基于 S&P500 指数成分股的高维资产组合优化实证研究

谢楚焓，数学科学学院，2101110085

2022 年 11 月 11 日

目录

1 背景介绍 **2**

2 数据说明 **4**

3 算法介绍 **4**

3.1 自助法增强 4

3.2 谱修正估计 5

3.2.1 方法一：矩估计法 6

3.2.2 方法二：最小二乘法 7

4 实证结果 **7**

4.1 样本、总体相关系数矩阵的谱分布估计 7

4.2 Markowitz 模型问题的最优估计 8

4.3 不同估计的回测结果 9

A 附录一：反立方密度的积分结论 **18**

1 背景介绍

Markowitz 模型是由 Harry Markowitz 于 1952 年提出的关于资产组合优化问题的一个经典模型，在现代金融理论中具有里程碑式的意义。其假设有 p 个金融资产 $\{s_1, \ldots, s_p\}$，对应的收益率向量 $\mathbf{x} = (x_1, \ldots, x_p)'$ 具有均值 $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_p)'$ 及协方差矩阵 $\mathrm{cov}(\mathbf{x}) = \boldsymbol{\Sigma} = (\sigma_{ij})$。投资者在这 p 个资产上进行资产配置，对应的配置权重 $\mathbf{c} = (c_1, \ldots, c_p)$ 满足 $\mathbf{c}'\mathbf{1} \leq 1$，并希望在最大化期望收益率 $R = \mathbf{c}'\boldsymbol{\mu}$ 的同时，将资产组合的方差 $r = \mathbf{c}'\boldsymbol{\Sigma}\mathbf{c}$ 控制在较低水平。具体而言，我们考虑如下优化问题：

$$\max_{\mathbf{c}} \mathbf{c}'\boldsymbol{\mu} \quad \text{subject to } \mathbf{c}'\mathbf{1} \leq 1 \text{ and } \mathbf{c}'\boldsymbol{\Sigma}\mathbf{c} \leq \sigma_0^2. \tag{1}$$

该优化问题具有解析解，由以下命题1.1给出。

命题 1.1. *对于优化问题(1)，*

1. *若* $\frac{\mathbf{1}'\boldsymbol{\Sigma}\boldsymbol{\mu}\sigma_0}{\sqrt{\boldsymbol{\mu}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}}} < 1$，*则最优的期望收益率* R *和对应的配置权重* $\mathbf{c}$ *分别为*

$$R^{(1)} = \sigma_0\sqrt{\boldsymbol{\mu}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}}, \quad \mathbf{c}_1 = \frac{\sigma_0}{\sqrt{\boldsymbol{\mu}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu};$$

2. *若* $\frac{\mathbf{1}'\boldsymbol{\Sigma}\boldsymbol{\mu}\sigma_0}{\sqrt{\boldsymbol{\mu}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}}} > 1$，*则最优的期望收益率* R *和对应的配置权重* $\mathbf{c}$ *分别为*

$$R^{(2)} = \frac{\mathbf{1}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}}{\mathbf{1}'\boldsymbol{\Sigma}^{-1}\mathbf{1}} + b\left(\boldsymbol{\mu}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} - \frac{(\mathbf{1}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu})^2}{\mathbf{1}'\boldsymbol{\Sigma}^{-1}\mathbf{1}}\right),$$
$$\mathbf{c}_2 = \frac{\boldsymbol{\Sigma}^{-1}\mathbf{1}}{\mathbf{1}'\boldsymbol{\Sigma}^{-1}\mathbf{1}} + b\left(\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} - \frac{\mathbf{1}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}}{\mathbf{1}'\boldsymbol{\Sigma}^{-1}\mathbf{1}}\boldsymbol{\Sigma}^{-1}\mathbf{1}\right),$$

其中

$$b = \sqrt{\frac{\mathbf{1}'\boldsymbol{\Sigma}^{-1}\mathbf{1}\sigma_0^2 - 1}{\boldsymbol{\mu}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\mathbf{1}'\boldsymbol{\Sigma}^{-1}\mathbf{1} - (\mathbf{1}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu})^2}}.$$

我们的目标是基于上述 Markowitz 模型，给出 p 个资产上的最优配置，并估计最优配置下的期望收益率。由于模型中的 $\boldsymbol{\mu}, \boldsymbol{\Sigma}$ 均未知，故需要根据资产历史数据进行估计。最简单的估计方法是利用 n 个时刻的样本均值、样本协方差，即

$$\bar{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i, \quad \mathbf{S} = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})'.$$

将其代入命题1.1的公式中，即得到 $R, \mathbf{c}$ 的 plug-in 估计

$$\hat{R}_p = \hat{\mathbf{c}}_p'\bar{\mathbf{x}}, \quad \hat{\mathbf{c}}_p = \begin{cases} \dfrac{\sigma_0\mathbf{S}^{-1}\bar{\mathbf{x}}}{\sqrt{\bar{\mathbf{x}}'\mathbf{S}^{-1}\bar{\mathbf{x}}}}, & \text{if } \dfrac{\sigma_0\mathbf{1}'\mathbf{S}^{-1}\bar{\mathbf{x}}}{\sqrt{\bar{\mathbf{x}}'\mathbf{S}^{-1}\bar{\mathbf{x}}}} < 1, \\ \dfrac{\mathbf{S}^{-1}\mathbf{1}}{\mathbf{1}'\mathbf{S}^{-1}\mathbf{1}} + \hat{b}\left(\mathbf{S}^{-1}\bar{\mathbf{x}} - \dfrac{\mathbf{1}'\mathbf{S}^{-1}\bar{\mathbf{x}}}{\mathbf{1}'\mathbf{S}^{-1}\mathbf{1}}\mathbf{S}^{-1}\mathbf{1}\right), & \text{if } \dfrac{\sigma_0\mathbf{1}'\mathbf{S}^{-1}\bar{\mathbf{x}}}{\sqrt{\bar{\mathbf{x}}'\mathbf{S}^{-1}\bar{\mathbf{x}}}} > 1, \end{cases} \tag{2}$$

其中

$$\hat{b} = \sqrt{\frac{\sigma_0^2 \mathbf{1}'\mathbf{S}^{-1}\mathbf{1} - 1}{(\bar{\mathbf{x}}'\mathbf{S}^{-1}\bar{\mathbf{x}})(\mathbf{1}'\mathbf{S}^{-1}\mathbf{1}) - (\mathbf{1}'\mathbf{S}^{-1}\bar{\mathbf{x}})^2}}.$$

然而许多实证研究表明，plug-in 估计得到的配置方案并不非常有效，其所获得的收益率有时甚至不如对所有资产做等权组合所获得的的收益率；这种现象也被称作 Markowitz optimization enigma。进一步的理论研究发现，plug-in 估计通常会高估真实值（见定理1.1），从而解释了这一现象。

定理 1.1. *假设历史数据 $\chi = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ 独立同分布，具有均值 $\boldsymbol{\mu}$ 和协方差矩阵 $\boldsymbol{\Sigma}$，且有有限的四阶矩。若当 $p, n \to \infty$ 时，$p/n \to y \in (0, 1)$，且下列极限存在：*

$$\frac{\mathbf{1}'\boldsymbol{\Sigma}^{-1}\mathbf{1}}{n} \to a_1, \quad \frac{\mathbf{1}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}}{n} \to a_2, \quad \frac{\boldsymbol{\mu}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}}{n} \to a_3.$$

