Problem Statement

Our goal is to construct a simulated stock portfolio using the components of the Standard & Poor's 500 Index, as well as United States Treasury bonds. Our approach is strongly influenced by the framework described in "Sparse and stable Markowitz portfolios" (Brodie et al., 2009). A Markowitz portfolio is one that uses numerical optimization to achieve a pre-specified level of returns with minimum variance. Sparsity means that most of our portfolio weights - the proportion of total capital allocated to a given asset - should be set to zero. That is, our portfolio should only contain a handful of 'active positions' at a given time. Stability means that we hope to achieve consistent returns month-to-month, avoiding large fluctuations. To construct our portfolios, we use publicly available historical daily prices for the components of the SP 500, as well as US Treasury bond yields.

Our main question of interest is whether we can consistently out-perform the SP 500 index - a weighted average of all 500 components - by achieving comparable returns with lower month-to-month volatility, using a sparse portfolio. We examine the risk/reward structure of portfolios with different amounts of short selling - the practice of selling borrowed shares of a stock in order to profit from an anticipated decline in the stock's price. We also explore different techniques for predicting future returns, based on past returns and other factors. In particular, we are curious about whether Twitter sentiment analysis can be useful in predicting returns.

Our project is relevant to the financial industry, and in particular to wealth management. There is a growing interest in passive or algorithmic fund management, which attempts to replace or supplement human decision-making with a quantitative algorithm. Existing ETFs like VFINX and VOO track the SP 500 (with fees in the neighborhood of 10-15 basis points) but many investors would likely be interested in a fund that also offers significantly reduced volatility. Since our goal is a sparse portfolio, our work may be of particular interest to small private investors. Such investors face non-negligible transaction costs in the form of brokerage fees, and must limit their number of active positions.

