沪深 300 股票相关矩阵分析与指数增强基金构建

作者：武亦文 程澄 吴舜奔

学号：23110190071 23110190009 25110190096

日期：2025-10-31

---

1. 研究目的

- 利用随机矩阵理论（Random Matrix Theory, RMT）对沪深 300 成分股之间的相关结构进行系统刻画，识别噪声与信息主成分，考察市场因子与行业，风格聚类等结构性特征。
- 在降噪后的相关/协方差矩阵基础上，构建面向沪深 300 的指数增强型投资组合，通过 β 因子的约束以方差最小化为目标优化投资仓位，获得可回测的“增强收益，降低风险”的表现。

---

2. 课题内容

- 利用上海股票个股价格的日线数据，计算个股价格变动的相关/协方差矩阵。并对相关矩阵进行特征值分析。
- 绘制相关矩阵的特征值谱并与随机矩阵的特征值谱进行比较。
- 对相关矩阵中与随机矩阵的特征值范围内外的特征值相对应的特征向量的分量进行分析，找到各自对应的个股的板块信息。
- 根据讲义中介绍的利用协方差矩阵构建投资组合的方法，以 30-50 个股数据构建基金，例如沪深 300 指数的ETF。比较使用完整或者去噪音后的协方差矩阵的结果。

---

3. 研究方法

3.0 数据源

- 标的范围：沪深 300 指数成分股（样本期：训练窗 2020-10-01 至 2023-10-01；回测窗 2023-10-02 至 2024-09-30）。
- 数据来源：指数及成分股数据来自 Baostock 抓取脚本 `stock_data_crawler_baostock.py`，已转存为与股票矩阵兼容的列格式，储存在 `stock_matrix` 文件夹下的 `hs300_2014-2024_matrix.csv` 和 `hs300_index_2014-2024_column.csv` 中。

3.1 相关与协方差矩阵

- 预处理：
  - 列删除缺失：仅保留在训练窗内完全无缺失的股票列。
  - 索引对齐：按日期对齐股票与指数序列。原始数据维度为 $(T+1)\times N$，数据存储类型为 `pandas` 的 `DataFrame` 格式。
- 收益矩阵：（维度为 $T\times N$）

- 对数收益率：
  $R_{t,i} = \log Y_{t,i} - \log Y_{t-1,i},$
  其中 $Y_{t,i}$ 是日期 t 第 i 支股票的收盘价。对数收益率用于计算相关/协方差矩阵与 RMT 进行对比。对数收益率矩阵由 `compute_profit_matrix` 函数计算。
- 普通收益率：
  $r_{t,i} = \frac{Y_{t,i}}{Y_{t-1,i}} - 1.$
  常规定义下的收益率用于回测以及计算夏普比率。普通收益率矩阵由 `compute_profit_matrix_simple` 函数计算。
- 两种收益率之间的关系：
  $R_{t,i} = \log(1 + r_{t,i}) \approx r_{t,i},$
  当收益率很小的时候，两种收益率近似相等。

- 相关矩阵：（维度为 $N \times N$）
  - 记收益矩阵 $R \in \mathbb{R}^{T\times N}$，按列计算均值 $\mu_i = (1/T)\sum_{t=1}^{T} r_{t,i}$ 与方差 $\sigma_i^2 = (1/T)\sum_{t=1}^{T}(r_{t,i} - \mu_i)^2$。令均值向量 $\mu = [\mu_1, \mu_2, \ldots, \mu_N] \in \mathbb{R}^{1\times N}$，标准差向量 $\sigma = [\sigma_1, \sigma_2, \ldots, \sigma_N] \in \mathbb{R}^{1\times N}$。
  - 对收益矩阵做标准化后得：
    $G = \frac{R-\mu}{\sigma}.$
    这里维度为 $1 \times N$ 的向量 μ 和 σ 会在程序中被自动扩展为维度为 $T \times N$ 的矩阵，与 R 的元素进行一一对应的运算，并非矩阵运算。这样做得到的矩阵 G 中的每个元素均值为0，标准差为1，方便后续与 chiral 型随机矩阵做比较。
  - 接下来即可计算相关矩阵：
    $C = \frac{1}{T}G^{\top}G.$
  - 这部分功能由 `compute_correlation_matrix` 函数实现。
- 协方差矩阵:（维度为 $N \times N$）
  - 直接计算法：
    $CV_{\text{direct}} = \frac{1}{T}(R-\mu)^{\top}(R-\mu).$
    在计算未去噪的协方差矩阵时可以用这个方法，比较简单直接。这部分由 `compute_covariance_matrix_direct` 函数计算。
  - 借助相关矩阵：用标准差对角矩阵 $D = \mathrm{diag}(\sigma)$ 与相关矩阵 C 复原协方差矩阵：
    $CV = D\,C\,D.$
    在计算去噪的协方差矩阵时，必须要用这个方法，因为我们首先要让相关矩阵跟 chiral 型随机矩阵做比较，去掉其中的一些本征值，而不是直接去掉协方差矩阵的一些本征值。这一部分由 `compute_covariance_matrix` 函数计算。