则在几乎处处的意义下，

$$\lim_{n\to\infty} \frac{\hat{R}_p}{\sqrt{n}} = \begin{cases} \sigma_0\sqrt{\gamma a_3} > \lim\limits_{n\to\infty} \dfrac{R^{(1)}}{\sqrt{n}} = \sigma_0\sqrt{a_3}, & \text{if } a_2 < 0, \\ \sigma_0\sqrt{\gamma(a_3 - a_2^2/a_1)} > \lim\limits_{n\to\infty} \dfrac{R^{(2)}}{\sqrt{n}} = \sigma_0\sqrt{(a_3 - a_2^2/a_1)}, & \text{if } a_2 > 0, \end{cases}$$

其中 $R^{(1)}, R^{(2)}$ 由命题1.1给出，$\gamma = \frac{1}{1-y} > 1$。

定理1.1说明，plug-in 估计在高维情形下是渐近有偏的。因此，我们需要利用高维统计、随机矩阵中的理论结果构造渐近无偏的估计。本文针对此问题，分别采用自助法增强（bootstrap enhancement）[Bai et al., 2009] 和谱修正估计（spectrum-corrected estimator）[Bai et al., 2013] 两种方法对 plug-in 估计进行修正，并在实际数据集上进行实证研究。同时，plug-in 估计的偏差很大程度上在于利用样本协方差估计总体协方差产生的偏误，因此刻画二者的统计性质也是研究重点。

研究目标 本文研究目标主要有以下三点：

1. 研究样本协方差 $\mathbf{S}$ 的谱分布等统计性质，并根据谱修正理论（具体见第3.2节）估计对应的总体协方差 $\boldsymbol{\Sigma}$；

2. 根据自助法增强和谱修正估计两种方法对 Markowitz 模型问题中的 plug-in 估计进行修正，并研究三种方法下的最优期望收益率估计与最优配置估计的性质；

3. 比较三种方法得到的下的最优配置估计在历史数据上的回测表现。

2 数据说明

本文采用 S&P500 指数成分股在 2013 年 2 月 8 日至 2018 年 2 月 7 日五年间每日的开盘价、收盘价等数据，由此构造日对数收益率，其公式为

$$\text{日对数收益率} = \ln\left(\frac{\text{当日收盘价}}{\text{上一收盘价}}\right).$$

原始数据可从https://www.kaggle.com/datasets/camnugent/sandp500?resource=download获得。使用不同的样本个数 n 时，默认取距离现在最近的 n 个交易日的数据（即从后往前取）。使用不同的成分股个数 p 时，我们从所有成分股中进行随机选取。以苹果公司（股票代码：AAPL）为例，其收益率随时间变化图及直方图如图1。可见，收益率在 ± 0.1 的范围内波动，且具有均值回归的特性；收益率分布较符合正态分布。

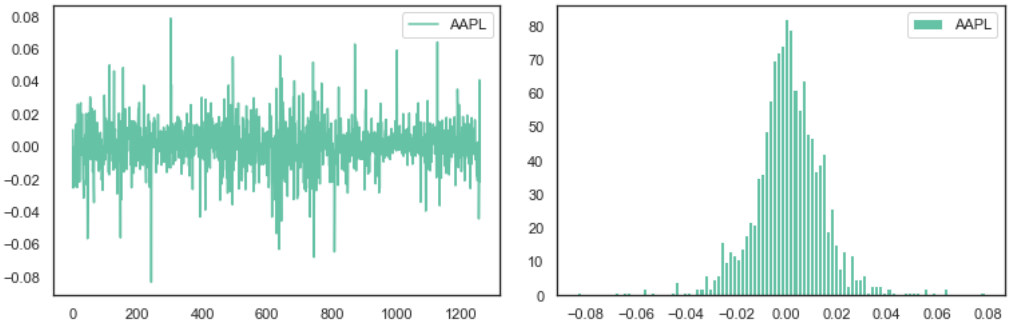

图 1: AAPL 收益率随时间变化图及直方图

3 算法介绍

3.1 自助法增强

我们考虑利用自助法（parametric bootstrap method）估计 $R - \hat{R}_p$。具体而言，我们从正态分布 $\mathcal{N}_p(\bar{\mathbf{x}}, \mathbf{S})$ 中重采样得到 $\chi^* = \{\mathbf{x}_1^*, \ldots, \mathbf{x}_n^*\}$，并利用该样本基于(2)得到“自助 plug-in”估计 $\hat{R}_p^*, \hat{\mathbf{c}}_p^*$。自助 plug-in 估计具有如下性质（见定理3.1）。

定理 3.1. *在定理1.1的条件下，有*

$$\sqrt{\gamma}(R - \hat{R}_p) \simeq \hat{R}_p - \hat{R}_p^*,$$

其中 γ 由定理1.1定义，R 为真实最优期望收益率，$\hat{R}_p, \hat{R}_p^*$ 分别为关于 R 的 plug-in 和自助 plug-in 估计。

基于定理3.1，我们可以得到利用自助法增强后的估计

$$\hat{R}_b = \hat{R}_p + \frac{1}{\sqrt{\gamma}}(\hat{R}_p - \hat{R}_p^*), \tag{3}$$

$$\hat{\mathbf{c}}_b = \hat{\mathbf{c}}_p + \frac{1}{\sqrt{\gamma}}(\hat{\mathbf{c}}_p - \hat{\mathbf{c}}_p^*). \tag{4}$$

实际中会重采样多次并将所得的 $\hat{R}_b, \hat{\mathbf{c}}_b$ 做平均得到最终估计。

3.2 谱修正估计

注意到，最优期望收益率 R 完全取决于关于 $\mathbf{\Sigma}$ 的二次型 $\mathbf{1}'\mathbf{\Sigma}^{-1}\mathbf{1}, \mathbf{1}'\mathbf{\Sigma}^{-1}\boldsymbol{\mu}, \boldsymbol{\mu}\mathbf{\Sigma}^{-1}\boldsymbol{\mu}$。当 $p, n \to \infty$ 时，$\mathbf{S}^{-1}$ 并不是 $\mathbf{\Sigma}^{-1}$ 的一个较好估计，导致引入较大误差。为了克服这种误差，通常采取的方法之一是利用 $\mathbf{S}$ 的极限谱分布（limiting spectral distribution, LSD）与 $\mathbf{\Sigma}$ 的总体谱分布（population spectral distribution, PSD）之间的关系得到 $\mathbf{\Sigma}$ 的一个谱分布估计，再构造经修正的协方差估计。

不妨假设各成分股的方差均为 1，从而协方差矩阵 $\mathbf{\Sigma}$ 同时也是相关系数矩阵。在谱修正估计（spectrum-corrected estimator）中，我们假设 $\mathbf{\Sigma}$ 的谱分解为 $\mathbf{\Sigma} = \mathbf{U}\mathbf{\Lambda}_p\mathbf{U}'$，且当 $p \to \infty$ 时，$\mathbf{\Sigma}$ 的谱分布 H_p 有极限 $H = \lim_{p\to\infty} H_p$。这里的 H 被称作总体谱分布（PSD）。考虑参数分布族 $\{H(\theta)\}_{\theta\in\Theta}$，其满足 $H = H(\theta_0)$。假设我们可以通过数据拟合得到参数估计 $\hat{\theta}$，则谱修正由以下定义3.1给出。估计 $\hat{\theta}$ 的方法见第3.2.1-3.2.2节。

