Use of Neural Networks (Autoencoding) for Stock Picking for the implementation of the Markovitz Portfolio Theory

Sai Kumar M 2014B3A70951H

Arvind A S 2014B3A70381H

ABSTRACT

One of the major obstacles faced in using Machine Learning in financial forecasts is the dimensionality problem (where the number of dimensions is too many – here the return of every past month is one dimension). Therefore, if stocks with smaller reconstruction errors (that is stocks which can be explained with fewer dimensions) are picked, then financial techniques tried on them should work better. The **Markovitz Theorem** gets affected by the same problem, this project considered whether better results can be obtained if stocks with smaller reconstruction error are used.

1. Introduction

1.1. Modern Portfolio Theory

Markovitz portfolio theory or the Modern portfolio theory hypothesizes that risk-averse investors can maximise return for a given level of risk or conversely minimise risk for a given level of return.

It suggests that it is possible to construct an efficient 'frontier' of optimal portfolios using historic data of a collection of stocks by varying risk and return at each point. This efficient portfolio would give the composition of stocks (weights) that maximise expected return for specified risk measured by standard deviation. The reduction in risk occurs through diversification of stocks as each stock's price is differently correlated with the price of the other. This reduces the effect of adverse price changes of one stock on the value of the portfolio to be minimal.

The R program used to generate the Markovitz weights does so by the Monte Carlo method [Raeside, D. E. (1976)]. The program creates half a million randomly weighted portfolios and for every target return, picks the one with the least variance. This portfolio, although not exactly the globally minimal variance portfolio, is close to the global minima by an order of 10^{-6}.

1.2. Neural Networks

Neural Networks have become very popular and effective in approximating target functions that discrete or real valued. It composed of layers of units, each unit representing a value. Neural networks send data through the network from one layer to another. The advantage of neural networks is in the fact that information is sent in a non-linear fashion through the presence of noon linear function like Relu (rectified linear unit) or tanhx. This insertion of non-linearity allows neural networks to model the real world accurately.

1.3. Autoencoding

Autoencoding is a technique that is used to convert high dimensional data into low dimensional data [see G. E. Hinton and R. R. Salakhutdinov (2006)]. It is a non-linear method of finding the number of dimensions (attributes) that can be reduced for any analysis without major loss in accuracy. This can be seen by calculating the reconstruction error. Autoencoding has been found to work better than Principal Component analysis for dimensionality reduction. This can be attributed to the fact that PCA is linear unlike Autoencoding.

In Autoencoding, the information is encoded into a hidden layer (in the neural network) with a reduced set of dimensions and then decoded back into the original

number of dimensions. The autoencoder follows a bottle neck architecture, where the corresponding first and last rows have the same number of dimensions in the neural network.

The encoding and decoding will lead to a loss in accuracy. The difference is the reconstruction error. There are several ways to calculate the reconstruction error including absolute difference, square of errors etc.

Therefore, we can infer that information with smaller reconstruction errors can be characterised with a smaller number of dimensions than those with larger reconstruction errors.

2. Methodology:

First, data is obtained in wide format and checked for inconsistencies in the data. Once the inconsistencies have been handled, companies are chosen that satisfy the survivorship bias for the period in which the training and the testing is to be done. In the current analysis stocks returns for the US stock market were taken. The training data on which the autoencoding was performed was from Jan 1980 to Dec 1989. The testing period was from Jan 1990 to Dec 1999.

Autoencoding (in this project a feed forward neural network with the bottle neck architecture was implemented) is performed over all such available companies. (that have consistent data and satisfy survivorship bias) The 30 companies with least reconstruction error are chosen to construct a portfolio called a **reconstructed portfolio**. Another 30 companies satisfying survivorship bias and not exclusive to the reconstructed portfolio are chosen randomly by a random generator function to construct a **random portfolio** for comparison.

As said above a normal bottle neck architecture was used in this project. It was a 3-layered neural network with the first and last layer have 120 units/ dimension and the middle-hidden layer being composed of only 80 units. Each layer utilised a Relu function that takes the max (0, x) for the value of a unit (where x is the value assigned to a unit from the previous layer)

Every cycle/window is for a period of 120 months and the Markovitz portfolio weights are calculated for every window. For example, the weights for the 121st month is calculated on the basis of the window from the 1st-120th month. Subsequently every month's portfolio weights are calculated on the basis of information present in the previous 120 months.

The weights are calculated for target returns ranging from 1.1-1.6% per month. The weights are calculated for both portfolios with constraints that all money is invested and no shorting of stock is allowed.

