

# 基于高频高维协方差矩阵收缩估计的 最小方差投资组合

李瑜<sup>1</sup>, 肖敏<sup>2,3</sup>, 明瑞星<sup>2</sup>

- (1. 浙江工商大学 心理健康教育中心, 浙江 杭州 310018;  
2. 浙江工商大学 统计与数学学院, 浙江 杭州 310018;  
3. 浙江工商大学 统计数据工程技术与应用协同创新中心, 浙江 杭州 310018)

**摘要:** 高频金融数据背景下金融资产收益率序列普遍存在微观噪声结构, 并且存在较为明显的重尾特征; 同时, 金融资产收益率的协方差矩阵具有高维性和稀疏性特征。基于预平均方法和 Huber 损失函数, 采用收缩估计方法, 得到高频金融数据背景下金融资产收益率的协方差矩阵的估计。模拟结果显示收缩估计方法有着较好的效果。此外, 以中国 A 股市场资产的高频数据为样本进行实证分析, 探究估计量在最小方差投资组合上的投资绩效。分析的结果显示: (1) 预平均方法可以剔除绝大部分微观结构噪声对协方差矩阵估计的影响; Huber 损失函数也可以减弱重尾现象对协方差矩阵估计的影响; (2) 收缩估计量均能更好地估计总体协方差矩阵, 并且在最小方差投资策略的比较中也拥有良好的投资绩效。

**关键词:** 高频数据; 高维; 微观结构噪声; 重尾; 最小投资组合

**中图分类号:** F830.59 **文献标志码:** A **文章编号:** 1000-2154(2023)03-0094-15

**DOI:** 10.14134/j.cnki.cn33-1336/f.2023.03.007

## Shrinkage Estimation of High Frequency and High Dimensional Covariance Matrix and Its Application in Minimum Variance Portfolio

LI Yu<sup>1</sup>, XIAO Min<sup>2,3</sup>, MING Ruixing<sup>2</sup>

- (1. Mental Health Education Center, Zhejiang Gongshang University, Hangzhou 310018, China;  
2. School of Statistics and Mathematics, Zhejiang Gongshang University, Hangzhou 310018, China;  
3. Collaborative Innovation Center of Statistical Data Engineering Technology & Application,  
Zhejiang Gongshang University, Hangzhou 310018, China)

**Abstract:** Microstructure noise and heavy tail phenomena are very common in financial asset return series under the background of high-frequency trading. At the same time, the covariance matrix of financial asset returns is characterized by high dimensionality and sparsity. Based on the pre average method with Huber loss function, a shrinkage estimation of the covariance matrix of financial asset returns in the context of high-frequency financial data is proposed. The simulation results show that the shrinkage esti-

收稿日期: 2022-10-08

**基金项目:** 浙江省社会科学规划课题“智媒体下 00 后大学生参与网络集群行为心理机制及对策研究”(22GXSZ001Z); 国家自然科学基金项目“基于有序数据的统计分析及其应用研究”(11971433); 浙江省教育厅一般项目“高频高维协方差矩阵的估计及其应用”(Y202045232); 浙江工商大学“数字+”学科建设管理项目“数据资产: 经济理论、价值核算、市场交易与政策创新(SZJ2022B004)”; 浙江省统计科学研究基地项目“高维情形下最小方差投资组合研究”(22TJD02)

**作者简介:** 李瑜, 女, 讲师, 心理学博士, 主要从事心理统计与测量研究; 肖敏(通讯作者), 女, 讲师, 经济学博士, 主要从事金融统计研究; 明瑞星, 男, 教授, 理学博士, 主要从事风险管理和数理统计研究。

mation performs well. Finally, high-frequency data of Chinese A-share stock market is applied for empirical analysis to explore the investment performance of the estimator on the minimum variance portfolio. The results show that: (1) The pre average method can eliminate the influence of most microstructure noises on covariance matrix estimation; Huber loss function can also weaken the influence of heavy tail on covariance matrix estimation; (2) Shrinkage estimators can better estimate the overall covariance matrix, and have the better performance in the comparison of minimum variance investment strategies.

**Key words:** high-frequency data; high dimension; microstructure noise; heavy tail; minimum variance portfolio (MVP)

## 一、引言

在 Markowitz 的均值—方差投资组合模型中,假设资产的预期平均收益率及其协方差均未知,Markowitz 提出分别利用历史数据计算的样本均值和样本协方差进行估计<sup>[1]</sup>。实证研究表明,风险资产的收益率分布具有尖峰厚尾的非正态特征<sup>[2,3]</sup>,用样本均值和样本方差作为平均收益率及其协方差的估计会导致实际风险偏高。此外,研究表明,预期收益率的估计比协方差的估计更困难,其带来的估计误差比协方差估计中的误差对投资组合绩效的影响更大<sup>[4]</sup>。为避免对预期收益率的估计,学者们研究最小方差投资组合(MVP)模型<sup>[5-6]</sup>。MVP 模型克服估计预期收益率的困难,得到收益率的投资组合有效前沿,在实际数据中有很好的应用效果<sup>[7]</sup>。实证分析表明,MVP 比一些基准的投资组合具有更高的收益率和更小的风险<sup>[8-10]</sup>。由于 MVP 模型只需对资产收益率的协方差矩阵进行估计,因此称之为对协方差矩阵或波动率建模。常见的波动率模型有自回归条件异方差(ARCH)模型、广义自回归条件异方差(GARCH)模型和随机波动(SV)模型等<sup>[11-13]</sup>,这些模型成功地捕捉了低频数据的波动性特征。

近年来,随着信息技术的不断革新,数据存储容量的进一步提升,现代社会的金融信息数据量呈现爆炸式增长,出现了以小时、分钟、秒实时存储的资产交易数据,这种类型的数据称为高频数据。投资组合中资产高频数据的出现,使得针对波动率的建模层出不穷。当投资组合中只有一种风险资产时,有学者从早期的波动率模型入手,考虑混频情形下的波动率建模<sup>[14-16]</sup>;有学者另辟蹊径,避开模型的限制,提出已实现波动率的测度方法<sup>[17-19]</sup>。当投资组合中有多种风险资产时,利用高频数据下的已实现协方差矩阵建模<sup>[20-21]</sup>。需要说明的是,一方面,相对低频数据而言,高频数据包含了大量价格信息,增大数据量更利于对协方差矩阵结构的理解以及投资组合优化的调整;另一方面,在使用高频数据时,由于市场价格买卖价差、交易的非同步性等原因会造成微观结构噪声的积累,会导致已实现协方差矩阵的估计效果不理想。去噪是解决微观结构噪声的常用办法,例如调整采样频率、子抽样法、已实现核估计和预平均法等<sup>[17-19,22-24]</sup>。另外,小波分析在各大领域也被证明为一种有效的去噪方法,且被广泛应用于金融时间序列数据中<sup>[25-28]</sup>。这些方法中,调整采样频率方法操作简单,但会造成大量信息损失。子抽样法通过将观测时间分成多个子时间段,然后求每个时间段的已实现波动率,最后对所有时间段的已实现波动率求平均作为原始时间已实现波动率的估计量,实现在没有缺失数据信息的前提下降低噪声的影响。已实现核估计方法是一种非参数的估计方法,通过对最优时间窗口内构造已实现波动率的核密度估计,即对不同时间窗口下的已实现波动率加权求和作为对波动率的估计量。预平均方法采取先平均处理原始数据,再通过调整参数的方法得到已实现协方差矩阵的估计。子抽样法不需要抛弃数据,但是该方法因为算法复杂,计算量巨大而不被广泛应用。已实现核估计依赖于核的选择,而且对于窗宽的选择非常敏感。小波方法属于频域领域常见的去噪方法,在含噪声强度不在同层次上的时间序列而言处理能力较弱。预平均方法简单实用,且相较于子抽样法而言,计算也相对简单。随着计算机的高速发展,预平均方法计算时间长的弊端在现阶段不再是一个重要问题。因此,预平均这种简单实用的降低噪声影响的方法被学者们普遍采用。

