Problem 1

Readme:
In this prompt, we will perform covariance matrix estimation and analyze the performance of the covariance matrix estimator.

The functionalities include:
1) TAQTradeReader: import the trade data
2) TAQQuotesReader: import the quote data
3) TAQFilter: extract S&P 500 stocks list and stock splitting information.
4) TAQAdjust: filter out only S&P 500 listed tickers and adjust the stock price and size for stock splitting/ stock buyback.
5) TAQProcess: process ticker by ticker with multi-threading to save the stocks 5-minute mid-quote returns in the matrix form. It will do the job of cleaning abnormal data points in the same time.
6) TAQCovariance: rolling test on covariance estimation, use three main schemes including Empirical, Clipped, and Optimal RIE to clean our covariance matrix. Computing optimized portfolio weighting from cleaned covariance matrix and weight factors and comparing out-of-sample portfolio vol are also included.

a) A description of the methodology you are using to compare the covariance matrix estimators. Please motivate any assumptions you make.
1. Introduction:

In this part, we construct the estimators with for different optimal Markowitz portfolios includes minimum variance optimization case, omniscient case, and random long-short predictors.

$$w := \frac{\hat{\Sigma}^{-1}g}{g^{*}\hat{\Sigma}^{-1}g}$$

where g is a vector of predictions and $\hat{\Sigma}$ is the cleaned covariance matrix. Different optimal portfolios mean different g for weighting.

1. The minimum variance portfolio, corresponding to $g_i = 1$ for any i in $[1, N]$
2. The omniscient case $g_i = \sqrt{N}\tilde{r}_{i,t}$ where $\tilde{r}_{i,t} = (P_{i,t+\tau} - P_{i,t})/P_{i,t}$ for any i in $[1, N]$
3. The mean-reverting case where g_i is inversely proportional to last return r_i for any i in $[1, N]$
4. Random long-short case $g_i = \sqrt{N}v$ where v is a random vector uniformly distributed on the unit sphere.

In this case, we standardize the sum of final portfolio weights to 1 which indicates in this case, across different strategies, we assume they have same capital and no external fund will be allowed.

2. Data Preparation

In this paper, we selected the S&P 500 which represents the most liquid stocks. We First use the same cleaning process as we did in previous two projects. We first adjusted tickers' price and size according to coefficients. Then we clean abnormal data points within TAQProcess. Since we only are about 5-min mid-quotes return, quotes data is the only source of data we needed here. We computed the all-time return matrix as final output and we saved it to local csv file, so that we do not need to waste time generating return matrix every time.

For the train and test splitting, the training set is normally represented by a T * N data matrix, where N stands for number of tickers in that training period and T stands for the time span over this training sample. One thing we want to make sure not to happen is the potential curse of dimensionality, which usually happens when we have less observations than our factors, meaning N >> T. So that we make sure Specifically, for this paper, we choose train window size to be more than 7 days and test to be 1 or 2 days. A little calculation behind is that since we use 5-min mid-quote return. we will have 78 observations every business day. In order for our T to be greater than N, we need at least 7 days' observations. We performed evaluation on rolling basis across the whole datasets. We showed the result of rolling test from 20070904 to 20070920. We got in total 13 days of full observations and rolling under window size ratio 10:2. If you are interested in seeing more generalized results about our rolling test, feel free to generate a full return matrix from 20070620 to 20070920 and do experiments on that data.

It is worth to notify that before we went straight into cleaning covariance matrix, we need to first perform normalization for the return matrix. We want to make sure that we ended up getting a stationary return matrix that is not affected by the scale of cross-sectional volatility. Firstly, we remove the sample mean of each asset. Secondly, we normalize each return by an estimate $\hat{\sigma}_{it}$ of its daily volatility $\tilde{r}_{i,t} = r_{it}/\hat{\sigma}_{it}$, where $\hat{\sigma}_{it} = \sqrt{\sum_j r_{it}^2}$. Finally, we further standardized each ticker return by its own volatility over time. Following those procedures, we successfully restored our return matrix into a stationary return matrix.

3. Model Assumption

The most important assumption in this paper is that $E(r_{i,t}) = 0$ for all stocks i and time t. Since we work on 5minutes frequency data, it is reasonable to assume that the data is normally distributed, so after we do the normalization, the $E(r_{i,t})$ should be very close to zero.