3.2 本征值分析与去噪

- Chiral 型随机矩阵：
  $C = \frac{1}{T}H^{\top}H,$
  其中随机矩阵 $H = [H_{i,j}] \in \mathbb{R}^{T\times N}$，$H_{i,j} \sim \mathcal{N}(0,1)$。
- 设维度比 $Q = T/N$，根据随机矩阵理论，在 $N \to \infty$，$T \to \infty$ 的极限下，Chiral 型随机矩阵的本征值有上下界：
  $\lambda_{\pm} = 1 + \frac{1}{Q} \pm 2\sqrt{\frac{1}{Q}},$
  并且满足概率分布：

$\rho(\lambda) = \frac{Q}{2\pi}\frac{\sqrt{(\lambda_+-\lambda)(\lambda-\lambda_-)}}{\lambda}$.
其每个本征向量的分量满足概率分布：
$p(u_{i,j}) = \frac{1}{\sqrt{2\pi}}\exp(-\frac{u_{i,j}^2}{2})$,
也即 $\mathcal{N}(0,1)$。

- 我们用本征分析函数 `spectral_decomposition` 对相关矩阵 C 的本征系统 $\{\lambda_i,\ \mathbf{u}_i\}_i$（按本征值从大到小排序）进行本征值分解：
$C = \sum_{i=1}^{N} \lambda_i\, \mathbf{u}_i\mathbf{u}_i^\top$.
- 之后我们可以在 `plot_eigensystem` 函数中将其与 chiral 型随机矩阵的本征值、本征向量的分布图进行比较，完成本课题的前三部分研究内容。
这里要特别注意，在比较本征向量的之前，我们要对本征向量做正规化：$\mathbf{u}_i \to \sqrt{N}\mathbf{u}_i$，这样才能使它的分量与 $\mathcal{N}(0,1)$ 处在同样的尺度下。
- 我们还可以利用 `denoising` 函数对相关矩阵实施本征值阈值截断，仅保留 $\lambda_i > \lambda_+$ 的本征对，令其他本征值为0，以构造去噪声的相关矩阵：
$C_{\rm pr} = \sum_{\lambda_i>\lambda_+} \lambda_i\, \mathbf{u}_i\mathbf{u}_i^\top$.
去噪后进行对称化并将对角线强制设为 1，以保证相关矩阵的对称性和单位对角性质。最后再通过
$CV_{\rm pr} = D\, C_{\rm pr}\, D$
重构去噪声的协方差矩阵，以供后续优化投资组合权重时使用。

3.3 股票筛选

- 在构建指数增强ETF时，我们要试图用少量的优质股票构建出比指数表现更好的投资组合，所以首先要对沪深 300 的成分股进行筛选。筛选的准则为以下两点。
- Pareto 非支配筛选：基于我们前面得到的所有股票的收益矩阵 R，计算出每支股票收益的标准差和期望 (σ_i, μ_i)，将其都画在 (σ, μ) 平面中，得到一个散点图。我们仅保留未被更高收益且更低波动的股票支配的样本。也即，若我们保留股票 i，那么不存在任何股票 j 满足：
$\mu_i \le \mu_j, \quad \sigma_i \ge \sigma_j.$
- Sharpe 填充：若通过 Pareto 非支配筛选所选出的股票数量少于目标 K（这里设置为40），则从剩余股票中按夏普比率 μ_i/σ_i 由高到低补齐至目标数。
- 这个模块由 `clean_profit_matrix` 函数实现。

