# 基于 AI 大语言模型的叙事资产定价研究

樊嘉诚　林建浩　李宗余

## 目 录

# 附录 I　文本数据与文本模型

附录 I 补充了新闻样本的选择标准，以及有关大语言模型和 LDA 主题模型的应用细节。

## I.1 新闻样本的选择

本文研究仅考察专业性的财经类媒体，主要出于以下几个方面的考虑。第一，本文强调媒体叙事信息会从宏观层面影响整个 A 股市场的股票价格，验证该逻辑仅需要使用对整个金融市场有全面性影响的新闻，例如与宏观经济状况、中央和地方政府政策、重要行业及公司变化等相关的新闻文本。第二，从数据专业性与可信度的角度看，虽然近年来以微信、微博、抖音、小红书为代表的自媒体发展迅速，传统纸质媒体的影响力有所下降，但相较于网络自媒体，纸质媒体受到更严格的道德和法律约束，其报道内容相对客观真实，从一定程度上确保了媒体数据的专业性与客观性，而且很多社交媒体的专业性信息实际上也是对专业财经媒体信息的转发和转化。第三，从数据可得性和规范性的角度看，社交媒体出现的时间相对较晚使实证研究的样本期较短，且社交媒体缺少规范的信息获取渠道，尤其是非公开数据具有较高的数据壁垒，不利于研究结果的后续推广和应用。

进一步地，我们对新闻样本的选择依据进行说明。本文财经媒体样本的选择综合了经济学顶级期刊文献，以及第三方媒体评比，最终确定了七家在我国具有广泛影响力、知名度和权威性较高的专业财经报纸作为媒体池。具体包括《证券日报》《证券时报》《金融时报》《第一财经日报》《21 世纪经济报道》《中国经营报》和《每日经济新闻》，涵盖了官方媒体和市场化媒体两类[1]。表 I1 总结了样本说明和选择依据，其中三家官方媒体均为证监会指定信息披露报刊（“七报一刊”），四家市场化媒体在多个评比中表现突出，如人民网研究院《2020 报纸融合传播指数报告》、胡润百富《2020 胡润中国最具影响力财经媒体榜》等，这些媒体在行业中占有较高地位和广泛影响力。更为重要的是，如 Qin et al.（2018）、You et al.（2018）等顶刊文献也采用了类似的财经媒体样本，证明了这些媒体的代表性和合理性。因此，选择这七家媒体作为样本不仅具备较高的权威性和代表性，也确保了研究的广泛适用性和科学性。

**表 I1　七家财经媒体的说明及选择依据**

| | 名称 | 主管单位 | 七报一刊 | QSW | YZZ | 慧科 | 人民网 | 胡润 |
|---|---|---|---|---|---|---|---|---|
| 1 | 证券日报 | 经济日报社 | ✓ | ✓ | ✓ | ✓ | | ✓ |
| 2 | 证券时报 | 人民日报社 | ✓ | ✓ | ✓ | ✓ | | ✓ |
| 3 | 金融时报 | 中国人民银行 | ✓ | ✓ | | ✓ | | |
| 4 | 第一财经日报 | 上海东方传媒集团 | | ✓ | ✓ | ✓ | ✓ | ✓ |
| 5 | 21 世纪经济报道 | 南方报业传媒集团 | | ✓ | ✓ | ✓ | | ✓ |
| 6 | 中国经营报 | 中国社会科学院 | | ✓ | ✓ | ✓ | | ✓ |
| 7 | 每日经济新闻 | 成都传媒集团 | | ✓ | | ✓ | ✓ | ✓ |

注：“七报一刊”为证监会指定信息披露报刊。“QSW”表示由 Qin、Strömberg 和 Wu 于 2018 年发表在 American Economic Review 的 Media Bias in China（Google 学术引用量为 259 次）。“YZZ”表示由 You、Zhang 和 Zhang 于 2018 年发表在 Review of Financial Studies 的 Who Captures the Power of the Pen?（Google 学术引用量为 314 次）。“慧科”表示由慧科讯业评选出最受关注的中文财经媒体；“人民网”表示人民网研究院评选的报纸融合指数排行榜前十媒体[2]；“胡润”表示由胡润百富评选的中国最具影响力财经媒体[3]。

最后，图 I1 汇总了数据筛选后各年度的新闻报道总量。从年度趋势来看，2008 年和 2009 年是财经新闻报道最为活跃的两个年份，有效新闻报道量达到峰值。从媒体分布来看，《每日

[1] 具体地，属于官方媒体的为《证券日报》《证券时报》和《金融时报》，属于市场化媒体的为《第一财经日报》《21 世纪经济报道》《中国经营报》和《每日经济新闻》。

[2] 参考链接：http://yjy.people.com.cn/n1/2021/0426/c244560-32088636.html。

[3] 参考链接：https://www.hurun.net/zh-CN/Info/Detail?num=65537CD76E20。

经济新闻》的报道量在大多数时间内位居七家媒体之首，尤其是在样本期的早期阶段，其报道数量显著领先。进一步对比官方媒体与市场化媒体的报道情况可以发现，在样本期初期，市场化媒体的报道量远超官方媒体，显示出其在财经信息传播中的主导地位。然而，近年来二者的报道量逐步趋于相当，表明官方媒体在财经新闻领域的影响力稳步提升。

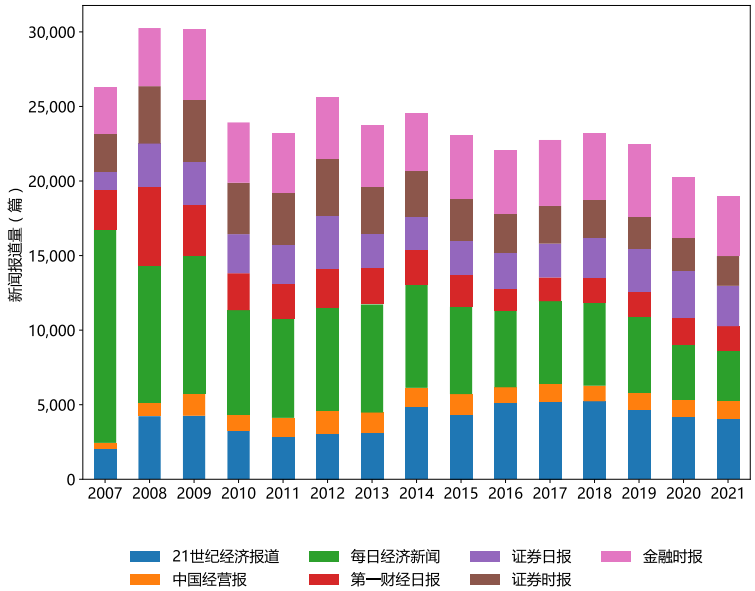


图 I1　七家财经媒体的年度新闻报道数量统计

I.2 关于大语言模型的补充说明

本文使用在 BERT 中文模型的基础上，通过 CoSENT（Cosine Sentence）方法进一步微调的预训练模型。BERT-CoSENT 模型在 BERT 基础上引入了余弦相似度损失函数，以优化句子级文本表示。尽管该模型更适用于句向量的提取，但一般而言也可以用于生成文档级别的向量表示。本文主要是基于以下三个方面的考虑：第一，BERT-CoSENT 模型在多个中文文本匹配任务中表现出色，显示了其在捕捉中文文本语义方面的有效性。根据评测结果，该模型在 ATEC、BQ、LCQMC、PAWSX、STS-B 等数据集上的平均 Spearman 系数超过 68%，表现优于基于 Sentence 方法的中文 BERT 和 RoBERTa 模型[4]。

第二，尽管该模型主要针对句子级别的文本分析，但通过适当的策略调整，比如将句子嵌入向量进行聚合（如池化技术）或层次化编码方法，可以有效地将其应用于文档级别的表示。这一方法在实践中已被验证可行，例如 Yang et al.（2016）、Liu and Lapata（2019）以及 Reimers and Gurevych（2019）等计算语言学研究均提供了支持。本文的应用同样参考了这一思路，以在确保输入 tokens 长度限制的前提下，尽可能地利用整篇文档的所有信息。

第三，在经济学领域，也有研究利用 BERT 模型构建长文档的嵌入向量。例如，许雪晨和田侃（2021）将个股财经新闻输入 BERT 模型，并通过[CLS]标记获取输出向量，结合 Bi-LSTM 和全连接层输出文本情感。刘青和肖柏高（2023）将专利文本输入 BERT 后得到代表文本整体语义的向量，然后添加线性分类层构建了一个专利文本分类模型，据此识别出与劳动节约型技术有关的专利。Chen et al.（2022）和 Tan et al.（2023）使用了一系列 BERT 类模型，均是采用先计算每个 token 的嵌入，再通过简单平均的方法生成文档级别的向量表征，最后将其用于回报预测等下游任务。这些文章表明，使用 BERT 模型构建文档嵌入在经济学研究中具有广泛性和普遍应用性。

[4] 参考文档：https://github.com/shibing624/text2vec/blob/master/docs/model_report.md。

此外，本文对超出 BERT 输入长度限制的文本进行了特殊处理。具体而言，我们通过正则表达式将正文文本按字符长度平均分割为若干块，每个块作为一个“段落”，长度不超过 256 个字符。例如，若文档长度为 600 字符，则分割为 3 个“段落”，每个约 200 字符。我们将这些“段落”直接输入 BERT-CoSENT 模型，并获得其 768 维的嵌入向量。为了获取整篇文章的信息表征（即文档层面的嵌入向量），并进一步降低字符分割、token 数量错位和语义不连贯等问题造成的信息损失风险。我们进行如下处理：

第一，以标题主导不同“段落”的语义聚合。对每一篇新闻，我们先单独计算标题的嵌入向量，作为首个块嵌入，然后将切割好的正文“段落”嵌入投影到标题嵌入的向量空间，再进行加权求和。目的是以标题向量作为锚点，通过投影将正文文本信息向标题靠拢，增强语义一致性，弱化段落分割误差对整体文档嵌入的影响，从而得到较为准确的文档信息表征。这样做一方面即使分割不完美，标题的强引导作用和池化策略仍能保持文档语义的稳定性；另一方面也过滤了长文本无关信息的干扰，强化与经济相关的核心叙事。这种做法在 NLP 领域较为常见。例如，在检索增强生成（Retrieval-Augmented Generation，RAG）任务中，模型通过检索相关文档并将其与查询结合来生成答案，以此聚合相关信息并增强生成内容的语义一致性。

第二，在段落中对每一个 token 的嵌入做均值池化，使局部信息丢失可以被全局平均补偿。LLM 模型的一个核心优势（相对传统词向量模型）是有效的噪声抑制和语义聚合，因为被切割断裂的 token（噪声）在整个段落中占比相对较低，有效 token（完整词句）在池化中占据主导，通过注意力机制对 token 嵌入进行加权平均后，噪声 token 对段落整体嵌入的影响较小，能有效保留段落核心语义。值得强调的是，本文采用均值池化主要基于以下两点考虑：一方面，Reimers and Gurevych（2019）的研究表明，相较于最大池化（max pooling）和[CLS]标记，均值池化（mean pooling）在不同优化任务中表现出更高的稳健性。另一方面，本文未对大语言模型本身进行改动，因此选择默认的均值池化方法对嵌入聚合。

I.3 关于 LDA 模型的补充说明

LDA 最早由 Blei et al.（2003）提出，是一种无监督的主题模型。该模型假设每篇文档由多个主题以不同概率混合而成，而每个主题则由特定的单词概率分布生成。有关 LDA 模型的数学描述与估计细节，Larsen and Thorsrud（2019）、Bybee et al.（2024）和 Hong et al.（2025）等文献已作详细介绍，本文不再赘述。

在使用 LDA 模型提取主题信息之前，需要对新闻文本数据进行必要的预处理，以将其转换为适用于模型训练的结构化数据。与已有文献一致，本文使用 Python 软件包 jieba 中文分词库进行分词，剔除数字、英文、标点符号、常见停用词（如“比如”“之前”等），以及无具体含义的词语（如“省”“其”等）。为提高分析效率，仅保留长度不少于 2 的词语，并筛选出至少出现于 50 篇文章中的词语。然后，我们基于预处理后的文本构建文档-词频矩阵（document-term matrix），并将其输入 LDA 模型，以获取每篇文章的主题分布。

LDA 模型在估计参数时需要指定超参数——主题的数量（$L$）。为了确定最优主题数，我们以 10 为增量，对主题数从 10 到 100 的一系列模型进行了估计。然后，选择最大化模型对数似然值（每个主题对数似然值之和）。另外，还计算了每个模型的困惑度（perplexity）作为稳健性检验。**图 I2** 报告了两种方法的结果，并表明 20～40 是一个最佳范围，最终选择$L = 30$作为拟合 LDA 模型的主题数，并开展正文中的一系列实验。

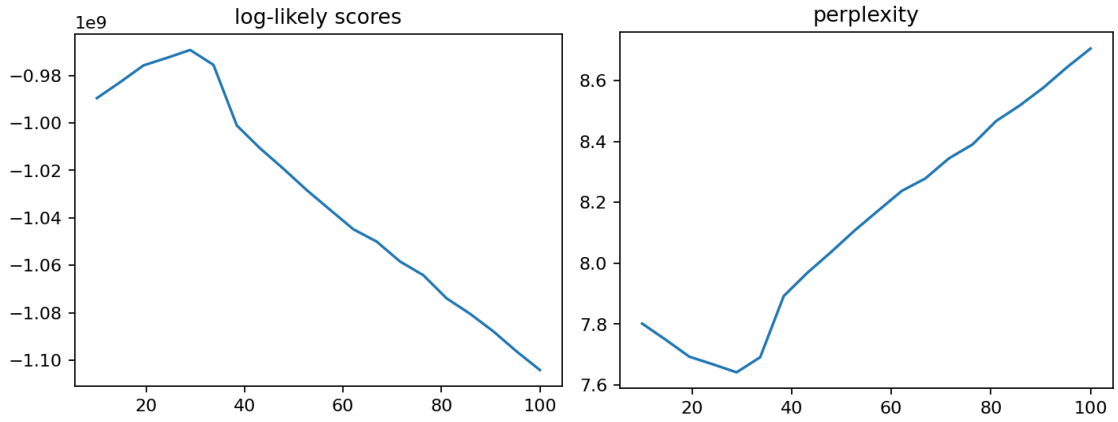


图 I2　LDA 模型的最优主题数选择

表 I2 展示了正文使用的 LDA 模型所识别出的 30 个叙事主题及其对应的前 10 大关键词。可以看出，这些主题广泛涵盖了经济、金融、政治和社会等多个维度，主要包括风险管理、经济发展、资本市场、货币政策、劳动力市场和房地产市场等宏观层面的核心话题。

