On the economic and statistical properties of high dimensional covariance estimation techniques: A comparative analysis of commodity portfolios*

Name student: Abinash Sant
Student ID number: 431421

ERASMUS UNIVERSITY ROTTERDAM
Erasmus School of Economics
Master Thesis Msc Data Science & Marketing Analytics

Supervisor: Dr. Eran Raviv
Second Assessor: Dr. Kathrin Gruber

27/05/2023

Abstract

In the pursuit of accurate high dimensional covariance estimation, many methods have been proposed to remedy the curse of dimensionality. This research compares these methods' performance and their susceptibility to statistical and economic factors to determine what method is most suitable for commodity portfolios. Results show that in every situation the optimal method is either Block-DECO or Non-Linear Shrinkage. These two methods are on the extreme opposites of the bias-variance trade-off and methods that compromise between the two consistently fall short.

*The views stated in this thesis are those of the author and not necessarily those of the supervisor, second assessor, Erasmus School of Economics or Erasmus University Rotterdam

Contents

1. Introduction

The covariance matrix is a representation of the relationships between the dispersion of variables in data. As such, it plays a key role in numerous statistical methods that have countless applications, ranging from dimensionality reduction and signal processing to volatility modeling of asset returns for portfolio management (Ledoit & Wolf, 2022). True covariance however, is a theoretical quantity. It requires complete information of all factors that influence it and thus cannot be observed directly. Thus, we are forced to rely on estimations that approximate the true covariance.

The default estimator for covariance estimation is the sample covariance matrix (SCM). Given data where the number of features is trivial compared to the number of observations, the sample covariance matrix suffices, but in applications such as portfolio management and signal processing this rarely happens, exposing its crucial shortcoming; its susceptibility to high dimensional data. As the dimensionality of data increases, the estimate becomes less accurate until the number of dimensions surpasses the sample size, where the SCM is no longer invertible, rendering it inadequate for many statistical procedures. (Jobson & Korkie, 1980)

As such, applications that demand high dimensional data such as portfolio optimization and signal processing look to more novel estimation methods. Over the last few decades, many such methods have been proposed and implemented successfully in a wide variety of applications, yet absence of conclusive comparative evidence implies that it is not known which method is most suitable in which situation. This thesis aims to compare several of the most noteworthy covariance estimation techniques. While this thesis is written with portfolio optimization as application, the methods discussed and analyzed have been used in numerous applications in signal processing, imaging, AI and medicine. With this, the objective is to improve understanding of these methods not just for data science and finance, but a wide range of disciplines.

The main research question for this thesis is:

How does the performance of different covariance estimation techniques differ, when applied in the context of portfolio optimization?

In order to dissect this further, a distinction must be made between the variables that exert possible influence on these results. We distinguish primarily between financial and statistical properties of statistical methods. Financial properties are those related to the economic processes reflected in the data such as the presence of recessions and high volatility periods. Statistical properties pertain to dimensionality, forecast horizon and estimation hyperparameters such as window length for covariance computation. The definition of performance will be derived from an aggregate of the methods' Sharpe Ratios across different simulated series to account for variance.

This leads to two additional null hypotheses that function as components of the overarching research question:

The effects of statistical properties are independent of estimation techniques

The effects of financial properties are independent of estimation techniques

If all methods are influenced in a similar manner by change in properties, then a single estimation technique should be the most suitable in almost every scenario, but if methods behave differently relative to changes in these properties, the optimal estimation method will depend on the statistical and economic properties.

Section 2 will contain the theoretical foundations of this research. This can be roughly divided into three parts. 1) an introduction of essential concepts such as covariance and modern portfolio theory, 2) academic literature and theory on the problems faced in covariance estimation for portfolio theory and 3) academic literature and theory on the estimation procedures that function as focal point of this research, which aim to solve the aforementioned problems. Section 3 will contain the methodology and description of the data, sampling methods and simulated series used. Section 4 presents and discusses the results of the analysis and Section 5 will provide concluding thoughts and reflections.

2. Theoretical Foundations

In section 2.1 we will first cover the fundamentals of covariance as a statistical concept and how it is utilized in portfolio optimization. Section 2.2 will provide an introduction to the Sample Covariance Matrix, the default go-to method for estimating covariance. Section 2.3 will highlight the downsides and shortcomings of this method, explaining why the Sample Covarince Matrix might not always be suitable. Sections 2.4 to 2.7 will discuss a number of possible alternatives to the Sample Covariance Matrix that mitigate or circumvent the issues raised in section 2.3.

2.1 Covariance and Modern Portfolio Theory

Covariance is a statistical measure of joint variability representing the degree to which deviation from the mean in one variable is paired with deviation from the mean in another variable. This value is positive if both variables diverge in the same direction (negative or positive) and is negative if the variables drift in opposite directions. Covariance has countless application, but for now only its relevance to portfolio theory will be discussed.

Modern Portfolio Theory (MPT) focuses on constructing Mean-Variance efficient portfolios by selecting weighted assets to either maximize returns under a volatility constraint, or minimize volatility under a returns constraint (Markowitz, 1952). These are known as efficient portfolios. Additionally the portfolio with the lowest volatility without any return constraints is known as the Global Minimum Variance Portfolio (GMVP).

In either scenario, minimizing volatility across multiple assets is done by means of diversification, where lower than expected returns on an asset are counteracted by higher than expected returns on a different asset to keep returns as stable as possible. In other terms, we seek assets that exhibit negative covariance.

For a portfolio of k assets, where μ is a k x 1 vector of asset return means, $\mathbf{x}$ is a $k \times 1$ vector of weights, $\mathbf{1}$ is $k \times 1$ vector where $\mathbf{1}_i = 1 \; \forall \; i \in \{1, \ldots, k\}$ and Σ is the $k \times k$ covariance matrix, the expected returns and volatility of the portfolio are given by

$$\mu_{p,x} = \mathbf{x}'\mu \tag{1}$$

and

$$\sigma^2_{p,x} = \mathbf{x}'\Sigma\mathbf{x} \tag{2}$$

Minimizing the variance of the GMVP by optimally selecting weights then solves

$$\begin{aligned} \min_x \sigma^2_{p,x} &= \mathbf{x}'\Sigma\mathbf{x} \; s.t. \; \mathbf{x}'\mathbf{1} = 1 \\ \min_x \sigma^2_{p,x} &= \mathbf{x}'\Sigma\mathbf{x} \; s.t. \; \mathbf{x}'\mathbf{1} = 1 \; \wedge \mu_{p,x} = \mu_p \\ \max_x \mu_{p,x} &= \mathbf{x}'\Sigma\mathbf{x} \; s.t. \; \mathbf{x}'\mathbf{1} = 1 \; \wedge \sigma^2_{p,x} = \sigma^2_p \end{aligned} \tag{3}$$

for the GMVP, fixed returns efficient portfolios and fixed volatility efficient portfolios respectively as described by Zivot (2016).

This presence of Σ in this optimization clarifies why an accurate calculation of asset weights requires an accurate estimate of the covariance matrix.

2.2 Sample Covariance Matrix

The first issue arises when we attempt to establish what is considered an accurate estimate. As mentioned before, true covariance is a theoretical quantity that cannot be observed. It would require complete and perfect knowledge of not only the entire population, but all external factors that exert influence. In this setting of financial assets that would include not only all financial assets, but also news, trade, weather, sociopolitical developments etc.

As a result we must settle for the next best alternative, which is the covariance of a simpler system of which we assume all knowledge is available. If it is assumed that a drawn sample is sufficiently representative of the population, then we treat this as a closed system of which to calculate the covariance.

This is known as the Sample Covariance Matrix (SCM) and is the default estimation procedure to such extent that it has become virtually synonymous with true covariance.

Given multivariate data with variables x and y and n observations,

$$Cov_{x,y} = \frac{\Sigma_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}, \tag{4}$$

In this equation, $(x_i - \bar{x})(y_i - \bar{y})$ is the product of the respective deviations from the mean, of which the expected value is computed using Bessel's correction to eliminate the bias caused by the non-zero correlation between individual observations and the sample mean (Rosenthal, 2015), (Rumsey, 2009).

The SCM of k variables will then yield a $k \times k$ matrix. Because $Cov_{i,i} = Var_i$ and $Cov_{i,j} = Cov_{j,i} \forall\ i, j \in \{1, \ldots, k\}$. this matrix is symmetric and contains the variances on the main diagonal written as

$$\Sigma = \begin{bmatrix} Var_1 & \ldots & Cov_{1,k} \\ \vdots & \ddots & \vdots \\ Cov_{k,1} & \ldots & Var_k \end{bmatrix} \tag{5}$$

2.3 The Curse of Dimensionality and Matrix Singularity

While the Sample Covariance Matrix is exceedingly useful and common, it is not flawless. For high dimensional data, the Sample Covariance Matrix struggles to yield comparable results across various data sets (Jobson & Korkie, 1980), which is just one manifestation of a group of statistical phenomena known as the Curse of Dimensionality.

In low dimensional feature spaces, it is common for an observation to be near the mean in every dimension, but with each dimension that is added, an opportunity arises to diverge in this new dimension. As the number of dimensions increases it becomes more likely that observations will find themselves in the extreme reaches of at least one dimension. Thus, as the number of dimensions keeps growing, every observation is bound to become a statistical outlier. Because covariances are scalar products in Euclidean space, they are dictated by the behavior of Euclidean distances in high dimensions as described by Aggarwal, Hinneburg & Keim (2001). and the ratios of distances between points converges to 1 as all observations become more equally spaced (Beyer, Goldstein, Ramakrishnan & Shaft, 1999).

As a result, high dimensional data requires vast amounts of data to form accurate estimates. for k dimensions, $\frac{k(k-1)}{2}$ covariances must be estimated. Even when working with only 20 assets, monthly return data over 16 years would yield only 192 observations to estimate 190 parameters.

This causes the covariance estimates to contain more extreme values, not because this is a reflection of reality, but because the estimation error is extreme (Ledoit & Wolf, 2003). An asset whose estimation error is extreme, might exhibit extremely high returns or low volatility. Given that this is considered favorable, a large portfolio weight will be assigned to this asset. Inversely, low estimation error assets that show more

conventional returns and volatility will receive fairly little weight. As a result, the portfolio will be dominated by the assets with the largest estimation error (Michaud, 1989).