定义 3.1. *设 $\mathbf{S} = \mathbf{V}\mathbf{D}_p\mathbf{V}'$ 为样本协方差 $\mathbf{S}$ 的谱分解，$\hat{\theta}$ 是 θ_0 的一个相合估计。则 $\widehat{\mathbf{\Sigma}}_s = \mathbf{V}\widehat{\mathbf{\Lambda}}_p\mathbf{V}'$ 被称作总体协方差 $\mathbf{\Sigma}$ 的谱修正。*

最后，用 $\widehat{\mathbf{\Sigma}}_s$ 代替(2)中的 $\mathbf{S}$，得到 $R, \mathbf{c}$ 的谱修正估计

$$\hat{R}_s = \hat{\mathbf{c}}_s'\bar{\mathbf{x}}, \quad \hat{\mathbf{c}}_s = \begin{cases} \dfrac{\sigma_0\widehat{\mathbf{\Sigma}}_s^{-1}\bar{\mathbf{x}}}{\sqrt{\bar{\mathbf{x}}'\widehat{\mathbf{\Sigma}}_s^{-1}\bar{\mathbf{x}}}}, & \text{if } \dfrac{\sigma_0\mathbf{1}'\widehat{\mathbf{\Sigma}}_s^{-1}\bar{\mathbf{x}}}{\sqrt{\bar{\mathbf{x}}'\widehat{\mathbf{\Sigma}}_s^{-1}\bar{\mathbf{x}}}} < 1, \\ \dfrac{\widehat{\mathbf{\Sigma}}_s^{-1}\mathbf{1}}{\mathbf{1}'\widehat{\mathbf{\Sigma}}_s^{-1}\mathbf{1}} + \hat{b}_s\left(\widehat{\mathbf{\Sigma}}_s^{-1}\bar{\mathbf{x}} - \dfrac{\mathbf{1}'\widehat{\mathbf{\Sigma}}_s^{-1}\bar{\mathbf{x}}}{\mathbf{1}'\widehat{\mathbf{\Sigma}}_s^{-1}\mathbf{1}}\widehat{\mathbf{\Sigma}}_s^{-1}\mathbf{1}\right), & \text{if } \dfrac{\sigma_0\mathbf{1}'\widehat{\mathbf{\Sigma}}_s^{-1}\bar{\mathbf{x}}}{\sqrt{\bar{\mathbf{x}}'\widehat{\mathbf{\Sigma}}_s^{-1}\bar{\mathbf{x}}}} > 1, \end{cases} \tag{5}$$

其中

$$\hat{b}_s = \sqrt{\frac{\sigma_0^2\mathbf{1}'\widehat{\mathbf{\Sigma}}_s^{-1}\mathbf{1} - 1}{(\bar{\mathbf{x}}'\widehat{\mathbf{\Sigma}}_s^{-1}\bar{\mathbf{x}})(\mathbf{1}'\widehat{\mathbf{\Sigma}}_s^{-1}\mathbf{1}) - (\mathbf{1}'\widehat{\mathbf{\Sigma}}_s^{-1}\bar{\mathbf{x}})^2}}.$$

3.2.1 方法一：矩估计法

矩估计法是一种估计 H 的方法，其利用了样本协方差 $\mathbf{S}$ 的 LSD $F_{y,H}$ 与总体协方差 $\mathbf{\Sigma}$ 的 PSD H 之间的关系（见引理3.1）。进一步假设

$$\mathbf{\Sigma} = \mathbf{U}\mathbf{\Lambda}_p\mathbf{U}', \quad \text{with } \mathbf{\Lambda}_p = \operatorname{diag}(\underbrace{\lambda_1, \ldots, \lambda_1}_{p_1}, \underbrace{\lambda_2, \ldots, \lambda_2}_{p_2}, \ldots, \underbrace{\lambda_L, \ldots, \lambda_L}_{p_L}),$$

其中 L 是一个给定的有限正整数，$p_1 + \cdots + p_L = p$；且当 $p \to \infty$ 时，$\mathbf{\Sigma}$ 的谱分布 H_p 有如下极限

$$H = \lim_{p\to\infty} H_p = \sum_{j=1}^{L} w_j \delta_{\lambda_j}, \quad \text{with } w_1 + \cdots + w_L = 1.$$

于是 $\theta = \{w_j, \lambda_j \colon 1 \le j \le L\}$ 是刻画 PSD 的参数。

引理 3.1. *LSD $F_{y,H}$ 的矩 $\alpha_j = \int x^j \mathrm{d}F_{y,H}(x)$ 与 PSD H 的矩 $\beta_j = \int t^j \mathrm{d}H(t)$ 之间有如下关系：*

$$\alpha_j = y^{-1} \sum_{j=i_1+2i_2+\cdots+ji_j, i_l\in\mathbb{N}} y^{i_1+i_2+\cdots+i_j} \beta_1^{i_1}\beta_2^{i_2}\ldots\beta_j^{i_j} \phi^{(j)}_{i_1,i_2,\ldots,i_j},$$

其中

$$\phi^{(j)}_{i_1,i_2,\ldots,i_j} = \frac{j!}{i_1!i_2!\ldots i_j!(j+1-(i_1+i_2+\cdots+i_j))!}.$$

取前 $2L-1$ 阶矩 $\{\alpha_j \colon 1 \le j \le 2L-1\}$ 和 $\{\beta_j \colon 1 \le j \le 2L-1\}$，由引理3.1知存在函数 $\Psi \colon \mathbb{R}^{2L-1} \to \mathbb{R}^{2L-1}$ 使得 $(\alpha_1, \ldots, \alpha_{2L-1}) = \Psi(\beta_1, \ldots, \beta_{2L-1})$。由矩的定义，$(\beta_1, \ldots, \beta_{2L-1})$ 可被表示为关于 θ 的函数

$$(\beta_1, \ldots, \beta_{2L-1}) = \left(\sum_{l=1}^{L} w_l\lambda_l, \sum_{l=1}^{L} w_l\lambda_l^2, \ldots, \sum_{l=1}^{L} w_l\lambda_l^{2L-1}\right) =: \Phi(\theta).$$

从而 $(\alpha_1, \ldots, \alpha_{2L-1}) = \Psi \circ \Phi(\theta)$。矩估计 $\hat{\theta}_{mom}$ 是矩方程 $(\hat{\alpha}_1, \ldots, \hat{\alpha}_{2L-1}) = \Psi \circ \Phi(\theta)$ 的解，其中 $\{\hat{\alpha}_l\}_{l=1}^{2L-1}$ 是 $\mathbf{S}$ 的样本谱分布（empirical spectral distribution, ESD）的矩

$$\hat{\alpha}_l = \frac{1}{p}\mathrm{tr}(\mathbf{S}^l) = \frac{1}{p}\sum_{j=1}^{p} \lambda_j(\mathbf{S})^l, \quad l = 1, \ldots, 2L-1.$$

3.2.2 方法二：最小二乘法

最小二乘法是另一种估计 H 的方法，其利用了随机矩阵理论中的 Silverstein 等式（见定理3.2）。

定理 3.2. *样本协方差 $\mathbf{S}$ 的 LSD $F_{y,H}$ 所对应的伴随 Stieltjes 变换 $\underline{s}(z)$ 满足如下 Silverstein 等式：*

$$z = -\frac{1}{\underline{s}(z)} + y\int \frac{t}{1+t\underline{s}(z)}\mathrm{d}H(t), \quad z \in \mathbb{C}\backslash\Gamma_F, \tag{6}$$

