扩散投资组合模型中的因子维度与偏差—方差权衡

AviBagchi* **MichaelTesfaye*** **Om Shastri**宾夕法尼亚
大学，费城，宾夕法尼亚州 19104，美国{
aviba,tesfaye,oshastri}@seas.upenn.edu

arXiv:2603.10385v1 [cs.CE] 11 Mar 2026

摘要

本文实现并评估了一种用于大规模股票数据上资产收益预测与投资组合构建的条件扩散模型。我们的方法对以公司特征（即因子）为条件的未来收益的完整分布进行建模，并利用所得条件矩构建投资组合。我们观察到明显的偏差—方差权衡：以过少因子为条件的模型会欠拟合，并产生过度分散的投资组合；而以过多因子为条件的模型会过拟合，导致不稳定且高度集中的配置，样本外表现较差。通过对因子维度进行消融实验，我们发现存在一个中间数量的因子，能够实现最佳泛化，并优于基准投资组合策略。

赛道： 行业与应用

1 引言

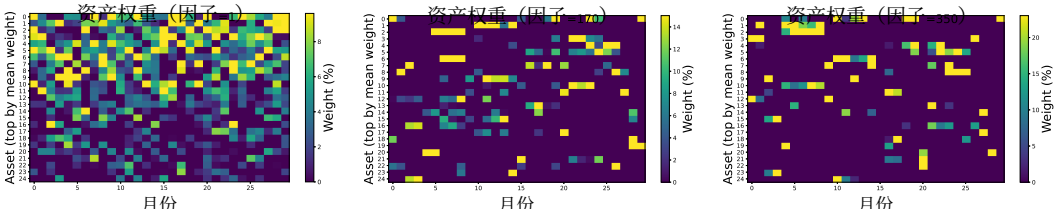
资产收益预测是量化金融中的一项基础问题。线性因子模型（Fama & French，1993年；2015年）为资产收益建模提供了一个易于处理的框架，但难以捕捉非线性和高阶市场动态。Chen 等人（2026年）提出了生成式方法，通过学习完整的条件收益分布而非点预测来建模。本文评估 Gao等人（2025年）提出的条件扩散框架，该框架基于可观测的公司特征（即因子）生成收益。我们表明，因子维度会在基于扩散的收益建模中引发明显的偏差—方差权衡：因子过少会导致欠拟合和过度多样化的投资组合，而因子过多则会产生高方差模型，导致配置过度集中。消融实验表明，存在一个能够优于基线策略的最优维度。我们使用来自沃顿研究数据服务（WRDS）的数据，并遵循 Jensen 等人（2023年）指定的流程。数据集详情和相关研究见附录（附录 A.1、附录 A.2）。

2 基于扩散的条件收益建模

我们遵循Gao等人（2025年）的做法，将资产收益预测表述为在给定可观测公司特征的条件下学习条件收益分布。 $R_{t+1} \in \mathbb{R}^N$ 表示时间段 $t = 1 \dots T$内观测到的 N项资产的收益向量，并令 $X_t = \{X_{i,j}\}_{i=1}^N$ 表示在时间 t观测到的相应资产层面特征（即因子）集合。假设收益满足 $R_{t+1} = f(X_t) + \epsilon_{t+1}$，其中 $f(\cdot)$是一个未知的、可能非线性的函数，而 ϵ_{t+1} 刻画与 t时已知信息无关的不可预测冲击。目标是学习完整的条件分布 $p(R_{t+1}|X_t)$，而不仅仅是条件均值。

为估计这一分布，我们采用条件去噪扩散概率模型（Ho 等人，2020年）。前向扩散过程通过在固定步数内逐步添加高斯噪声来扰乱观测收益，将数据转化为各向同性高斯分布。随后训练神经网络，通过预测每个扩散

*同等贡献。



(a) 低容量（高偏差） (b) 中等容量 (c) 高容量（高方差）

图 1：200个样本的热力图展示了在不同因子维度下学习得到的月度资产权重（按整个时间段内平均配置排名前 25 的资产）。低容量模型（左）广泛分配权重，体现出欠拟合和高偏差。中等容量模型（中）将配置集中于持续性信号，表明因子得到了有效利用。高容量模型（右）产生稀疏且不稳定的配置，这与过拟合和高方差相符。

步中添加的噪声，在给定特征 X_t的条件下逆转这一过程。反向扩散过程通过扩散Transformer架构实现。我们遵循Gao等人（2025年）对Peebles与Xie（2023年）所作的改进，将每项资产表示为一个词元，并通过自注意力层捕捉资产之间的横截面相依性。通过自适应归一化层，在词元层面对公司特征进行局部条件化。这种方法使每项资产的去噪动力学取决于其自身特征，同时仍能对资产间的联合收益行为进行建模。训练完成后，模型针对每个时期从条件分布 $p(R_{t+1}|x_t)$中生成蒙特卡洛样本，用于估计收益的条件均值和协方差，并将其作为投资组合构建流程（即均值—方差优化）的输入。

3 结果

ˆ 每个月 t，我们估计下一月份收益的条件均值向量 ˆμ_t 和协方差矩阵 Σ_t。随后，我们通过求解约束均值-方差优化问题，计算仅做多投资组合权重：max $\omega^\top$ ˆ$\mu_t - \frac{\gamma}{2}\omega^\top$ ˆ$\Sigma_t\omega$，约束条件为 $1^\top\omega = 1$和 $\omega \geq 0$，其中 $\bar{\mu}$为期望收益向量， Σ 为收益协方差矩阵， $\gamma = 100$ 为风险厌恶参数， ω 为投资组合权重（马科维茨，1952年）。我们遵循 Gao等人（2025年），将扩散因子投资组合与三个较简单的基准投资组合进行比较（附录 A.3）。

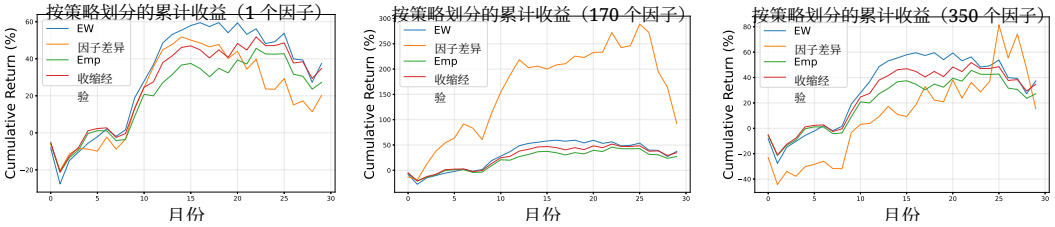
当 k较小时，投资组合权重在各资产之间相对分散，这反映出低容量模型能够生成广泛分散的配置（图 1）。随着 k 增加，权重分布变得更加集中，较大的头寸被配置于较少的资产上（图 1）。我们发现，适度分散投资组合 (b) 的累计收益优于等权、经验和收缩经验投资组合，而低容量 (a) ($k = 1$) 和高容量 (c) ($k = 350$) 投资组合则未能做到这一点（图 2）。完整的消融实验结果见附录 B.1 和附录 B.2。未来的研究应依据 Chen 等人（2026年；2023年）的框架评估这些结果；该框架在分数估计过程中通过得分分解隐式学习低维因子结构，从而无需显式进行因子选择（附录 B.3）。

参考文献

Nicola Borri、Denis Chetverikov、Yukun Liu 和 Aleh Tsyvinski。《采用高阶资产定价因子的前向选择法玛-麦克白回归》，2025年。网址：https://arxiv.org/abs/ 2503.23501。

Mark M Carhart。《共同基金业绩的持续性研究》。《style id='1'>《金融学杂志》，52(1):57–82，1997年。

Minshuo Chen、Kaixuan Huang、Tuo Zhao 和 Mengdi Wang。低维数据上扩散模型的得分近似、估计与分布恢复，2023年。网址：https: //arxiv.org/abs/2302.07194。



(a) 低容量（高偏差）　(b) 中等容量　(c) 高容量（高方差）

图2：四种策略在测试月份中，基于200个样本计算的投资组合累计收益。通过模型容量展示偏差—方差权衡。低容量模型（左）欠拟合数据，表现出高偏差。中等容量模型（中）实现了良好的偏差–方差平衡和最佳泛化效果（我们在图17中通过更大的样本量验证了这一点）。表达能力过强的模型（右）过拟合，表现出高方差和较低的样本外稳定性。

Minshuo Chen、Renyuan Xu、Yumin Xu 和 Ruixun Zhang。扩散因子模型：生成具有因子结构的高维收益，2026年。网址：https://arxiv.org/abs/ 2504.06566。

Eugene F Fama 和 Kenneth R French。股票和债券收益中的共同风险因子。《金融经济学杂志》，33(1):3–56，1993年。

Eugene F Fama 和 Kenneth R French。《五因子资产定价模型》。《金融经济学杂志》，116(1):1–22，2015年。

Jianqing Fan、Yingying Fan 和 Jinchi Lv。使用因子模型进行高维协方差矩阵估计。《计量经济学杂志》，147(1):186–197，2008年。

Jianqing Fan、Yuan Liao 和 Martina Mincheva。通过主正交补阈值化进行大规模协方差估计，2013年。网址 https://arxiv.org/abs/1201.0175。

Guanhao Feng、Wei Lan、Hansheng Wang 和 Jun Zhang。资产定价模型的选择与检验：逐步法，2026年。网址 https://arxiv.org/abs/2601.10279。

Xuefeng Gao、Mengying He 和 Xuedong He。用于投资组合优化的基于因子的条件扩散模型，2025年。网址 https://arxiv.org/abs/2509.22088。