表 I2　LDA 模型识别的 30 个主题

| 序号 | 主题名称 | 前 10 大关键词 |
|---|---|---|
| 1 | 风险管理 | 风险，监管，金融，保险，管理，业务，机构，行业，发展，产品 |
| 2 | 区域发展 | 产业，发展，经济，建设，上海，规划，全国，成都，深圳，中心 |
| 3 | 金融服务 | 金融，服务，企业，支持，贷款，发展，业务，融资，金融机构，中小企业 |
| 4 | 公司财务 | 公司，股东，股权，业绩，净利润，股份，业务，披露，资产，持有 |
| 5 | 能源环境 | 疫情，绿色，环保，环境，企业，排放，影响，粮食，生产，工作 |
| 6 | 法律合规 | 信息，相关，法律，工作，情况，部门，违规，人员，案件，合同 |
| 7 | 经济发展 | 经济，不是，政府，发展，不能，很多，国家，解决，都是，利益 |
| 8 | 美国经济 | 美国，经济，全球，美联储，预期，黄金，加息，复苏，影响，风险 |
| 9 | 消费市场 | 品牌，汽车，产品，销售，消费者，销量，消费，电商，行业，渠道 |
| 10 | 文化传媒 | 文化，告诉，上海，都是，媒体，很多，工作，当地，公司，发现 |
| 11 | 网络平台 | 平台，服务，互联网，技术，用户，行业，业务，科技，发展，网络 |
| 12 | 经济改革 | 发展，改革，经济，创新，推进，建设，推动，工作，资本，体系 |
| 13 | 资本运作 | 企业，公司，行业，资本，投资，业务，并购，重组，收购，国企 |
| 14 | 投资管理 | 投资，基金，产品，资产，收益，理财，投资者，资金，规模，管理 |
| 15 | 市场行情 | 板块，投资者，资金，上涨，下跌，行情，估值，行业，涨幅，反弹 |
| 16 | 经济增长 | 增长，同比，经济，增速，消费，下降，投资，企业，收入，出口 |
| 17 | 国际合作 | 合作，全球，国家，贸易，经济，发展，世界，国际，投资，出口 |
| 18 | 资本市场 | 交易，投资者，资本，发行，期货，监管，券商，融资，租赁，业务 |
| 19 | 农村发展 | 农业，信托，旅游，农民，扶贫，土地，产业，发展，贫困，农户 |
| 20 | 政府政策 | 改革，政府，政策，财政，地方，收入，企业，制度，部门，试点 |
| 21 | 银行业务 | 银行，贷款，资产，业务，信贷，银行业，商业银行，不良，利率，存款 |
| 22 | 地方融资 | 融资，债券，资金，债务，地方，发行，政府，风险，信用 |
| 23 | 外汇市场 | 人民币，汇率，香港，外资，境外，升值，跨境，资本，投资，规模 |
| 24 | 基础建设 | 项目，建设，投资，工程，能源，铁路，基础设施，规划，交通，内地 |
| 25 | 房地产市场 | 房地产，土地，房价，住房，调控，房企，项目，政策，价格，开发 |
| 26 | 劳动力市场 | 工作，教育，就业，人才，员工，人口，人员，培训，学生，工资 |
| 27 | 物价水平 | 价格，上涨，煤炭，产能，行业，钢铁，石油，原油，企业，需求 |
| 28 | 生产制造 | 产品，企业，技术，生产，行业，研发，公司，医疗，设备，制造 |
| 29 | 货币政策 | 政策，经济，央行，货币政策，流动性，利率，货币，调整，压力，信贷 |
| 30 | 全球经济 | 美国，欧盟，英国，欧洲，德国，希腊，经济，国家，总统，政治 |

# 附录 II ICAPM 理论与收益率模型

附录 II 补充了一些 ICAPM 理论和条件因子模型中的公式推导，以及实证中估计模型参数的算法细节。

II.1 ICAPM 中叙事新息$z$、状态变量$\boldsymbol{x}$与定价因子$f$的关系推导

在方程（2）中，假设高维的叙事新息$z_\tau$中蕴含着影响投资者行为和预期的低维状态变量$\boldsymbol{x}_\tau$，其之间的关系通过一个线性矩阵$\widetilde{\boldsymbol{A}}$来刻画。但是，系数矩阵$\widetilde{\boldsymbol{A}}$、状态变量$\boldsymbol{x}_\tau$均不可观测，为了描述状态变量与资产价格的关系，根据 ICAPM 理论，资产超额收益可表示为[5]

$$r_{i,t+1} = \beta_{i,t}^{(m)} f_{t+1}^{(m)} + \beta_{i,t}^{(h)} f_{t+1}^{(h)} \tag{II.1}$$

其中，第一项为市场因子（及暴露），第二项为 Merton 所定义的对冲项（hedge），是投资者根据自身的状态变量（经济增长衰退，信贷扩张、收缩等）而做的对冲组合，用来对冲经济环境的变化。针对第一项，本文沿用 Bybee et al.（2023）的设定，不指定叙事定价因子的某一维度来对应市场因子，而是认为所有维度均对市场因子和状态变量因子有贡献。尽管 Merton（1973）中的 ICAPM 因子（$f_{t+1}^{(m)}, f_{t+1}^{(h)}$）都是 1 维的 SDF，叙事定价因子为$K$维，但 Huberman and Kandel（1987）指出了多因子模型与随机折现因子（SDF）的关联，即多因子模型是对定价核（Pricing Kernel）的线性逼近，其 MVE 组合在有效前沿上与 SDF 等价。第二项$f_{t+1}^{(h)}$受状态变量影响，即$\beta_{i,t}^{(h)}$由状态变量决定。我们不假设状态变量和定价因子一一对应（例如，不认为$\beta_{i,t}$的第一个维度对应$\beta_{i,t}^{(m)}$），而是存在更复杂的关系，需要进一步通过数据识别。在加入了这个假设后，本文不再区分$\beta_{i,t}^{(m)}$和$\beta_{i,t}^{(h)}$，而是通过 IPCA 模型统一识别$\beta_{i,t}$。

具体而言，根据 ICAPM 理论，状态变量$\boldsymbol{x}_t$决定了因子暴露$\beta_{i,t}$，若$\boldsymbol{x}_t$可观测，那可以按文献中经典做法 Fama-Macbeth 回归（下称 FMB 回归），用方程（II.2）中的$\beta_{i,t}^{(x)}$来估计$\beta_{i,t}$：

$$r_{i,t+1} = \beta_{i,t}^{(x)} x_{t+1} + u_{i,t+1} \tag{II.2}$$

但是在本文设定中$\boldsymbol{x}_t$不可观测，FMB 回归无法进行。所以考虑先从$\boldsymbol{x}_t$中提取能够影响资产价格的部分信息，即影响$f_t$的部分，这个关系可以通过以下回归方程（II.3）来表达：

$$x_t = \boldsymbol{B} f_t + v_t \tag{II.3}$$

其中$v_t \perp f_t$，系数矩阵$\boldsymbol{B}$不一定是单位阵，将（II.3）代入（II.2），此时$\beta_{i,t}$可表示为

$$\beta_{i,t}^{(x)} \boldsymbol{B} = \frac{Cov\left(r_{i,t+1}, f_{t+1}\right)}{Var(f_{t+1})} \triangleq \beta_{i,t} \tag{II.4}$$

注意到方程（II.4）右边的等号是$\beta_{i,t}$定义，左边恰好是 IPCA 模型能够识别的方程形式[6]，这也是要考虑使用 IPCA 方法来估计参数的原因。但是，IPCA 要求$\beta_{i,t}^{(x)}$是可观测的工具，但此时$\beta_{i,t}^{(x)}$仍不可观测，为估计方程（3）的$\beta_{i,t}$，将方程（II.3）代入正文方程（2）得到：

$$z_t = \widetilde{\boldsymbol{A}} \boldsymbol{B} f_t + \widetilde{\boldsymbol{A}} v_t + \eta_t \tag{II.5}$$

并记：$\mathbf{A}_{L\times K} \triangleq \widetilde{\boldsymbol{A}} \boldsymbol{B}, g_t \triangleq \widetilde{\boldsymbol{A}} v_t + \eta_t$，便得到正文方程（4）。进一步将方程（4）代入$Cov\left(r_{i,t+1}, z_{t+1}\right)$得到

$$Cov\left(r_{i,t+1}, z_{t+1}\right) = Cov\left(r_{i,t+1}, \boldsymbol{A} f_{t+1} + g_{t+1}\right) = Cov\left(r_{i,t+1}, f_{t+1}\right) \boldsymbol{A}^T \tag{II.6}$$

将方程（II.4）代入方程（II.6），并定义：$\boldsymbol{\Sigma}_{ff} \triangleq Cov_t(f_{t+1})$，得到：

[5] 在因子模型中我们要用日频收益率来构建月频的因子收益率，故这里我们将状态变量也写成月频形式。

[6] Kelly et al.（2019）中描述的 IPCA 估计方程为：$r_{i,t+1} = c_{i,t} \boldsymbol{\Gamma} f_{t+1} + \epsilon_{i,t+1}$，其中$c_{i,t}$为个股层面的代理变量，Kelly 等人将其称为“工具”（instrument），$\boldsymbol{\Gamma}$为系数矩阵。IPCA 的目的是借助可观测的“工具”（instrument）来识别资产收益率中不可观测的因子结构。

$$Cov\left(r_{i,t+1}, f_{t+1}\right) = \beta_{i,t}^{(x)} \boldsymbol{B}\boldsymbol{\Sigma}_{ff} = Cov\left(r_{i,t+1}, z_{t+1}\right)\boldsymbol{A}(\boldsymbol{A}^T\boldsymbol{A})^{-1} \tag{II.7}$$

整理可得：

$$\beta_{i,t}^{(x)}\boldsymbol{B} = Cov\left(r_{i,t+1}, z_{t+1}\right)\boldsymbol{A}(\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{\Sigma}_{ff}^{-1} \tag{II.8}$$

记$cov_{i,t} \triangleq Cov_t\left(r_{i,t+1}, z_{t+1}\right), \tilde{\boldsymbol{\Gamma}} \triangleq \boldsymbol{A}(\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{\Sigma}_{ff}^{-1}$，便得到正文中的 IPCA 估计方程（5）。

II.2 估计方程数值算法细节

本文按以下步骤估计上述模型。

第一步，估计个股的工具$cov_{i,t}$。具体而言，对于每只股票$i$和月份$t$，利用日度新闻叙事新息$z_\tau$和个股收益率$r_{i,\tau}$计算样本协方差：

$$\widehat{cov}_{i,t} \triangleq \sum_{\tau} \kappa(\tau; t) r_{i,\tau} z_\tau^T - \left(\sum_{\tau} \kappa(\tau; t) r_{i,\tau}\right)\left(\sum_{\tau} \kappa(\tau; t) z_\tau^T\right) \tag{II.9}$$

其中，$\widehat{cov}_{i,t}$是一个$1 \times L$的行向量，$\kappa(\tau; t)$是一个对每日观测值进行加权的函数。考虑到月初的新闻影响相对月末小很多，考虑到投资者的有限注意力与信息衰减机制，对样本进行加权处理，加权核函数设定如下：

$$\kappa(\tau; t) \coloneqq \frac{\xi^{(\tau_t - \tau)}}{\sum_{\tau \le \tau_t} \xi^{(\tau_t - \tau)}} = \frac{\xi^{(\tau_t - \tau)}}{\frac{(1-\xi^{\tau_t})}{(1-\xi)}} = \frac{(1-\xi)\xi^{\tau_t}}{1-\xi^{\tau_t}}\xi^{\tau}, \qquad \tau \le \tau_t \tag{II.10}$$

其中，$\xi$为权重衰减系数，在本文中取 0.99，$\tau_t$为第$t$个月的最后一天（如 3 月则：$t = 3, \tau_t = 31$），$\tau$为第$t$个月的第$\tau$天。

第二步，由于叙事新息具有高维性，但并非所有维度都与资产收益率相关。因此，通过分组 LASSO 惩罚项引入稀疏性，优化如下目标函数：

$$\min_{\boldsymbol{\Gamma}, f_t} \frac{1}{2}\sum_{i,t\in\mathbb{S}}\left(r_{i,t} - c_{i,t-1}\boldsymbol{\Gamma} f_t\right)^2 + \lambda N_{\mathbb{S}} \sum_{l=0}^{L} \sigma_l^c \|\Gamma_l\|_2 + \sum_{t\in\mathbb{S}} \|f_t\|_2^2 \tag{II.11}$$

其中，$\mathbb{S}$为训练数据集，$c_{i,t} \triangleq \left[1, \widehat{cov}_{i,t}\right]$，$\boldsymbol{\Gamma} \triangleq [\Gamma_0; \tilde{\boldsymbol{\Gamma}}] = [\Gamma_0; \Gamma_1; \dots; \Gamma_L]$为$(L+1) \times K$的系数矩阵；$\Gamma_0$为常数项，目的在允许不同因子有不同的截距项。第二项$\|\Gamma_l\|_2 \triangleq \sqrt{\Gamma_{l,1}^2 + \cdots + \Gamma_{l,K}^2}$为分组 LASSO 正则化的 L2 范数，$\lambda$为惩罚项系数，$N_{\mathbb{S}}$为训练集样本数，$\sigma_l^c$为第$l$列的标准差[7]。第三项对$f_t$正则化，以保证数值解的合理性。惩罚项的$\lambda$超参数根据最大化因子模型的夏普比率调优：

$$\lambda_{\mathbb{S}}^* = \arg\max_\lambda SR(\lambda; \mathbb{S}) \triangleq \sqrt{\hat{\mu}_f^T \hat{\boldsymbol{\Sigma}}_{ff}^{-1} \hat{\mu}_f} \tag{II.12}$$

其中，$\hat{\mu}_f$和$\hat{\boldsymbol{\Sigma}}_{ff}$为$f_t$的样本均值和样本协方差矩阵。

接下来，对于上述优化问题沿用 Bybee et al.（2023）的交替正则最小二乘算法（Alternating Regularized Least Squares，ALRS），它在最小化$\boldsymbol{\Gamma}$同时保持$\{f_t\}$固定和最小化$\{f_t\}$同时保持$\boldsymbol{\Gamma}$固定之间交替进行，直到$\boldsymbol{\Gamma}$和$\{f_t\}$收敛。ARLS 算法可以看作是区块坐标下降算法的一种特例，这两个参数就是两个“块”，当目标函数值下降很小（或满足一阶条件）时算法结束。其中，求解$\boldsymbol{\Gamma}$的子问题是在$\{i, t\}$面板数据上的分组 LASSO 回归：

$$\min_{\boldsymbol{\Gamma}} \frac{1}{2}\sum_{i,t\in\mathbb{S}}\left(r_{i,t} - c_{i,t-1}\boldsymbol{\Gamma} f_t\right)^2 + \lambda N_{\mathbb{S}} \sum_{l=0}^{L} \sigma_l^c \|\boldsymbol{\Gamma}_l\|_2 \tag{II.13}$$

将截面样本按列堆叠成向量有：

[7] 具体地，每一个$i$的$\widehat{cov}_{i,t-1}$为 1 行$L$列向量，$N$个资产则形成$N \times L$的矩阵，$\sigma_l^c$为第$l$列的标准差。

$$\min_{\mathrm{vec}(\boldsymbol{\Gamma})} \frac{1}{N_{\mathbb{S}}}\|r_t-(\boldsymbol{C}_{t-1}\otimes f_t^T)\mathrm{vec}(\boldsymbol{\Gamma})\|_2^2+\lambda\sum_{l=0}^{L}2\sigma_l^c\|\boldsymbol{\Gamma}_l\|_2 \tag{II.14}$$

其中，$r_t:N_t\times 1$，$\boldsymbol{C}_{t-1}:N_t\times(L+1)$，$f_t:K\times 1$，$\mathrm{vec}(\boldsymbol{\Gamma}):(L+1)K\times 1$表示向量化的$\boldsymbol{\Gamma}$，$\otimes$表示克罗内克直积。使用 Yang and Zou（2015）的算法数值求解上述分组 LASSO 回归问题。接着，在得到$\boldsymbol{\Gamma}$后求解$\{f_t\}$子问题：

$$\min_{\boldsymbol{\Gamma}} \frac{1}{2}\|r_t-\boldsymbol{C}_{t-1}\boldsymbol{\Gamma}f_t\|_2^2+\|f_t\|_2^2 \tag{II.15}$$

这是一个截面岭回归问题，有解析解：

$$f_t=(\boldsymbol{\Gamma}^T\boldsymbol{C}_{t-1}^T\boldsymbol{C}_{t-1}\boldsymbol{\Gamma}+2\mathbb{I}_K)^{-1}\boldsymbol{\Gamma}^T\boldsymbol{C}_{t-1}^T r_t \tag{II.16}$$

其中，$f_t$是由单个资产构建的可交易投资组合的收益率。只要$\boldsymbol{C}_{t-1}$和$\boldsymbol{\Gamma}$是用 $t$ 月初之前的可观测数据估计的，则投资组合的权重是事前可得的。从而在不考虑更新$\boldsymbol{\Gamma}$的情况下，全样本估计的$f_t$也可以作为样本外因子收益率使用。

第三步，在得到$\tilde{\boldsymbol{\Gamma}}$和$\boldsymbol{\Sigma}_{ff}$的估计值后，联立方程（2）和方程（5）可以得到状态变量的估计：

$$x_t=(\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{A}^T z_t=\boldsymbol{\Sigma}_{ff}\tilde{\boldsymbol{\Gamma}}^T z_t \tag{II.17}$$

# 附录 III　实证分析的补充结果

附录 III 展示了正文未包含的其他实证分析结果。

## III.1 模型训练结果与描述性统计

**图 III1** 和**图 III2** 报告了惩罚项系数$\lambda$变化对模型估计结果和拟合优度的影响，主要包括三个指标：①总体$R^2 \triangleq 1-\sum_{i,t}\left(r_{i,t}-c_{i,t-1}\mathbf{\Gamma}f_t\right)^2/\sum_{i,t}{r_{i,t}}^2$，即因子所解释的收益率占已实现收益率的比率；②因子 MVE 组合的年化夏普比率$SR(\lambda;\mathbb{S}) \triangleq \sqrt{12\,\hat{\mu}_f^T\hat{\mathbf{\Sigma}}_{ff}^{-1}\hat{\mu}_f}$，用于对超参数$\lambda$进行调优；③模型选择的嵌入向量维度，定义为$\mathbf{\Gamma}$中不全为零的行的个数。