其中 H 是总体协方差 $\mathbf{\Sigma}$ 的 PSD，Γ_F 是 $F_{y,H}$ 的支撑集。

对于样本协方差 $\mathbf{S}$ 及其 ESD F_n，其伴随 Stieltjes 变换为

$$\underline{s}_n(u) = -\frac{1-p/n}{u} + \frac{1}{n}\sum_{l=1}^{p}\frac{1}{\lambda_l(\mathbf{S})-u}. \tag{7}$$

于是最小二乘法由以下三步构成：

1. 选择一个实数点集 $\{u_1,\ldots,u_m\}$，其中 $m \geq 2L-1$；
2. 对每个 u_j，根据(7)计算 $\underline{s}_n(u_j)$ 并代入(6)右边，得到

$$\hat{u}_j(\underline{s}_{nj},\theta) := -\frac{1}{\underline{s}_n(u_j)} + \frac{p}{n}\int \frac{t\mathrm{d}H(t,\theta)}{1+t\underline{s}_n(u_j)}; \tag{8}$$

3. 得到最小二乘估计

$$\hat{\theta}_{lse} = \underset{\theta}{\operatorname{argmin}}\sum_{j=1}^{m}(u_j - \hat{u}_j(\underline{s}_{nj},\theta))^2.$$

4 实证结果

4.1 样本、总体相关系数矩阵的谱分布估计

我们取 $n = 500, 1000$ 和 $p = 10, 50, 200, 400$，分别在不同情形下对数据的相关系数矩阵（经标准化后的 $\mathbf{S}$）做谱分解，得到特征值如图2。我们发现，在所有情况下总有部分特征值非常大。这在 $p \ll n$ 时较好理解，大特征值代表其对应的特征向量在 PCA 中解释大部分方差，从而只需要少数维度即可保留原数据大部分信息，达到降维的效果。但在 $p \sim n$ 时，该现象违反随机矩阵理论：$n, p \to \infty$ 时，$\mathbf{S}$ 的谱分布收敛到一个具有连续支撑集的分布，这代表不应有离群的特征值出现。因此成分股数据的相关性本身存在更精细的结构，其常被建模为尖刺模型（spiked model），

离群特征值被称作尖刺（spike）。本文不再深入讨论这类模型，但在后续估计总体谱分布时，需要把这部分离群特征值删去再进行拟合。图2中的黄线代表 $p/20$ 的分位点，其右侧基本涵盖了所有离群特征值。

接下来，我们利用第3.2节中的方法对总体谱分布进行估计。所选取参数分布的概率密度为反立方密度（inverse cubic density）

$$h(t \mid \alpha) = \frac{c}{(t-a)^3}\mathbb{I}(t \geq \alpha), \quad 0 \leq \alpha < 1,$$

其中 $c = 2(1-\alpha)^2, a = 2\alpha - 1$。当 $\alpha \to 1$ 时 PSD 为 $H = \delta_1$，对应的 LSD 为 M-P 分布。我们用最小二乘法（见第3.2.2节）拟合得到 $\hat{\alpha}$。其中，(8)右边的积分项可由引理A.1得到。

实验发现，所删去的离群特征值个数对 $\hat{\alpha}$ 的取值有较大影响。以 $n = 500$ 和 $p = 50, 200, 400$ 为例，删去个数与 $\hat{\alpha}$ 的关系如图3。取删去个数为 $\lfloor p/20 \rfloor$，分别得到 $p = 50, 200, 400$ 时的估计

$$\hat{\alpha}_{50} = 0.78, \quad \hat{\alpha}_{200} = 0.50, \quad \hat{\alpha}_{400} = 0.32.$$

进一步地，我们将 $p = 400$ 时样本相关系数的谱分布（ESD）与估计得到的总体谱分布（PSD）$h(t \mid \hat{\alpha}_{400})$ 所对应的极限谱分布（LSD）、M-P 分布进行对比，如图4。图中 ESD 曲线利用高斯核密度估计（KDE）得到，取带宽为 0.04；$h(t \mid \hat{\alpha}_{400})$ 对应的 LSD 曲线通过数值方法解出，具体步骤可参见 [Yao et al., 2015] 第 2.4.2 节。可以发现，经拟合的 LSD（橙线）比 M-P 分布（蓝线）更接近 ESD（绿线），因而能更好地刻画成分股间相关系数的结构。

将 $p = 50, 200, 400$ 时样本相关系数的谱分布（ESD）和估计得到的总体谱分布（PSD，也即 $h(t \mid \hat{\alpha})$ 的值）进行对比，如图5。可以发现，当 $p \sim n$ 时（图中示例为 $p/n = 0.1, 0.4, 0.8$），$\mathbf{S}$ 与 $\mathbf{\Sigma}$ 对应的相关系数矩阵在极限意义下的谱分布相差较远，这也充分说明了进行谱修正的必要性。

4.2 Markowitz 模型问题的最优估计

下面我们研究 Markowitz 模型问题(1)中的最优值 $R, \mathbf{c}$ 的估计。以下恒取 $\sigma_0 = 1$。

首先给定 $n = 500$，取 $p = 10, 20, \ldots, 450$ 进行实验。图6展示了 plug-in 估计 $\hat{R}_p$ 与分别重复 $1, 10, 30$ 次做平均后的 bootstrap 估计 $\hat{R}_b$。可以发现，就最优期望收益率 R 而言，bootstrap 估计低于 plug-in 估计，有效地克服了 plug-in 所存在的 over-estimation 问题；同时，增加 bootstrap 重复次数再做平均使得 bootstrap 估计随 p 变化曲线（橙线）更光滑，可以有效降低估计方差。图7展示了 plug-in 估计 $\hat{R}_p$ 与 spectrum-corrected 估计 $\hat{R}_s$，其中后者的 PSD 用反立方密度 $h(t \mid \alpha)$ 进行修正，

且考虑三种情况：$\alpha = 0$，$\alpha = 1$（此时各成分股渐近独立，样本相关系数矩阵对应的 LSD 即为 M-P 分布），以及 $\alpha = \hat{\alpha}_p$（利用第4.1节的方法拟合，删去的离群特征值个数为 $\lfloor p/20 \rfloor$）。事实上，由图3可知，$\alpha = 0$ 对应着不删去大特征值进行拟合的情况，$\alpha = 1$ 对应着删去过多大特征值进行拟合的情况。从图7的结果来看，$\alpha = \hat{\alpha}_p$ 给出了最小的估计，其在数值上也最接近 bootstrap 估计；$\alpha = 0$ 给出的估计曲线（蓝线）波动较大，说明不删去离群特征值会造成估计不准确、方差较大的问题；$\alpha = 1$ 给出的估计介于二者之间。

接下来，我们给定 $n = 500$，取 $p = 50, 100, 200, 400$ 进行 200 次重复实验，对比得到的 $\hat{R}_p, \hat{R}_b, \hat{R}_s$，结果如图8。每一个重复实验中，bootstrap 估计为重采样 30 次取平均的结果，spectrum-corrected 估计为 $\alpha = \hat{\alpha}_p$ 下的估计结果。可以发现，plug-in 估计永远是最大的，spectrum-corrected 估计次之，bootstrap 估计最小。此外，随着 p 逐渐增大，plug-in 估计与另外二者的差距越来越大，即 over-estimation 现象愈发明显。