The companies in both these portfolios are held for the next 120 months with rebalancing at the end of every month for each target return.

3. Data Description

The US financial market was chosen for the study due to availability of reliable market data for long periods of time. For the purpose of the study, the years 1980-2000 were chosen. The US market made an average gain of 17.99% annually or roughly 1.38% monthly during this period. Although this period was accentuated with four crises, namely in 1980, 1987, 1989, 1999-2000, the study should not be affected by these as these factors would be accounted for in the creation of portfolios.

4. Results and discussion:

4.1. Companies selected by the Autoencoding process

Running our dataset through the autoencoding process and selecting the companies with the least reconstruction errors, we obtain the following companies. The companies tabulated below are ordered in increasing order of reconstruction errors. The average market cap of these companies over the entire period was $ 934 million. Thirty companies were chosen to create the portfolio and of the thirty, 23 companies are in the financial sector and of the remaining 7, one is in the telecommunications industry and the rest are in Energy. This gives us a key insight that companies with low reconstruction errors happen to be mostly in financial services, implying that this industry has some characteristics that make its stock movements and price predictable. This could be because financial institutions are highly regulated and must comply with US Federal laws on risk and liquidity, thereby keeping the stock fairly stable and predictable. Or it could also be each of these institutions ability to hedge risks internally that keeps them from experiencing high uncertainty. Similarly, Energy companies must also have such characteristics. It could possibly be the effect of the predictability of energy demand that kept prices from experiencing high uncertainty. Since the companies chosen in the random portfolio have no choosing criterion, the industries to which they belong are irrelevant. The results are shown in Table 1 in the appendix.

The average return of the companies in the reconstructed portfolio was 1.32% monthly return and the average standard deviation is 4.26%. For the companies in the random portfolio the average monthly return is slightly higher at 1.51% and the average standard deviation is nearly twice at 9.76%. The statistics related to the individual companies are given in Table 2 and Table 3 in the appendix.

Table 4. Mean and St. deviation of monthly returns from 1980 to 1990

Random portfolio		Reconstructed portfolio	
Mean Return (%)	St. Deviation	Mean Return (%)	St. Deviation
1.32	4.26	1.51	9.76

4.2. Returns of the portfolios

From the Markovitz portfolios created for both sets of stocks (chosen by reconstructed error and randomly) and reweighting of the portfolio after every rolling period, we obtain the following results over the chosen 10-year testing period (1990 to 2000).

Table 5. Mean and standard deviation of stocks in both portfolios

Target return (%)	Random portfolio		Reconstructed portfolio	
	Mean Return (%)	Standard Deviation	Mean Return (%)	Standard Deviation
1.1	0.5798465	4.9901117	0.7311146	3.3589989
1.2	1.5361861	4.5553901	0.4990832	3.0747824
1.3	1.2441181	4.5315598	1.0132301	3.2022942
1.4	1.237607	5.0320513	0.9493012	2.8919891
1.5	0.9998721	4.8865048	1.0767933	3.3254934
1.6	1.708661	5.3315468	1.1980237	3.2839328

Firstly, with increasing target return, an increase in the mean return of the Markovitz portfolios in both the randomly selected stocks and Reconstructed portfolio is expected. But due to some idiosyncrasies of the data in the selected time period, a non-conformity to expectations is seen.

Next, the means of the random portfolio are greater than the reconstructed portfolio (for 4 out of the 6 target returns) but the standard deviations are also significantly greater for all the target returns. This holds consistently with our expectation that the reconstructed portfolio has better predictability characteristics and should perform better than randomly selected stocks on average.

Table 6. Median monthly Return for the 2 portfolios

Target return (%)	Median	
	Random (%)	Reconstructed (%)
1.1	0.692002	0.8985117
1.2	1.1847676	0.6227762
1.3	1.4984569	0.6699698
1.4	1.4221413	0.6667867
1.5	1.3501493	0.7848497
1.6	1.4691266	0.8510625

Comparing the median returns of both the randomly selected and reconstructed stocks, we observe a non-uniform pattern of increasing and decreasing medians. This could be because of the nature of stocks selected. At this juncture, it is unclear as to why there isn't an increasing trend in the medians.

