Construction of Optimal Portfolio Strategies

Zhao Jin & Tao Yu

Mentor: Jingyi Zhu

December 12, 2014

Abstract

One of common practices in financial industry to calculate correlation between different stock returns is to use daily closed prices. In this project, we propose new methods to compute correlation matrix that uses real stock data based on transaction level. We use both Brownian bridge and linear interpolation to interpolate prices of all stocks at any given specific time.

1 Introduction

The question of constructing an optimal stock portfolio has been a very important topic in financial investing, and various propositions have been developed since 1950s [1]. Here we intend to propose and test some strategies with a unique approach. There is simply no universal optimal portfolio that is appropriate for every investor, since everyone has a different investment goal with a unique risk preference. For example, young investors may be willing to take more risk to pursue higher return than senior retirees. Even for investors with same age, income, and education background, they would still have quite different attitudes regarding risk tolerance. With this in mind, we are trying to develop a strategy designed for certain risk characteristics that an investor is willing to take, and expect to achieve a maximum return.

In addition to people's different risk aversion degree, it is common that different investors who bear the same risk would end up with different returns. It has been proven that a diversified portfolio containing various individual stocks is able to reduce unsystematic risk without decreasing expected return. In other words, for an individual stock and a diversified portfolio that have the same expected return, the portfolio always has a lower standard deviation (risk).

To construct such a portfolio, the key step is to find the correlation matrix of stocks in the portfolio. The most common way in the finance industry to find correlation between stocks is to use daily closed prices. However we would like to propose an alternative way to calculate correlation coefficient since we believe that using daily closed prices would lose much information for high frequency trading given that hundreds or thousands of transactions happened in a trading day. Alternatively, we computed correlation coefficients using stock data based on a transaction level. However this method results in asynchronous prices for different stocks because the timeline of prices did not match for any two stock pair. We solved this problem using both Brownian Bridge and linear interpolation to interpolate prices of all stocks at any given specific time. Soon we found that the Brownian bridge and linear interpolation would introduce errors if the degree to which timelines of two stocks prices do not match is relatively high. To measure the degree to which timelines of two stocks prices match, we employed kernel density smoothing as a criterion to select stocks that can be used to compute correlation using our method.

The project report is organized as follows. In Section 2, we model stock price movements by using geometric Brownian motion, from which we can estimate the mean and variance of an individual stock return. Section 3 will present how we use kernel density smoothing and both Brownian bridge and linear interpolation to calculate correlation matrix. In addition, we will conduct a simulation to compare the simulation of Brownian bridge and linear interpolation in Section 3. Section 4 will discuss the results of our method compared to results of using daily closed price data in terms of portfolio performance.

2 Model Selection

1. Let S, be the price of the stock; μ, the drift rate of S; and σ, the standard deviation of the stock's returns. The price of a stock follows a geometric Brownian motion, to be more specific. That is

   $$\frac{dS}{S} = \mu dt + \sigma dW_t$$

   where W_t is a standard Brownian motion following one of the propositiones: $W_t - W_s \sim \mathcal{N}(0, t-s)$.
   Also, for two Brownian motions, we can estimate their correlation from the equation:

   $$Cov(W^1_{t_2} - W^1_{t_1},\ W^2_{t_2} - W^2_{t_1}) = \rho(t_2 - t_1)$$

2. $S_{t\geq 0}$ follows a geometric Brownian motion with drift μ and volatility σ. This means that for all t, s ≥ 0 the random variable $\frac{S_{i+1}}{S_i}$ is independent of all prices up to time t and for $\log S_{i+1} - \log S_i$ is a normal random variable which satisfies:

   $$\log S_{i+1} - \log S_i = (\mu - \frac{1}{2}\sigma^2)(t_{i+1} - t_i) + \sigma(W_{t_{i+1}} - W_{t_i})$$

   Let R be the log or continuous return and $R_{i+1} = \log S_{i+1} - \log S_i$, and $M = \mu - \frac{1}{2}\sigma^2$, then:

   $$R_{i+1} \sim \mathcal{N}(M\Delta t_{i+1}, \sigma^2 \Delta t_{i+1})$$

3. To estimate drift μ and volatility σ in MLEs for each stock, we use:

$$\frac{R_{i+1}}{\sqrt{\Delta t_{i+1}}} \sim \mathcal{N}(M\sqrt{\Delta t_{i+1}},\ \sigma^2)$$

then let $H = \frac{R_{i+1}}{\sqrt{\Delta t_{i+1}}}$, and $F = \sqrt{\Delta t_{i+1}}$. We can use the linear regression to estimate mean and variance for each individual stock.