最后，我们比较三种估计方法得到的最优配置 $\hat{\mathbf{c}}_p, \hat{\mathbf{c}}_b, \hat{\mathbf{c}}_s$ 的权重分布，结果如图9。可以发现，从权重分布的范围来讲，plug-in 估计最大，bootstrap 估计次之，spectrum-corrected 估计最小。此外，当 p 越大时三种估计的权重分布范围也越大。

4.3 不同估计的回测结果

最后，我们利用得到的最优配置估计在历史数据上的回测结果。给定 $n = 500$ 和 $p = 400$，即随机选定 400 支成分股，在每一个回测日利用当日之前（包括或不包括当日）的 500 个交易日的数据构建最优配置估计，并在回测日持有对应的资产组合。选取的回测时长为 2018 年 2 月 7 日及之前的 500 个交易日。我们假设资产无交易成本，并据此计算其间每日的累计收益率。

我们考虑两种设定：样本内回测和样本外回测。在样本内回测中，每个交易日的资产组合配置利用**包括**当日的前 500 个交易日数据得到（注意这是一种“作弊”的回测设定，其在构建当日资产组合时已经利用到了当日的收益率信息；但其有助于与样本外回测进行比较）。最终的累计收益率随时间变化图如图10（上）。可以发现，三种估计的样本内回测结果均表现出色，其中 plug-in 估计给出的最优配置具有最高的累计收益率。

由于三种最优配置估计中都有异常大的权重（如图9），例如，± 40 的权重意味着需要以 40 倍的本金做多或做空某支成分股，这显然不符合实际交易情况，因此我们考虑将三种最优权重投影到单纯形 $\{\mathbf{c}\colon 0 \le \mathbf{c}_j \le 1, \sum_{j=1}^{p} \mathbf{c}_j = 1\}$ 上，得到三种新的资产组合。此外，我们再引入一个“等权”的资产组合，即每支成分股的权重均为 $1/p$。这四种资产组合的累计收益率随时间变化图如图10（中）。可见，经投影

过后的资产组合仍有较好的表现，以 spectrum-corrected 估计对应的资产组合为最佳。考虑到等权组合的累计收益率大致可以刻画 S&P500 指数本身的走势，也即市场的系统性走势，因此将该累计收益率从其他三种资产组合对应的累计收益率中剔除，得到图10（下）。其表明在样本内回测中，三种估计所对应资产组合的表现均优于市场表现，实验结果符合直觉。

下面考虑样本外回测。在样本外回测中，每个交易日的资产组合配置利用**不包括**当日的前 500 个交易日数据得到。实验结果如图11，每幅子图的含义同上。可以发现，三种最优配置估计的样本外回测结果要远差于样本内回测结果。具体而言，图11（上）中每条累计收益率曲线波动均较大，且大部分时间累计收益率均为负。plug-in 估计和 bootstrap 估计在第 80 日左右有剧烈波动，而 spectrum-corrected 估计相对而言更为平缓，波动率较小。

经投影过后的三种资产组合均有较好的表现（如图11（中））。然而在剔除等权组合的累计收益率之后，三种资产组合的累计收益率均为负（如图11（下））。上述现象说明，三种最优配置估计无论是否投影到单纯形上，其实际表现并不出色（至少并不优于市场表现）。为达到更好的效果，可能需要对算法和数据做进一步的改进，例如加入除成分股收益率外的其他因子数据，或直接在优化问题(1)中加入“配置权重 $\mathbf{c}$ 属于单纯形”的约束条件等等。

