Xie_Kexin_01484052.pdf

by Kexin Xie

Submission date: 30-Aug-2022 03:59PM (UTC+0100)
Submission ID: 185540902
File name: Xie_Kexin_01484052.pdf (10.26M)
Word count: 21820
Character count: 96184

Imperial College London

Imperial College London

Department of Mathematics

Ranking of Covariance Forecasts by Robust Loss Functions

Author: Kexin Xie (CID: 01484052)

A thesis submitted for the degree of

MSc in Mathematics and Finance, 2021-2022

Declaration

The work contained in this thesis is my own work unless otherwise stated.

Acknowledgements

I would like to express my gratitude to my supervisors Dr Mikko Pakkanen and Professor Richard Martin for their generous assistance and valuable advice. Thank you for always providing timely help and guidance whenever needed.

I would also like to thank all the teaching staff of the MSc Mathematics and Finance programme at Imperial College for providing such a rewarding experience. In this special year of blended learning mode, you did a great job in delivering the courses.

I had such great luck in getting to know and work with my brilliant classmates in this programme. Thank you so much for your help and guidance, and I learned a lot from doing coursework with you.

Finally, words fail to express my love and thanks to my loving and supporting family, especially my parents. Your patience and unconditional love always help me overcome challenges and make me a better person.

Abstract

The forecast of the joint covariance matrix of financial asset returns suffers from the difficulty of the true covariance matrix being unobservable. Work has been done in the one-dimensional case by evaluating and comparing the performances of conditional variance forecasts using a "robust" loss function and volatility proxies from observable data. This thesis extends to multidimensional setting and aims to find a suitable loss function to rank the covariance forecasts and prove its relevant properties so that the ranking of competing covariance forecasts is robust to the noise in the covariance proxy, i.e. the ranking is consistent whether forecasts are evaluated using true conditional covariance matrix or an unbiased proxy of it. The comparison method is then applied to simulated and historical financial returns and we test the statistical significance of the difference in the predictive accuracy of different covariance forecasts to arrive at a ranking of competing forecasts.

Contents

List of Figures

List of Tables

Introduction

Both univariate and multivariate volatility(the multivariate volatility is referred to as covariance in this thesis) of financial asset returns can be seen as a measure of risk and plays an important role in portfolio management and asset pricing. It is therefore of great interest to forecast it. In the literature, forecasts are usually evaluated either using statistical loss functions such as in [6] or by economic significance in a global minimum-variance portfolio framework such as in [3]. Sometimes both approaches are employed in forecast evaluation [19]. We focus on evaluation by statistical loss functions.

However, volatility/covariance is unobservable even after the realization of the returns. This means we have to substitute a proxy for the true conditional volatility/covariance into the loss functions. As shown in [4], there can be distortions in the rankings of competing forecasts if noisy proxies are used in forecast comparison. Such distortions will result in an inferior model being chosen as the optimal model for the forecast task. Patton [15] addressed this problem in forecasts of univariate volatility by introducing sufficient and necessary conditions satisfied by the loss functions in order for the ranking to be consistent whether the loss values are computed using the volatility proxy or the unobservable true conditional volatility. Patton [15] also derived functional forms of such a class of loss functions which he called "robust".

Normally, robustness describes the statistical property that an estimator is not sensitive to outliers, and this is the most widely used definition in the literature of covariance forecasts [21]. However, we adopt a slight abuse of the word "robust" following the definition in Patton's work [15]. In this thesis, this refers to the consistency in the rankings of competing forecasts, whether the evaluation is based on the true conditional covariance or an unbiased proxy of it.

Patton and Sheppard [14] gave a multivariate analogue to [15], and Patton's work [15] in univariate volatility was extended by Laurent et al. [7] to the multivariate scenario. This paper used the terminology "consistency" for what we call "robustness" in Patton's work and this thesis. Besides exploring the influence on the distortions of ranking caused by the level of noise in the covariance proxy, it mainly proposed a generalized sufficient and necessary functional form for a class of Bregman distance measures that ensure "robustness" of the ranking.

This thesis in particular studies one such loss function that was not studied in detail in [7]. The assumption that the loss function is uniquely minimised at the true conditional covariance matrix was directly proposed without much discussion

in [7], but this thesis explicitly proves such property and most importantly, the robustness of the loss function is proved from the first principals without high-level assumptions as in Laurent et al. [7].

The thesis is structured as follows. Chapter 1 introduces several main forecast schemes used in covariance forecasts. Several technical points such as the shrinkage of the covariance matrix and random matrix theory filtering are presented in detail. Chapter 2 continues the technical setup by elaborating on how we evaluate different forecasts and the testing procedure we apply to the loss values. Out of several widely-used testing procedures, we choose the Diebold and Mariano test [2] (referred to as DM test in the rest of the thesis) for its ease of implementation and effectiveness in identifying different predictive accuracy. This chapter also contains the most important theoretical results of the thesis– we propose a special case of robust loss functions and prove several important relevant properties. Then we are ready to apply this loss function with the chosen covariance proxy and testing procedure to a variety of simulated return data sets and real returns of stocks and Treasury yield returns, in Chapter 3 and 4 respectively. The robustness of our chosen loss function is verified with simulated return data set, and we are able to identify top performers of covariance forecasts based on the DM test results on all the return data sets. Chapter 5 concludes the thesis.

Chapter 1

Covariance Matrix Forecasts

In this chapter, we introduce several reliable estimation/forecast schemes of the conditional covariance matrix of financial return time series. These include the multivariate version of the exponentially-weighted moving average (EWMA) scheme, Ledoit and Wolf's shrinkage covariance matrix estimators [10, 9, 8], and the filtering method [5, 16] based on random matrix theory.

Suppose we have time series of log returns of N financial assets, each of length T. These time series are arranged in an $N \times T$ matrix $\boldsymbol{X}$. The tth column of $\boldsymbol{X}$ is denoted as $\boldsymbol{X}_t$, and is treated as an observation of a multi-dimensional variable of dimension N. Let $(\mathscr{F}_t)_{t\in\mathbb{Z}}$ be the filtration generated by the process $(\boldsymbol{X}_t)_{t\in\mathbb{Z}}$. We are interested in estimating the covariance matrix $\boldsymbol{C}_{T+1}$ of these N financial assets given $\mathscr{F}_T$. And the conditional covariance matrix is defined as

$$\boldsymbol{C}_{T+1} := \mathbb{E}\left[(\boldsymbol{X}_{T+1} - \mathbb{E}[\boldsymbol{X}_{T+1}|\mathscr{F}_T])(\boldsymbol{X}_{T+1} - \mathbb{E}[\boldsymbol{X}_{T+1}|\mathscr{F}_T])^\intercal|\mathscr{F}_T\right].$$

Meanwhile, we can also interpret this estimation as a forecast for the conditional covariance matrix at time $T+1$ given information at T. The reference to estimation and forecast is therefore interchangeable in the rest of this thesis.

To begin with, we look at the easily-computed and unbiased approach – the sample covariance matrix, which is defined as

$$\boldsymbol{S}_{N,T} = \frac{1}{T}\sum_{t=1}^{T}\left(\boldsymbol{X}_t - \overline{\boldsymbol{X}}\right)\left(\boldsymbol{X}_t - \overline{\boldsymbol{X}}\right)^\intercal,$$

where $\overline{\boldsymbol{X}} = \frac{1}{T}\sum_{t=1}^{T}\boldsymbol{X}_t$ is the sample mean. It is necessary that $N \leq T$ for $\boldsymbol{S}_{N,T}$ to be invertible.

In the application of covariance matrix estimation such as mean-variance portfolio optimization in the framework of Markowitz [11], the number of assets can be large while the length of time series T is limited by lack of data and the non-stationary nature of financial returns. The performance of the sample covariance matrix estimator becomes poor when the ratio $q := N/T$ becomes large. Meanwhile, this estimator is not robust to noise or outliers, further corrupting its performance. In particular, as shown in [13, Chapter 4, pages 57-60], the closer q is to 1, the more likely that $\boldsymbol{S}_{N,T}$ gives eigenvalues very close to zero, which can make the inversion

of $\boldsymbol{S}_{N,T}$ numerically impossible. There are other poor performances [17, Chapter 20, pages 322-327] of $\boldsymbol{S}_{N,T}$ such as the poor out-of-sample realized risk when applied to the Markowitz optimal portfolio [11]. Therefore, we need to consider more accurate estimators that take into account the heteroskedasticity of volatility and correlations, seek to attenuate estimation noise and avoid ill-conditioning (that is, inversion hugely amplifies the estimation error).

1.1 Multivariate EWMA Scheme

The exposition in this section is partly based on [12] with slightly different notations. The multivariate version of the exponentially-weighted moving average (EWMA) scheme is an extension of the univariate version. It takes into account the heteroskedasticity of covariance and the non-stationarity of financial series. Compared with GARCH modelling, the EWMA scheme is more straightforward and it usually produces quite close results compared with formal multivariate GARCH modelling [12, Chapter 9, page 340].

Let $(\boldsymbol{X}_t)_{t\in\mathbb{Z}}$ be a multivariate process of dimension N with zero conditional mean.(This assumption is justifiable by the nearly zero sample mean of stock returns.) The conditional covariance matrix at time $t+1$ is written as

$$\boldsymbol{C}_{t+1} := \mathbb{E}[\boldsymbol{X}_{t+1}\boldsymbol{X}_{t+1}^{\intercal}|\mathscr{F}_t].$$

An EWMA forecast $\boldsymbol{H}_{t+1}$ of $\boldsymbol{C}_{t+1}$ given $\mathscr{F}_t$ is computed recursively by

$$\boldsymbol{H}_{t+1} = \alpha\boldsymbol{H}_t + (1-\alpha)\boldsymbol{X}_t\boldsymbol{X}_t^{\intercal}, \tag{1.1.1}$$

where α is a positive number usually slightly smaller than 1. The forecast consists of a weighted sum of one-step-earlier forecast $\boldsymbol{H}_t$ and a conditionally unbiased estimator $\hat{\boldsymbol{C}}_t := \boldsymbol{X}_t\boldsymbol{X}_t^{\intercal}$ of the current covariance (which we refer to as covariance proxy at time t from now), so that the forecast reacts to new information at each step. A more rigorous definition of the equally weighted moving average forecast is

$$\boldsymbol{H}_{t+1} = \frac{1-\alpha}{1-\alpha^{t+1}}\sum_{k=0}^{t}\alpha^k\boldsymbol{X}_{t-k}\boldsymbol{X}_{t-k}^{\intercal}, \tag{1.1.2}$$

which makes sure the weights assigned to each covariance proxy add up to 1. In the limit of $t\longrightarrow\infty$, the denominator vanishes, and (1.1.2) is asymptotically

$$\boldsymbol{H}_{t+1} \approx (1-\alpha)\sum_{k=0}^{t}\alpha^k\boldsymbol{X}_{t-k}\boldsymbol{X}_{t-k}^{\intercal},$$

and this approximation is used in applications (consistent with (1.1.1)) as long as t is a sufficiently big integer. Denote the weight as $w_k := (1-\alpha)\alpha^k$, where k is the number of days back from t, then the exponentially weighted average of look back period as $t\to\infty$ is

$$\overline{T}(\alpha) := \sum_{k=0}^{\infty} kw_k = \frac{\alpha}{1-\alpha}. \tag{1.1.3}$$

$\overline{T}(\alpha)$ is increasing in α, and we can interpret alpha as a measure of the length of the effective look-back window. The larger α, the longer the look-back period, and more memory of past covariances are retained. In its application, too big α values limit the ability of the EWMA scheme to capture the non-stationarity of financial time series, while too small values waste data and make the forecast of covariance matrix numerically singular. Therefore, we wish to find the optimal α in application to data.

1.2 Ledoit and Wolf's Shrinkage Approach

The exposition is based on work by Ledoit and Wolf [10, 9, 8]. Let $\boldsymbol{X} \in \mathbb{R}^{N\times T}$ denote T observations of a vector of N random variables. For the shrinkage analysis, we assume these T observations are independent and identically distributed through time. The aim is to come up with a reliable covariance estimator of the true covariance $\boldsymbol{C}$ given information in $\boldsymbol{X}$.

1.2.1 Why Shrinkage and Choice of Shrinkage Targets

Despite being unbiased, the sample covariance matrix (SCM) contains too much estimation error when the ratio of the number of assets to the number of observations becomes large. On the other hand, a very structured estimator such as the single-factor model of Sharpe [18] is potentially misspecified and severely biased. There is a clear trade-off between the bias and the estimation error. To reach a compromise, we consider a weighted average of the SCM and a structured estimator (referred to as the shrinkage target) as an estimator of the covariance matrix, a technique known as linear shrinkage. Denote the shrinkage target by $\boldsymbol{F}$ and the SCM by $\boldsymbol{S}$ we can write our shrinkage estimator as

$$\hat{\boldsymbol{C}}_{Shrink} := \delta\boldsymbol{F} + (1-\delta)\boldsymbol{S} \tag{1.2.1}$$

where $\delta \in (0,1)$ is called the shrinkage intensity and is interpreted as the weight given to the structured estimator $\boldsymbol{F}$. The closer δ is to 1, the more structure is introduced to $\hat{\boldsymbol{C}}_{Shrink}$ and vice versa. This approach effectively pulls extreme coefficients in $\boldsymbol{S}$ towards more central ones to reduce estimation error, in particular, the extreme eigenvalues of $\boldsymbol{S}$ will be adjusted to avoid the computational singularity. It is, therefore, well-conditioned and won't break down even when $N > T$. Asymptotically, the linear combination yields a more accurate estimator than either the sample covariance matrix or the structured estimator.

Next, we discuss three types of structured estimators as the shrinkage target:

- **The Identity Model**

  This model [10] assumes that all variances are the same and all covariances are zero, then the shrinkage target is a scalar multiple of the identity matrix so that $\boldsymbol{F}_{\mathcal{I}} = \nu\boldsymbol{I}$, where ν is a positive scalar.

- **The Equi-Correlation Model**

  This model [9] treats all pairwise correlations to be equal, and the constant correlation is estimated by averaging over sample correlations between each pair. It makes more sense than simply assuming all assets are uncorrelated. The shrinkage target has sample variances on the diagonal. Let s_{ij} denote the entries of $\boldsymbol{S}$, then the sample correlation between asset i and asset j and the correlation estimate are expressed as

$$r_{ij} = \frac{s_{ij}}{\sqrt{s_{ii}s_{jj}}} \qquad \text{and} \qquad \bar{r} = \frac{2}{N(N-1)} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} r_{ij}$$

  respectively. Our constant correlation shrinkage target $\boldsymbol{F}_{\mathcal{CC}}$ satisfies

$$f_{ii} = s_{ii} \qquad \text{and} \qquad f_{ij} = \bar{r}\sqrt{s_{ii}s_{jj}} \ \text{ when } i \neq j$$

  The equi-correlation model incorporates more structure from the available data compared to the over-simplified identity model.

- **The Market Index Model**

  This model [8] adopts a different angle and derives the structure from a single-index model. It is suggested by Sharpe [18] and assumes that the asset returns follow the expression:

$$x_{it} = \alpha_i + \beta_i x_{0t} + \epsilon_{it}$$

  where x_{it} denote the entries of $\boldsymbol{X}$, x_{0t} represents the return of the market index at time t and ϵ_{it} are residuals that are uncorrelated across asset and time, and also uncorrelated with the market return. For each asset, variance of residual is constant and is denoted by $Var(\epsilon_{it}) = \delta_i$. We can write the vector of N variables as $\boldsymbol{x} = [x_1, x_2, \ldots, x_N]^\intercal$, $\boldsymbol{\beta} = [\beta_1, \beta_2, \ldots, \beta_N]^\intercal$, and $\boldsymbol{\epsilon} = [\epsilon_1, \epsilon_2, \ldots, \epsilon_N]^\intercal$, then we have

$$\boldsymbol{x} = \boldsymbol{\beta} x_0 + \boldsymbol{\epsilon}$$

  Then, the theoretical covariance matrix of $\boldsymbol{x}$ is

$$\boldsymbol{\Phi} := Var(x_0)\boldsymbol{\beta}\boldsymbol{\beta}^\intercal + \boldsymbol{\Delta}$$

  where $\boldsymbol{\Delta}$ is a diagonal matrix with ith entry on the diagonal equal to δ_i. Note that we assume $\boldsymbol{\Phi} \neq \boldsymbol{C}$, otherwise the structured estimator is asymptotically unbiased.

  To get an estimate of the unobservable $\boldsymbol{\Phi}$, we carry out a univariate linear regression of returns of stock i against the market index. The coefficients from the regression give an estimate of β_i, and δ_i is calculated as the empirical variance of the regression residuals. The market index variance is computed as the sample variance. Let $\hat{\sigma}_0$ denote the estimator of $Var(x_0)$, and $\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\Delta}}$ denote the estimators mentioned above, we get the market shrinkage target

$$\boldsymbol{F}_{\mathcal{M}} = \hat{\sigma}_0 \hat{\boldsymbol{\beta}} \hat{\boldsymbol{\beta}}^\intercal + \hat{\boldsymbol{\Delta}}.$$

Using the single-index model avoids the debatable selection of multiple factors, and the market index can be any broad-based market index. In the implementation, we take the equally-weighted market factor by averaging returns over the N assets.

1.2.2 Optimal Shrinkage Intensity

The key question is: which shrinkage intensity should be chosen for the best accuracy of the shrinkage estimator? Ledoit and Wolf [10, 9, 8] quantified the problem by finding δ that minimizes the expected Frobenius norm between the shrinkage estimator and the true covariance matrix $\boldsymbol{C}$.

Definition 1.2.1 (Frobenius Norm)**.** The Frobenius norm of an $N \times N$ symmetric matrix $\boldsymbol{Z}$ with entries $(z_{ij})_{i,j=1,\cdot,N}$ and eigenvalues $(\lambda_i)_{i=1,\cdots,N}$ is defined by:

$$\|\boldsymbol{Z}\|^2 = \text{Trace}\left(\boldsymbol{Z}^2\right) = \sum_{i=1}^{N}\sum_{j=1}^{N} z_{ij}^2 = \sum_{i=1}^{N} \lambda_i^2.$$

With the above definition, we consider the quadratic loss function:

$$L(\delta) = \|\delta\boldsymbol{F} + (1-\delta)\boldsymbol{S} - \boldsymbol{C}\|^2.$$

The smaller the loss, the more accurate the estimator is. The optimal shrinkage intensity is therefore

$$\delta^* = \underset{\delta\in(0,1)}{\arg\min}\, \mathbb{E}[L(\delta)]$$

and the shrinkage estimator is given by

$$\hat{\boldsymbol{C}}^*_{Shrink} := \delta^*\boldsymbol{F} + (1-\delta^*)\boldsymbol{S}.$$

First order condition of the risk function $\mathbf{R}(\delta) := \mathbb{E}[L(\delta)]$ gives the optimal shrinkage intensity and the second order condition verifies it is a minimum. However, δ^* depends on the true covariance $\boldsymbol{C}$, of which we don't have the knowledge. Fortunately, we can find a consistent estimator of δ^*. Notice that the asymptotic framework of the identity shrinkage target [10] is different from those of the other two shrinkage targets [9, 8]. The former adopts a framework called general asymptotics where the number of assets N is allowed to go to infinity while the ratio N/T remains bounded, whereas the latter assumes N is fixed and finite and T tends to infinity. The form of $\hat{\boldsymbol{C}}^*$ can be unified though, with detailed expression in [8].

When applied to data, the forecast is carried out in a rolling-window manner, and the sample covariance matrix of a certain period is shrunk towards the target to give the forecast for the next day. It is shown in [10] that the consistent estimator of δ^* shows good behaviours when $N \geq 20$ and $T \geq 20$. Hence, applying the shrinkage estimator to sufficiently large finite samples doesn't affect the results. It remains to determine what is the optimal look-back period for best forecast performance, which can be quantified in the next chapter.

1.3 Random Matrix Theory Filtering

If we calculate the empirical cross-correlation matrix of financial asset returns with finite length, a lot of noise is involved. Although noise can be neglected as the length of observations goes to infinity, the accuracy of the estimation will be negatively affected by the non-stationarity of cross-correlations, and the historical data have limited length. Hence, we cannot get rid of noise by using an arbitrarily long time window. While Ledoit and Wolf's shrinkage approach blends a structured estimator with a noisy but (asymptotically) unbiased estimator, the filtering technique introduced below employs results from the random matrix theory to statistically analyse how much information and noise is contained in the correlation matrix and produce a filtered version of it.

Let's first give a brief overview of the random matrix theory background.

Definition 1.3.1 (Wishart Matrices)**.** Suppose we have an $N \times T$ matrix $\boldsymbol{R}$ whose columns are vectors drawn independently from a multivariate Gaussian distribution with zero mean and true (or population) covariance matrix $\boldsymbol{C}$. Then we call

$$\boldsymbol{S} = \frac{1}{T}\boldsymbol{R}\boldsymbol{R}^T$$

a Wishart matrix.