对于 NF-LLM 模型而言，当惩罚项$\lambda$较小时，模型可能在样本内过拟合，此时尽管$R^2$较高，但夏普比率未达到最优，说明未能有效估计叙事因子结构；当$\lambda$增加时，夏普比率先升后降，随后因惩罚过重而下降，表明模型未能充分吸收文本信息。此外，随着$\lambda$的增大，模型对稀疏性的约束加强，导致保留的嵌入向量维度减少。进一步地，对比不同因子数量$K$的模型可以发现：当$K=1$时，夏普比率几乎不变，说明模型未能有效利用叙事信息[8]；$K=2$时，夏普比率提升至约 0.20；$K=3$时，最优夏普比率显著提高，超过 0.55。随后继续增加$K$虽有所提升，但边际效应递减且对超参数相对敏感。

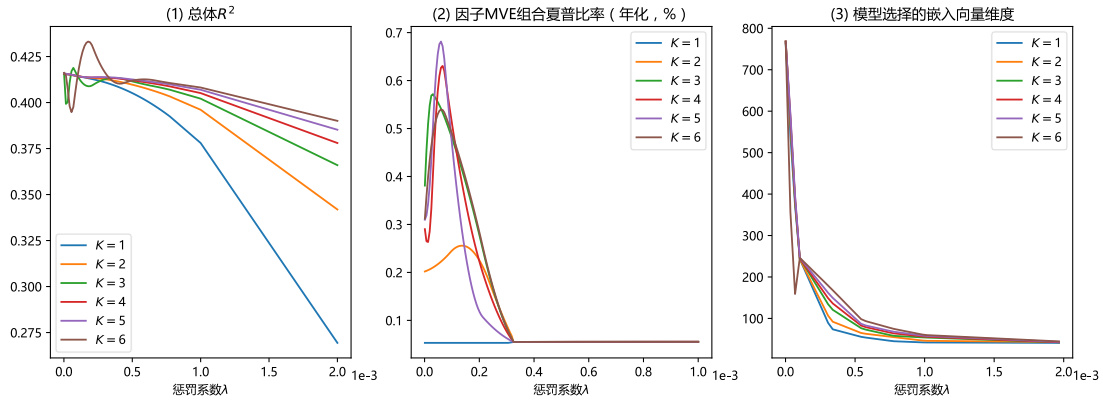


**图 III1　NF-LLM 模型的全样本训练结果**

注：样本期为 2008 年 1 月至 2021 年 12 月，下同。

进一步地，对于 NF-LDA 模型而言，我们发现相比于 LLMs，LDA 模型的拟合效果随着惩罚系数$\lambda$的增大迅速衰减。从第一个子图可以看到，$K\geq 4$的模型在$\lambda=10^{-4}$时已经无法正常拟合。这是因为惩罚系数相对较高，数值算法收敛性较差。与此同时，子图 2 显示$K\geq 4$时夏普比率的波动较为明显，而$K\leq 2$时尽管加大惩罚对$R^2$影响不大，但夏普比率却一直处于较低水平（0.1 左右）。总体来看，因子数$K=3$的结果较为稳定，因此我们选取该设定报告主要的检验结果。

[8] 由于叙事新息是高维的，第一个因子不受$\lambda$的影响表明该因子的估计结果不依赖文本信息，Bybee et al.（2023）同样有类似发现，这一结果意味着模型未能得到有效拟合。

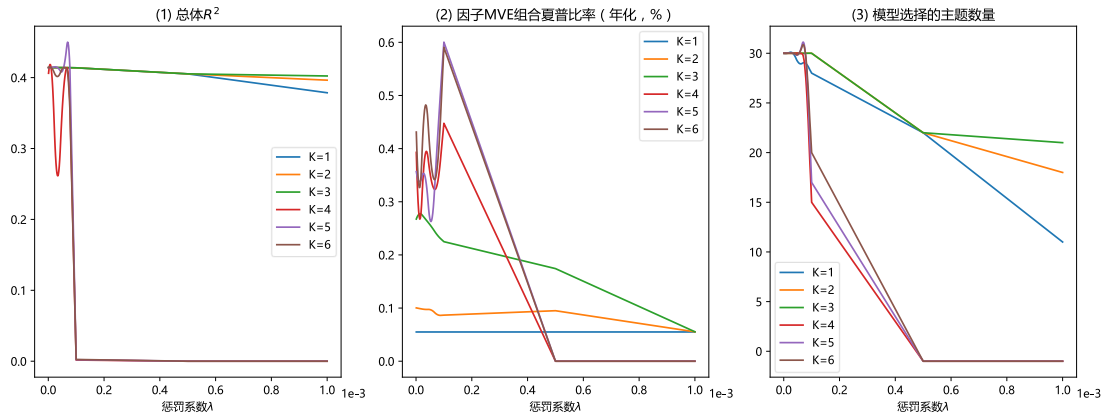


**图 1112　NF-LDA 模型的全样本训练结果**

进一步地，**表 1111** 报告了 NF-LLM 和 NF-LDA 模型的叙事因子以及基准因子（FFC6）的描述统计。在 Panel A 中，各因子间相关系数绝对值最高不超过 0.50，正负分布均匀，说明不同 NF-LLM 因子的特征和作用迥异，整体来看因子模型不会出现共线性过强的情况。Panel B 报告了 NF-LDA 叙事因子的描述性统计。结果与 NF-LLM 因子接近，但波动性相对更高、因子间彼此的相关性更低。Panel C 报告了 FFC6 因子模型的描述统计作为参考，发现在 A 股市场中市场因子（MKT）波动较大，与文献结果基本一致。此外，Panel C 还报告了 NF-LLM 叙事因子与基准因子之间的相关性，绝对值最大不超过 0.25，初步说明二者相关性不高，叙事因子存在额外信息。需要强调的，由于不可观测因子结构在同时估计暴露和因子收益率的时候存在旋转不确定性（Rotational Indeterminacy），单个因子的识别不是唯一的。

**表 1111　叙事因子与基准因子的描述统计**

Panel A：NF-LLM 叙事因子

| | | | 相关系数 | | | | | |
|---|---|---|---|---|---|---|---|---|
| 因子 | 均值(%) | 标准差(%) | $f_1$ | $f_2$ | $f_3$ | $f_4$ | $f_5$ | $f_6$ |
| $f_1$ | 0.01 | 0.50 | 1 | -0.05 | -0.29 | 0.13 | -0.34 | -0.02 |
| $f_2$ | 0.01 | 0.43 | -0.05 | 1 | 0.02 | 0.27 | 0.18 | -0.33 |
| $f_3$ | -0.03 | 0.44 | -0.29 | 0.02 | 1 | -0.07 | -0.16 | -0.29 |
| $f_4$ | 0.01 | 0.44 | 0.13 | 0.27 | -0.07 | 1 | -0.32 | 0.13 |
| $f_5$ | -0.05 | 0.45 | -0.34 | 0.18 | -0.16 | -0.32 | 1 | 0.50 |
| $f_6$ | -0.04 | 0.52 | -0.02 | -0.33 | -0.29 | 0.13 | 0.50 | 1 |

Panel B：NF-LDA 叙事因子

| | | | 相关系数 | | | | | |
|---|---|---|---|---|---|---|---|---|
| 因子 | 均值(%) | 标准差(%) | $f_1$ | $f_2$ | $f_3$ | $f_4$ | $f_5$ | $f_6$ |
| $f_1$ | -0.05 | 0.48 | 1.00 | 0.11 | -0.01 | -0.07 | 0.04 | -0.01 |
| $f_2$ | 0.02 | 0.55 | 0.11 | 1.00 | -0.01 | -0.04 | -0.01 | 0.07 |
| $f_3$ | -0.03 | 0.56 | -0.01 | -0.01 | 1.00 | 0.03 | -0.07 | -0.02 |
| $f_4$ | 0.02 | 0.50 | -0.07 | -0.04 | 0.03 | 1.00 | 0.11 | -0.08 |
| $f_5$ | -0.06 | 0.52 | 0.04 | -0.01 | -0.07 | 0.11 | 1.00 | 0.00 |
| $f_6$ | 0.01 | 0.53 | -0.01 | 0.07 | -0.02 | -0.08 | 0.00 | 1.00 |

Panel C：基准 FFC6 因子

| | | | 相关系数 | | | | | |
|---|---|---|---|---|---|---|---|---|
| 因子 | 均值(%) | 标准差(%) | $f_1$ | $f_2$ | $f_3$ | $f_4$ | $f_5$ | $f_6$ |
| MKT | -0.06 | 7.67 | -0.00 | 0.06 | 0.08 | 0.07 | -0.05 | -0.04 |
| SMB | 0.90 | 3.80 | -0.04 | 0.11 | -0.10 | 0.08 | 0.02 | 0.07 |
| HML | -0.02 | 2.27 | 0.06 | 0.01 | -0.03 | 0.06 | 0.05 | 0.04 |
| UMD | -0.05 | 1.77 | 0.09 | 0.17 | 0.02 | -0.04 | -0.05 | 0.00 |
| RMW | -0.01 | 1.20 | 0.03 | 0.04 | -0.16 | 0.17 | -0.01 | 0.22 |
| CMA | 0.48 | 5.30 | -0.08 | 0.05 | -0.04 | 0.04 | -0.05 | 0.01 |

III.2 状态变量与叙事因子的经济属性

**图 III3** 报告了 NF-LLM 和 NF-LDA 模型提取的叙事状态变量定价核$x^{MVE}$的时间序列（以$K=3$为例）。可以看出，两种叙事因子模型估计的状态变量与消费者信心指数（CCI）都显著的正相关，但由 NF-LLM 估计出的状态变量$x_{LLM}^{MVE}$在相关性和趋势上更接近 CCI。

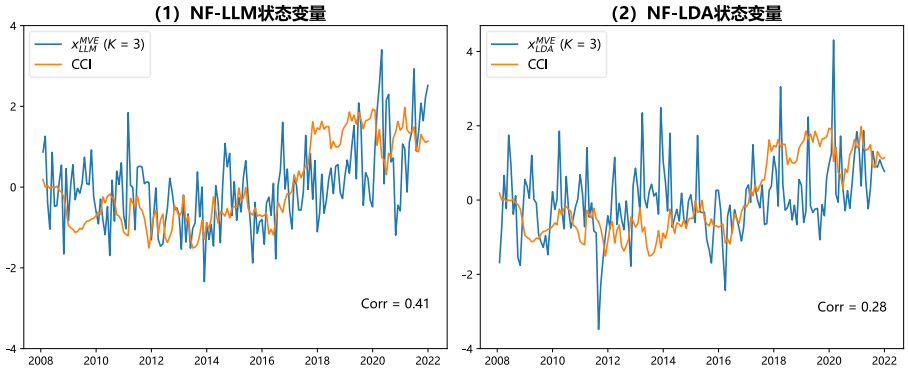


图 III3　叙事因子模型的状态变量定价核（$\mathbf{x}^{\mathbf{MVE}}$）

注：CCI 为消费者信心指数，相关系数（Corr）为斯皮尔曼秩相关系数。样本期为 2008 年 1 月至 2019 年 12 月，下同。

此外，本文在扩张时序窗口中进一步检验这些叙事状态变量对宏观经济指标的预测能力，以探究新闻文本中是否包含对预测宏观经济指标有价值的前瞻性信息。具体而言，参考 Bybee et al.（2023）的做法，将不同的宏观经济变量$y_t$对状态变量的定价核$x_t^{MVE}$进行预测回归。对于每个宏观变量$Macro_t$，预测其未来$h$期的累积值：

$$\sum_{s=0}^{h} Macro_{t+s} = a_h + b_h\left(\frac{x_t^{MVE}}{std(x_t^{MVE})}\right) + e_t^{(h)} \tag{III.1}$$

其中，求和符号表示预测目标在未来$h$期的累积值，$std(x_t^{MVE})$表示定价核的标准差，因而关注的是标准化系数$b_h$，即状态变量的 1 个标准差变化对宏观经济变量的累积影响。

**图 III4** 报告了基于 NF-LLM 提取的状态变量定价核对宏观经济变量的预测回归结果（即系数$b_h$），其中预测期限$h$为 0～24 个月。整体而言，与 ICAPM 理论的结果相符，发现定价核$x_{LLM}^{MVE}$具有顺周期性：$x_{LLM}^{MVE}$的增加为“好消息”，伴随着经济预期的好转、产出和通胀的增加等。具体地，第一个子图表明$x_{LLM}^{MVE}$对 CCI 具有显著的正向影响，并且持续时间长达 2 年以上。从其他子图也可以看出，$x_{LLM}^{MVE}$正向地预测了未来的产出、消费和投资的增长，以及通胀的上升。

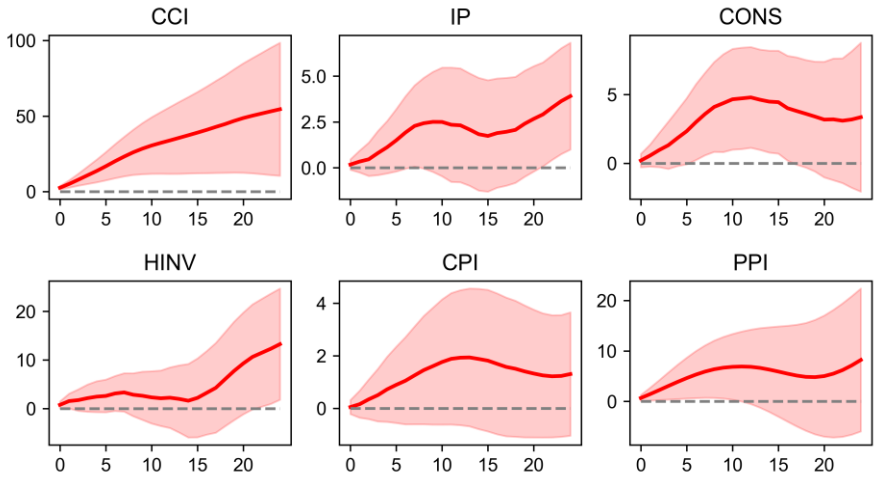


图 III4　NF-LLM 的状态变量定价核对宏观经济变量的预测能力

注：该图报告了预测回归的结果。每个子图对应一个不同的预测目标，横轴是 0～24 个月的窗口长度$h$，纵轴是估计系数$b_h$，阴影区域表示 90%的置信区间。CCI 为消费者信心指数，IP 为规模以上工业增加值同比，CONS 为社会消费品零售总额同比，HINV 为房地产开发投资同比，CPI 和 PPI 分别为消费者价格指数同比和生产者价格指数同比。为减少疫情数据结构性变化的影响，样本期为 2008 年 1 月至 2019 年 12 月，下同。

作为对比，**图 III5** 报告了 NF-LDA 定价核$x_{LDA}^{MVE}$对宏观变量的预测结果，发现其脉冲响应在趋势上与**图 III4** 一致，但置信区间更大。例如，$x_{LDA}^{MVE}$对消费者信心指数也具有显著为正的预测能力。然而，其对产出、消费和通胀的影响系数在方向性和显著性方面均不及$x_{LLM}^{MVE}$。综上，本文的结果表明在状态变量影响宏观基本面的检验方面，基于 LLMs 构建的叙事资产定价模型相较于 LDA 这类传统文本分析方法具有优越性。