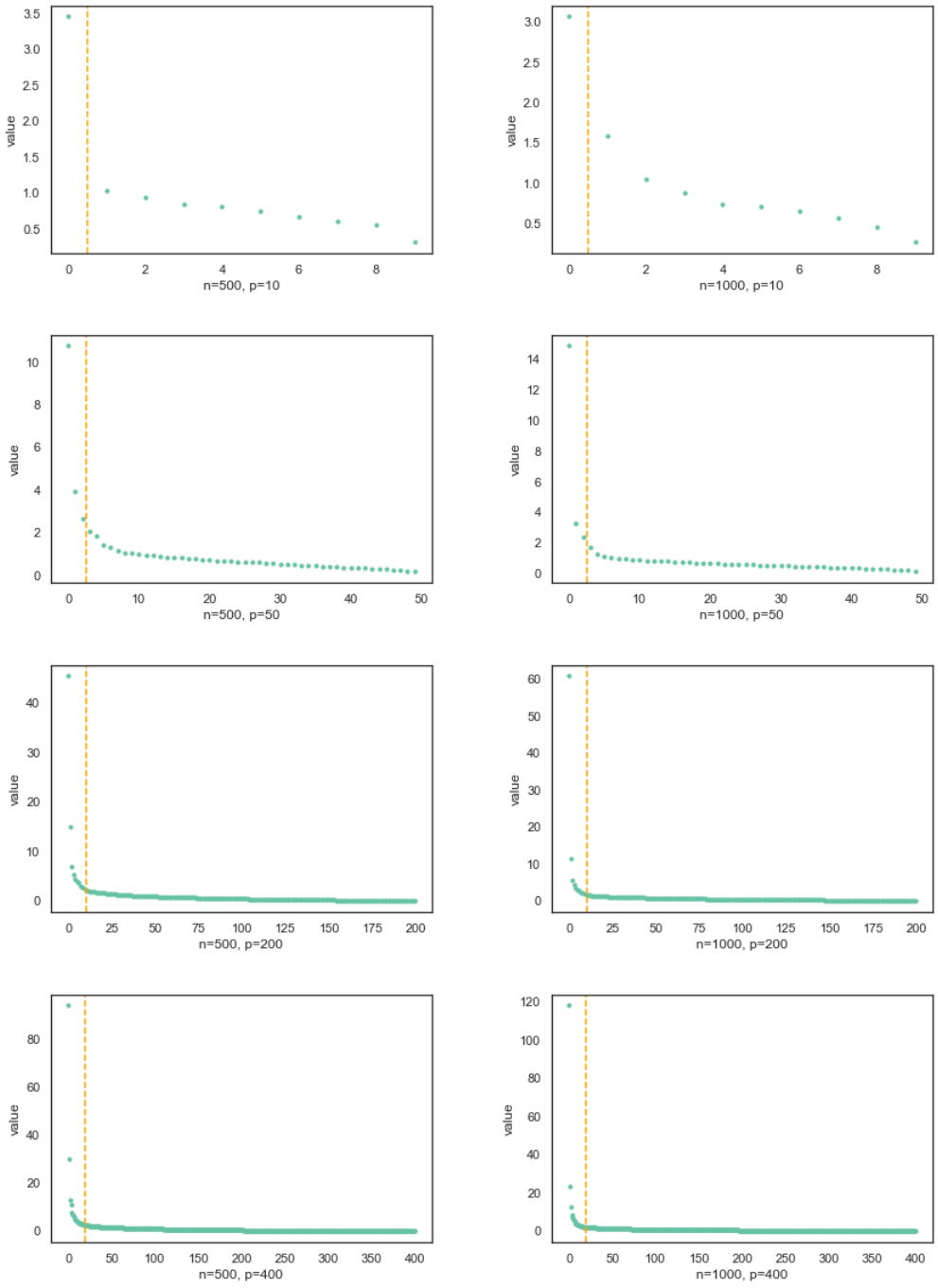

图 2: 特征值图，黄线为 $p/20$ 的分位点

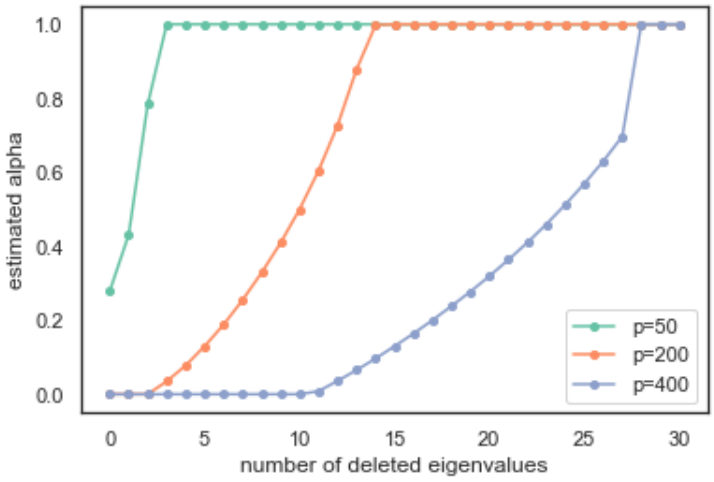

图 3: 删去个数与 $\hat{\alpha}$ 的关系图

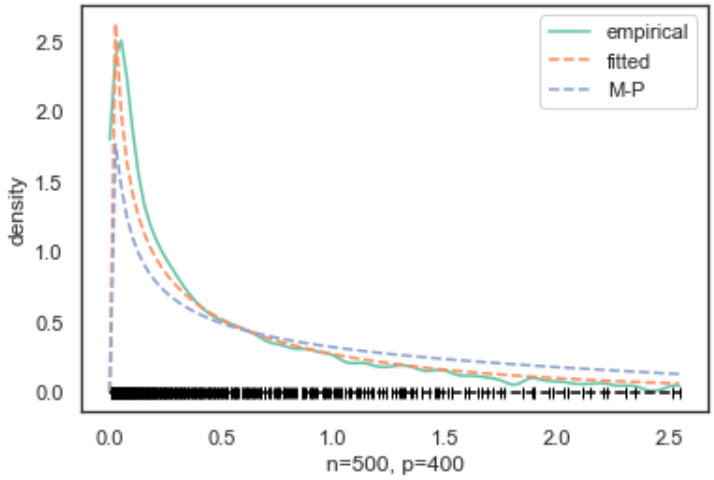

图 4: 样本谱分布 ESD、$h(t \mid \hat{\alpha}_{400})$ 对应的极限谱分布 LSD、M-P 分布

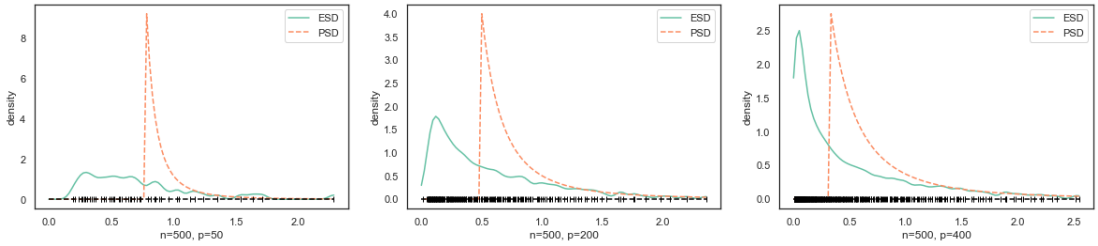

图 5: 样本谱分布 ESD、估计得到的总体谱分布 PSD

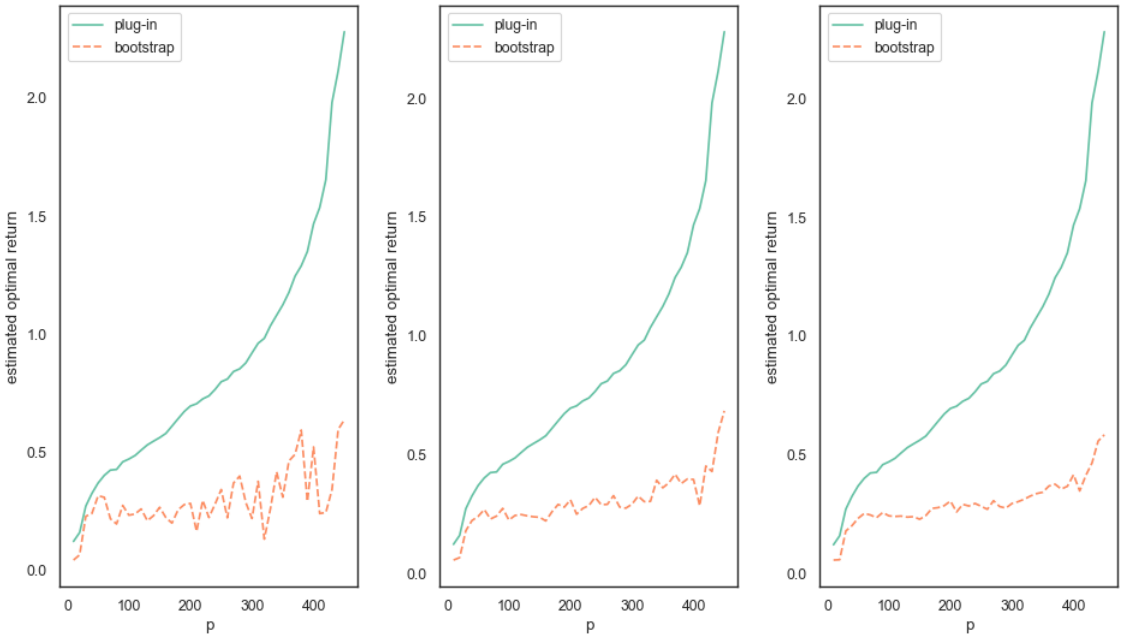

图 6: $\hat{R}_p$ 与 $\hat{R}_b$ 的对比图，从左至右 bootstrap 估计重复次数分别为 $1, 10, 30$

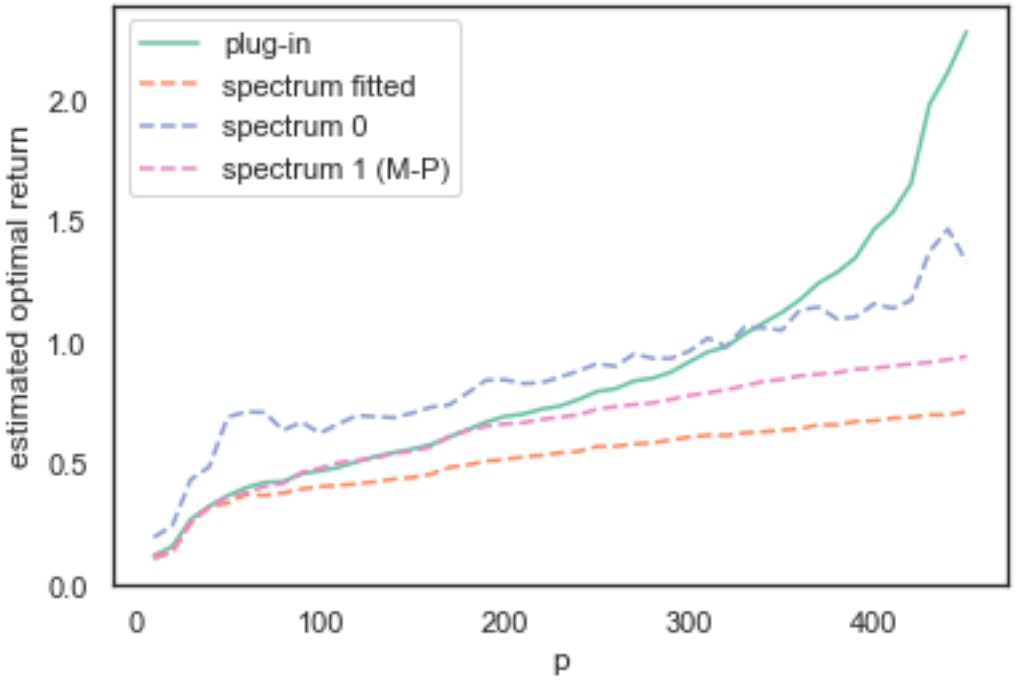

