Final Project Report: Index Portfolio Construction

Tung Luu, Khoa Trinh, Nova Nguyen

May 5, 2023

1 Abstract

The Markowitz mean variance-return stock optimization model is one of the most widely used portfolio construction methods, with an emphasis on calculating the distribution of weights across stocks in an existing portfolio based on the historical price performance of each asset. Our project aims to address some shortcomings of the Markowitz model such as the requirement for an existing portfolio where the investors already have in mind which companies they want to invest in. Non-linear optimization of the Markowitz model also tends to generate impractically small positions of some of the available stocks in order to achieve the optimal return or variance level that cannot be meaningfully exercised in real setting. Instead, we propose building a portfolio with the goal of mirroring the performance of an index by choosing a smaller number of stocks from that index, with the assumption that the risk and return performance of the base index would carry over to our chosen portfolio.

2 Introduction

We chose the S&P500 index as the basis for our portfolio, which tracks the performance of the 500 biggest publicly listed companies in the U.S. stock market whose products and services are varied across a range of industries. The S&P500 is considered the best indica-

tor of the U.S. market and economy health among popular indices such as the NASDAQ Composite, the NYSE Composite, and even the Dow Jones Industrial. Therefore, the S&P500 provides the foundation for relatively diversified risks with good return performance. Investing based on the S&P500 is considered one of the most secure investment methods for individuals seeking long-term investments, and there are 2 ways to achieve this: buying an S&P500 index fund or buying all 500 stocks in the index in equal quantity.

However, for short-term investors who tend to adjust their stock inventory based on market conditions or who employ active quantitative trading, keeping track of all 500 stocks is humanly impossible and inefficient, as it requires constant monitoring of both the assets and the financial performance of the underlying companies. It also incurs large transaction costs in the form of income tax every time the investor sells off any amount of their stock holding. On the other hand, buying an S&P500 exchange-traded fund (ETF) is basically the same as buying a single stock with the same level of return as the aggregate return of the 500 stocks, therefore it is not diversified by itself. For example, buying an S&P500 ETF means no flexibility to cut losses on certain stocks that are performing considerably worse than others in the S&P500. Therefore, ideally, we will want to have a portfolio of more than 1 stock and much less than 500 stocks.

To address this issue, we developed a large-scale deterministic model to reduce the number of stocks in our portfolio while maintaining a similar return performance as that of the S&P500. The idea is to cluster the S&P500 into completely separate groups of stocks that have similar price movements, or stocks that have high covariance in other words. All stocks in the S&P500 can only belong to one cluster with no overlap. For each of these clusters, we will then choose one stock with the highest sum of covariance of itself with all stocks in the cluster to put in our index portfolio. The number of clusters is the number of q stocks we want in our portfolio.

After choosing stocks to be included in the portfolio, the next step is to decide how

much money we should invest in each stock. As discussed above, the scenario where we most closely resemble the S&P500 performance is when buying all 500 stocks in equal quantity. To replicate this, we assigned the fraction of the total price of all stocks in each cluster out of the total price of all 500 stocks as the weight of its corresponding representative in our portfolio. Finally, we also want to identify the point of diminishing returns where purchasing additional stocks becomes unnecessary and to determine the optimal number of stocks to include in the portfolio. Results showed that the portfolio return reached a constant value after purchasing more than 7 stocks. This suggests that a portfolio of only 7 stocks can represent the performance of the S&P500.

3 Data Collection

We obtained the list of all stocks listed in the S&P500 by web scraping `Wikipedia`. Information on historical daily price data of these stocks is accessible on `Python` via the `yfinance` package. Two sets of data were created for analysis: a training set and a testing set. The training dataset consisted of values for all 500 stocks in the S&P500, ranging from 2017-12-31 to 2022-12-31. This time range is used to account for stock performance both before and after the start of Covid-19, which contains sharp and volatile movements that deviate from normal.

The training set is used to determine which stocks to be included in the portfolio and their corresponding weight. Exact same values and conditions will then be evaluated on the testing dataset, which was generated using values from 2023-01-01 up to the most recent date of the model run, 2023-05-04. The original list contains 503 companies, and after removing stocks with missing data because they were not yet listed by the start date of the training set, our model samples a total of 487 stocks.