This dimensionality problem is most severe once the number of dimensions equals or exceeds the number of observations in the data. Consider a matrix $X_{m\times n}$ with m n-variate observations. If $m = n$, its Covariance Matrix Σ is a $n\times n$ square matrix. The complication arises with the centering of the data values, as covariance is computed relative to the mean as shown in equation 4. Because $\Sigma(x_i - \bar{x}) = 0$ by construction, only $n-1$ rank 1 matrices are required, and the final matrix is deterministic. Thus, rank(Σ) cannot exceed $n-1$. As a result, the covariance matrix is singular and its inverse Σ^{-1} is undefined. Because solving the constrained optimizations of equation (3) by means of a Lagrangian multiplier requires Σ^{-1} as described in Chapter 12 of Zivot (2016), a sample covariance matrix cannot be computed when $m \geq n$. Thus, for both cases where $m \geq n$ and where m is commensurate with n, alternative methods of estimation are a necessity.

2.4 GARCH-based Estimators

In search of alternative ways to estimate covariance, we first turn to the estimation of the individual time-varying volatility of a time-series. For financial assets, the GARCH (Generalized AutoRegressive Conditional Heteroscedasticity) model is among the most popular and successful methods, and functions as the basis for several covariance estimation techniques.

The concept behind GARCH is that an observation of high returns is more likely caused by a probable event under high volatility than by an improbable event under low volatility. The returns are decomposed into two parts, the mean and the innovation. The latter of these is then further split into the square root of volatility and a standard normal distribution.

$$r_t = \mu_t + \epsilon_t = \mu_t + \sigma_t z_t \tag{6}$$

where $z_t = N(0,1)$.

Assuming $\mu_t = 0$, this yields

$$r_t = \epsilon_t = \sigma_t z_t \tag{7}$$

The volatility is than defined as

$$\sigma_t^2 = \omega + \sum_{p=1}^{P} \alpha_p \sigma_{t-p}^2 \; + \sum_{q=1}^{Q} \beta_q r_{t-q}^2 \tag{8}$$

for a (P,Q) GARCH model with P autoregressive terms and Q moving average terms. Given our earlier assumptions that the data is demeaned and z_t has unit variance, the volatility of the returns r_t are defined as σ_t^2 to model the returns volatility of each individual asset (Engle, 1982). While there are multivariate adaptations of this model such as those proposed by Bollerslev, Engle & Woolridge (1988), these require the estimation of large amounts of parameters through Maximum Likelihood Estimation, making these methods impractical for the high dimensional settings discussed here.

2.4.1 Constant Conditional Correlation

Building forth on these aforementioned multivariate GARCH models, restriction were imposed on the structure of the model in order to simplify the estimation procedure and shorten the computation. Constant Conditional Correlation (CCC) achieves this by using the correlation matrix R. The covariance matrix is related to the correlation matrix by the equation

$$\hat{\Sigma} = D_t R D_t \tag{9}$$

where $D_t = \text{diag}\{\sqrt{h_{i,t}}\}$ and $h_{i,t}$ is the volatility estimate of asset i at time t (Bollerslev ,1990). Even though the volatility estimates are computed differently for each t it should be noted that the correlation matrix R is time invariant and is computed only once. While the model is simpler and computationally feasible in high dimensions, the assumption that correlation remains constant over time becomes likely to be violated as the time span of the data increases. As such a less restrictive variant was created.

2.4.2 Dynamic Conditional Correlation

Dynamic Conditional Correlation (DCC) is closely related to CCC. Its main difference is that the static correlation matrix R that is used in CCC is now replaced with a dynamic time-variant correlation matrix. This changes the model to

$$\hat{\Sigma} = D_t R_t D_t. \tag{10}$$

This allows the model to account for changes in correlation by using a rolling window to compute correlations over a smaller time frame that experiences little to no interference from distant events (Engle & Sheppard, 2001), (Engle, 2002).

2.4.3 Dynamic Equicorrelation

In order to further reduce the computational demands for high dimensional data, Dynamic Equicorrelation (DECO) was introduced by Engle & Kelly (2012). The idea of DECO is to assume that all pairwise correlations between time series at a given time t are equal. This allows for a far simpler computation of inverses and determinants, circumventing the issues presented by dimensionality.

While the method is computationally efficient, it imposes a very rigid structure. As such it is important to understand whether this equicorrelation assumption is realistic, and what happens if it is violated. Engle & Kelly (2012) show that if DCC provides a consistent estimator, then the DECO estimator is also efficient by extension, even if the equicorrelation assumption is violated.

The form of the equation is identical to that of DCC, except the dynamic correlation matrix R_t is replaced by

$$R_t = (1 - \rho_t) I_n + \rho_t J_n \tag{11}$$

where ρ_t is the equicorrelation, I_n is an n-dimensional identity matrix and J_n is an $n \times n$ square matrix of where all elements are 1.

2.4.4 Block-DECO

Despite the fact that DECO is a consistent estimator even if the equicorrelation assumption is violated, its estimation can potentially be improved by loosening the imposed equicorrelation structure. Block-DECO applies equicorrelation in groups of time series known as blocks instead of across the entire set of time series. This allows for variation in equicorrelation between industries for instance. A K-block Block-DECO model is specified using

$$R_t = \begin{bmatrix} (1-\rho_{1,1,t})I_{n_1} & 0 & \dots \\ 0 & \ddots & 0 \\ \vdots & 0 & (1-\rho_{K,K,t})I_{n_K} \end{bmatrix} + \begin{bmatrix} \rho_{1,1,t}J_{n_1} & \rho_{1,2,t}J_{n_1 \times n_2} & \dots \\ \rho_{2,1,t}J_{n_2 \times n_1} & \ddots & \\ \vdots & & \rho_{K,K,t}J_{n_K} \end{bmatrix} \tag{12}$$

where n_l is the number of dimensions in block $l \in 1, \ldots, K$ and $\rho_{i,j,t}$ is the average of the DCC correlation matrix in each block (Engle & Kelly, 2012) defined as

$$\rho = \frac{\sigma^2 - \sum_{j=1}^{n} w_j^2 s_j^2}{\sum_{i \neq j} w_i w_j s_i s_j} \tag{13}$$

The form of equation 12 ensures that the computed matrix remains semi positive definite. It should be noted that blocks consisting of a single time series cannot be computed, as the denominator of equation 13 requires input of different time series.

2.5 Shrinkage Estimators

Besides the GARCH-based estimators described thus far, there are also methods that do not bypass the SCM, but manipulate it. As discussed in previous sections, the curse of dimensionality causes the SCM to contain extreme values with large errors, regardless of its lack of bias. Shrinkage is the process of pulling values back to more centered values of the matrix based on how extreme they are, effectively compressing the matrix to impose structure as proposed by Stein (1956), (Touloumis, 2015).

2.5.1 Linear Shrinkage

The simplest form of shrinkage is outlined by Ledoit & Wolf (2003). Instead of using the high error SCM or a structured estimator with a high bias such as a single factor model, a compromise between two extremes is achieved by means of a convex combination of the two, resulting in the form

$$\hat{\Sigma} = \delta F \ + (1 - \delta) S \tag{14}$$

where F is the shrinkage target, S is the SCM and δ is the shrinkage constant where $0 > \delta > 1$.

First a shrinkage target is selected, such as the Identity matrix, the single-factor model by Sharpe (1963) or the CCC model. Then the optimal value for the shrinkage constant is computed by minimizing a loss function defined as the Frobenius norm of $\hat{\Sigma} - \Sigma$ as defined in Appendix B of Ledoit & Wolf (2003). Note that the same shrinkage intensity δ^* that is derived is used for the shrinkage of every eigenvalue. The resulting estimate is an adjusted form of the SCM whose values have been pulled towards a more structured matrix in order to counteract the extreme errors that plague high dimensional estimates.

2.5.2 Non-Linear Shrinkage

The problem posed by linear shrinkage methods, is that they require prior knowledge on the covariance to establish the shrinkage target. In 2011, Ledoit & Peche proposed an estimator that does not require any prior knowledge of the covariance, and instead relies solely on the distribution of sample eigenvalues.

The Spectral Decomposition of the SCM is

$$S_T = Q_T \Lambda_T Q_T^{-1} \tag{15}$$

where the columns of Q are the eigenvectors of S_T and Λ is a diagonal matrix with the corresponding eigenvalues on the main diagonal. We can then express the shrinkage estimate as

$$\hat{\Sigma}_T^* = Q_T \Delta_T^* Q_T^{-1} \tag{16}$$

Where Δ_T^* is a diagonal matrix not with the eigenvalues, but with the convex combinations of the SCM eigenvalues and the mean of the population eigenvalues as entries as defined by

$$\begin{aligned} \Delta_T^* &= \text{diag}(\delta_{T,1}^*, \ldots, \delta_{T,N}^*) \\ \delta_{T,i}^* &= \gamma_T^* \mu_T \ + (1 - \gamma_T^*) \lambda_{T,i} \end{aligned} \tag{17}$$

Because the shrinkage parameter γ is now separated from the eigenvalues and eigenvectors, they can be changed independently without endangering the positive-definite properties of the matrix, allowing for the use of different shrinkage intensities for different sample eigenvalues.

2.6 Non-Linear Kernel Estimation

Non-Linear Kernel Estimation also uses the Spectral Decomposition, but instead of obtaining an estimate by means of shrinkage, it uses Kernel Principal Component Analysis (KPCA).

2.6.1 Principal Component Analysis

Principal Component Analysis (Pearson, 1901) is a dimensionality reduction method. It creates a principal component by taking the linear combination of original features with the highest variance and uses it as new basis. For subsequent principal components, the same method is used with as requirement that the principal components are orthogonal.

After this process has reached the same number of principal components as the original number of features, all of the variance in the data is explained, but now by a new set of variables that are sorted by their eigenvalues and variance explained. This allows for the dimensionality of the data to be reduced by only taking the first n components to maximize the variance with n features, after which the data is rotated to align with the principal component axes.

For this thesis, the step of selecting n components will be omitted and the method will be utilized as an analysis tool for estimating covariance through rotation, given that we do not wish to reduce the dimensionality.

2.6.2 Kernels & the Kernel Trick

The linearity of data is relative to its dimensionality. Data generation and distinction could be non-linear in two dimensions, but linear in three or more dimensions. This is the main concept behind the Support Vector Machine (SVM) and other kernel methods. This means that if we want to use a linear method like PCA to make non-linear estimates, this is possible, but requires us to connect the input space to a high dimensional space using a non-linear feature map.