In the case when the cross correlation matrix of the columns of $\boldsymbol{R}$ is the identity matrix (i.e. the variables are uncorrelated), if we keep the ratio $Q := T/N$ fixed while sending T and N to infinity, the density of the eigenvalues of $\boldsymbol{S}$ is given by

$$\rho(\lambda) = \frac{Q}{2\pi\sigma^2}\frac{\sqrt{(\lambda_+ - \lambda)(\lambda_- - \lambda)}}{\lambda} \tag{1.3.1}$$

where σ is the standard deviation of the variables in the columns of $\boldsymbol{R}$ and the maximum and minimum eigenvalues are given by

$$\lambda_\pm = \sigma^2\left(1 \pm \sqrt{\frac{1}{Q}}\right)^2. \tag{1.3.2}$$

$\rho(\lambda)$ is known as the Marčenko-Pastur density. Without loss of generality, we focus on studying the correlation matrix instead of the covariance matrix. (It is straightforward to convert between covariance matrix and correlation matrix.) Because the diagonal of the correlation matrix consists of ones, it's equivalent to assume that every asset has $\sigma = 1$. Note that the Gaussian assumption of Wishart matrix can be relaxed [16] when using the asymptotic density of eigenvalues of $\boldsymbol{S}$, and we still have eigenvalue density of sample correlation matrix $\boldsymbol{S}$ numerically in good agreement with equation (1.3.1).

1.3.1 Finding the Noise Band

Now consider a matrix $\boldsymbol{X} \in \mathbb{R}^{N\times T}$ that denotes T observations of returns of N financial assets. We calculate the normalised return of asset i at time t as

$$\boldsymbol{G}_{it} := \frac{\boldsymbol{X}_{it} - \overline{\boldsymbol{X}}_i}{\sigma_i}$$

where $\overline{\boldsymbol{X}} := \frac{1}{T}\sum_{i=1}^{T} \boldsymbol{X}_t$ is the sample mean vector and σ_i is the sample standard deviation of asset i. Define

$$\boldsymbol{P} := \frac{1}{T}\boldsymbol{G}\boldsymbol{G}^T,$$

then $\boldsymbol{P}$ is the sample correlation matrix. If $\boldsymbol{P}$ is a random matrix (with components of $\boldsymbol{G}$ having zero mean and unit variance and mutually uncorrelated), then the eigenvalue density of $\boldsymbol{P}$ when $N, T \longrightarrow \infty$ and the ratio $Q = T/N >= 1$ fixed is given by equation (1.3.1) with $\sigma = 1$, and we write it as

$$\rho_{rm}(\lambda) = \frac{Q}{2\pi}\frac{\sqrt{(\lambda_+ - \lambda)(\lambda_- - \lambda)}}{\lambda}. \tag{1.3.3}$$

The empirical density of eigenvalues of $\boldsymbol{P}$, denoted by $\rho(\lambda)$, is then compared with the theoretical density $\rho_{rm}(\lambda)$. Some deviations of $\rho(\lambda)$ from $\rho_{rm}(\lambda)$ will occur, and this indicates real historical returns are not completely random, and they contain some information distinguishable from the noise. The purpose of filtering is to extract those information and filter out the noise by reconstructing the correlation matrix. Also, by the results in [16], we know that the deviation of empirical density from theoretical one (especially the largest few eigenvalues) doesn't result from finite values of T and N, it is therefore safe to regard the large eigenvalues of $\boldsymbol{P}$ as carrying information. However, we need to determine quantitatively a noise band outside of which eigenvalues of $\boldsymbol{P}$ are considered non-random.

The process of finding the noise band is best illustrated through an example. Let's take the normalised log returns of 300 stocks of S&P500 during the years 2018-2022. We have $N = 300$ and $T = 1110$, so $Q = T/N = 3.7$. The RMT (random matrix theory) boundaries are given by equation (1.3.2) with $[\lambda_-, \lambda_+] = [0.23, 2.31]$ and $\rho_{rm}(\lambda)$ plotted in a black dashed line in the below Figure 1.1. This plot only shows $\rho(\lambda)$ for $\lambda \in (0, \lambda_+ + 1)$ to compare the random bulk of eigenvalues. The eigenvalues of $\boldsymbol{P}$ are concentrated in bulk between 0 and 1, with a small proportion of much larger eigenvalues. In particular, the upper bound is $\lambda_+ = 2.31$, while the biggest eigenvalue is 127.49, which is about 55 times of λ_+. This is the most obvious deviation from the random matrix theory results. Indeed, its corresponding eigenvector represents an influence common to all stocks [16], often referred to as the "market factor".

Meanwhile, we can see that $\rho_{rm}(\lambda)$ doesn't quite fit $\rho(\lambda)$, this means the parameters $Q = 3.7$ and $\sigma = 1$ do not give the Marčenko-Pastur density that matches $\rho(\lambda)$. By minimizing the squared loss, we find a best fit with $\hat{Q} = 2.08, \hat{\sigma} = 0.59$ as shown by the red curve and denoted as $\hat{\rho}_{rm}(\lambda)$. We can interpret $\hat{\rho}_{rm}(\lambda)$ as the density contributing to the randomness in the correlation matrix eigenvalues. The best fit density has theoretical RMT boundaries $[\hat{\lambda}_-, \hat{\lambda}_+] = [0.03, 0.99]$, covering 88% of the spectrum of $\boldsymbol{P}$. The fitted density agrees with the bulk of empirical eigenvalues satisfactorily, but this is not sufficient to draw the conclusion that eigenvalues spectrum covered in the RMT boundaries are random and only those that deviate carry information. Luckily, results from [16] confirm that the statistics of the bulk of empirical eigenvalues is consistent with those of a real symmetric random matrix. Therefore, the information carried by the genuine correlation exists in the eigenvalues that deviate from the RMT boundaries, which we refer to as the noise

band. A fraction of $\hat{\sigma}^2 = 0.34$ of variance is explained by eigenvalues corresponding to random noise.

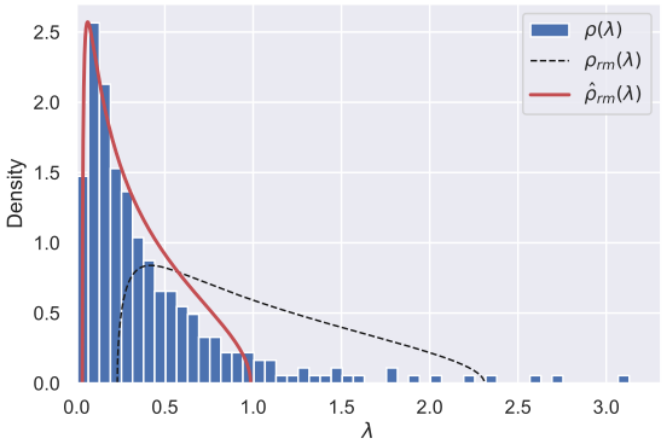

Figure 1.1: The theoretical distribution $\rho_{rm}(\lambda)$, the empirical density $\rho(\lambda)$ of $\boldsymbol{P}$ and the fitted theoretical distribution $\hat{\rho}_{rm}(\lambda)$ in the range $\lambda \in (0, \lambda_+ + 1)$, where $\boldsymbol{P}$ is formed from $N = 300$ stocks in S&P500 stock index from 2018 to 2022

It is tempting to keep only the eigenvalues above $\hat{\lambda}_+$, but the finiteness effect of N and T can affect the upper bound of the noise band. We can see from Figure 1.1 that beyond the right end of the red curve, there exists a continuation of the spectrum until about 1.5, and this continuation looks like part of the random bulk. Indeed, summing up eigenvalues larger than 1.5 divided by the sum of all eigenvalues of $\boldsymbol{P}$ yields that 0.67 of variance is explained by eigenvalues larger than 1.5, which closely matches that a fraction of 0.34 of variance is explained by noise. (The fractions 0.67 and 0.34 add up to 1.01, very close to 1.) The lower bound of the noise band is very close to zero (about 0.03), there are only 3 empirical eigenvalues below it. Because the eigenvalues of the correlation matrix can be interpreted as the variance of a portfolio of stocks (with normalised returns), a portfolio with nearly zero in-sample variance is very unrealistic, and it is very likely to be a result of noise. Hence, we adjust the lower bound of the noise band to be zero and the upper bound is modified to be the eigenvalue such that the eigenvalues above it explain the amount of variance that the best fit Marčenko-Pastur density fails to explain. Denote this upper bound by λ^* and the eigenvalues of $\boldsymbol{P}$ as $\lambda_{\boldsymbol{P}}$, it is defined as

$$\lambda^* := \max \{ \lambda : \sum_{\lambda_{\boldsymbol{P}} \geq \lambda} \lambda_{\boldsymbol{P}} \geq 1 - \hat{\sigma}^2 \}. \tag{1.3.4}$$

Therefore, our final noise band is $[0, \max\{\hat{\lambda}_+, \lambda^*\}]$, and we are ready to perform the filtering process, which essentially modifies the eigenvalues in the noise band while preserving the trace of $\boldsymbol{P}$.

1.3.2 Implementation of Filtering

The real symmetric correlation matrix $\boldsymbol{P}$ can be decomposed as

$$\boldsymbol{P} = \mathbf{Q}\mathbf{D}\mathbf{Q}^\intercal$$

where columns of $\mathbf{Q}$ are normalised eigenvectors of $\boldsymbol{P}$ denoted as $(\boldsymbol{v}_i)_{i=1,\cdots,N}$, and $\mathbf{D}$ is a diagonal matrix with eigenvalues $(\lambda_i)_{i=1,\cdots,N}$ on the diagonal and the eigenvalues satisfy $\lambda_1 \geq \cdots \geq \lambda_N$. The above decomposition can also be written as

$$\boldsymbol{P} = \sum_{i=1}^{N} \lambda_i \boldsymbol{v}_i \boldsymbol{v}_i^\intercal.$$

After finding the noise band, we proceed with two ways of filtering.

The first method is from Laloux et al.[5] (referred to as the LCPB method), who propose that because eigenvalues below λ^* are only a result of random noise, we shouldn't distinguish between them. Hence, we replace each eigenvalue in the noise band with the average of them. By doing so, the trace of $\boldsymbol{P}$ is preserved. More precisely, suppose $\lambda_{k-1} > \lambda^*$ and $\lambda_k \leq \lambda^*$ for some $2 \leq k \leq N$, let

$$\overline{\lambda} := \frac{1}{N-k+1} \sum_{i=k}^{N} \lambda_i,$$

we have the filtered correlation matrix

$$\boldsymbol{P}_{LCPB} := \sum_{i=1}^{k-1} \lambda_i \boldsymbol{v}_i \boldsymbol{v}_i^\intercal + \sum_{i=k}^{N} \overline{\lambda} \boldsymbol{v}_i \boldsymbol{v}_i^\intercal. \tag{1.3.5}$$

Strictly speaking, $\boldsymbol{P}_{LCPB}$ is not a correlation matrix with diagonal of 1's, rather, it can be seen as the correlation between clean and noisy time series.

The second method comes from Plerou et al.[16] (referred to as the PG+ method) This procedure replaces $\lambda_k, \cdots, \lambda_N$ by zero because the random eigenvalues and eigenvectors shouldn't contribute to the genuine correlation. i.e. we construct

$$\boldsymbol{P}'_{PG+} := \sum_{i=1}^{k-1} \lambda_i \boldsymbol{v}_i \boldsymbol{v}_i^\intercal.$$

Then, the diagonal elements of $\boldsymbol{P}'_{PG+}$ are all replaced by 1's to give the filtered correlation matrix $\boldsymbol{P}_{PG+}$.

To sum up, given a sample correlation matrix $\boldsymbol{P}$ with the ratio Q, we implement the filtering procedure as follows:

1. Find the eigenvalues and eigenvectors of $\boldsymbol{P}$
2. Calculate the maximum eigenvalue (1.3.2) using the given Q and $\sigma = 1$
3. Only keep the small eigenvalues in $[0, \lambda_+ + 1]$ and find the empirical distribution of these eigenvalues for fitting a best fit Marčenko-Pastur distribution. The best fit is quantified by minimizing squared loss from real Marčenko-Pastur distribution. Obtain the fitted $\hat{Q}$ and $\hat{\sigma}$.

4. Find the noise band $[0, \max\{\hat{\lambda}_+], \lambda^*\}]$ and modify the corresponding eigenvalues to get $\boldsymbol{P}_{LCPB}$ or $\boldsymbol{P}_{PG+}$.

Both filtered correlation matrices will be converted back to covariance matrix in the covariance forecast schemes. The random matrix theory filtering will be applied to both the sample covariance matrix forecast and EWMA forecast to reduce the negative effect of noise. Most of the time, the effects of the two types of filtering are comparable, and we use the PG+ method in our implementation of filtering in Chapters 3 and 4.

Chapter 2

Forecast Performance Evaluation

The performances of covariance forecasts in the previous chapter are evaluated by a class of loss functions, from which we define the optimal forecast. Then, we introduce the Diebold-Mariano test [2] to compare the performances of different forecasts. The loss functions have to be "robust" [15] for the DM test applied to data to reflect the infeasible ranking, which is the ranking obtained using the unobservable true covariance matrix. Our choice of the loss function in a multi-dimensional scenario is motivated by the class of robust and homogeneous loss functions in Patton's work [15], and we will prove our loss function is indeed robust with the assumption also justified.

2.1 The Optimal Forecast

Under the same setting of chapter 1, we are interested in forecasting the conditional covariance matrix $\boldsymbol{C}_t$ at time t given $\mathscr{F}_{t-1}$, where $\boldsymbol{C}_t$ is defined as

$$\boldsymbol{C}_t := \mathbb{E}[\boldsymbol{X}_t \boldsymbol{X}_t^\intercal | \mathscr{F}_{t-1}].$$

Note that we assume the multi-dimensional variables have zero conditional mean, i.e.

$$\mathbb{E}[\boldsymbol{X}_t | \mathscr{F}_{t-1}] = \boldsymbol{0}.$$

However, the true conditional covariance matrix of financial returns is not observable, we cannot find the forecast error directly. Therefore, we need a proxy $\hat{\boldsymbol{C}}_t$ of it as a standard for comparison such that $\mathbb{E}_{t-1}[\hat{\boldsymbol{C}}_t] = \boldsymbol{C}_t$, where $\mathbb{E}_{t-1}[\cdot]$ is the abbreviation for $\mathbb{E}[\cdot|\mathscr{F}_{t-1}]$. An example can be $\hat{\boldsymbol{C}}_t := \boldsymbol{X}_t \boldsymbol{X}_t^\intercal$.

Using the information in $\mathscr{F}_{t-1}$, we come up with a forecast $\boldsymbol{H}_t$ of $\boldsymbol{C}_t$. $\boldsymbol{H}_t$ is $\mathscr{F}_{t-1}$-measurable and is a real symmetric positive definite matrix. By convention, we denote the set of $N \times N$ real symmetric matrices as $\mathbf{S}^N$, the set of symmetric positive semi-definite matrices as $\mathbf{S}_+^N$, and the set of symmetric positive definite matrices as $\mathbf{S}_{++}^N$. To assess the predictive accuracy of the forecast, we need a loss function

$$L : \mathbf{S}_{++}^N \times \mathbf{S}_{++}^N \to \mathbb{R}_+,$$

where $\mathbb{R}_+$ represents the positive real line, and the optimal forecast is defined as

$$\boldsymbol{H}_t^* := \underset{\substack{\boldsymbol{H}_t \in \mathbf{S}_{++}^N \\ \boldsymbol{H}_t \text{ is } \mathscr{F}_{t-1}-\text{measurable}}}{\arg\min} \mathbb{E}_{t-1}[L(\boldsymbol{C}_t, \boldsymbol{H}_t)]. \tag{2.1.1}$$

A smaller value of the loss function should represent a smaller distance of $\boldsymbol{H}_t$ from $\boldsymbol{C}_t$. We want the forecast to be as close to the true covariance as possible, therefore, it is desired that the expected loss $\mathbb{E}_{t-1}[L(\boldsymbol{C}_t, \boldsymbol{H}_t)]$ is minimised when the forecast equals the variable of interest, which is the conditional covariance. i.e. we want

$$\boldsymbol{H}_t^* = \boldsymbol{C}_t. \tag{2.1.2}$$

Note that $\boldsymbol{H}_t$ and $\boldsymbol{C}_t$ are both $\mathcal{F}_{t-1}$-measurable, we actually have

$$\mathbb{E}_{t-1}[L(\boldsymbol{C}_t, \boldsymbol{H}_t)] = L(\boldsymbol{C}_t, \boldsymbol{H}_t).$$

By (2.1.1) and (2.1.2), we can derive a necessary condition for the optimal forecast to be the true conditional covariance:

$$L(\boldsymbol{C}, \cdot)(\boldsymbol{H}) \text{ is minimised at } \boldsymbol{H} = \boldsymbol{C}$$

The notation $L(\boldsymbol{C}, \cdot)(\boldsymbol{H})$ means that we treat the argument $\boldsymbol{C}$ as fixed and $\boldsymbol{H}$ is the only variable.

However, due to the unobservable nature of $\boldsymbol{C}_t$, we cannot calculate $L(\boldsymbol{C}_t, \boldsymbol{H}_t)$. Instead, we put the conditionally unbiased estimator $\hat{\boldsymbol{C}}_t$ in the loss function. In practice, we define the optimal forecast for a given loss function and proxy as

$$\boldsymbol{H}_t^* := \underset{\substack{\boldsymbol{H}_t \in \mathbf{S}_{++}^N \\ \boldsymbol{H}_t \text{ is } \mathcal{F}_{t-1}-\text{measurable}}}{\arg\min} \mathbb{E}_{t-1}[L(\hat{\boldsymbol{C}}_t, \boldsymbol{H}_t)]. \tag{2.1.3}$$

When mentioning $\boldsymbol{H}_t^*$ below, we refer to the one defined in (2.1.3). Still, we want the optimal forecast derived using (2.1.3) to be the true covariance matrix $\boldsymbol{C}_t$.

2.2 Diebold-Mariano Test

We have quantified the performance of a forecast by the expected loss, and when we have two different forecasts, we can compare them based on the expected loss.

Let's suppose we have two forecasts $\boldsymbol{H}_{1,t}$ and $\boldsymbol{H}_{2,t}$. Let $u_{i,t} := L(\boldsymbol{C}_t, \boldsymbol{H}_{i,t})$ for $i = 1, 2$, and define $d_t := u_{1,t} - u_{2,t}$ as the loss differential. The presence of noise makes it hard to tell which of $u_{1,t}$ and $u_{2,t}$ is bigger, so we perform a statistical test on finite samples as proposed by Diebold and Mariano [2], which we refer to as the DM test. The null hypothesis of the DM test is

$$H_0 : \mathbb{E}[d_t] = 0,$$

which means that on average, there is no difference between the loss values of the two forecasts. i.e. the two forecasts statistically have equal predictive accuracy.

Suppose we have a sample of loss differential series $\{d_t\}_{t=1}^T$ of length T. Denote its sample mean by

$$\overline{d} := \frac{1}{T}\sum_{t=1}^{T} d_t = \frac{1}{T}\sum_{t=1}^{T}(u_{1,t} - u_{2,t}).$$

Let the population mean of the loss differential be μ. Under the assumption of the loss differential series being covariance stationary and short memory, the sample mean $\overline{d}$ can be shown to have the following asymptotic distribution

$$\sqrt{T}(\overline{d} - \mu) \xrightarrow{d.} \mathcal{N}(0, \sigma_d^2)$$

as the sample size goes to infinity. Taking into account of the effect of serial correlation of d_t, we have

$$\sigma_d^2 = \sum_{\tau=-\infty}^{\infty} \gamma_d(\tau) \qquad \text{and} \qquad \gamma_d(\tau) = \mathbb{E}[(d_t - \mu)(d_{t-\tau} - \mu)]$$

where $\gamma_d(\tau)$ is the auto-correlation function of d_t of lag τ. While σ_d is unknown, we use a consistent estimator in the finite sample. And when we only consider the one-step-ahead forecast, we can ignore serial correlation and take

$$\hat{\sigma}_d^2 = \gamma_d(0) = \frac{1}{T}\sum_{t=1}^{T}(d_t - \overline{d})^2.$$

Under the null hypothesis of $\mu = 0$, the test statistic

$$S_1 = \frac{\sqrt{T}\,\overline{d}}{\hat{\sigma}_d} \tag{2.2.1}$$

is asymptotically $\mathcal{N}(0, 1)$. Note that this is a two-sided test. Hence, we reject H_0 at a confidence level of $1 - \alpha$ and conclude that the two forecasts have different predictive accuracy when $S_1 > z_{\frac{\alpha}{2}}$ or $S_1 < -z_{\frac{\alpha}{2}}$, where $z_{\frac{\alpha}{2}} := \Phi^{-1}(1 - \frac{\alpha}{2})$ is the $1 - \frac{\alpha}{2}$ quantile of the standard normal distribution. In particular, if $S_1 > z_{\frac{\alpha}{2}}$, it means on average $u_{1,t} > u_{2,t}$ and $\boldsymbol{H}_{2,t}$ is a more accurate forecast. Vice versa if $S_1 < -z_{\frac{\alpha}{2}}$.

When applying the DM test to the loss differential series, we need to check whether the assumptions of covariance stationarity and short memory are satisfied. To test covariance stationarity, we can first plot the series and visualise it. If visualization cannot provide obvious evidence of non-stationarity, we try the Kwiatkowski-Phillips-Schmidt-Shin (KPSS) Test[1] whose null hypothesis is that the time series considered is stationary around a deterministic trend. Regarding the short memory property, we plot the empirical auto-correlation function to see the rate of decay of the serial correlation of d_t. Rapid decay such as exponential decay or sudden drop to zero after a certain time lag serves as good evidence of a short memory process. However, it's likely that one or both assumptions are not satisfactorily met, in that case, the conclusion from the DM test will be less reliable.