We also want to make sure that our return matrix did not show strong autocorrelation between data points. In another words, we chose 5-min mid-quote return in a way that the 5-min lag will hopefully erase the potential bid-ask bounce hidden inside in quotes data. If there exists any autocorrelation between observations in time scale, our covariance matrix will generate disastrous result since we assume fake independency between observations. Since we tested autocorrelation for trades and quotes data in project 1 and 5-min is the best trade-off between information ratio and possible autocorrelation.

4. Evaluation Framework

In this paper, followed the framework of BBP-Risk2016. We compared 3 different covariance matrix estimators performance including the

i) empirical covariance
ii) clipped covariance matrix
iii) optimal shrinkage

Empirical covariance is corresponding to linear shrinkage method where we manipulated the eigenvectors and eigenvalues linearly.

Clipped covariance is corresponding to clipping eigenvalue shrinkage method where we compute a threshold for eigenvalue clipping using Random Matrix Theory and we average out the rest eigenvalues so that the trace of original eigen matrix matches up with the new cleaned one's.

Optimal covariance is corresponding to rotational invariance estimator method where we realized a way to actually compute the overlap between true and sample eigenvalues.

For evaluation purpose, we choose the R-square which represents the out of sample covariance matrix volatility and the bias statistics which represents the variance of expected value with respect to the risk factors.

Since the $E(r_{i,t}) = 0$, R-squared which represents the variance of expected returns, can be written as:

$$\mathcal{R}^2{}_{(t,w)} = \frac{1}{T_{out}} \sum_{\tau=t+1}^{t+T_{out}} (\sum_{i=1}^{N} w_i X_{i\tau})^2$$

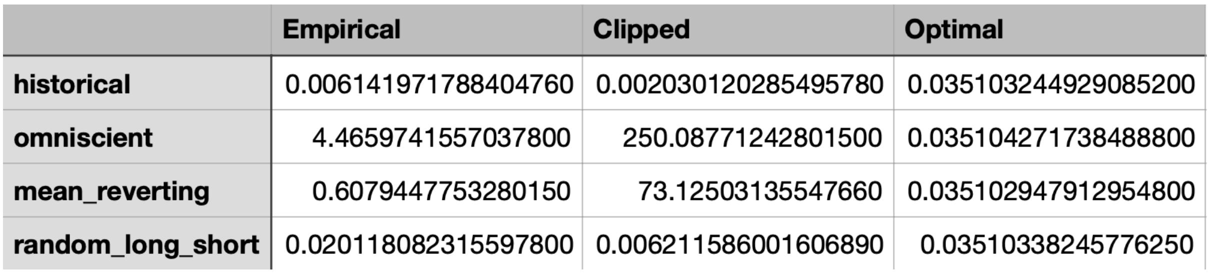
From Axioma paper, we introduced the concept of the bias statistics, the bias statistics is defined as the standard deviation of the standardized returns.

$$\text{Bias Statistics} = \sqrt{var(Z_t)} \text{ with } Z_t = r_t/\sigma_t$$

In addition, we check the null hypothesis that the model is unbiased with 95% confidence interval that whether the bias statistics in the $[1 - \sqrt{2/T}, 1 + \sqrt{2/T}\ ]$.

For the later parts, we will analyze the performance across different weights and shrinkage methods combination to find the optimal solutions based on the out-of-sample average volatility and the bias statistics.

b) Provide an analysis of the covariance matrix and their performance



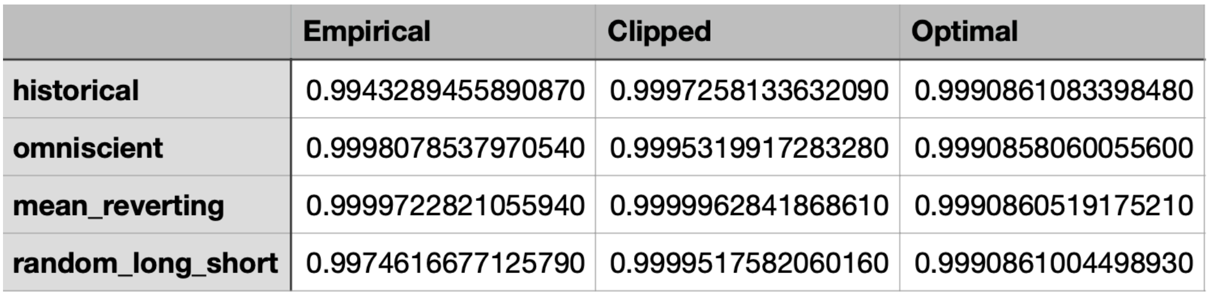
	Empirical	**Clipped**	**Optimal**
historical	0.006141971788404760	0.002030120285495780	0.035103244929085200
omniscient	4.4659741557037800	250.08771242801500	0.035104271738488800
mean_reverting	0.6079447753280150	73.12503135547660	0.035102947912954800
random_long_short	0.020118082315597800	0.006211586001606890	0.03510338245776250