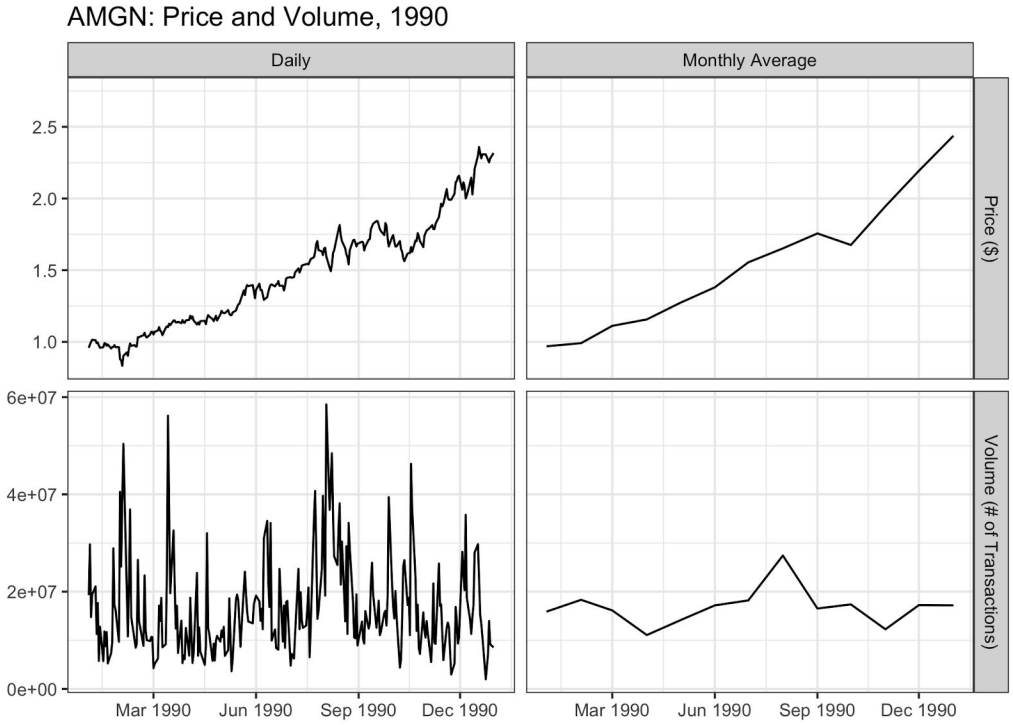
Data Description

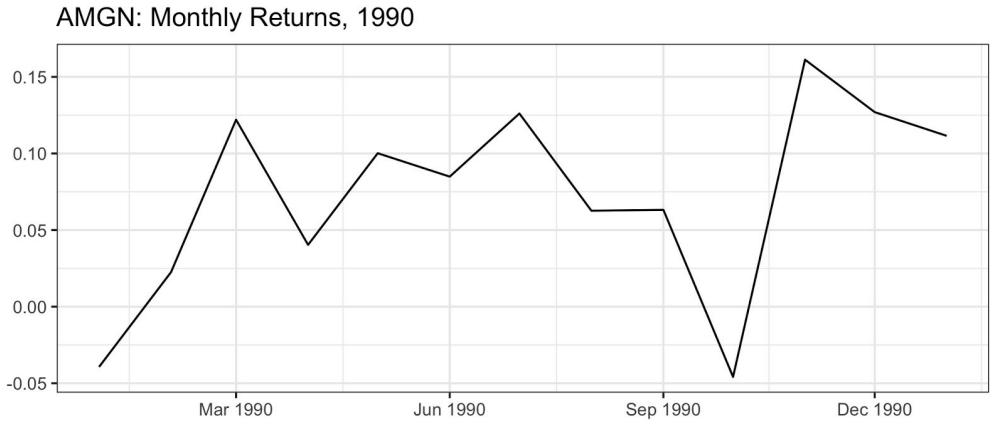
Yahoo! Finance maintains a publicly searchable database of historical prices, including all current components of the SP 500. We used a bash script to download the historical prices for each stock, starting either in January 1980 or the date when the stock joined the SP 500, whichever came later. Since there is some turnover in the SP 500, our data for the 1980s, 90s, and aughts has fewer than 500 different stocks. That is not a problem, however, since our main goal is to make predictions for the years 2013 onward. Yahoo! Finance also has data on dividends (small quarterly payments that some companies issue to their shareholders) and volume (the total number of shares traded on a given day).

The raw data for each stock is a csv file that includes the ticker symbol (e.g. AAPL for Apple), the date, volume, opening price, closing price, and adjusted closing price. The adjusted price takes stock splits into account; when a company's value rises significantly over time, they sometimes split each share into multiple shares. Where applicable, there is a separate csv file with the date and value of dividends. In total, we have approximately 3.5 million rows of raw data, requiring about 200 megabytes of hard drive space.

The most important variables is the adjusted closing price, which is used to calculate returns. Before conducting our analysis, we aggregated the prices and volume for each stock by month. Returns for a given month are calculated as the percent change in the monthly price, compared to the previous month. A return of 0.05 in month t means that the average price rose 5% from month $t-1$ to month t. The

total returns of our portfolio in a given year are a function of 1. the returns of each stock during each month of that year and 2. the weight we assign to each stock. The volatility of our portfolio in a given year can be measured by the variance of its returns during each month of that year.





Methods

Regrettably, investors lack routine access to crystal balls. Before we can optimize a portfolio, we need to predict future returns. Brodie et al. do this by using a simple average of each stock's returns in the past 12 months. We use this method as a baseline. We will develop a more sophisticated prediction model, and assess whether this yields more accurate results. For a given year, we store the predicted monthly returns in a matrix **R** with 12 rows and N columns, where N is the number of stocks.