随着金融业的发展,可配置资产的种类不断增加,很多投资组合中风险资产的个数非常庞大,即高维投资组合。针对投资组合中同时出现的高频和高维问题,学者们建立了高频环境下的高维最小方差投资组合模型。对于该模型中投资组合风险的估计可以采用直接估计方法和间接估计方法。直接估计方法延续波

动率模型的时间序列建模思路,通过直接构建多元时间序列模型用于预测基于高频日内收益计算的可观测的最小方差投资组合(MVP)权重的高维向量,它避免了对高维协方差矩阵的估计来逃避维数灾难和噪声的影响<sup>[29]</sup>。事实上,该方法适用于短期时间内的权重预测,对于中长期投资而言,该方法会因为模型的限制而出现失误。间接估计方法采用两步估计方法,首先估计高频高维已实现的协方差矩阵,然后将得到的估计代入 MVP 公式<sup>[30-33]</sup>。间接估计方法中最关键的技术是估计高频数据下的高维协方差矩阵。传统的样本协方差矩阵计算的最小方差投资组合容易受参数不确定性的影响,会导致投资组合收益率的不稳定<sup>[34-35]</sup>。因此,寻找高频数据下的高维协方差矩阵的优良估计具有重要的理论与现实意义。

近20年来,高维协方差矩阵的估计是统计学和计量经济学研究的热点问题之一。常见的估计高维协方差矩阵方法有正则化方法、收缩方法和参数化建模等<sup>[36-57]</sup>。正则化方法,如 banding 方法、thresholding 方法和 tapering 方法等,它基于协方差矩阵具有稀疏性的假设,通过减少待估计参数达到避免维数灾难的目的。参数化模型方法是通过建立线性模型,因子模型等统计模型来描述变量之间的相关关系,使得总体协方差矩阵是这些变量对应的协方差矩阵,从而利用统计模型的特点来估计总体协方差矩阵。收缩估计(算术型收缩和几何型收缩)将样本协方差矩阵和目标矩阵的平均值作为总体协方差矩阵的估计,收缩估计的本质是让样本协方差矩阵朝着目标矩阵收缩,这里的目标矩阵可以是经验协方差矩阵,也可以是人为定制的矩阵(人为定制的矩阵具有结构简单、能刻画变量间重要关系等特点)。收缩估计有如下优点。第一,不直接对总体协方差矩阵假设具有某种结构,而是将先验信息集中在目标矩阵中,得到的最终估计具有稳健性。第二,通过选择损失函数可求得最优收缩参数,得到的最终估计具有最小的风险。第三,最终的估计矩阵有着很好的解释<sup>[51,53-55,57]</sup>。

对于高频情形下的高维协方差矩阵估计,学者建议将去噪处理以后的样本协方差矩阵与高维协方差矩阵的估计量相结合来进行估计<sup>[58-61]</sup>。这种处理方法忽略了在高频环境下金融资产收益率分布明显呈现重尾这一现象。事实上,采用 Huber 损失函数可以减弱重尾现象带来的影响<sup>[62-63]</sup>。基于此,本文提出高频数据的高维协方差矩阵的收缩估计量,并将之应用于最小方差投资组合中。

本研究的主要贡献如下。第一,运用预平均方法对金融资产收益率数据的已实现协方差矩阵进行处理,解决高频数据情形下微观结构的噪声影响。第二,采用 Huber 损失函数来缓解高频金融数据存在的重尾影响。第三,将剔除了受噪声和重尾影响的已实现协方差估计量与高维协方差矩阵的收缩估计量结合,得到新的已实现协方差收缩估计量。第四,将新的已实现协方差收缩估计量引入最小方差投资组合中,得到高频数据下高维最优投资组合权重。第五,通过数据模拟和实证分析,采用平均绝对误差的评价标准,比较了基于样本协方差矩阵计算的已实现协方差收缩估计量、基于预平均法计算的已实现协方差收缩估计量和基于 Huber 损失函数的预平均法计算的已实现协方差收缩估计量在协方差矩阵的估计上的效果;通过夏普比、组合收益和组合波动等指标综合比较了已实现协方差收缩估计量在最小方差投资组合绩效上的表现。

## 二、理论与模型

本小节主要是在高维协方差矩阵的收缩估计量的基础上,提出高频高维已实现协方差矩阵的收缩估计量,进一步应用于最小方差投资组合中,提出高频数据的高维最小方差投资组合的最优权重。其中,在估计高频高维已实现协方差矩阵时,需要利用基于 Huber 损失函数的预平均方法来弱化收益率序列存在的微观结构噪声和重尾的影响。

### (一) 基于 Huber 损失函数的预平均方法

考虑证券交易数据,以间隔5分钟为观测时间点,假设资产  $i$  在第  $t$  天第  $k$  个观测点的对数价格为  $Y_i(t_k)$ ,假定观测到的资产的对数价格受微观结构噪声影响,则

$$Y_i(t_k) = X_i(t_k) + \varepsilon_i(t_k) \quad (1)$$

其中  $i = 1, 2, \dots, p, k = 0, 1, 2, \dots, n - 1$ ;  $p$  指资产维度,  $n$  指观测次数,  $\varepsilon_i(t_k)$  表示噪声,  $X_i(t_k)$  则为资产  $i$  真实的对数价格, 假设  $\{X_i(t), t \geq 0\}$  是均值平稳过程。对于资产  $i$ , 假设  $\{\varepsilon_i(t_k), k = 0, 1, 2, \dots, n - 1\}$  是独立同分布的, 且  $E(\varepsilon_i(t_k)) = 0, \text{Var}(\varepsilon_i(t_k)) = \eta_{ii}$ ; 进一步假设  $\varepsilon_i(\cdot)$  与  $X_i(\cdot)$  也相互独立。易知资产  $i$  在第  $t$  天第  $k$  个观测点的对数收益率为

$$Y_i(t_{k+1}) - Y_i(t_k) = X_i(t_{k+1}) - X_i(t_k) + \varepsilon_i(t_{k+1}) - \varepsilon_i(t_k), k = 0, 1, 2, \dots, n - 1 \quad (2)$$

从式(2)知, 微观噪声的方差会影响收益率序列的波动, 观测到的收益率序列的波动并非真实的波动。可以用如下的预平均方法来消除微观噪声结果影响。令

$$Z_i(t_k) = \sum_{l=0}^{k_n-1} g\left(\frac{l}{k_n}\right) \{Y_i(t_{k+l+1}) - Y_i(t_{k+l})\} \quad (3)$$

其中,  $k = 0, 1, \dots, n - k_n, k_n = C_k n^{1/2}$  为带宽参数,  $C_k$  为其中的调整参数, 与  $p$  和  $n$  无关,  $k_n$  通过选择预平均处理后的已实现波动率的渐近方差最小准则得出;  $g(\cdot)$  为权重函数, 满足条件

$$\begin{cases} g(0) = 0, \\ g(1) = 1, \\ \int_0^1 \{g(t)\}^2 dt \geq 0 \end{cases} \quad (4)$$

且分段连续可微, 在本文中, 选择

$$g(x) = \min\{x, 1 - x\} \quad (5)$$

这样,  $Z_i(t_k)$  为剔除噪声影响的数据。虽然  $Z_i(t_k)$  不受噪声方差影响, 但是其存在重尾现象, 从而使得协方差矩阵估计表现差。利用 Huber 函数作为损失函数以减弱重尾现象带来的影响, 通过最小化 Huber 损失函数可得如下的估计量

$$\hat{\theta}_{ija} = \operatorname{argmin}_{\theta} \sum_{k=0}^{n-k_n} l_{a_{ij}} \left[ \frac{n - k_n + 1}{\varphi k_n} Z_i(t_k) Z_j(t_k) - \theta \right] \quad (6)$$