3.4 投资组合权重优化

- 设权重向量 $\omega = [\omega_1, \omega_2, \ldots, \omega_N] \in \mathbb{R}^{1\times N}$，我们接下来试图通过投资组合中成分股 β 因子的约束将其优化。
- 个股相对指数的 β 因子：
$\beta_i = \frac{\mathrm{Cov}(r_i, r_m)}{\sigma_M^2}$,
其中 r_i 是股票 i 的普通日收益率，r_M 是沪深 300 指数的普通日收益率，$\mathrm{Cov}(r_i, r_m)$ 是它们在训练窗时间内的协方差，σ_M 是沪深 300 指数普通日收益率在训练窗时间内的标准差。
- 令 $\beta = [\beta_1, \beta_2, \ldots, \beta_N] \in \mathbb{R}^{1\times N}$ 为 β 因子向量，记 $u = [1, 1, \ldots, 1] \in \mathbb{R}^{1\times N}$ 为一个元素全为 1 的行向量，CV 为（去噪或未去噪）协方差矩阵。记：
$uu = uCV^{-1}u^\top, \qquad ub = uCV^{-1}\beta^\top,$
$bu = \beta CV^{-1}u^\top, \qquad bb = \beta CV^{-1}\beta^\top.$
- 在限制条件 $\beta\omega^\top = u\omega^\top = 1$ 下，我们根据拉格朗日乘子法得出最优权重：

$\omega = \frac{(bb-ub)\,uCV^{-1}+(uu-bu)\,\beta CV^{-1}}{uu\cdot bb-ub\cdot bu}$.

- 对比两套协方差矩阵 CV：未去噪 CV_{direct} 与去噪 CV_{pr}，分别代入上式得到两套权重 ω 与 ω_{pr}。
- 这部分功能由 `optimize_portfolio` 实现。

3.5 回测与评估

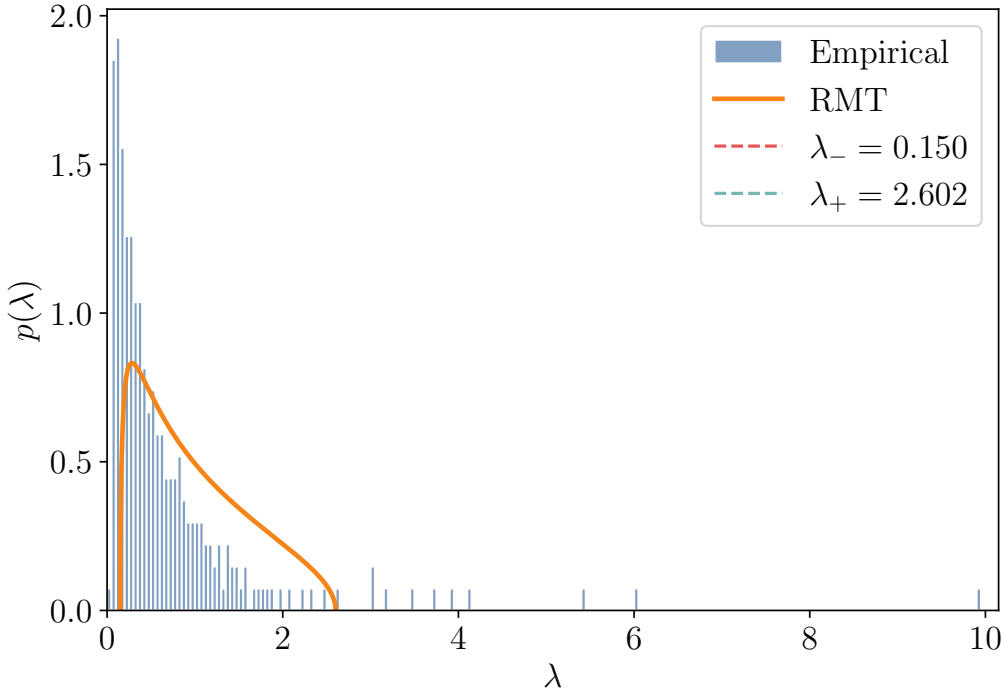
- 策略：在回测起点以初始资产 $\text{NAV}_0 = 100$ 分别按照权重 ω 与 ω_{pr} 建仓并采用“buy-and-hold”策略至窗口末尾，期间不进行调仓。为了与指数收益率做对比，我们直接调取回测窗口内每个交易日的指数数值 I_t。
- 回测时间内的累计收益曲线：
  - 第 t 个交易日的基金净值：
    $\text{NAV}_t = 100 \cdot \sum_i \omega_i \frac{Y_{t,i}}{Y_{0,i}}$.
  - 第 t 个交易日的基金累计收益：
    $\text{FundCumRet}_t = \frac{\text{NAV}_t}{\text{NAV}_0} - 1$.
  - 第 t 个交易日的指数累计收益：
    $\text{IndexCumRet}_t = \frac{I_t}{I_0} - 1$.
- 夏普比率计算：
  - 第 t 个交易日的基金日收益：
    $r_t = \frac{\text{NAV}_t}{\text{NAV}_{t-1}} - 1$,
  - 第 t 个交易日的指数日收益：
    $r_t^{\text{index}} = \frac{I_t}{I_{t-1}} - 1$.
    分别计算 50/100/150/200 日滚动均值 μ，总体标准差 σ 和夏普比率 $R_s = \mu/\sigma$；超额夏普 $R_I = \mu_{\text{ex}}/\sigma_{\text{ex}}$ 用超额日收益率 $r_t - r_t^{\text{index}}$ 计算。
- 回测模块打包为 `backtest_buy_and_hold` 函数。