An increase in the dimensionality of data usually comes paired with an exponential increase in required computational power to make calculations with that data. As such, it would be unrealistic to map data to very large numbers of dimensions for this purpose. This is where the kernel trick appears.

The kernel functions from which the kernel trick owes its name allow for calculations in high dimensional and infinite dimensional feature spaces implicitly, meaning the data does not need to be translated to these higher dimensions in order to perform the calculations.

Firstly, in order for the trick to operate, confirmation is needed that the required rules of calculus are valid in this high dimensional space. This requires the vector space to be complete, such as $\mathbb{R}$, and have an inner product that maps vector multiplication to scalars such as the dot product. Under these circumstances, this is called a Hilbert space, denoted by $\mathcal{H}$ (von Neumann, 1930).

If $\mathcal{X}$ is the input data $x \in \mathcal{X}$ and $f(x)$ maps these to $\mathbb{R}$ using the functions $f \in \mathcal{H}$ that are in the Hilbert space, the function $\mathcal{L}_x$ is defined as

$$\mathcal{L}_x : \mathcal{H} \to \mathbb{R}, \text{where } \mathcal{L}_x(f) = f(x) \; \forall \; f \in \mathcal{H}. \tag{18}$$

This function $\mathcal{L}_x$ takes on the value of x for all $f \in \mathcal{H}$. Some of these functions possess the reproducing property. If for every $f \in \mathcal{H}$, there exists a unique $K_x \in \mathcal{H}$ such that $\langle f, K_x \rangle = f(x)$, then $\mathcal{L}_x$ is considered continuous for every $f \in \mathcal{H}$ and has the reproducing property (Riesz, 1907 & 1909).

As the name implies, This means there are two ways of reaching $f(x)$, through the function $\mathcal{L}_x(f) = f(x)$, and a reproduction through the inner product of f and a unique function K_x. If this property is present in $\mathcal{L}_x$ for every $x \in \mathcal{X}$, then the corresponding $\mathcal{H}$ is a reproducing kernel Hilbert space (RKHS).

Here it should be noted that $\mathcal{L}_x$ is continuous at K_x by definition because $K_x \in \mathcal{H}$. Hence at a different point $y \in \mathcal{X}$,

$$\mathcal{L}_y(K_x) = K_x(y) = \langle K_x, K_y \rangle \tag{19}$$

Because this holds for every value of $x \in \mathcal{X}$,

$$K(x, y) = \langle K_x, K_y \rangle. \tag{20}$$

Note that K is independent of input values and this single kernel can be used for all functions $f \in \mathcal{H}$, not unlike a master key that fits in any lock. This is the reproducing kernel that enables the kernel trick.

Without the kernel trick, data needs to be transformed from $\mathcal{X}$ to the high or infinite dimensional $\mathcal{H}$ using a feature map, before inner products are computed in this high or infinite dimensional space. However, if a reproducing kernel can be established, the original data does not need to be transformed, as the reproducing kernel $K(x, y)$ takes values from $\mathcal{X}$ as input.

2.6.3 Kernel Principal Component Analysis

Combining the kernel trick with PCA creates a version that can account for non-linear interactions because Kernel Principal Component Analysis (KPCA) does not yield principal components as linear combinations in the input space, but projections in the kernel feature space onto the principal components (Schölkopf et al., 1998).

The kernels used for KPCA are the Gaussian kernel and the polynomial kernel up to a degree of $q = 5$. These options were selected as they are commonly utilized in machine learning (Peng et al., 2019).

2.7 Random Matrix Filtering

Random Matrix Theory (RMT) was pioneered by Dyson & Mehta in the 1960's for the purpose of energy levels in nuclear physics, but has since been applied successfully to finance (Mehta & Dyson, 1963). The main concept of Random Matrix Filtering for portfolio optimization is that the distribution of eigenvalues of a random matrix depends on the distribution of the random variable from which the matrix elements are sampled, and that there are theoretical boundaries to the eigenvalues of a random matrix. Vice versa, this means that if the eigenvalues of a matrix fall outside of these boundaries, there is sufficient indication that the matrix deviates from the random distribution and thus contains information. Additionally, this information would be contained in the eigenvalues outside the boundaries, while noise is contained in the eigenvalues that lie within the boundaries (Marčenko & Pastur, 1967),

The Seminal work by Marčenko and Pastur shows that for a $m \times n$ random matrix with mean 0 and finite variance, the theoretical bounds for the eigenvalues are defined by

$$\lambda_{\pm} = \sigma^2 \left(1 \pm \sqrt{\frac{m}{n}}\right)^2 \tag{21}$$

where σ^2 is the finite variance of the matrix, λ are the eigenvalues, and that $\frac{m}{n}$ converges to $\lambda \in (0, \infty)$ as $m, n \to \infty$.

Spectral Decomposition of a covariance matrix as in equation 15 isolates the eigenvalues on the main diagonal of Λ. This allows for evaluation of which eigenvalues are greater than the upper bound as defined by the Marčenko-Pastur Distribution. The eigenvalues that fall below the upper bound are considered to generate noise. The filtering process consists of taking all noisy eigenvalues, and replacing them with their mean so that all noisy eigenvalues are equal, yet the trace remains unchanged to ensure positive definiteness. Λ in the Spectral Decomposition is then replaced with the filtered version Λ^* and the filtered covariance matrix is computed as $Q\Lambda^*Q^{-1}$.

3. Data & Methodology

In section 3.1 an overview of the data, its acquisition and key characteristics is provided. Section 3.2 explains the process of using this data to create simulated series with which the research will be conducted. Section 3.3 discusses the methods through which the simulated series are tuned individually, and the different specifications that are accounted for such as time period and window length. At last section 3.4 will explain how the acquired simulated series will be used to generate and evaluate results.

3.1 Description

The original data set used for this thesis was provided by Dr. Raviv, Erasmus University Rotterdam. The dataset contains 31 time series of commodity returns, spanning the period January 3rd 1990 to July 2nd 2021. The early period of the data set does not contain all the commodities. Maintaining zero-value data would lead to perfect multicollinearity. While it is possible to split the data in periods that are separated at points where data becomes available, this would lead to fluctuation in the number of commodities, obstructing the integrity of results drawn from a different number of available variables. Additionally, the predictive power of data this far back is questionable. To avoid incomplete data, only the period from September 1st 1998 onwards is used, as all commodities are available from this point.

Of the 31 commodities, most fit within a predefined category relating to food, oils or metals. The main exception to this is the natural gas commodity. Its behavior is unique among the commodities and causes complications during the computations of Block-DECO, as it relies on correlations between different time series within a block, which are nonexistent given that it is the only series within its block. With this in mind, it has been removed after careful consideration. While this provides a less accurate representation of reality, its removal from both test and training data for all methods will allow for fair comparison between different methods.

Training and test data are split according to the most recent available point that allows for the forecast e.g. for a 1 year forecast, the last year of data will be used as test data, with the remainder before that as training data.

3.2 Simulation

In order to account for idiosyncrasies in the data, multiple data sets will be simulated based on the original data provided that maintain the economic characteristics of the original, but allow random occurrences to even out across these data sets.

Simulating a commodity data set across its entire duration at once is not possible. This is because the variance of returns is not constant over time, and the simulations produce constant variance. This will yield a monotonous simulated series with constant variance that is effectively useless for any analysis.

Thus, the commodity returns time series are divided into segments separated by their volatility as implied by squared returns (Engle, 1982). These change points are estimated by a divisive hierarchical clustering algorithm devised by James & Matteson (2013, 2014). This non-parametric method clusters adjacent observations in a time series by their variance. This divides the time series into segments of approximately constant variance. Because this method demands a prespecified number of break points, a decision needs to be made how many points will be used. For this, natural numbers were used in ascending order to test whether all created segments were deemed significantly different from one another by using Levene's Test for equality of variances (Levene, 1960). This is done separately for each time series, and as such the placement and quantity of break points will differ across the commodities.

Figure 1a: Segment break points for CO1

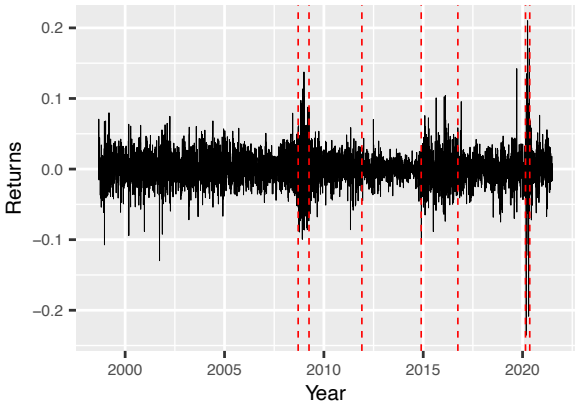
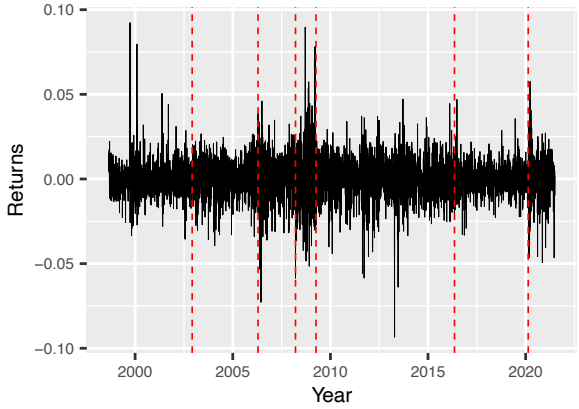


Figure 1b: Segment break points for GC1



Simulation will be performed using The Johnson's S_U distribution. This distribution is a transformation of the normal distribution that has been proven successful for portfolio management (Sin-Yi Tsai, 2011). The transformation is defined as

$$z = \gamma + \delta \sinh^{-1}(\frac{x - \xi}{\lambda}) \tag{22}$$

where $z \sim \mathcal{N}(0, 1)$ (Johnson, 1949) and the estimation of parameters is performed from quantiles according to the procedure described by Wheeler (1980).

After splitting each commodity time series into segments with approximately constant variance, Johnson's S_U Distribution is used to simulate a segment 10 times. For instance, the HG1 Copper commodity has 5 segments, each of which will have 10 simulated versions. To create a full simulated HG1 commodity time series, the segments are concatenated again.