其中,  $i, j = 1, 2, \dots, p, k = 0, 1, 2, \dots, n - k_n, \varphi = \frac{1}{k_n} \sum_{l=0}^{k_n-1} \left\{ g\left(\frac{l}{k_n}\right) \right\}^2, a_{ij}$  为截断参数,  $l_a$  为 Huber 损失函数, 满足

$$l_a = \begin{cases} 2a|x| - a^2, & x > a, \\ x^2, & x \leq a \end{cases} \quad (7)$$

由于该方法得到的估计量  $\hat{\theta}_{ija}$  是有偏估计, 需要对其进行纠偏, 纠偏后的估计量为

$$\hat{\Gamma}_{ij} = \hat{\theta}_{ija} - \frac{n - k_{n+1}}{\varphi k_n} \zeta \hat{\eta}_{ij} \mathbf{1}\{i = j\} \quad (8)$$

其中,

$$\zeta = \sum_{l=0}^{k_n-1} \left\{ g\left(\frac{l+1}{k_n}\right) - g\left(\frac{l}{k_n}\right) \right\}^2 \quad (9)$$

$$\hat{\eta}_{ij} = \operatorname{argmin}_{\theta} \sum_{k=0}^n l_{a_{ij}} \left\{ \frac{[Y_i(t_{k+l+1}) - Y_i(t_{k+l})]^2}{2 - \theta} \right\} \quad (10)$$

$\hat{\eta}_{ij}$  指的是微观结构噪声方差的估计量。这样,  $\hat{\Gamma}$  为剔除了微观结构噪声影响和缓解了受重尾影响的已实现协方差矩阵  $\Gamma$  的估计量。

## (二) 高频高维已实现协方差矩阵的收缩估计量

收缩估计起源于 Stein(1956)<sup>[50]</sup>, 他对样本均值进行收缩得到的估计量 (James-Stein 估计量) 是非容许估计, 它比样本均值具有更小的风险函数。在估计高维协方差矩阵的情形下, 样本协方差矩阵  $S$  是奇异矩阵, Ledoit 和 Wolf(2004)<sup>[53]</sup> 建议采用如下的算术型收缩估计

$$\hat{\Sigma}_A = \lambda T + (1 - \lambda) S \quad (11)$$

其中,  $T$  是某个特定的目标矩阵 (常用的目标矩阵多元单位阵, 样本协方差矩阵的算术平均值作为对



角目标矩阵等),  $\lambda$  为收缩参数(该参数在 Stein 损失函数最小时取得)。算术型收缩估计量具有可解释性及渐近相合性等优良统计性质, 并且有很好的预测效果。

在基因微阵列数据分析中, 由于基因类型差异很大, 方差的估计量表现不稳定, Tong 和 Wang(2007)<sup>[55]</sup> 提出采用几何型收缩估计估计协方差矩阵

$$\widehat{\Sigma}_G = T^{-\alpha} S^{1-\alpha} \quad (12)$$

其中  $\alpha$  为收缩参数。理论和模拟证明几何型收缩估计量是一种具有更小风险函数的稳健估计。

受到高维协方差矩阵的收缩估计启示, 本文提出高频高维已实现协方差矩阵的几何型收缩估计量和算术型收缩估计量。注意到高频协方差矩阵的估计矩阵  $\widehat{\Gamma}$  不一定是正定矩阵, 这给构建几何型收缩估计量带来困难, 本文采用扰动的技巧得到高频高维已实现协方差矩阵的几何型收缩估计量。

**1. 高频高维已实现协方差矩阵的几何型收缩估计量。**由于高维情形下已实现协方差矩阵不一定满足正定性要求, 在本文中构建高频高维已实现协方差矩阵的收缩估计量如下。

$$\widehat{\Sigma}_G = (T + \delta I)^\alpha (\widehat{\Gamma} + \delta I)^{1-\alpha} - \delta I \quad (13)$$

其中,  $T = \text{diag}(\Pi_{i=1}^p \widehat{\Gamma}_{ii})^{1/p}$  是目标矩阵,  $\widehat{\Gamma}$  为样本已实现协方差,  $I$  为单位矩阵,  $\delta$  为调整参数, 其作用是使已实现协方差矩阵满足正定性的要求,  $\alpha$  为收缩参数。选择 Log-Euclidean 损失函数作为确定最优参数的损失函数

$$L = \|\log(\widehat{\Sigma}_G + \delta I) - \log(\Sigma + \delta I)\|_F^2 = \|\log((T + \delta I)^\alpha (\widehat{\Gamma} + \delta I)^{1-\alpha}) - \log(\Sigma + \delta I)\|_F^2 \quad (14)$$

在该损失函数下, 平均风险为

$$\begin{aligned} R(\alpha) &= \sum_{i=1}^p E \|\log(\widehat{\Sigma}_G + \delta I) - \log(\Sigma + \delta I)\|_F^2 \\ &= \sum_{i=1}^p E \{ u_i [\alpha \log(\lambda_{T_i} + \delta) + (1 - \alpha) \log(\lambda_{\sigma_i} + \delta)] u_i^T - v_i \log(\lambda_i + \delta) v_i^T \}^2 \end{aligned} \quad (15)$$

其中,  $\lambda_{T_i}$ ,  $\lambda_{\sigma_i}$  和  $\lambda_i$ ,  $i = 1, 2, \dots, p$  分别是目标矩阵  $T$ , 已实现协方差矩阵  $\widehat{\Gamma}$  和总体协方差矩阵  $\Sigma$  的特征根,  $u_i$  和  $v_i$  分别是  $\lambda_{\sigma_i}$  和  $\lambda_i$  对应的特征向量。根据上述损失函数, 可得最优收缩参数为

$$\alpha^* = \max\{0, \min\{1, \widehat{\alpha}\}\} \quad (16)$$

其中

$$\widehat{\alpha} = \frac{\sum_{i=1}^p E [u_i \log(\lambda_{T_i} + \delta) u_i^T - u_i \log(\lambda_{\sigma_i} + \delta) u_i^T] [v_i \log(\lambda_i + \delta) v_i^T - u_i \log(\lambda_{\sigma_i} + \delta) u_i^T]}{\sum_{i=1}^p E [u_i \log(\lambda_{T_i} + \delta) u_i^T - u_i \log(\lambda_{\sigma_i} + \delta) u_i^T]^2} \quad (17)$$

**2. 高频高维已实现协方差矩阵的算术型收缩估计量。**构建高频高维已实现协方差矩阵的算术收缩估计量如下。

$$\widehat{\Sigma}_A = \lambda T + (1 - \lambda) \widehat{\Gamma} \quad (18)$$

即

$$\widehat{\sigma}_{a,ij} = \lambda \times t_{ij} + (1 - \lambda) \widehat{\sigma}_{ij} \quad (19)$$

其中,  $T = \text{diag}(\sum_{i=1}^p \widehat{\Gamma}_{ii}/p)$  是目标矩阵,  $\widehat{\Gamma}$  为样本已实现协方差,  $\lambda$  为收缩参数。选择 Stein 损失函数作为确定最优参数的损失函数

$$L = \|\widehat{\Sigma}_A - \Sigma\|_F^2 = \|\lambda T + (1 - \lambda) \widehat{\Gamma} - \Sigma\|_F^2 \quad (20)$$

在该损失函数下, 平均风险为

$$R(\lambda) = E \left\{ \sum_{i=1}^p [\lambda * t_{ij} + (1 - \lambda) \widehat{\sigma}_{ij} - \sigma_{ij}]^2 \right\} \quad (21)$$

得最优参数为

$$\lambda^* = \max\{0, \min\{1, \widehat{\lambda}\}\} \quad (22)$$

其中,

$$\tilde{\lambda} = \frac{\sum_{i=1}^p \sum_{j=1}^p [\text{var}(\hat{\sigma}_{ij}) - \text{cov}(t_{ij} - \hat{\sigma}_{ij}) - \text{Bias}(\hat{\sigma}_{ij}) \cdot E(t_{ij} - \hat{\sigma}_{ij})]}{\sum_{i=1}^p \sum_{j=1}^p E[(t_{ij} - \hat{\sigma}_{ij})^2]} \quad (23)$$

