A Kernel Density Based Approach to Portfolio Optimization

Lawrence Liu

May 2020

1 Introduction

Portfolio optimization can be characterized as a constrained optimization problem, where we want to maximize the expected return of a portfolio while minimizing the risk. The risk is usually measured by the standard deviation of the portfolio. If we allow margin trading, meaning that we can borrow money to invest, and thus have a position larger than our initial capital, then the optimal portfolio becomes a scaling of the portfolio that maximizes the Sharpe ratio, this is what we call the efficient frontier. We define the sharpe ratio as the ratio of the expected return of the portfolio to the standard deviation of the portfolio.

$$\text{Sharpe Ratio} = \frac{\mu_p}{\sigma_p} \tag{1}$$

Where μ_p is the expected return of the portfolio, and σ_p is the standard deviation of the portfolio. If we take the μ_p as the average day to day return of the portfolio, and σ_p as the standard deviation of the day to day returns, then we can also define the annualized Sharpe ratio as

$$\text{Sharpe Ratio} = \frac{\mu_p}{\sigma_p}\sqrt{252} \tag{2}$$

Where 252 is the number of trading days in a year. From now on we use the annualized Sharpe ratio and all further references to Sharpe ratio will be the annualized Sharpe ratio.

1.1 Modern Portfolio Theory

The most common approach to create the portfolio that maximizes the Sharpe ratio is Modern Portfolio Theory (MPT). MPT was first introduced by Harry Markowitz in 1952 [?]. Let us define the sample mean of the returns of the assets as μ, and the sample covariance matrix of the returns of the assets as Σ. Then μ_p and σ_p can be written as

$$\mu_p = w^T \mu \tag{3}$$

$$\sigma_p = \sqrt{w^T \Sigma w} \tag{4}$$

Where w is the weight vector of the portfolio. The optimization problem can then be written as

$$\begin{aligned} \underset{w}{\text{maximize}} \quad & \frac{w^T \mu}{\sqrt{w^T \Sigma w}} \\ \text{subject to} \quad & \sum_{i=1}^{n} w_i = 1 \\ & w_i \geq 0 \quad \forall i \in \{1, \dots, n\} \end{aligned} \tag{5}$$

Where n is the number of assets in the portfolio. This is a fractional programming problem which we can solve in the following manner.[?] If we let $y = \alpha w$, where $\alpha = 1^T y$ is a scalar, then the problem becomes

$$\begin{aligned} \underset{y}{\text{minimize}} \quad & y^T \Sigma y \\ \text{subject to} \quad & \mu^T y = 1 \\ & y_i \geq 0 \quad \forall i \in \{1, \dots, n\} \end{aligned} \tag{6}$$

This is a convex optimization problem, which we can solve with convex optimization methods. We then get that the optimal weights are given by $w^* = \frac{y^*}{1^T y^*}$, where y^* is the optimal solution to the convex optimization problem.

At the central core of MPT are two assumptions.

- The returns of the assets are normally distributed.
- The returns of the assets are stationary.

In this paper we will first show that the returns of the assets are not normally distributed. Then we will propose a new approach to portfolio optimization that does not rely on this assumption. We will show that this method provides a superior sharpe ratio compared to MPT.

2 Data

We used the data from the S&P 500 index in a 10 year period from 2010-01-01 to 2020-01-01. The data was downloaded from Yahoo Finance. Because certain companies were not traded during the entire period, we only used the companies that were traded during the entire period. As a result we had 428 companies in our dataset.

We used the adjusted close price of the stocks to calculate the returns. The data was split into a training set and a test set. the training set was the first 10 years of the data, and the test set was the last year of the data. We fit our model on the training set, and evaluated it on the test set.

Non Gaussanity of the Returns

To evaluate the assumption that the returns of the assets are normally distributed, let us introduce two measures, the skewness and the kurtosis of a distribution. We define the skewness as

$$\gamma = \frac{\mu_3}{\sigma^3} \tag{7}$$

Where $\mu_3 = \mathbb{E}[(X - \mu)^3]$ is the third central moment of the distribution, and σ is the standard deviation of the distribution. We define the kurtosis as

$$\kappa = \frac{\mu_4}{\sigma^4} \tag{8}$$

Where $\mu_4 = \mathbb{E}[(X - \mu)^4]$ is the fourth central moment of the distribution, and σ is the standard deviation of the distribution. An intuitive way to think about the skewness is that it measures the asymmetry of the distribution. If the distribution is symmetric, then the skewness is zero. Likewise an intuitive way to think about the kurtosis is that it measures the thickness of the tails of the distribution. If the distribution has the same tails as a normal distribution, then the kurtosis is 3. Therefore we can define the excess kurtosis as $\kappa - 3$. This is an important measure that we need to account for when we are modeling the returns of the assets, since the tails of the distribution can cause large losses, that if we use a normal distribution to model the returns, we will serverly underestimate the probability of.

