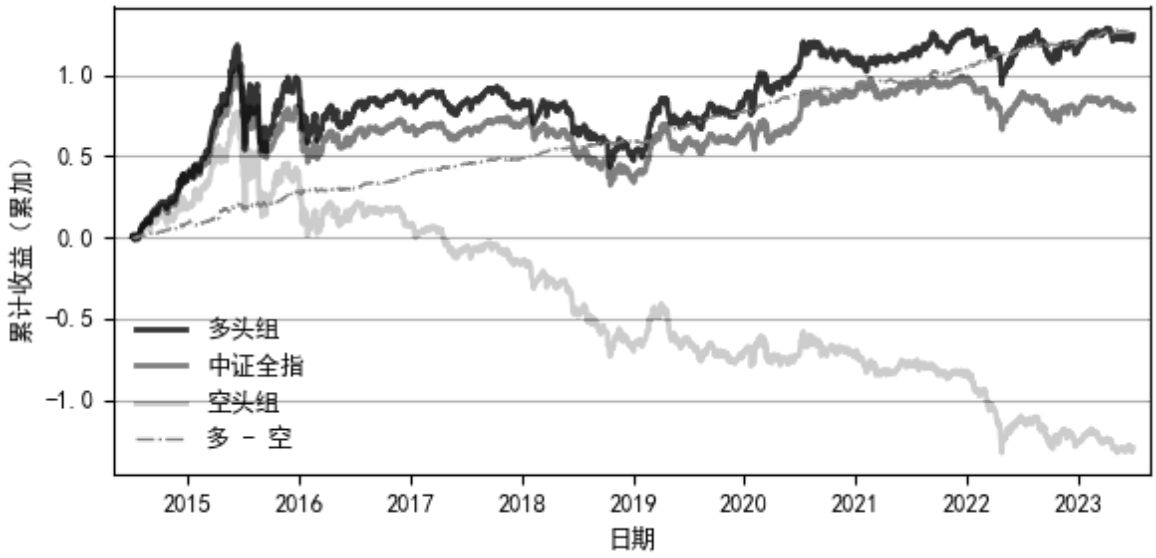


图 C.12　W2V 相似度牵引因子（行业、市值中性）5 分组多头/多空策略回测表现

表 C.7　W2V 相似度牵引因子（行业、市值中性）5 分组回测“多头组”历史表现

半年度	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
2014-H2	71.0%	3.70	-1.2%	5.2%	65.6%	35.2%
2015-H1	109.5%	2.67	31.6%	34.7%	65.5%	39.9%
2015-H2	6.7%	0.09	1.1%	65.8%	56.0%	51.3%
2016-H1	-30.7%	-0.66	1.9%	60.2%	51.7%	44.1%
2016-H2	10.0%	0.63	3.8%	20.1%	58.9%	36.6%
2017-H1	-1.4%	-0.09	-0.9%	14.9%	51.3%	27.7%
2017-H2	-4.8%	-0.31	-10.3%	13.1%	53.6%	36.2%
2018-H1	-27.4%	-1.21	9.7%	28.5%	50.4%	30.2%
2018-H2	-35.1%	-1.37	-6.7%	48.3%	51.6%	38.6%
2019-H1	51.9%	1.99	3.9%	37.2%	53.4%	41.2%
2019-H2	15.9%	0.90	5.9%	18.9%	54.8%	29.3%
2020-H1	46.1%	1.45	29.9%	14.1%	59.8%	34.4%
2020-H2	5.7%	0.26	-25.9%	13.8%	54.8%	44.4%
2021-H1	10.5%	0.65	-0.5%	18.5%	55.1%	36.2%
2021-H2	25.1%	1.53	18.5%	11.7%	57.6%	32.4%
2022-H1	-10.5%	-0.40	11.6%	33.4%	54.7%	27.2%
2022-H2	-15.2%	-0.75	3.6%	20.2%	52.0%	34.7%
2023-H1	21.2%	1.49	18.0%	10.9%	58.5%	32.5%
全历史	13.7%	0.46	5.0%	65.8%	55.8%	36.3%