4 Methodology

4.1 Metric

Our model uses the following metric:

$$p_{ij} = \text{similarity between stock i and j}$$

In our model, p_{ij} is the covariance between prices of stock i and j over the period of time in the train set. The more similar two stocks i and j, the larger the value of p_{ij} is. We calculate covariance to measure the similarity between stocks.

4.2 Decision Variables

- y_j is binary variable indicating if we pick stock j from n stocks for our index fund. $y_j = 1$ if we pick stock j for our index fund, 0 otherwise.

$$\begin{cases} y_j = 1 & \text{if we pick stock } j \text{ for the index fund} \\ y_j = 0 & \text{otherwise} \end{cases}$$

- x_{ij} is binary variable indicating if j is the most similar stock to i. In other words, x_{ij} indicates if we can represent stock i using stock j in **our index fund**. $x_{ij} = 1$ if j is the most similar stock to i, 0 otherwise. Note that, if $x_{ij} = 1$, then stock j **MUST** be in our index fund, or $y_j = 1$, because we will need to use stock j to represent stock i.

$$\begin{cases} x_{ij} = 1 & \text{if } j \text{ is the most similar stock to } i \\ x_{ij} = 0 & \text{otherwise} \end{cases}$$

4.3 Objective

Our model aims to cluster all n stocks into different groups of similar stocks (maybe these stocks belong to one industry, or belong to one company). For each group, we will

select some representatives to be included in our index fund. Therefore, the objective of our model is maximizing the similarity between n stocks and their representatives in our index fund, which are q stocks that we pick.

$$\text{Maximize} \quad \sum_{i=1}^{n}\sum_{j=1}^{n} \rho_{ij}x_{ij}$$

With this objective function, our model is trying to pick q stocks such that each stock captures the **maximum similarity with a specific group** in the original pool of n stocks. In addition, for each stock j that we pick, we will find all stocks that are most similar to it to include in its cluster, so that later we can represent all those stocks using only stock j in our index fund.

4.4 Constraints

$$\begin{aligned} \text{subject to} \quad & \sum_{j=1}^{n} y_j = q \\ & \sum_{j=1}^{n} x_{ij} = 1 \qquad \text{for } i = 1, \ldots, n \\ & x_{ij} \leq y_j \qquad \text{for } i = 1, \ldots, n; j = 1, \ldots, n \end{aligned}$$

- The first constraint makes sure that our model to pick exactly q stocks from n stocks.
- The second constraint makes sure that each stock i has exactly one representative stock j in our index fund.
- The third constraint makes sure that each stock i can be represented by stock j **only if** stock j is chosen for our index fund.

4.5 Stock Weight

Once the model has been solved and a set of q stocks has been selected for the portfolio, a weight w_j is calculated for each j in the portfolio:

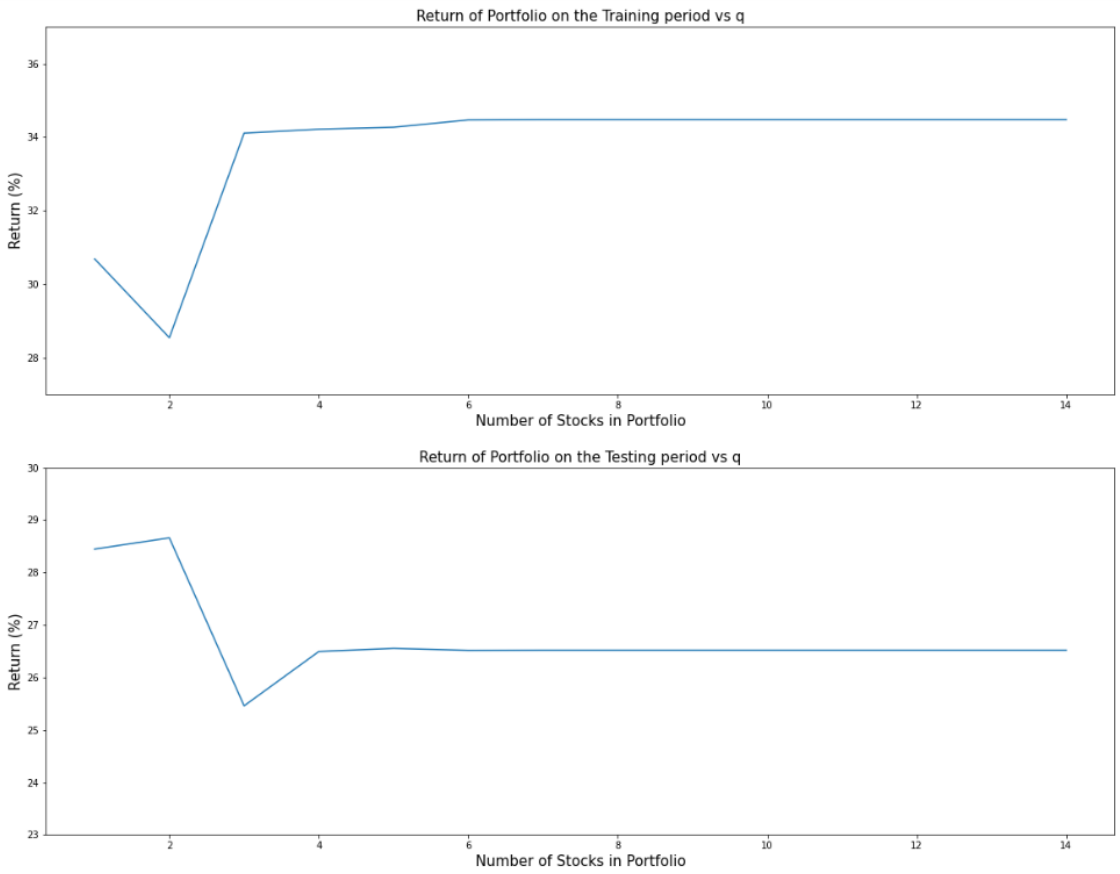
$$w_j = \frac{\sum_{i=1}^{n} V_i x_{ij}}{\sum_{i=1}^{n} V_i} \text{ for } i = 1, \ldots, n$$

where V_i is the price of stock i at the time of buying. So w_j is the fraction of the sum of total market value of all stocks i represented by stock j out of all stocks in the original sample of n stocks.

5 Result

We ran our model for a range of q from 1 to 14 and found that as q increases, the level of return stabilizes around 34.47%. This is very close to the aggregate return if you buy all 500 stocks of the S&P500 which is 36.74% for the same 5 year period. Therefore, we can safely conclude that our model does reasonably satisfy its aim of mirroring the performance of the base index. Replicating the same conditions, same list of stocks chosen and weight of each, on the test set gives us a stabilized return level of 26.52%, which is higher than the actual S&P500 level of 15.6%.

We also compared our returns to that of popular S&P500 ETFs in the market. Both the Schwab, Fidelity and Vanguard 500 Index Fund give about 41-42% profit on the training period. While all 3 perform better than our portfolio on the training set, they give around 6.5% on the test set which is lower than the return of our portfolio and on the lower side compared to actual performance of the 500 stocks. The reason for this deviation of ETFs from the underlying index is because they are traded assets whose value includes investor sentiments that are not reflected in the 500 stock prices alone.



For $q \geq 7$, the returns of our training set portfolio became unchanged. Looking further into concrete results behind each q, we found that every additional stock beyond a portfolio of 7 stocks does not actually represent any of the original 488 sample of stocks.

The table below lists all stocks chosen for a portfolio of 9 stocks ($q = 9$), the weight of each stock, the cluster of all stocks each represents, and the total number of stocks in their respective cluster. As we can see, while A and AAL were chosen by the model ($y_j = 1$), they were chosen purely to satisfy the constraint that the number of stocks chosen must be 9 ($\sum y_j = 9$) even though there are no stock i that is most similar to it (no $x_{ij} = 1$). In other words, since we are not actually buying any amount of either A or AAL, they can be excluded from our portfolio.