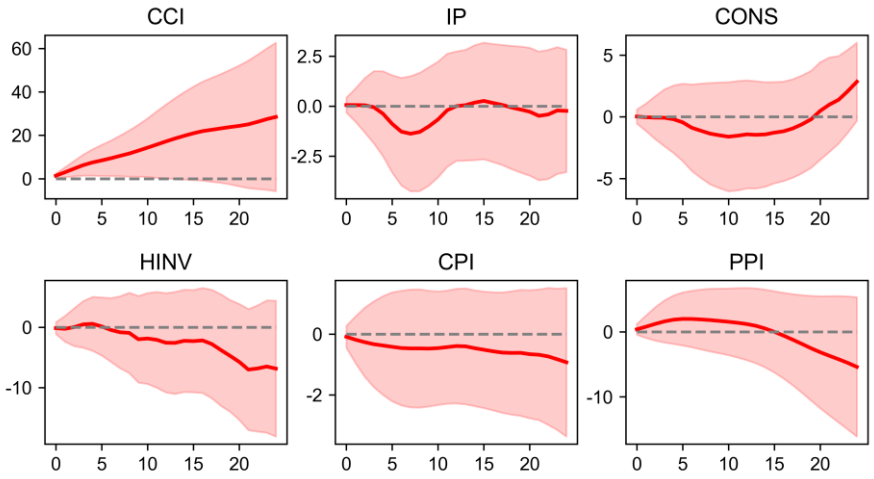


**图 III5　NF-LDA 的状态变量定价核对宏观经济变量的预测能力**

针对叙事定价因子的性质，本文考察其相较于传统 FFC 因子是否具有额外的超额收益，以及其在传统因子上的暴露程度。具体地，我们将叙事因子的 MVE 组合收益率（$f^{MVE}$）作为被解释变量，传统的 Fama-French 五因子以及 Carhart 动量因子作为解释变量，逐一进行时序回归：

$$f_t^{MVE} = \alpha + \beta_{MKT} MKT_t + \beta_{SMB} SMB_t + \beta_{HML} HML_t + \beta_{UMD} UMD_t + \beta_{RMW} RMW_t + \beta_{CMA} CMA_t + e_{v,t} \quad \text{(III.2)}$$

其中，截距项（$\alpha$）表示叙事因子的超额收益，斜率系数$\beta$体现了叙事因子在传统因子上的暴露程度。

**表 III2** 的 Panel A 和 B 分别报告了 NF-LLM 模型的叙事因子（$f_{LLM}^{MVE}$）和 NF-LDA 模型的叙事因子（$f_{LDA}^{MVE}$）的回归结果，其中第$K$列表示由前$K$个因子张成的 MVE 组合。在超额收益方面，当因子数量较少（$K=1$或2）时，模型因欠拟合未能充分吸收文本信息，$f_{LLM}^{MVE}$和$f_{LDA}^{MVE}$相较于 FFC6 因子未能产生显著的超额收益[9]。然而，在$K \geq 3$时，NF-LLM 模型的超额收益显著为正，且在经济意义上同样重要。例如，在控制 FFC6 因子后，基于 LLMs 的三因子模型年化超额回报达 0.84%（=0.07%×12），而五因子模型可提升至 1.20%。对于 NF-LDA 模型而言，研究发现其在绝大多数情形下相对 FFC6 模型没有显著为正的$\alpha$，并且其数值在经济意义上也相对有限。

在因子暴露方面，NF-LDA 的叙事因子主要在 Fama-French 三因子（MKT、SMB 和 HML）

[9] 由于$K=1$时叙事因子的估计结果与惩罚高维文本信息的系数$\lambda$几乎无关，所以 NF-LLM 和 NF-LDA 估计出的第一个叙事因子是高度接近的。当$K=2$时，尽管两个模型估计出的第二个叙事因子并不相同，但由于第一个因子的高度相关性，两个因子的线性 MVE 组合也较为相关，导致$\alpha$的估计结果相似。

上具有显著暴露，而 NF-LLM 的因子不仅与三因子相关，还对动量（UMD）和盈利（RMW）因子具有显著载荷，表明其能够捕捉更广泛的股票特征风险。此外，传统因子对叙事因子的解释力度较低，回归$R^2$普遍较小，同时 GRS 统计量较大，这表明 FFC6 因子无法充分捕捉叙事因子的收益特征。综上，可以认为 NF-LLM 因子相较于 FFC6 定价因子具有显著的增量信息，而 NF-LDA 模型在 A 股市场的表现相对一般。

**表 III2　叙事因子在传统 FFC 因子上的超额收益和暴露程度**

| Panel A：因变量$f_{LLM}^{MVE}$ | | | | | | |
|---|---|---|---|---|---|---|
| $K$ | 1 | 2 | 3 | 4 | 5 | 6 |
| $\alpha$(%) | -0.02 | 0.22 | 0.07 | 0.07 | 0.10 | 0.03 |
| | (-0.02) | (0.62) | (2.18) | (2.48) | (2.79) | (1.68) |
| $\beta_{MKT}$ | 1.99 | 1.07 | 0.11 | 0.12 | 0.07 | 0.02 |
| | (2.41) | (2.55) | (1.74) | (2.49) | (1.50) | (0.90) |
| $\beta_{SMB}$ | 3.54 | 1.91 | 0.19 | 0.21 | 0.14 | 0.03 |
| | (2.51) | (2.67) | (1.76) | (2.47) | (1.65) | (0.93) |
| $\beta_{HML}$ | -2.10 | -1.13 | -0.10 | -0.12 | -0.07 | -0.02 |
| | (-2.24) | (-2.37) | (-1.46) | (-2.11) | (-1.43) | (-0.85) |
| $\beta_{UMD}$ | -3.31 | -1.96 | -0.34 | -0.36 | -0.33 | -0.01 |
| | (-1.45) | (-1.68) | (-2.75) | (-3.18) | (-2.34) | (-0.23) |
| $\beta_{RMW}$ | 2.72 | 1.58 | 0.24 | 0.25 | 0.28 | -0.00 |
| | (0.89) | (1.05) | (2.59) | (2.62) | (2.02) | (-0.03) |
| $\beta_{CMA}$ | 1.56 | 0.76 | -0.03 | 0.01 | -0.04 | 0.02 |
| | (0.90) | (0.90) | (-0.48) | (0.12) | (-0.63) | (0.42) |
| $R^2$ | 0.091 | 0.093 | 0.053 | 0.068 | 0.103 | 0.018 |
| GRS | | | 333.57 | | | |
| ($p$ 值) | | | (0.00) | | | |
| Panel B：因变量$f_{LDA}^{MVE}$ | | | | | | |
| $K$ | 1 | 2 | 3 | 4 | 5 | 6 |
| $\alpha$(%) | -0.02 | 0.15 | 0.04 | 0.03 | 0.03 | 0.03 |
| | (-0.02) | (0.42) | (0.46) | (0.87) | (1.29) | (1.80) |
| $\beta_{MKT}$ | 1.99 | 1.00 | 0.25 | -0.02 | -0.06 | -0.04 |
| | (2.42) | (2.44) | (2.34) | (-0.77) | (-2.70) | (-2.14) |
| $\beta_{SMB}$ | 3.54 | 1.79 | 0.44 | -0.03 | -0.11 | -0.07 |
| | (2.51) | (2.55) | (2.43) | (-0.67) | (-2.83) | (-2.24) |
| $\beta_{HML}$ | -2.10 | -1.06 | -0.26 | 0.01 | 0.05 | 0.03 |
| | (-2.24) | (-2.28) | (-2.17) | (0.18) | (2.05) | (1.46) |
| $\beta_{UMD}$ | -3.31 | -1.63 | -0.40 | 0.08 | 0.08 | 0.07 |
| | (-1.45) | (-1.43) | (-1.35) | (1.19) | (1.68) | (1.64) |
| $\beta_{RMW}$ | 2.72 | 1.40 | 0.37 | 0.01 | 0.04 | 0.02 |
| | (0.89) | (0.90) | (0.92) | (0.21) | (0.85) | (0.47) |
| $\beta_{CMA}$ | 1.56 | 0.74 | 0.16 | -0.11 | -0.03 | -0.04 |
| | (0.90) | (0.86) | (0.73) | (-1.44) | (-0.88) | (-1.12) |
| $R^2$ | 0.091 | 0.094 | 0.091 | 0.044 | 0.074 | 0.047 |
| GRS | | | 11.17 | | | |
| ($p$ 值) | | | (0.00) | | | |

注：该表报告叙事因子模型（NF-LLM 和 NF-LDA）的定价因子相较于 FFC6 因子的超额收益$\alpha$和暴露程度$\beta$，其中叙事因子个数$K = 1, \cdots, 6$。Panel A 中因变量为 NF-LLM 叙事因子的 MVE 组合（$f_{LLM}^{MVE}$），Panel B 中因变量为 NF-LDA 叙事因子的 MVE 组合（$f_{LDA}^{MVE}$）。$t$统计量通过 Newey and West（1987）标准误计算。样本期为 2008 年 1 月至 2021 年 12 月。

III.3 异象检验的补充分析

本文从两个维度考察叙事因子模型的异质性表现[10]：一是时序维度，即在不同市场状态（牛熊市、投资者情绪高涨或低落）下对比模型的解释能力；二是横截面维度，即分析不同类别的异象是否因经济叙事的变化而导致定价效果的差异。

在时序异质性中，本文首先区分不同的市场状态，然后将数据按时间划分为两类样本分别进行回归分析[11]。图 III 6 汇总了因子模型在不同市场状态下的资产异象检验结果，为清晰对比不同因子模型的表现，本文选取各模型中解释能力最优的版本进行分析。子图（1）显示，各类因子模型在牛市期间的定价效果普遍优于熊市，表现为平均$|t(\alpha)|$值更低、显著比例更小。其中，NF-LLM 模型在熊市中仍保持稳健的定价能力，展现出较强的跨周期适应性。相比之下，NF-LDA 模型虽在牛市中表现良好，但在熊市中其效果未能明显优于 FFC 和 PC 模型。子图（2）表明，FFC 因子模型在投资者情绪高涨时定价效果较弱，这与 Stambaugh et al.（2012）的发现一致。然而，叙事因子模型在此时对异象的解释能力更强，可能源于新闻叙事中蕴含了一定的市场情绪信息（姜富伟等，2021），在情绪高涨时能有效缓解市场中的错误定价。值得注意的是，NF-LLM 模型在不同情绪状态下均表现最佳，说明基于 LLMs 的嵌入向量更能捕捉文本中的情绪变化。综合来看，NF-LLM 在不同市场状态下均展现出稳健且优越的定价能力。

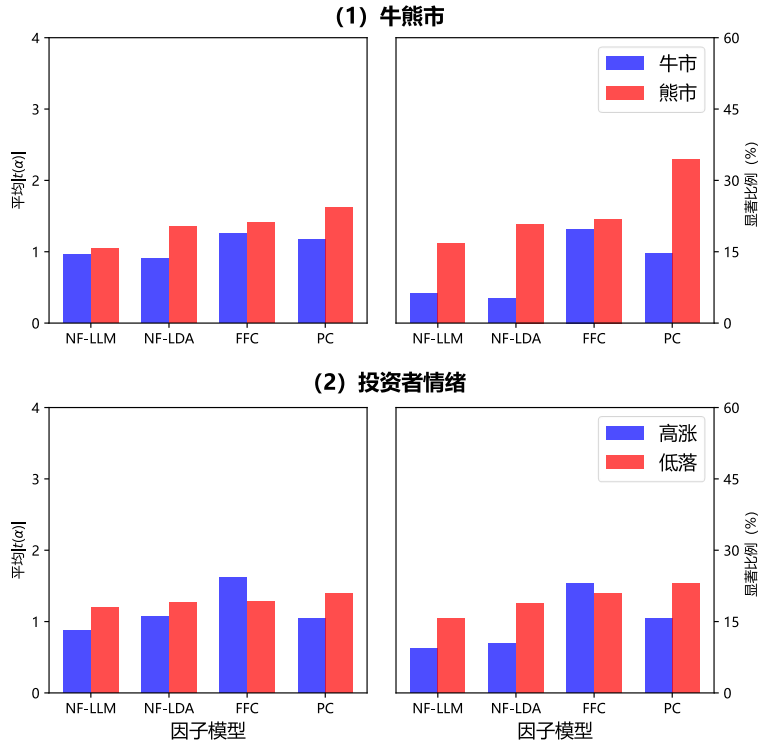


**图 III 6　不同市场状态下因子模型对异象的平均解释能力**

注：上半部分基于牛熊市划分，下半部分基于投资者情绪高低分组。横轴为四类因子模型，包括叙事因子模型（NF-LLM 和 NF-LDA）、基准因子模型（FFC）和统计因子模型（PC）。各模型均选择了具有最佳解释能力的版本；样本期为 2008 年 1 月至 2018 年 11 月，下同。

[10] 感谢审稿专家提出关于资产异象检验异质性分析的宝贵建议。

[11] 对于牛熊市的识别，本文主要参考国内文献的划分结果（肖峻，2013；隋新玉和王云清，2020），并结合 Pagan and Sossounov（2003）提出的波峰-波谷规则进行辅助判断。在本文 2008 年 1 月至 2018 年 11 月的异象检验样本期内，2008-11 – 2009-12、2010-07 – 2011-04、2013-07 – 2015-05、2016-03 – 2018-01 为牛市阶段，其余为熊市阶段。对于投资者情绪高低的判断，我们使用易志高和茅宁（2009）构造的投资者情绪指数（CICSI），数据来源为 CSMAR。参考 Da et al.（2011）的分组标准，若 CICSI 大于过去 8 个月的中位数则判定为投资者情绪高涨时期；反之，则为投资者情绪低落时期。

在横截面异质性中，本文根据李斌等（2019）的分类，将 96 个异象因子划分为六组，分别统计各因子模型对不同类别异象的解释程度。为便于展示，每组选取了解释能力最优的模型进行分析，具体结果见图 III 7。主要结论如下：第一，FFC 因子模型在价值、成长和盈利类异象上的定价能力较强，叙事因子模型在这些类别中的表现同样良好。第二，在交易摩擦、动量和财务流动性等传统因子模型难以解释的异象上，叙事因子模型显著提升了定价效果[12]。第三，整体而言，NF-LLM 模型的解释能力优于 NF-LDA，尤其在动量和财务流动性类异象上，进一步突显了 LLMs 在叙事资产定价研究中的优势。第四，相较于统计因子模型（PC），NF-LLM 模型在定价效果上更为出色，特别是在成长和盈利类异象上。总体来看，基于 LLMs 提取新闻文本信息构建的定价因子，能够全面改善传统定价模型对异象的解释效果。

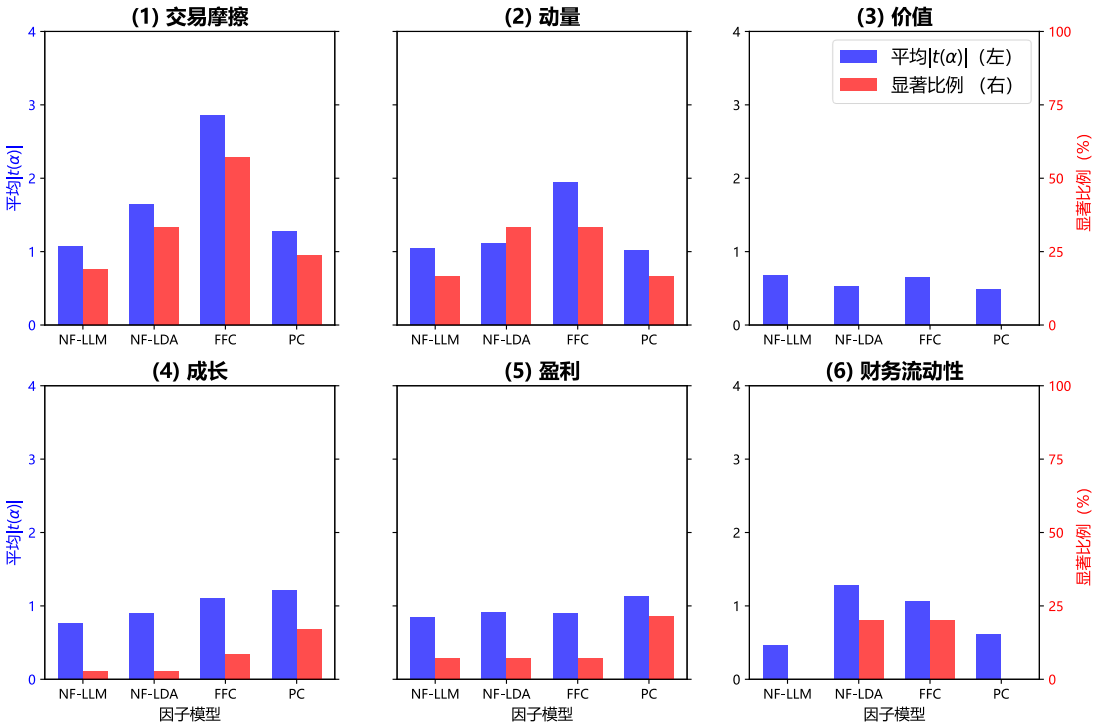


**图 III 7　因子模型对六组异象的平均解释能力**

此外，图 III 8 展示了因子模型对每个异象的解释能力，即异象超额收益率的显著性（$|t(\alpha)|$）。横轴为不同的因子模型，纵轴为 96 个异象，其名称及其具体含义可参考李斌等（2019）附录中的详细说明。“NF-LLM1”指的是只包含一个因子的 NF-LLM 模型，以此类推。颜色越深表示$|t(\alpha)|$值越大，即该异象的超额收益在给定模型下仍具有较高的统计显著性，说明模型对该异象的解释力度较弱；相反，颜色较浅则表明模型能够更有效地解释该异象。总体来看，叙事因子模型在异象的解释效果上最为突出，其$|t|$值显著低于其他基准模型。具体而言，对于传统因子模型和统计因子模型难以解释的交易摩擦类异象，如交易换手率波动率（*std_turn*）、交易额波动率（*std_vol*）和标准化换手率（*LM*），叙事因子模型，尤其是$K = 5\sim6$的 NF-LLM 模型，能够提供更有效的解释。