---

4. 研究结果

4.1 相关矩阵本征值谱

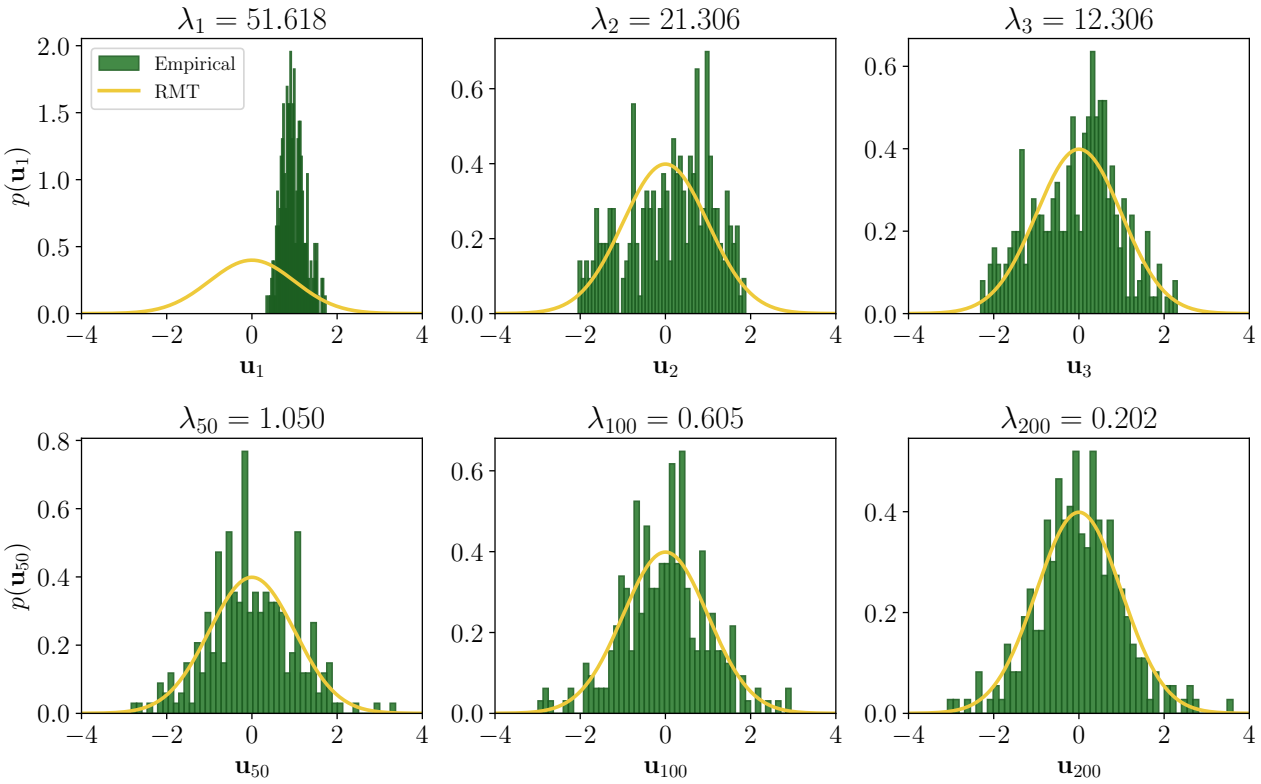
- 训练窗 2020-10-01 至 2023-10-01 内，沪深 300 指数中有 $N = 273$ 支股票数据完整，因此 $Q = T/N = 726/273 = 2.659$，对应随机矩阵本征值上界 $\lambda_+ = 2.602$。与随机矩阵的本征值谱进行对比后，我们发现实际股票的相关矩阵本征值分布更为分散，λ_+ 右侧和 λ_- 左侧都存在许多本征值。



图中最右边的本征值为第四大本征值，前三大本征值因为与主体偏离太大，不在图中展示。前三大本征值分别为 $\lambda_1 = 51.618$，$\lambda_2 = 21.306$，$\lambda_3 = 12.306$。