3 Measure and Methods

1. Before we calculate correlations of real stock data, we need to select stocks that we can apply our method. We assigned a measure from zero to one to the degree to which timelines of two stocks prices match. The way we calculated the Kernel density function estimator for any pair of stocks is following:

$$\hat{f}_h(x) = \frac{1}{n}\sum_{i=1}^{n} K_h(x - x_i) = \frac{1}{nh}\sum_{i=1}^{n} K\Big(\frac{x - x_i}{h}\Big)$$
$$, where\ K(\frac{t - t_i}{\varepsilon}) = e^{-\frac{(t-t_i)^2}{2\varepsilon^2}}$$

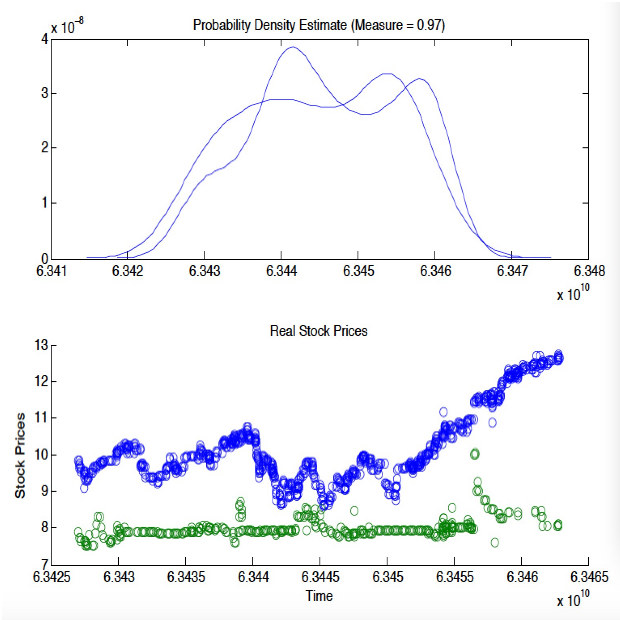
Then we assign a measure to the degree to which two stocks match as following:

$$M(S^1, S^2) = \frac{\int_{t_1}^{t_2} f_1(t) f_2(t) \mathrm{d}t}{\|f_1\|_2 \|f_2\|_2}$$

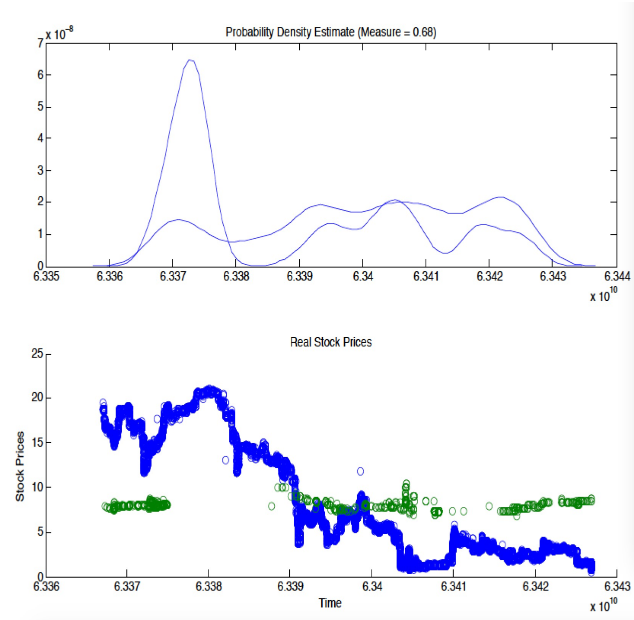
Figure 1 shows the result of kernel density function estimations based on different two pairs of real stocks. One pair has a high measure and the other has a low one, and our cut-off is 0.8, meaning if any pair of stocks has a measure below 0.8 we are not going to calculate its correlation. As you can see, for the low-measure pair there is a significant period in which one stock had transactions and the other did not have any trading activity, indicating that using Brownian bridge would introduce many errors since these two stocks do not match well.