Each of the 10 simulations will have idiosyncrasies in the distribution estimation, so to ensure that a simulated series is not based on a single distribution estimate, the simulation that is used to reconnect the series is random. The five segments from the simulated HG1 series would for instance be from simulation 4,3,9,9 and 1 respectively to form semi-randomized simulated HG1 time series.

This process is repeated ten times for all commodities, resulting in 10 simulated series for each commodity. To then form a full simulated data set, one of the ten simulated series is randomly drawn for each commodity to get a full-sized simulated data set. This random draw is repeated 20 times, to form 20 full-sized simulated data sets.

It should be noted that this procedure does not allow for a randomized validation data set. The autocorrelation processes in the data require chronological order to be maintained within segments of constant variance, any potential validation data would therefore be sampled at a different time period than the true test data.

Because the simulated data sets are based on data from the same time period as the original data, it does not allow for extrapolation into the future but mainly serves the purpose of ensuring robustness to random sampling errors.

The covariance estimation methods described previously are all tested with and without random matrix filtering (RMF), and for three types of constructed portfolios; an efficient portfolio with a returns constraint set at the highest mean asset return of the training data, an efficient portfolio with a returns constraint set at an equally weighted portfolio return of the training data, and the global minimum variance portfolio.

3.3 Parameter Optimization

For several estimation procedures, a window length parameter is required. While all these methods are performed with window lengths of 30, 60 and 90 days respectively, the short window is prone to highly fluctuating results. An optimal window length of n days is determined for each method in the range of 30 to 3000 days by constructing the corresponding portfolio type and optimizing the Sharpe ratio on the test data. Note that this optimal window length can fluctuate wildly and will not be consistent across the 20 simulated series, its main purpose of including it is not to function as variable, but as measure of how well hyperparameters from a previous period can be extrapolated. If the optimal window length of a previous period is nowhere near optimal in the subsequent period, it can be concluded that extrapolating hyperparameters to future time periods is of no use. Vice versa, if the optimal hyperparameters are consistent it could indicate the data is suitable for predictive analyses.

When RMF is applied, the eigenvalue boundaries are computed on variance. By default, the window length over which this variance is computed is set equal to the window length of the estimation procedure itself, but is optimized independently in the case the procedure does not require a window length parameter such as is the case for CCC and DECO.

The testing procedure is performed with the historical training data up to June 31st 2018, 2019 and 2020 respectively. With the forecast horizon and the test data consisting of the remainder until June 31st 2021. For all three points, the procedures are performed with a forecast horizon of one year, and multiple years whenever possible to account for both the effects of the specific time period and the forecast horizon at large. For all methods, the risk free rate utilized is the annualized rate on a US treasury bill with the corresponding start time and a duration equal to the forecast horizon.

3.4 Testing

The testing procedure consists of covariance estimation with the method of choice using the previously obtained hyperparameters. The method is applied to the 20 simulated data sets at the various time periods and forecast horizons to compute the portfolios, which are then evaluated using the cumulative product returns, the Sharpe ratio and the volatility. The parameters of the 20 simulated data sets are then aggregated to both remove idiosyncrasies as well as quantify the measure of variance between data sets. The key parameters used to assess performance will be the mean Sharpe ratio $\bar{S}$ across the 20 simulated data sets, its standard deviation σ_S and the ratio $\frac{\bar{S}}{\sigma_S}$, the Sharpe Ratio Coefficient of Variation (SRCV Henceforth). The purpose of this metric is to not just measure the performance of the method on a data set, but to also get an indication of how robust the method is to randomness in the data.

To analyze the effect of the parameters on the performance measures, Elastic Net penalized regression will be used. This choice was made after consideration of the number of factor levels used in the regression. While LASSO regression would aid in variable selection by shrinking coefficients to 0, it is also unstable and highly dependent on the data. By using an Elastic Net whose parameters are optimized using 10-fold cross validation, the optimal combination of Ridge and LASSO will be selected based on performance.

4. Results

Section 4.1 will go over the results of all methods that have been discussed, whereas section 4.2 and 4.3 further analyze the two most successful methods the see what sets them apart.

4.1 General Results

Table 1: Table of coefficients of Elastic Net Regression of factor variables on the Sharpe Ratio Coefficient of Variation (SRCV) across 20 simulated data sets by estimation method (all values displayed in this table are independent from other factors and relative to the intercept)

	Factor variables	SCM	CCC	DCC	DECO	BLOCK-DECO	Linear Shrinkage
	Intercept	0.917	1.128	0.929	1.064	1.317	1.187
1	Start July 1st 2019	−0.589	−0.879	−0.781	−0.847	−1.180	−0.897
2	Start July 1st 2018	−0.593	−0.904	−0.795	−0.861	−1.202	−0.927
3	GMVP	0.200	0.276	0.170	0.289	0.278	0.283
4	Equally Weighted Portfolio	0.181	0.283	0.174	0.283	0.289	0.292
5	2 year forecast horizon	0.444	0.612	0.430	0.655	0.646	0.604
6	3 year forecast horizon	0.439	0.604	0.403	0.600	0.615	0.592
7	Window length 60 days	−0.026		0.030			
8	Window length 90 days	−0.015		0.053			−0.026
9	Window length optimized	0.055		0.049			−0.004
10	RMF	0.210	0.072	0.252	0.002	−0.040	0.050

Table 2: Table of coefficients of Elastic Net Regression of factor variables on the Sharpe Ratio Coefficient of Variation (SRCV) across 20 simulated data sets by estimation method (all values displayed in this table are independent from other factors and relative to the intercept) (Cont'd)

	Factor variables	Non-Linear Shrinkage	KPCA Gaussian	KPCA POLY
	Intercept	1.015	1.132	1.002
1	Start July 1st 2019	−0.591	−0.917	−0.858
2	Start July 1st 2018	−0.638	−0.933	−0.955
3	GMVP	0.383	0.286	0.280
4	Equally Weighted Portfolio	0.394	0.277	0.266
5	2 year forecast horizon	0.457	0.477	0.387
6	3 year forecast horizon	0.477	0.479	0.364
7	Window length 60 days		0.021	
8	Window length 90 days	−0.024	−0.041	
9	Window length optimized	−0.001	−0.091	
10	RMF	0.035	0.093	0.095

The omitted reference levels that have been used in the computation of the intercept are "Start July 1st 2020", "Maximized Asset return Portfolio", "Window length 30 days" and "Window length 30 days". The factor level "Window length none" has been omitted due to multicollinearity with methods CCC, DECO and Block-DECO that don't utilize window length. Values omitted in the methods Linear Shrinkage, Non-Linear Shrinkage and Polynomial KPCA indicate that the coefficient was not significant and thus removed by the Elastic Net variable selection process.

One of the main insights of this table is that certain traits appear to to be consistent across every method. The SRCV decreases as the start date of the test data is pushed back in time, and the SRCV increases with an increase in forecast horizon. Additionally, portfolios based on the highest mean asset return in the training data (reference level) is heavily outperformed by both GMVP and Equally weighted portfolios.

When combining all potential combinations of factor levels, it is found that for 1 year forecast horizons starting on July 1st 2020, Block-DECO outperforms every other method regardless of portfolio type, window length or RMF. For every other combination of start time and forecast horizon, Non-Linear Shrinkage yields the highest SRCV, again regardless of other factors.

It should be noted here that while Non-Linear Shrinkage is still competitive in performance in the scenarios where Block-DECO is the best, Block-DECO performs poorly outside of the 1 year forecast starting in 2020.

To further analyze what sets these two successful methods apart from the rest, a visual representation of the previous table is provided. Figure 2 below contains figures plotting SRCV on the x-axis against RMF, start time, portfolio type, forecast horizon and window length respectively on the y-axis for all methods.