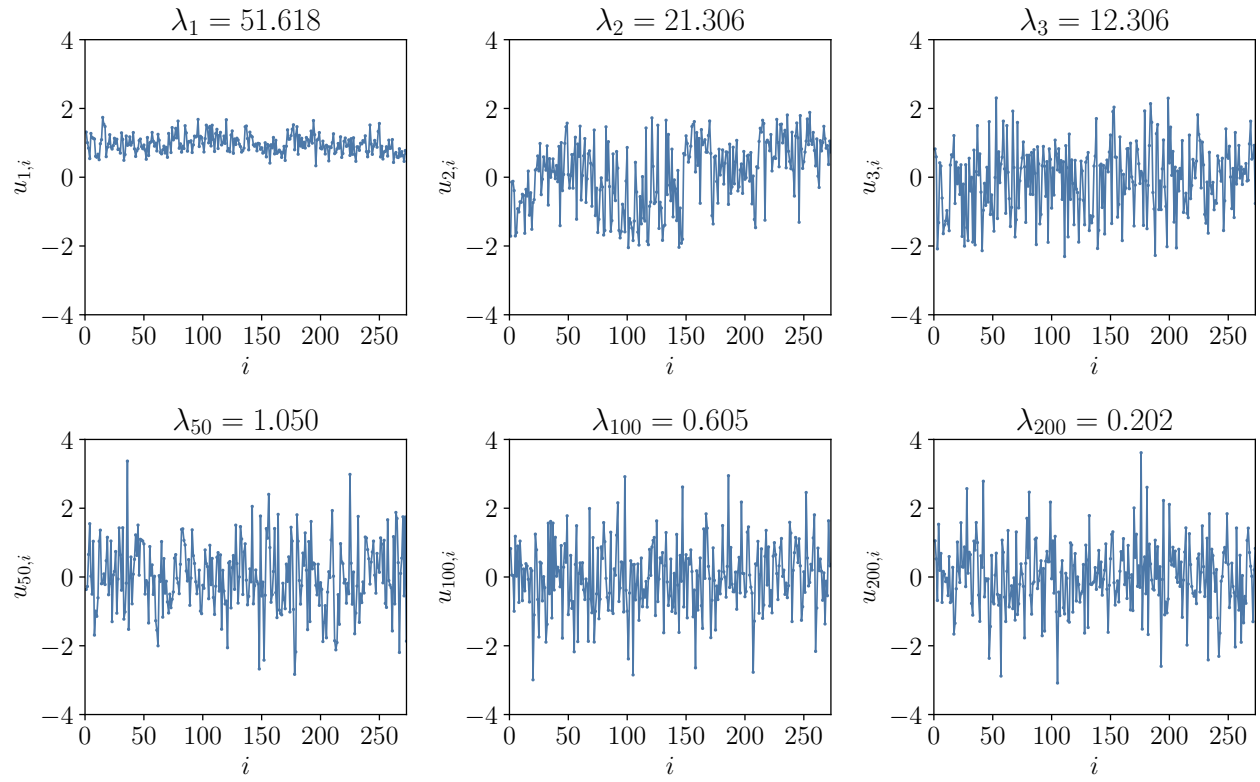
4.2 相关矩阵本征向量谱

- 接下来我们将对前三大本征值，以及第五十，一百，一百五十大本征值对应的本征向量进行分析。首先绘制本征向量分量的分布图，并与随机矩阵的本征向量分量分布，也就是与标准正态分布 $\mathcal{N}(0,1)$ 做比较。



我们发现处于 λ_+ 右侧的前三大本征值对应的本征向量分量分布与随机矩阵偏差极大，它们看起来比正态分布更为集中。而在 $[\lambda_-, \lambda_+]$ 范围内的第五十，一百，一百五十大本征值对应的本征向量分量分布基本符合随机矩阵的分布。

- 接下来我们这六个本征向量的分量都展示出来，每个分量对应每支股票，试图分析其对应的板块信息。



我们发现第一大本征值对应的本征向量分量都为正。作为第一大主成分，其本征向量分量对应市场因子。此外，不同于第五十，一百，一百五十大本征值对应的本征向量分量呈现出的高斯白噪声形状，可以观察到第二大和第三大本征值对应的本征向量分量存在簇集现象。分析这两个本征向量的分量，可以挖掘出行业和板块的聚类现象。我们节选了一些观察到的股票以做展示：