表 C.8　W2V 相似度牵引因子（行业、市值中性）5 分组回测“多 - 空”组历史表现

半年度	年化收益	夏普率	年化超额收益	最大回撤	日度胜率
2014-H2	17.5%	4.86	-54.7%	0.7%	56.8%
2015-H1	18.6%	3.07	-59.4%	2.5%	52.9%
2015-H2	19.2%	3.63	13.6%	3.1%	64.0%
2016-H1	4.6%	1.02	37.2%	1.8%	57.5%
2016-H2	15.6%	6.20	9.3%	1.0%	64.5%
2017-H1	13.4%	5.40	13.9%	0.7%	65.5%
2017-H2	6.0%	1.92	0.5%	2.5%	52.8%
2018-H1	14.6%	6.64	51.7%	0.6%	63.9%
2018-H2	5.4%	1.68	33.8%	1.6%	58.1%
2019-H1	21.0%	5.53	-26.9%	1.8%	66.9%
2019-H2	15.9%	7.24	5.9%	0.6%	68.3%
2020-H1	23.4%	5.70	7.2%	1.3%	68.4%
2020-H2	2.4%	0.52	-29.3%	3.0%	46.8%
2021-H1	11.3%	2.81	0.4%	3.7%	55.1%
2021-H2	17.5%	3.85	10.9%	2.2%	61.6%
2022-H1	25.5%	4.90	47.6%	1.4%	62.4%
2022-H2	6.9%	1.51	25.7%	1.2%	53.6%
2023-H1	12.9%	3.80	9.7%	1.6%	61.0%
全历史	13.9%	3.47	5.3%	3.7%	60.0%

附录 D　因子检验相关结果

表 D.1　每日因子方差膨胀系数（VIF）分布（2015.1-2023.6）

因子	截面数	均值	标准差	最小值	25%分位数	中位数	75%分位数	最大值
beta	2064	1.71	0.46	1.07	1.39	1.6	1.9	6.54
市值	2064	1.76	0.27	1.36	1.57	1.75	1.88	4.41
动量	2064	1.49	0.3	1.02	1.31	1.43	1.61	5.03
残余波动率	2064	2.17	0.35	1.35	1.93	2.13	2.34	5.7
价值	2064	1.66	0.29	1.17	1.47	1.63	1.82	2.66
非线性市值	2064	1.39	0.31	1.04	1.17	1.35	1.55	4.53
盈利	2064	1.52	0.13	1.3	1.43	1.51	1.61	2.18
流动性	2064	1.96	0.31	1.08	1.77	1.96	2.16	3.41
杠杆	2064	1.26	0.06	1.14	1.2	1.26	1.29	1.58
成长	2064	1.15	0.09	1.02	1.08	1.12	1.21	1.48
截面最大值	2064	2.28	0.38	1.56	2.04	2.22	2.45	6.54

表 D.2　日度 rank IC 和 ICIR（2015.1-2023.6）

因子	rank IC 均值	ICIR	日均样本数
人工关联度牵引因子	2.00%	0.1930	1400.4
人工关联度牵引因子（行业中性）	1.78%	0.2449	1400.4
W2V 相似度牵引因子	2.46%	0.1851	1759.8
W2V 相似度牵引因子（行业中性）	2.39%	0.2593	1759.8
beta（BETA）	0.06%	0.0029	3327.0
市值（SIZE）	-1.84%	-0.1200	3354.9
动量（MOMENTUM）	-6.40%	-0.4006	3354.8
残余波动率（RESVOL）	-4.73%	-0.2952	3354.8
价值（BTOP）	-2.62%	-0.1730	3350.1
非线性市值（NLSIZE）	1.79%	0.2158	3354.9
盈利（EARNYILD）	1.97%	0.1720	3354.3
流动性（LIQUIDTY）	-6.81%	-0.4270	3354.5
杠杆（LEVERAGE）	-0.38%	-0.0357	3353.4

表 D.3　各年度因子 rank IC 情况

平均 rank IC	2015	2016	2017	2018	2019	2020	2021	2022	2023-H1
人工	4.1%	3.0%	1.4%	0.9%	2.0%	0.9%	1.3%	2.6%	1.6%
ICIR	0.31	0.28	0.15	0.11	0.29	0.09	0.13	0.19	0.18
人工.I	3.7%	3.1%	1.2%	0.7%	1.9%	1.1%	0.9%	1.8%	1.4%
ICIR	0.40	0.40	0.18	0.11	0.35	0.17	0.11	0.22	0.28
W2V	3.6%	2.8%	2.4%	1.6%	2.0%	1.7%	2.5%	2.8%	3.0%
ICIR	0.24	0.20	0.19	0.12	0.19	0.16	0.20	0.16	0.23
W2V.I	3.8%	3.1%	2.3%	1.5%	2.0%	1.7%	2.3%	2.1%	2.8%
ICIR	0.39	0.32	0.25	0.16	0.24	0.25	0.25	0.20	0.31
beta	3.1%	0.3%	-1.1%	0.6%	0.2%	-0.2%	-1.2%	-1.3%	-0.0%
ICIR	0.18	0.01	-0.05	0.03	0.01	-0.01	-0.07	-0.07	-0.00
市值	-6.1%	-4.1%	1.2%	-0.6%	-0.7%	-0.1%	-1.9%	-1.9%	-2.7%
ICIR	-0.39	-0.30	0.07	-0.04	-0.06	-0.01	-0.12	-0.11	-0.17
动量	-8.0%	-8.3%	-5.8%	-6.9%	-8.0%	-4.7%	-5.5%	-5.2%	-3.9%
ICIR	-0.43	-0.54	-0.34	-0.44	-0.62	-0.30	-0.36	-0.31	-0.27
残余波动率	-1.1%	-4.5%	-4.4%	-3.6%	-5.9%	-5.2%	-5.8%	-6.6%	-6.5%
ICIR	-0.06	-0.26	-0.28	-0.21	-0.40	-0.32	-0.40	-0.47	-0.48
价值	-1.2%	-3.5%	-2.9%	-2.0%	-1.9%	-1.0%	-3.3%	-4.2%	-4.4%
ICIR	-0.06	-0.23	-0.23	-0.15	-0.18	-0.06	-0.20	-0.24	-0.35
非线性市值	3.0%	3.9%	-0.3%	1.2%	1.5%	1.2%	1.5%	2.1%	2.1%
ICIR	0.40	0.57	-0.03	0.13	0.22	0.17	0.17	0.22	0.29
盈利	-0.1%	1.7%	3.6%	2.2%	2.3%	2.1%	1.8%	2.0%	2.3%
ICIR	-0.01	0.15	0.37	0.21	0.19	0.18	0.18	0.17	0.21
流动性	-4.9%	-7.9%	-6.7%	-6.8%	-8.0%	-6.7%	-6.4%	-7.5%	-5.9%
ICIR	-0.29	-0.44	-0.44	-0.40	-0.50	-0.40	-0.44	-0.56	-0.41
杠杆	-2.1%	-0.1%	0.5%	-0.6%	-0.7%	-0.5%	0.1%	0.5%	-0.7%
ICIR	-0.15	-0.01	0.04	-0.06	-0.10	-0.05	0.01	0.04	-0.06
成长	0.3%	-0.2%	1.9%	1.0%	1.5%	1.5%	0.7%	0.4%	0.1%
ICIR	0.05	-0.04	0.32	0.15	0.24	0.22	0.09	0.04	0.01