在实证中,由于式(23)当中仍有未知参数,将未知参数的矩估计量代入式(23),得

$$\hat{\lambda} = \frac{\sum_{i=1}^p \sum_{j=1}^p [\widehat{\text{var}}(\hat{\sigma}_{ij}) - \widehat{\text{cov}}(t_{ij} - \hat{\sigma}_{ij}) - \text{Bias}(\hat{\sigma}_{ij}) \cdot E(t_{ij} - \hat{\sigma}_{ij})]}{\sum_{i=1}^p \sum_{j=1}^p [(t_{ij} - \hat{\sigma}_{ij})^2]} \quad (24)$$

### (三) 最小方差投资组合

考虑高频数据的高维最小方差投资组合。假设该投资组合有  $p$  个证券资产,其收益率序列为  $R = (R_1, R_2, \dots, R_p)^T$ ,协方差矩阵为  $\Sigma$ 。最小方差投资组合的目标是最小化如下目标函数:

$$\text{argmin}_w w^T \Sigma w \quad (25)$$

$w$  满足约束条件

$$w^T \mathbf{1} = 1 \quad (26)$$

其中, $w = (w_1, w_2, \dots, w_p)^T$  代表投资在各个资产上的权重, $\mathbf{1} = (1, 1, \dots, 1)^T$  表示  $p$  维全1向量。上述优化问题的最优解为

$$w_{opt} = \frac{\Sigma^{-1} \mathbf{1}}{\mathbf{1}^T \Sigma^{-1} \mathbf{1}} \quad (27)$$

将高频高维已实现协方差矩阵的收缩估计量代入式(27),可得高频数据的高维最小方差投资组合的最优权重分别为

$$\hat{w}_G = \frac{\hat{\Sigma}_G^{-1} \mathbf{1}}{\mathbf{1}^T \hat{\Sigma}_G^{-1} \mathbf{1}} \quad (28)$$

和

$$\hat{w}_A = \frac{\hat{\Sigma}_A^{-1} \mathbf{1}}{\mathbf{1}^T \hat{\Sigma}_A^{-1} \mathbf{1}} \quad (29)$$

## 三、模拟分析

本节采用模拟分析来评价本文提出的高频高维的协方差矩阵的收缩估计量的表现。比较样本已实现协方差矩阵、预平均已实现协方差矩阵、经过 Huber 损失函数处理的预平均已实现协方差矩阵、已实现协方差矩阵几何型收缩估计和已实现协方差矩阵算术型收缩估计在估计总体协方差矩阵时的表现。

### (一) 数据生成过程

建立  $p$  个资产对数价格的随机模型:

$$dX_t = \mu_t dt + \sigma_t dW_t \quad (30)$$

其中, $i = 1, 2, \dots, p, \mu_t = (\mu_{1t}, \mu_{2t}, \dots, \mu_{pt})^T$  为漂移向量,表示资产对数价格随时间变化的期望值, $W_t = (W_{1t}, W_{2t}, \dots, W_{pt})^T$  为独立的布朗运动, $\sigma_t$  为波动率,为  $p$  维向量。用  $\varepsilon_t = (\varepsilon_{1t}, \varepsilon_{2t}, \dots, \varepsilon_{pt})^T$  表示噪声, $\varepsilon_t$  满足独立同分布的正态分布,均值为0,取方差为0.01<sup>2</sup>。在生成的资产对数价格上加上独立同分布的微弱噪声结构,即模拟生成了现实资产交易过程中能观测到的价格信息。

### (二) 参数设置

在模拟过程中,生成对数价格的过程中的参数随机生成,需要设定的参数主要是维数  $p$  和样本容量  $n$ ,以及调整参数  $\delta$ 。涉及 Huber 损失函数处理重尾影响时所需计算时长情况,本文选择最大资产维数为500维,样本容量最大为300。 $p$  依次取100,200,300,500, $n$  依次取50,100,150,200,300,分别搭配成高维情

形,不考虑  $n > p$  的情形。调整参数  $\delta$  的主要作用在于调整待估协方差矩阵的非奇异现象,因此,为了减小对最终的估计结果的影响,取调整参数  $\delta = 0.0001$ 。

### (三) 评价标准

本小节的主要目的是度量各种估计量在估计总体协方差矩阵的表现,首先评价的是预平均方法是否能够有效地消除噪声对已实现协方差矩阵的影响的估计,其次评价 Huber 损失函数是否能够有效地减弱部分重尾现象,最后评价在上述基础上提出的高频高维的协方差矩阵的收缩估计量的表现。常用的评价标准有均方误差(MSE),平均绝对误差(MAE)等评价指标,MSE 是参数估计值与真实值之差平方的期望值,MAE 是绝对误差的平均值,相比较而言 MSE 会对高差异惩罚更多,而 MAE 能更好地反映误差的实际情况。因此,本文选择平均绝对误差作为评价标准,其定义为

$$MAE_m = E|\hat{\Sigma} - \Sigma| \quad (31)$$

其中, $\hat{\Sigma}$ 表示总体协方差阵的估计量, $m$ 对应不同的估计量, $\Sigma$ 表示真实的协方差矩阵,在模拟过程中,缺乏真实的协方差矩阵,本文采用 bootstrap 抽样,产生真实的协方差矩阵。

### (四) 模拟分析结果

记直接计算的已实现协方差矩阵为  $S$ ,经过预平均处理后得到的已实现协方差矩阵估计量为  $\hat{\Sigma}_2$ ,经过预平均和 Huber 损失函数处理后得到的已实现协方差矩阵估计量为  $\hat{\Sigma}_3$ ,以  $\hat{\Sigma}_3$  对角线上元素几何平均值作为目标矩阵,进而获得的高频高维已实现协方差矩阵的几何型收缩估计量为  $\hat{\Sigma}_c$ ,以  $\hat{\Sigma}_3$  对角线上元素算术平均值作为目标矩阵,进而获得的高频高维已实现协方差矩阵的算术型收缩估计量为  $\hat{\Sigma}_A$ 。MAE<sub>1</sub>、MAE<sub>2</sub>、MAE<sub>3</sub>、MAE<sub>4</sub>、MAE<sub>5</sub> 分别依次描述这五种估计量的平均绝对误差。具体的对应关系如表1所示。

表1 各个估计量及其记号

估计量	$S$	$\hat{\Sigma}_2$	$\hat{\Sigma}_3$	$\hat{\Sigma}_c$	$\hat{\Sigma}_A$
平均绝对误差	MAE <sub>1</sub>	MAE <sub>2</sub>	MAE <sub>3</sub>	MAE <sub>4</sub>	MAE <sub>5</sub>

根据模拟结果,得到不同维度与不同样本容量组合下的平均绝对误差,如表2所示。

从表2可知,不同维度与不同样本容量组合下,都有  $MAE_4 < MAE_5 < MAE_3 < MAE_2 < MAE_1$ ,其中 MAE<sub>4</sub> 的值始终是最小的,这意味着在高频数据环境下,高频高维已实现协方差矩阵的几何型收缩估计量在 MAE 准则下表现最优。其次,样本协方差矩阵  $S$  表现最差,说明在高维情况下需要改变传统的估计方法是必要的。此外,通过表2还容易得出预平均处理确实可以剔除噪声的影响,经过预平均处理后得到的已实现协方差矩阵估计量的平均绝对误差,比直接计算的已实现协方差矩阵得到的平均绝对误差都要小。经过预平均和 Huber 损失函数处理后得到的已实现协方差矩阵估计量在  $\hat{\Sigma}_2$  的基础上更加靠近真实的协方差矩阵。