Richard C. Grinold 和 Ronald N. Kahn。主动投资组合管理：实现卓越收益并控制风险的定量方法。麦格劳-希尔，第2版，2000年。

Shihao Gu、Bryan Kelly 和 Dacheng Xiu。通过机器学习进行实证资产定价。《金融研究评论》，33(5):2223–2273，2020年。

Jonathan Ho、Ajay Jain 和 Pieter Abbeel。去噪扩散概率模型，2020年。网址 https://arxiv.org/abs/2006.11239。

William James 和 Charles Stein。二次损失下的估计。载于 Samuel Kotz 和 Norman L. Johnson（编），统计学的突破：基础与基本理论，第443–460页。施普林格，纽约，1992年。原始论文发表于1961年，本文为其重印版。

Theis IngerslevJensen、Bryan Kelly 和 Lasse Heje Pedersen。金融学是否存在重复研究危机？《金融学杂志》，78(5):2465–2518，2023年。

Chen Jin 和 Ankush Agarwal。使用生成式扩散模型预测隐含波动率曲面，2025年。网址：https://arxiv.org/abs/2511.07571。

Bryan T. Kelly、Seth Pruitt 与 Yinan Su。《特征即协方差：风险与收益的统一模型》。《金融经济学杂志》，134(3):501–524，2019年。doi: 10.1016/j.jfineco.2019. 05.010。URL https://www.nber.org/papers/w24540。美国国家经济研究局工作论文版本可供获取；正式发表版本见《金融经济学杂志》。

Gihun Kim、Sun-Yong Choi 与 Yeoneung Kim。一种面向金融时间序列、经由几何布朗运动的基于扩散的生成模型，2025年。URL https://arxiv.org/abs/2507. 19003。

Olivier Ledoit 和 Michael Wolf。《亲爱的，我缩小了样本协方差矩阵》。《投资组合管理杂志》，30(4):110–119，2004年。Harry Markowitz。投资组合选择。《金融学杂志》，7(1):77–91，1952年。

Caspar Meijer 和 Lydia Y. Chen。扩散模型在时间序列预测中的兴起，2024年。网址：https://arxiv.org/abs/2401.03006。

William Peebles 和 Saining Xie。使用变换器构建可扩展的扩散模型，2023年。网址：https: // arxiv.org/abs/2212.09748。

Kashif Rasul、Calvin Seward、Ingmar Schuster 和 Roland Vollgraf。用于多变量概率时间序列预测的自回归去噪扩散模型，2021年。网址：https:// arxiv.org/abs/2101.12072。

Barr Rosenberg 与 Vijay Marathe。证券收益中协方差的市场外成分。《金融与数量分析杂志》，1974年。Lifeng Shen 和 James Kwok。用于时间序列预测的非自回归条件扩散模型，2023年。网址 https://arxiv.org/abs/2306.05043。Yang Song、Jascha Sohl-Dickstein、Diederik P. Kingma、Abhishek Kumar、Stefano Ermon 和 Ben Poole。通过随机微分方程进行基于分数的生成建模，2021年。网址 https://arxiv.org/abs/2011.13456。Chen Su、Zhengzhou Cai、Yuanhe Tian、Zhuochao Chang、Zihong Zheng 和 Yan Song。用于时间序列预测的扩散模型：综述，2025年。网址 https://arxiv.org/abs/2507. 14507。Tomonori Takahashi 和 Takayuki Mizuno。利用扩散模型生成合成金融时间序列，2024年。网址 https://arxiv.org/abs/2410.18897。

Yuki Tanaka、Ryuji Hashimoto、TakehiroTakayanagi、Zhe Piao、Yuri Murayama 与 Kiyoshi Izumi。Cofindiff：面向时间序列生成的可控金融扩散模型，2025年。URL https: //arxiv.org/abs/2503.04164。

Yusuke Tashiro、Jiaming Song、Yang Song 与 Stefano Ermon。Csdi：用于概率时间序列插补的条件式基于分数的扩散模型，2021年。URL https://arxiv.org/abs/2107.03502。Zhuohan Wang 与 Carmine Ventre。基于扩散模型的金融时间序列去噪器，2024年。URL https://arxiv.org/abs/2409.02138。

A 附录

A.1 数据

我们的分析使用由 Jensen、Kelly、Pederson 构建并通过沃顿研究数据服务（WRDS）分发的全球因子数据。该数据集整合了来自 CRSP 和 Compustat 的信息，以提供公开交易股票的公司层面特征与收益的综合面板。该数据集涵盖从2010年1月至2025年2月的数据。这些数据包含400多项特征，其构建遵循（Jensen 等人，2023年）中记录的程序。

为使 WRDS 因子数据与扩散框架保持一致，我们采用标准的横截面预处理，并构建固定形状的月度面板。我们将样本限定为美国普通股，并通过将实际收益率向前移一期来定义下月收益率这一预测目标。我们在横截面上对收益进行缩尾处理以减轻异常值的影响，同时使用横截面均值填补公司特征

，并在每个月份内进行标准化和截尾处理。对于每个月份，我们保留固定数量的资产，并将所得数据整理为特征张量和收益张量，其维度分别为 $(T,\ N,\ K)$ 和 $(T,\ N)$，其中 T 表示月份数，N 表示资产数，K 表示公司层面特征数。我们使用 $T = 150$、$N = 200$、$K = 350$。这些张量作为条件扩散模型的输入。

A.2 相关工作

扩散模型：扩散模型通过逐步向数据中加入噪声使其受到扰动，并训练神经网络逆转这一过程，从而学习复杂的数据分布。模型通过得分匹配学习得分函数，使其能够从噪声出发，经由迭代去噪还原数据，进而进行采样（Ho 等人，2020年；Song 等人，2021年）。对于时间序列，先前研究将扩散模型用于：（i）作为预测的条件情景生成器，或（ii）作为处理缺失数据问题的条件模型（Rasul 等人，2021年；Tashiro 等人，2021年）。综述总结了基于扩散模型的时间序列预测在评估方面的注意事项（Meijer & Chen，2024年；Su 等人，2025年）。

在金融学中，扩散模型主要用作收益的条件情景生成器，也有若干论文着重探讨可控性或金融特定的噪声结构。Shen 等人提出了一种非自回归条件扩散模型，根据历史数据生成未来时间序列轨迹，可用于金融收益预测（Shen & Kwok，2023年）。Tanaka 等人着重研究金融时间序列的可控条件生成，通过添加显式控制，引导采样轨迹呈现所需属性（Tanaka 等人，2025年）。Kim 等人修改扩散模型的前向加噪过程，以反映金融结构（例如异方差性和乘性噪声），旨在生成更真实的合成动态并改进条件采样（Kim 等人，2025年）。Wang 等人研究面向金融领域的去噪器，Takahashi 等人则提出了旨在生成合成金融时间序列的方法，并采用金融特定的建模选择（Wang & Ventre，2024年；Takahashi & Mizuno，2024年）。除了收益／路径生成之外，Jin 等人还在以期权为中心的场景中应用扩散模型，通过预测隐含波动率曲面实现这一目标（Jin & Agarwal，2025年）。

因子模型 因子模型是投资组合构建的标准框架，广泛采用的设定包括 Fama & French（1993年；2015年）和 Carhart（1997年）。现代投资组合风险系统以这一框架为基础，通过 Rosenberg & Marathe（1974年）的方法估计因子暴露和协方差结构，并结合 Grinold & Kahn（2000年）在量化投资组合构建方面的实践发展。因子模型的一个局限在于高维环境下存在估计误差（Ledoit & Wolf，2004年；Fan 等人，2008年；2013年）。实证资产定价研究表明，大规模的公司特征面板能够提高收益预测能力，但也会带来冗余和模型选择方面的挑战（Gu 等人，2020年；Kelly 等人，2019年）。随着所提出因子数量不断增加，近期研究强调了系统性检验、维度控制，以及高维因子空间中的过拟合风险（Feng 等人，2026年；Borri 等人，2025年）。

A.3 投资组合构建

在交易成本设定中，我们将线性交易成本纳入目标函数。投资组合收益通过投资组合权重与实际收益率的内积计算，并使用平均收益、波动率和年化夏普比率来概括性能。

- **等权重（EW）**采用统一权重，且不估计收益矩。
- **经验法（Emp）**使用滚动窗口，根据历史收益直接估计均值和协方差。
- **收缩经验法（ShrEmp）**采用协方差收缩以提高稳定性，同时保留样本协方差均值（James 与 Stein，1992年）。

我们使用上述条件扩散模型，对下一期收益的条件分布进行矩估计。从该分布中抽取的蒙特卡洛样本用于估计收益的条件均值和协方差。结果图中使用200个样本。

B 补充图

令 k 表示因子数量。我们对 $k \in \{1, 3, 6, 11, 18, 30, 48, 75, 115, 170, 240, 300, 350\}$ 进行消融实验。

B.1 累计收益

本节中的图展示了随着因子数量 k 增加，性能如何变化。当 k 较小时，模型受到过度约束，无法捕捉数据中的足够结构。这种高偏差状态会导致欠拟合：累计收益紧密跟随基线，或表现低于基线。随着 k 增加，性能得到改善，模型开始捕捉有意义的关系。当 $k \geq 18$ 时，模型开始超越基线。然而，当 k非常大时，性能再次恶化。模型进入高方差状态，此时新增因子主要是在拟合噪声，而不是信号。这体现为样本外表现下降，收益再次无法超越基线。