图 7: $\hat{R}_p$ 与 $\hat{R}_s$ 的对比图

图 8: $\hat{R}_p, \hat{R}_b, \hat{R}_s$ 的重复实验对比图，从上至下分别为 $p = 50, 100, 200, 400$ 的情况

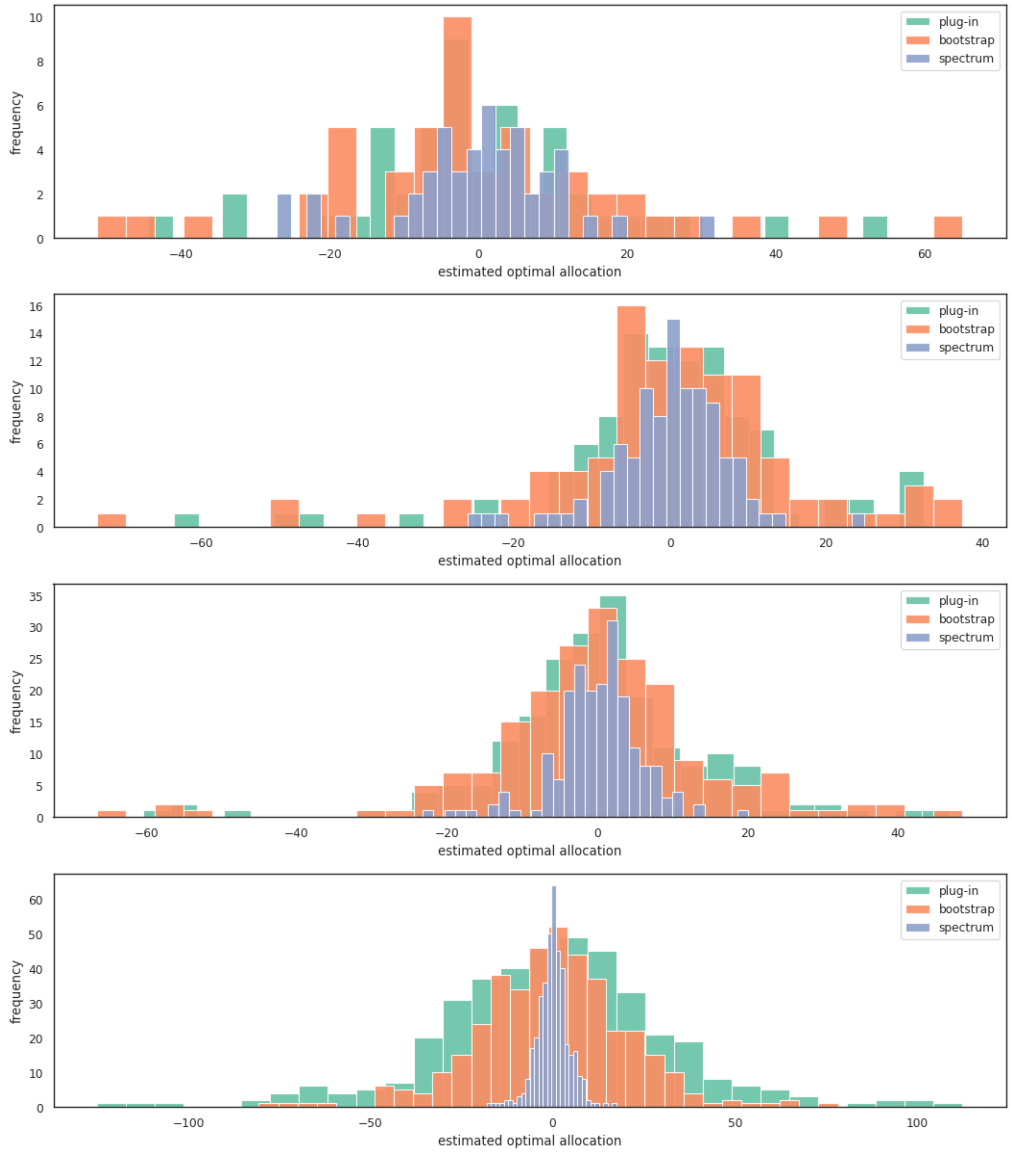

图 9: $\hat{\mathbf{c}}_p, \hat{\mathbf{c}}_b, \hat{\mathbf{c}}_s$ 的权重分布对比图，从上至下分别为 $p = 50, 100, 200, 400$ 的情况

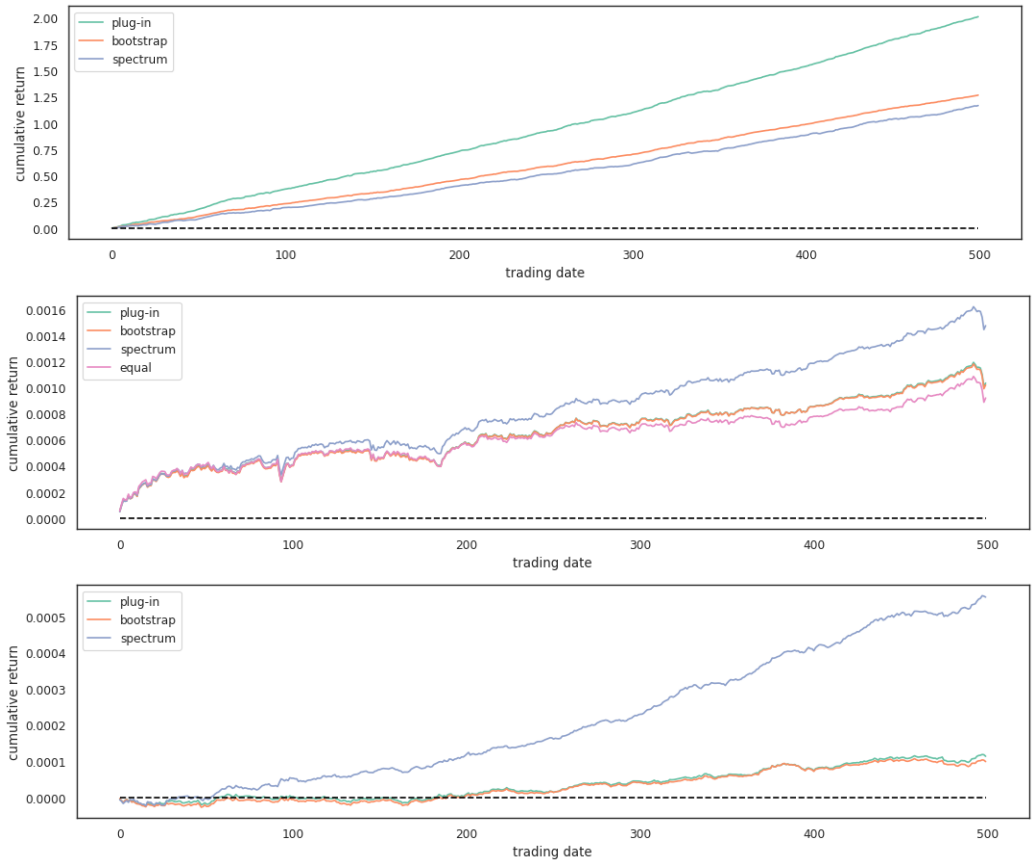
plug-in
bootstrap
spectrum
cumulative return
trading date
plug-in
bootstrap
spectrum
equal
cumulative return
trading date
plug-in
bootstrap
spectrum
cumulative return
trading date