Figure 2a: Comparison of RMF effects on SRCV

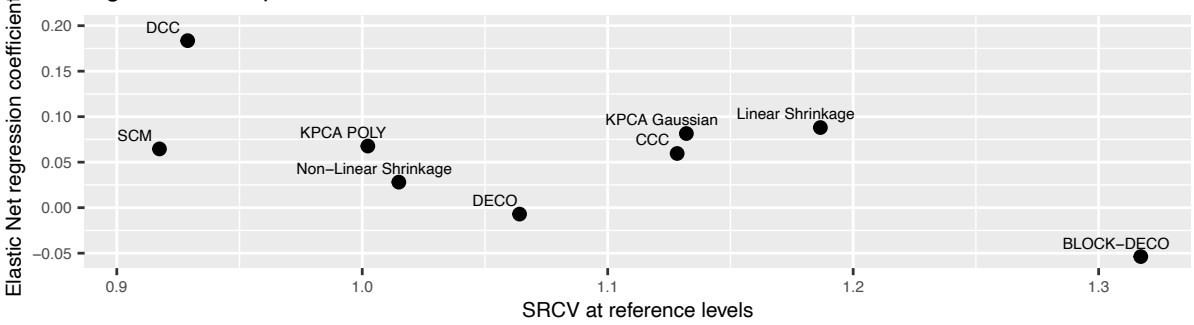
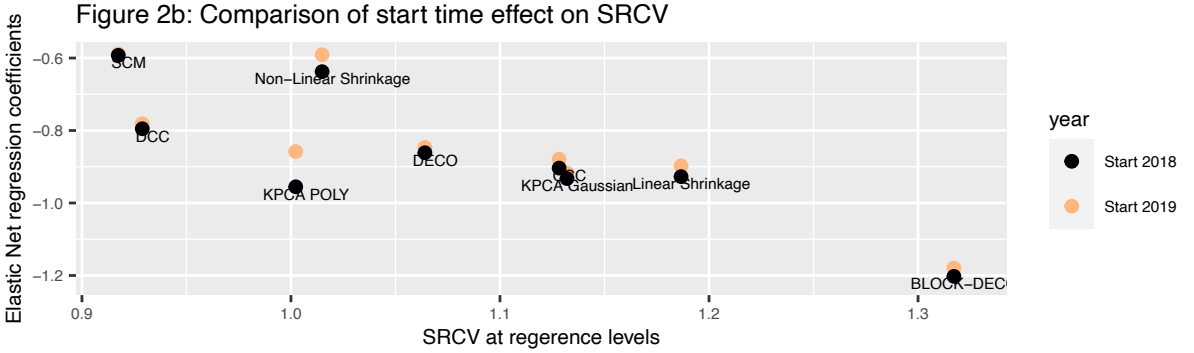
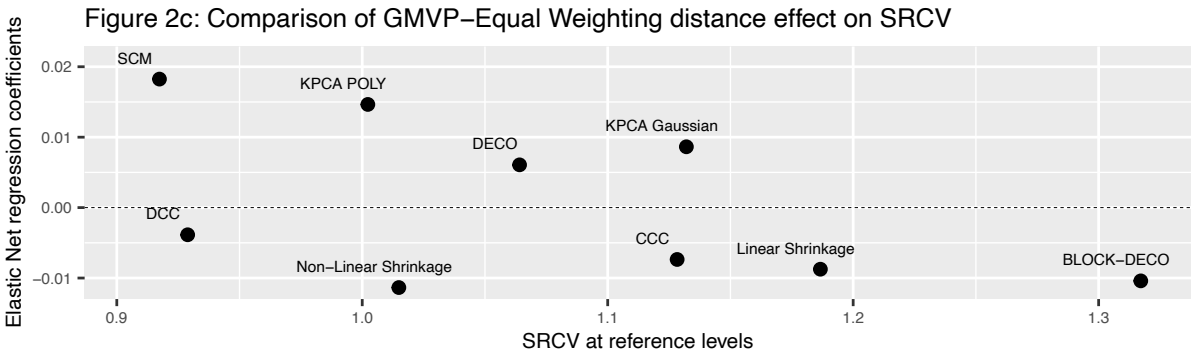
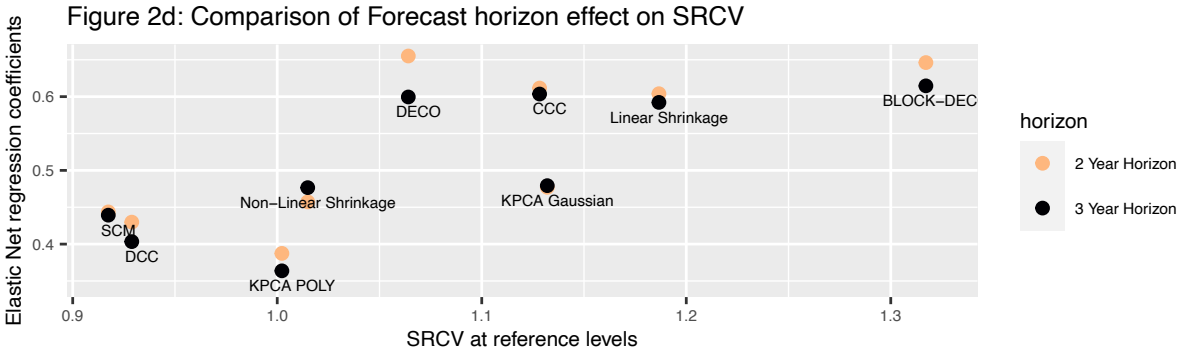
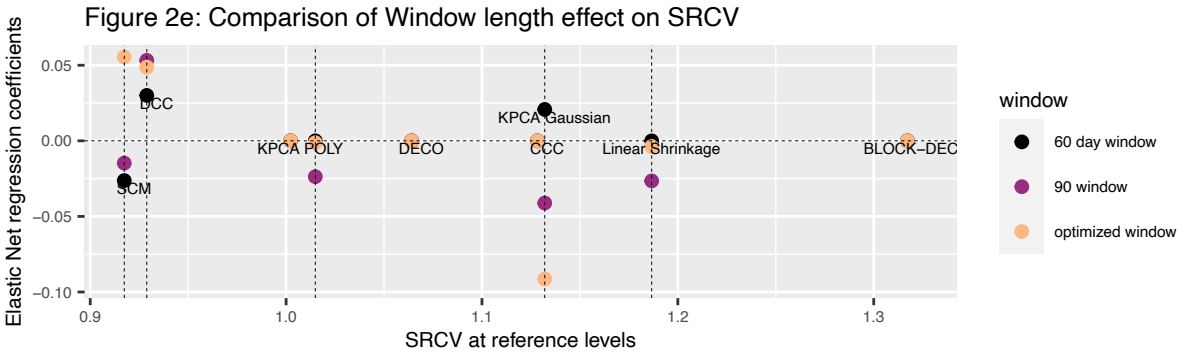


Figure 2b: Comparison of start time effect on SRCV









Using Ordinary Least Squares (OLS), it is found that start time is considered statistically significant at all conventional significance levels. Both the horizon factor levels for two and three years are significant at $\alpha = 0.05$. Moreover, RMF beta is significant at a 10% level. The window lengths as well as portfolio types do not yield significant results. All OLS models have been checked for homoscedasticity, correlation in the error terms and normally distributed errors for the purpose of being the Best Linear Unbiased Estimator (BLUE) as well as being suitable for significance testing. Given the non-normally distributed errors of optimized window length regression, Generalized Least Squares (GLS) has been used to correct for this (Johnston, 1972)

From this it is deduced that RMF, start time and horizon are not independent of SRCV, whereas we cannot reject the hypotheses that portfolio type and window length Elastic Net coefficients are independent of SRCV based performance, even when measured across all methods. The significance of RMF and its relation with SRCV should be treated with caution as it is heavily reliant on the observations of DCC and Block-DECO.

Additionally, Polynomial KPCA is the only method that exhibits a notable difference between the coefficients for a start time in 2018 and 2019. DECO and Block-DECO show considerable difference in coefficients for two and three year forecast horizons, indicating these methods' relatively high sensitivity to start time and forecast horizon respectively.

The most important results that can be concluded from table 2 is that start time, forecast horizon and RMF are strongly related to SRCV, whereas portfolio type and forecast horizon are not. It should also be noted that the two best performing methods, Block-DECO and Non-Linear Shrinkage, tend to find themselves on approximately opposite sides of the plots in figure 2.

When regarding the two best performing methods specifically, we find that Block-DECOs performance can be attributed mainly to its very high SRCV at reference level. While it still yields high positive coefficients for longer forecast horizons, this is counteracted by the very large negative coefficients for start time. The drop in performance when the start time is pushed back could be explained by overfitting, but this would not explain the large positive effect of a longer forecast horizon.

For further analysis of the well performing methods' performance the SRCV will be dissected into its two key components, the mean and standard deviation of the Sharpe Ratio across all 20 simulated data sets. This yields the following visual representations.

4.2 Block-DECO Results

Figure 3a: Start time effect on Block–DECO Sharpe Ratio

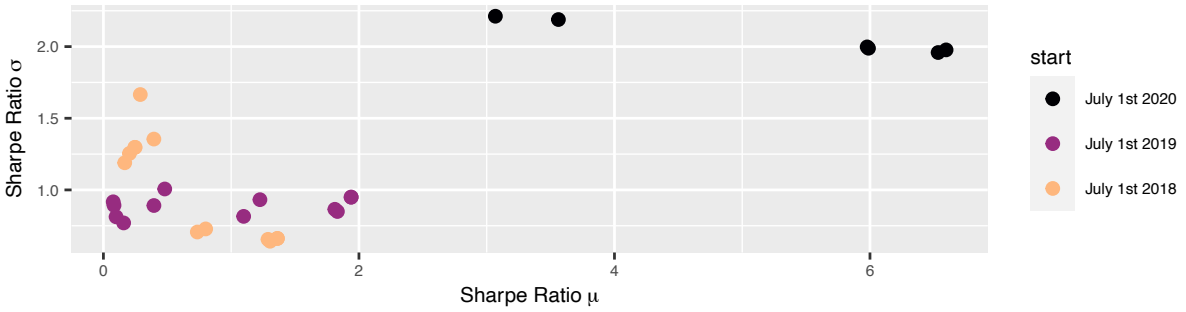


Figure 3b: Forecast horizon effect on Block–DECO Sharpe Ratio

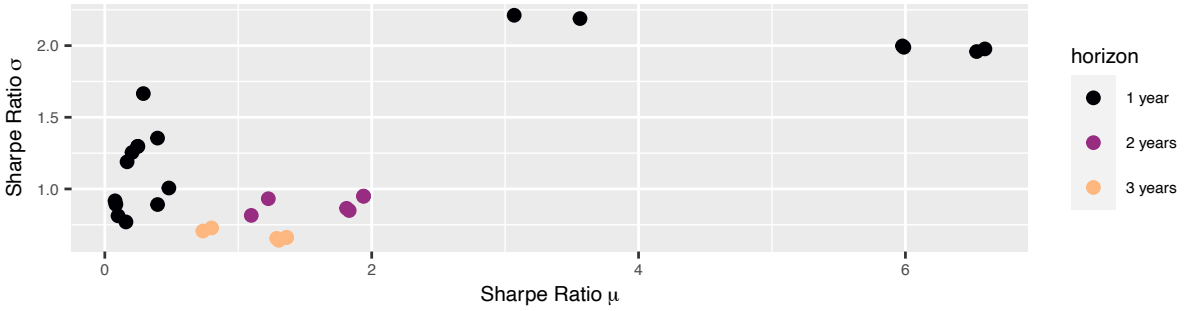


Figure 3c: Portfolio effect on Block–DECO Sharpe Ratio

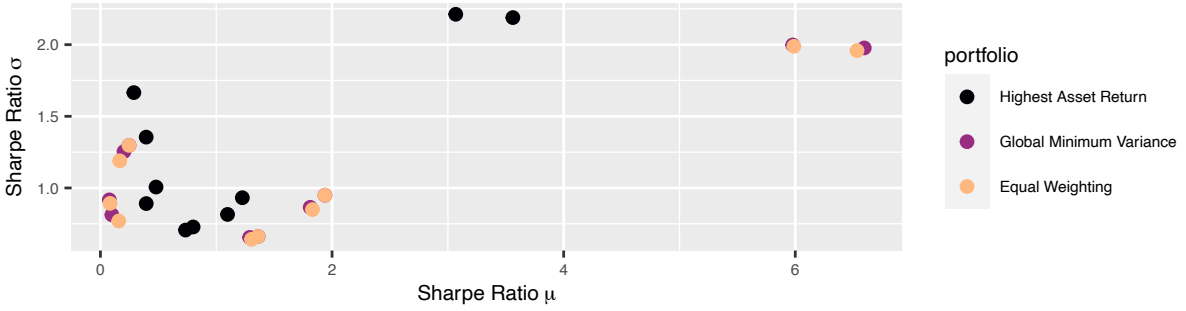
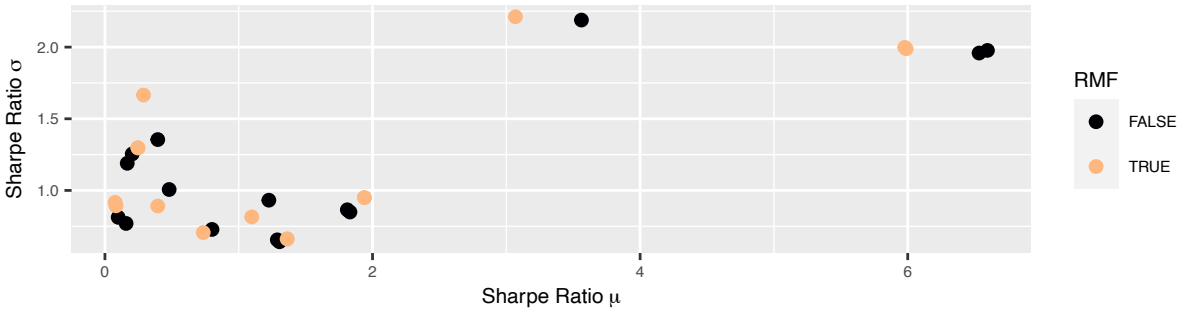