2.3 Robust Loss Functions

When we introduce the DM test, we defined $u_{i,t} := L(\boldsymbol{C}_t.\boldsymbol{H}_{i,t})$. In the actual implementation, we can only use $\hat{\boldsymbol{C}}_t$. By true ranking of the forecasts, we refer to the

[1] https://en.wikipedia.org/wiki/KPSS_test

ranking based on the expected loss $\mathbb{E}[L(\boldsymbol{C}_t, \boldsymbol{H}_{i,t})]$. While the true ranking is not directly available, we can work out the proxy ranking by $\mathbb{E}[L(\hat{\boldsymbol{C}}_t, \boldsymbol{H}_{i,t})]$, with the mean estimated by the sample mean of a series of forecasts. The crucial question is that whether the ranking of two forecasts obtained by comparing the expected loss using the covariance proxy $\hat{\boldsymbol{C}}_t$ is consistent with the true ranking using the true covariance matrix. In another word, if we conclude from the DM test with d_t calculated using $\hat{\boldsymbol{C}}_t$, we need the same conclusion to hold when the true covariance $\boldsymbol{C}_t$ is used. This consistency must be satisfied for the comparison between different forecasts to make sense, as the accuracy is measured as a distance from the true covariance matrix. The consistency, however, may be violated as discussed in [4]. In order to avoid distorted ranking when evaluating forecasts, it is crucial to choose a sensible loss function.

2.3.1 Definition and Properties

To formally address this problem, we define loss functions that are "robust".

Definition 2.3.1 (Robust Loss Function)**.** A loss function $L : \mathbf{S}^N_{++} \times \mathbf{S}^N_{++} \to \mathbb{R}_+$ is robust if, for any two $\mathcal{F}_{t-1}$-measurable forecasts $\boldsymbol{H}_{1,t}$ and $\boldsymbol{H}_{2,t}$, the following is satisfied:

$$\begin{aligned} E\left[L\left(\boldsymbol{C}_t, \boldsymbol{H}_{1,t}\right)\right] &\gtreqless E\left[L\left(\boldsymbol{C}_t, \boldsymbol{H}_{2,t}\right)\right] \\ &\iff \\ E\left[L\left(\hat{\boldsymbol{C}}_t, \boldsymbol{H}_{1,t}\right)\right] &\gtreqless E\left[L\left(\hat{\boldsymbol{C}}_t, \boldsymbol{H}_{2,t}\right)\right] \end{aligned} \tag{2.3.1}$$

for any $\hat{\boldsymbol{C}}_t$ such that $\mathbb{E}_{t-1}[\hat{\boldsymbol{C}}_t] = \boldsymbol{C}_t$.

This idea of robust loss function was introduced and elaborated by Patton [15], whose work focuses on forecasts of one-dimensional conditional volatility. We extend the idea to the multi-dimensional conditional covariances. Conventionally, robustness describes the property of an estimator that is not sensitive to the presence of outliers, while this definition emphasizes the robustness of forecast ranking to the noise in the covariance proxy $\hat{\boldsymbol{C}}_t$.

In the implementation of the DM test, we use $\hat{u}_{i,t} := L(\hat{\boldsymbol{C}}_t.\boldsymbol{H}_{i,t})$ and $\hat{d}_t := \hat{u}_{1,t} - \hat{u}_{2,t}$. We are testing the null hypothesis $\hat{H}_0 : \mathbb{E}[\hat{d}_t] = 0$, and if the null hypothesis is rejected, we conclude that

$$\mathbb{E}[\hat{d}_t] \lesseqgtr 0 \quad \Leftrightarrow \quad \mathbb{E}\left[L(\hat{\boldsymbol{C}}_t, \boldsymbol{H}_{1,t})\right] \lesseqgtr \mathbb{E}\left[L(\hat{\boldsymbol{C}}_t, \boldsymbol{H}_{2,t})\right]$$

Thanks to the loss function L being robust, the property (2.3.1) leads to the conclusion about the infeasible ranking with respect to the distance from the true covariance matrix:

$$\mathbb{E}[\hat{d}_t] \lesseqgtr 0 \quad \Leftrightarrow \quad \mathbb{E}\left[L(\boldsymbol{C}_t, \boldsymbol{H}_{1,t})\right] \lesseqgtr \mathbb{E}\left[L(\boldsymbol{C}_t, \boldsymbol{H}_{2,t})\right]$$

It follows from (2.3.1) that a necessary condition for a loss function to be robust to noisy proxy is that the optimal forecast $\boldsymbol{H}^*_t$ defined in (2.1.3) is the true conditional covariance matrix. Recall that in section 2.1, we put a restriction on our loss function:

Assumption 2.3.2. $L(\boldsymbol{C}, \cdot)(\boldsymbol{H})$ is minimised at $\boldsymbol{H} = \boldsymbol{C}$ and the minimiser is unique.

Under this assumption, we propose and prove the following:

Proposition 2.3.3. *Let assumption 2.3.2 hold, if the loss function is robust in the sense of definition 2.3.1, then the optimal forecast $\boldsymbol{H}_t^*$ under the loss function is the true conditional covariance matrix $\boldsymbol{C}_t$.*

Proof. Let $\tilde{\boldsymbol{H}}_t$ be any $\mathscr{F}_{t-1}$-measurable forecast, by definition of the optimal forecast given in (2.1.3), we have

$$\mathbb{E}_{t-1}\left[L(\hat{\boldsymbol{C}}_t, \boldsymbol{H}_t^*)\right] \leq \mathbb{E}_{t-1}\left[L(\hat{\boldsymbol{C}}_t, \tilde{\boldsymbol{H}}_t)\right]$$

Taking expectation on both sides and using tower property gives

$$\mathbb{E}\left[L(\hat{\boldsymbol{C}}_t, \boldsymbol{H}_t^*)\right] \leq \mathbb{E}\left[L(\hat{\boldsymbol{C}}_t, \tilde{\boldsymbol{H}}_t)\right]$$

L is a robust loss function, by (2.3.1), we know

$$\mathbb{E}\left[L(\boldsymbol{C}_t, \boldsymbol{H}_t^*)\right] \leq \mathbb{E}\left[L(\boldsymbol{C}_t, \tilde{\boldsymbol{H}}_t)\right] \tag{2.3.2}$$

By assumption, we know $L(\boldsymbol{C}, \boldsymbol{H})$ has a unique minimum when $\boldsymbol{H} = \boldsymbol{C}$. So

$$L(\boldsymbol{C}_t, \boldsymbol{C}_t) \leq L(\boldsymbol{C}_t, \boldsymbol{H}_t^*) \quad \text{and} \quad \mathbb{E}[L(\boldsymbol{C}_t, \boldsymbol{C}_t)] \leq \mathbb{E}[L(\boldsymbol{C}_t, \boldsymbol{H}_t^*)]$$

Also, $\boldsymbol{C}_t$ is $\mathscr{F}_{t-1}$-measurable, setting $\tilde{\boldsymbol{H}}_t = \boldsymbol{C}_t$ in (2.3.2) gives

$$\mathbb{E}\left[L(\boldsymbol{C}_t, \boldsymbol{H}_t^*)\right] \leq \mathbb{E}\left[L(\boldsymbol{C}_t, \boldsymbol{C}_t)\right]$$

Hence, we must have

$$\mathbb{E}\left[L(\boldsymbol{C}_t, \boldsymbol{H}_t^*)\right] = \mathbb{E}\left[L(\boldsymbol{C}_t, \boldsymbol{C}_t)\right] \quad \text{and} \quad \boldsymbol{H}_t^* = \boldsymbol{C}_t.$$

□

One might wonder about the exact form of robust loss functions. We start with one dimension when $N = 1$. In one dimensional setting where we denote the volatility proxy as $\hat{\sigma}^2$ and h for the forecast, common robust loss functions include the MSE and QLIKE loss functions, which are given by

$$\text{MSE}: \quad L(\hat{\sigma}^2, h) = (\hat{\sigma}^2 - h)^2$$

$$\text{QLIKE}: \quad L(\hat{\sigma}^2, h) = \log h + \frac{\hat{\sigma}^2}{h}$$

Patton [15] showed that the entire subset of robust and homogeneous loss functions when $N = 1$ is indexed by a scalar parameter b. We present the results below.

Definition 2.3.4. A loss function L is homogeneous of order k if for some $k \in \mathbb{R}$,

$$L(\alpha\hat{\sigma}^2, \alpha h) = \alpha^k L(\hat{\sigma}^2, h)$$

holds for any $\alpha > 0$.

Proposition 2.3.5. *Let $\mathscr{H}$ be a compact subset of $\mathbb{R}_{++}$, which is the positive part of the real line. The entire subset of robust and homogeneous loss functions $L : \mathbb{R}_+ \times \mathscr{H} \to \mathbb{R}_+$ is given by*

$$L\left(\hat{\sigma}^2, h; b\right) = \begin{cases} \frac{1}{(b+1)(b+2)}\left(\hat{\sigma}^{2b+4} - h^{b+2}\right) - \frac{1}{b+1}h^{b+1}\left(\hat{\sigma}^2 - h\right), & \text{for } b \notin \{-1, -2\} \\ h - \hat{\sigma}^2 + \hat{\sigma}^2 \log \frac{\hat{\sigma}^2}{h}, & \text{for } b = -1 \\ \frac{\hat{\sigma}^2}{h} - \log \frac{\hat{\sigma}^2}{h} - 1, & \text{for } b = -2 \end{cases}$$

2.3.2 Choice of Loss Function

We extend the results in proposition 2.3.5 to the multidimensional scenario by replacing $\hat{\sigma}^2$ with $\hat{\boldsymbol{C}}$ and replacing h with $\boldsymbol{H}$. In particular, we choose $b = -\frac{3}{2}$, then we obtain the expression

$$\begin{aligned}
&-4(\hat{\boldsymbol{C}}^{\frac{1}{2}} - \boldsymbol{H}^{\frac{1}{2}}) + 2\boldsymbol{H}^{-\frac{1}{2}}(\hat{\boldsymbol{C}} - \boldsymbol{H}) \\
&= -4\hat{\boldsymbol{C}}^{\frac{1}{2}} + 2(\boldsymbol{H}^{\frac{1}{2}} + \boldsymbol{H}^{-\frac{1}{2}}\hat{\boldsymbol{C}})
\end{aligned}$$

Remark 2.3.6. In general, a matrix can have several square roots. The positive definite square root of a symmetric positive definite matrix is well-defined and unique. Suppose the eigendecomposition of $\boldsymbol{H}$ is

$$\boldsymbol{H} = \boldsymbol{Q}\boldsymbol{D}\boldsymbol{Q}^{\intercal},$$

where columns of $\boldsymbol{Q}$ are normalised orthogonal eigenvectors and $\boldsymbol{D}$ has a diagonal of corresponding eigenvalues with zeros elsewhere. Positive definiteness of $\boldsymbol{H}$ indicates that all eigenvalues are positive, so we take

$$\boldsymbol{D}^{\frac{1}{2}} = \begin{bmatrix} \sqrt{\lambda_1} & 0 & 0 & \cdots & 0 \\ 0 & \sqrt{\lambda_2} & 0 & \cdots & 0 \\ \vdots & & \ddots & & \vdots \\ 0 & \cdots & 0 & \sqrt{\lambda_{N-1}} & 0 \\ 0 & \cdots & 0 & 0 & \sqrt{\lambda_N} \end{bmatrix}$$

and define

$$\boldsymbol{H}^{\frac{1}{2}} := \boldsymbol{Q}\boldsymbol{D}^{\frac{1}{2}}\boldsymbol{Q}^{\intercal}$$

By orthogonality of $\boldsymbol{Q}$, this yields

$$\boldsymbol{H}^{\frac{1}{2}}\boldsymbol{H}^{\frac{1}{2}} = \boldsymbol{Q}\boldsymbol{D}^{\frac{1}{2}}\boldsymbol{Q}^{\intercal}\boldsymbol{Q}\boldsymbol{D}^{\frac{1}{2}}\boldsymbol{Q}^{\intercal} = \boldsymbol{Q}\boldsymbol{D}^{\frac{1}{2}}\boldsymbol{D}^{\frac{1}{2}}\boldsymbol{Q}^{\intercal} = \boldsymbol{H}.$$

The loss function maps matrices to the positive real line, hence, we need a mapping $f : \mathbf{S}^N_{++} \mapsto \mathbb{R}_+$ that measures the distance in some sense. We choose f to be the trace function because

- It is a linear operator which is easy to manipulate;
- It represents the sum of all eigenvalues;
- As we will show later, it makes L satisfy assumption 2.3.2.

Therefore, we consider a loss function of the form

$$\begin{aligned}
L(\hat{\boldsymbol{C}}, \boldsymbol{H}) &= \boldsymbol{Tr}(-4\hat{\boldsymbol{C}}^{\frac{1}{2}} + 2(\boldsymbol{H}^{\frac{1}{2}} + \boldsymbol{H}^{-\frac{1}{2}}\hat{\boldsymbol{C}})) \\
&= -4\boldsymbol{Tr}(\hat{\boldsymbol{C}}^{\frac{1}{2}}) + 2\boldsymbol{Tr}(\boldsymbol{H}^{\frac{1}{2}} + \boldsymbol{H}^{-\frac{1}{2}}\hat{\boldsymbol{C}}))
\end{aligned} \tag{2.3.3}$$

Discard the first term $-4\boldsymbol{Tr}(\hat{\boldsymbol{C}}^{\frac{1}{2}})$ which doesn't depend on $\boldsymbol{H}$, and scale down the second term by a factor of a half, we arrive at

$$L(\hat{\boldsymbol{C}}, \boldsymbol{H}) = \boldsymbol{Tr}(\boldsymbol{H}^{\frac{1}{2}} + \boldsymbol{H}^{-\frac{1}{2}}\hat{\boldsymbol{C}}) \tag{2.3.4}$$

Changing from (2.3.3) to (2.3.4) doesn't affect the measure of accuracy and the choice of optimal forecast.

To verify that our loss function is indeed robust and the assumption of unique minimum at $\boldsymbol{H} = \boldsymbol{C}$ is satisfied, we present the following proofs.

Proposition 2.3.7. *The loss function* $L(\hat{\boldsymbol{C}}, \boldsymbol{H}) = \boldsymbol{Tr}(\boldsymbol{H}^{\frac{1}{2}} + \boldsymbol{H}^{-\frac{1}{2}}\hat{\boldsymbol{C}})$ *is robust and homogeneous.*

Proof. Let $\hat{\boldsymbol{C}}_t$ be a conditionally unbiased estimator of $\boldsymbol{C}_t$ such that $\mathbb{E}_{t-1}[\hat{\boldsymbol{C}}_t] = \boldsymbol{C}_t$. Then we can decompose the proxy $\hat{\boldsymbol{C}}_t$ as

$$\hat{\boldsymbol{C}}_t = \boldsymbol{C}_t + \boldsymbol{\epsilon}_t$$

where $\boldsymbol{\epsilon}_t \in \mathbf{S}^N$ and $E_{t-1}[\boldsymbol{\epsilon}_t] = \mathbf{0}$, i.e. the conditional mean of the estimation error is a zero matrix. Substituting the above decomposition into the expected loss gives

$$\begin{aligned}
\mathbb{E}[L(\hat{\boldsymbol{C}}_t, \boldsymbol{H}_t)] &= \mathbb{E}[\boldsymbol{Tr}(\boldsymbol{H}_t^{\frac{1}{2}} + \boldsymbol{H}_t^{-\frac{1}{2}}\hat{\boldsymbol{C}})] \\
&= \mathbb{E}[\boldsymbol{Tr}(\boldsymbol{H}_t^{\frac{1}{2}} + \boldsymbol{H}_t^{-\frac{1}{2}}(\boldsymbol{C}_t + \boldsymbol{\epsilon}_t)] \\
&= \mathbb{E}[\boldsymbol{Tr}(\boldsymbol{H}_t^{\frac{1}{2}} + \boldsymbol{H}_t^{-\frac{1}{2}}\boldsymbol{C}_t)] + \mathbb{E}[\boldsymbol{Tr}(\boldsymbol{H}_t^{-\frac{1}{2}}\boldsymbol{\epsilon}_t)],
\end{aligned}$$

where in the last line we used the linearity of expectation and the trace function. The first term is equal to $\mathbb{E}[L(\boldsymbol{C}_t, \boldsymbol{H}_t)]$, while the second term can be simplified by the fact that expectation and trace commute (which is true because trace is a linear operator):

$$\begin{aligned}
\mathbb{E}\left[\boldsymbol{Tr}(\boldsymbol{H}_t^{-\frac{1}{2}}\boldsymbol{\epsilon}_t)\right] &= \boldsymbol{Tr}\left(\mathbb{E}[\boldsymbol{H}_t^{-\frac{1}{2}}\boldsymbol{\epsilon}_t]\right) \\
&= \boldsymbol{Tr}\left(\mathbb{E}\left[\mathbb{E}_{t-1}[\boldsymbol{H}_t^{-\frac{1}{2}}\boldsymbol{\epsilon}_t]\right]\right) \\
&= \boldsymbol{Tr}\left(\mathbb{E}\left[\boldsymbol{H}_t^{-\frac{1}{2}}\underbrace{\mathbb{E}_{t-1}[\boldsymbol{\epsilon}_t]}_{\mathbf{0}}\right]\right) \\
&= \boldsymbol{Tr}(\mathbf{0}) = 0
\end{aligned}$$

where we make use of the tower property in the second line and that $\boldsymbol{H}_t$ is $\mathscr{F}_{t-1}$-measurable in the third line. Therefore, we arrive at

$$\mathbb{E}[L(\hat{\boldsymbol{C}}_t, \boldsymbol{H}_t)] = \mathbb{E}[L(\boldsymbol{C}_t, \boldsymbol{H}_t)].$$

Robustness follows immediately by (2.3.1).

Homogeneity is straightforward by the linearity of both the trace operator and the expectation:

$$L(\boldsymbol{\alpha}\boldsymbol{C}, \boldsymbol{\alpha}\boldsymbol{H}) = (\boldsymbol{\alpha}\boldsymbol{H}_t)^{\frac{1}{2}} + (\boldsymbol{\alpha}\boldsymbol{H}_t)^{-\frac{1}{2}}(\boldsymbol{\alpha}\boldsymbol{C}_t) = \boldsymbol{\alpha}^{\frac{1}{2}}L(\boldsymbol{C}, \boldsymbol{H})$$

for any $\boldsymbol{\alpha} \in \mathbf{S}_{++}^N$ and we identify $k = \frac{1}{2}$. □

Proposition 2.3.8. $L(\boldsymbol{C}, \cdot\,)(\boldsymbol{H}) = \boldsymbol{Tr}(\boldsymbol{H}^{\frac{1}{2}} + \boldsymbol{H}^{-\frac{1}{2}}\boldsymbol{C})$ *has a unique minimum at* $\boldsymbol{H} = \boldsymbol{C}$.

Proof. Following remark 2.3.6, we can define the power of $\boldsymbol{H}$ as

$$\boldsymbol{H}^{\alpha} := \boldsymbol{Q}\boldsymbol{D}^{\alpha}\boldsymbol{Q}^{\intercal}$$

for any $\alpha \in \mathbb{R}$. This is well-defined because the eigenvalues of $\boldsymbol{H}$ are all positive.

We employ the idea of completing squares. Note that both $\boldsymbol{H}^{\alpha}$ and $\boldsymbol{C}^{\beta}$ are symmetric for any $\alpha, \beta \in \mathbb{R}$, we can write

$$\begin{aligned}
\boldsymbol{A} :&= \boldsymbol{H}^{\frac{1}{2}} + \boldsymbol{H}^{-\frac{1}{2}}\boldsymbol{C} \\
&= \boldsymbol{H}^{\frac{1}{2}} - 2\boldsymbol{C}^{\frac{1}{2}} + \boldsymbol{H}^{-\frac{1}{2}}\boldsymbol{C} + 2\boldsymbol{C}^{\frac{1}{2}} \\
&= \left(\boldsymbol{H}^{\frac{1}{4}} - \boldsymbol{H}^{-\frac{1}{4}}\boldsymbol{C}^{\frac{1}{2}}\right)^{\intercal}\left(\boldsymbol{H}^{\frac{1}{4}} - \boldsymbol{H}^{-\frac{1}{4}}\boldsymbol{C}^{\frac{1}{2}}\right) - \boldsymbol{C}^{\frac{1}{2}}\boldsymbol{H}^{-\frac{1}{2}}\boldsymbol{C}^{\frac{1}{2}} + \boldsymbol{H}^{-\frac{1}{2}}\boldsymbol{C} + 2\boldsymbol{C}^{\frac{1}{2}}.
\end{aligned}$$

Let $\boldsymbol{B} := \boldsymbol{H}^{\frac{1}{4}} - \boldsymbol{H}^{-\frac{1}{4}}\boldsymbol{C}^{\frac{1}{2}}$, then applying the trace gives

$$L = \boldsymbol{Tr}(\boldsymbol{A}) = \boldsymbol{Tr}\left(\boldsymbol{B}^{\intercal}\boldsymbol{B}\right) + 2\boldsymbol{Tr}(\boldsymbol{C}^{\frac{1}{2}}).$$

The two middle terms vanish because

$$\boldsymbol{Tr}(\boldsymbol{C}^{\frac{1}{2}}\boldsymbol{H}^{-\frac{1}{2}}\boldsymbol{C}^{\frac{1}{2}}) = \boldsymbol{Tr}(\boldsymbol{H}^{-\frac{1}{2}}\boldsymbol{C}).$$

The second term is independent of $\boldsymbol{H}$, while the first term is non-negative because $\boldsymbol{B}^{\intercal}\boldsymbol{B}$ is positive semi-definite. The loss function is minimised with respect to $\boldsymbol{H}$ if and only if the first term is zero, i.e.

$$\boldsymbol{Tr}\left(\boldsymbol{B}^{\intercal}\boldsymbol{B}\right) = 0 \quad \iff \quad \boldsymbol{B}^{\intercal}\boldsymbol{B} = \boldsymbol{0}$$

This further yields

$$\boldsymbol{B} = \boldsymbol{0} \quad \iff \quad \boldsymbol{H} = \boldsymbol{C}$$

Hence, the unique minimum occurs at $\boldsymbol{H} = \boldsymbol{C}$.