图 10: 样本内回测结果

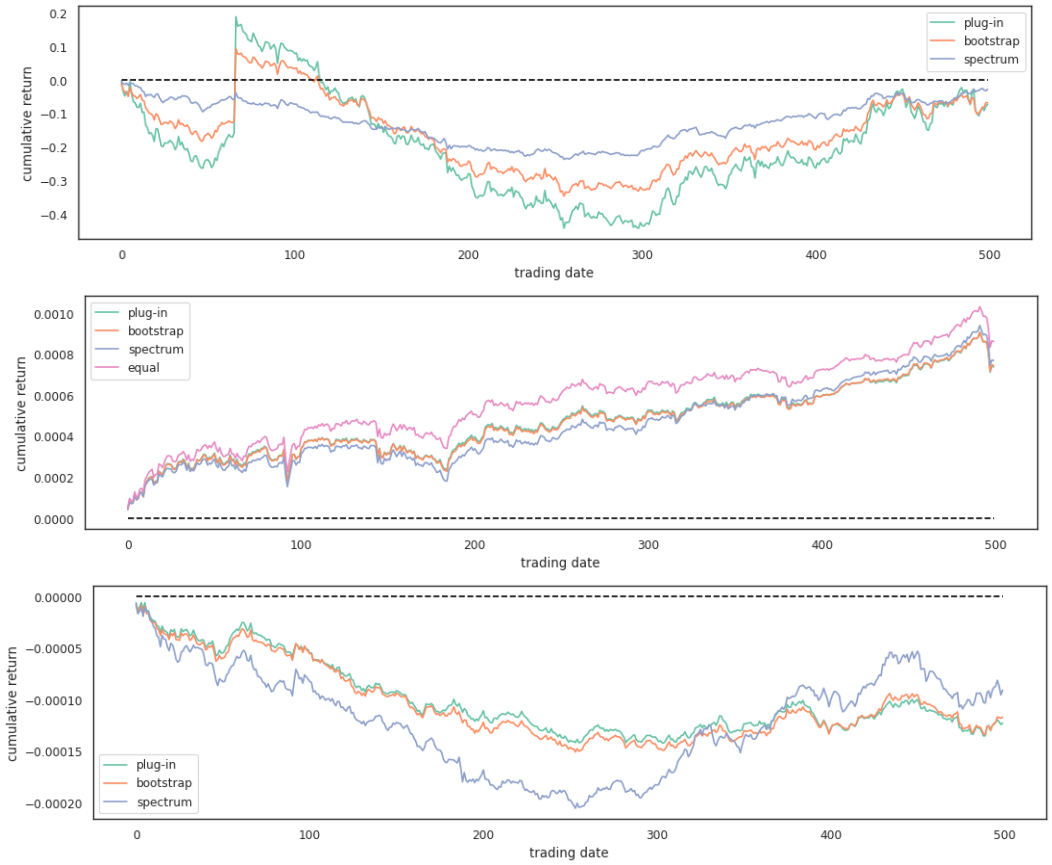
plug-in
bootstrap
spectrum
cumulative return
trading date
plug-in
bootstrap
spectrum
equal
cumulative return
trading date
plug-in
bootstrap
spectrum
cumulative return
trading date

图 11: 样本外回测结果

A 附录一：反立方密度的积分结论

考虑反立方密度

$$h(t \mid \alpha) = \frac{c}{(t-a)^3}\mathbb{I}(t \geq \alpha), \quad 0 \leq \alpha < 1,$$

其中 $c = 2(1-\alpha)^2, a = 2\alpha - 1$。关于 $h(t \mid \alpha)$ 的以下结论成立。

引理 A.1. 对任意 $s > 0$ 或 $s < -1/\alpha$，有

$$\begin{aligned}\int \frac{t}{1+ts} h(t \mid \alpha)\mathrm{d}t &= \int_\alpha^\infty \frac{ct}{(t-a)^3(1+ts)}\mathrm{d}t \\ &= \frac{cs}{(1+sa)^3}\ln\left(\frac{1-\alpha}{\alpha+1/s}\right) + \frac{2(1-\alpha)}{(1+sa)^2} + \frac{a}{1+sa}.\end{aligned}$$

证明.

$$\begin{aligned}\int_\alpha^\infty \frac{ct}{(t-a)^3(1+ts)}\mathrm{d}t &= \int_{1-\alpha}^\infty \frac{c(u+a)}{u^3(1+(u+a)s)}\mathrm{d}u \\ &= \frac{c}{s}\int_{1-\alpha}^\infty \frac{u+a}{u^3(u+a+1/s)}\mathrm{d}u \\ &= \frac{c}{s}\int_{1-\alpha}^\infty \Bigg\{ -\frac{s^2}{(1+sa)^3}\frac{1}{u} + \frac{s}{(1+sa)^2}\frac{1}{u^2} + \frac{sa}{1+sa}\frac{1}{u^3} \\ &\qquad + \frac{s^2}{(1+sa)^3}\frac{1}{u+a+1/s}\Bigg\}\mathrm{d}u \\ &= \frac{c}{s}\left[\frac{s^2}{(1+sa)^3}\ln\left(\frac{1-\alpha}{\alpha+1/s}\right) + \frac{s}{(1+sa)^2}\frac{1}{1-\alpha} + \frac{sa}{1+sa}\frac{1}{2(1-\alpha)^2}\right] \\ &= \frac{cs}{(1+sa)^3}\ln\left(\frac{1-\alpha}{\alpha+1/s}\right) + \frac{2(1-\alpha)}{(1+sa)^2} + \frac{a}{1+sa}.\end{aligned}$$

□

参考文献

[Bai et al., 2013] Bai, Z., Li, H., and Wong, W.-K. (2013). The best estimation for high-dimensional markowitz mean-variance optimization.

[Bai et al., 2009] Bai, Z., Liu, H., and Wong, W.-K. (2009). Enhancement of the applicability of markowitz's portfolio optimization by utilizing random matrix theory. *Mathematical Finance: An International Journal of Mathematics, Statistics and Financial Economics*, 19(4):639–667.

[Yao et al., 2015] Yao, J., Zheng, S., and Bai, Z. (2015). *Sample covariance matrices and high-dimensional data analysis*. Cambridge University Press Cambridge.