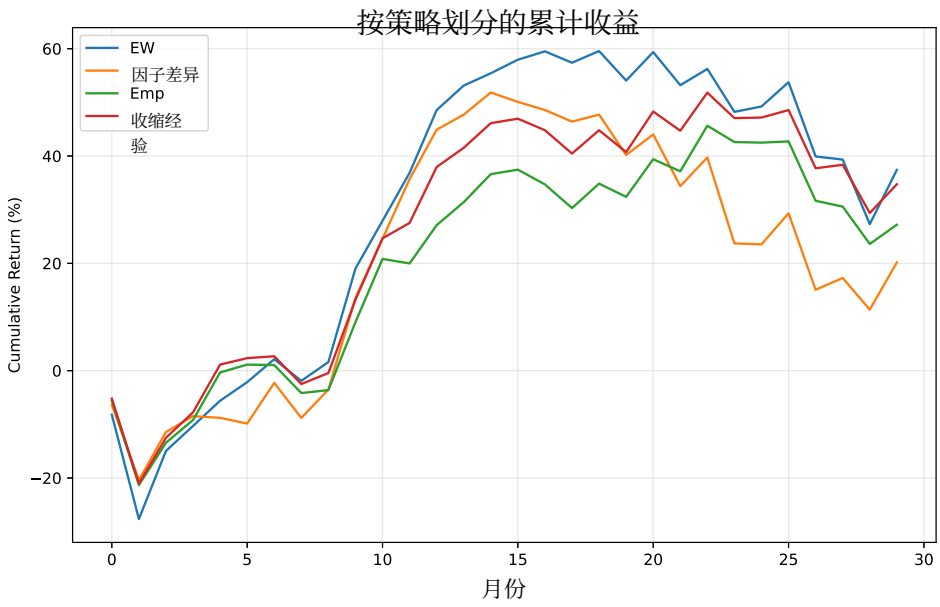


图3：累计收益（$k = 1$）

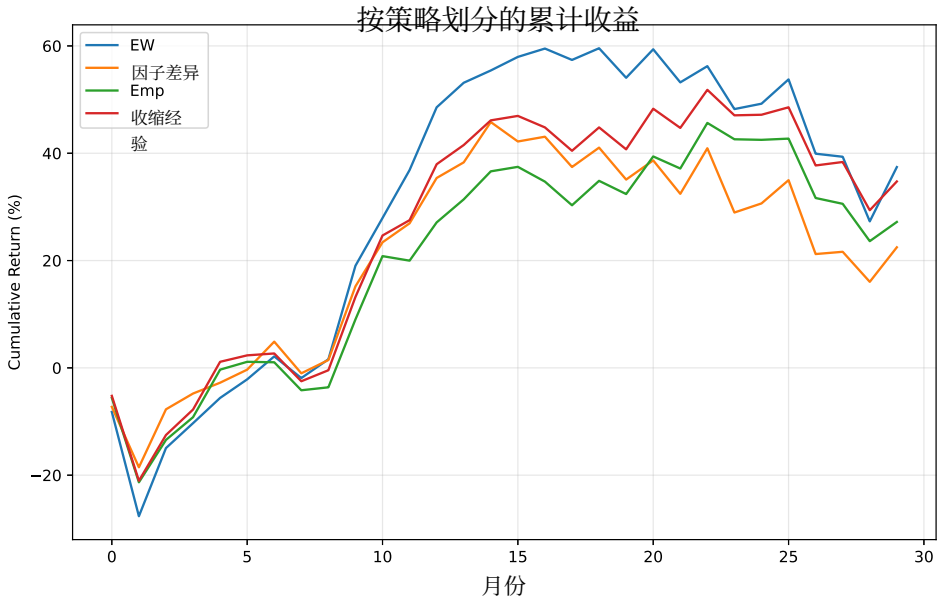


图 4：累计收益（$k = 3$）

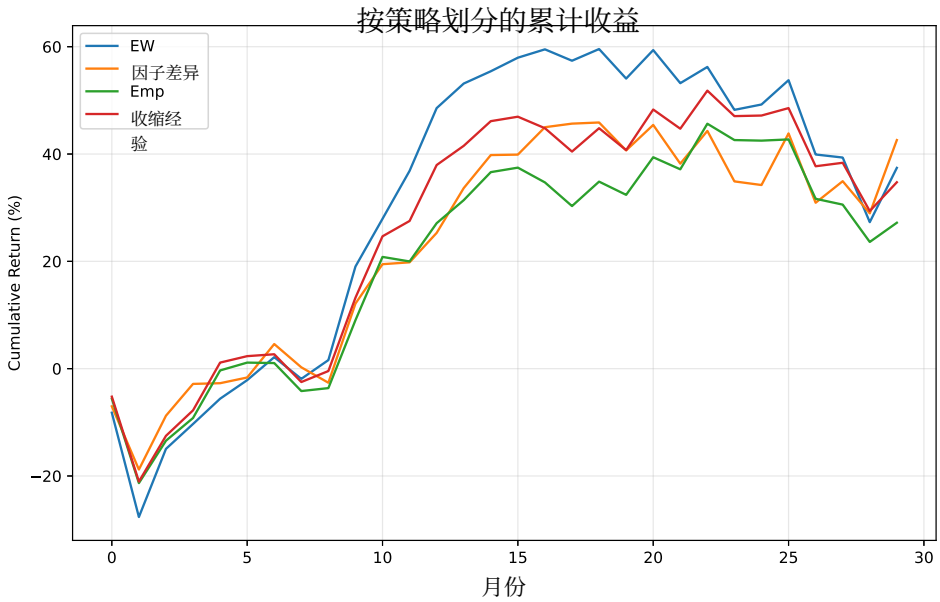


图 5：累计收益（$k = 6$）

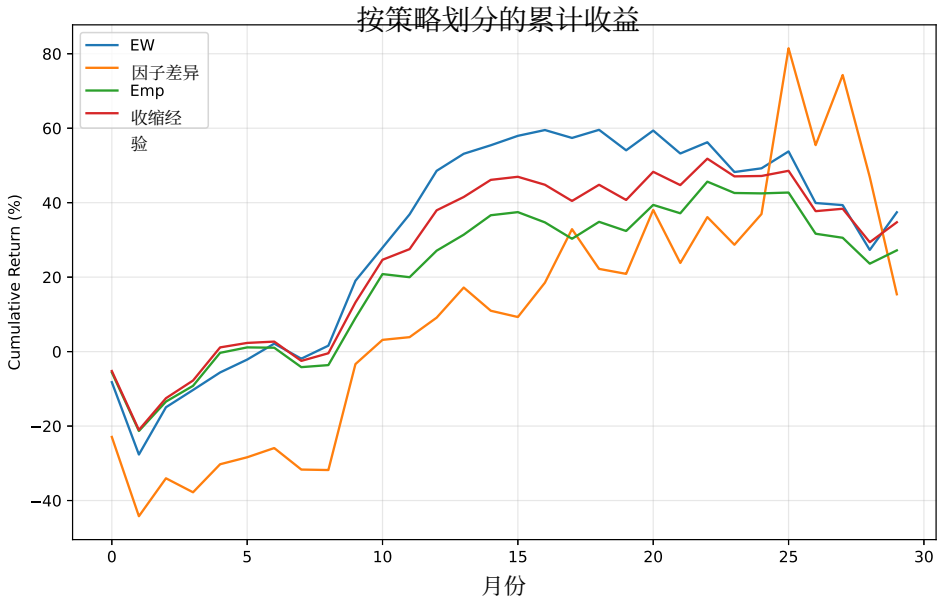


图6：累计收益（$k = 10$）

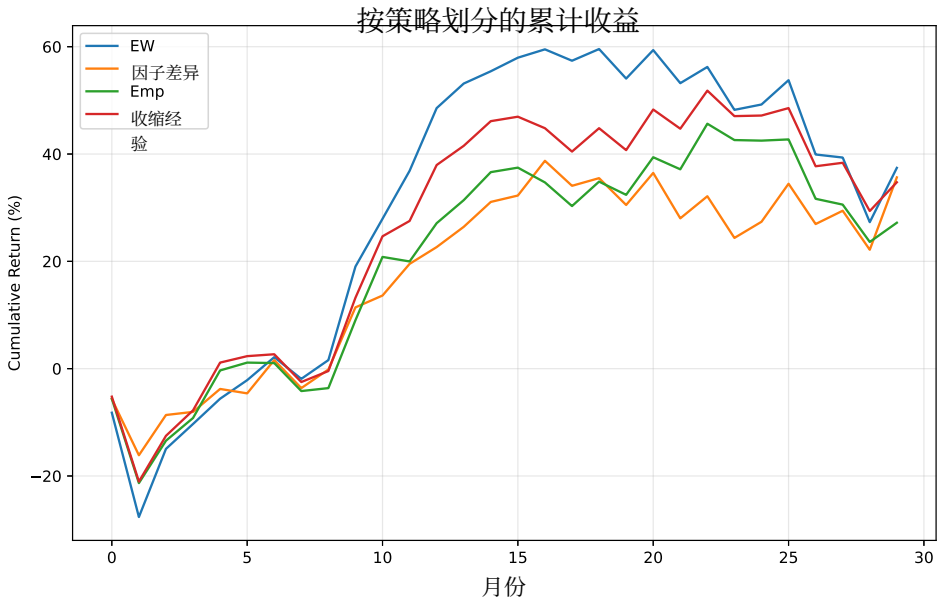


图7：累计收益（$k = 11$）

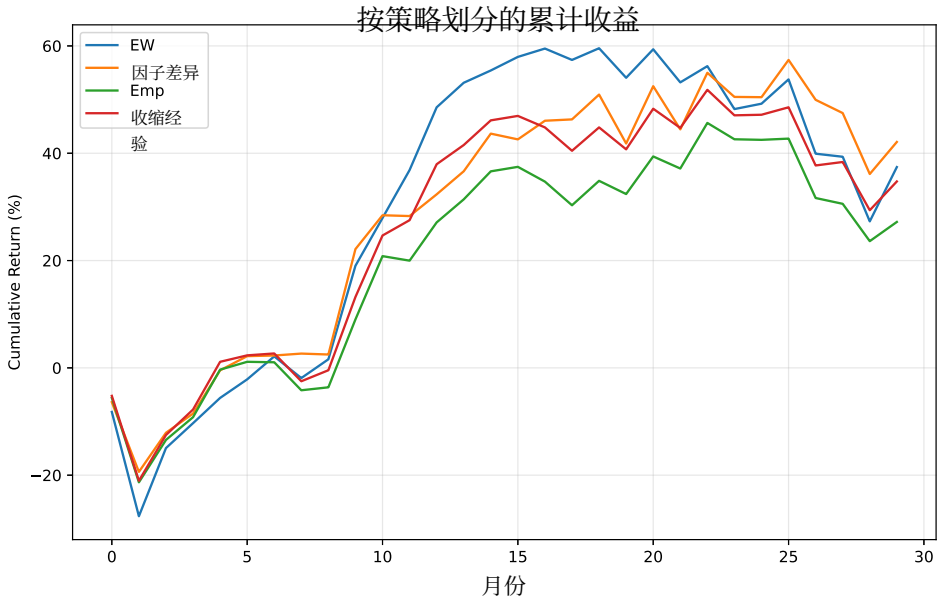


图 8：累计收益（$k = 18$）

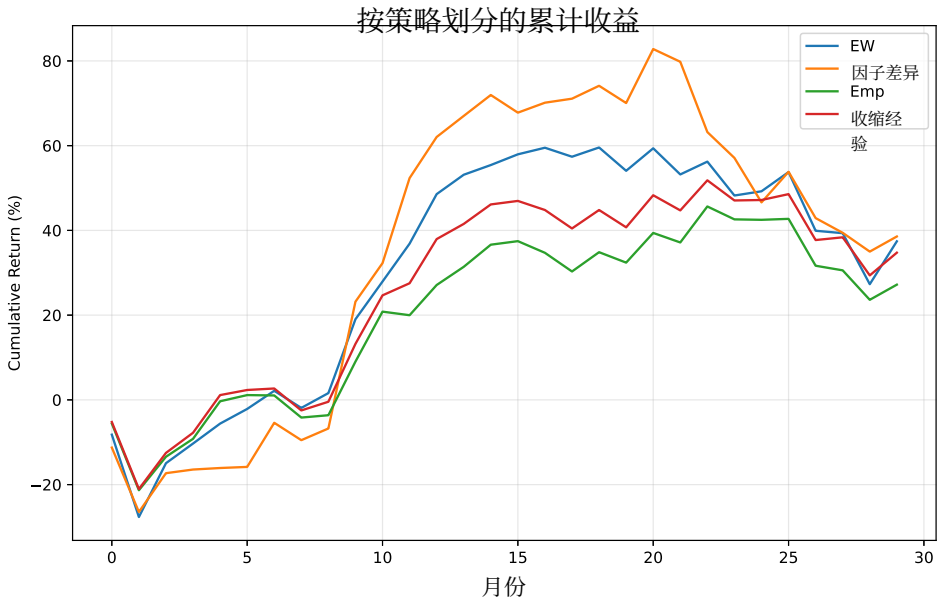