Ticker	Weight	Cluster	Num Stocks Rep
A	0.000000	[]	0
AAL	0.000000	[]	0
AZO	0.039443	[APA, ATO, CAH, CF, CNP, COP, CTRA, CVX, DVN, ...	30
BA	0.043437	[AAL, BA, BXP, CCL, DAL, DXC, HII, KMI, LVS, N...	25
BIIB	0.007760	[BIIB, GILD]	2
BKNG	0.057749	[ALK, BEN, BK, BKR, C, CMA, DD, DISH, EIX, FRT...	36
MKTX	0.002935	[INCY, VZ]	2
NFLX	0.000890	[INTC]	1
NVR	0.847785	[A, AAP, AAPL, ABBV, ABT, ACGL, ACN, ADBE, ADI...	391

In fact, while the hard cut-off point beyond which additional stocks do not have any weight is 7, we found that the threshold for additional stocks to start having impractically small weights (the problem that we are trying to avoid with the Markowitz model) is 5. Beyond this point, every new stock represents only about 2 to 3 stocks which amounts to a weight of a decimal fraction out of the entire portfolio. While we can actually invest meaningful amount into these stocks given enough money, the change in return doing this would bring does not affect much of the entire portfolio. This is shown in the graph above where fluctuation in the return level decreases drastically for $q \geq 5$.

At any number of q, NVR stock (NVR Inc., a home construction and real estate company) holds the majority of weight, upwards of about 85% of the entire portfolio and consistently represents most of the S&P500. This is followed by BKNG (Booking Holdings, travel agency and search engine company), AZO (AutoZone, automotive retailer) and BA (Boeing, aircraft manufacturer), with each making up 5% of the portfolio. While we have a good mix of differentiated industries with these stocks, our portfolio is rather skewed in weight, and we cannot conclude that each company chosen is representative of the industry they belong to because real estate only makes up 2.5% in market cap of the S&P500. We are also missing core industries such as banking, healthcare and energy.

However, this result is interesting because the market cap of companies in our portfolio are inversely proportional to the amount invested. As of May 2023, NVR ranks rather low in terms of market cap, 343 out of 503 companies, AZO at 136, BKNG and BA at markedly higher of 76 and 66 respectively. Yet, this small to mid-sized cap company, relative to the S&P500, was enough to mirror the returns of the entire index given appropriate investment weight. Therefore, what we lack in industry diversification, our portfolio makes up for it in terms of market-cap diversification where companies with higher market caps are considered blue chips or safer and less volatile.

6 Future Work

While we were able to identify the point of diminishing marginal return of the effect of the number of stocks in the portfolio on its return ($q = 7$), there is one other input item that we can work on modeling without hand-picking: the range of time for the training set. We only had 1 training period for this project, and while our return level matches up rather well with the actual S&P500 performance, this sample size is not at all enough to conclude that the model works well for all training period. Theoretically speaking, a longer period for the training set would provide us with more data to mirror actual index performance more closely. However, at a 5 year period, we needed to exclude nearly 20 newer companies and as the sample period increases, we would need to exclude even more. This means past composition of the S&P500 does not necessarily reflect well on its future composition and performance. For future iteration, we would like to work on identifying a balance between these two factors. Another option is to develop a model that can work with differing time series for each company.

Our model also has room to be more robust by adding more variables and constraints depending on the investing style. For example, we can add information regarding the industry of each company as variable, and require the model to choose companies from

at least a fixed number of industries for more risk diversification. We can also incorporate return and variance of the chosen portfolio as a constraint with an upper and lower bound level for the investor to customize their desired level of portfolio performance.

References

[1] Cornuejols, G., Tutuncu, R. (2006, January). Optimization methods in finance, from https://cs.brown.edu/courses/cs1951g/slides.html

[2] S&P 500 Companies by Weight from https://www.slickcharts.com/sp500

[3] The Top 25 Stocks in the S&P 500 from https://www.investopedia.com/ask/answers/08/find-stocks-in-sp500.asp