2. To measure the correlation between two stocks with homogeneous timeline, we could let:

$$h_i = (\mu_A - \frac{1}{2}\sigma_A^2)\Delta t_i + \sigma_A \Delta W_A^i$$
$$f_i = (\mu_B - \frac{1}{2}\sigma_B^2)\Delta t_i + \sigma_B \Delta W_B^i$$



(a) Image of Two Stocks With High Measure Result



(b) Image of Two Stocks With Low Measure Result

Figure 1: Density Estimate Measure of Two Stocks

Then the correlation between these two could be calculated by:

$$\begin{aligned}\rho &= \frac{cov(h_i,\ f_i)}{\sqrt{var(h_i)}\sqrt{var(f_i)}}\\ &= \frac{cov(\Delta W_A^i,\ \Delta W_B^i)\sigma_A\sigma_B}{\sigma_A\sigma_B\Delta t_i}\\ &= \frac{1}{\Delta t_i}cov(\Delta W_A^i,\ \Delta W_B^i)\\ &= \frac{1}{\Delta t_i}E[\Delta W_A^i,\ \Delta W_B^i]\end{aligned}$$

Finally, the correlation of two stock could be given from the following formula:

$$\rho = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{\Delta t_i}E[\frac{h_i-(\mu_A-\frac{1}{2}\sigma_A^2)\Delta t_i}{\sigma_A}\times\frac{f_i-(\mu_B-\frac{1}{2}\sigma_B^2)\Delta t_i}{\sigma_B}]$$

3. Naturaly, we can not find two stocks with the same timeline in the real world. Thus, we have to interpolate prices of all stocks at any given specific time. To do this, we first consider using liner interpolation, which is, for the value x in the known interval $(x_0,\ x_1)$, the value y is given from the equation:

$$y = y_0 + (y_1 - y_0)\frac{x - x_0}{x_1 - x_0}$$

However the problem with using linear interpolation is that the log differences of any pair of interpolated prices are not independent with each other. To avoid this issue, we also utilized Brownian Bridge to interpolate prices.
LetW_t be a Brownian motion. Fix $T > 0, a \in \mathbb{R}, and b \in \mathbb{R}$. We define the Brownian bridge from $a to b$ on $[0, T]$ to be the process:

$$X_t^{a->b} = a + \frac{(b-a)t}{T} + X_t, 0 \leq t \leq T, where X_t = W_t - \frac{t}{T} W_T$$

4 Simulation and Numerical Results

1. Simulated Data

To see whether there is a significant difference in the results of correlation between using linear interpolation and Brownian bridge, we conducted a series of simulations. In the simulation, we generate two correlated Brownian motion processes simulating real stock prices on a transaction level and then we employed both linear interpolation and Brownian bridge to interpolate prices to calculate correlation.
In figure 2, it presents the results of simulation. For each path, we randomly generated 20,000 points to simulate real stock prices and interpolated 10, 100, 1,000, and 10,000 prices in the regular grids for the purpose of calculating correlation. And given a number of interpolated prices, we generated 500 paths to calculate the mean and standard deviation of correlations. Our simulation results indicates that there does not exit a substantial difference between linear interpolation and Brownian bridge. But we still used Brownian bridge to calculate correlations of real stock data.