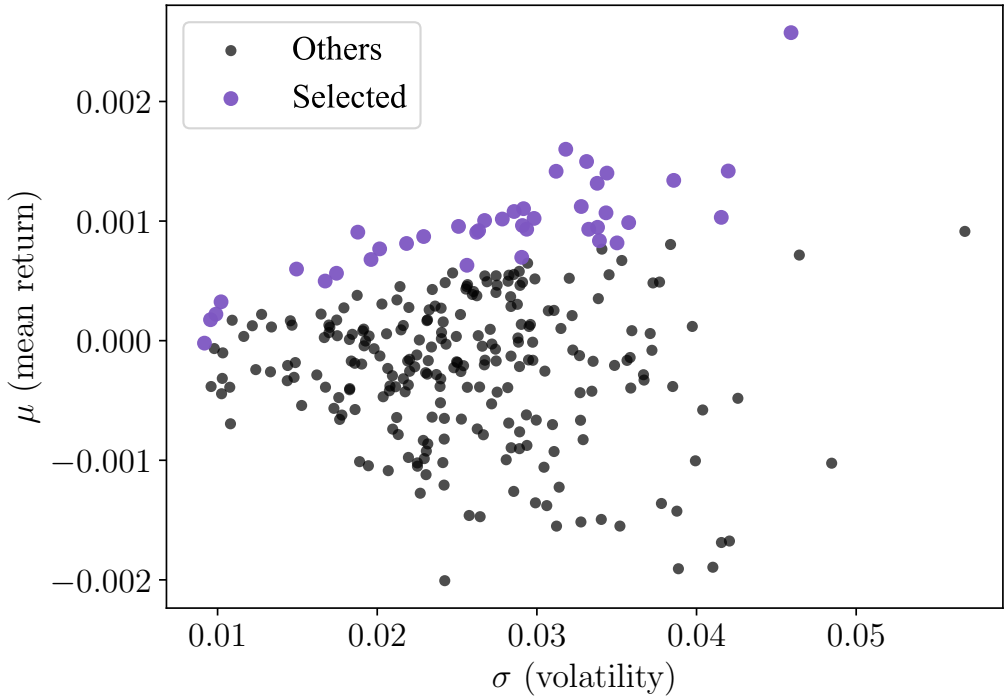
$\mathbf{u}_{2,+}$ (编号)	$\mathbf{u}_{2,+}$ (行业)	$\mathbf{u}_{2,-}$ (编号)	$\mathbf{u}_{2,-}$ (行业)	$\mathbf{u}_{3,+}$ (编号)	$\mathbf{u}_{3,+}$ (行业)	$\mathbf{u}_{3,-}$ (编号)	$\mathbf{u}_{3,-}$ (行业)
sh.601398	银行	sz.002709	电子	sh.600690	电气设备	sz.002463	电子
sh.601939	银行	sz.300782	电子	sz.000333	电气设备	sh.603019	电子
sh.601328	银行	sh.600460	电子	sh.601012	电气设备	sz.300502	电子
sh.601288	银行	sz.002049	电子	sh.600438	电气设备	sz.300308	电子
sh.601988	银行	sz.002180	电子	sz.300014	电气设备	sz.000938	电子
sh.601169	银行	sh.603501	电子	sz.000651	电气设备	sh.688041	电子
sh.600019	有色金属	sz.002459	电气设备	sz.002142	银行	sh.601985	电力
sz.003816	电力	sh.600276	医药	sz.000786	有色金属	sh.601288	银行
sh.601658	银行	sz.002938	电子	sz.002459	电气设备	sz.000977	电子
sh.600015	银行	sz.002475	电子	sz.300274	电气设备	sz.300394	电子

从表格中可以看出，第二大本征向量正分量 $\mathbf{u}_{2,+}$ 对应的行业多为银行和证券等红利板块，而负分量 $\mathbf{u}_{2,-}$ 对应的行业则有很多是电子和半导体等科技板块。第三大本征向量正分量 $\mathbf{u}_{3,+}$ 对应的行业中有很多电气设备，而负分量 $\mathbf{u}_{3,-}$ 还是对应电子和半导体等科技板块。