表2 不同样本容量与资产维数的平均绝对误差

$(n, p)$	MAE <sub>1</sub>	MAE <sub>2</sub>	MAE <sub>3</sub>	MAE <sub>4</sub>	MAE <sub>5</sub>
(50, 100)	0.4369	0.2235	0.1921	0.1316	0.1831
(50, 200)	0.4892	0.2578	0.2165	0.1598	0.2057
(50, 300)	1.4657	0.7795	0.7345	0.6054	0.7214
(50, 500)	6.8641	3.5812	3.5017	3.4273	3.4816
(100, 200)	0.9899	0.5326	0.5107	0.4268	0.5004
(100, 300)	1.2674	0.6853	0.6277	0.5538	0.6135
(100, 500)	6.9378	3.8745	3.6923	3.4704	3.6773
(200, 300)	3.7952	1.8714	1.8472	1.6059	1.8047
(200, 500)	7.5927	3.9316	3.7253	3.6078	3.7974
(300, 500)	7.8865	4.6783	4.5804	4.4516	4.5497

高频高维已实现协方差矩阵的算术型收缩估计量虽然不及几何型收缩估计量,但仍然是五个估计量中表现相对优异的。纵向看表2,如果固定样本容量,变化资产维数,那么无论是何种估计量的平均绝对误差,都会随着维数的增加而增加,具体的走势展示在图1中。为了展示各个估计量的平均绝对误差的差异性,图1的纵坐标采用了底数为2的对数处理。 $MAE_1$ 明显地处于图1的最高位置,该结果进一步说明,对于高频高维的数据而言,若直接采用样本协方差矩阵作为总体协方差矩阵的估计是不合理的。 $MAE_4$ 则位于图形的最底端,即高频高维的几何收缩估计量效果最佳。如果固定资产维数,变化样本容量,那么平均绝对误差也会随着样本容量的增加而增加,与图1呈现同样的走势和形态,在此不再赘述。

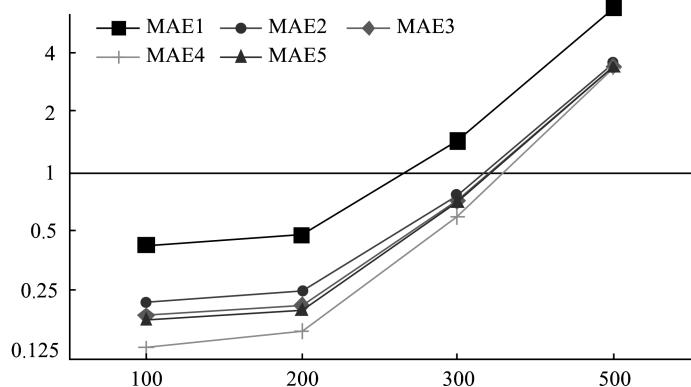


图1 样本容量50情形下样本维数变化下的平均绝对误差比较

## 四、实证分析

本小节主要是实证分析高维高频协方差的几何型收缩估计量与高维高频协方差的算术型收缩估计量在最小方差投资组合中的表现。具体过程为选取实证数据,并进行初步的资产选择,最后对比这两个估计量在最小方差投资组合中的表现。

### (一) 数据筛选与描述

研究对象为中国A股市场的股票,研究周期分别选择新冠疫情发生前后两个周期,为2021年9月6日至2022年8月22日和2017年1月3日至2017年11月13日,数据来源于同花顺高频交易数据库。风险厌恶型投资者选择风险较小的稳健型股票进行投资,抑或选择债券进行投资,这种情形下的投资组合的协方差矩阵可能是非正定矩阵,本文不考虑这种情况,只考虑投资组合的协方差矩阵是正定矩阵的情形。当投资组合的协方差矩阵为正定矩阵时,此时投资者被视为风险偏好型投资者。对于风险偏好型投资者,往往会选择具有超额收益率的股票投资组合和风险较大的股票投资组合进行投资。

金融市场中各风险资产的收益率受到诸多共同因素影响,通常通过建立因子模型来描述这些因素的影响,著名的Fama-French三因子模型和五因子模型是应用最为广泛的因子模型。Fama-French五因子模型由Fama-French三因子模型发展而来。它综合考虑了系统风险、账面市值比、市值规模因子、盈利因子和投资因子对基金业绩的影响,能够更为全面地评价策略业绩,并且更为有效地衡量策略通过主动投资管理取得超额收益的能力。另外,五因子模型对三因子模型难以解释的组合——小市值—低盈利、小市值—低投资和中等偏大的市值组合具有较强的解释能力,故文章选择了五因子模型作为挑选投资证券的工具。本文先建立股票收益率的五因子模型,然后通过所建立的五因子模型来选择股票的投资组合。由于在不同的时期不同的个股能有效解释收益率的因子不同(衡量因子的有效性指标称为信息比),某些因子的有效性会随着时间的推移或者因子种类的转化而递减,但是如果将所有能解释股票收益率的因子纳入模型,通过计算得出模型的常数项为正的股票,那么该股票具有超额收益率。据此,将各股票在因子模型中计算出来



的常数项按照从大到小的次序进行排列就可以选出具有超额收益率的股票。考虑经典的 Fama-French 五因子模型

$$R_t = \alpha + a_1 MAR_t + a_2 SMB_t + a_3 HML_t + a_4 RMW + a_5 CMA + \epsilon_t \quad (32)$$

五个因子 *MAR*、*SMB*、*HML*、*RMW* 和 *CMA* 分别对应市场因子、规模因子、账面市值比因子、盈利因子和投资因子。Fama-French 五因子数据来自国泰安数据库。剔除交易日期缺失的股票后,使用 Fama-French 五因子模型对剩余支证券前100个交易日的收益率序列进行回归得到相应的 Alpha,对每只股票的 Alpha 进行排序,取出 Alpha 大于0的股票。其后,每隔30个交易日进行一次调仓,挑选出 Alpha 大于0的证券。其滚动过程如图2所示。

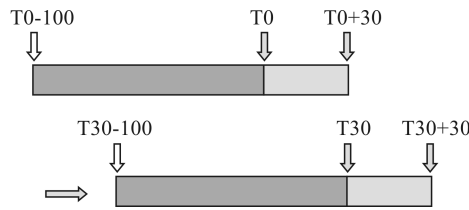


图2 选股周期示意图

将滚动期间有最大可能性获得超额收益的股票挑选出,最终满足条件的股票有33只,具体股票名称和代码如表3所示。为保持一致性,故新冠发生前的周期也选择33只证券。具体股票名称和代码如表4中所示。

表3 新冠疫情发生后周期具有超额收益率的33只股票投资组合的股票代码

000007	000525	000546	000733	000875	000887	002003
002025	002049	002180	002182	002240	002459	002484
002549	002779	002840	002850	002865	300073	300260
300416	300438	300498	300571	300645	300769	600011
600188	600233	600563	600571	600765		

表4 新冠疫情发生前周期具有超额收益率的33只股票投资组合的股票代码

002805	000898	600179	600231	002497	000778	600782
300059	300214	002016	600192	000705	300157	600256
000875	000683	600549	300182	000852	600808	600396
000825	000951	002422	600717	000761	000088	300249
002645	600790	300003	002452	300377		

以002049为例,绘制该股票在选择时间段的收益率时序图,图3给出了该股票在这210天每五分钟的收益率时序图。从图中可以看出收益率的波动性受到噪声的影响,呈现波动性。

用方差作为风险度量,秉承“高风险高收益”原则,按照回报率的方差大小排序,为与第一种情形保持一致,故也挑选出方差最大的33只证券作为投资对象,具体的股票名称和代码如表5所示。新冠疫情发生前的周期挑选的证券具体的股票名称和代码如表6所示。

表5 新冠疫情发生后周期风险最大的33只股票投资组合的股票代码

300199	300432	300343	300584	300595	300619	300642
300677	300681	300745	300769	300772	300809	300969
600753	000056	000537	000422	000432	000812	000829
000838	002006	002326	002349	002374	002453	002487
002529	002607	002865	300052	300077		

表6 新冠疫情发生前周期风险最大的33支股票投资组合的股票代码

300531	002131	300562	300534	300304	002620	300194
300379	002783	002649	300555	300546	300502	300352
300279	300522	002354	300403	300062	300515	300519
300569	002793	300487	002281	300539	300471	002169
002703	300493	300382	300581	002429		