Figure 3d: RMF effect on Block–DECO Sharpe Ratio



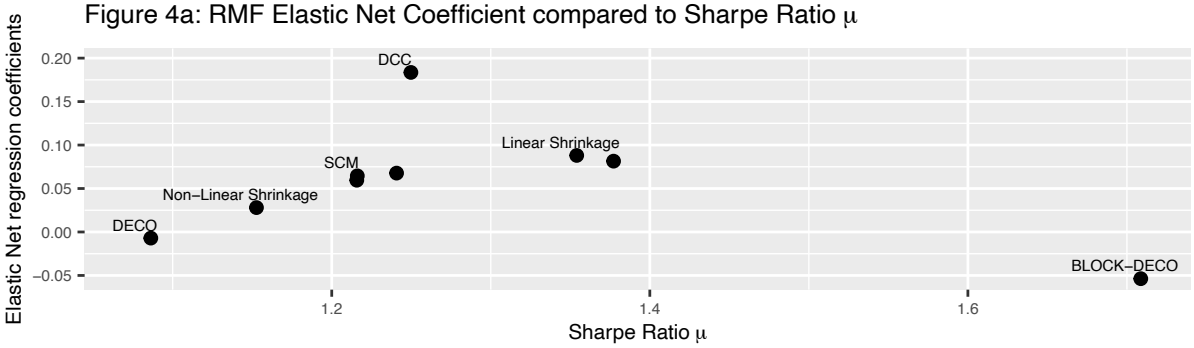
While all methods show a decrease in performance as the start time is shifted back, Block-DECO sees the biggest change as portrayed in both figure 2b and table 1(2). From being the best performing method at a start time in 2020, it becomes the 2nd worst performing method by SRCV when the start period is pushed back 1 year. Compared to the other levels, the 2018 start period portfolios also show unusually high variation in Sharpe Ratio σ.

Looking at table 1(cont'd 2), and figure 2a we see that Block-DECO is the only method for which RMF does not yield a significant increase in SRCV. While the difference seems negligible for portfolios starting in 2018 and 2019, the 2020 observations display a decrease in mean Sharpe Ratio with RMF, without reducing the volatility as observed in figure 3a. The Highest Asset Return portfolio underperforms drastically yielding lower Sharpe Ratio means at a volatility compared to the other portfolios. This difference is particularly notable in the observations starting in 2020 as is concluded from figure 3a and 3c.

The most important observations here are that the performance of Block-DECO in the 1 year horizon portfolios starting in 2020 seems driven by exceptionally high Sharpe Ratio means to justify the high Sharpe Ratio σ as shown in figure 3b. Increasing the forecast horizon has a large positive effect on performance by means of suppressing the volatility. This is caused not by economic circumstances, but by the larger sample size and is consistent with other methods. Moving the start time of the portfolio to years prior trades an exceptional Sharpe Ratio mean for a mediocre one, resulting in poor performance

Additional caution must be exercised due to the fact that the recent period from early 2020 onwards saw the largest returns and volatility in commodity prices of the entire utilized data set, caused by the economic effects of the Coronavirus pandemic. The large difference in economic circumstances between time period means that any method prone to overfitting, would have this shortcoming amplified due to time-variant properties of the data.

Figure 4a: RMF Elastic Net Coefficient compared to Sharpe Ratio μ



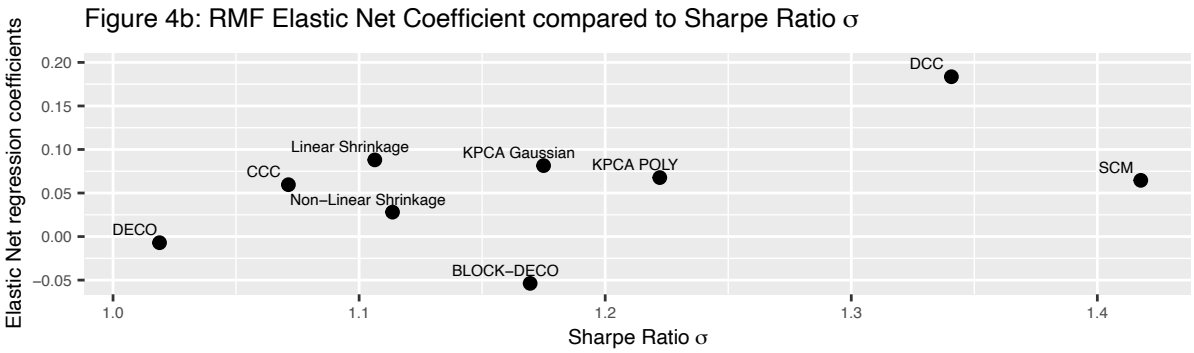
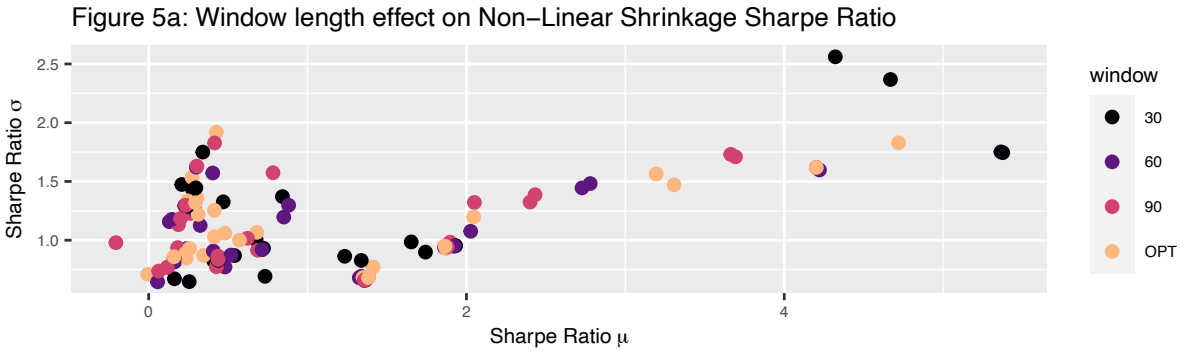


Figure 4a and 4b show the unusual effect of RMF on Block-DECO when SRCV is decomposed into its two primary components. Given that the decrease in Sharpe Ratio μ caused by RMF is most noticeable in the highly volatile period in 2020 and RMFs process of reducing noise in the data, it seems that the noise reduction is redundant for Block-DECO.