4.3 指数增强型基金的构建

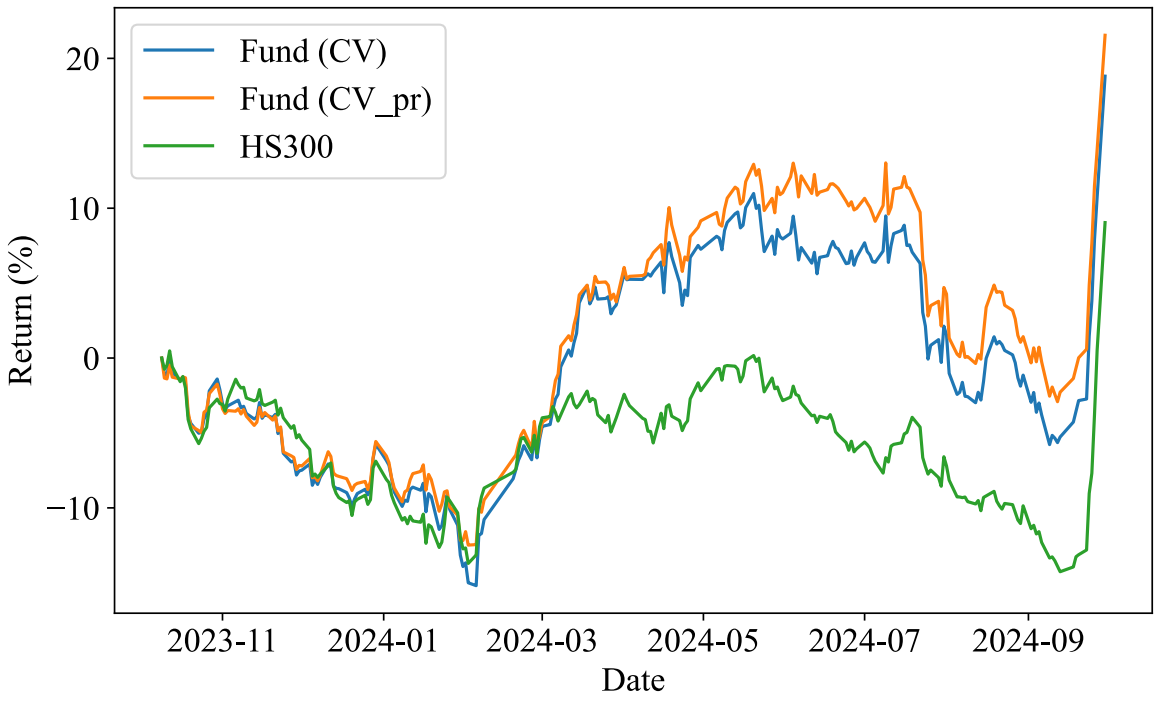
- 选股集：在沪深 300 指数的 273 支在训练窗内数据完整的成分股中，我们进一步通过 Pareto 前沿严选出 10~30 支样本，再通过 Sharpe 填充补齐至 40 支。我们绘制了散点图（紫色为精选集合，黑色为其它），

用于观察风险-收益的关系与边界形态。



散点图显示入选样本在低波动高收益区域分布更集中。此外，我们还可以观察到散点的边界大致形成了 Markowitz Bullet 的形状，而我们精选的股票基本都处于有效界限的附近，这为其超越沪深 300 奠定了基础。

- 基于 CV_{direct} 构造的基金：
  - 最优权重 ω 的分量通常在 $0.1\% \sim 10\%$ 的数量级，最大不超过 25%。
- 基于 CV_{pr} 构造的基金：
  - 去噪声：在计算出 40 支精选股票的相关矩阵 C 之后，我们以 $\lambda_i > \lambda_+$ 截断后重构 C_{pr}，以削弱噪声本征模式，保留少数信息主成分。进一步设置单位对角后，C_{pr} 符合了随机变量与自身的关联系数等于 1 的定义，并且重新变得满秩，使在接下来的优化权重过程中对去噪协方差矩阵 $CV_{\text{pr}} = D\,C_{\text{pr}}\,D$ 的求逆更稳健。
  - 最优权重 ω_{pr} 在 C_{pr} 添加单位对角限制后，权重幅度与稳定性与 ω 可比。
- 回测：我们在回测窗 2023-10-02 至 2024-09-30 内，对比了基于 CV_{direct} 和 CV_{pr} 构建的两只基金与沪深 300 指数的收益率与波动率。
  - 我们首先做三线同图的累计收益曲线。



可以看到，在熊市和牛市中，我们的两只基金都表现出了良好的指数跟随效果。在牛市中，我们的两只基金能获得可观的超额收益，并且控制了波动，提高了夏普比率。另外，正如我们所预期的，基于去噪协方差矩阵构造的基金几乎完胜了我们的另一只基金，收益更高的同时波动更小。