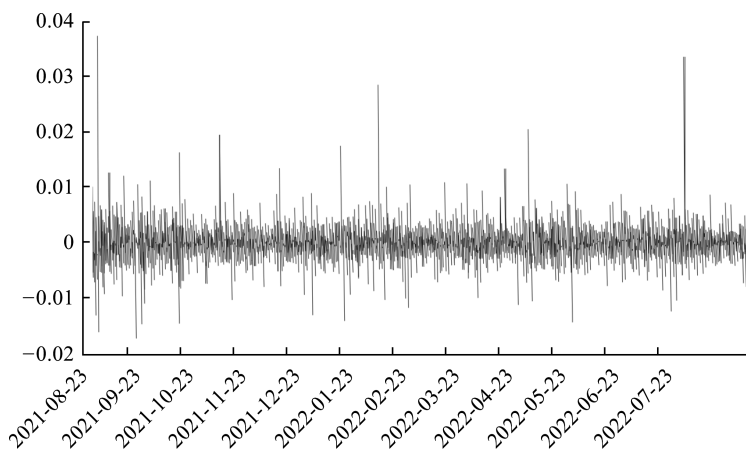


图3 股票002049收益率时间序列图

以300595为例,绘制该股票在选择时间段的收益率时序图,图4给出该股票在这210天每五分钟的收益率时序图。从图中可以很直观地看出收益率走势基本在0附近,并且在一段时间内表现急剧的波动,在接下来一段时间又相对平稳,这也说明了该收益率序列存在一定的噪声影响。

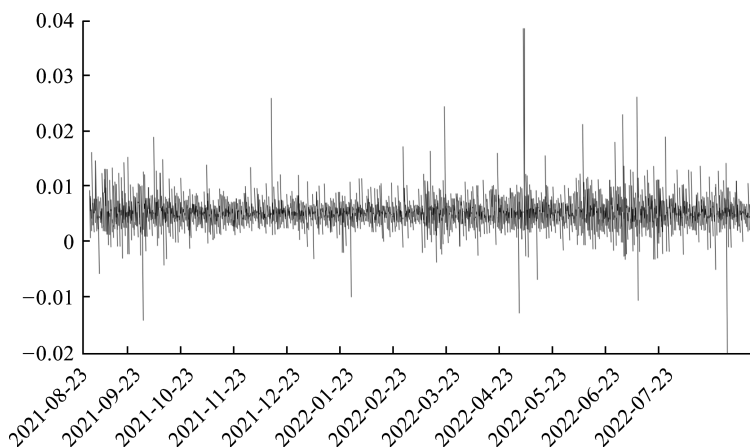


图4 股票300595收益率时间序列图

## (二) 最小投资组合收益比较

对挑选出两个周期两种情形下的33只股票进行投资,以5分钟为一次观测点,每个交易日的观测次数为48次,总交易日期为210天,共10080个数据,每隔30个交易日,进行一次投资优化调整。

在最小方差投资组合模型中,需要估计总体协方差矩阵,在本文中,选择四种不同类型的协方差矩阵的估计量,构成四种不同的投资策略,从组合收益、组合波动和夏普比率等角度比较了这四种策略的投资情况。另外为说明投资策略的稳健性,特从中选择最有可能获得超额收益的前30只证券和前20只证券,以及方差最大的前30只证券和前20只证券也做同样的策略选择。

沿用上一小节的记号,记直接计算的已实现协方差矩阵为  $S$ ,基于  $S$  计算的权重  $\hat{w}_s$ ;经过预平均和 Huber 损失函数处理后得到的已实现协方差矩阵估计量为  $\hat{\Sigma}_3$ ,基于估计量  $\hat{\Sigma}_3$  计算的权重  $\hat{w}_h$ ;以  $\hat{\Sigma}_3$  对角线上元素几何平均值作为目标矩阵,进而获得的高频高维已实现协方差矩阵的几何型收缩估计量为  $\hat{\Sigma}_c$ ,基于几何型收缩估计量  $\hat{\Sigma}_c$  计算的投资权重  $\hat{w}_c$ ;以  $\hat{\Sigma}_3$  对角线上元素算术平均值作为目标矩阵,进而获得的高频高维已实现协方差矩阵的算术型收缩估计量为  $\hat{\Sigma}_A$ ,基于算术型收缩估计量  $\hat{\Sigma}_A$  计算的投资权重  $\hat{w}_A$ 。

表7和表8展示了新冠疫情发生前后周期具有超额收益率的股票投资组合的各个投资策略的绩效情况。

表7 新冠疫情发生后周期具有超额收益率的33只股票投资组合的投资策略比较

投资证券个数	投资权重	组合收益	组合波动	夏普比率
n = 33	$\hat{w}_s$	0.000415	0.003208	0.129364
	$\hat{w}_h$	0.000433	0.003057	0.141642
	$\hat{w}_c$	0.000508	0.002843	0.178684
	$\hat{w}_A$	0.000465	0.002953	0.157467
n = 30	$\hat{w}_s$	0.001212	0.061498	0.019708
	$\hat{w}_h$	0.013817	0.051959	0.265921
	$\hat{w}_c$	0.047617	0.044102	1.079702
	$\hat{w}_A$	0.032621	0.057774	0.564631
n = 20	$\hat{w}_s$	0.017250	0.093202	0.185082
	$\hat{w}_h$	0.026353	0.090819	0.290171
	$\hat{w}_c$	0.032812	0.043221	0.759168
	$\hat{w}_A$	0.020553	0.045825	0.448511

表8 新冠疫情发生前周期具有超额收益率的33只股票投资组合的投资策略比较

投资证券个数	投资权重	组合收益	组合波动	夏普比率
n = 33	$\hat{w}_s$	0.022572	0.014129	0.825785
	$\hat{w}_h$	0.023985	0.027334	1.697572
	$\hat{w}_c$	0.035345	0.007914	4.466136
	$\hat{w}_A$	0.025751	0.010588	2.432093
n = 30	$\hat{w}_s$	0.006939	0.056455	0.122912
	$\hat{w}_h$	0.019977	0.011939	1.673256
	$\hat{w}_c$	0.023280	0.006775	3.436162
	$\hat{w}_A$	0.020042	0.006958	2.880425
n = 20	$\hat{w}_s$	0.023546	0.036061	0.652949
	$\hat{w}_h$	0.014509	0.014786	0.981266
	$\hat{w}_c$	0.040245	0.014656	2.745974
	$\hat{w}_A$	0.014538	0.014784	0.983360

通过表7和表8不难发现:基于高维高频协方差矩阵收缩估计量的两类资产组合的组合波动要小于基于样本协方差矩阵以及仅仅处理了噪声和重尾情形的估计量的资产组合。对于基于高维高频协方差矩阵的收缩估计量的两类资产组合中,几何型收缩估计量构造的投资策略表现最优,其对应的组合波动最小,夏普比率最高。该结果也说明,噪声和重尾对高频高维协方差矩阵的估计及构造投资策略的影响不容忽视。

表9和表10展示新冠疫情发生前后两个周期风险最大的33只股票投资组合的各个投资策略的绩效情况。

表9 新冠疫情发生后周期风险最大的33只股票投资组合的投资策略比较

投资证券个数	投资权重	组合收益	组合波动	夏普比率
n = 33	$\hat{w}_s$	0.002247	0.017525	0.128217
	$\hat{w}_h$	0.002615	0.013166	0.198618
	$\hat{w}_G$	0.003028	0.007218	0.419507
	$\hat{w}_A$	0.002686	0.009217	0.291418
n = 30	$\hat{w}_s$	0.012995	0.128978	0.100754
	$\hat{w}_h$	0.019053	0.060186	0.316569
	$\hat{w}_G$	0.144073	0.039241	3.671492
	$\hat{w}_A$	0.019331	0.058666	0.329509
n = 20	$\hat{w}_s$	0.040070	0.380079	0.105425
	$\hat{w}_h$	0.073789	0.274186	0.26912
	$\hat{w}_G$	0.162398	0.165956	0.978561
	$\hat{w}_A$	0.086707	0.256880	0.337539