□

Chapter 3

Forecasts Applied to Simulated Data

In this chapter, we run various forecasts on simulated return data. For each forecast, we first determine the parameter (for example, we decide on a most suitable look back period when using the Ledoit and Wolf's shrinkage approach, and the optimal value of α when EWMA forecast is employed.) Results from running the forecasts include the loss value series over the forecast period and the averaged loss values for each forecast method with covariance proxy and with true covariance matrix. The loss differentials are passed into the DM test to come up with a ranking of the forecasts.

3.1 Simulation Schemes

We use two sets of simulations. The first one borrows the idea from the single-index model by Sharpe [18], in which we assume the simulated index returns are independent and identically distributed across time. The second involves the Generalized AutoRegressive Conditional Heteroskedasticity (GARCH) process that models volatility clustering and persistence in financial returns so that the simulated returns have a non-trivial serial correlation and evolving covariance matrix. The details are elaborated as follows.

3.1.1 IID Simulation

This simulation framework is built upon the single-index model (The Diagonal Model) proposed by Sharpe [18], which is also mentioned in section 1.2.1 in our introduction of the Market Shrinkage. Suppose we have N stocks, each has a return time series of length T. Let x_i be the log return of the ith stock, it is assumed that

$$x_i = \alpha_i + \beta_i x_0 + \epsilon_i,$$

where α_i, β_i are deterministic parameters, x_0 is the market index return and ϵ_i is the noise term with zero mean. Different stock returns are related only through the common market index, which is thought to have a major impact on stock returns.

If we also include the time evolution of the returns, we have

$$x_{it} = \alpha_i + \beta_i x_{0t} + \epsilon_{it}.$$

Note that the noise ϵ_{it} is uncorrelated across time, stock, and the market index. The market index x_{0t} is also assumed to be i.i.d. through time.

To simulate the stock return time series, we first simulate the market index x_0. Financial returns tend to have fat tails and market index returns are very close to zero on average, so let the index follow a scaled t-distribution with a degree of freedom ν_0.

$$x_{0t} \sim \gamma_0 t_{\nu_0}$$

To determine the scaling factor γ_0, we take a reference index, the S&P500 index log returns from 2012 to 2022 and find the sample variance $\hat{\sigma}_0^2$ of the 10-year time series. It is desired that the variance of our constructed market index matches $\hat{\sigma}_0^2$. Also, we know $\text{Var}(x_{0t}) = \gamma_0^2 \frac{\nu_0}{\nu_0 - 2}$ Then, we choose

$$\gamma_0 = \frac{\hat{\sigma}_0}{\sqrt{\frac{\nu_0}{\nu_0 - 2}}} \tag{3.1.1}$$

Next, we simulate the market index by generating an i.i.d. sequence of length T of t-distributed random variables of the degree of freedom ν_0 multiplied by γ_0. The degree of freedom is chosen as $\nu_0 = 5$ in the actual simulation. Denote the simulated market index series as $\tilde{x}_0$.

Then, we talk about how to simulate individual stock returns. There are two major steps:

1. Determine the values of α_i, β_i for $i = 1, \cdots, N$;
2. Specify a distribution for ϵ_i for $i = 1, \cdots, N$, and simulate the noise time series.

For 1, We take recent 10 years of stock log returns of the first N stocks listed in the S&P500 stock index, and for each stock, run a regression of the returns against the S&P500 index log returns. Results from the regression show the fitted $\hat{\alpha}_i$ of the order 10e-4, hence we can treat the stock log returns to have zero mean. Besides, the sample variance of the fitted residual $\hat{\epsilon}_{it} := x_{it} - \hat{\alpha}_i - \hat{\beta}_i x_{0t}$ is also calculated.

For 2, suppose the noise term follows a scaled t-distribution as well, that is, for $i = 1, \cdots, N$,

$$\epsilon_{it} \sim \gamma_i t_{\nu_i}$$

we take $\nu_i := \nu_1$ to be constant across stocks to facilitate the data simulation, then $\text{Var}(\epsilon_{it}) = \gamma_i^2 \frac{\nu_1}{\nu_1 - 2}$. The choice of scaling factors γ_i is similar to the choice of γ_0, which makes the variance of the simulated noise time series match that of the fitted residual $\hat{\epsilon}_i$. The simulated noise term for stock i at time t is denoted by $\tilde{\epsilon}_{it}$, and the whole time series of length T is constructed by multiplying γ_i with a simulated i.i.d. sequence of t-distributed random variables, the degree of freedom of which is ν_1.

Putting everything together, we have the simulated return time series:

$$\tilde{x}_{it} = \hat{\beta}_i \tilde{x}_{0t} + \tilde{\epsilon}_{it}$$

for $i = 1, \cdots, N$ and $t = 1, \cdots, T$. Note that we omit the $\hat{\alpha}_i$ term because it is negligible in magnitude.

Let $\tilde{\boldsymbol{x}}_t = [\tilde{x}_{1t}, \tilde{x}_{2t}, \ldots, \tilde{x}_{Nt}]^\intercal$, $\hat{\boldsymbol{\beta}} = [\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_N]^\intercal$, and $\tilde{\boldsymbol{\epsilon}}_t = [\tilde{\epsilon}_{1t}, \tilde{\epsilon}_{2t}, \ldots, \tilde{\epsilon}_{Nt}]^\intercal$, we can then put the individual stock returns into a multivariate vector

$$\tilde{\boldsymbol{x}}_t = \hat{\boldsymbol{\beta}} \tilde{x}_{0t} + \tilde{\boldsymbol{\epsilon}}_t.$$

Eventually, we have an i.i.d. sequence of multivariate random variables, each has zero mean and contemporaneous covariance

$$\boldsymbol{C}_t = \mathbb{E}[\tilde{\boldsymbol{x}}_t \tilde{\boldsymbol{x}}_t^\intercal] = \mathbb{E}[(\hat{\boldsymbol{\beta}} \tilde{x}_{0t} + \tilde{\boldsymbol{\epsilon}}_t)(\hat{\boldsymbol{\beta}} \tilde{x}_{0t} + \tilde{\boldsymbol{\epsilon}}_t)^\intercal] = \hat{\boldsymbol{\beta}} \hat{\boldsymbol{\beta}}^\intercal \mathrm{Var}(\tilde{x}_{0t}) + \mathbb{E}[\tilde{\boldsymbol{\epsilon}}_t \tilde{\boldsymbol{\epsilon}}_t^\intercal]$$

By (3.1.1), we have $\mathrm{Var}(\tilde{x}_{0t}) = \hat{\sigma}_0^2$. Let $\sigma_1^2 := \frac{\nu_1}{\nu_1 - 2}$, we can write

$$\mathbb{E}[\tilde{\boldsymbol{\epsilon}}_t \tilde{\boldsymbol{\epsilon}}_t^\intercal] = \sigma_1^2 \begin{bmatrix} \gamma_1^2 & 0 & 0 & \cdots & 0 \\ 0 & \gamma_2^2 & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & \gamma_{N-1}^2 & 0 \\ 0 & 0 & \cdots & 0 & \gamma_N^2 \end{bmatrix} := \sigma_1^2 \boldsymbol{\Gamma}$$

Therefore, we get an expression for the true covariance matrix

$$\boldsymbol{C} := \boldsymbol{C}_t = \hat{\boldsymbol{\beta}} \hat{\boldsymbol{\beta}}^\intercal \hat{\sigma}_0^2 + \sigma_1^2 \boldsymbol{\Gamma} \tag{3.1.2}$$

and this is constant in time. The expression (3.1.2) is useful, because in running the covariance forecasts, the true covariance is known. Hence, the forecasts can be compared against the true covariance in the loss function.

The advantage of this simulation model lies in its ease of implementation. However, the i.i.d. returns give a covariance matrix constant in time. This lack of dynamic in the covariance doesn't agree with stylised facts in financial returns as it fails to explain volatility clustering and persistence. To incorporate this feature in the simulated returns, we employ a more complicated tool – GARCH simulation.

3.1.2 GARCH Simulation

One-Dimensional Process

Definition 3.1.1. We say a process $(X_t)_{t \in \mathbb{Z}}$ is strictly stationary if $(X_{t_1}, \ldots, X_{t_n})$ have equal distribution to $(X_{t_1+k}, \ldots, X_{t_n+k})$ for any $t_1, \ldots, t_n \in \mathbb{Z}, k \in \mathbb{Z}$.

Definition 3.1.2. Let $(X_t)_{t \in \mathbb{Z}}$ be a square-integrable process which is i.i.d. through time, then it is called a strict white noise. Besides, if $\mathbb{E}[X_t] = 0$ and $\mathrm{Var}(X_t) = \sigma^2$, we denote $(X_t)_{t \in \mathbb{Z}} \sim \mathrm{SWN}(0, \sigma^2)$.

Definition 3.1.3. Let $(X_t)_{t\in\mathbb{Z}}$ be a strictly stationary process. If we can write

$$X_t = \sigma_t Z_t, \quad \sigma_t^2 = \alpha_0 + \sum_{i=1}^{p} \alpha_i X_{t-i}^2 + \sum_{j=1}^{q} \beta_j \sigma_{t-j}^2, \quad t \in \mathbb{Z},$$

where $\alpha_0 > 0, \alpha_1, \ldots, \alpha_p \geq 0$, and $\beta_1, \ldots, \beta_p \geq 0$, for some $(Z_t)_{t\in\mathbb{Z}} \sim \text{SWN}(0,1)$ and $(\sigma_t)_{t\in\mathbb{Z}}$ that is strictly stationary and positive-valued, then we say $(X_t)_{t\in\mathbb{Z}}$ is a GARCH(p,q) process.

In financial applications, we usually model returns $(X_t)_{t\in\mathbb{Z}}$ using a GARCH(1,1) process, so that

$$X_t = \sigma_t Z_t$$

for some $(Z_t)_{t\in\mathbb{Z}} \sim \text{SWN}(0,1)$, and σ_t satisfies

$$\sigma_t^2 = \alpha_0 + \alpha_1 X_{t-1}^2 + \beta_1 \sigma_{t-1}^2 \tag{3.1.3}$$

From the above expression (3.1.3), we can see that σ_t is $\mathscr{F}_{t-1}$-measurable. The conditional mean and variance of the process are

$$\mathbb{E}[X_t|\mathscr{F}_{t-1}] = \sigma_t \mathbb{E}[Z_t|\mathscr{F}_{t-1}] = \sigma_t \mathbb{E}[Z_t] = 0$$

$$\text{Var}[X_t|\mathscr{F}_{t-1}] = \mathbb{E}[X_t^2|\mathscr{F}_{t-1}] - \mathbb{E}[X_t|\mathscr{F}_{t-1}]^2 = \sigma_t^2 \mathbb{E}[Z_t^2|\mathscr{F}_{t-1}] = \sigma_t^2 \mathbb{E}[Z_t^2] = \sigma_t^2$$

In [20, Part 3, slides 86-94], it has been shown in model diagnostics[1] that assuming a t-distribution for F_Z makes the fitted standardised residuals $\hat{Z}_t := \frac{X_t}{\hat{\sigma}}$ behave most like i.i.d. samples from the chosen F_Z. Given a return time series $X_0, X_1, \cdots, X_T$, the parameters α_0, α_1 and β_1 in (3.1.3) are estimated using maximum likelihood after specifying a distribution F_Z for $(Z_t)_{t\in\mathbb{Z}} \sim \text{SWN}(0,1)$. Once we have the fitted parameters and proxy for the starting value $\hat{\sigma}_0^2$, the fitted conditional variance $\hat{\sigma}_t^2$ can be worked out recursively

$$\hat{\sigma}_t^2 = \hat{\alpha}_0 + \hat{\alpha}_1 X_{t-1}^2 + \hat{\beta}_1 \hat{\sigma}_{t-1}^2$$

for $t = 1, \cdots, T$. These are called the GARCH-fitted volatilities. Suppose we now want to forecast the conditional volatility at $T+1$, this one-step-forward forecast comes naturally as

$$\hat{\sigma}_{T+1}^2 = \hat{\alpha}_0 + \hat{\alpha}_1 X_T^2 + \hat{\beta}_1 \hat{\sigma}_T^2.$$

We can also simulate returns following a GARCH model of fitted parameters. For example, we have the fitted degree of freedom for the t-distributed white noise, so that we can simulate an independent sample Z_{T+1}, then the simulated return at $T+1$ is

$$X_{T+1} := \hat{\sigma}_{T+1} Z_{T+1}.$$

We can continue to find the simulated $\hat{\sigma}_{T+2}$ using (3.1.3), simulate another independent Z_{T+2} and so on. This way we have simulated returns that are serially correlated. In implementation, both the fitting and simulation of GARCH process are done using the `arch5.3.1`[2] package of `Python`.

[1] through theoretical pdf plotted with histogram, Q-Q plot and the ACF plot

[2] https://bashtage.github.io/arch/univariate/univariate.html

Multi-Dimensional Process

When considering multiple assets, we deal with multivariate time series modelled by a multivariate stochastic process $(\boldsymbol{X}_t)_{t\in\mathbb{Z}}$, where each $\boldsymbol{X}_t$ is an N-dimensional random vector. We extend the definitions for the one-dimensional process to give the following key definition:

Definition 3.1.4. Let $(\boldsymbol{Z}_t)_{t\in\mathbb{Z}}$ be an N-dimensional strict white noise with zero mean and contemporaneous covariance $\mathbb{E}[\boldsymbol{Z}_t\boldsymbol{Z}_t^\intercal] = \boldsymbol{I}_N$. Then, we say $(\boldsymbol{X}_t)_{t\in\mathbb{Z}}$ is a multivariate GARCH process if it is strictly stationary and satisfies

$$\boldsymbol{X}_t = \boldsymbol{C}_t^{\frac{1}{2}}\boldsymbol{Z}_t$$

where $\boldsymbol{C}_t$ is a symmetric positive definite matrix that is random, and it is measurable with respect to $\mathscr{F}_{t-1} = \sigma\{\boldsymbol{X}_1, \cdots, \boldsymbol{X}_{t-1}\}$.

The conditional mean and covariance are given respectively by

$$\mathbb{E}[\boldsymbol{X}_t|\mathscr{F}_{t-1}] = \mathbb{E}[\boldsymbol{C}_t^{\frac{1}{2}}\boldsymbol{Z}_t|\mathscr{F}_{t-1}] = \boldsymbol{C}_t^{\frac{1}{2}}\mathbb{E}[\boldsymbol{Z}_t|\mathscr{F}_{t-1}] = \boldsymbol{0}$$

$$\mathbf{Cov}[\boldsymbol{X}_t|\mathscr{F}_{t-1}] = \mathbb{E}[\boldsymbol{X}_t\boldsymbol{X}_t^\intercal|\mathscr{F}_{t-1}] = \boldsymbol{C}_t^{\frac{1}{2}}\underbrace{\mathbb{E}[\boldsymbol{Z}_t\boldsymbol{Z}_t^\intercal]}_{\boldsymbol{I}_N}\boldsymbol{C}_t^{\frac{1}{2}} = \boldsymbol{C}_t$$

Let $\boldsymbol{X} \in \mathbb{R}^{N\times T}$ represent returns of N stocks over a period of T days, where the columns $(\boldsymbol{X}_t)_{t=1,\cdots,T}$ are assumed to follow a multivariate GARCH process. The white noise $(\boldsymbol{Z}_t)_{t=1,\cdots,T}$ are assumed to follow a multivariate t-distribution with components being i.i.d. t-distributed random variables. We decompose the conditional covariance into individual stock's conditional variances and the conditional correlation matrix $\boldsymbol{P}_t$ as

$$\boldsymbol{C}_t = \boldsymbol{D}_t\boldsymbol{P}_t\boldsymbol{D}_t$$

where $\boldsymbol{D}_t = \mathrm{diag}(\sigma_{1,t}, \cdots, \sigma_{N,t})$ consists of standard deviations of individual stock returns. Note that $\sigma_{i,t}^2 = \boldsymbol{C}_{t,ii}$. The individual stock returns can be modelled by one-dimensional GARCH process discussed above, for the modelling of $\boldsymbol{P}_t$, we focus on the constant conditional correlation model.

- **Equi-Correlation**
  We assume all stocks have equal correlation, similar to the constant correlation model used for the shrinkage target in section 1.2.1. Given a matrix $\boldsymbol{X}$ of historical returns of N assets and T days, we can calculate the sample correlation matrix. The constant correlation $\overline{r}$ is taken as the average of all pairwise correlations, and we use

$$\tilde{\boldsymbol{P}}_E := \begin{bmatrix} 1 & \overline{r} & \overline{r} & \cdots & \overline{r} \\ \overline{r} & 1 & \overline{r} & \cdots & \overline{r} \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \overline{r} & \cdots & \overline{r} & 1 & \overline{r} \\ \overline{r} & \overline{r} & \cdots & \overline{r} & 1 \end{bmatrix}$$

  in the simulation of stock returns.

- **Constant Correlation**
  This model is based on the CCC model of Bollerslev[1]. Assume $\boldsymbol{P}_t := \boldsymbol{P}_C$ is constant, we can write

$$\boldsymbol{X}_t = \boldsymbol{C}_t^{\frac{1}{2}} \boldsymbol{Z}_t = \boldsymbol{D}_t \underbrace{\boldsymbol{P}_C^{\frac{1}{2}} \boldsymbol{Z}_t}_{:= \boldsymbol{M}_t}$$

  where $\boldsymbol{M}_t$ is a strict white noise with contemporaneous covariance $\boldsymbol{P}_C$. The following steps are taken to obtain an estimate of $\boldsymbol{P}_C$.

  1. Fit GARCH(1,1) process to each of $\boldsymbol{X}_t$'s components and we come up with a sequence of fitted volatilities $(\hat{\sigma}_{i,t})_{t=1,\cdots,T}$ for $i = 1, \cdots, N$, and the corresponding diagonal matrix of fitted volatility is $\hat{\boldsymbol{D}}_t$.
  2. Compute the devolatised process $\hat{\boldsymbol{Y}}_t := \hat{\boldsymbol{D}}_t^{-1} \boldsymbol{X}_t$ for $t = 1, \cdots, N$. If the individual components indeed follow GARCH(1,1) process, the devolatised process should be a strict white noise with covariance $\boldsymbol{P}_C$.
  3. Estimate $\boldsymbol{P}_C$ by the sample covariance matrix of $\tilde{\boldsymbol{P}}_C = \frac{1}{T} \sum_{t=1}^{T} \hat{\boldsymbol{Y}}_t \hat{\boldsymbol{Y}}_t^{\intercal}$.

Simulation Procedure

1. Take the most recent five years of returns of the first N stocks listed in the S&P500 stock index and fit a GARCH(1,1) process to the return time series of each individual stock.[3]

2. The fitting results give the fitted residuals and parameters, which can be passed into the `simulate`[4] function to give simulated GARCH process of length T. And these give the simulated diagonal matrices $\hat{\boldsymbol{D}}_t = \text{diag}(\hat{\sigma}_{1,t}, \cdots, \hat{\sigma}_{N,t})$.

3. Calculate the estimated correlation matrix $\tilde{\boldsymbol{P}}_E$ or $\tilde{\boldsymbol{P}}_C$ depending on which correlation model we choose. Both correlation estimates are constant, which is not realistic in modelling actual stock returns, but still serves as an effective way to include serial correlation in simulated returns.

4. Simulate the N-dimension multivariate t-distributed $\tilde{\boldsymbol{Z}}_t$ with each component an i.i.d. t-distributed random variable of a specified degree of freedom, which is chosen to be 10 in our implementation.

5. The simulated returns are

$$\tilde{\boldsymbol{X}}_t = \hat{\boldsymbol{D}}_t \tilde{\boldsymbol{P}}^{\frac{1}{2}} \tilde{\boldsymbol{Z}}_t$$

   for $t = 1, \cdots, T$. where $\tilde{\boldsymbol{P}}$ is either $\tilde{\boldsymbol{P}}_E$ or $\tilde{\boldsymbol{P}}_C$.

[3] The stock returns have a sample mean very close to zero, so we fitted a zero mean ARCH model(which is equivalent to a GARCH model).

[4] https://bashtage.github.io/arch/univariate/univariate_volatility_modeling.html#Simulation

3.2 Implementation of Forecasts

3.2.1 Details of Implementation

We choose $N = 100, 200, 300$ and 400, and the simulation length is $T = 2000$, which is approximately 10 years of data. The forecast schemes applied are

- Sample Covariance Matrix(referred to as SCM) with and without filtering[5];
- EWMA forecast with and without filtering;
- Ledoit and Wolf shrinkage estimators[6] with the identity target, the equi-correlation target and the market index target.