表 D.4　Fama-MacBeth 回归结果（不同样本股池，2015.1-2023.6）

	(1)	(2)	(3)	(4)	(5)
人工	0.0181				
	[5.39]				
人工.I		0.0157			
		[5.92]			
W2V			0.0236		
			[5.01]		
W2V.I				0.0209	
				[6.03]	
beta	0.0225	0.0225	0.0292	0.0278	0.0416
	[2.80]	[2.77]	[3.74]	[3.54]	[5.97]
市值	-0.0225	-0.0223	-0.0244	-0.0256	-0.0289
	[-3.60]	[-3.53]	[-4.04]	[-4.20]	[-4.77]
动量	-0.0351	-0.0354	-0.0329	-0.0326	-0.0506
	[-4.40]	[-4.44]	[-4.04]	[-4.02]	[-6.95]
残余波动率	0.0461	0.0462	0.0413	0.0411	0.0245
	[6.87]	[6.81]	[6.35]	[6.26]	[3.74]
价值	0.0005	0.0004	0.0021	0.0010	0.0006
	[0.07]	[0.06]	[0.34]	[0.15]	[0.12]
非线性市值	0.0085	0.0085	0.0122	0.0119	0.0207
	[2.63]	[2.66]	[3.88]	[3.83]	[6.98]
盈利	0.0050	0.0049	0.0056	0.0051	0.0049
	[1.20]	[1.19]	[1.50]	[1.37]	[1.46]
流动性	-0.0990	-0.0978	-0.1059	-0.1057	-0.1191
	[-9.36]	[-9.88]	[-11.06]	[-10.90]	[-11.48]
杠杆	-0.0045	-0.0047	-0.0056	-0.0053	-0.0052
	[-1.24]	[-1.27]	[-1.59]	[-1.45]	[-1.40]
成长	0.0244	0.0244	0.0262	0.0263	0.0267
	[7.38]	[7.36]	[8.93]	[8.87]	[11.80]
平均观测数	1400.2	1400.2	1758.6	1758.6	3321.5
平均 Adj.R2	0.1249	0.1241	0.1196	0.1178	0.0941

注：中括号内为 *t* 值

致谢

岭外音书断，经冬复历春。本科时感受着学位论文和研究报告的差别，如今这两种文体之间或许还得加入商业分析报告、上机实验报告、求职答辩报告，以及 GAI 代我之口连缀而成的文字排列。四年复两年，心境能够一以贯之，自知难得。

感谢金融学系刘玉珍老师在论文完成过程中耐心且专业的指导。从论文选题到形成终稿，正是老师的悉心帮助才使得最初宽泛的研究想法逐步落实成具体的研究问题和应用场景。还要感谢王汉生、贾金柱、王立威等学校和商业分析项目的老师，他们向我提示着研究和应用领域的最新进展，启发我拥抱变化，独立思考，运用新的工具方法解决重要问题。此外，还要感谢身在远方的几位友人同我分享本文相关课题的新进展。关于这些课题的研究兴趣最初来自三年前的胡俊峰老师的课堂，感谢三年来的自己。最后，还想感谢这段时间陪伴我的家人、同事，还有这台历经沧桑临危受命的 ThinkPad T480s 电脑。