图 9：累计收益（$k = 30$）

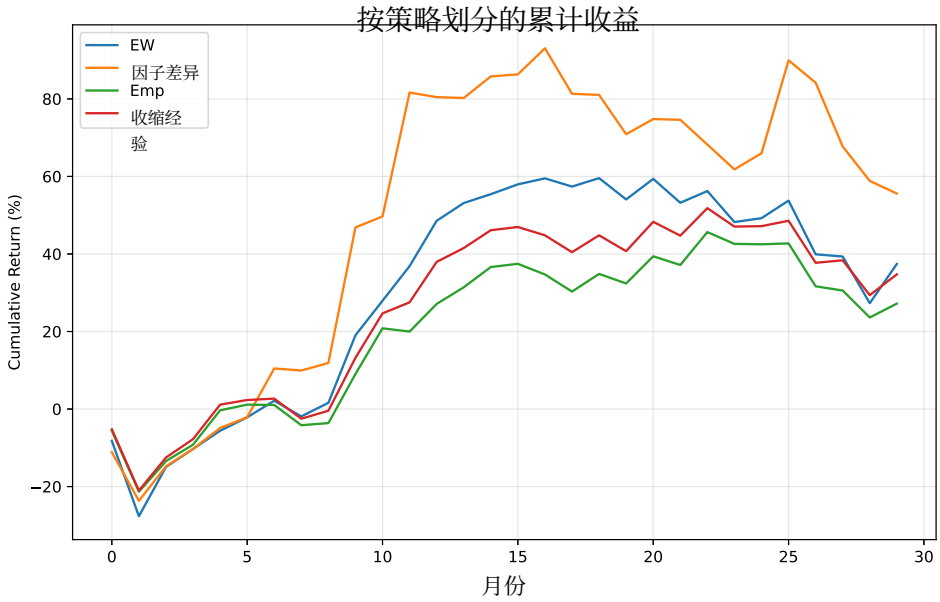


图 10：累计收益（$k = 48$）

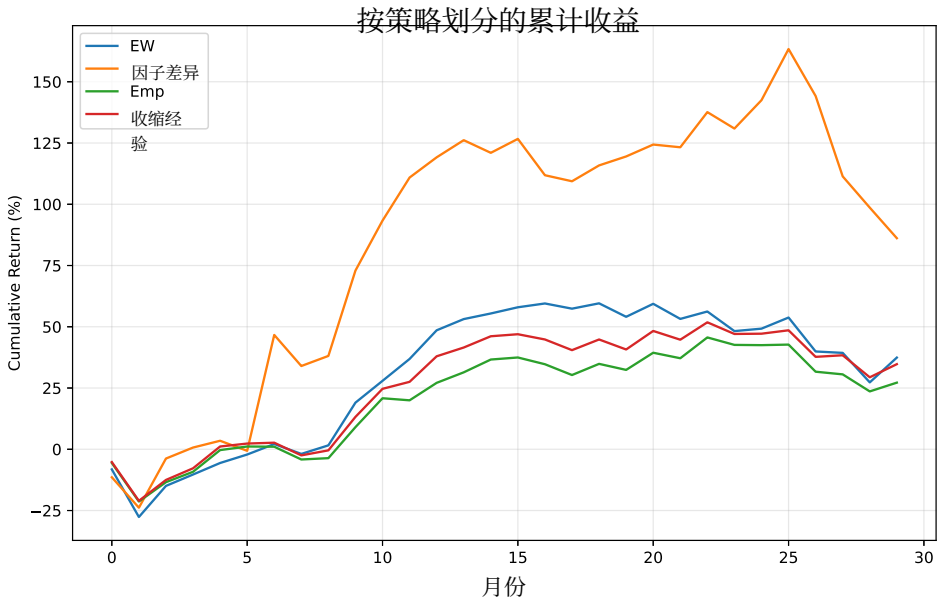


图 11：累计收益（$k = 75$）

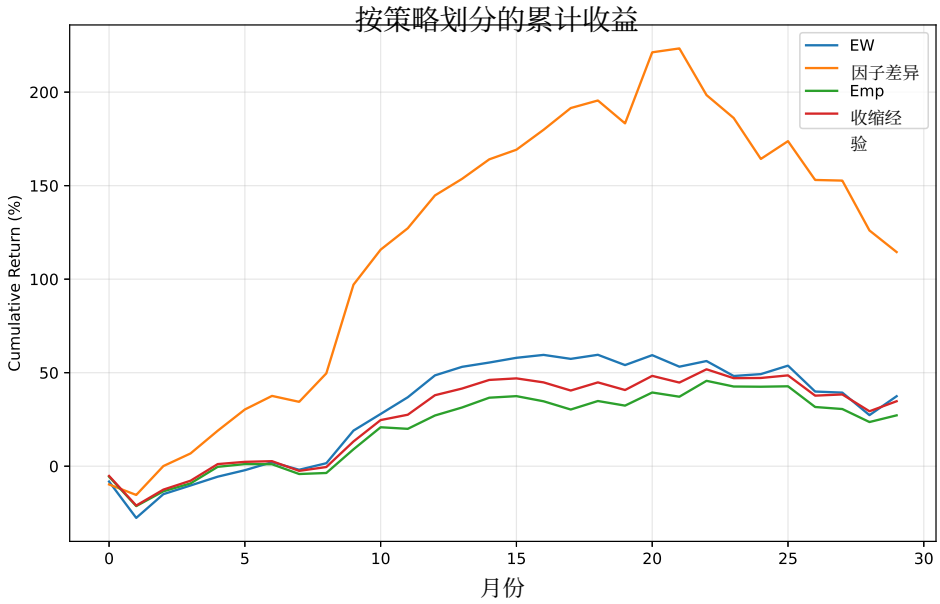


图 12：累计收益（$k = 115$）

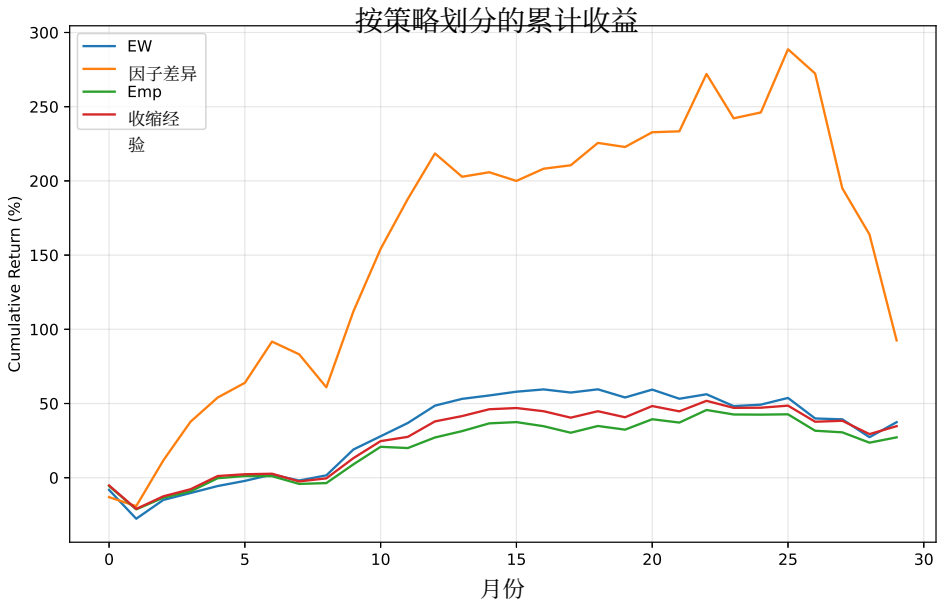


图 13：累计收益（$k = 170$）

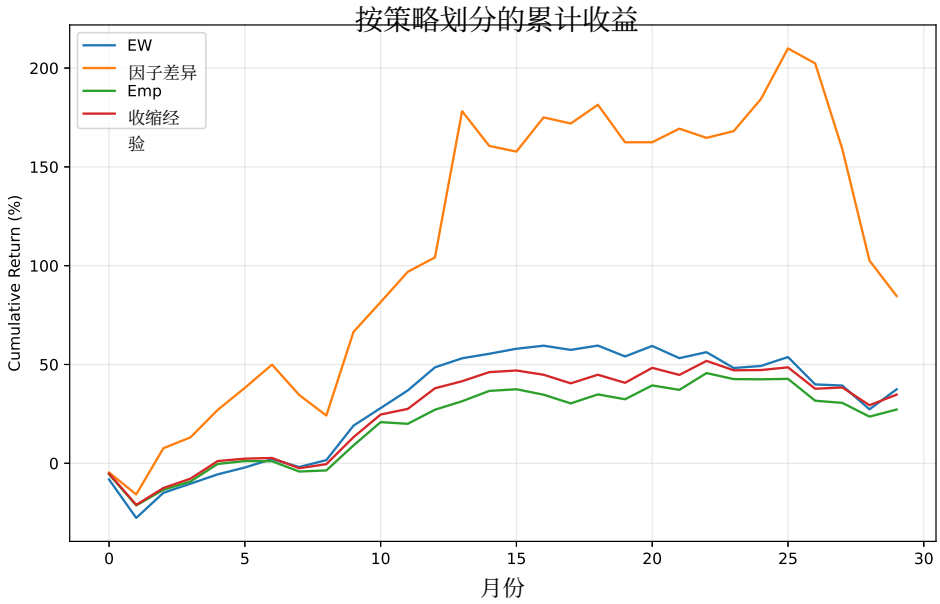


图14：累计收益（$k = 240$）

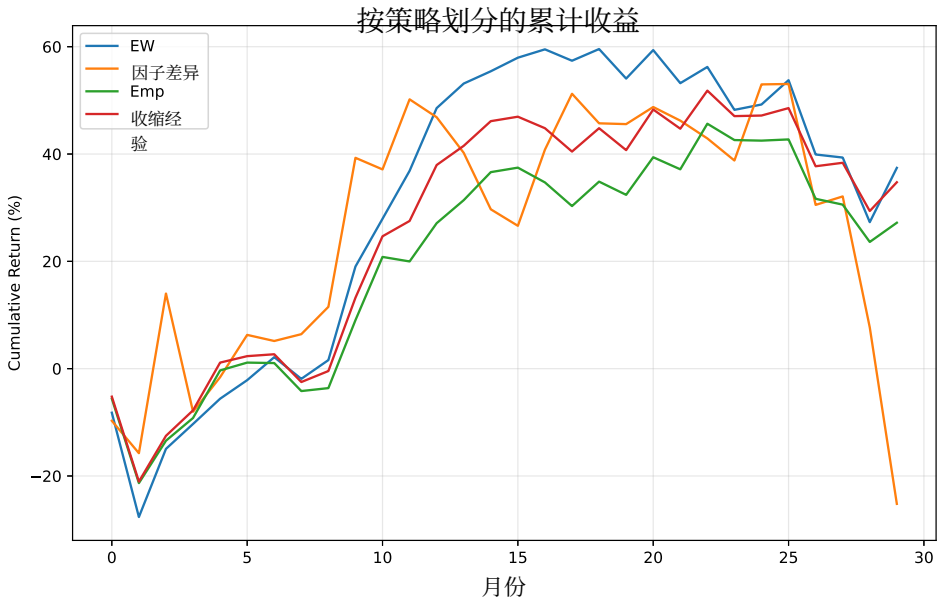