Table 7. The RMSE of the monthly returns when compared to the target return

Target return (%)	Root mean square error	
	Random	Reconstructed
1.1	30.206129	13.7027402
1.2	25.0375204	11.9346857
1.3	24.6457886	12.4043104
1.4	30.4174945	10.2800763
1.5	28.9536681	13.4856124
1.6	34.1246387	13.1349597

Comparing root mean square errors of the two types of portfolios created, it can be observed that the RMSE error of the Random portfolio is much higher**, almost twice as high** as that of the reconstructed portfolio. This is a direct result of the autoencoding process which selects highly predictable stocks, thereby reducing the error significantly.

Table 8. Mean difference test between returns of reconstructed and random portfolio over 1990 - 2000

Target return	t-statistic	p-value
1.1	0.2743	0.7841
1.2	-2.0585	0.0408
1.3	-0.4540	0.6504
1.4	-0.5419	0.5885
1.5	0.1420	0.8872
1.6	-0.8896	0.3748

Comparing the obtained mean returns for each target return for both portfolios by means of a **mean difference test**, we obtain the results tabulated above. The test statistics prove all mean returns obtained as insignificant at the 5% confidence level except at a target rate of 1.2 %. A plausible explanation of the insignificance of the mean returns could be the high magnitude of the standard deviations, which are of the same order as the mean return but higher. Due to the insignificance of these mean returns, further evidence to conclusively prove that reconstruction error on autoencoding is a good stock picking criterion is necessitated. Additionally, it has been observed that the average market cap of the companies in the reconstructed portfolio is larger than that of the random portfolio. This shows that the variability of larger stocks is inherently lower than smaller stocks.

5. Graphs:

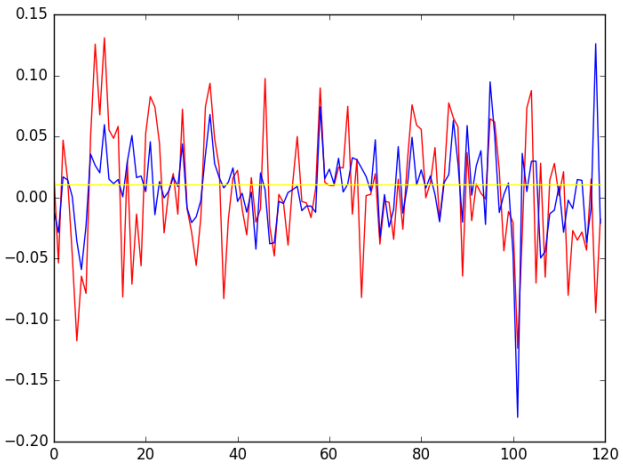
5.1. Line graphs depicting the returns over 120 months

Blue - returns for reconstructed portfolio.