For a normal distribution we have that $\gamma = 0$ and $\kappa = 3$. Thus we have plotted in figure **??** the skewness and the excess kurtosis of the returns of the assets. of the assets with the highest and lowest

$|\gamma|$ and $|\kappa - 3|$.

TODO: Add figure

3 Method

Our method consists of two parts. The first part is to model the returns of a set of assets with a nonparametric Kernel Density Estimation with a multivariate Gaussian Kernel. The second part is to isolate the assets into smaller subsets that are highly correlated with itself.

3.1 Kernel Density Estimation

Effectively with Kernel Density Estimation what we try to do is with a kernel function, we smooth a normalized histogram of the data. The kernel function is a function that is centered around a point, and is symmetric around that point. Mathematically, we can write the kernel density estimation for the the *pdf* of a random variable X as

$$\hat{f}_\theta(x) = \frac{1}{n}\sum_{i=1}^{n} K_\theta(x - x_i') \tag{9}$$

where K_h is the kernel function, and θ are the parameters of the Kernel, and x_i' is the ith training point. In our case x_i' is a m dimensional vector representing the daily change for each of the m stocks in our dataset. We assume that that each kernel function is normalized $\int_{\mathbb{R}^d} K_h(x)dx = 1$

3.1.1 Optimizing the Kernel Parameters

We would want to minimize the weighted log likelihood function, ie we would want to minimize:

$$\mathcal{L}(x_1, \ldots, x_k) = -\sum_{i=1}^{k} \log(\hat{f}_\theta(x_i)) \tag{10}$$

Where $x_1, \ldots, x_k$ are the test points.

Now let us restrict our considerations to the case of a gaussian multivariate kernel. We have that

$$K_\Sigma(x) = \frac{1}{\sqrt{(2\pi)^d|\Sigma|}} \exp\left(-\frac{1}{2}x^T\Sigma^{-1}x\right)$$

Where Σ is the covariance matrix for the kernel. Because the covariance matrix is symmetric positive semidefinite we can express Σ as $\Sigma = R^T R$ through Cholesky decomposition. We have that the derivative of the weighted log likelihood function is given by:

$$\frac{\partial}{\partial R}\mathcal{L}(x_1, \ldots, x_k) = -\sum_{i=1}^{k} \frac{1}{\hat{f}_\theta(x_i)} \frac{\partial \hat{f}_\theta(x_i)}{\partial R}$$

We have that

$$\begin{aligned}\frac{\partial|\Sigma|}{\partial R} &= \frac{\partial|R^T R|}{\partial R} \\ &= 2|\Sigma|R^{-T}\end{aligned}$$

Therefore we have that

$$\frac{\partial}{\partial R}\frac{1}{\sqrt{(2\pi)^d|\Sigma|}} = -R^{-T}\frac{1}{\sqrt{(2\pi)^d|\Sigma|}} \tag{11}$$

We also have that:

$$\frac{\partial}{\partial R}e^{\frac{1}{2}x^T\Sigma^{-1}x} = \frac{1}{2}e^{\frac{1}{2}x^T\Sigma^{-1}x}\frac{\partial}{\partial R}x^T(R^{-1}R^{-T})x \tag{12}$$

$\frac{\partial}{\partial R}x^T(R^{-1}R^{-T})x$ is very difficult to calculate, so we must approximate it. First we note that for a function $f(\mathbf{x})$ that takes in a vector $\mathbf{x}$ we have that the first order taylor expansion is given by:

$$f(\mathbf{x}+\Delta\mathbf{x}) \approx f(\mathbf{x}) + \frac{\partial f(\mathbf{x})}{\partial \mathbf{x}}\Delta\mathbf{x} \tag{13}$$

Where $\frac{\partial f(\mathbf{x})}{\partial \mathbf{x}}$ is a $1\times d$ vector, if $\mathbf{x}$ is a d dimensional vector. Therefore we argue that a generalization to a function of a matrix $\mathbf{X}$ is given by:

$$f(\mathbf{X}+\Delta\mathbf{X}) \approx f(\mathbf{X}) + \mathbf{1}^T\left(\frac{\partial f(\mathbf{X})}{\partial \mathbf{X}} \circ \Delta\mathbf{X}\right)\mathbf{1} \tag{14}$$