图15：累计收益（$k = 300$）

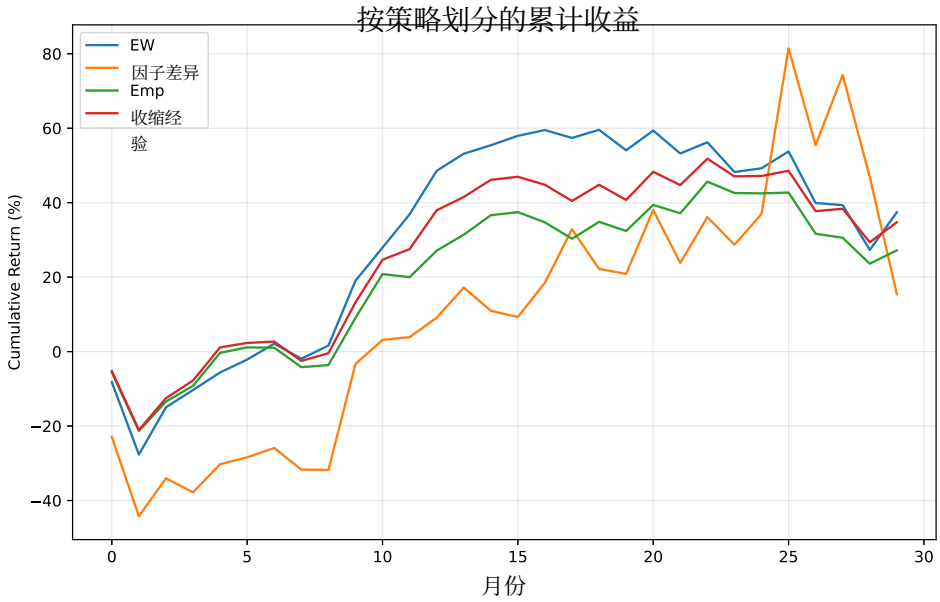


图16：累计收益（$k = 350$）

我们通过将样本数量增加到 1000，验证了 $k = 170$ 的性能，并再次观察到其优于基线。

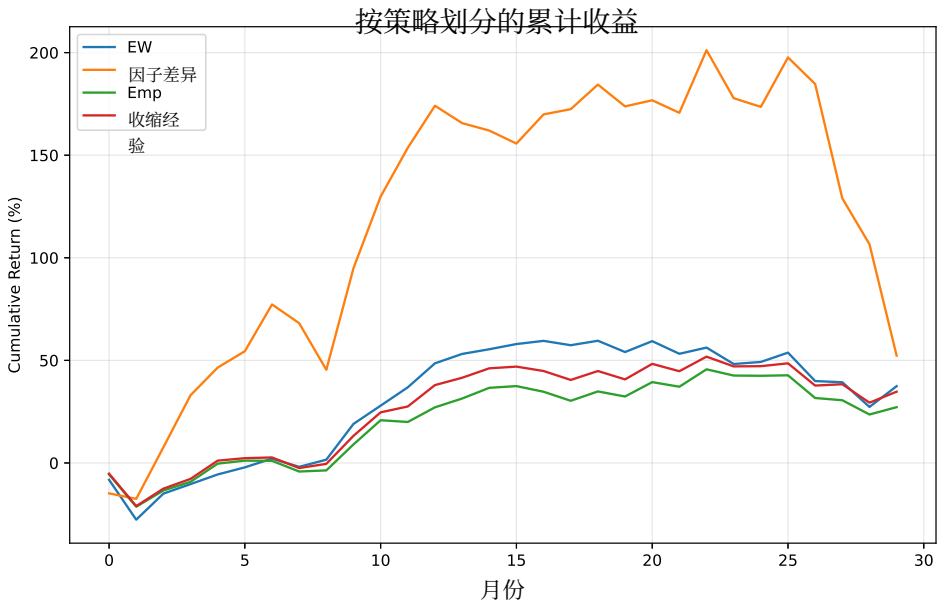


图17：累计收益（$k = 170$）

B.2 投资组合权重

本节展示了增加因子数量 k 如何逐步改变学习得到的投资组合的结构。当 k 较小时，权重广泛分布于多种资产之中，表明配置较为分散，与欠参数化模型相符。随着 k 增长，投资组合权重变得日益集中，较大的权重被分配给较少的一部分资产。这种集中趋势表明，容量更高的模型能够识别更具体的信号，但也会变得对噪声更加敏感，从而导致过度专门化配置，并降低样本外性能。