表10 新冠疫情发生前周期风险最大的33只股票投资组合的投资策略比较

投资证券个数	投资权重	组合收益	组合波动	夏普比率
n = 33	$\hat{w}_s$	0.012411	0.145022	0.085580
	$\hat{w}_h$	0.083547	0.070456	1.185804
	$\hat{w}_G$	0.154950	0.015668	9.889584
	$\hat{w}_A$	0.022645	0.068976	0.328303
n = 30	$\hat{w}_s$	0.010785	0.474198	0.022744
	$\hat{w}_h$	0.131068	0.385752	0.339773
	$\hat{w}_G$	0.255026	0.142591	1.788514
	$\hat{w}_A$	0.239991	0.289994	0.827572
n = 20	$\hat{w}_s$	0.077791	0.660396	0.117794
	$\hat{w}_h$	0.206687	0.480712	0.429960
	$\hat{w}_G$	0.377069	0.245743	1.534404
	$\hat{w}_A$	0.212039	0.272067	0.779363

表9和表10与表7和表8有相同的结论,即基于高维高频协方差矩阵收缩估计量的投资策略的绩效最佳,且以基于几何型收缩估计量的投资策略表现最优。将表5与表6对比,可发现选择投资对象不同,获得的组合收益和组合波动也完全不同。各种策略的投资结果也表明,表7和表9展示的组合收益和组合波动远远小于表8和表10展示的组合收益和组合波动。该结果也验证了高风险资产对应高收益的论断。

## 五、结论与展望

在高频金融数据建模与高维资产组合的条件下,考虑到金融资产价格的波动率会因为非同步性交易与买卖价差等因素的影响而普遍存在的微观结构噪声,以及金融数据资产收益率序列存在的重尾特征,本文基于 Huber 损失函数和预平均方法对高维协方差矩阵的收缩估计量进行拓展,提出高频高维的协方差矩阵的收缩估计量,提升了该估计量对真实协方差矩阵的估计效果。通过模拟分析,评价了本文提出的估计量的估计效果。以中国上证 A 股的股票资产为投资对象,选择了两种风险偏好型投资组合的高频数据作为样本,用以上估计量进行最小方差投资组合的实证研究。本文的模拟分析和实证结果及相关结论有如下几点。



首先,模拟分析结果表明,本文提出的高频高维协方差矩阵收缩估计量均能够更好地估计总体协方差矩阵,由此还可以提升其在最小方差投资组合中的投资效果,且高频高维协方差的几何型收缩估计在模拟分析时表现最佳,而算术型收缩估计次之。另外,预平均方法确实可以较好地剔除微观结构噪声的影响,Huber 损失函数也确实可以减弱重尾现象对估计结果的影响。

其次,实证分析结果表明,在最小方差投资组合中,本文提出的高频高维协方差矩阵的收缩估计量可以改进以往的投资策略,从而达到提升投资绩效的作用。无论从组合收益、组合波动还是夏普比率的角度来做绩效分析,基于高频高维协方差的收缩估计量构成的投资策略表现最佳,其中又以几何型收缩估计量构建的投资策略表现最优。

最后,从两种不同类型的资产的投资绩效来看,两种不同类型的绩效验证了高风险高收益的论断,且无论哪种类型的资产,最小方差投资组合策略都可以挑选出收益高风险小的投资组合进行投资。

综合以上研究结论,本文最主要的贡献在于进一步扩展和丰富了高频高维协方差矩阵的估计方法,引入了微观结构噪声和重尾两个特征来刻画原有的协方差矩阵的特征,并证实改进后的估计量能够更加准确地估计总体的协方差矩阵,可以为投资组合的优化以及风险管理提供参考。本文提出的估计量是用于高频金融数据背景下的高维投资组合优化,建模的高频交易是建立在5分钟交易数据之上,在之后的研究中,可探讨更换更高频次的交易频率,与量化交易策略结合,获取更高的投资回报。

#### 参考文献:

- [1] MARKOWITZ H M. Portfolio selection[J]. The Journal of Finance, 1952, 7(1): 77-91.
- [2] MANDELROT B. The variation of certain speculative prices[J]. Journal of Business, 1963, 36(4): 394-419.
- [3] FAMA E F. Portfolio analysis in a stable paretian market[J]. Management Science, 1965, 11(3): 404-419.
- [4] MERTON R C. On estimating the expected return on the market: an exploratory investigation [J]. Journal of Financial Economics, 1980, 8(4): 323-361.
- [5] JAGANNATHAN R, MA T. Risk reduction in large portfolios: why imposing the wrong constraints helps[J]. Journal of Finance, 2003, 58(4): 1651-1684.
- [6] CANDELON B, HURLIN C, TOKPAVI S. Sampling error and double shrinkage estimation of minimum variance portfolios[J]. Journal of Empirical Finance, 2012, 19(4): 511-527.
- [7] DEMIGUEL V, GARLAPPI L, NOGALES F J, et al. Optimal versus naive diversification: how inefficient. Is the 1/N portfolio strategy? [J]. Review of Financial Studies, 2007, 22(5): 1915-1953.
- [8] HAUGEN R A, BAKER N L. The efficient market inefficiency of capitalization-weighted stock portfolios[J]. Journal of Portfolio Management, 1991, 17(3): 35-40.
- [9] CHAN L K, KARCESKI J, LAKONISHOK J. On portfolio optimization: forecasting covariances and choosing the risk model[J]. Review of Financial Studies, 1999, 12(5): 937-974.
- [10] CLARKE R G, SILVA H D, THORLEY S. Minimum-variance portfolios in the U. S. equity market [J]. Journal of Portfolio Management, 2006, 33(1): 10-24.
- [11] ENGLE R F. Autoregressive conditional heteroscedasticity with estimates of the variance of U. K. inflation [J]. Econometrica, 1982, 50(4): 987-1007.
- [12] BOLLERSLEV T. Generalized autoregressive conditional heteroskedasticity [J]. Journal of Econometrics, 1986, 31(3): 307-327.
- [13] CLARK P K. A subordinated stochastic process model with finite variance for security prices [J]. Econometrica, 1973, 41(1): 135-155.
- [14] 王天一, 黄卓. 高频数据波动率建模——基于厚尾分布的 Realized GARCH 模型 [J]. 数量经济技术经济研究, 2012(5): 149-161.
- [15] 于孝建, 王秀花. 基于混频已实现 GARCH 模型的波动预测与 VaR 度量 [J]. 统计研究, 2018(1): 104-116.
- [16] 郭宝才, 项琳. 基于跳跃, 好坏波动率的混频已实现 EGARCH 模型的波动率预测与风险度量 [J]. 商业经济与管理, 2022(5): 79-97.