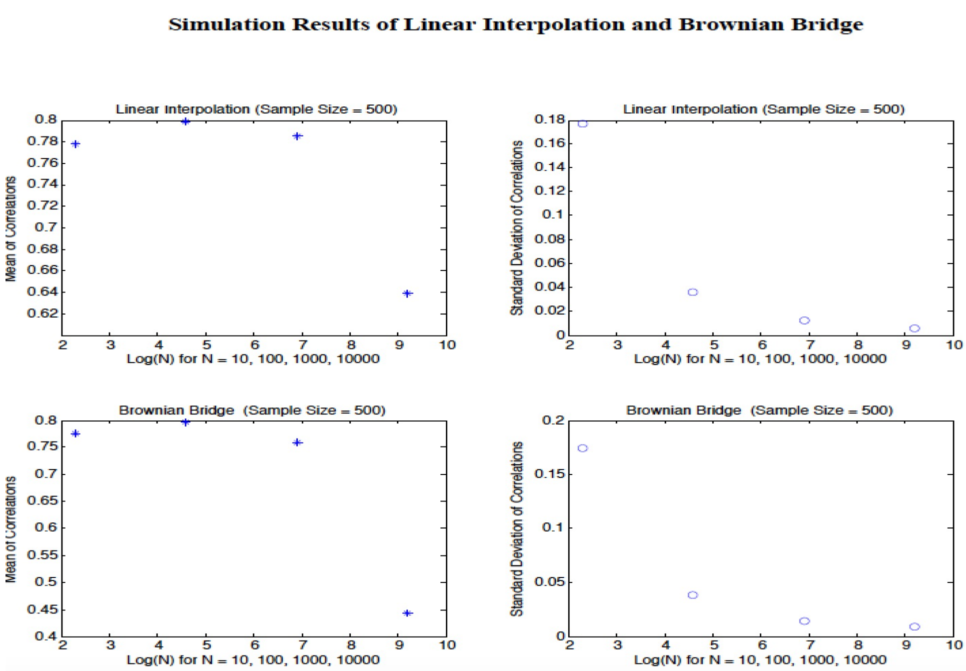


Figure 2: Simulation Results of Linear Interpolation and Brownian Bridge

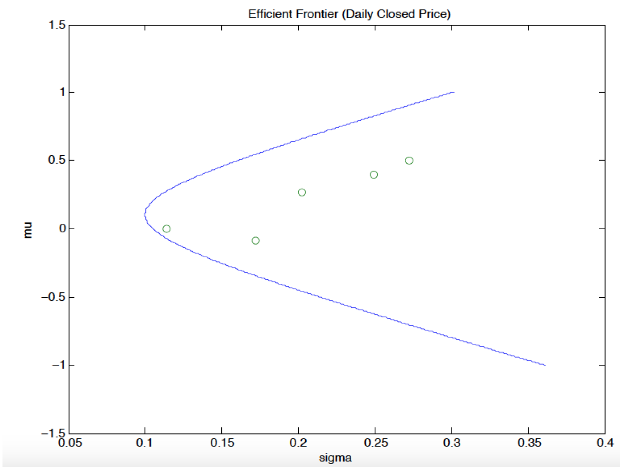
2. Real Data

Figure 3 provides the result of the correlation matrix of five stocks (AT&T, Tesla, Cisco, Ford, and Huntsman) based on the Brownian bridge, linear interpolation, and daily closed prices. For Brownian bridge, we calculated the correlation coefficient because we believe Brownian motion only works for a short time period. So for a pair of stocks, we calculated correlations for each month and take the average of monthly correlations as the final correlation. We also calculated the correlation matrix for these five stocks based on linear interpolation and daily closed prices. Table 1 provides results of these three correlation matrixes from January 1st, 2008 to March 31st, 2008, three months period. We also conducted hypothesis tests to see whether all correlation coefficients are statistically significant from zero. From Table 1, as we can see, all correlation coefficients based on our method are less than those based on daily closed prices.