Our portfolio is expressed as **w**, a vector of weights or allocations with length N. The *ith* element represents the proportion of total capital invested in the *ith* stock. By convention, total amount of capital is

one unit: $\mathbf{w}^T \mathbf{1}_N = 1$. The average predicted returns for each stock are in **mu-hat**, an N-vector obtained by averaging the N columns of **R**. We choose *p,* the target monthly return that we would like to achieve. A portfolio that achieves this return satisfies $\mathbf{w}^T$ **mu-hat** $= p$. The predicted variability of our portfolio can be expressed as the sum of squared deviations of monthly returns from the target return: $|| p \mathbf{1}_T - \mathbf{w}^T \mathbf{R} ||_2^2$, where T is the number of time periods (12). Brodie et al. suggest adding a penalty term based on the L1 norm of the portfolio weights. This stabilizes the optimization problem, which can be very unstable if returns are collinear, as they often are. It also discourages high levels of short selling, and encourages a sparse solution. We can choose different levels of sparsity by changing the tuning parameter tau.

The optimal portfolio **w-hat** can thus be expressed as follows (Brodie et al.):

$$\mathbf{w\text{-}hat} = \arg\min_w [\, || p \mathbf{1}_T - \mathbf{w}^T \mathbf{R} ||_2^2 \; + \; \text{tau} \, ||\mathbf{w}||_1 \,] \text{, subject to the constraint } \mathbf{A}^T \mathbf{w} = \mathbf{a},$$

where **A** is a 2 by N matrix whose first row is **mu-hat** and second row is $\mathbf{1}_N$, and **a** is a vector of length 2, whose first element is *p* and second element is 1. The expression **A w** $=$ **a** means that **w** achieves the specified return *p* and represents 1 total unit of capital. The optimal solution can be approximated by moving the constraint into the objective function. The tuning parameter epsilon represents the trade-off between adhering to the constraints and achieving low variability.

$$\mathbf{w\text{-}hat} \approx \arg\min_w [\, \text{epsilon} \, || p \mathbf{1}_T - \mathbf{w}^T \mathbf{R} ||_2^2 \; + \; \text{tau} \, ||\mathbf{w}||_1 \; + \; ||\mathbf{A}\,\mathbf{w} - \mathrm{a}||_2^2 \,],$$

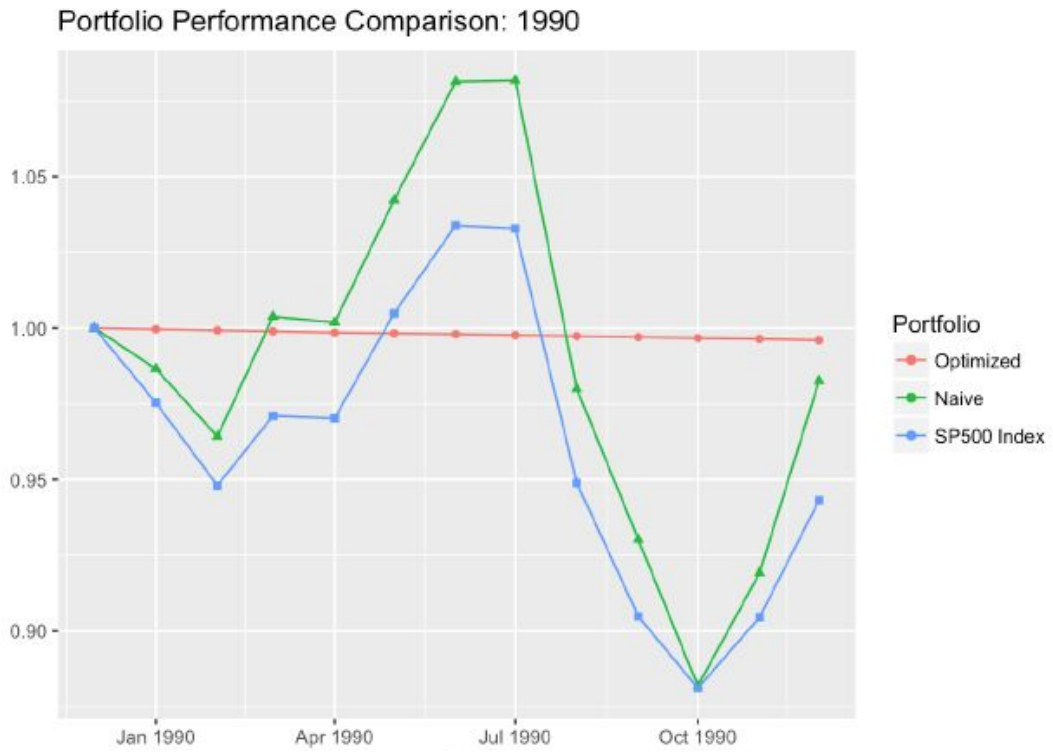
where the term $||\mathbf{A}\,\mathbf{w} - \mathrm{a}||_2^2$ measures the extent to which the constraint is satisfied. Our current results are based on the approximate solution, but we hope to implement the precise solution in the future.