Both the SCM forecast and the shrinkage forecasts are carried out on a rolling window basis, so they require an input value for the look-back period, which is the number of past observations used in the sample covariance matrix calculation. The length of simulated data is T, and let's denote the look back period as l, then the forecasts start on the $l+1$th day and the whole forecast period is $T-l$ days. During the forecast period, the forecast $\boldsymbol{H}_t$ on day t is produced using information up until day $t-1$, and the covariance proxy is taken to be $\hat{\boldsymbol{C}}_t := \boldsymbol{X}_t \boldsymbol{X}_t^\intercal$. The value of the loss function given by (2.3.4) is calculated and stored for each day in the forecast period, and the averaged loss value is also recorded. The algorithm is outlined as follows:

Algorithm 1: The algorithm to compute rolling-window forecasts using T observations and look back period l, note that $l < T$ is required.

Input: look back period l, return matrix $\boldsymbol{X} \in \mathbb{R}^{N \times T}$ (the shrinkage target type if forecast method is shrinkage)

Output: A vector of loss values and averaged loss value

for $t = l+1, \cdots, T$ **do**

- $\boldsymbol{X}_{obs} = \boldsymbol{X}[:, t-l : t-1]$
- `/* This denotes the` $(t-l)th$ `to` $(t-1)th$ `column of` $\boldsymbol{X}$ `*/`
- $\hat{\boldsymbol{C}}_t = \boldsymbol{X}[:,t]\boldsymbol{X}[:,t]^\intercal$
- **if** *SCM Forecast* **then**
  - $\boldsymbol{H}_t = \frac{1}{l}\boldsymbol{X}_{obs}\boldsymbol{X}_{obs}^\intercal$
- **else**
  - $\boldsymbol{H}_t =$ `shrinkage`$(\boldsymbol{X}_{obs})$
- $L_t = L(\hat{\boldsymbol{C}}_t, \boldsymbol{H}_t)$
- Append the value L_t to the loss value series

Averaged loss value $\overline{L}(\hat{\boldsymbol{C}}) = \frac{1}{T-l}\sum_{t=l+1}^{T} L(\hat{\boldsymbol{C}}_t, \boldsymbol{H}_t)$ is the mean of the loss value series

Theoretically, larger l will denoise the sample covariance better, while at the same time diminish the dynamics of an evolving covariance by estimating it assuming it's

[5]The effects of LCPB and PG+ filtering are comparable, hence, we stick the PG+ type of filtering in our implementation.

[6]Source codes for implementing the three shrinkage methods are from https://github.com/pald22/covShrinkage, and we adapted the output format.

constant over a long period of time. Also, we need to take into account the limited availability of data for some stocks(such as newly listed ones). Hence, we are motivated to find an optimal look-back period. Run the SCM forecast on the set of IID simulated returns of 100 stocks, using a look-back period of 250 to 950 days, with increments of 50 days. We found the loss values are monotonically decreasing as seen from Figure 3.1. Indeed, the SCM is an asymptotically unbiased estimator of the true covariance matrix when $l \to \infty$. But practical limitations won't allow too long look back period. We take $l = 500$, an equivalence of 2-year data.

If we run the shrinkage forecast with a market index target on the same set of data and look back periods, we notice from the plot 3.1 that, although fluctuating, the loss values are decreasing as look back period increases. We also use $l = 500$ as a balance.

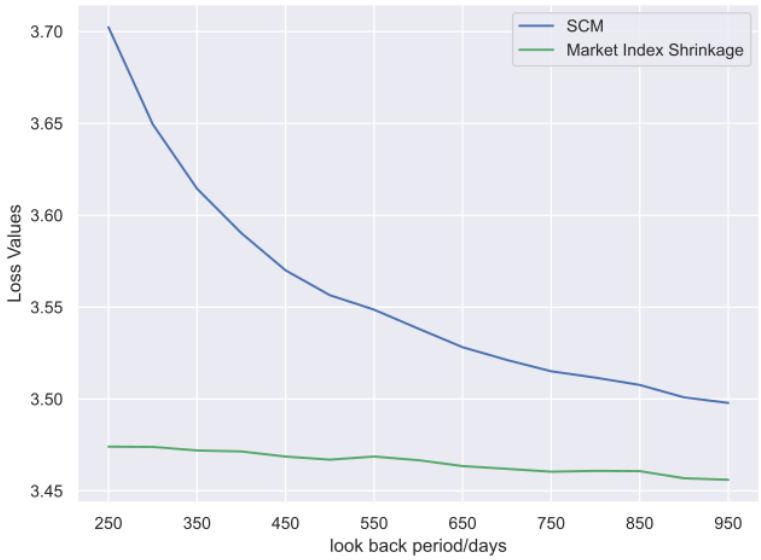

Figure 3.1: Averaged loss values when applying the SCM forecast and Market Shrinkage forecast to an IID simulated return data set of $N = 100$, plotted against look back periods from 250 to 950 days, increment at 50 days

In the actual implementation of the market shrinkage[7], for any asset type, the market index is estimated as the equally-weighted return of the N assets. Instead of doing regression, it is easier to find the covariance vector between the N assets and the market index, calculate the outer product of this covariance vector, and then divide by the variance of the equally-weighted market index. Finally, we replace the diagonal with the diagonal of the sample covariance matrix.

[7]see https://github.com/pald22/covShrinkage/blob/main/covMarket.py for source codes

For the EWMA forecast scheme, it remains the question to determine an optimal value of α. We employ the `scipy.optimize.minimize`[8] function of `Python` to find the α that minimizes the averaged loss value. Using the IID simulated returns of 100 stocks, we get the optimal $\hat{\alpha} = 0.998954$ with the effective look back period 955 as given by (1.1.3), which is much longer than we can take in practice. A plot of the averaged loss values against α's in the vicinity of $\hat{\alpha}$ is shown below in Figure 3.2. We then use $\hat{\alpha}$ in running the EWMA forecast. Moreover, the EWMA forecast needs an initial $\boldsymbol{H}_0$ and $\hat{\boldsymbol{C}}_0$ to kick off the initial forecast. We use the first 500 days of data to find the SCM as $\boldsymbol{H}_0$, and the covariance proxy on day 500 as $\hat{\boldsymbol{C}}_0$. The forecast starts on day 501 until the last day in the sample.

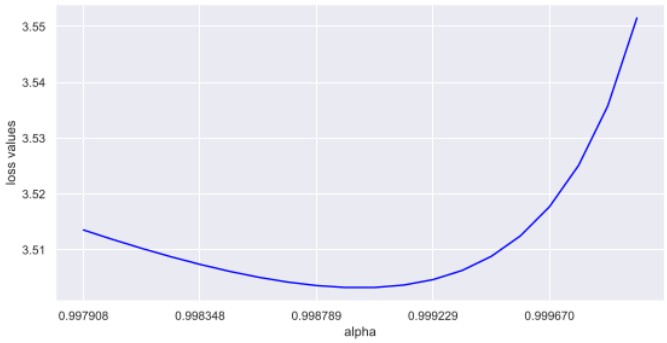

Figure 3.2: Averaged loss values when applying EWMA forecast to an IID simulated return data set of 100 stocks with α's near $\hat{\alpha}$, plotted against the corresponding α's

For the SCM forecast, adding filtering is straightforward, as we only need to apply filtering with the process detailed in section 1.3.2 to each $\boldsymbol{H}_t$. However, for the EWMA forecast, it is a little tricky. We need to distinguish between the forecast that is filtered and then put into the loss function and the forecast that is used to update the next step forecast. Failing to do so will give an inaccurate update that leads to undesirably large loss values.

In the above set-up, all the forecast schemes have a forecast period of 1500 days, and the loss series of length 1500 can be passed into the DM test to tell which forecast is statistically better, hence yielding a ranking of the various forecasts. We don't apply filtering to the shrinkage estimators because they both serve the purpose of noise reduction. The results are presented for different simulated data sets.

3.2.2 Results for IID Simulated data

The IID assumption of financial asset returns doesn't quite agree with real data, however, it is commonly used [8] thanks to its statistical tractability. For IID simulated data, we can obtain the true covariance matrix $\boldsymbol{C}$ (3.1.2) which is constant in

[8]https://docs.scipy.org/doc/scipy/reference/generated/scipy.optimize.minimize.html

time, so we also run the forecast with $\boldsymbol{C}$ in $L(\boldsymbol{C}, \boldsymbol{H}_t)$.

We first look at the optimal α values and the corresponding effective look-back periods of the EWMA forecast.

N	100	200	300	400
$\hat{\alpha}$	0.998909	0.998907	0.998881	0.998877
$\overline{T}(\hat{\alpha})$	916	914	893	890

Table 3.1: Optimal values of α's that minimize the averaged loss value $\overline{L}(\hat{\boldsymbol{C}})$ when EWMA forecast is applied to IID simulated returns

As the number of assets increases, the optimal α decreases, which means in the equation (1.1.1), the forecast $\boldsymbol{H}_{t+1}$ depends less on the previous forecast $\boldsymbol{H}_t$, and more weight is put on the new information at t. This is counter-intuitive because larger N will bring more noise to the covariance proxy $\hat{\boldsymbol{C}}_t = \boldsymbol{X}_t \boldsymbol{X}_t^\intercal$, whose weight should be lowered to achieve a more accurate forecast. We would be expecting increasing effective look-back periods, while the opposite is observed.

The averaged loss values are presented in the tables below:

	$\overline{L}(\hat{\boldsymbol{C}})$	$\overline{L}(\boldsymbol{C})$
SCM	3.548	3.564
SCM + filtering	3.459	3.473
EWMA	3.496	3.509
EWMA + filtering	3.456	3.469
Identity Shrinkage	3.533	3.549
EquiCorrelation Shrinkage	3.516	3.531
Market Shrinkage	3.460	3.474

Table 3.2: Averaged loss values over a forecast period of 1500 days when different forecast schemes are applied to an IID simulated return data set of $N = 100$

	$\overline{L}(\hat{\boldsymbol{C}})$	$\overline{L}(\boldsymbol{C})$
SCM	7.763	7.754
SCM + filtering	7.279	7.272
EWMA	7.466	7.460
EWMA + filtering	7.274	7.267
Identity Shrinkage	7.639	7.631
EquiCorrelation Shrinkage	7.537	7.531
Market Shrinkage	7.281	7.274

Table 3.3: Averaged loss values over a forecast period of 1500 days when different forecast schemes are applied to an IID simulated return data set of $N = 200$

The diagram 3.3 illustrates the averaged loss values $\overline{L}(\hat{\boldsymbol{C}})$ for different forecast schemes when N varies. Because we fixed the look back period $l = 500$, as N

	$\overline{L}(\hat{\boldsymbol{C}})$	$\overline{L}(\boldsymbol{C})$
SCM	12.214	12.235
SCM + filtering	10.763	10.780
EWMA	11.254	11.273
EWMA + filtering	10.755	10.771
Identity Shrinkage	11.737	11.757
EquiCorrelation Shrinkage	11.388	11.408
Market Shrinkage	10.764	10.781

Table 3.4: Averaged loss values over a forecast period of 1500 days when different forecast schemes are applied to an IID simulated return data set of $N = 300$

	$\overline{L}(\hat{\boldsymbol{C}})$	$\overline{L}(\boldsymbol{C})$
SCM	18.738	18.705
SCM + filtering	14.773	14.780
EWMA	15.819	15.815
EWMA + filtering	14.752	14.759
Identity Shrinkage	16.803	16.784
EquiCorrelation Shrinkage	16.013	16.003
Market Shrinkage	14.762	14.769

Table 3.5: Averaged loss values over a forecast period of 1500 days when different forecast schemes are applied to an IID simulated return data set of $N = 400$

increases, the ratio $Q := l/N$ decreases, and random noise distorts the forecasts from true covariance more, giving larger loss values. Obviously, the SCM forecast is most likely to be influenced by noise.

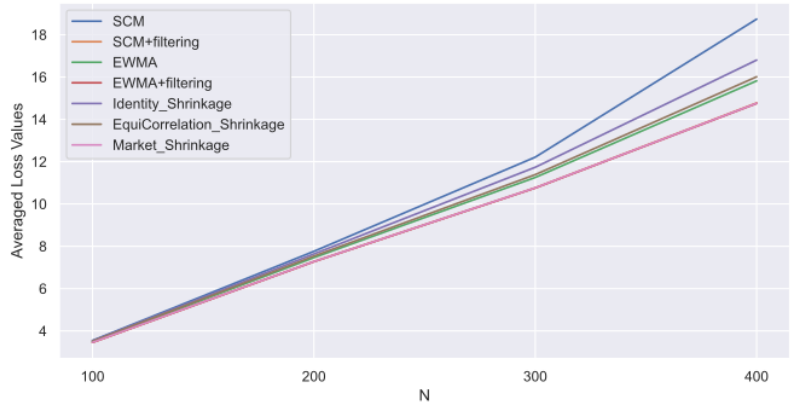

Figure 3.3: The averaged loss values $\overline{L}(\hat{\boldsymbol{C}})$ plotted against N when various forecasts are applied to IID simulated returns

We can immediately see some patterns in the above tables:

- The SCM forecast produces the largest averaged loss values, and is the worst forecast as expected;
- Filtering effectively reduces noise in the forecast and gives smaller loss values;
- Among the three shrinkage schemes, the Market Shrinkage performs much better than the other two. This makes sense because the data set used for forecasts is simulated under the market index model, so the target $\boldsymbol{F}_{\mathcal{M}}$ best describes the structure of the return data.

The ranking of various forecasts by $\overline{L}(\hat{\boldsymbol{C}})$ in an ascending order are in table 3.6. And the ranking by $\overline{L}(\boldsymbol{C})$ is the same. We can see that when N varies, the ranking is quite consistent, with the SCM with filtering and Market Shrinkage being the only exception.

N	100	200	300	400
SCM	7	7	7	7
SCM + filtering	2	2	2	3
EWMA	4	4	4	4
EWMA + filtering	1	1	1	1
Identity Shrinkage	6	6	6	6
EquiCorrelation Shrinkage	5	5	5	5
Market Shrinkage	3	3	3	2

Table 3.6: Ranking of forecasts applied to IID simulated returns by the averaged loss values

The next step is to determine whether the difference between $\overline{L}(\hat{\boldsymbol{C}})$ of different forecasts is statistically significant, utilising the DM test introduced in section 2.2. The same is repeated for the difference between $\overline{L}(\boldsymbol{C})$ to check if the DM ranking[9] is consistent using covariance proxy and true covariance, which should be satisfied when our loss function is robust. The DM test statistic (2.2.1) is asymptotically normal, and a forecast period of 1500 days is large enough for the asymptotic S_1 to work. For loss series produced using covariance proxy and using true covariance, we carry out the DM test of every pair of loss series by calculating the DM test statistic S_1 (2.2.1) of the loss differential of the two loss series. Note that the 97.5%-quantile of a standard normal variable is 1.96, because the DM test is a two-tailed test, we will consider ± 1.96 as critical values of DM statistic so that the null hypothesis is rejected at a significance level of 5% when $S_1 > 1.96$ or $S_1 < -1.96$. We present the DM test statistics with covariance proxy $\hat{\boldsymbol{C}}_t$ for $N = 100, 200, 300$ and 400 in tables 3.8, 3.9, 3.10 and 3.11 respectively. The same is done with true covariance in tables 3.12, 3.13, 3.14 and 3.15. In these tables, for brevity, we denote different forecast schemes by capital letters, with the mapping shown in table 3.7.

We interpret the presentation of DM test statistics as explained below:

[9]The ranking based on results of the DM test

SCM	SCM+ filtering	EWMA	EWMA+ filtering	Identity Shrinkage	Equi-Correlation Shrinkage	Market Shrinkage
A	B	C	D	E	F	G

Table 3.7: Mapping of forecast schemes to capital letters

- The forecast schemes in each column are $\boldsymbol{H}_{1,t}$ and the ones in each row are $\boldsymbol{H}_{2,t}$.
- For example, the test statistic at row B and column C is the normalised mean of $\hat{d}_t^{C,B} = \hat{u}_{C,t} - \hat{u}_{B,t}$. Hence, a value larger than 1.96 tells us we can reject the null hypothesis at 5% confidence level and conclude that the loss values using forecast C are statistically larger than that of B, so B is a better forecast at that particular value of N. This conclusion is visualized by colouring the corresponding cells in red.
- Vice versa, a test statistic smaller than -1.96 leads to the conclusion that C is a better forecast. This is coloured green.
- Otherwise, we cannot reject the null hypothesis and cannot tell which one is more accurate. This is marked by a yellow cell.
- Recall that the assumption of stationarity needs to be fulfilled for the DM test to be applicable, therefore, we plot the loss differential series of each pair of loss series and each N. We take a p-value of 0.05 as the critical value, and any smaller values make us reject the null hypothesis of the KPSS test that the observed loss series is stationary. These loss differential series are less suitable for the DM test and are marked in red in the plot. We present in the appendix some plots of loss differentials $\hat{d}_t = \hat{u}_{1,t} - \hat{u}_{2,t}$ of every pair of forecasts. The loss differentials in the 7×7 plots corresponds to the DM test statistic values in the tables in implementation of the forecasts in chapters 3 and 4.

	A	B	C	D	E	F	G
A		-55.26	-38.92	-54.02	-53.02	-65.18	-61.05
B	55.26		31.15	-5.54	51.35	45.48	2.46
C	38.92	-31.15		-36.13	30.05	17.77	-33.22
D	54.02	5.54	36.13		49.65	44.06	6.94
E	53.02	-51.35	-30.05	-49.65		-51.87	-57.64
F	65.18	-45.48	-17.77	-44.06	51.87		-52.57
G	61.05	-2.46	33.22	-6.94	57.64	52.57	

Table 3.8: The DM test statistics using $\hat{u}_{i,t} = L(\hat{\boldsymbol{C}}_t, \boldsymbol{H}_{i,t})$ and $\hat{d}_t = \hat{u}_{1,t} - \hat{u}_{2,t}$ on IID simulated returns when $N = 100$

From the tables of DM test using covariance proxy, we can see that all of the null hypotheses are rejected with very large test statistics. For those tests with the true covariance matrix, the test statistics are larger, which means that the difference in predictive accuracy is more statistically significant. We determine the DM ranking

	A	B	C	D	E	F	G
A		-107.52	-71.36	-106.32	-120.03	-127.00	-111.76
B	107.52		54.16	-7.41	97.85	87.60	6.85
C	71.36	-54.16		-56.89	47.45	22.41	-55.09
D	106.32	7.41	56.89		96.10	85.83	9.94
E	120.03	-97.85	-47.45	-96.10		-106.01	-102.86
F	127.00	-87.60	-22.41	-85.82	106.01		-93.60
G	111.76	-6.85	55.09	-9.94	102.86	93.60	

Table 3.9: The DM test statistics using $\hat{u}_{i,t} = L(\hat{\boldsymbol{C}}_t, \boldsymbol{H}_{i,t})$ and $\hat{d}_t = \hat{u}_{1,t} - \hat{u}_{2,t}$ on IID simulated returns when $N = 200$

	A	B	C	D	E	F	G
A		-159.50	-95.27	-159.07	-166.07	-173.57	-161.45
B	159.50		56.78	-10.59	145.84	129.73	3.64
C	95.27	-56.78		-58.32	54.38	16.57	-56.86
D	159.07	10.59	58.32		144.98	128.29	11.21
E	166.07	-145.84	-54.38	-144.98		-158.87	-148.52
F	173.57	-129.73	-16.57	-128.29	158.87		-133.35
G	161.45	-3.64	56.86	-11.21	148.52	133.35	

Table 3.10: The DM test statistics using $\hat{u}_{i,t} = L(\hat{\boldsymbol{C}}_t, \boldsymbol{H}_{i,t})$ and $\hat{d}_t = \hat{u}_{1,t} - \hat{u}_{2,t}$ on IID simulated returns when $N = 300$

	A	B	C	D	E	F	G
A		-202.59	-117.03	-202.94	-188.33	-201.11	-203.27
B	202.59		51.33	-19.65	188.97	173.14	-21.78
C	117.03	-51.33		-52.55	48.13	9.78	-51.67
D	202.94	19.65	52.55		188.92	172.98	10.88
E	188.33	-188.97	-48.13	-188.92		-195.37	-190.14
F	201.11	-173.14	-9.78	-172.98	195.37		-175.13
G	203.27	21.78	51.67	-10.88	190.14	175.13	

Table 3.11: The DM test statistics using $\hat{u}_{i,t} = L(\hat{\boldsymbol{C}}_t, \boldsymbol{H}_{i,t})$ and $\hat{d}_t = \hat{u}_{1,t} - \hat{u}_{2,t}$ on IID simulated returns when $N = 400$

	A	B	C	D	E	F	G
A		-898.65	-113.66	-931.68	-800.61	-730.91	-1700.46
B	898.65		75.97	-37.07	807.25	718.05	6.87
C	113.66	-75.97		-94.56	82.54	46.67	-73.80
D	931.68	37.07	94.56		810.90	789.15	53.47
E	841.75	-807.25	-82.54	-810.90		-497.20	-1735.14
F	730.91	-718.05	-46.67	-789.15	497.20		-2059.97
G	1700.46	-6.87	73.80	-53.47	1735.14	2059.97	

Table 3.12: The DM test statistics using $u_{i,t} = L(\boldsymbol{C}_t, \boldsymbol{H}_{i,t})$ and $d_t = u_{1,t} - u_{2,t}$ on IID simulated returns when $N = 100$

by comparing the number of cells of each colour in every column. For a particular value of N, a green cell in a column means that the forecast of the column is more

	A	B	C	D	E	F	G
A		-3360.32	-119.86	-3779.22	-1126.35	-866.41	-3394.26
B	3360.32		75.31	-59.32	2817.86	932.42	47.15
C	119.86	-75.31		-79.73	69.40	30.13	-74.29
D	3779.22	59.33	79.73		3174.14	1134.22	84.62
E	1126.35	-2817.86	-69.40	-3174.14		-511.06	-3027.40
F	866.41	-932.42	-30.13	-1134.22	511.06		-947.69
G	3394.26	-47.15	74.29	-84.62	3027.40	947.69	

Table 3.13: The DM test statistics using $u_{i,t} = L(\boldsymbol{C}_t, \boldsymbol{H}_{i,t})$ and $d_t = u_{1,t} - u_{2,t}$ on IID simulated returns when $N = 200$

	A	B	C	D	E	F	G
A		-3094.40	-130.88	-3145.15	-966.99	-1419.93	-3122.65
B	3094.40		66.25	-75.15	2253.05	1274.69	17.32
C	130.88	-66.25		-68.40	64.74	18.65	-65.97
D	3145.15	75.15	68.40		2250.63	1408.06	77.94
E	966.99	-2253.05	-64.74	-2250.63		-882.87	-2260.35
F	1419.93	-1274.69	-18.65	-1408.06	882.87		-1262.14
G	3122.65	-17.32	65.97	-77.94	2260.35	1262.14	

Table 3.14: The DM test statistics using $u_{i,t} = L(\boldsymbol{C}_t, \boldsymbol{H}_{i,t})$ and $d_t = u_{1,t} - u_{2,t}$ on IID simulated returns when $N = 300$

	A	B	C	D	E	F	G
A		-2767.51	-150.11	-2778.71	-1052.22	-1427.51	-2956.45
B	2767.51		55.68	-65.25	2025.70	1203.32	-33.98
C	150.11	-55.68		-57.07	52.24	10.11	-55.96
D	2778.71	65.25	57.07		2125.18	1330.95	52.65
E	1052.22	-2025.70	-52.24	-2125.18		-1391.83	-2029.62
F	1427.51	-1203.32	-10.11	-1330.95	1391.83		-1281.52
G	2956.45	33.98	55.96	-52.65	2029.62	1281.52	

Table 3.15: The DM test statistics using $u_{i,t} = L(\boldsymbol{C}_t, \boldsymbol{H}_{i,t})$ and $d_t = u_{1,t} - u_{2,t}$ on IID simulated returns when $N = 400$

accurate than the forecast of the row, a yellow cell is a tie, while a red cell means the column forecast is less accurate. Using this principle, we arrive at the DM ranking of various forecasts based on loss values against covariance proxy and true covariance. The DM ranking for both scenarios agree completely with the ranking by averaged loss values in table 3.6.