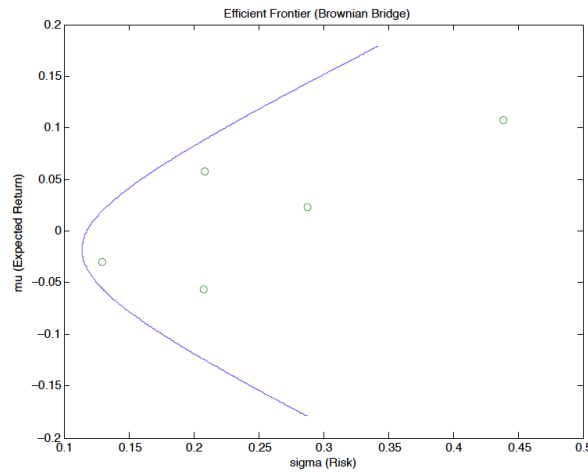
	Correlation Matrix Based on Real Stock Data Using Brownian Bridge				
	Cisco	Ford	Huntsman	Jet Blue	AT&T
Cisco	1				
Ford	0.3402**	1			
Huntsman	0.3182**	0.2277*	1		
Jet Blue	0.3265**	0.2807*	0.2315*	1	
AT&T	0.4268**	0.3452**	0.3307*	0.3454**	1
	Correlation Matrix Based on Real Stock Data Using Linear Interplotation				
	Cisco	Ford	Huntsman	Jet Blue	AT&T
Cisco	1				
Ford	0.2823**	1			
Huntsman	0.4675**	0.3348*	1		
Jet Blue	0.3093*	0.1912*	0.0661**	1	
AT&T	0.3348**	0.4048**	0.4099**	0.4952**	1
	Correlation Matrix Based on Daily Closed Stock Data				
	Cisco	Ford	Huntsman	Jet Blue	AT&T
Cisco	1				
Ford	0.4504**	1			
Huntsman	0.359**	0.3811**	1		
Jet Blue	0.3992**	0.4276**	0.3106**	1	
AT&T	0.5601**	0.4105*	0.3398*	0.4229**	1

*represents 5% significant level ** represents 1% significant level

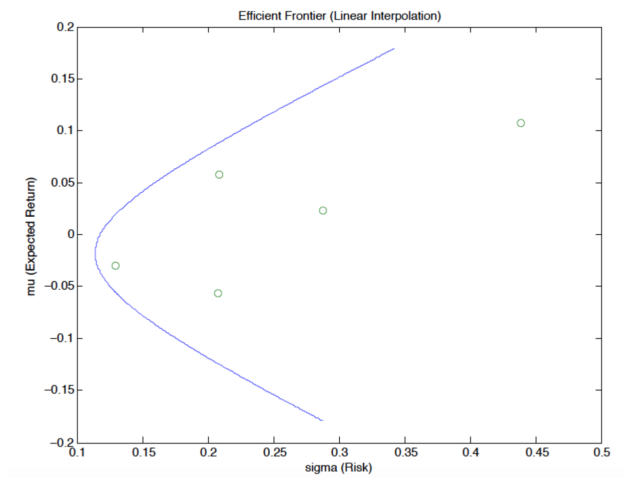
Figure 3: Correlation Matrix Based on Real and Daily Closed Stock Data



(a) Efficient Frontier(Daily Closed Price)



(b) Efficient Frontier(Brownian Bridge)



(c) Efficient Frontier(Linear Interpolation)

Figure 4: Efficient Frontier for a Portfolio with Five Stocks

3. Portfolio Performance

Figure 4 provides mean-variance efficient frontier of back test based on the correlation matrices in Table 1. Each efficient frontier line represents a portfolio's expected return on y-axis and risk (standard deviation of return) on x-axis. Each circle stands for the performance for an individual stock. As we can see, for a given return portfolio always has lower risk (standard deviation) compared to individual stock. This shows a risk reduction for a portfolio due to the correlation of each pair of stocks. On the one hand, although different correlation matrix gives us different efficient frontier, at this stage we are unable to draw the conclusion whether using one method could yield a superior result since back test is not a good indicator to evaluate a particular method. On the other hand, the substantial difference in the correlation matrices shows the meaning to calculate correlation coefficient in a more detailed level, especially for high frequency trading activities.

5 Future Works

In our simulation (Figure 2), one thing caught our eyes is that there seems to be a optimal ratio of number of interpolated points over the number of actual trading points. In the Figure 2, when the number of interpolated points is 100, it yields the best result. Neither more nor less points have a comparable results. This reminds us that there might be a potential relationship between the number of interpolated points and estimated correlation coefficient.

In addition, another way to calculate correlation coefficient based on the real stock data is to compute the weighted average, based on the trading volume, of stock prices for a regular time grid. And we generate new stock price using the average price at each midpoint of the time grid. We are interested in whether this method would yield significantly different results than the results of interpolation.

6 Acknowlegements

Some computational techniques were under the help of professor Jeff Phillips from School of Computing, University of Utah.

References

[1] Harry Markowitz. Portfolio selection*. *The journal of finance*, 7(1):77–91, 1952.