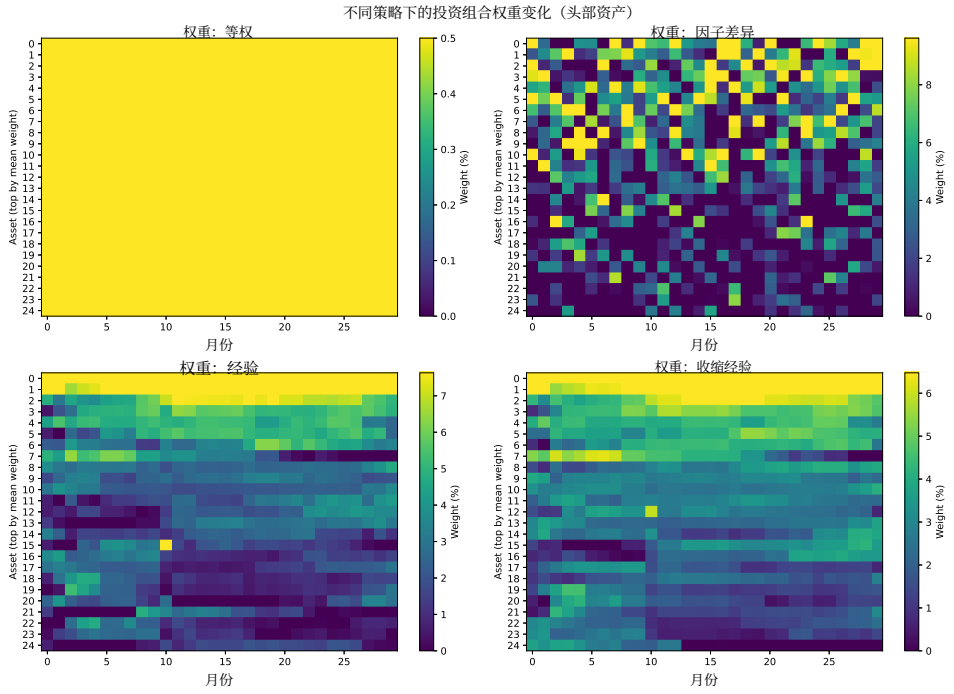


图 18：权重热力图（$k = 1$）

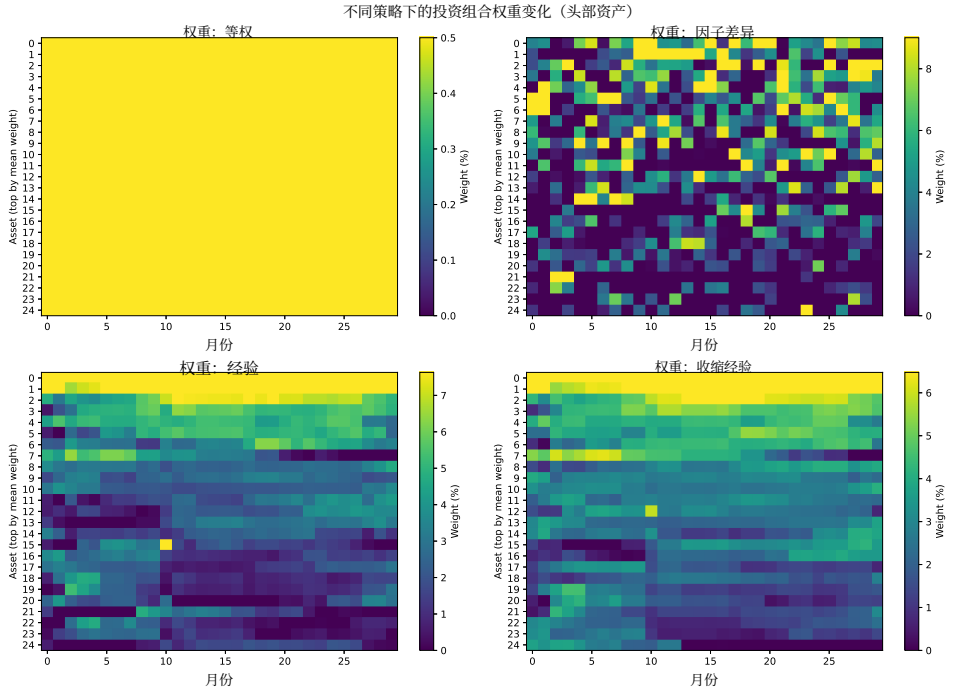


图 19：权重热力图（$k = 3$）

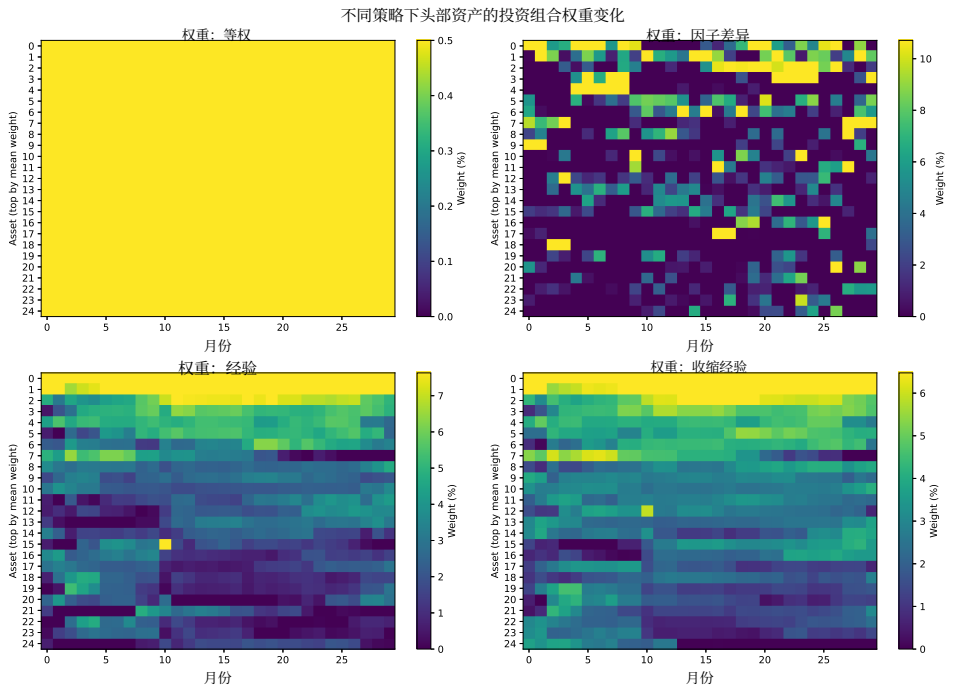


图 20：权重热力图（$k = 6$）

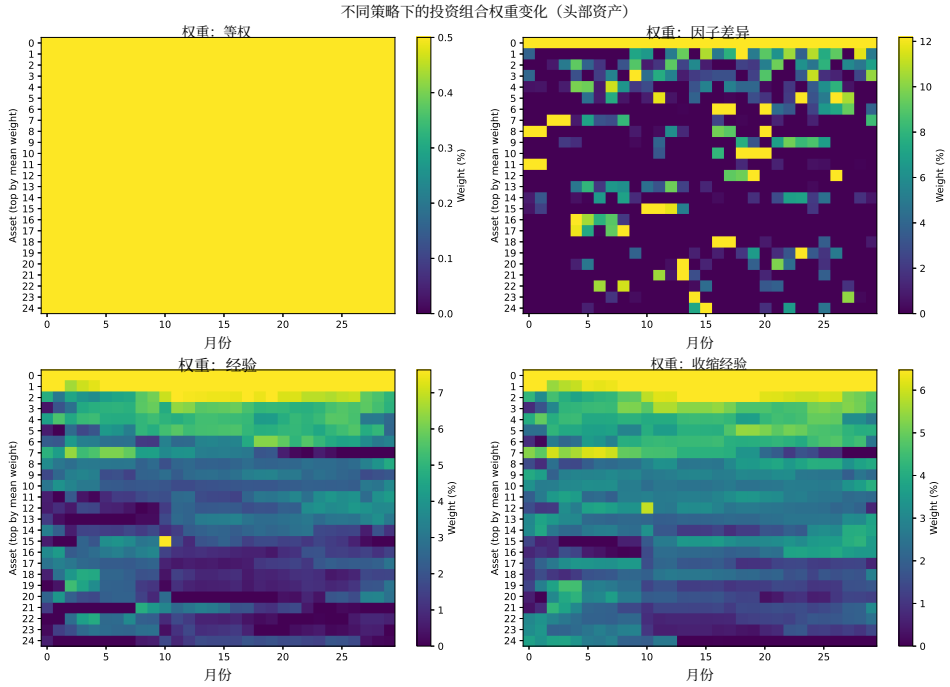


图 21：权重热力图（$k = 11$）

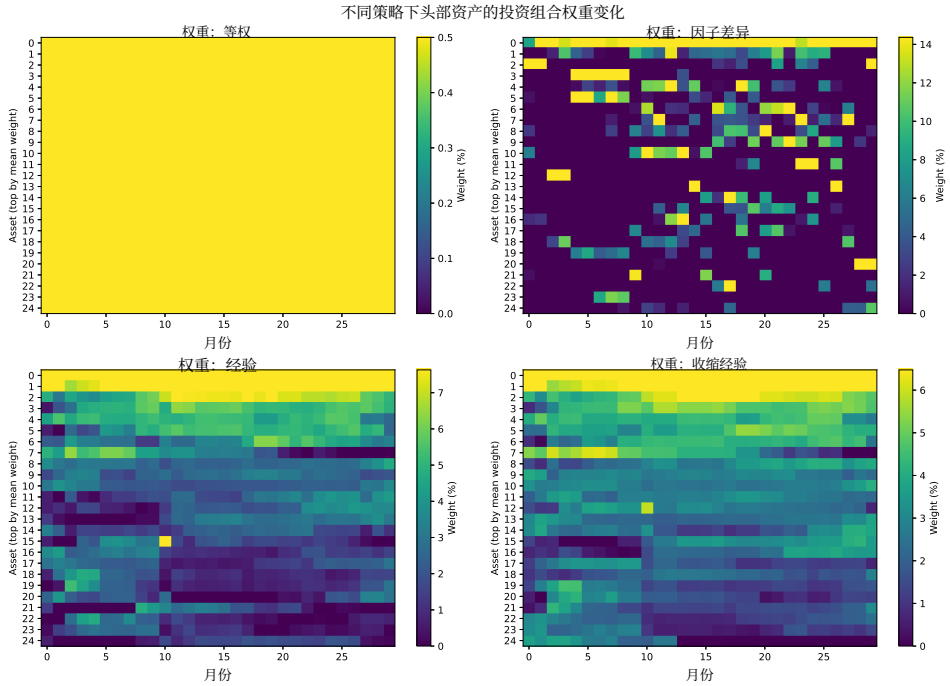


图 22：权重热力图（$k = 18$）

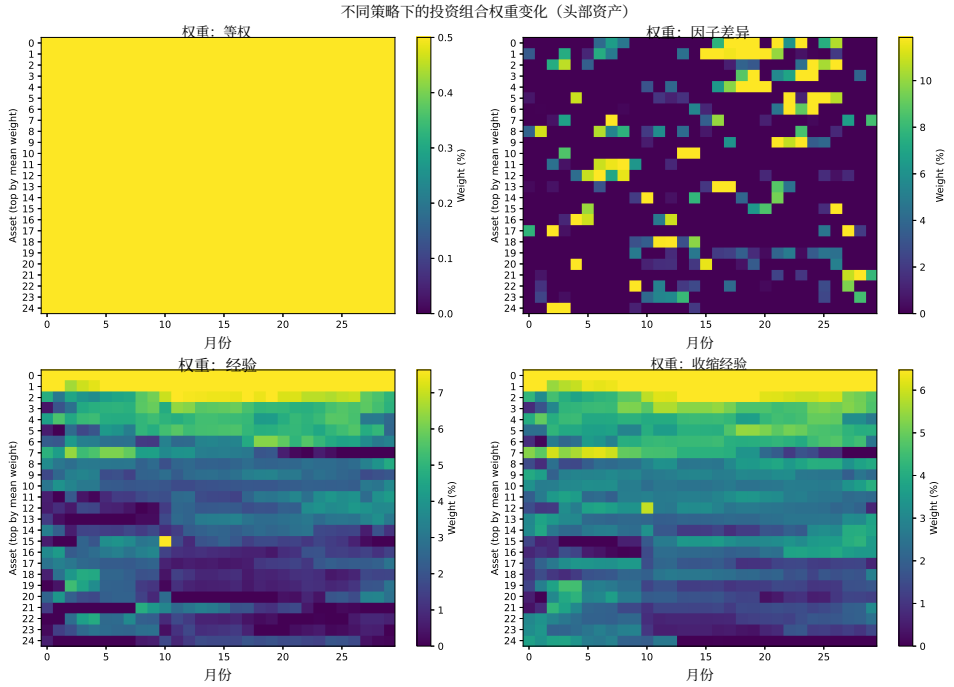


图 23：权重热力图（$k = 30$）