For every pair of forecasts, we can very confidently conclude the DM ranking, which is consistent for both the covariance proxy and true covariance, verifying that our loss function is indeed robust. Also, we can see that the DM ranking remains almost unchanged when N changes.

The conclusion for the IID simulated returns is that the EWMA forecast with filtering is the most accurate forecast, followed by the SCM forecast with filtering/Market Shrinkage, and then the EWMA forecast, EquiCorrelation Shrinkage, Identity Shrinkage and SCM forecast is the worst forecast.

3.2.3 Results for Equi-Correlational GARCH (ECG) simulated data

The optimal α values and the corresponding effective look-back periods for the EWMA forecast are shown in the following table.

N	100	200	300	400
$\hat{\alpha}$	0.998801	0.998842	0.998852	0.998841
$\overline{T}(\hat{\alpha})$	833	862	870	862

Table 3.16: Optimal values of α's that minimize the averaged loss value $\overline{L}(\hat{\boldsymbol{C}})$ when EWMA forecast is applied to ECG simulated returns

The $\hat{\alpha}$ for the ECG simulated returns are fluctuating when N increases, but slightly smaller than those for the IID simulated returns.

The averaged loss values are presented in the table 3.17 below. Note that we don't include the loss values against the true covariance matrix, because the data is simulated based on auto-regressive processes, and there is no known distribution that the returns follow.

N	100	200	300	400
SCM	31.146	66.035	106.031	160.262
SCM + filtering	30.270	61.460	92.305	123.452
EWMA	30.629	63.245	96.930	133.178
EWMA + filtering	30.268	61.475	92.295	123.424
Identity Shrinkage	30.425	61.927	92.945	124.360
EquiCorrelation Shrinkage	30.244	61.434	92.266	123.392
Market Shrinkage	30.287	61.484	92.317	123.450

Table 3.17: Averaged loss values $\overline{L}(\hat{\boldsymbol{C}})$ over a forecast period of 1500 days when different forecast schemes are applied to an ECG simulated return data set of varying Ns

The loss values are much bigger than the IID simulated data, but we are more focused on the relative magnitudes of the loss values and the absolute values can be adjusted by scaling the returns. Some patterns to notice are

- SCM forecast gives the largest averaged loss values and is again the worst forecast.
- As N increases, the averaged loss values increase almost linearly except for the SCM forecast whose loss value grows faster than the linear rate. This also indicates it is the forecast most influenced by noise.
- Filtering still works as an effective way of noise reduction.

- This time the EquiCorrelation Shrinkage performs the best out of the three shrinkage targets because the original return data is simulated using the Equi-Correlational model. The next best is Market Shrinkage while the worst one is Identity Shrinkage which incorporates the least structure.

The ranking by averaged loss values in ascending order for different N's is in table 3.18. Applying the DM test to the relevant pairs of loss series gives the following results in tables 3.19, 3.20, 3.21 and 3.22.

N	100	200	300	400
SCM	7	7	7	7
SCM + filtering	3	2	3	4
EWMA	6	6	6	6
EWMA + filtering	2	3	2	2
Identity Shrinkage	5	5	5	5
EquiCorrelation Shrinkage	1	1	1	1
Market Shrinkagee	4	4	4	3

Table 3.18: Ranking of forecasts applied to ECG simulated returns by the averaged loss values

	A	B	C	D	E	F	G
A		-54.92	-39.56	-52.93	-49.37	-57.15	-56.72
B	54.92		31.86	-0.55	17.53	-16.90	12.30
C	39.56	-31.86		-34.77	-18.15	-34.72	-31.61
D	52.93	0.55	34.77		18.04	-5.99	4.85
E	49.37	-17.53	18.15	-18.04		-20.67	-16.33
F	57.15	16.90	34.72	5.99	20.67		22.51
G	56.72	-12.30	31.61	-4.85	16.33	-22.51	

Table 3.19: The DM test statistics using $\hat{u}_{i,t} = L(\hat{\boldsymbol{C}}_t, \boldsymbol{H}_{i,t})$ and $\hat{d}_t = \hat{u}_{1,t} - \hat{u}_{2,t}$ on ECG simulated returns when $N = 100$

	A	B	C	D	E	F	G
A		-107.35	-71.77	-106.03	-101.46	-108.19	-108.15
B	107.35		52.63	2.27	24.89	-17.23	15.90
C	71.77	-52.63		-53.81	-39.79	-53.18	-52.11
D	106.03	-2.27	53.81		26.39	-6.17	1.39
E	101.46	-24.89	39.79	-26.39		-26.20	-24.04
F	108.19	17.23	53.18	6.17	26.20		24.44
G	108.15	-15.90	52.11	-1.39	24.04	-24.44	

Table 3.20: The DM test statistics using $\hat{u}_{i,t} = L(\hat{\boldsymbol{C}}_t, \boldsymbol{H}_{i,t})$ and $\hat{d}_t = \hat{u}_{1,t} - \hat{u}_{2,t}$ on ECG simulated returns when $N = 200$

Based on the results in the tables of DM test statistics, we come up with the DM ranking of forecasts in table **??**, with the different ranking from averaged loss values ranking marked in red. We can see that, statistically, we are unable to tell

	A	B	C	D	E	F	G
A		-158.64	-94.57	-157.80	-154.85	-159.00	-159.19
B	158.64		58.91	-1.34	33.79	-18.05	5.67
C	94.57	-58.91		-60.00	-50.64	-59.48	-58.90
D	157.80	1.34	60.00		35.86	-3.74	2.97
E	154.85	-33.79	50.64	-35.86		-35.75	-33.41
F	159.00	18.05	59.48	3.74	35.75		24.09
G	159.19	-5.67	58.90	-2.97	33.41	-24.09	

Table 3.21: The DM test statistics using $\hat{u}_{i,t} = L(\hat{\boldsymbol{C}}_t, \boldsymbol{H}_{i,t})$ and $\hat{d}_t = \hat{u}_{1,t} - \hat{u}_{2,t}$ on ECG simulated returns when $N = 300$

	A	B	C	D	E	F	G
A		-190.13	-114.63	-189.64	-188.59	-190.47	-190.49
B	190.12		52.20	-3.17	38.00	-21.89	-0.55
C	114.63	-52.20		-52.56	-47.07	-52.40	-52.13
D	189.64	3.17	52.56		41.81	-3.73	3.01
E	188.59	-38.00	47.07	-41.81		-40.56	-38.27
F	190.47	21.89	52.40	3.73	40.56		26.43
G	190.49	0.55	52.13	-3.01	38.27	-26.43	

Table 3.22: The DM test statistics using $\hat{u}_{i,t} = L(\hat{\boldsymbol{C}}_t, \boldsymbol{H}_{i,t})$ and $\hat{d}_t = \hat{u}_{1,t} - \hat{u}_{2,t}$ on ECG simulated returns when $N = 400$

N	100	200	300	400
SCM	7	7	7	7
SCM + filtering	2/3	2	2/3	3/4
EWMA	6	6	6	6
EWMA + filtering	2/3	3/4	2/3	2
Identity Shrinkage	5	5	5	5
EquiCorrelation Shrinkage	1	1	1	1
Market Shrinkage	4	3/4	4	3/4

Table 3.23: The DM ranking of forecasts applied to ECG simulated returns

- whether SCM forecast with filtering or EWMA forecast with filtering is a better one when $N = 100, 300$;
- whether EWMA forecast with filtering and Market Shrinkage is a better forecast when $N = 200$;
- whether SCM forecast with filtering and Market Shrinkage is a better forecast when $N = 400$.

Nevertheless, the rest DM rankings are clear and consistent with the averaged loss values ranking. In general, the EquiCorrelation Shrinkage is the most accurate forecast, followed by SCM with filtering/EWMA with filtering, then Market Shrinkage, Identity Shrinkage, and EWMA forecast, with SCM at the bottom.

3.2.4 Results for Constant Correlational GARCH(CCG) simulated data

The optimal α values and the corresponding effective look-back periods for the EWMA forecast are shown in table 3.24. The $\hat{\alpha}$ for the CCG simulated returns increases when N increases and the effective look back period also increases. This is expected as explained in the results for IID simulated data.

N	100	200	300	400
$\hat{\alpha}$	0.998655	0.998771	0.998780	0.998807
$\overline{T}(\hat{\alpha})$	742	813	818	837

Table 3.24: Optimal values of α's that minimize the averaged loss value $\overline{L}(\hat{\boldsymbol{C}})$ when EWMA forecast is applied to CCC simulated returns

The averaged loss values are presented in the table 3.25 below. Note that we don't include the loss values against true covariance for the same reason as the ECG simulated data.

N	100	200	300	400
SCM	25.840	51.380	78.981	113.273
SCM + filtering	25.949	50.352	74.536	100.472
EWMA	25.478	49.499	73.245	97.381
EWMA + filtering	25.878	49.998	73.530	96.968
Identity Shrinkage	25.751	50.674	76.230	102.961
EquiCorrelation Shrinkage	25.633	50.025	74.097	98.246
Market Shrinkage	25.622	49.995	73.851	97.845

Table 3.25: Averaged loss values $\overline{L}(\hat{\boldsymbol{C}})$ over a forecast period of 1500 days when different forecast schemes are applied to a CCG simulated return data set of varying Ns

The ranking by averaged loss values in ascending order are in table 3.26. Applying the DM test to the relevant pairs of loss series gives results presented in tables 3.27, 3.28, 3.29 and 3.30. Most of the DM tests have null hypothesis rejected and the DM ranking is the same as the ranking by averaged loss values in table 3.26. Sometimes, we cannot tell the DM ranking of two forecasts directly from the single test between them, for example, when $N = 200$ in table 3.28, the DM test of EquiCorrelation Shrinkage and EWMA forecast with filtering gives $S_1 = 1.12$, so we cannot reject the null hypothesis. However, EquiCorrelation Shrinkage has 3 green cells, 1 yellow cell and 2 red cells, while EWMA forecast with filtering has 1 more yellow cell and 1 fewer red cell. This means that, statistically, EWMA forecast with filtering has a tie with two forecast schemes, meanwhile, EquiCorrelation Shrinkage has one tie and one loss. It is therefore fair to say EWMA forecast with filtering is more accurate.

N	100	200	300	400
SCM	5	7	7	7
SCM + filtering	7	5	5	5
EWMA	1	1	1	2
EWMA + filtering	6	3	2	1
Identity Shrinkage	4	6	6	6
EquiCorrelation Shrinkage	3	4	4	4
Market Shrinkage	2	2	3	3

Table 3.26: Ranking of forecasts applied to CCG simulated returns by the averaged loss values and the DM ranking are the same

	A	B	C	D	E	F	G
A		6.54	-38.81	2.32	-50.81	-50.03	-49.97
B	-6.54		-28.77	-6.46	-12.47	-21.56	-22.72
C	38.81	28.77		27.56	31.63	18.98	17.59
D	-2.32	6.46	-27.56		-8.16	-17.01	-18.12
E	50.81	12.47	-31.63	8.16		-43.12	-43.57
F	50.03	21.56	-18.98	17.01	43.12		-8.59
G	49.97	22.72	-17.59	18.12	43.57	8.59	

Table 3.27: The DM test statistics using $\hat{u}_{i,t} = L(\hat{\boldsymbol{C}}_t, \boldsymbol{H}_{i,t})$ and $\hat{d}_t = \hat{u}_{1,t} - \hat{u}_{2,t}$ on CCG simulated returns when $N = 100$

	A	B	C	D	E	F	G
A		-32.34	-70.29	-42.88	-108.64	-102.99	-104.08
B	32.34		-27.08	-20.71	11.54	-13.62	-15.02
C	70.29	27.08		18.38	49.45	23.56	22.31
D	42.88	20.71	-18.38		23.97	1.12	-0.12
E	108.64	-11.54	-49.45	-23.97		-89.38	-91.82
F	102.99	13.62	-23.56	-1.12	89.38		-14.63
G	104.08	15.02	-22.31	0.12	91.82	14.63	

Table 3.28: The DM test statistics using $\hat{u}_{i,t} = L(\hat{\boldsymbol{C}}_t, \boldsymbol{H}_{i,t})$ and $\hat{d}_t = \hat{u}_{1,t} - \hat{u}_{2,t}$ on CCG simulated returns when $N = 200$

	A	B	C	D	E	F	G
A		-57.54	-90.34	-86.04	-140.32	-138.05	-137.22
B	57.54		-16.16	-16.10	25.03	-7.04	-11.07
C	90.34	16.16		5.34	55.46	16.32	11.72
D	86.04	16.10	-5.34		55.87	15.07	8.84
E	140.32	-25.03	-55.46	-55.87		-126.07	-125.37
F	138.05	7.04	-16.32	-15.07	126.07		-60.65
G	137.22	11.07	-11.72	-8.84	125.37	60.65	

Table 3.29: The DM test statistics using $\hat{u}_{i,t} = L(\hat{\boldsymbol{C}}_t, \boldsymbol{H}_{i,t})$ and $\hat{d}_t = \hat{u}_{1,t} - \hat{u}_{2,t}$ on CCG simulated returns when $N = 300$

	A	B	C	D	E	F	G
A		-126.04	-111.38	-139.87	-173.67	-177.63	-177.02
B	126.04		-25.72	-61.62	44.95	-49.92	-59.98
C	111.38	25.72		-3.85	49.16	7.41	4.05
D	139.87	61.62	3.85		89.13	26.30	18.94
E	173.67	-44.95	-49.16	-89.13		-157.09	-155.47
F	177.63	49.92	-7.41	-26.30	157.09		-49.91
G	177.02	59.98	-4.05	-18.94	155.47	49.91	

Table 3.30: The DM test statistics using $\hat{u}_{i,t} = L(\hat{\boldsymbol{C}}_t, \boldsymbol{H}_{i,t})$ and $\hat{d}_t = \hat{u}_{1,t} - \hat{u}_{2,t}$ on CCG simulated returns when $N = 400$

Unlike the IID and ECG simulated returns, where the forecast ranking by the DM test isn't significantly influenced by the number of assets N, the DM ranking can be N-dependent when different forecast schemes are applied to the CCG simulated returns. And we can see that:

- The EWMA forecast with filtering isn't as powerful as with the previous two simulated data sets, but its predictive accuracy improves as N increases. Because there is more noise and filtering's role in noise reduction becomes more important.
- The EWMA forecast and Market Shrinkage are the top forecasts for various N, with the EWMA forecast having a slightly smaller loss value than the Market Shrinkage.
- Among the three shrinkage targets, the Market Shrinkage is the most accurate, followed by EquiCorrelation Shrinkage and then Identity Shrinkage. This is the same as the IID simulated data. It shows that, without the special structure of equal correlation in data simulation, the Market Shrinkage target best captures the useful structure in the return data.

In general, for the CCG simulated returns, it is best to use EWMA forecast or Market Shrinkage, and only apply filtering when N gets larger.

Chapter 4

Forecasts Applied to Historical Data

In this chapter, we apply various forecast schemes to the historical returns of the S&P500 listed stocks and explore the ranking of different forecasts. We also explore the covariance between US Treasury yield returns at the 1, 2, 3, 5, 7, 10, 20, and 30 years maturity. Note that while we calculate log returns for the stocks, we directly calculate the difference between adjacent yields at a daily frequency.

4.1 S&P500 Stock Data

We use the data from 01-01-2017 to 06-01-2022 of the first $N = 100, 200, 300$ and 400 stocks listed in the S&P500 index and calculate their log returns. The total forecast period is 861 days when we keep using 500 days as the look-back period just as in the previous chapter. The optimal α values and the corresponding effective look-back periods for the EWMA forecast are shown in the following table.

N	100	200	300	400
$\hat{\alpha}$	0.996649	0.997805	0.998146	0.998333
$\overline{T}(\hat{\alpha})$	297	455	538	599

Table 4.1: Optimal values of α's that minimize the averaged loss value $\overline{L}(\hat{\boldsymbol{C}})$ when EWMA forecast is applied to historical stock returns

As N increases, there is more noise in the covariance proxy $\hat{\boldsymbol{C}}_t$, the weighting of which is therefore reduced to achieve a more accurate forecast, and this is done by increasing α. Meanwhile, the effective look-back period increases, which makes sense because more assets are involved. Also, we notice that the $\hat{\alpha}$ of the historical stock returns are smaller than those of the simulated returns.

Next, we show the averaged loss values calculated using covariance proxy and the ranking by the averaged loss values in tables 4.2 and 4.3. The test statistics for the DM tests are presented in the tables 4.4, 4.5, 4.6 and 4.7 for $N = 100, 200, 300$ and 400 respectively.