- [17] ANDERSEN T G, BOLLERSLEV T, DIEBOLD F X. The distribution of exchange rate volatility[J]. *Journal of the American Statistical Association*, 1999, 96(453): 42–55.
- [18] ANDERSEN T G, BOLLERSLEV T, DIEBOLD F X. Exchange rate returns standardized by realized volatility are (nearly) gaussian[J]. *Multinational Finance Journal*, 2000, 4(3/4): 159–179.
- [19] ANDERSEN T G, BOLLERSLEV T. Modeling and forecasting realized volatility[J]. *Econometrica*, 2003, 71(2): 579–625.
- [20] 徐正国, 张世英. 多维高频数据的“已实现”波动建模研究[J]. *系统工程学报*, 2006(1): 6–11.
- [21] 徐正国, 张世英. 高频时间序列的改进“已实现”波动特性与建模[J]. *系统工程学报*, 2005(4): 344–350.
- [22] ZHANG L, MYKLAND P A, AT-SAHALIA Y. A tale of two times scales: determining integrated volatility with noisy high frequency data[J]. *Journal of the American Statistical Association*, 2005, 100(472): 1394–1411.
- [23] BARNDORFF-NIELSEN O E, HANSEN P R, LUNDE A, et al. Designing realized kernels to measure the ex-post variation of equity prices in the presence of noise[J]. *Econometrica*, 2008, 76(6): 1481–1536.
- [24] JACOD J, LI Y, MYKLAND P A, et al. Microstructure noise in the continuous case[J]. *Stochastic Processes and Their Applications*, 2009, 119(7): 2249–2276.
- [25] RHIF M, BEN A A, FARAH I R, et al. Wavelet transform application for/in non-stationary time-series analysis: a review[J]. *Applied Sciences*, 2019, 9(7): 1345.
- [26] JI G, YU J, HU K, et al. An adaptive feature selection schema using improved technical indicators for predicting stock price movements[J]. *Expert Systems With Applications*, 2022, 200: 116941.
- [27] YAN H, OUYANG H. Financial time series prediction based on deep learning[J]. *Wireless Personal Communications*, 2018, 102(2): 683–700.
- [28] WU D, WANG X, WU S. A hybrid method based on extreme learning machine and wavelet transform denoising for stock prediction[J]. *Entropy*, 2021, 23(4): 440.
- [29] VASYL G, BASTIAN G. Modeling and forecasting realized portfolio weights[J]. *Journal of Banking and Finance*, 2022, 138: 1–15.
- [30] CALLOT L, KOCK A, MEDEIROS M C. Modeling and forecasting large realized covariance matrices and portfolio choice[J]. *Journal of Applied Econometrics*, 2017, 32(1): 140–158.
- [31] DENARD G, LEDOIT O, WOLF W. Factor models for portfolio selection in large dimensions: the good, the better and the ugly[J]. *Journal of Financial Econometrics*, 2021, 19(2): 236–257.
- [32] GRIBISCH B, HARTKOPF J P, LIESENFELD R. Factor state-space models for high-dimensional realized covariance matrices of asset returns[J]. *Journal of Empirical Finance*, 2020, 55: 1–20.
- [33] HAFNER C M, LINTON O, TANG H. Estimation of a multiplicative correlation structure in the large dimensional case[J]. *Journal of Econometrics*, 2020, 217(2): 431–470.
- [34] DEMIGUEL V, GARLAPPI L, NOGALES F J, et al. A generalized approach to portfolio optimization: improving performance by constraining portfolio norms[J]. *Management Science*, 2009, 55(5): 798–812.
- [35] BAI J, SHI S. Estimating high dimensional covariance matrices and its applications[J]. *Annals of Economics and Finance*, 2011, 12(2): 199–215.
- [36] BICKEL P J, LEVINA E. Regularized estimation of large covariance matrices[J]. *Annals of Statistics*, 2008, 36(1): 199–227.
- [37] BICKEL P J, LEVINA E. Covariance regularization by thresholding[J]. *Annals of Statistics*, 2008, 36(6): 2577–2604.
- [38] BICKEL P J, LEVINA E. Some theory for Fisher’s linear discriminant function, “naive bayes”, and some alternatives when there are many more variables than observations[J]. *Bernoulli*, 2004, 10(6): 989–1010.
- [39] FURRER R, BENGTTSSON T. Estimation of high-dimensional prior and posterior covariance matrices in kalman filter variants[J]. *Journal of Multivariate Analysis*, 2007, 98(2): 227–255.
- [40] CAI T T, ZHANG C H, ZHOU H H. Optimal rates of convergence for covariance matrix estimation[J]. *Annals of Statistics*, 2010, 38(4): 2118–2144.
- [41] KARONI N E. Operator norm consistent estimation of large-dimensional sparse covariance matrices[J]. *Annals of Statistics*, 2008, 36(6): 2717–2756.
- [42] DONOHO D L, JOHNSTONE I M. Threshold selection for wavelet shrinkage of noisy data[EB/OL]//(1994–06–03)[2022–09–10]. <https://ieeexplore.ieee.org/document/412133>.

- [43] ROTHMAN A, LEVINA E, ZHU J. Generalized thresholding of large covariance matrices[J]. *Journal of the American Statistical Association*, 2009, 104(485): 177-186.
- [44] CAI T T, LIU W. Adaptive thresholding for sparse covariance matrix estimation[J]. *Journal of the American Statistical Association*, 2011, 106(494): 672-684.
- [45] POURAHMADI M. Covariance estimation: the GLM and regularization perspectives[J]. *Statistical Science*, 2011, 26(3): 369-387.
- [46] FAN J, FAN Y, LV J. High dimensional covariance matrix estimation using a factor model[J]. *Journal of Econometrics*, 2008, 147(1): 186-197.
- [47] FAN J, LIAO Y, MINCHEVA M. High-dimensional covariance matrix estimation in approximate factor models[J]. *Annals of Statistics*, 2011, 39(6): 3320-3356.
- [48] FAN J, LIAO Y, MINCHEVA M. Large covariance estimation by thresholding principal orthogonal complements[J]. *Journal of the Royal Statistical Society*, 2013, 75(4): 603-680.
- [49] FAN J, LIAO Y, LIU H. An overview of the estimation of large covariance and precision matrices[J]. *Econometrics Journal*, 2016, 19(1): 1-32.
- [50] STEIN C. Inadmissibility of the usual estimator for the mean of a multivariate normal distribution[EB/OL]. (1956-06-30)[2022-09-10]. <https://doi.org/10.1525/9780520313880-018>.
- [51] SCHÄFER J, STRMMER K. A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics[J]. *Statistical Applications in Genetics and Molecular Biology*, 2005, 4(1): 1-32.
- [52] CHEN Y, WIESEL A, ELDAR Y C, et al. Shrinkage algorithms for MMSE covariance estimation[J]. *IEEE Transactions on Signal Processing*, 2010, 58(10): 5016-5029.
- [53] LEDOIT O, WOLF M. A well-conditioned estimator for large-dimensional covariance matrices[J]. *Journal of Multivariate Analysis*, 2004, 88(2): 365-411.
- [54] CUI X, HWANG J T G, QIU J, et al. Improved statistical tests for differential gene expression by shrinking variance components estimates[J]. *Biostatistics*, 2005, 6(1): 59-75.
- [55] TONG T J, WANG Y B. Optimal shrinkage estimation of variances with applications to microarray data analysis[J]. *Journal of the American Statistical Association*, 2007, 102(102): 113-122.
- [56] XIAO M, CHEN T, MING R X, et al. Optimal estimation for power of variance with application to gene-set testing[J]. *系统科学与信息学报(英文版)*, 2020(6): 549-564.
- [57] LI Y, XIAO M, MING R X, et al. A family of flexible shrinkage estimators for the variances of high-dimensional gene expressions[J]. *Communication in Statistics-Simulation and Computation*, 2022, 51(11): 6755-6772.
- [58] WANG Y, ZOU J. Vast volatility matrix estimation for high-frequency financial data[J]. *Annals of Statistics*, 2010, 38(2): 943-978.
- [59] AÏT-SAHALIA Y, XIU D. Using principal component analysis to estimate a high dimensional factor model with high-frequency data[J]. *Journal of Econometrics*, 2017, 201(2): 384-399.
- [60] XIA N, ZHENG X. On the inference about the spectral distribution of high-dimensional covariance matrix based on high-frequency noisy observations[J]. *Annals of Statistics*, 2018, 46(2): 500-525.
- [61] CAI T T, HU J, LI Y, et al. High-dimensional minimum variance portfolio estimation based on high-frequency data[J]. *Journal of Econometrics*, 2020, 214(2-3): 482-494.
- [62] FAN J, KIM D. Robust high-dimensional volatility matrix estimation for high-frequency factor model[J]. *Journal of the American Statistical Association*, 2017, 113(523): 1268-1283.
- [63] 陈雨馨. 基于高频数据高维协方差矩阵的估计及应用[D]. 杭州: 浙江工商大学统计学院, 2020.