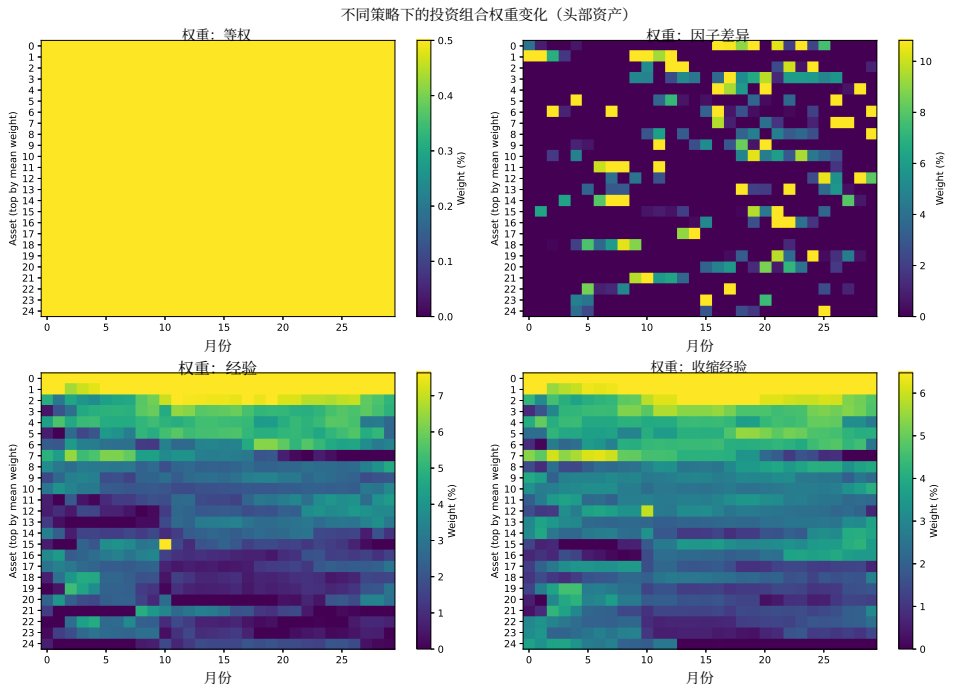


图 24：权重热力图（$k = 48$）

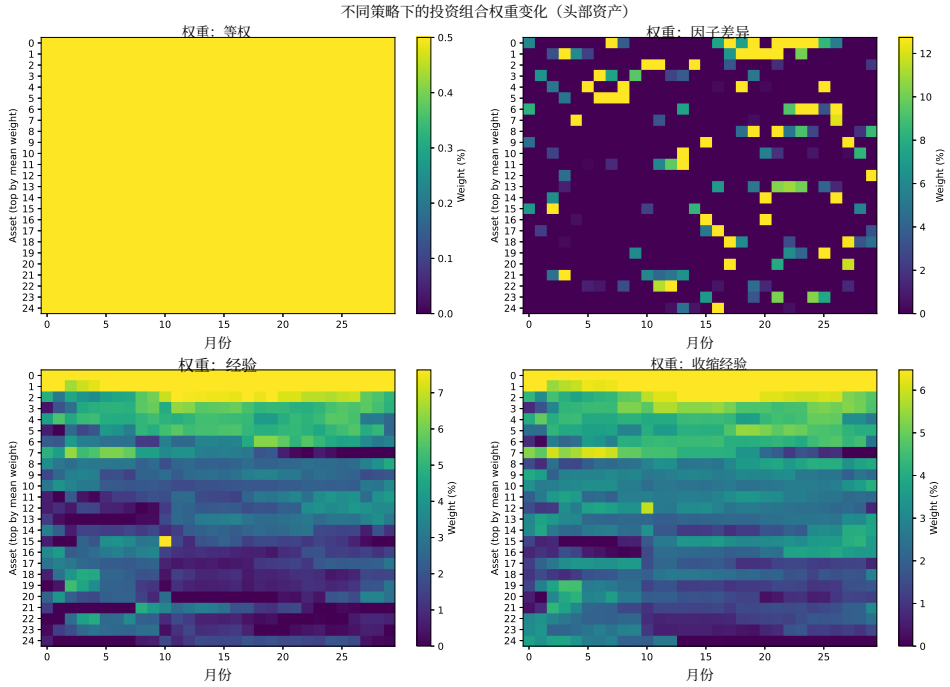


图 25：权重热力图（$k = 75$）

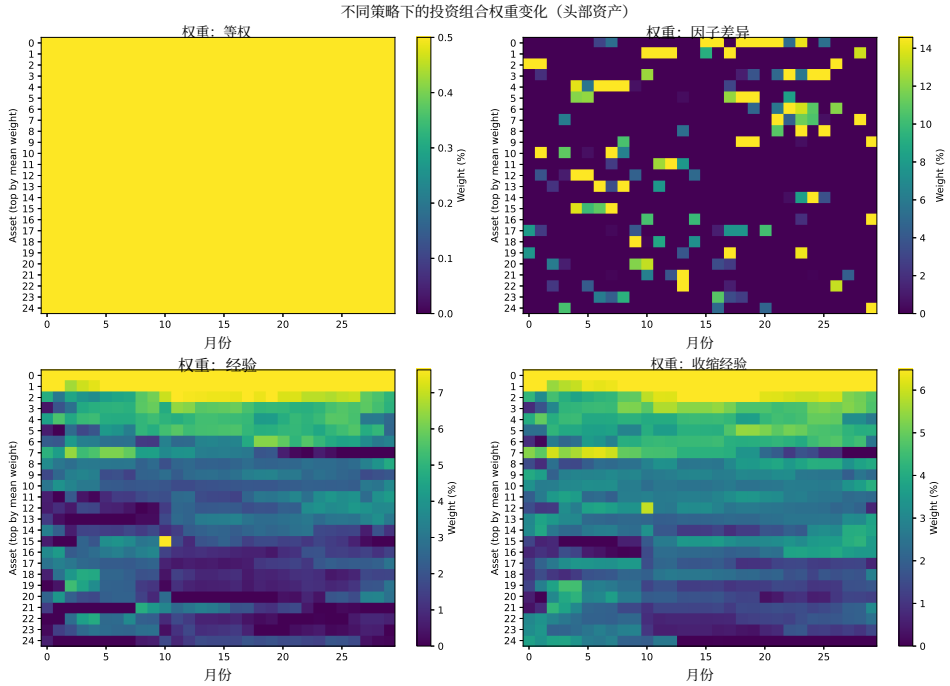


图 26：权重热力图（$k = 115$）

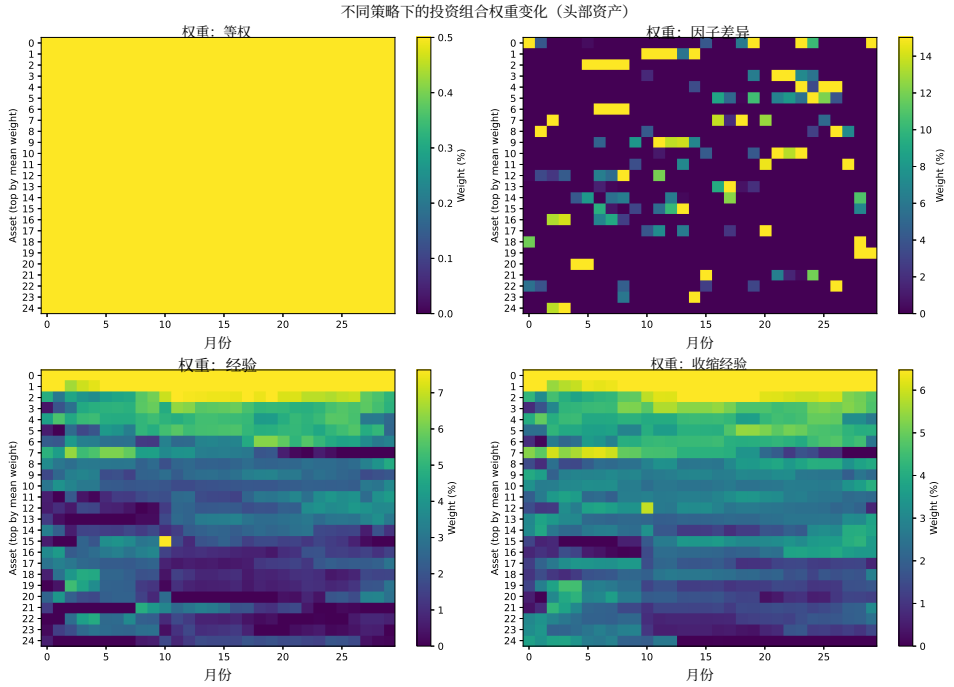


图 27：权重热力图（$k = 170$）

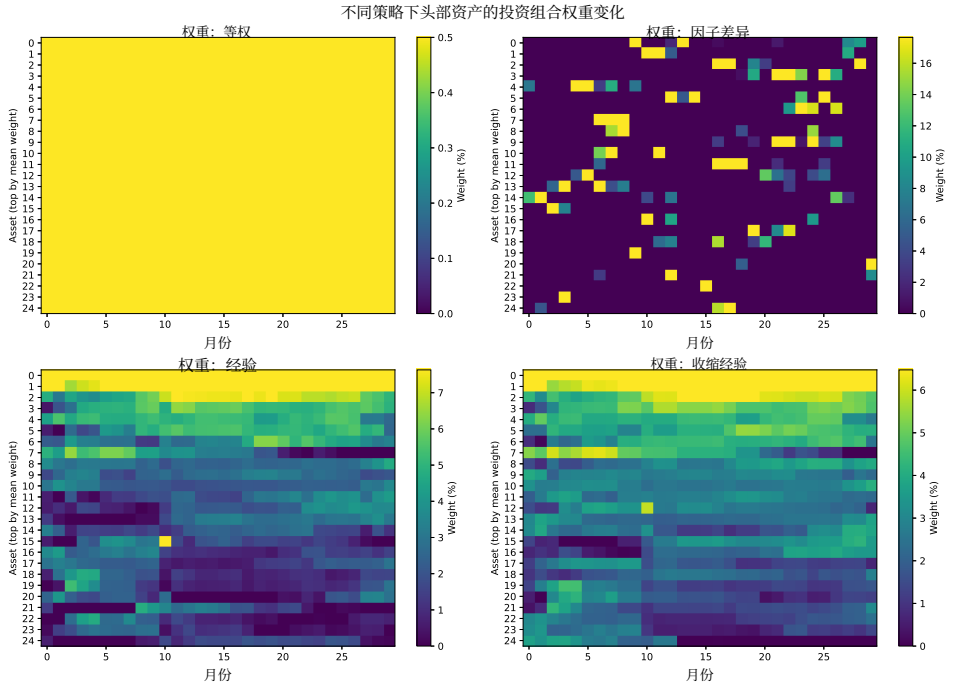


图 28：权重热力图（$k = 240$）

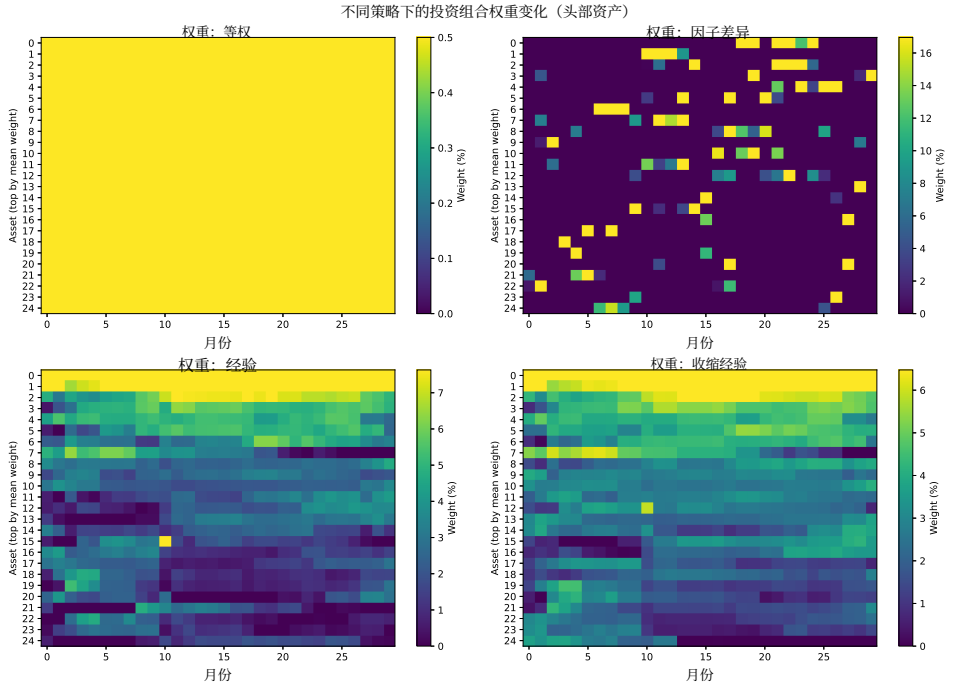


图29：权重热力图（$k = 300$）