N	100	200	300	400
SCM	3.672	7.487	11.632	17.641
SCM + filtering	3.640	7.152	10.512	14.903
EWMA	3.595	7.219	10.811	15.075
EWMA + filtering	3.599	7.044	10.306	13.972
Identity Shrinkage	3.595	7.051	10.411	14.197
EquiCorrelation Shrinkage	3.579	7.112	10.349	13.941
Market Shrinkage	3.585	7.026	10.290	13.864

Table 4.2: Averaged loss values $\overline{L}(\hat{\boldsymbol{C}})$ over a forecast period of 861 days when different forecast schemes are applied to historical return data set of varying Ns

N	100	200	300	400
SCM	7	7	7	7
SCM + filtering	6	5	5	5
EWMA	4	6	6	6
EWMA + filtering	5	2	2	3
Identity Shrinkage	3	4	4	4
EquiCorrelation Shrinkage	1	1	3	2
Market Shrinkage	2	3	1	1

Table 4.3: Ranking of forecasts applied to historical stock returns by the averaged loss values

	A	B	C	D	E	F	G
A		-4.81	-12.28	-5.87	-6.52	-7.07	-8.88
B	11.90		-5.93	-4.06	-4.98	-6.32	-7.85
C	12.28	5.93		0.37	-0.06	-1.62	-1.50
D	5.87	4.06	-0.37		-0.38	-2.03	1.62
E	6.52	4.98	0.06	0.38		-4.91	-2.52
F	7.07	6.32	1.62	2.03	4.91		1.48
G	8.88	7.85	1.50	1.62	2.52	-1.48	

Table 4.4: The DM test statistics using $\hat{u}_{i,t} = L(\hat{\boldsymbol{C}}_t, \boldsymbol{H}_{i,t})$ and $\hat{d}_t = \hat{u}_{1,t} - \hat{u}_{2,t}$ on historical stock returns when $N = 100$

	A	B	C	D	E	F	G
A		-11.90	-21.40	-17.91	-9.36	-9.91	-11.78
B	11.90		3.33	-10.01	-2.04	-5.26	-6.27
C	21.40	-3.33		-11.35	-3.59	-5.29	-6.13
D	17.91	10.01	11.35		3.25	-0.70	0.44
E	9.36	2.04	3.59	-3.25		-10.43	-8.20
F	9.91	5.26	5.29	0.70	10.43		2.41
G	11.78	6.27	6.13	-0.44	8.20	-2.41	

Table 4.5: The DM test statistics using $\hat{u}_{i,t} = L(\hat{\boldsymbol{C}}_t, \boldsymbol{H}_{i,t})$ and $\hat{d}_t = \hat{u}_{1,t} - \hat{u}_{2,t}$ on historical stock returns when $N = 200$

	A	B	C	D	E	F	G
A		-15.98	-21.06	-19.38	-12.29	-12.36	-13.81
B	15.98		7.91	-11.91	-2.76	-3.84	-6.60
C	21.06	-7.91		-14.54	-6.23	-6.76	-8.37
D	19.38	11.91	14.54		2.70	0.99	-0.45
E	12.29	2.76	6.23	-2.70		-5.73	-12.84
F	12.36	3.84	6.76	-0.99	5.73		-4.90
G	13.81	6.60	8.37	0.45	12.84	4.90	

Table 4.6: The DM test statistics using $\hat{u}_{i,t} = L(\hat{\boldsymbol{C}}_t, \boldsymbol{H}_{i,t})$ and $\hat{d}_t = \hat{u}_{1,t} - \hat{u}_{2,t}$ on historical stock returns when $N = 300$

	A	B	C	D	E	F	G
A		-17.28	-18.29	-18.04	-13.92	-13.97	-14.86
B	17.28		3.68	-16.36	-6.86	-8.05	-9.75
C	18.29	-3.68		-15.76	-7.63	-8.65	-9.98
D	18.04	16.36	15.76		4.01	-0.43	-1.84
E	13.92	6.86	7.63	-4.01		-12.04	-17.79
F	13.97	8.05	8.65	0.43	12.04		-3.90
G	14.86	9.75	9.98	1.84	17.79	3.90	

Table 4.7: The DM test statistics using $\hat{u}_{i,t} = L(\hat{\boldsymbol{C}}_t, \boldsymbol{H}_{i,t})$ and $\hat{d}_t = \hat{u}_{1,t} - \hat{u}_{2,t}$ on historical stock returns when $N = 400$

By comparing the proportions of green, yellow and red cells for each column, we can reach a conclusion about the DM ranking of various forecasts at different N's, which is shown in table 4.8. The bold numbers in red mark the difference from the ranks by averaged loss values in table 4.3.

N	100	200	300	400
SCM	7	7	7	7
SCM + filtering	6	5	5	5
EWMA	**3**	6	6	6
EWMA + filtering	**4**	2	2	**2**
Identity Shrinkage	**5**	4	4	4
EquiCorrelation Shrinkage	1	1	3	**3**
Market Shrinkage	2	3	1	1

Table 4.8: The DM ranking of forecasts applied to historical stock returns

Although the ranking varies as N changes, the top performers are Market Shrinkage, EquiCorrelation Shrinkage and EWMA forecast with filtering regardless of the number of stocks. Recall that from Chapter 3, we have EWMA forecast filtering as the most accurate forecast with IID simulated returns; EquiCorrelation Shrinkage as the best for ECG simulated returns, and EWMA forecast with filtering is the best when N gets large in CCG simulated returns. Market Shrinkage constantly performs well except for a slightly worse ranking with ECG simulated returns. Therefore, it is not surprising to get results like this for the historical return data. This also indicates that all three simulation models capture part of the actual market performance

of stocks.

4.2 US Treasury Yield Data

Treasury bonds, Treasury bills, and Treasury notes are all government-issued fixed-income securities.[1] The different names are for different maturities. For convenience, we call them Treasury securities. We use the US Treasury yield data[2] at constant maturity of 1, 2, 3, 5, 7, 10, 20 and 30 years from 1980-01-02 to 2022-08-11. The only exception is the 20-year Treasury yield, which is not available from the beginning of 1987 until October 1993. Therefore, we use the period 1993-10-01 to 2022-08-11 for the 20-year Treasury yield. When applying forecasts, we use two different data sets, one includes the 20-year Treasury yield from 1993 to 2022, and the other excludes the 20-year Treasury yield and all other yields dating from 1980 to 2022. There are 8 and 7 assets in these two data sets respectively. The return from the treasury yield is calculated directly as the absolute return. The optimal α's for the EWMA forecast and the corresponding effective look-back periods are shown below.

N	7	8
$\hat{\alpha}$	0.970837	0.970507
$\overline{T}(\hat{\alpha})$	33.3	32.9

Table 4.9: Optimal values of α's that minimize the averaged loss value $\overline{L}(\hat{\boldsymbol{C}})$ when EWMA forecast is applied to Treasury yield returns

The values of $\hat{\alpha}$ and the effective look-back periods are much smaller than those of the stock returns. This is because the Treasury securities are traded less frequently, and their yields contain much less noise. Moreover, unlike stock data, the Treasury yields are available over decades of time, we can therefore double our look-back period from 500 to 1000 when applying various forecasts of the covariance matrix. The averaged loss values calculated using covariance proxy and the ranking by the averaged loss values are shown in table 4.10. The test statistics for the DM tests are presented in the tables 4.11, 4.12 without and with the 20-year yield returns respectively. We also present the DM ranking in table 4.13.

The averaged loss values are much smaller than those calculated with the stock returns, indicating that the covariance forecasts of the Treasury yield returns are more accurate and less affected by noise. Therefore, we can observe that filtering doesn't work and worsens the original forecasts by SCM and EWMA. The DM ranking shows some difference with and without the 20-year yield returns, but the EWMA forecast is the most accurate for both data sets. We can also conclude statistically that filtering indeed worsens our forecasts, giving larger loss values. Therefore, our conclusion for the covariance forecast of the Treasury yield returns covariance matrix is that it is best to apply the EWMA forecast and avoid the use

[1] See "https://www.investopedia.com/ask/answers/033115/what-are-differences-between-treasury-bond-and-treasury-note-and-treasury-bill-tbill.asp" for detailed explanation

[2] For example, the 10-year Treasury yield is the "Market Yield on U.S. Treasury Securities at 10-Year Constant Maturity, Quoted on an Investment Basis" downloaded from "https://fred.stlouisfed.org/series/DGS10"

N	7		8	
SCM	0.5226	4	0.5199	4
SCM + filtering	0.6143	7	0.5510	7
EWMA	0.4895	1	0.4916	1
EWMA + filtering	0.5364	2	0.5144	6
Identity Shrinkage	0.5222	5	0.5209	2
EquiCorrelation Shrinkage	0.5227	6	0.5215	5
Market Shrinkage	0.5225	3	0.5199	3

Table 4.10: Averaged loss values $\overline{L}(\hat{\boldsymbol{C}})$ when different forecast schemes are applied to Treasury yield returns, with the ranking in the adjacent column on the right

	A	B	C	D	E	F	G
A		28.28	-20.67	-3.16	3.86	6.14	-3.32
B	-28.28		-30.63	-20.47	-28.83	-29.24	-28.35
C	20.67	30.63		22.42	21.57	21.84	20.66
D	3.16	20.47	-22.42		3.87	4.24	3.15
E	-3.86	28.83	-21.57	-3.87		8.22	-3.97
F	-6.14	29.24	-21.84	-4.24	-8.22		-6.27
G	3.32	28.35	-20.66	-3.15	3.97	6.27	

Table 4.11: The DM test statistics using $\hat{u}_{i,t} = L(\hat{\boldsymbol{C}}_t, \boldsymbol{H}_{i,t})$ and $\hat{d}_t = \hat{u}_{1,t} - \hat{u}_{2,t}$ on Treasury yield returns without the 20-year Treasury security

	A	B	C	D	E	F	G
A		29.98	-16.88	5.04	-1.13	0.22	-2.80
B	-29.98		-29.59	-19.35	-30.56	-30.58	-30.04
C	16.88	29.59		22.79	16.94	17.02	16.86
D	-5.04	19.35	-22.79		-5.42	-5.24	-5.07
E	1.13	30.56	-16.94	5.42		8.69	1.04
F	-0.22	30.58	-17.02	5.24	-8.69		-0.35
G	2.80	30.04	-16.86	5.07	-1.04	0.35	

Table 4.12: The DM test statistics using $\hat{u}_{i,t} = L(\hat{\boldsymbol{C}}_t, \boldsymbol{H}_{i,t})$ and $\hat{d}_t = \hat{u}_{1,t} - \hat{u}_{2,t}$ on Treasury yield returns with the 20-year Treasury Security

N	7	8
SCM	4	4/5
SCM + filtering	7	7
EWMA	1	1
EWMA + filtering	2	6
Identity Shrinkage	5	2/3
EquiCorrelation Shrinkage	6	4/5
Market Shrinkage	3	2/3

Table 4.13: The DM ranking of forecasts applied to Treasury yield returns

of filtering. The next best choice is Market Shrinkage.

Chapter 5

Conclusions and Discussions

This thesis studies the ranking of competing forecasts of the conditional covariance matrix of financial returns by a robust loss function (2.3.4).

$$L(\hat{\boldsymbol{C}}, \boldsymbol{H}) = \boldsymbol{Tr}(\boldsymbol{H}^{\frac{1}{2}} + \boldsymbol{H}^{-\frac{1}{2}}\hat{\boldsymbol{C}})$$

where $\hat{\boldsymbol{C}}$ is the conditionally unbiased covariance proxy and $\boldsymbol{H}$ is the covariance forecast. The loss function evaluates the distance of the forecasts from the true covariance matrix and it comes from a special case of Patton's [15] class of robust and homogeneous loss functions, and we extended the inputs into the loss function from univariate to multivariate. We then proved its robustness as defined in 2.3.1. and also proved that the above loss function is uniquely minimised when the forecast equals the true conditional covariance matrix. Then, we verified its robustness using the IID simulated returns, by showing that the DM ranking of forecasts based on the covariance proxy and the true conditional covariance agree with each other. With the robustness property, this loss function can be used to evaluate various covariance forecasts and most importantly, we can select the forecast closest to the true conditional covariance.

The following table summarizes the top forecasts of covariance for different return data sets:

	1	2	3
IID Simulated Returns	EWMA + filtering	Market Shrinkage/ SCM + filtering	EWMA
ECG Simulated Returns	EquiCorrelation Shrinkage	SCM + filtering/ EWMA + filtering	Market Shrinkage
CCG Simulated Returns	EWMA/ EWMA + filtering (when N is large)	Market Shrinkage	EquiCorrelation Shrinkage
Historical Stock Returns	Market Shrinkage	EquiCorrelation Shrinkage	EWMA + filtering
Treasury yield Returns	EWMA	Market Shrinkage	N.A.

Table 5.1: The most accurate forecast schemes for different return data set

From the above table, we can clearly see that the most accurate forecasts are EWMA/EWMA with filtering, Market Shrinkage and EquiCorrelation Shrinkage. When the return data set has certain features, some forecasts tend to do better than others.

- When the number of assets is more than 200, filtering is statistically significant in noise reduction and improvement of forecast accuracy. This can be verified if we look at the DM ranking of forecasts applied to CCG simulated returns in table 3.26 and to historical stock returns in table 4.8, where the rank of SCM with filtering and EWMA with filtering are both much better when $N \geq 200$ compared to when $N = 100$. In particular, the ranking of EWMA forecast with filtering increases from the bottom to the top as N increases from 100 to 400 in table 3.26. When the asset number is small and when the returns are in nature not very noisy (for example, the Treasury securities are traded less frequently than the stocks), filtering actually worsens the forecast by treating useful information as noise.

- Market Shrinkage is a good forecast regardless of the structure of the return data sets. This is because it uses the average of all assets involved as the market index and always captures a good amount of structured information in the return data set. Notice that Market Shrinkage is the best forecast when applied to historical stock returns.

- Although the EWMA forecast provides a very good forecast in some return data sets (e.g the CCG simulated returns, historical returns when $N = 100$ and the Treasury yield returns), it can be near the bottom for some other data sets. The behaviour is quite unstable compared to Market Shrinkage. If we are not sure whether the asset number is big enough for filtering to be beneficial, it is the most reliable to just apply the Market Shrinkage.

- EquiCorrelation Shrinkage's rank depends on the degree to which the actual correlations between the assets are equal. When the returns are simulated assuming an equal correlation i.e. the ECG simulated returns, EquiCorrelation Shrinkage naturally provides the best forecast. Interestingly, the EquiCorrelation Shrinkage applied to historical stock returns also gives good accuracy, indicating that for the most frequently traded stocks, the correlations between them don't differ much.

- For the US Treasury yield returns, filtering is not suitable at all. EWMA and Market Shrinkage are the best choices, while the others are all not so good, the third most accurate forecast is therefore not applicable.

The structure of actual financial returns can change over time and can be quite unpredictable. The simulated data sets above are far from an accurate replication of the real stock returns, and the ranking of various forecasts applied to them just gives us a list of potentially well-performing forecasts.

Even though we have performed the same procedure with some actual financial returns like the stock returns and Treasury yield returns, we didn't explore how the forecast accuracy can be influenced by certain time periods in the economy. The

performance of different forecasts might differ in calm and volatile markets. And this can be a further direction of investigation.

In general, by introducing the multivariate version of the robust loss function, the work of this thesis does provide a reliable way to measure the loss of a covariance forecast/estimate against the unobservable true covariance matrix. Although a significant proportion of the loss differentials don't seem to satisfy the stationary assumption of the DM test, most test statistics values exceed the critical value of ± 1.96 by a large amount. This already shows the loss differentials deviate much from zero, which is a fair justification of different predictive accuracy even if the test statistic S_1 aren't asymptotically standard normal.

Appendix A

Loss Differentials Plots

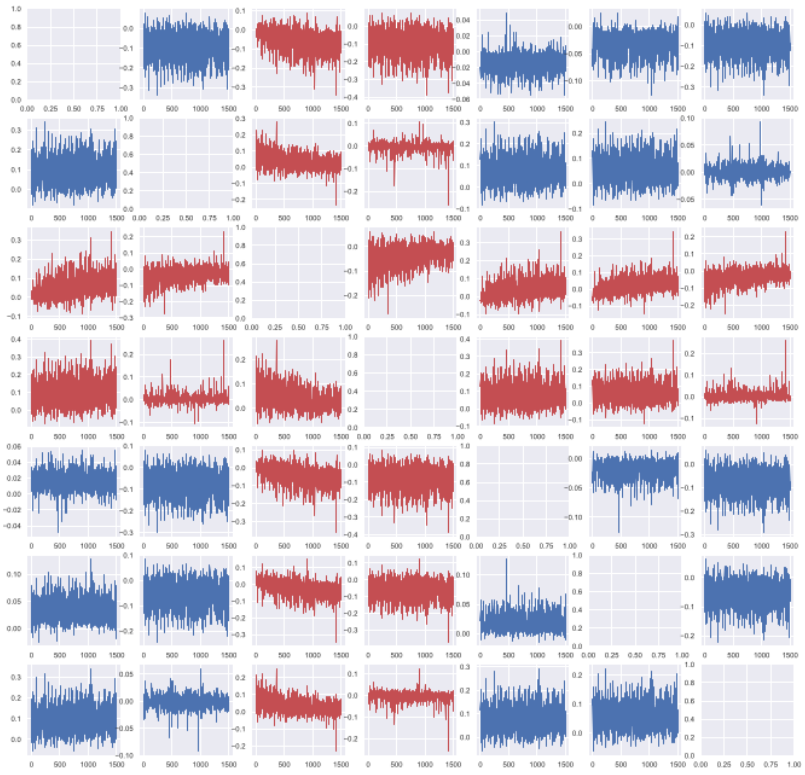

Figure A.1: Plots of loss differentials of every pair of the seven forecasts calculated using the covariance proxy when applied to IID simulated returns of 100 assets

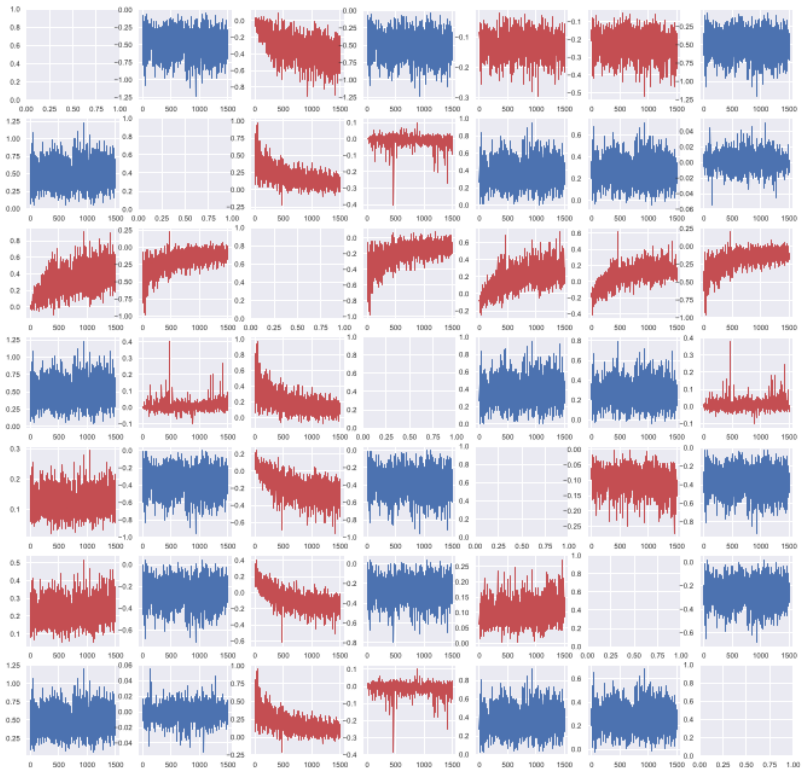

Figure A.2: Plots of loss differentials of every pair of the seven forecasts calculated using the covariance proxy when applied to IID simulated returns of 200 assets

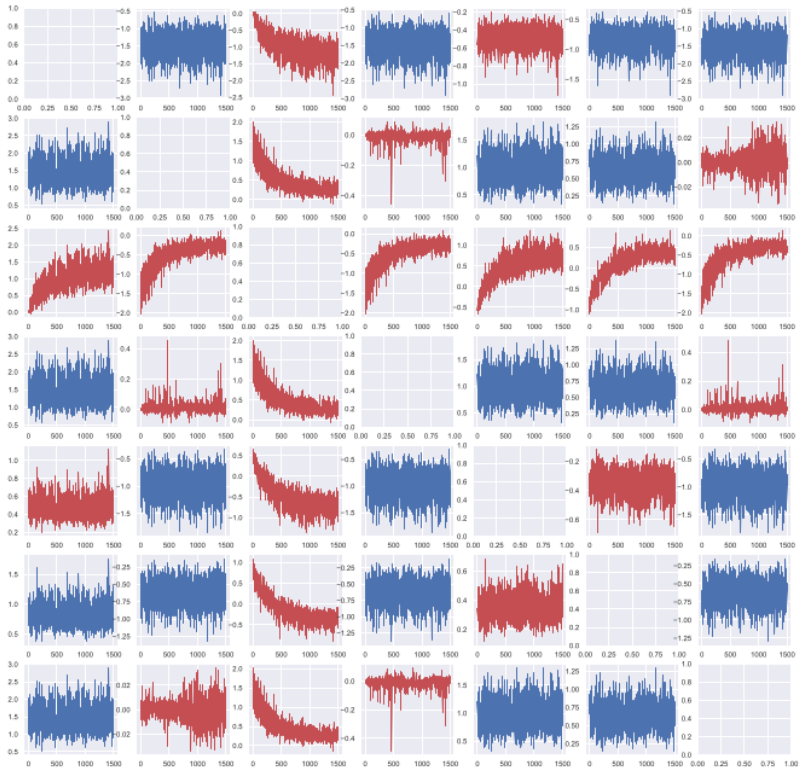

Figure A.3: Plots of loss differentials of every pair of the seven forecasts calculated using the covariance proxy when applied to IID simulated returns of 300 assets

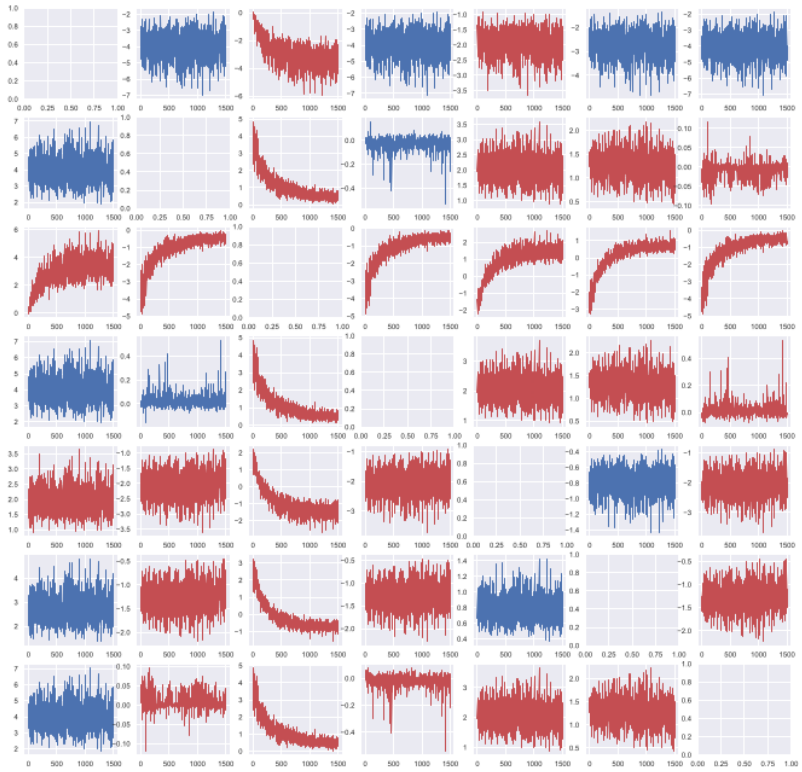

Figure A.4: Plots of loss differentials of every pair of the seven forecasts calculated using the covariance proxy when applied to IID simulated returns of 400 assets