[12] 大量文献表明，这些异象通常源于市场不完全有效，反映了投资者行为偏差、套利限制或信息不对称等错误定价因素。如前文所述，新闻文本的定价效力可能与错误定价理论密切相关，这也解释了叙事因子模型在这类异象上的优越表现。


**图 III 8　每个异象的显著性水平（$|t(\alpha)|$）**

III.4 高频叙事因子模型

在高频（周度）模型训练前，本文对输入数据做出一些必要调整。一是每年仅保留前 52 周的样本。这一操作主要考虑到实践操作的便利性，同时避免了数据量的过度损失。二是考虑到一周内的交易日数量有限[13]，仍使用月内数据计算的样本协方差$cov_{i,t}$，并通过线性插值生成周度数据。完成上述调整后，本文在周度层面训练叙事因子模型，并得到相应的周度因子估计值，然后进行同样的资产异象检验和样本外交易策略[14]。

正文的结果表明，周度叙事因子模型在整体上对 96 个异象的平均解释能力优于月度模型。那么，在不同类型的异象中，周度模型是否依然占优？图 III 9 的结果给出了肯定的答案。首先，从平均显著性水平（$|t(\alpha)|$）来看，周度 NF-LDA 模型在绝大多数类型中均低于相应的月度模型。更重要的是，月度 NF-LDA 模型对交易摩擦、动量和财务流动性异象的解释能力相对较弱，而周度模型则显著提升了对这些异象的定价效果。其次，与月度 NF-LLM 模型相比，周度 NF-LDA 模型在交易摩擦和动量类别中的解释能力进一步增强。这一结果可归因于高频

[13] 在月度模型中，$cov_{i,t}$的估计值是利用日度的叙事新息和个股收益率数据计算当月内的样本协方差而得的。然而，由于周内通常只有 5 个交易日，且部分周的交易日甚至少于 3 天，直接计算叙事信息与股票收益率的样本协方差可能导致较大误差。

[14] 在资产异象检验方面，由于叙事因子为周度频率，而异象收益率是月度的，我们参考混频数据抽样（MIDAS）模型的思想，先将周度因子通过月内平均的方式转换为月度频率，然后再对异象进行回归分析。交易策略的构造方式与月度模型保持一致，计算年化夏普比率时年化系数取$\sqrt{52}$。

数据在捕捉短期市场动态和交易行为上的优势（Nakamura and Steinsson，2018；陈贞竹等，2023）。这是因为交易摩擦和动量异象通常与短期市场波动及投资者行为密切相关，而周度频率的叙事因子模型能够更及时地整合新闻叙事信息，从而提高对这些异象的解释能力。

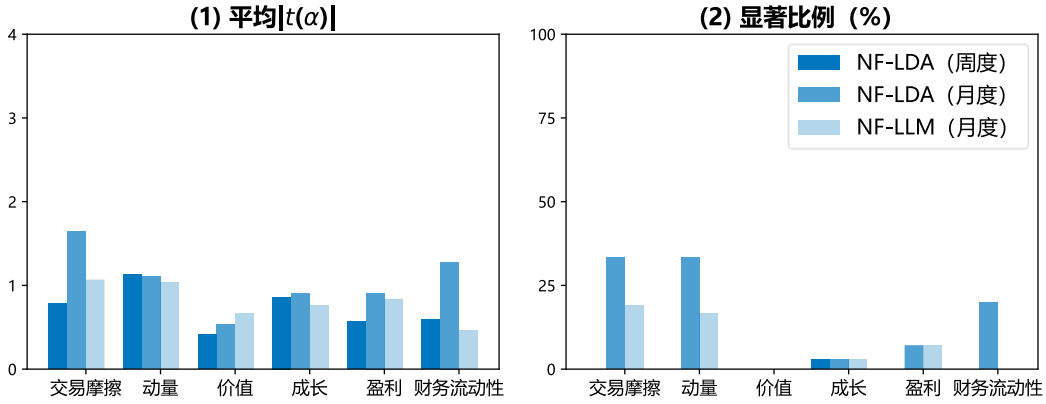


**图 III 9　周度与月度叙事因子模型的异象分组检验**

III.5 叙事因子与传统因子的联合模型

正文发现，叙事文本的加入为资产定价提供了额外的信息增量，显著增强了基于传统股票特征的因子模型的解释能力。这些发现引发了一个重要的理论问题：文本信息与股票特征之间究竟是互补关系还是替代关系？从理论基础来看，股票特征因子的构建依托于风险补偿与错误定价的双重支柱[15]，而文本叙事视角能够为这一理论体系提供更丰富的解释维度。一方面，文本信息可以反映市场对未来经济状况的预期，揭示经济波动、行业趋势和政策调整等系统性风险来源，进而补充传统因子对风险溢价的刻画。另一方面，文本信息能够影响投资者的预期和感知，通过放大市场情绪波动或强化认知偏差，导致资产价格的错误定价，这为行为金融相关的因子提供了新的解释维度。从指标构建来看，文本信息与股票特征信息的结合可以体现在量化市场情绪和微观主体预期等“软信息”，从而对传统股票特征所反映的财务指标、估值水平和公司规模等“硬信息”形成有效补充。

在本部分，为检验本文所构建叙事因子（NF）在资产定价中的稳健性与增量解释能力，我们在正文 NF-LLM 模型基础上，进一步引入 NF-LDA 模型构建的叙事因子，并与经典的 FFC6 因子模型进行联合建模。联合模型的资产定价结果在表 III 3 中报告。在 Panel A 中，随着纳入的叙事因子个数（$K$）增加，各联合模型的样本外交易策略表现整体呈现上升趋势。尤其在$K = 4$时，夏普比率基本达到最大值。NF-LLM 与 FFC6 的联合模型其定价效果稳定优于 NF-LDA 的相应模型，印证了 NF-LLM 方法在资产定价建模中的优势。

在 Panel B 中，我们补充了联合因子模型的资产异象检验结果。通过与正文中基准模型（如 MKT、FF3、FF5 等）的对比分析，可以发现联合模型的定价能力相较于基准因子模型得到明显提升。与之前一致，基于 NF-LDA 的联合模型整体上仍逊色于基于 NF-LLM 的联合模型。上述结果共同表明，叙事文本的加入为资产定价提供了额外的信息增量，显著增强了基于传统股票特征的因子模型的解释能力。

[15] 例如，在经典的 FFC 六因子中，市场因子反映了系统性风险补偿，Fama-French 因子捕捉了与特定投资风格相关的系统性风险溢价，而动量因子则与错误定价或行为偏差有关。

**表 III 3　联合因子模型的资产定价效果**

Panel A：NF-LDA 与 FFC 的联合模型（交易策略）

| $K$ | 0 | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|---|
| 仅 NF-LDA | | 0.06 | 0.27 | 0.34 | 0.88 | 0.81 | 0.74 |
| NF + CAPM | 0.18 | 0.28 | 0.38 | 0.35 | 0.83 | 0.57 | 0.67 |
| NF + FF3 | 0.04 | 0.38 | 0.45 | 0.38 | 0.81 | 0.52 | 0.60 |
| NF + FF5 | 0.24 | 0.39 | 0.54 | 0.72 | 0.90 | 0.67 | 0.79 |
| NF + FFC6 | 0.27 | 0.42 | 0.57 | 0.76 | 0.86 | 0.70 | 0.83 |

Panel B：资产异象检验

| 基准因子 | 叙事因子 | NF-LDA | | | | NF-LLM | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | 平均$\alpha$ | 平均$\lvert t(\alpha)\rvert$ | 显著比例 | GRS | 平均$\alpha$ | 平均$\lvert t(\alpha)\rvert$ | 显著比例 | GRS |
| MKT | NF1 | 0.15 | 1.57 | 22.92 | 6.91 | 0.15 | 1.57 | 22.92 | 6.91 |
| | NF2 | 0.13 | 1.23 | 18.75 | 4.75 | 0.13 | 1.31 | 18.75 | 5.11 |
| | NF3 | 0.14 | 1.18 | 15.62 | 4.42 | 0.08 | 1.24 | 19.79 | 3.65 |
| | NF4 | 0.14 | 1.19 | 15.62 | 4.24 | 0.07 | 1.25 | 18.75 | 2.71 |
| | NF5 | 0.15 | 1.21 | 16.67 | 4.61 | 0.01 | 0.91 | 10.42 | 2.38 |
| | NF6 | 0.15 | 1.17 | 17.71 | 4.26 | 0.02 | 0.88 | 10.42 | 2.32 |
| FF3 | NF1 | 0.06 | 1.90 | 31.25 | 6.48 | 0.06 | 1.90 | 31.25 | 6.48 |
| | NF2 | 0.08 | 1.65 | 26.04 | 4.55 | 0.05 | 1.62 | 26.04 | 4.83 |
| | NF3 | 0.08 | 1.49 | 23.96 | 4.21 | 0.02 | 1.65 | 28.12 | 3.49 |
| | NF4 | 0.08 | 1.51 | 23.96 | 4.06 | 0.01 | 1.72 | 33.33 | 2.57 |
| | NF5 | 0.08 | 1.52 | 22.92 | 4.61 | 0.05 | 1.18 | 16.67 | 2.31 |
| | NF6 | 0.09 | 1.52 | 22.92 | 4.23 | 0.07 | 1.16 | 15.62 | 2.25 |
| FF5 | NF1 | 0.08 | 2.14 | 39.58 | 6.62 | 0.08 | 2.14 | 39.58 | 6.62 |
| | NF2 | 0.10 | 1.94 | 35.42 | 4.73 | 0.08 | 1.96 | 31.25 | 4.65 |
| | NF3 | 0.09 | 1.75 | 31.25 | 4.43 | 0.06 | 1.92 | 34.38 | 3.40 |
| | NF4 | 0.10 | 1.76 | 33.33 | 4.20 | 0.06 | 1.86 | 37.50 | 2.55 |
| | NF5 | 0.10 | 1.73 | 33.33 | 4.49 | 0.03 | 1.24 | 17.71 | 2.21 |
| | NF6 | 0.12 | 1.77 | 32.29 | 4.10 | 0.05 | 1.22 | 16.67 | 2.15 |
| FFC6 | NF1 | 0.08 | 2.32 | 43.75 | 6.49 | 0.08 | 2.32 | 43.75 | 6.49 |
| | NF2 | 0.10 | 2.04 | 35.42 | 4.68 | 0.07 | 2.10 | 31.25 | 4.54 |
| | NF3 | 0.09 | 1.88 | 32.29 | 4.40 | 0.06 | 1.99 | 34.38 | 3.31 |
| | NF4 | 0.10 | 1.86 | 34.38 | 4.17 | 0.05 | 1.98 | 36.46 | 2.47 |
| | NF5 | 0.10 | 1.81 | 35.42 | 4.40 | 0.07 | 1.18 | 14.58 | 2.14 |
| | NF6 | 0.11 | 1.87 | 34.38 | 4.01 | 0.08 | 1.14 | 14.58 | 2.08 |

III.6 叙事主题与收益率信息

在标准资产定价框架中，股票的未预期收益可以表示为未来现金流（Cash flow，CF）信息与贴现率（Discount rate，DR）信息之差。大量金融学文献考虑了股市对两类收益率信息的反应，以及影响这两类信息的潜在因素。新闻报道作为金融市场中信息传递的重要渠道，在市场定价过程中发挥关键作用。然而，新闻叙事究竟如何影响资产收益率？是通过改变投资者对公司未来现金流的预期，还是通过调整市场贴现率来影响风险定价？为了回答这一问题，本文尝试性地考察叙事主题与现金流和贴现率信息的关系，明确其对资产收益率影响的作用机制。

首先，我们在市场层面上估计两类收益率信息[16]。沿用 Campbell（1991）和 Campbell and Vuolteenaho（2004）等文献的经典做法，本文使用向量自回归（VAR）模型将股票超额收益率$r_t$分解为两个时间序列——现金流信息$N_{CF,t}$和贴现率信息$N_{DR,t}$。具体地，假设数据由如下一阶 VAR 系统生成：

$$y_t = a + \mathbf{B}y_{t-1} + u_t \tag{III.3}$$

[16] 由于新闻属于总体层面的宏观新闻，为确保匹配性，我们同样在市场层面提取收益率信息。

其中，$y_t$是$m \times 1$维状态向量（state vector），其第一个元素为$r_t^m$；$a$和$\mathbf{B}$分别为$m \times 1$维常数向量和$m \times m$维系数矩阵，$u_t$是$m \times 1$维冲击向量。现金流信息和贴现率信息是同期冲击向量$u_t$的线性函数：

$$N_{CF,t} = (e1' + e1'\boldsymbol{\lambda})u_t \tag{III.4}$$

$$N_{DR,t} = e1'\boldsymbol{\lambda}u_t \tag{III.5}$$

其中，向量$e1$中的第一个元素为 1，其余元素均为 0；矩阵$\boldsymbol{\lambda} \equiv \rho\mathbf{B}(\mathbf{I} - \rho\mathbf{B})^{-1}$将冲击向量映射到两类信息，$\mathbf{I}$为$m \times m$维单位矩阵。

参考 Campbell and Vuolteenaho（2004）、Hecht and Vuolteenaho（2006）和 Wu et al.（2021），本文使用上证指数（或深证成指）的超额收益率、对数市盈率、对数股息支付率和期限溢价作为状态向量[17]。数据来源为 CSMAR，样本期为 2006 年 12 月至 2021 年 12 月，以保证收益率信息的第一个估计值可以从 2007 年 1 月开始。在估计过程中，首先利用 OLS 方法获得上述四变量 VAR 系统的系数矩阵$\mathbf{B}$和冲击项$u_t$的估计；随后参考 Campbell and Vuolteenaho（2004）的研究，设定参数$\rho = 0.95^{1/12}$以计算映射矩阵$\boldsymbol{\lambda}$；最后，基于公式（III.4）和（III.5）分别计算现金流信息$N_{CF}$和贴现率信息$N_{DR}$。

为考察叙事主题与两类收益率信息的关系，本文建立如下回归方程：

$$N_{j,t+h} = b_{j,0} + \sum_{l=1}^{L} b_{j,l} topic_{l,t} + \epsilon_{j,t},\ j \in \{CF, DR\} \tag{III.6}$$

其中，$topic_{l,t},\ l = 1, \cdots, L$为主题$l$在月度$t$的分布频率[18]，主题总数$L = 30$。在模型设定方面，本文针对预测期限（$h = 0, 1, 3$和$12$）分别构建回归方程，样本区间覆盖 2007 年 1 月至 2021 年 12 月共计 180 个月度观测样本。由于主题序列的维度相对较高且并非所有主题都与收益率信息相关，我们参考 Larsen et al.（2021）的做法，采用 LASSO 回归对式（III.6）进行估计，通过 5 折交叉验证确定最优正则化参数，并将所有变量标准化。

表 III 4 汇总了 LDA 所识别的叙事主题与现金流信息和贴现率信息的 LASSO 回归结果，其中“非零系数比例”为回归系数中非零元素（$b_{j,l} \neq 0$）的个数，“调整 $R^2$”是对非零系数变量集进行 post-LASSO 程序计算而得，“关键主题比例”为系数非零的关键叙事主题个数，“偏 $R^2$（关键主题）”衡量了关键主题对 post-LASSO 拟合效果的贡献程度。总体来看，叙事主题与现金流信息更为相关，这可能是因为与宏观经济运行、企业盈利状况等基本面相关的宏观经济叙事更易被新闻文本直接捕捉。这也说明新闻叙事通过反映宏观经济和市场状态，改变投资者对公司未来现金流的预期，从而影响资产价格。