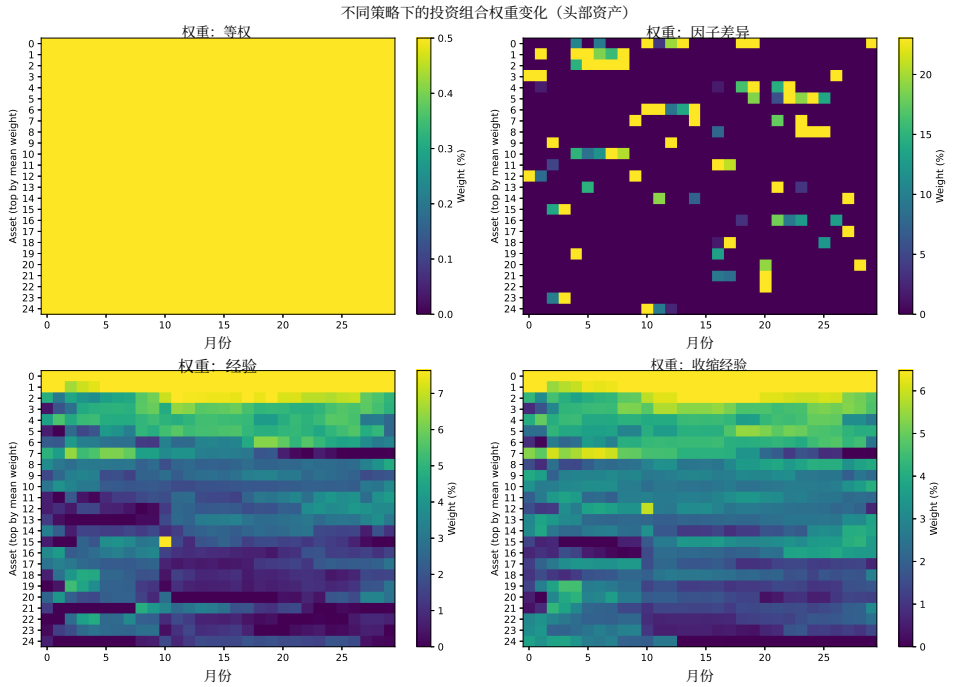


图30：权重热力图（$k = 350$）

B.3 隐式因子建模

图31展示了 Chen 等人（2023年）提出的分解方法，其中观测轨迹被分解为两个部分：捕捉系统性结构的低维线性子空间，以及表示特异性变化的正交分量。投影到该线性子空间上可以

被解释为因子的演化，而正交空间则捕捉残差噪声。与显式指定因子不同，分数估计通过对沿两个方向的数据密度梯度进行模型建模，隐式地学习这一分解。这一视角促成了隐式因子建模：扩散模型通过分数学习直接恢复低维结构，而无需预先定义因子暴露。未来研究应评估这一方法能否达到本文因子维度消融后的结果。

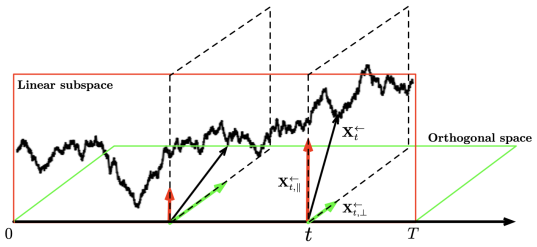


图31：来自 Chen 等人（2023年）的图像