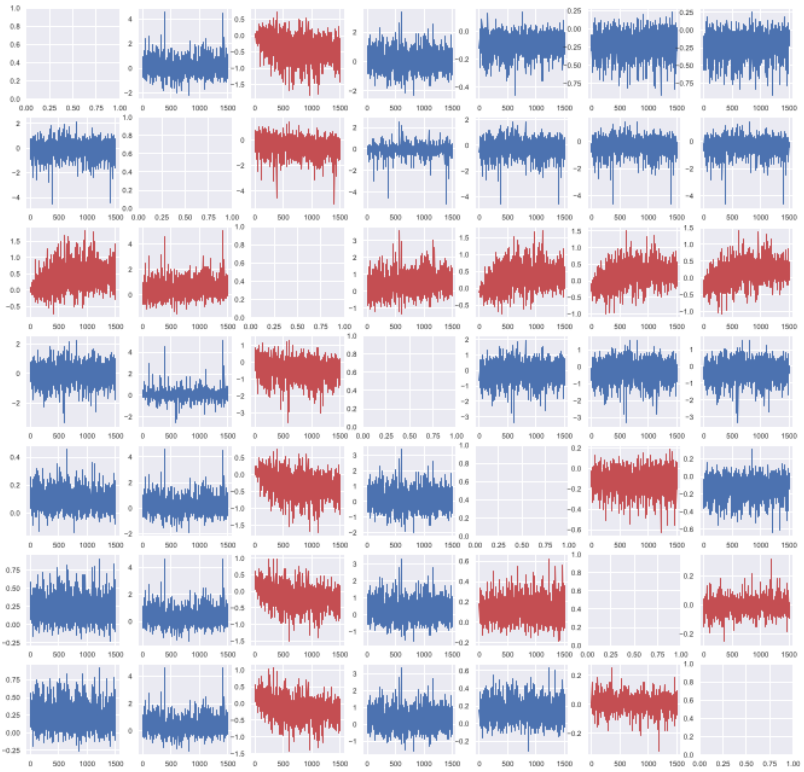

Figure A.5: Plots of loss differentials of every pair of the seven forecasts calculated using the covariance proxy when applied to CCG simulated returns of 100 assets

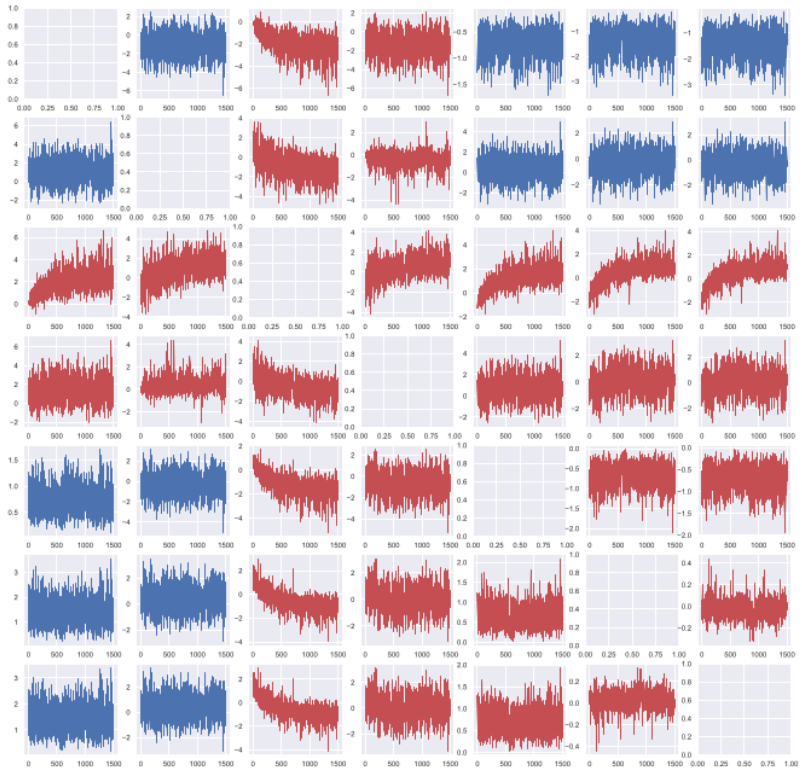

Figure A.6: Plots of loss differentials of every pair of the seven forecasts calculated using the covariance proxy when applied to CCG simulated returns of 200 assets

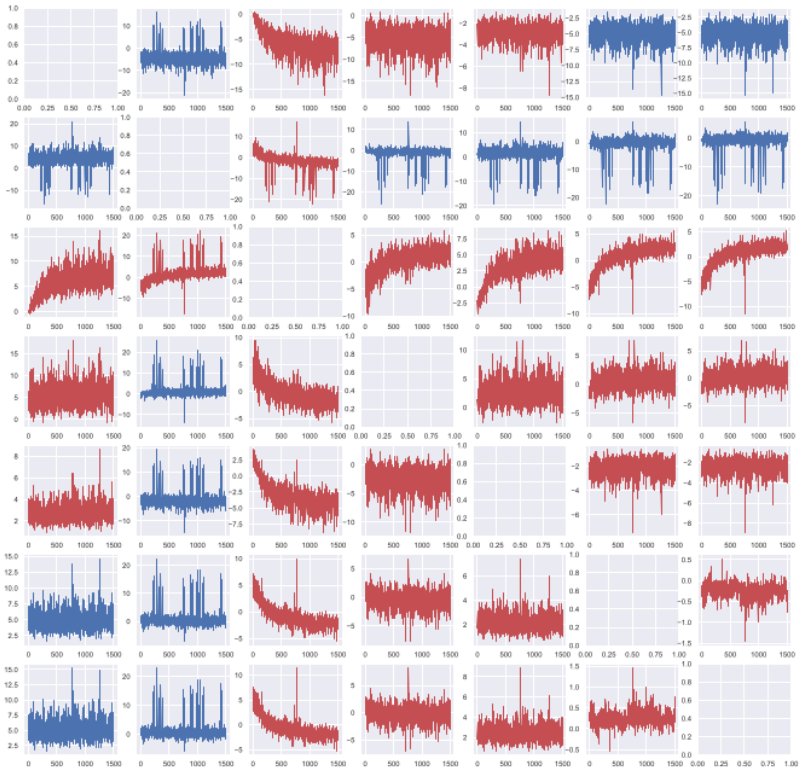

Figure A.7: Plots of loss differentials of every pair of the seven forecasts calculated using the covariance proxy when applied to CCG simulated returns of 300 assets

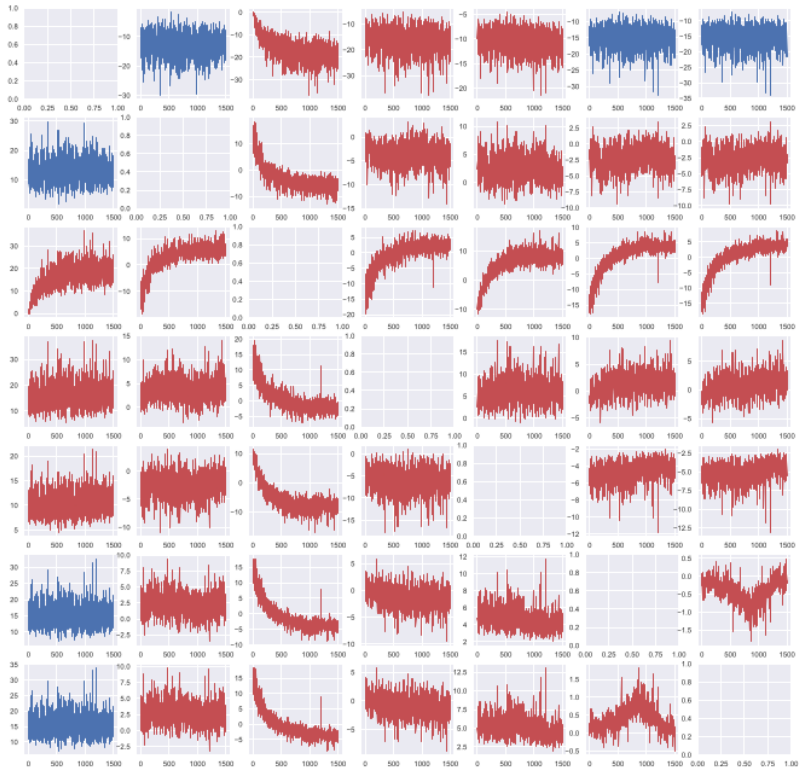

Figure A.8: Plots of loss differentials of every pair of the seven forecasts calculated using the covariance proxy when applied to CCG simulated returns of 400 assets

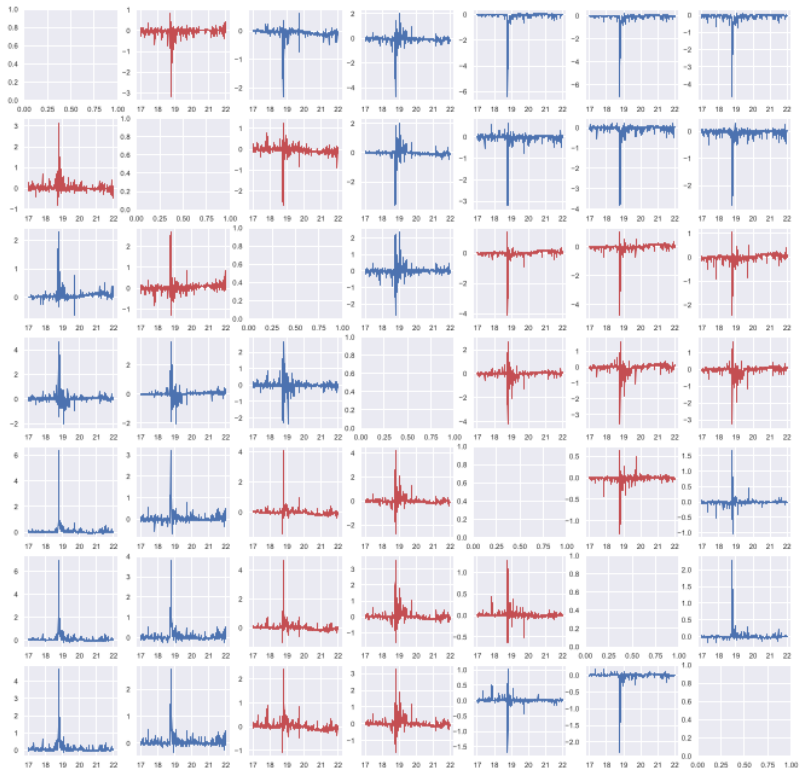

Figure A.9: Plots of loss differentials of every pair of the seven forecasts calculated using the covariance proxy when applied to historical stock returns from 2017-01-01 to 2022-01-06 of the top 100 stocks in S&P500 index

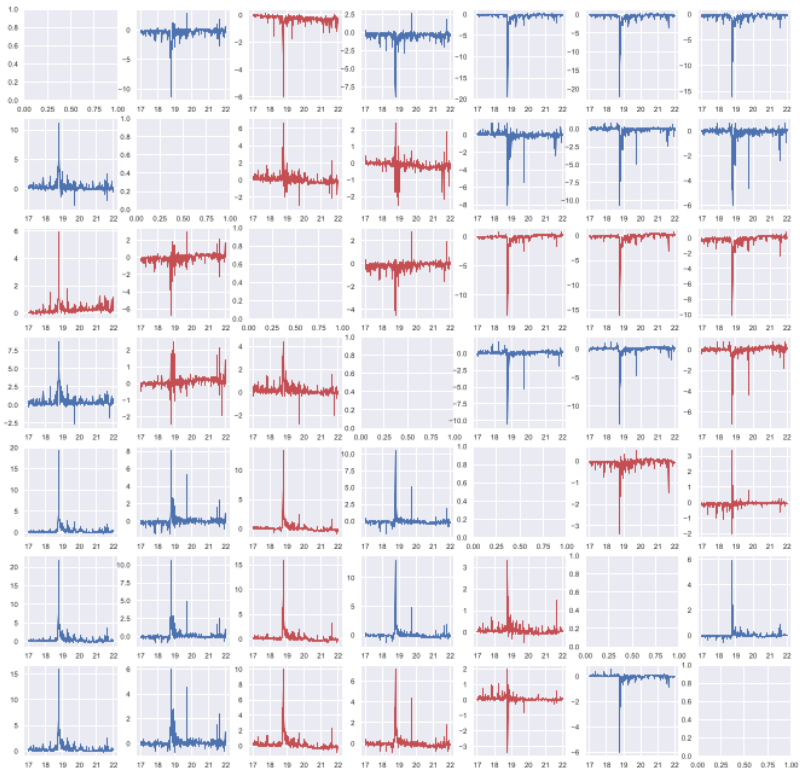

Figure A.10: Plots of loss differentials of every pair of the seven forecasts calculated using the covariance proxy when applied to historical stock returns from 2017-01-01 to 2022-01-06 of the top 200 stocks in S&P500 index

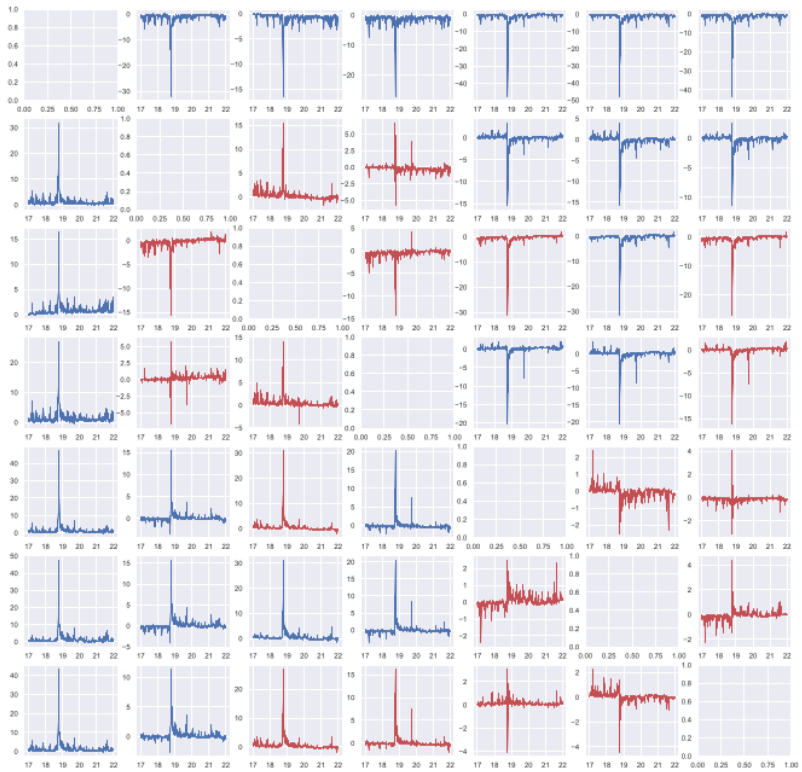

Figure A.11: Plots of loss differentials of every pair of the seven forecasts calculated using the covariance proxy when applied to historical stock returns from 2017-01-01 to 2022-01-06 of the top 300 stocks in S&P500 index

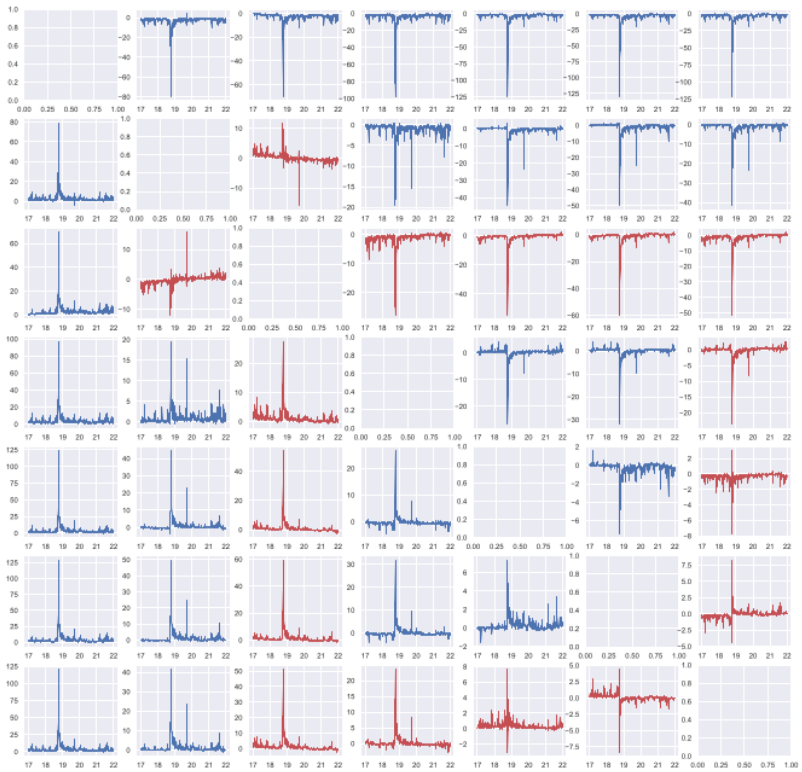

Figure A.12: Plots of loss differentials of every pair of the seven forecasts calculated using the covariance proxy when applied to historical stock returns from 2017-01-01 to 2022-01-06 of the top 400 stocks in S&P500 index

Bibliography

[1] Tim Bollerslev. Modelling the coherence in short-run nominal exchange rates: A multivariate generalized arch model. *The Review of Economics and Statistics*, 72(3):498–505, 1990.

[2] Francis Diebold and Roberto Mariano. Comparing predictive accuracy. *Journal of Business & Economic Statistics*, 13(3):253–63, 1995.

[3] Yingjie Dong and Yiu-Kuen Tse. Forecasting large covariance matrix with high-frequency data using factor approach for the correlation matrix. *Economics Letters*, 195:109465, 2020.

[4] Peter Hansen and Asger Lunde. Consistent ranking of volatility models. *Journal of Econometrics*, 131(1-2):97–121, 2006.

[5] Laurent Laloux, Pierre Cizeau, Marc Potters, and Jean-Philippe Bouchaud. Random matrix theory and financial correlations. *International Journal of Theoretical and Applied Finance (IJTAF)*, 03(03):391–397, 2000.

[6] Sébastien Laurent, Jeroen V. K. Rombouts, and Francesco Violante. On the forecasting accuracy of multivariate garch models. *Journal of Applied Econometrics*, 27(6):934–955, 2012.

[7] Sébastien Laurent, Jeroen V.K. Rombouts, and Francesco Violante. On loss functions and ranking forecasting performances of multivariate volatility models. *Journal of Econometrics*, 173(1):1–10, 2013.

[8] Olivier Ledoit and Michael Wolf. Improved estimation of the covariance matrix of stock returns with an application to portfolio selection. *Journal of Empirical Finance*, 10:603–621, 01 2001.

[9] Olivier Ledoit and Michael Wolf. Honey, i shrunk the sample covariance matrix. *The Journal of Portfolio Management*, 30(4):110–119, 2004.

[10] Olivier Ledoit and Michael Wolf. A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis*, 88(2):365–411, 2004.

[11] Harry M. Markowitz. *Portfolio Selection: Efficient Diversification of Investments*. Yale University Press, 1959.

[12] Alexander J. McNeil, Rüdiger Frey, and Paul Embrechts. *Quantitative Risk Management: Concepts, Techniques and Tools Revised edition*. Number 10496 in Economics Books. Princeton University Press, 2015.

[13] M. Pakkanen. Math97133 advanced statistical finance lecture notes. 03 2020.

[14] Andrew Patton and Kevin Sheppard. Evaluating volatility and correlation forecasts. *Handbook of Financial Time Series*, pages 801–838, 2009.

[15] Andrew J. Patton. Volatility forecast comparison using imperfect volatility proxies. *Journal of Econometrics*, 160(1):246–256, 2011.

[16] Vasiliki Plerou, Parameswaran Gopikrishnan, Bernd Rosenow, Luís Amaral, Thomas Guhr, and H. Stanley. A random matrix approach to cross-correlations in financial data. *Physical review. E, Statistical, nonlinear, and soft matter physics*, 65:066126, 07 2002.

[17] Marc Potters and Jean-Philippe Bouchaud. *A First Course in Random Matrix Theory: for Physicists, Engineers and Data Scientists*. Cambridge University Press, 2020.

[18] William Sharpe. A simplified model for portfolio analysis. *Management Science*, 9(2):277–293, 1963.

[19] Efthymia Symitsi, Lazaros Symeonidis, Apostolos Kourtis, and Raphael Markellos. Covariance forecasting in equity markets. *Journal of Banking Finance*, 96:153–168, 2018.

[20] L. Sánchez-Betancourt. Math97108 quantitative risk management lecture slides. 2021.

[21] Liusha Yang, Romain Couillet, and Matthew R. McKay. Minimum variance portfolio optimization with robust shrinkage covariance estimation. In *2014 48th Asilomar Conference on Signals, Systems and Computers*, pages 1326–1330, 2014.