此外，NF-LDA 模型识别的关键叙事主题具有重要作用。以 Panel A 中上证指数的结果为例，在筛选的 12 个非零系数主题中，关键主题占比达 42%（≈5/12），高于全部主题中关键主题的占比（10/30，即三分之一）；在 post-LASSO 回归中，关键主题的偏 $R^2$ 与总体的调整 $R^2$ 几乎相当，表明关键主题对现金流信息解释力的边际贡献接近于全部主题的联合效应。Panel B 至 D 的结果表明上述发现对于不同期限的回归也是稳健的，即叙事主题对 A 股市场的现金流信息展现出一定的长期解释能力。

[17] 市盈率和股息支付率采用市值加权，期限溢价的计算方式为 10 年期国债收益率减去 3 月期国债收益率。

[18] 由于 LDA 输出的是文档层面的统计指标，本文采用算术平均方法对一个月内的所有新闻文本主题分布进行聚合计算以得到月度值。

**表 III 4 叙事主题与两类收益率信息的 LASSO 回归结果**

| | 上证指数 | | 深证成指 | |
|---|---|---|---|---|
| | 现金流信息$N_{CF}$ | 贴现率信息$N_{DR}$ | 现金流信息$N_{CF}$ | 贴现率信息$N_{DR}$ |
| Panel A: $h=0$ | | | | |
| 非零系数个数 | 12 | 1 | 11 | 0 |
| 调整 $R^2$ | 0.15 | 0.03 | 0.14 | 0 |
| 关键主题个数 | 5 | 0 | 5 | 0 |
| 偏 $R^2$（关键主题） | 0.15 | 0 | 0.07 | 0 |
| Panel B: $h=1$ | | | | |
| 非零系数个数 | 2 | 0 | 6 | 1 |
| 调整 $R^2$ | 0.10 | 0 | 0.10 | 0.02 |
| 关键主题个数 | 1 | 0 | 3 | 0 |
| 偏 $R^2$（关键主题） | 0.08 | 0 | 0.08 | 0 |
| Panel C: $h=3$ | | | | |
| 非零系数个数 | 1 | 1 | 1 | 1 |
| 调整 $R^2$ | 0.01 | 0.02 | 0.02 | 0.02 |
| 关键主题个数 | 0 | 0 | 0 | 0 |
| 偏 $R^2$（关键主题） | 0 | 0 | 0 | 0 |
| Panel D: $h=12$ | | | | |
| 非零系数个数 | 3 | 0 | 4 | 1 |
| 调整 $R^2$ | 0.06 | 0 | 0.06 | 0.05 |
| 关键主题个数 | 2 | 0 | 2 | 0 |
| 偏 $R^2$（关键主题） | 0.06 | 0 | 0.04 | 0 |

III.7 稳健性检验

（1）控制嵌入向量中写作风格的影响

由于本文使用 BERT 模型提取整篇新闻文档的嵌入向量，而非各个词语的词语向量，因此文档的写作风格、报道偏好和语言习惯等因素可能会在嵌入向量中被隐式捕捉，从而影响其对经济叙事信息的表征。为解决上述问题，我们借鉴 Shapiro et al.（2022）、张一帆等（2023）的思路，对嵌入向量$\theta_{j,m,\tau}^{k}, k=1,\cdots,768$的各个维度分别进行时间-媒体层面的双向固定效应回归：

$$\theta_{j,m,\tau}^{k}=fe_{\tau(j)}^{k}+fe_{m(j)}^{k}+\epsilon_{j,m,\tau}^{k} \tag{III.7}$$

其中，$j,m$和$\tau$分别表示文档、媒体和时间（日），$k$表示嵌入向量的某一维度，$m(j)$对应文档$j$由媒体$m$发布，$\tau(j)$对应文档$j$的发布时间，$\epsilon_{j,m,\tau}^{k}$为残差项[19]。本文假设媒体固定效应$fe_{m(j)}^{k}$吸收了不同媒体写作风格的差异，将时间固定效应$fe_{\tau(j)}^{k}$作为日度层面所有新闻文本信息的聚合（aggregation）。其合理性在于写作风格因素与媒体自身特征密切相关，而与市场和经济环境变化相关的信息则通过时间固定效应得以保留。进一步地，在$k$维度上将其张成向量，从而得到剔除写作风格因素的文档嵌入$\tilde{\theta}_{\tau}\equiv[fe_{\tau}^{1},\cdots,fe_{\tau}^{768}]$。最后，利用$\tilde{\theta}_{\tau}$重新估算月度协变量（工具）$cov_{it}$并再次训练 NF-LLM 模型。

表 III 5 对比了剔除写作风格因素后的 NF-LLM 模型与基准模型的资产定价效果。数据显示，剔除写作风格因素后的模型在异象检验中平均$|t(\alpha)|$和显著比例普遍更低（例如，其三因子模型已优于 FFC6 模型），表明模型在异象解释方面具有更强的能力。交易策略的检验同样稳健：年化夏普比率最高可提升至 1.12，展现出更稳健的风险收益表现。因此，在剔除写作风格因素后，NF-LLM 模型依然保持良好的定价效果，甚至在部分情况下优于基准模型。

[19] 在本文的样本中，共涉及 7 家不同的财经媒体，故$m=1,\cdots,7$。在实际操作中，由于全样本面板数据量较大，我们选择逐月对式（III.7）进行回归，这也与协变量$cov_{it}$基于月度数据估算的做法一致。

**表 III 5　剔除写作风格因素后的 NF-LLM 模型的资产定价效果**

| | NF-LLM（剔除写作风格因素） | | | NF-LLM（基准） | | |
|---|---|---|---|---|---|---|
| | 异象检验 | | 交易策略 | 异象检验 | | 交易策略 |
| $K$ | 平均$\|t(\alpha)\|$ | 显著比例（%） | 年化夏普比率 | 平均$\|t(\alpha)\|$ | 显著比例（%） | 年化夏普比率 |
| 1 | 1.55 | 22.92 | 0.00 | 1.58 | 22.92 | 0.06 |
| 2 | 1.37 | 19.79 | 0.71 | 1.29 | 17.71 | 0.68 |
| 3 | 0.73 | 4.17 | 0.80 | 1.24 | 19.79 | 0.86 |
| 4 | 0.83 | 3.12 | 1.12 | 1.23 | 18.75 | 0.90 |
| 5 | 0.83 | 4.17 | 0.66 | 0.88 | 10.42 | 0.86 |
| 6 | 0.83 | 4.17 | 0.99 | 0.86 | 9.38 | 0.95 |

（2）NF-LDA 模型主题数的影响

进一步地，我们验证 NF-LLM 相对于 NF-LDA 具有优越性不受 LDA 主题数量的影响。为避免计算的复杂度，选择文献中广泛使用的主题数（如 80，120 和 180）拟合 LDA 模型，然后构建相应的叙事因子模型。表 III 6 报告了使用不同主题数构建的 NF-LDA 模型的样本外交易策略表现。结果表明：其一，即使选择不同主题数构建 LDA 模型，NF-LLM 均优于 NF-LDA，这表明本文的主要结论是稳健的。其二，基准结果（$L=30$）的 NF-LDA 已经可以得到较高的夏普比率，而继续增大主题数量并不能在很大程度上改善 NF-LDA 的定价效果（甚至往往会恶化），这验证了本文选择的主题数量具有合理性。

**表 III 6　不同主题数 NF-LDA 模型的最优样本外年化夏普比率**

| | NF-LLM | NF-LDA | | | |
|---|---|---|---|---|---|
| 嵌入向量维度/主题数（$L$） | 768（基准） | 30（基准） | 80 | 120 | 180 |
| 最优夏普比率 | 0.95 | 0.88 | 0.74 | 0.73 | 0.88 |

（3）其他基准定价因子

表 III 7 至表 III 10 报告了叙事因子模型（NF-LLM 和 NF-LDA）相较于 FFC6、中国四因子（CH4）、$q$ 因子（HXZ4）以及长短期行为三因子（DHS3）的检验结果，其中 Panel A 的因变量为叙事因子 MVE 组合，Panel B 的因变量为基准因子。

表 III 7 的 Panel A 与正文完全一致，结果显示为 NF-LLM 模型相较于基准因子模型有额外信息。Panel B 从另一个方向支撑了上述结论，可以发现所有 FFC6 因子对 NF-LLM 叙事因子（$K=1\sim6$）均不显著；更重要的是，GRS 检验联合检验不显著，即 FFC6 在叙事因子模型中不存在无法解释系统性风险溢价。

从表 III 8 的 Panel A 中可以看到，NF-LDA 模型在$K=1\sim6$相对 CH4 模型均没有显著的正$\alpha$；而 NF-LLM 模型尽管在$K=1\sim4$时相对 CH4 没有显著的正$\alpha$，但在$K\geq5$后有显著的正$\alpha$。在 Panel B 中，CH4 因子相对于叙事因子也不存在无法解释系统性风险溢价。

进一步地，表 III 9 以 Hou et al.（2015）的四因子模型（$q$-factor model，HXZ4）作为比较基准。在 Panel A 中，在$K=5$时 NF-LLM 相对 HXZ4 有显著的正$\alpha$，且 GRS 统计量表明$\alpha$在$K=1\sim6$的 MVE 组合中联合显著不为 0。尽管以 HXZ4 为基准，NF 模型的优势相对弱一些，但在部分设定下相对基准仍有增量信息。在 Panel B 中，可以看出所有的 HXZ4 因子相对 NF-LLM 均没有显著的正$\alpha$，且 GRS 统计量指示联合不显著，进一步说明叙事因子相对 HXZ4 有边际定价能力的提高。

最后，与 Daniel et al.（2020）提出的长短期行为三因子模型（DHS3）相比（表 III 10），在 Panel A 中，NF-LLM 五因子相对于 DHS3 因子表现出显著的正$\alpha$，且 GRS 统计量表明$\alpha$值联合显著不为 0。在 Panel B 中，所有 DHS3 因子相对于 NF-LLM 均未表现出显著的正$\alpha$，且 GRS 统计量指示联合不显著，进一步表明叙事因子相较于 DHS3 在边际定价能力上有所提升。

**表 III 7　叙事因子对 FFC6 因子的定价表现**

| | Panel A　因变量：叙事因子的 MVE 组合 | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | LDA | | | | | | LLM | | | | | |
| $f_{1\sim i}^{MVE}$ | 1 | 2 | 3 | 4 | 5 | 6 | 1 | 2 | 3 | 4 | 5 | 6 |
| $\alpha(\%)$ | -0.02 | 0.15 | 0.04 | 0.03 | 0.03 | 0.03 | -0.02 | 0.22 | 0.07 | 0.07 | 0.10 | 0.03 |
| | (-0.02) | (0.42) | (0.46) | (0.87) | (1.29) | (1.80) | (-0.02) | (0.62) | (2.18) | (2.48) | (2.79) | (1.68) |
| $\beta_{MKT}$ | 1.99 | 1.00 | 0.25 | -0.02 | -0.06 | -0.04 | 1.99 | 1.07 | 0.11 | 0.12 | 0.07 | 0.02 |
| | (2.42) | (2.44) | (2.34) | (-0.77) | (-2.7) | (-2.14) | (2.41) | (2.55) | (1.74) | (2.49) | (1.50) | (0.90) |
| $\beta_{SMB}$ | 3.54 | 1.79 | 0.44 | -0.03 | -0.11 | -0.07 | 3.54 | 1.91 | 0.19 | 0.21 | 0.14 | 0.03 |
| | (2.51) | (2.55) | (2.43) | (-0.67) | (-2.83) | (-2.24) | (2.51) | (2.67) | (1.76) | (2.47) | (1.65) | (0.93) |
| $\beta_{HML}$ | -2.10 | -1.06 | -0.26 | 0.01 | 0.05 | 0.03 | -2.10 | -1.13 | -0.10 | -0.12 | -0.07 | -0.02 |
| | (-2.24) | (-2.28) | (-2.17) | (0.18) | (2.05) | (1.46) | (-2.24) | (-2.37) | (-1.46) | (-2.11) | (-1.43) | (-0.85) |
| $\beta_{UMD}$ | -3.31 | -1.63 | -0.40 | 0.08 | 0.08 | 0.07 | -3.31 | -1.96 | -0.34 | -0.36 | -0.33 | -0.01 |
| | (-1.45) | (-1.43) | (-1.35) | (1.19) | (1.68) | (1.64) | (-1.45) | (-1.68) | (-2.75) | (-3.18) | (-2.34) | (-0.23) |
| $\beta_{RMW}$ | 2.72 | 1.40 | 0.37 | 0.01 | 0.04 | 0.02 | 2.72 | 1.58 | 0.24 | 0.25 | 0.28 | -0.00 |
| | (0.89) | (0.90) | (0.92) | (0.21) | (0.85) | (0.47) | (0.89) | (1.05) | (2.59) | (2.62) | (2.02) | (-0.03) |
| $\beta_{CMA}$ | 1.56 | 0.74 | 0.16 | -0.11 | -0.03 | -0.04 | 1.56 | 0.76 | -0.03 | 0.01 | -0.04 | 0.02 |
| | (0.90) | (0.86) | (0.73) | (-1.44) | (-0.88) | (-1.12) | (0.90) | (0.90) | (-0.48) | (0.12) | (-0.63) | (0.42) |
| $R^2$ | 0.091 | 0.094 | 0.091 | 0.044 | 0.074 | 0.047 | 0.091 | 0.093 | 0.053 | 0.068 | 0.103 | 0.018 |
| GRS | 11.17 | | | | | | 333.57 | | | | | |
| ($p$ 值) | (0.00) | | | | | | (0.00) | | | | | |

（续表）

Panel B　因变量：FFC6

| | LDA | | | | | | LLM | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | MKT | SMB | HML | UMD | RMW | CMA | MKT | SMB | HML | UMD | RMW | CMA |
| $\alpha$(%) | 0.47 | -0.05 | 0.33 | 0.07 | -0.04 | 0.01 | 1.12 | -0.46 | 0.59 | -0.42 | -0.43 | -0.40 |
| | (0.74) | (-0.11) | (1.27) | (0.59) | (-0.37) | (0.10) | (0.42) | (-0.32) | (0.91) | (-1.13) | (-1.14) | (-1.13) |
| $\beta_{f1}$ | 3.13 | -2.02 | -0.30 | -0.10 | -0.11 | -0.02 | 4.51 | -2.86 | 0.83 | -1.99 | -1.56 | -1.52 |
| | (1.81) | (-2.27) | (-0.35) | (-0.48) | (-0.60) | (-0.10) | (0.49) | (-0.62) | (0.36) | (-1.50) | (-1.19) | (-1.24) |
| $\beta_{f2}$ | -6.45 | 3.43 | -0.86 | 0.61 | 0.69 | 0.70 | -9.32 | 6.00 | -1.66 | 4.18 | 3.28 | 3.19 |
| | (-2.17) | (2.40) | (-0.63) | (1.56) | (2.07) | (2.65) | (-0.49) | (0.62) | (-0.35) | (1.52) | (1.21) | (1.26) |
| $\beta_{f3}$ | 0.93 | 2.57 | 5.86 | -1.41 | -1.63 | -2.41 | -4.39 | 2.55 | -0.77 | 0.95 | 0.26 | 0.19 |
| | (0.07) | (0.46) | (1.27) | (-1.04) | (-1.22) | (-1.70) | (-0.55) | (0.60) | (-0.52) | (1.31) | (0.48) | (0.46) |
| $\beta_{f4}$ | -4.32 | 1.81 | -1.00 | -0.33 | -0.43 | -0.44 | 12.78 | -7.62 | 2.27 | -4.10 | -2.67 | -2.52 |
| | (-1.74) | (1.48) | (-1.32) | (-1.10) | (-1.69) | (-1.99) | (0.63) | (-0.71) | (0.57) | (-1.72) | (-1.26) | (-1.31) |
| $\beta_{f5}$ | -5.33 | 1.84 | -1.42 | 0.10 | 0.41 | 0.27 | -0.70 | -0.10 | 0.53 | -1.84 | -1.21 | -1.20 |
| | (-1.46) | (1.06) | (-1.05) | (0.22) | (0.94) | (0.76) | (-0.09) | (-0.03) | (0.25) | (-1.76) | (-1.18) | (-1.28) |
| $\beta_{f6}$ | 12.49 | -6.74 | -0.06 | 0.91 | 0.75 | 0.97 | 2.05 | -0.21 | 0.72 | 0.54 | 0.15 | 0.10 |
| | (1.68) | (-1.89) | (-0.02) | (1.15) | (0.98) | (1.29) | (0.44) | (-0.09) | (0.58) | (1.10) | (0.35) | (0.27) |
| $R^2$ | 0.057 | 0.069 | 0.057 | 0.085 | 0.115 | 0.123 | 0.048 | 0.044 | 0.023 | 0.118 | 0.099 | 0.103 |
| GRS | 2.57 | | | | | | 0.84 | | | | | |
| ($p$ 值) | (0.02) | | | | | | (0.54) | | | | | |