4.3 Non-Linear Shrinkage Results



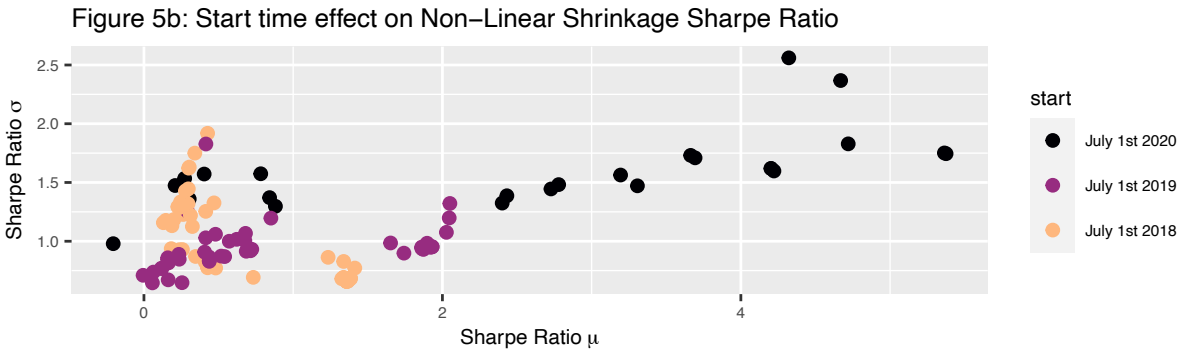


Figure 5c: Forecast horizon effect on Non−Linear Shrinkage Sharpe Ratio

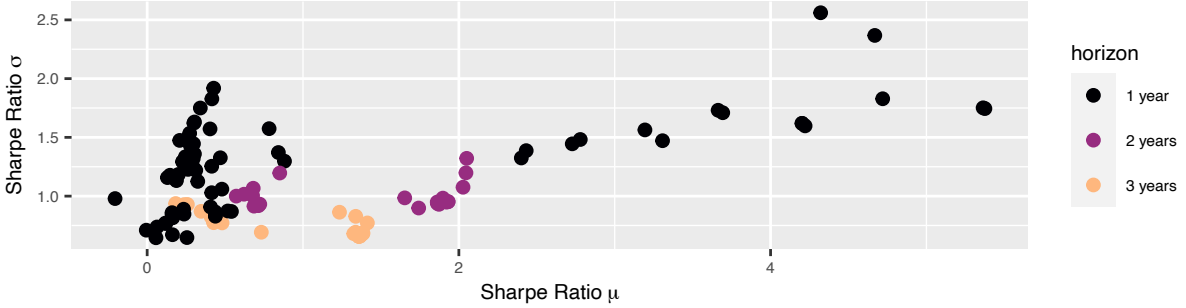


Figure 5d: Portfolio effect on Non−Linear Shrinkage Sharpe Ratio

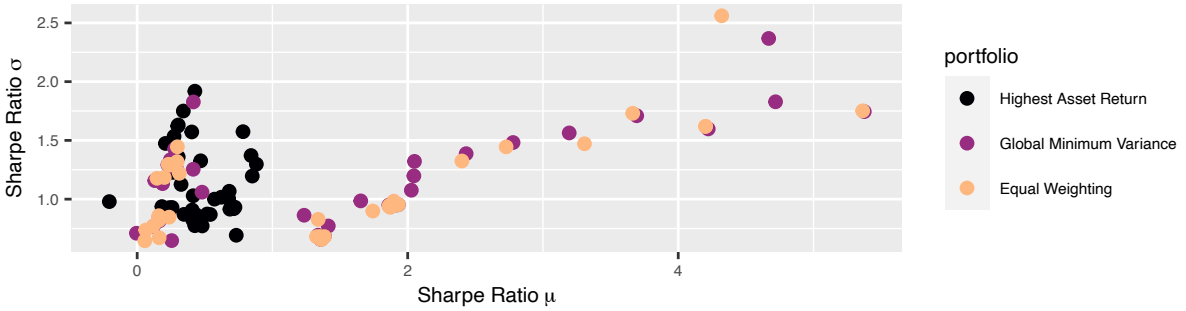
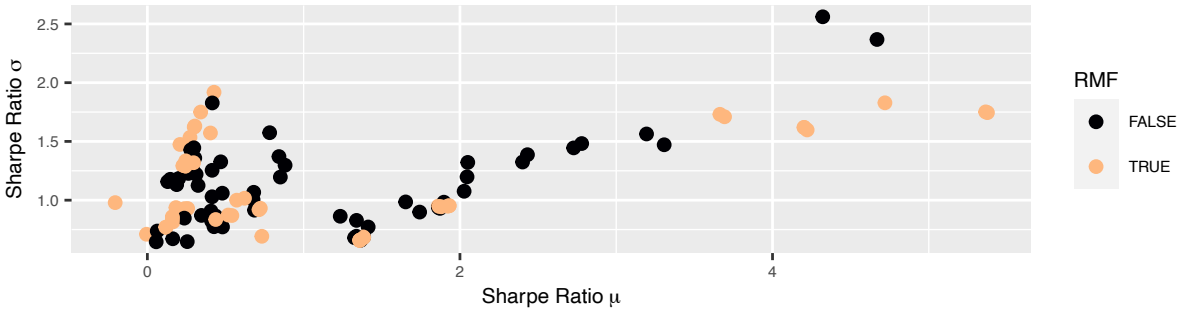


Figure 5e: RMF effect on Non−Linear Shrinkage Sharpe Ratio



Firstly, Non-Linear Shrinkage has considerably more observations by virtue of having a defined window length that can be adjusted. While all window length factors except 60 days are significant, the coefficients are small and negligible for the overall performance.

For Non-Linear Shrinkage, an earlier start time decreases both the mean Sharpe Ratio as well as its volatility, which is consistent with every other method. Compared to Block-DECO however, this difference is not nearly as defined. Whereas Block-DECO saw 2020 start times be completely separated from the other factor levels, Non-Linear Shrinkage shows considerable overlap between the three different start times. Still, the 2020 start time yields the highest volatility on average and the highest variation in the mean across different specifications, causing relatively poor performance. Like Block-DECO, the 2019 start data contains less variability in Sharpe Ratio σ than the 2018 portfolios.

The effect of the forecast horizon on SRCV and its components seems to mirror that of Block-DECO, with a longer horizon causing a reduction in mean Sharpe Ratio, but a proportionately greater reduction in volatility causing and increase in SRCV.

The influence exerted by portfolio type is comparable too, with GMVP and Equally Weighed portfolios heavily outperforming Highest Asset Return portfolios. Again this discrepancy grows to several magnitudes for a recent start time of July 1st 2020.

The effect of RMF is notably different compared to Block-DECO. While the effect is again most prominent in the 2020 start time observations, applying RMF seems to reduce volatility by a large margin. At low mean Sharpe Ratios, this difference seems nonexistent, but as the Mean Sharpe Ratio increases, the effect of RMF on volatility grows gradually. This is highly consistent with the observations across methods from figure 4a.

Most important here is that regardless of the sudden economic extremities in 2020 and onward, Non-Linear Shrinkage exhibits one of the smallest absolute start time coefficients and its change is less radical than Block-DECO, indicating its relative robustness. Any change in temporal parameters had significant negative effects on performance, but least so for Non-Linear Shrinkage. Another way of defining what sets this method apart is by comparing it with its closest relative, Linear Shrinkage. When looking at previous figures comparing methods, Linear Shrinkage does not show any irregularities in performance. It is in many cases however, the method whose results match closest with Block-DECO.

5. Conclusion

Section 5.1 will conclude the research and its results and findings. In section 5.2 the shortcomings and limitations of this research are brought forth, and section 5.3 discusses the potential of further research based on this.

5.1 Conclusion of results

Circling back to the research hypotheses, the question posed at the beginning of the research was how the performance of different covariance estimation techniques differ, which was then split into the effects of statistical and economic factors. These effects are quantified by building portfolios for each estimation method. The portfolios are constructed with different parameter values for statistical and economic factors and are then evaluated using the Sharpe Ratio. This is performed for all 20 simulated series to arrive at the Coefficient of Variation of the Sharpe Ratio. This value and its components then function as dependent variable in Elastic Net regressions to observe the influence that each factor and model hyperparameter has on the overall performance.

From this it is concluded that some statistical properties exert profound influence on the performance of estimation, whereas others do not. Meanwhile, economic properties seem to be the main driving force in performance discrepancies.

Window length used in the estimation procedure is negligible. For the methods that utilize a window length parameter, the effect of this is occasionally statistically significant, but also exceedingly small. Its inconsistencies also means it cannot be claimed that one window length systematically outperforms others.

Portfolio type is largely inconsequential as well. Portfolios based on maximum single asset returns were abandoned due to exceedingly poor performance. While equal weighting portfolios seem to outperform GMVP for methods with high SRCV, this effect is small and not statistically significant.

Random Matrix Filtering has a positive effects on covariance estimation with the exception of Block-DECO. Its positive effects of both increasing mean and reducing variance become weaker as the estimation method already performs well, and are strongest for underperforming techniques. As such it functions to bring the results of different methods closer to each other, but has no effect on what method is optimal in a given situation.

Adjusting the forecast horizon yields similar results for each method, with both the Sharpe Ratio μ and σ decreasing, accompanied by a large reduction in variance.

The start time of the data portfolio, however, walks the line between statistical and economic properties as it is intertwined with the most striking economic property of the data, which is a large increase in commodity returns variance starting in 2020. The degree to which pushing back the start data negatively affects

performance differs wildly across methods. Accounting for all possible scenarios, any estimation starting July 1st 2020 with a one year forecast horizon will perform best when using Block-DECO for covariance estimation, whereas for every other scenario Non-Linear Shrinkage yields the best results.

The extreme change in economic circumstances polarizes performance and the two aforementioned methods stand diametrically opposed to one another. Non-Linear Shrinkage is very flexible in its structure, resulting in wide applicability in many scenarios, even those that differ significantly. This is most apparent when compared with its linear counterpart which is not remotely as robust to extrapolation to different economic environments.

On the other side, Block-DECO reduces variance by introducing heavy bias. It has a very strong superimposed structure and uniquely does not benefit from the RMF noise filtering. Its rigid structure and the fact that its blocks are defined by economic behavior, grant this method remarkable resilience to extreme economic environments. It does however seem that this comes with the trade-off that the model is prone to overfitting. While performance in these volatile economic periods is very good, extrapolation to different economic scenarios is unsuccessful.

In conclusion, statistical factors heavily impact the performance of covariance estimation overall, but do not affect which method is best in a given situation. This almost exclusively determined by economic factors, the techniques robustness to change and its ability to deal with extreme and noisy data. Overall Non-Linear Shrinkage performs best for its fluidity and general application, but Block-DECO outperforms every other method should the economic circumstances be extreme or unusual. Both methods fall on opposite sides of the spectrum in terms of the bias-variance trade-off while compromising methods in between are suboptimal in any given situation.

5.2 Shortcomings

Despite my attempts to ensure accuracy, statistical robustness and economic applicability, several compromises have been made.

Firstly, the research has been conducted with a relatively small sample size, both in number of commodities as in number of simulated data sets. While results are subject to hypothesis testing that accounts for sample sizes, it could be that results have been deemed statistically insignificant by virtue of the sample size being too small to draw definitive conclusions. In addition, some methods require more hyperparameters to be tuned than others, most notably window length. This creates a discrepancy between the number of observations available for each method, affecting comparison.

Several means have been used to homogenize calculations across methods. Returns and interest rates have all been calculated on the premise that a year contains 252 trading days. While this might not always be completely accurate, these assumptions were used in the calculations of the Sharpe Ratios, which were used strictly for comparison. All methods should be affected equally by any error in specification and thus is should only affect the outcome in absolute terms, and not relative performance this research is focused on.

Another point of contention is the use of penalized regression. As previously explained this was used for the sake of variable selection due to the large number of factor levels. To facilitate the accuracy of penalized regression, bias is introduced, potentially hindering the interpretability of results.

Additionally, the method of evaluating results in this research was through the performance of portfolios. While this is ultimately the end goal and most important result, the performance of covariance estimation procedures could also be measured through metrics such as matrix distance. This has been omitted for two reasons. Firstly, it would measure how closely a portfolio corresponds with an existing portfolio, instead of optimizing performance and secondly, the matrix that is used to calculate distance would need to be an accurate representation of true covariance. Given the context of this research, using the realized SCM as 'true' covariance seems questionable at best.

Lastly, the decision to have the data sets and predictions start on July 1st instead of January 1st might seem odd. This was done after considering the end of the data set. Starting with January 1st would leave

the most recent 6 months of data available unused. Using these six months for annualized results would require drastic assumptions that could not be verified. Here it is important to note that the data used in this research was an unusual one. The Coronavirus pandemic heavily influenced economic developments and caused considerable fluctuation in the most recent years of the data set. While this serves to highlight any problems estimation techniques might have with volatile data or large changes, it means that any data that would be used as reference for 'normal' economic circumstances is relatively outdated.