Results

Predicting returns and optimizing a portfolio based on those returns are two separate challenges. Since we are still in the process of fine-tuning our prediction technique, we decided to use actual returns in 1990 instead of predicted returns as a proof-of-concept for our optimization method. These results can be interpreted as a proxy for what would be achieved if we had highly accurate predictions of monthly returns.

We solved the above optimization problem with the 'conjugate gradient' iterative method, implemented in the R optim() function. We tried different values of the tuning parameters tau and epsilon, ranging from 10^{-10} to 10^3. Some of these yielded portfolios that were unacceptable, either because they involved excess short-selling ($||\mathbf{w}||_1 > 2$) or because they were more volatile than the 'naive portfolio' which assigns equal weight (N^{-1}) to all assets. After discarding the 'unacceptable' portfolios, we chose the portfolio that achieved the lowest variance. Interestingly, this is also the sparsest portfolio: 20% of all weights are 0, and 89% of total capital is allocated to the top 40 assets.

This plot shows the performance of the portfolio, compared to the SP 500 index and the 'naive' portfolio. In a year when the SP 500 displayed high volatility and declined in value, our portfolio was stable and largely preserved its value.



	Average Monthly Return	Variance of Monthly Returns
Optimized Portfolio	-0.0003352	5.344e-09
Naive Portfolio	-0.0003627	0.002373
SP 500 Index	-0.004194	0.001443

Conclusions

It is trivial to choose a stock portfolio with the benefit of perfect hindsight. Although that is essentially what we did, our success is still meaningful. Our optimization algorithm identified a solution that achieved better returns with much lower variance, compared to the SP 500 index over the same period. This serves as a proof of concept for our approach to portfolio construction. Our immediate next step will be to repeat the optimization using actual predictions, instead of the true returns.

One major challenge was solving the precise, instead of approximate, optimization problem. To our knowledge, base R includes a function for optimization with linear inequality constraints, but not linear equality constraints. We will explore other optimization techniques in order to use the precise implementation, including an approach suggested by Brodie et al. We are hopeful that this might lead to a solution that is more sparse: our current solution is only 20% sparse, as discussed above, and there is

significant room for improvement. We will also try to achieve greater sparsity by making fine-grained changes to the tuning parameters.

We plan to incorporate dividends into the model. For simplicity, we will combine the dividends with the existing returns data: whenever a given stock issues dividends, this will be reflected as a higher return for that month. Additionally, we will update our portfolio to include US Treasury bonds. As a first step, we will allocate a fixed 10% of total capital to bonds with a mix of maturity dates; at the end of each year, we will re-balance our portfolio to keep the portion invested in bonds constant. Time permitting, we will try to develop a method to choose the bond allocation dynamically, in response to predicted stock market conditions; if we anticipate high stock returns, we will invest less in stocks, and vice versa.

Another topic for future exploration is the creation of higher-level prediction features from day-level data. For example, if a short-term moving average of a stock's price crosses above a longer-term moving average, this is known as a 'golden-cross,' and can be a signal that prices will rise, especially if accompanied by an increase in trading volume. When the reverse happens, it is called a 'death cross.'

References

"Sparse and stable Markowitz portfolios" Joshua Brodie, Ingrid Daubechies, Christine De Mol, Domenico Giannone, and Ignace Loris, 2009
http://www.investopedia.com/terms/g/goldencross.asp