**表 III 8　叙事因子对 CH4 因子的定价表现**

| | Panel A　因变量：叙事因子的 MVE 组合 | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | LDA | | | | | | LLM | | | | | |
| $f_{1\sim i}^{MVE}$ | 1 | 2 | 3 | 4 | 5 | 6 | 1 | 2 | 3 | 4 | 5 | 6 |
| $\alpha(\%)$ | -0.68 | -0.04 | -0.01 | 0.03 | 0.02 | 0.04 | -0.68 | -0.00 | 0.03 | 0.03 | 0.07 | 0.03 |
| | (-0.95) | (-0.11) | (-0.07) | (0.94) | (1.08) | (1.94) | (-0.95) | (-0.01) | (1.01) | (1.15) | (3.08) | (2.01) |
| $\beta_{MKT}$ | 0.74 | 0.38 | 0.10 | -0.00 | -0.00 | -0.00 | 0.74 | 0.38 | 0.04 | 0.05 | -0.01 | -0.0 |
| | (1.9) | (1.94) | (1.92) | (-0.34) | (-0.38) | (-0.3) | (1.9) | (1.94) | (1.93) | (2.09) | (-0.57) | (-0.46) |
| $\beta_{VMG}$ | -0.68 | -0.36 | -0.09 | 0.03 | 0.02 | 0.01 | -0.68 | -0.32 | 0.03 | 0.02 | -0.02 | -0.01 |
| | (-1.64) | (-1.74) | (-1.71) | (1.52) | (1.61) | (0.77) | (-1.64) | (-1.61) | (1.75) | (0.97) | (-0.79) | (-1.0) |
| $\beta_{SMB}$ | 2.56 | 1.51 | 0.39 | 0.05 | -0.00 | 0.02 | 2.56 | 1.34 | 0.12 | 0.14 | 0.05 | 0.02 |
| | (1.35) | (1.6) | (1.6) | (0.96) | (-0.01) | (0.52) | (1.35) | (1.41) | (1.13) | (1.41) | (0.67) | (0.32) |
| $\beta_{PMO}$ | -0.55 | -0.50 | -0.14 | -0.11 | -0.04 | -0.04 | -0.55 | -0.34 | -0.09 | -0.10 | -0.05 | -0.01 |
| | (-0.32) | (-0.57) | (-0.62) | (-1.8) | (-0.64) | (-0.98) | (-0.32) | (-0.38) | (-0.75) | (-0.9) | (-0.58) | (-0.24) |
| $R^2$ | 0.030 | 0.031 | 0.030 | 0.043 | 0.017 | 0.010 | 0.030 | 0.030 | 0.031 | 0.031 | 0.036 | 0.015 |
| GRS | 52.02 | | | | | | 888.08 | | | | | |
| ($p$ 值) | (0.00) | | | | | | (0.00) | | | | | |

（续表）

Panel B　因变量：CH4

| | LDA | | | | LLM | | | |
|---|---|---|---|---|---|---|---|---|
| | MKT | VMG | SMB | PMO | MKT | VMG | SMB | PMO |
| $\alpha$(%) | 1.34 | -0.33 | -0.48 | -0.27 | 3.36 | -1.18 | -1.47 | -1.31 |
| | (1.59) | (-0.67) | (-1.10) | (-0.76) | (0.80) | (-0.44) | (-0.62) | (-0.66) |
| $\beta_{f1}$ | 3.20 | -1.42 | -1.61 | -1.12 | 6.30 | -2.74 | -3.14 | -2.78 |
| | (1.91) | (-1.39) | (-2.03) | (-1.68) | (0.85) | (-0.58) | (-0.76) | (-0.8) |
| $\beta_{f2}$ | -6.36 | 3.21 | 3.30 | 2.49 | -13.07 | 5.64 | 6.54 | 5.79 |
| | (-2.26) | (1.76) | (2.25) | (2.08) | (-0.84) | (0.58) | (0.76) | (0.80) |
| $\beta_{f3}$ | -0.09 | -1.45 | -0.27 | -0.93 | -5.04 | 2.73 | 2.81 | 2.49 |
| | (-0.01) | (-0.18) | (-0.04) | (-0.17) | (-0.74) | (0.67) | (0.74) | (0.79) |
| $\beta_{f4}$ | -4.13 | 3.14 | 2.17 | 1.74 | 15.90 | -6.68 | -7.93 | -6.97 |
| | (-1.62) | (1.85) | (1.46) | (1.42) | (0.97) | (-0.68) | (-0.87) | (-0.91) |
| $\beta_{f5}$ | -5.54 | 3.28 | 2.74 | 1.92 | 0.85 | -0.57 | -0.64 | -0.75 |
| | (-1.62) | (1.65) | (1.70) | (1.43) | (0.12) | (-0.13) | (-0.17) | (-0.24) |
| $\beta_{f6}$ | 12.60 | -7.41 | -6.46 | -4.74 | 2.39 | -2.05 | -1.35 | -1.12 |
| | (1.74) | (-1.59) | (-1.80) | (-1.57) | (0.52) | (-0.74) | (-0.56) | (-0.59) |
| $R^2$ | 0.058 | 0.053 | 0.053 | 0.043 | 0.050 | 0.035 | 0.042 | 0.041 |
| GRS | | | 3.07 | | | | 0.92 | |
| ($p$ 值) | | | (0.02) | | | | (0.45) | |

**表 III 9　叙事因子对 HXZ4 因子的定价表现**

| | Panel A　因变量：叙事因子的 MVE 组合 | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | LDA | | | | | | LLM | | | | | |
| $f_{1\sim i}^{MVE}$ | 1 | 2 | 3 | 4 | 5 | 6 | 1 | 2 | 3 | 4 | 5 | 6 |
| $\alpha(\%)$ | -0.57 | 0.02 | 0.01 | 0.04 | 0.03 | 0.04 | -0.57 | 0.05 | 0.03 | 0.03 | 0.08 | 0.03 |
| | (-0.75) | (0.05) | (0.09) | (1.16) | (1.58) | (2.30) | (-0.75) | (0.14) | (1.14) | (1.31) | (2.82) | (1.83) |
| $\beta_{MKT}$ | 0.47 | 0.22 | 0.05 | -0.01 | 0.01 | -0.00 | 0.47 | 0.25 | 0.03 | 0.04 | -0.01 | -0.00 |
| | (0.55) | (0.51) | (0.44) | (-0.46) | (0.37) | (-0.23) | (0.55) | (0.59) | (1.13) | (1.54) | (-0.37) | (-0.23) |
| $\beta_{ME}$ | 0.91 | 0.44 | 0.09 | -0.00 | 0.02 | -0.00 | 0.91 | 0.49 | 0.07 | 0.07 | -0.00 | -0.01 |
| | (0.53) | (0.51) | (0.43) | (-0.11) | (0.41) | (-0.14) | (0.53) | (0.59) | (1.20) | (1.54) | (-0.02) | (-0.38) |
| $\beta_{INV}$ | -0.45 | -0.17 | -0.03 | -0.01 | -0.02 | 0.01 | -0.45 | -0.28 | -0.11 | -0.10 | 0.02 | 0.02 |
| | (-0.24) | (-0.18) | (-0.11) | (-0.18) | (-0.44) | (0.26) | (-0.24) | (-0.30) | (-1.67) | (-1.63) | (0.29) | (0.95) |
| $\beta_{ROE}$ | 0.22 | 0.05 | 0.00 | -0.01 | -0.02 | -0.04 | 0.22 | 0.15 | 0.12 | 0.09 | -0.02 | -0.03 |
| | (0.20) | (0.10) | (0.03) | (-0.16) | (-0.73) | (-1.36) | (0.20) | (0.28) | (2.18) | (1.63) | (-0.41) | (-1.25) |
| $R^2$ | 0.017 | 0.018 | 0.017 | 0.034 | 0.045 | 0.039 | 0.017 | 0.017 | 0.026 | 0.017 | 0.033 | 0.020 |
| GRS | 50.68 | | | | | | 896.44 | | | | | |
| ($p$ 值) | (0.00) | | | | | | (0.00) | | | | | |

（续表）

Panel B　因变量：HXZ4

| | LDA | | | | LLM | | | |
|---|---|---|---|---|---|---|---|---|
| | MKT | ME | INV | ROE | MKT | ME | INV | ROE |
| $\alpha$(%) | 1.53 | -0.69 | 0.08 | 0.22 | 3.67 | -1.92 | 0.13 | 0.59 |
| | (1.80) | (-1.12) | (0.15) | (0.76) | (0.90) | (-0.93) | (0.13) | (0.86) |
| $\beta_{f1}$ | 3.09 | -1.94 | -0.66 | -0.23 | 6.42 | -3.82 | -0.45 | 0.50 |
| | (1.85) | (-2.10) | (-0.75) | (-0.45) | (0.88) | (-1.07) | (-0.27) | (0.44) |
| $\beta_{f2}$ | -5.98 | 3.50 | 0.10 | -0.45 | -13.28 | 7.97 | 1.00 | -1.01 |
| | (-2.22) | (2.43) | (0.07) | (-0.59) | (-0.87) | (1.07) | (0.29) | (-0.42) |
| $\beta_{f3}$ | -0.61 | 1.59 | 4.91 | 3.58 | -5.02 | 2.82 | 0.97 | 0.85 |
| | (-0.05) | (0.29) | (1.21) | (1.27) | (-0.77) | (0.83) | (0.85) | (1.03) |
| $\beta_{f4}$ | -4.27 | 1.33 | -0.63 | 0.07 | 16.01 | -9.45 | -1.97 | -0.05 |
| | (-1.85) | (1.06) | (-0.85) | (0.15) | (1.01) | (-1.15) | (-0.68) | (-0.02) |
| $\beta_{f5}$ | -5.27 | 2.12 | -0.67 | -0.47 | 0.93 | -0.82 | -0.12 | 0.30 |
| | (-1.48) | (1.24) | (-0.57) | (-0.68) | (0.14) | (-0.27) | (-0.09) | (0.35) |
| $\beta_{f6}$ | 12.51 | -6.27 | -1.54 | -1.74 | 2.53 | -0.70 | 0.54 | 0.07 |
| | (1.70) | (-1.83) | (-0.55) | (-0.99) | (0.59) | (-0.36) | (0.48) | (0.08) |
| $R^2$ | 0.060 | 0.057 | 0.055 | 0.059 | 0.054 | 0.049 | 0.017 | 0.020 |
| GRS | | | 2.00 | | | | 0.99 | |
| ($p$ 值) | | | (0.10) | | | | (0.42) | |

**表 III 10　叙事因子对 DHS3 因子的定价表现**

| | Panel A　因变量：叙事因子的 MVE 组合 | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | LDA | | | | | | LLM | | | | | |
| $f_{1\sim i}^{MVE}$ | 1 | 2 | 3 | 4 | 5 | 6 | 1 | 2 | 3 | 4 | 5 | 6 |
| $\alpha(\%)$ | -0.40 | 0.10 | 0.03 | 0.03 | 0.02 | 0.03 | -0.40 | 0.14 | 0.04 | 0.04 | 0.08 | 0.03 |
| | (-0.61) | (0.31) | (0.34) | (0.95) | (1.16) | (1.96) | (-0.61) | (0.42) | (1.49) | (1.68) | (2.93) | (1.85) |
| $\beta_{MKT}$ | 0.64 | 0.30 | 0.07 | -0.04 | -0.01 | -0.02 | 0.64 | 0.30 | -0.03 | -0.02 | -0.01 | 0.02 |
| | (1.77) | (1.61) | (1.48) | (-1.99) | (-0.67) | (-1.78) | (1.77) | (1.70) | (-1.19) | (-0.79) | (-0.44) | (1.98) |
| $\beta_{FIN}$ | 0.40 | 0.19 | 0.04 | -0.04 | -0.01 | -0.02 | 0.40 | 0.17 | -0.05 | -0.04 | -0.00 | 0.02 |
| | (0.93) | (0.86) | (0.73) | (-1.53) | (-0.29) | (-1.26) | (0.93) | (0.80) | (-1.59) | (-1.38) | (-0.06) | (1.92) |
| $\beta_{PEAD}$ | 2.10 | 1.02 | 0.25 | -0.10 | -0.04 | -0.05 | 2.10 | 1.05 | -0.01 | 0.02 | 0.02 | 0.03 |
| | (1.91) | (1.78) | (1.73) | (-2.28) | (-1.15) | (-2.42) | (1.91) | (1.89) | (-0.11) | (0.51) | (1.79) | (1.56) |
| $R^2$ | 0.055 | 0.051 | 0.048 | 0.08 | 0.019 | 0.041 | 0.055 | 0.057 | 0.017 | 0.028 | 0.035 | 0.025 |
| GRS | 51.96 | | | | | | 921.23 | | | | | |
| ($p$ 值) | (0.00) | | | | | | (0.00) | | | | | |

（续表）

Panel B　因变量：DHS3

| | LDA | | | LLM | | |
|---|---|---|---|---|---|---|
| | MKT | FIN | PEAD | MKT | FIN | PEAD |
| $\alpha(\%)$ | 1.53 | -1.03 | -0.11 | 3.67 | -1.92 | -1.05 |
| | (1.80) | (-2.18) | (-0.40) | (0.90) | (-0.77) | (-1.41) |
| $\beta_{f1}$ | 3.09 | -1.93 | -0.31 | 6.42 | -3.28 | -1.79 |
| | (1.85) | (-1.98) | (-0.66) | (0.88) | (-0.74) | (-1.40) |
| $\beta_{f2}$ | -5.98 | 3.85 | 0.86 | -13.28 | 6.79 | 3.75 |
| | (-2.22) | (2.51) | (0.92) | (-0.87) | (0.73) | (1.40) |
| $\beta_{f3}$ | -0.61 | -0.13 | -0.77 | -5.02 | 3.67 | 0.48 |
| | (-0.05) | (-0.02) | (-0.28) | (-0.77) | (0.94) | (0.32) |
| $\beta_{f4}$ | -4.27 | 2.49 | 0.27 | 16.01 | -9.89 | -3.21 |
| | (-1.85) | (1.66) | (0.40) | (1.01) | (-1.02) | (-1.09) |
| $\beta_{f5}$ | -5.27 | 3.69 | 0.72 | 0.93 | -0.41 | -0.54 |
| | (-1.48) | (1.73) | (0.73) | (0.14) | (-0.10) | (-0.43) |
| $\beta_{f6}$ | 12.51 | -7.72 | -1.96 | 2.53 | -0.61 | -0.41 |
| | (1.70) | (-1.90) | (-0.84) | (0.59) | (-0.24) | (-0.34) |
| $R^2$ | 0.06 | 0.07 | 0.034 | 0.054 | 0.056 | 0.053 |
| GRS | | 1.42 | | | 0.65 | |
| ($p$ 值) | | (0.24) | | | (0.58) | |

# 附录 IV　统计因子模型

附录 IV 展示了统计因子模型的构造方式、基本性质与实证结果。

## IV.1 数据与方法

本文参考 Kozak et al.（2018）的做法，构造了基于股票收益率信息的统计主成分因子模型，并将其作为基准与叙事因子模型进行比较。具体而言，使用 5×5 的规模-账面市值比特征组合，数据来源为锐思数据库，样本期为 2007 年 1 月至 2021 年 12 月。与主流文献一致，本文采用流通市值加权计算组合收益率，而使用总市值加权的结果类似。此外，为了与文章中使用的因子模型频率保持一致，我们在基准结果中采用月度数据，并将对日度组合收益率进行 PCA 的结果作为稳健性。具体的因子构造过程如下[20]：