5.3 Further Research

One avenue of further research that would improve the applicability is introducing bias into the simulated test data. All 20 simulated data sets used for this research were, albeit with randomness, modeled after the original realized data. Adding more categories of simulations such as simulations with manually increased volatility could yield more insight into the effects of volatility on method performance given that this would allow it to be completely independent of temporal variables such as start time or forecast horizon, which currently make different forecast horizon results incompatible for comparison.

Another potential improvement would lie in the inclusion of more options for start periods and forecast horizons. Given the data set has a definitive end, the forecast horizon is limited by the start time. As such, long forecast horizon portfolios have less observations because they have to start at an earlier date. While any additional options would face the problem that the data becomes increasingly more outdated, it would aid in statistical veracity.

Additional performance measures would also be a valuable addition to this research. While Sharpe Ratio is the most common measure for portfolio performance, and the use of its coefficient of variation accounts for its fluctuation across the simulated data sets, there are numerous other performance measures such as the Treynor Ratio or Sortino Ratio whose different computation could uncover new insights in covariance estimation performance.

The method used to generate the simulated data sets for this research was the Johnson's S_U Distribution. While considered suitable for this situation, it would be insightful to conduct the same procedures with data sets generated through bootstrapping, or other simulation methods to study how much the simulation method affects results.

Newer, more novel and advanced methods for covariance estimation are constantly being developed and only a portion of them have been covered in this research. The addition of any techniques would be useful to create a more extensive and complete overview and comparison of covariance estimation.

6. References

Aggarwal, C. C., Hinneburg, A., & Keim, D. A. (2001). "On the surprising behavior of distance metrics in high dimensional space". *Database Theory—ICDT 2001: 8th International Conference London*, 420-434.

Bauwens, L., Laurent, S. & Rombouts, J. V. K. (2006). "Multivariate GARCH models: A survey". *Journal of Applied Econometrics*, 21, 79-109.

Beyer, K., Goldstein, J., Ramakrishnan, R., & Shaft, U. (1999). "When is "nearest neighbor" meaningful?". *Database Theory—ICDT'99: 7th International Conference Jerusalem*, 217-235.

Bien, J., & Tibshirani, R. J. (2011). "Sparse estimation of a covariance matrix". *Biometrika*, 98 (4), 807-820.

Bollerslev, T. (1990). "Modelling the coherence in short-run nominal exchange rates: a multivariate generalized ARCH model". *The review of economics and statistics*, 498-505.

Bollerslev, T., Engle, R. F., & Wooldridge, J. M. (1988). "A capital asset pricing model with time-varying covariances". *Journal of political Economy*, 96 (1), 116-131.

Byun, J., Tan, G., Normand, J. & Spector, K. (2007). "Introducing the JPMorgan Commodity Curve Index (JPMCCI)". *J.P. Morgan Securities Global Commodities Research*, https://www.jpmorgan.com/content/dam/jpm/global/disclosures/us/CCI_Introduction.pdf

Chen, C. (1976). "On information and distance measures, error bounds, and feature selection". *Information Sciences*, 10 (2) 159-173.

Engle, R. F. (1982). "Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation". *Econometrica*, 987-1007.

Engle, R. F. (2002). "Dynamic conditional correlation: A simple class of multivariate generalized autoregressive conditional heteroskedasticity models". *Journal of Business & Economic Statistics*, 20 (3), 339-350.

Engle, R. & Kelly, B. (2012). "Dynamic equicorrelation". *Journal of Business & Economic Statistics*, 30 (2), 212-228.

Engle, R. F., Ledoit, O., & Wolf, M. (2019). "Large dynamic covariance matrices". *Journal of Business & Economic Statistics*, 37 (2), 363-375.

Engle, R. F. & Sheppard, K. (2001). "Theoretical and empirical properties of dynamic conditional correlation multivariate GARCH". *National Bureau of Economic Research*, working paper 8554, DOI 10.3386/w8554

Jagannathan, R., & Ma, T. (2003). "Risk reduction in large portfolios: Why imposing the wrong constraints helps". *The Journal of Finance*, 58(4), 1651-1683.

James, N. A., & Matteson, D. S. (2013). "ecp: An R package for nonparametric multiple change point analysis of multivariate data". *Journal of Statistical Software*, 62 (7), 1-25. https://doi.org/10.18637/jss.v062.i07

Jobson, J. D., & Korkie, B. M. (1980). "Estimation for Markowitz efficient portfolios". *Journal of the American Statistical Association*, 75 (371), 544-554.

Jobson, J. D., & Korkie, B. M. (1981). "Performance hypothesis testing with the Sharpe and Treynor measures". *Journal of Finance*, 36 (4), 889-908.

Johnson, N. L. (1949). "Systems of frequency curves generated by methods of translation". *Biometrika*, 36 (1/2), 149-176.

Johnston, J. (1972). "Generalized Least Squares". *Econometric Methods*. New York: McGraw-Hill, 208-242.

Ledoit, O., & Péché, S. (2011). "Eigenvectors of some large sample covariance matrix ensembles". *Probability Theory and Related Fields*, 151 (1-2), 233-264.

Ledoit, O., & Wolf, M. (2003). "Improved estimation of the covariance matrix of stock returns with an application to portfolio selection". *Journal of empirical finance*, 10(5), 603-621.

Ledoit, O., & Wolf, M. (2004). "A well-conditioned estimator for large-dimensional covariance matrices". *Journal of multivariate analysis*, 88 (2), 365-411.

Ledoit, O., & Wolf, M. (2004). "Honey, I shrunk the sample covariance matrix". *The Journal of Portfolio Management*, 30 (4), 110-119.

Ledoit, O., & Wolf, M. (2012). "Nonlinear shrinkage estimation of large-dimensional covariance matrices". *Annals of Statistics*, 40 (2), 1024-1060.

Ledoit, O., & Wolf, M. (2020). "Analytical nonlinear shrinkage of large-dimensional covariance matrices". *Annals of Statistics*, 48 (5), 3043-3065.

Ledoit, O., & Wolf, M. (2022). "The power of (non-) linear shrinking: A review and guide to covariance matrix estimation". *Journal of Financial Econometrics*, 20 (1), 187-218.

Levene, H. (1960). "Robust tests for equality of variances". *Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling*. Stanford University Press. pp. 278–292.

Livan, G., Novaes, M., & Vivo, P. (2018). Introduction to random matrices. *Cornell University* https://doi.org/10.48550/arXiv.1712.07903

Markowitz, H. (1952). "The utility of wealth". *Journal of political Economy*, 60 (2), 151-158.

Marchenko, V. A., & Pastur, L. A. (1967). "Distribution of eigenvalues for some sets of random matrices". *Matematicheskii Sbornik*, 114 (4), 507-536.

Matteson, D. S., & James, N. A. (2014). "A nonparametric approach for multiple change point analysis of multivariate data". *Journal of the American Statistical Association*, 109 (505), 334-345.

May, W. L. & Johnson, W. D. (1998). "On the singularity of the covariance matrix for estimates of multinomial proportions". *National Library of Medicine*. https://pubmed.ncbi.nlm.nih.gov/9598426/

Mehta, M. L., & Dyson, F. J. (1963). "Statistical theory of the energy levels of complex systems". *Journal of Mathematical Physics*, 4 (5), 713-719.

Michaud, R. O. (1989). "The Markowitz optimization enigma: Is 'optimized'optimal?". *Financial analysts journal*, 45 (1), 31-42.

von Neumann, J (1929), "Allgemeine Eigenwerttheorie Hermitescher Funktionaloperatoren", *Mathematische Annalen*, 102 (1): 49–131, doi:10.1007/BF01782338, S2CID 121249803.

Ng, I. S. (2018), "Reproducing Kernel Hilbert Spaces & Machine Learning". https://ngilshie.github.io/jekyll/update/2018/02/01/RKHS.html

Pearson, K. (1901). "On lines and planes of closest fit to systems of points in space". *The London, Edinburgh, and Dublin philosophical magazine and journal of science*, 2 (11), 559-572.

Peng, Y., Albuquerque, P. H. M., do Nascimento, I. F., & Machado, J. V. F. (2019). "Between nonlinearities, complexity, and noises: An application on portfolio selection using kernel principal component analysis". *Entropy*, 21 (4), 376.

Riesz, F. (1907). "Sur une espèce de géométrie analytique des systèmes de fonctions sommables". *Comptes rendus de l'Académie des Sciences*, 144, 1409–1411.

Riesz, F. (1909). "Sur les opérations fonctionnelles linéaires". *Comptes rendus de l'Académie des Sciences*. 149, 974–977.

Rosenthal, J. S. (2015). "The Kids are Alright: Divide by n when estimating variance". *Institute of Mathematical Statistics* https://imstat.org/2015/11/17/the-kids-are-alright-divide-by-n-when-estimating-variance/

Rumsey, D. J. (2009). "Let's just eliminate the variance". *Journal of Statistics Education*, 17 (3).

Sin-Yi Tsai, C. (2011). "The real world is not normal". *Morningstar*. http://morningstardirect.morningstar.com/clientcomm/iss/tsai_real_world_not_normal.pdf

Schölkopf, B., Smola, A., & Müller, K. R. (1998). "Nonlinear component analysis as a kernel eigenvalue problem". *Neural computation*, 10 (5), 1299-1319.

Sharpe, W. F. (1963). "A simplified model for portfolio analysis". *Management science*, 9 (2), 277-293.

Stein, C.M. (1956). "Inadmissibility of the Usual Estimator for the Mean of a Multivariate Normal Distribution". *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, 1954-1955, I, 197-206. University California Press, Berkeley.

Touloumis, A. (2015). "Nonparametric Stein-type shrinkage covariance matrix estimators in high-dimensional settings". *Computational Statistics & Data Analysis*, 83, 251-261

Wheeler, R. E. (1980). "Quantile estimators of Johnson curve parameters". *Biometrika*, 725-728.

Zivot, E. (2016). "Introduction to Computational Finance and Financial Econometrics with R". https://bookdown.org/compfinezbook/introcompfinr/ https://faculty.washington.edu/ezivot/econ424/portfolioTheoryMatrix.pdf