From the R-squared results table, we can see that for minimum variance optimization portfolio, the clipped methods work the best. The clipped covariance estimation “clips” the eigenvalues of an empirical covariance matrix by using PCA to choose appropriate K eigenvalues in order to provide a cleaned matrix. For MVO portfolio, since the weights is sensitive to the covariance computation. The clipped methods provide a cleaned matrix for this case.

For omniscient and random long short case, we can find that the optimal methods constantly perform better than the empirical methods. It indicates that the cleaning scheme works. Since optimal methods provide a convex way to correct the systematic bias for small eigenvalues. One of the potential reasons will be that the error of empirical methods mainly comes from some large extreme wrong estimations. After correcting for the systematic bias, the results become way better.

Another part needs to be highlighted is that all the volatility for the random long short portfolio is small. The primary reason may be the portfolio weights are generated randomly. It this case, when we generate portfolios, it brings some rebalancing and diversification to the whole portfolio which may help decrease the volatility.

From the bias estimation table, we can see that across different methods. The 95% confidence interval is [0.8875, 1.1125].
where $T = (T_{total} - N/q)/T_{out}$. Since the majority of value fall in the 95% confidence intervals, we can't reject the null hypothesis that the model is unbiased.



	Empirical	Clipped	Optimal
historical	0.9943289455890870	0.9997258133632090	0.9990861083398480
omniscient	0.9998078537970540	0.9995319917283280	0.9990858060055600
mean_reverting	0.9999722821055940	0.9999962841868610	0.9990860519175210
random_long_short	0.9974616677125790	0.9999517582060160	0.9990861004498930

In addition, from the table we can see, the majority of the bias statistics value is close to 1, which indicates that the risk forecasts are accurate.

d) Provide a summary of your findings from c as a set of recommendations of which covariance estimators to use.

We will recommend the optimal methods for covariance estimators. From the table in c) we can see, optimal estimators provide a consistent way for covariance estimate which results in small out-of-sample volatility and accurate risk forecasts.

In this optimal shrinkage methods, it will compute a cleaned optimal shrinkage, rotationally invariant estimators (RIE) of the true correlation matrix based on noisy in-sample estimate. The RIE estimators replace the eigenvalues from the spectrum of E by an estimation of the true one. Intuitively, it will correct for the systematic downward bias in small eigenvalues.

Among all the methods we discussed above, we will not recommend the empirical uncleaned covariance estimators since the empirical methods is sensitive to the estimation errors and the computed results is not robust.

e) Implement the covariance matrix estimator described in "Exponential Weighting and Random-Matrix-Theory-Based Filtering of Financial Covariance Matrices for Portfolio Optimization" by Szilard Pafka, Marc Potters and Imre Kondor. Then evaluate and compare its performance with your other 3 estimators from above. What do you find?

As in the script TAQCovariance, we implement the exponential weighting methods in the paper. The main idea is that the covariance estimators take on exponential weighted to smooth out the daily values.

$C_{ij} = \frac{1-\alpha}{1-\alpha^T} \sum_{k=0}^{T-1} \alpha^k x_{ik} x_{jk}$

with exponentially weighted random matrix given by:

$$C_{ij} = \sum_{k=0}^{\infty} H_{ik} H_{jk}$$

which H_{ik} having a k-dependent variance $\sigma_k^2 = \sigma^2(1-\alpha)\alpha^k$. In this case, higher α will indicates a larger emphasize on the past observations.

with the table computed above, we can see from results that EWMA increases the out-of-sample volatility but with a smaller bias statistic. We think the primary reason is that the variance is usually accumulated for short time intervals across the trading horizons. The EWMA smooth out the results which results in more accurate calculation of the standardized returns but also attributes to the variance accumulations.