- 进一步地，我们在多个时间窗口下给出了两只基金的日化夏普比率与超额日化夏普比率。若想得到年化/区间化夏普比率，可以乘以根号下的窗口时间 $\sqrt{T}$ 得到。下表汇总了六组指标在 50/100/150/200 日窗口下的结果（μ 为窗口内日均收益，σ 为窗口内日总体标准差，$R_s = \mu/\sigma$ 为日化夏普比率；μ_{ex} 为窗口内超额日均收益，σ_{ex} 为窗口内超额日收益总体标准差，$R_I = \mu_{\text{ex}}/\sigma_{\text{ex}}$ 为超额日化夏普比率）。

  基于 CV_{direct} 的基金：

时间	$\mu(\%)$	$\sigma(\%)$	R_s	$\mu_{\text{ex}}(\%)$	$\sigma_{\text{ex}}(\%)$	R_I
50日	0.215	1.766	0.122	-0.065	0.704	-0.092
100日	0.113	1.423	0.079	0.003	0.700	0.004
150日	0.171	1.305	0.131	0.063	0.705	0.090
200日	0.131	1.280	0.102	0.050	0.658	0.075

  基于 CV_{pr} 的基金：

时间	$\mu(\%)$	$\sigma(\%)$	R_s	$\mu_{\text{ex}}(\%)$	$\sigma_{\text{ex}}(\%)$	R_I
50日	0.191	1.736	0.110	-0.089	0.846	-0.105
100日	0.120	1.394	0.086	0.011	0.798	0.013
150日	0.178	1.267	0.140	0.070	0.806	0.087
200日	0.140	1.216	0.115	0.058	0.763	0.076

5. 讨论与考察

- 负权重的出现：在构建基金时，我们发现程序计算出的最优权重向量中有很多元素为负数。这个现象并不意外，因为我们的限制条件仅有 $\beta\omega^\top = u\omega^\top = 1$。相比于存在正权限制 $\omega_i \geq 0$ 的投资组合，负权重的出现意味着我们需要按照仓位做空一些股票，用做空卖出时得到的资金再额外买入一些股票；并在结算时，买回做空的股票，卖出做多的股票。但在实际的市场中，这样的操作未必容易实现。或许我们还需观察加上正权限制后得到的基金的表现。
- 去噪时的对角约束：将去噪相关矩阵对角强制为 1 是保持相关矩阵性质的必要步骤，也在实践上提升了协方差矩阵的逆矩阵的数值稳定性。如果去掉这一步，计算出的最优权重将爆炸，虽然求和为 1，但其中每个分量的绝对值都远远大于 1，有很多很大的正数和很大的负数，相当于上了很多做空和做多的杠杆。
- 回测结果的稳定性：虽然在累积收益率曲线上我们已经观察到我们所构建的基金相对沪深 300 指数的显著优势，以及去噪操作带来的进一步提升，但夏普比率的计算过程还存在问题。
    - 我们展示的是日化夏普比率，但即使转化为区间化夏普比率，也没有观察到它随着窗口时间增长的收敛情况。我们认为可能是时间长度还不够，应该增加到 300 日和 500 日再观察。
    - 其次，日收益率也没有和累积收益曲线对应上。至少从图中看，前 50 日所有基金都处于持续亏损状态，而表格中给出的收益率的数值却是一个较大的正数。所以后续还需进行计算程序的检查。
    - 此外，回测窗口、选股目标数 K、Sharpe 填充都会影响最终表现。之后还应进行多窗口、多起点与灵敏度分析。

---

6. 结论

- 我们对沪深 300 指数成分股进行了关联矩阵的本征分析，从其主成分中观察到了市场因子和潜在的行业结构。
- 我们从沪深 300 指数成分股中精选出 40 支优质股票，分别通过完整和去噪声的协方差矩阵构建了两只指数增强基金。

---

7. 参考文献

1. V. Plerou, P. Gopikrishnan, B. Rosenow, L. A. N. Amaral, T. Guhr, and H. E. Stanley, Random matrix approach to cross correlations in financial data, Phys. Rev. E 65, 066126 (2002).

附录 A：项目链接

本项目的完整代码可在Github 开源仓库中获取。

附录 B：主要符号

- 时间长度：T；股票数量：N。
- 股价矩阵：Y；收益矩阵：R；相关矩阵：C；协方差矩阵：CV；标准差矩阵：D。
- 去噪相关矩阵：C_{pr}；去噪协方差矩阵：CV_{pr}

附录 C：成员量化贡献

姓名	贡献度	具体贡献
武亦文	50%	计算相关矩阵并做本征分析；基于协方差矩阵构建指数增强基金
程澄	30%	基于本征向量分析行业聚类现象
吴舜奔	20%	前期收集股票数据