首先，对 25 个组合收益率$r_{p,t}$的协方差矩阵进行特征值分解，得到一组特征值和对应的特征向量。然后，按照特征值大小提取前五个主成分（PC）作为统计因子，其中每个因子$k = 1,\cdots,5$都可以表示为 25 个组合收益率的线性组合：

$$PC_{k,t} = \sum_{p=1}^{25} \pi_{k,p} r_{p,t}，\ p = 1,\cdots,25 \tag{IV.1}$$

其中，$\pi_{k,p}$为组合$p$在因子$k$上的特征向量权重（或因子载荷系数）。

在样本外交易策略方面，为确保可比性并避免潜在的前瞻性偏误（look-ahead bias），统计因子也采用滚动窗口的方式构造。具体而言，以三年期为训练窗口对 25 个投资组合收益率进行 PCA，获得各个主成分对应的特征向量权重$\hat{\pi}_{k,p}, k = 1,\cdots,5$；然后将样本外组合收益率$r_p, p = 1,\cdots,25$和权重$\pi_{k,p}$进行加总计算，预测未来一年内主成分的样本外估计值：

$$PC_{k,t+m}^{OOS} = \sum_{p=1}^{25} \hat{\pi}_{k,p} r_{p,t+m}，\ m = 1,\cdots,12 \tag{IV.2}$$

其中，初始训练样本为 2008 年 1 月至 2010 年 12 月，第一个样本外预测对应于 2011 年 1 月，最后得到 11 年共 132 个月度的样本外主成分时间序列。

## IV.2 基本性质

图 IV 1 展示了在全样本月度数据下，前三个主成分因子（$PC_1$至$PC_3$）的特征向量权重。与 Kozak et al.（2018）的发现接近，前三个主成分因子体现出了和 FF3 因子相似的特性：第一个主成分因子（$PC_1$）是一个水平因子（level factor），对所有投资组合都具有相同的权重（与市场因子 MKT 相似）。另外两个则分别与规模因子（SMB）和价值因子（HML）有关，例如第二个主成分因子（$PC_2$）对小盘股权重相对较低，但在大盘股中具有高权重（恰好与 SMB 相反）；第三个主成分因子（$PC_3$）对高账面市值比股票具有正权重，对低账面市值比股票则具有负权重（与 HML 相似）[21]。

[20] 在构建基于日度组合收益率的 PC 因子时，我们保持与月度数据一致。

[21] 然而，在中国股市中，第二、第三主成分并未完全对应单一因子维度。例如，在第二主成分的权重中，可以观察到部分价值因子的特征；而在第三主成分的权重中，则同样体现出一定的规模因子特征。这一现象在日度数据的结果中仍然存在，表明规模-账面市值比的前三个主成分中，规模效应与价值效应并未被完全区分。这一结果既不同于美国市场，也在一定程度上偏离了 PCA 主成分彼此正交的理论性质。本文对此现象不作进一步探讨，未来研究可对其深入分析。

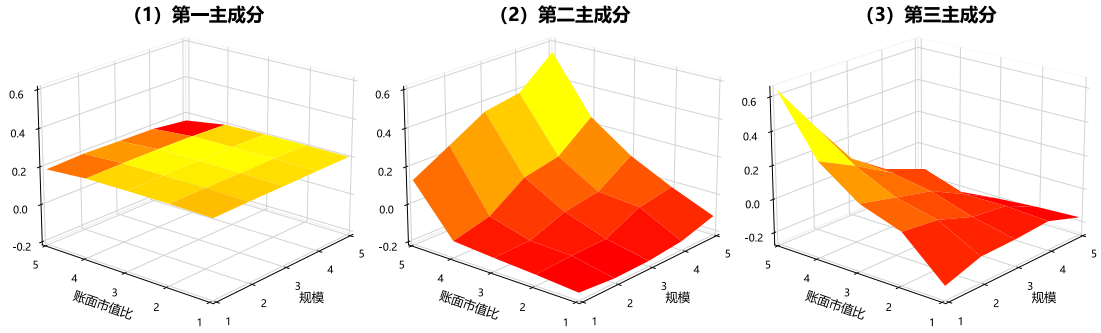


**图 IV 1　前三个主成分因子的特征向量权重（月度数据）**

图 IV 2 展示了在全样本日度数据下，前三个主成分因子的特征向量权重，同样呈现出与 FF3 因子相似的特征：$PC_1$作为水平因子（level factor），对所有投资组合赋予相近权重，类似于市场因子（MKT）；$PC_2$在规模维度上与小盘股呈现低权重，而在大盘股中权重较高，与规模因子（SMB）的表现相反；$PC_3$则对高账面市值比股票赋予正权重，对低账面市值比股票赋予负权重，与价值因子（HML）相似。使用日度收益率的结果与月度数据基本一致，表明主成分因子的特征在不同时间尺度上具有稳定性。

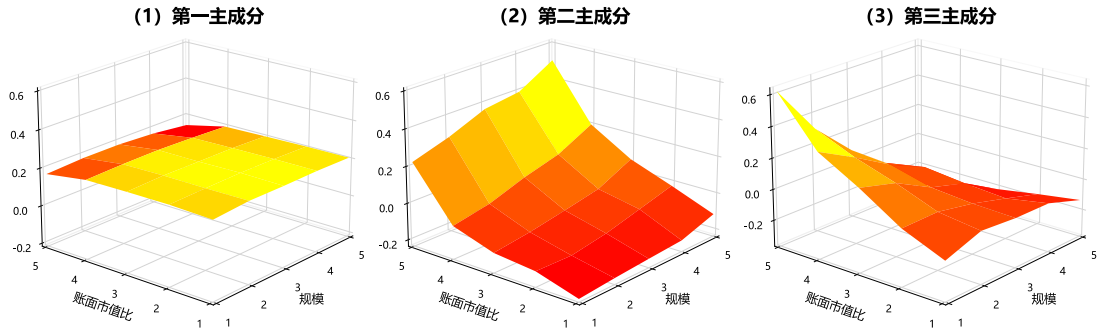


**图 IV 2　前三个主成分因子的特征向量权重（日度数据）**

IV.3 实证结果

表 IV 1 汇总了统计因子模型的定价效果，其中 Panel A 为正文汇报时使用的基准结果，Panel B 至 D 为使用不同加权方式或数据频率的稳健性分析。在资产异象检验方面，如 Panel A 所示，统计因子模型在因子数量较少时表现一般，不及基准的 Fama-French 因子模型和单因子 PC 模型。但是，随着主成分数量的增加，其对异象的解释能力也逐步提升。性能最强的 PC5 模型能够解释约 85%的资产异象，超额收益的平均显著性水平仅有 1.12，GRS 联合检验的统计量也仅为 2.30。因此，统计方法在捕捉资产异象方面具有一定的价值和应用潜力。在交易策略方面，PC 模型的样本外表现与因子数量呈现非单调的正 U 型关系，其中单因子 PC 模型已能实现较高的夏普比率。这可能是由于 PCA 方法在较短窗口的样本内拟合不足，未能有效识别真正能够预测样本外收益率的因子结构，导致当因子数量增加时，部分因子可能更多地捕捉噪声，从而削弱其样本外表现。Panel B 显示出使用总市值加权的结果类似。

若采用更为高频的日度数据（Panel C 和 D），PC 模型在主成分数量超过 2 时，对资产异象的解释能力显著提升。然而，在交易策略表现方面，主成分因子的夏普比率仍呈现非单调关系，且整体表现不及叙事因子模型。这表明，尽管日度数据下 PC 模型在捕捉资产异象

方面具有一定优势，但其样本外预测能力仍受限。

**表 IV 1　统计因子模型的定价效果**

| | Panel A：流通市值加权，月度数据（基准结果） | | | | |
|---|---|---|---|---|---|
| | 异象检验 | | | | 交易策略 |
| 因子模型 | 平均$\alpha$（%） | 平均$\|t(\alpha)\|$ | 显著比例（%） | GRS | 年化夏普比率 |
| PC1 | 0.24 | 2.02 | 38.54 | 5.41 | 0.43 |
| PC2 | 0.99 | 3.60 | 57.29 | 10.27 | 0.20 |
| PC3 | 1.08 | 2.04 | 39.58 | 4.01 | -0.01 |
| PC4 | 0.32 | 1.29 | 22.92 | 2.86 | 0.28 |
| PC5 | 0.14 | 1.12 | 15.62 | 2.30 | 0.40 |
| | Panel B：总市值加权，月度数据（稳健性 1） | | | | |
| PC1 | 0.24 | 2.03 | 38.54 | 5.42 | 0.42 |
| PC2 | 0.97 | 3.44 | 55.21 | 8.38 | 0.19 |
| PC3 | 1.25 | 2.41 | 46.88 | 4.84 | -0.02 |
| PC4 | 0.48 | 1.50 | 28.12 | 3.14 | 0.34 |
| PC5 | 0.31 | 1.21 | 17.71 | 2.61 | 0.43 |
| | Panel C：流通市值加权，日度数据（稳健性 2） | | | | |
| PC1 | 0.83 | 1.57 | 30.21 | 1.80 | 0.48 |
| PC2 | 0.28 | 1.12 | 17.71 | 1.55 | 0.22 |
| PC3 | 0.07 | 0.90 | 10.42 | 1.49 | -0.01 |
| PC4 | 0.09 | 0.95 | 10.42 | 1.46 | -0.00 |
| PC5 | 0.10 | 0.96 | 11.46 | 1.44 | 0.33 |
| | Panel D：总市值加权，日度数据（稳健性 3） | | | | |
| PC1 | 0.84 | 1.57 | 30.21 | 1.81 | 0.46 |
| PC2 | 0.26 | 1.09 | 17.71 | 1.56 | 0.22 |
| PC3 | 0.04 | 0.89 | 10.42 | 1.50 | 0.22 |
| PC4 | 0.10 | 0.94 | 11.46 | 1.40 | 0.31 |
| PC5 | 0.06 | 0.94 | 9.38 | 1.44 | 0.58 |

# 参考文献


[1] Blei, D. M., A. Y. Ng, and M. I. Jordan, “Latent Dirichlet Allocation”, *Journal of Machine Learning Research*, 2003, 3, 993-1022.

[2] Bybee, L., B. Kelly, and Y. Su, “Narrative Asset Pricing: Interpretable Systematic Risk Factors from News Text”, *The Review of Financial Studies*, 2023, 36(12), 4759-4787.

[3] Bybee, L., B. Kelly, A. Manela, and D. Xiu, “Business News and Business Cycles”, *The Journal of Finance*, 2024, 79(5), 3105-3147.

[4] Campbell, J. Y., “A Variance Decomposition for Stock Returns”, *The Economic Journal*, 1991, 101(405), 157-179.

[5] Campbell, J. Y., and T. Vuolteenaho, “Bad Beta, Good Beta”, *American Economic Review*, 2004, 94(5), 1249-1275.

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

[7] 陈贞竹、李力、余昌华，“我国货币政策传导效率及信号效应研究——基于金融市场高频识别的视角”，《经济学》(季刊)，2023 年第 1 期，第 56－73 页。

[8] Da, Z., J. Engelberg, and P. Gao, “In Search of Attention”, *The Journal of Finance,* 2011, 66(5), 1461-1499.

[9] Daniel, K., D. Hirshleifer, and L. Sun, “Short- and Long-Horizon Behavioral Factors”, *The Review of Financial Studies*, 2020, 33(4), 1673-1736.

[10] Hecht, P., and T. Vuolteenaho, “Explaining Returns with Cash-Flow Proxies”, *The Review of Financial Studies*, 2006, 19(1), 159-194.

[11] Hong, Y., F. Jiang, L. Meng, and B. Xue, “Forecasting Inflation Using Economic Narratives”, *Journal of Business & Economic Statistics*, 2025, 43(1), 216-231.

[12] Hou, K., C. Xue, and L. Zhang, “Digesting Anomalies: An Investment Approach”, *The Review of Financial Studies*, 2015, 28(3), 650-705.

[13] Huberman, G., and S. Kandel, “Mean-Variance Spanning”, *The Journal of Finance*, 1987, 42(4), 873-888.

[14] 姜富伟、孟令超、唐国豪，“媒体文本情绪与股票回报预测”，《经济学》(季刊)，2021 年第 4 期，第 1323－1344 页。

[15] Kelly, B. T., S. Pruitt, and Y. Su, “Characteristics Are Covariances: A Unified Model of Risk and Return”, *Journal of Financial Economics*, 2019, 134(3), 501-524.

[16] Kozak, S., S. Nagel, and S. Santosh, “Interpreting Factor Models”, *The Journal of Finance*, 2018, 73(3), 1183-1223.

[17] Larsen, V. H., and L. A. Thorsrud, “The Value of News for Economic Developments”, *Journal of Econometrics*, 2019, 210(1), 203-218.

[18] Larsen, V. H., L. A. Thorsrud, and J. Zhulanova, “News-Driven Inflation Expectations and Information Rigidities”, *Journal of Monetary Economics*, 2021, 117, 507-520.

[19] 李斌、邵新月、李玥阳，“机器学习驱动的基本面量化投资研究”，《中国工业经济》，2019 年第 8 期，第 61－79 页。

[20] 刘青、肖柏高，“劳动力成本与劳动节约型技术创新——来自 AI 语言模型和专利文本的证据”，《经济研究》，2023 年第 2 期，第 74－90 页。

[21] Liu, Y., and M. Lapata, “Text Summarization with Pretrained Encoders”, in Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Hong Kong, China: Association for Computational Linguistics, 2019.

[22] Merton, R. C., “An Intertemporal Capital Asset Pricing Model”, *Econometrica*, 1973, 41(5), 867-887.

[23] Nakamura, E., and J. Steinsson, “High-Frequency Identification of Monetary Non-Neutrality: The Information Effect”, *The Quarterly Journal of Economics*, 2018, 133(3), 1283-1330.

[24] Newey, W. K., and K. D. West, “A Simple, Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix”, *Econometrica*, 1987, 55(3), 703-708.

[25] Pagan, A. R., and K. A. Sossounov, “A Simple Framework for Analysing Bull and Bear Markets”, *Journal of Applied Econometrics*, 2003, 18(1), 23-46.

[26] Qin, B., D. Strömberg, and Y. Wu, “Media Bias in China”, *American Economic Review*, 2018, 108(9), 2442-2476.

[27] Reimers, N., and I. Gurevych, “Sentence-BERT: Sentence Embeddings Using Siamese BERT-Networks”, in Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Hong Kong, China: Association for Computational Linguistics, 2019.

[28] Shapiro, A. H., M. Sudhof, and D. J. Wilson, “Measuring News Sentiment”, *Journal of Econometrics*, 2022, 228(2), 221-243.

[29] Stambaugh, R. F., J. Yu, and Y. Yuan, “The Short of It: Investor Sentiment and Anomalies”, *Journal of Financial Economics*, 2012, 104(2), 288-302.

[30] 隋新玉、王云清，“中国股市周期与经济周期的比较”，《统计与决策》，2020 年第 17 期，第 129－133 页。

[31] Tan, L., H. Wu, and X. Zhang, *Large Language Models and Return Prediction in China*. Rochester, NY, 2023.

[32] Wu, M., K. Ohk, and K. Ko, “Does Cash-Flow News Play a Better Role Than Discount-Rate News? Evidence from Global Regional Stock Markets”, *Journal of International Money and Finance*, 2021, 110, 102267.
[33] 肖峻，“股市周期与基金投资者的选择”，《经济学》(季刊)，2013 年第 41 期，第 1299－1320 页。
[34] 许雪晨、田侃，“一种基于金融文本情感分析的股票指数预测新方法”，《数量经济技术经济研究》，2021 年第 12 期，第 124－145 页。
[35] Yang, Y., and H. Zou, “A Fast Unified Algorithm for Solving Group-Lasso Penalized Learning Problems”, *Statistics and Computing*, 2015, 25(6), 1129-1141.
[36] Yang, Z., D. Yang, C. Dyer, X. He, A. Smola, and E. Hovy, “Hierarchical Attention Networks for Document Classification”, in Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. San Diego, California: Association for Computational Linguistics, 2016.
[37] 易志高、茅宁，“中国股市投资者情绪测量研究：CICSI 的构建”，《金融研究》，2009 年第 11 期，第 174－184 页。
[38] You, J., B. Zhang, and L. Zhang, “Who Captures the Power of the Pen?”, *The Review of Financial Studies*, 2018, 31(1), 43-96.
[39] 张一帆、林建浩、樊嘉诚，“新闻文本大数据与消费增速实时预测——基于叙事经济学的视角”，《金融研究》，2023 年第 5 期，第 152－169 页。