Where $\circ$ is the Hadamard product, and $\mathbf{1}$ is a $d\times 1$ vector of ones. We note that $\mathbf{1}^T\left(\frac{\partial f(\mathbf{X})}{\partial \mathbf{X}} \circ \Delta\mathbf{X}\right)\mathbf{1}$ equals to $\operatorname{tr}\left(\left(\frac{\partial f(\mathbf{X})}{\partial \mathbf{X}}\right)^T\mathbf{X}\right)$. We have for our specific case:

$$x^T((R+\delta R)^{-1}(R+\delta R)^{-T})x \approx x^T(R^{-1}R^{-T})x + \operatorname{tr}\left(\left(\frac{\partial(x^TR^{-1}R^{-T})x}{\partial R}\right)^T\delta R\right)$$

We note that for small pertubations, $(R+\delta R)^{-1} \approx R^{-1} - R^{-1}\delta RR^{-1}$, and therefore:

$$x^T(R^{-1} - R^{-1}\delta RR^{-1})(R^{-T} - R^{-T}\delta R^TR^{-T})x \approx x^T(R^{-1}R^{-T})x + \operatorname{tr}\left(\left(\frac{\partial(x^TR^{-1}R^{-T})x}{\partial R}\right)^T\delta R\right)$$

Only keeping the zeroth order and first order terms, we have that:

$$\begin{aligned}x^T(R^{-1}R^{-T})x - x^T\left(R^{-1}\delta RR^{-1}R^{-T} + R^{-1}R^{-T}\delta R^TR^{-T}\right)x \approx& x^T(R^{-1}R^{-T})x \\ &+ \operatorname{tr}\left(\left(\frac{\partial(x^TR^{-1}R^{-T})x}{\partial R}\right)^T\delta R\right)\end{aligned}$$

$$-x^T\left(R^{-1}\delta RR^{-1}R^{-T} + R^{-1}R^{-T}\delta R^TR^{-T}\right)x \approx \operatorname{tr}\left(\left(\frac{\partial(x^TR^{-1}R^{-T})x}{\partial R}\right)^T\delta R\right) \tag{15}$$

Because the left side is a scalar, we can apply an trace operator to both sides, and noting that $R^{-1}R^{-T} = \Sigma^{-1}$, we have that:

$$\begin{aligned}-\operatorname{tr}\left(x^T\left(R^{-1}\delta R\Sigma^{-1} + \Sigma^{-1}\delta R^TR^{-T}\right)x\right) &\approx \operatorname{tr}\left(\left(\frac{\partial(x^TR^{-1}R^{-T})x}{\partial R}\right)^T\delta R\right) \\ -\operatorname{tr}\left(x^TR^{-1}\delta R\Sigma^{-1}x\right) - \operatorname{tr}\left(x^T\Sigma^{-1}\delta R^TR^{-T}x\right) &\approx \\ -\operatorname{tr}\left(x^TR^{-1}\delta R\Sigma^{-1}x\right) - \operatorname{tr}\left(x^TR^{-1}\delta R\Sigma^{-T}x\right) &\approx \\ -2\operatorname{tr}\left(x^TR^{-1}\delta R\Sigma^{-1}x\right) &\approx \\ -2\operatorname{tr}\left(\Sigma^{-1}xx^TR^{-1}\delta R\right) &\approx \operatorname{tr}\left(\left(\frac{\partial(x^TR^{-1}R^{-T})x}{\partial R}\right)^T\delta R\right)\end{aligned}$$

Therefore we can see that

$$\frac{\partial}{\partial R}x^T(R^{-1}R^{-T})x \approx -2\Sigma^{-1}xx^TR^{-1} \tag{16}$$

Therefore we have that:

$$\frac{\partial}{\partial R}e^{\frac{1}{2}x^T\Sigma^{-1}x} \approx -e^{\frac{1}{2}x^T\Sigma^{-1}x}\Sigma^{-1}xx^TR^{-1} \tag{17}$$

Therefore we have that

$$\frac{\partial}{\partial R}K_\Sigma(x) \approx -K_\Sigma(x)R^{-T} - K_\Sigma(x)\Sigma^{-1}xx^TR^{-1} \tag{18}$$

And thus we have:

$$\frac{\partial}{\partial R}\hat{f}(x) \approx -\sum_{i=1}^{n} K_\theta(x - x_i') \left(R^{-T} - \Sigma^{-1}(x - x_i')(x - x_i')^T R^{-1}\right) \tag{19}$$

Therefore we have that:

$$\frac{\partial}{\partial R}\mathcal{L}(x_1, \ldots, x_k) \approx \sum_{i=1}^{k} \frac{1}{\hat{f}(x_i)} \sum_{j=1}^{n} K_\theta(x_i - x_j') \left(R^{-T} - \Sigma^{-1}(x_i - x_j')(x_i - x_j')^T R^{-1}\right) \tag{20}$$

From this we can just optimize R using stochastic gradient descent, or other more advanced methods such as ADAM.

3.1.2 Obtaining a More Accurate Estimate of the Covariance Matrix