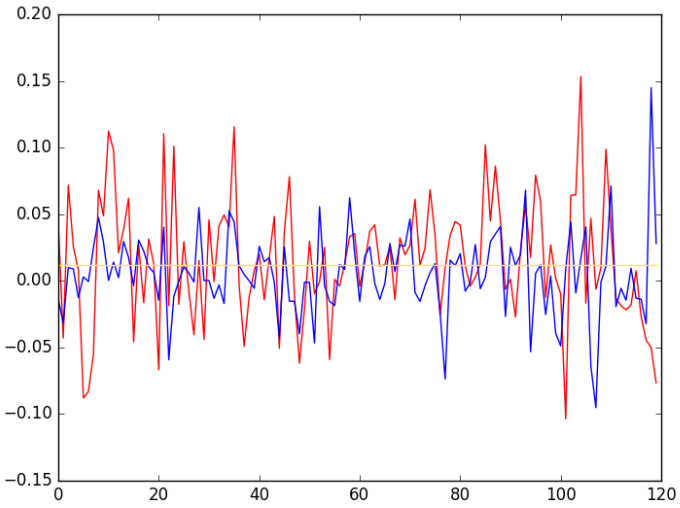
Red – returns for random portfolio

Yellow – Target return

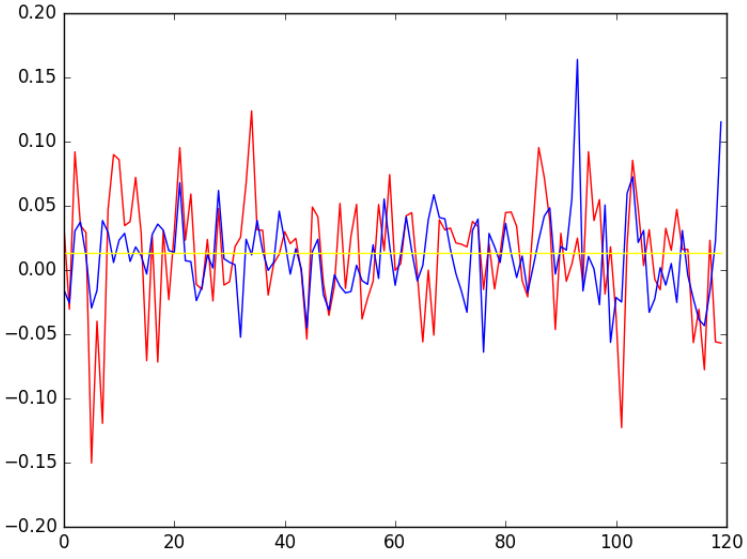
1.1% expected returns-



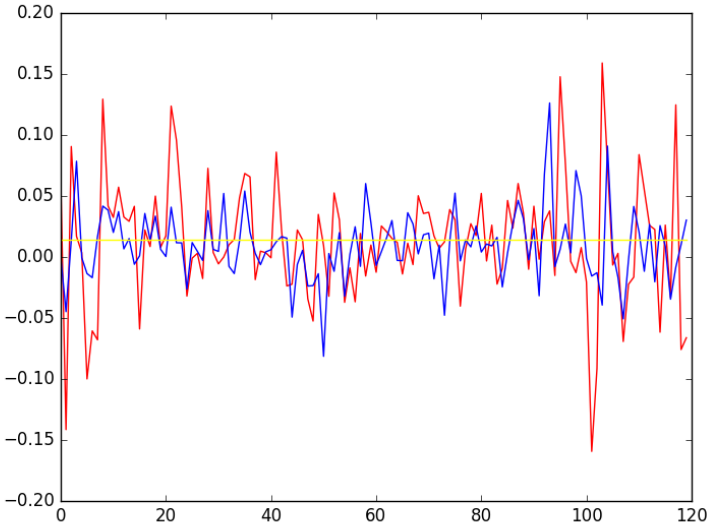
1.2% expected return-



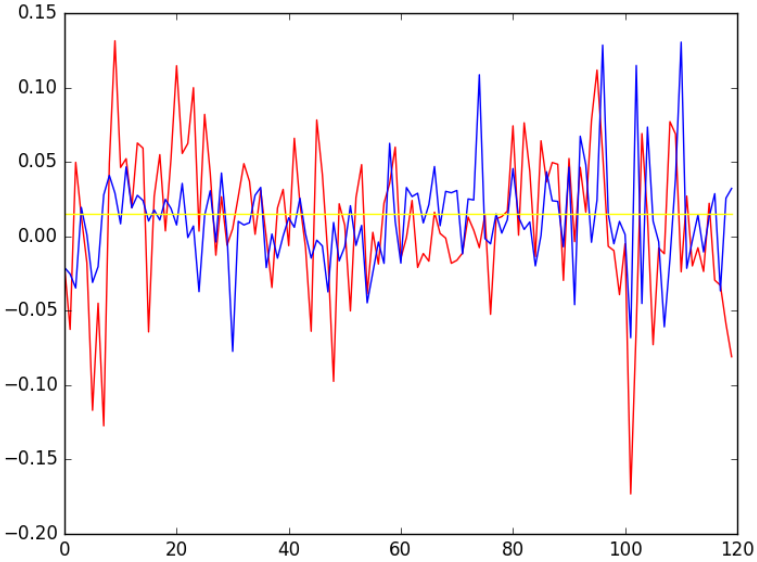
1.3% expected return-



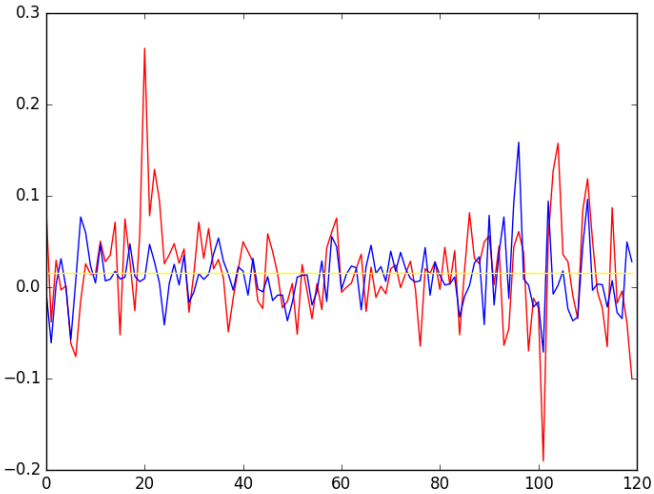
1.4% expected return-



1.5% expected return-



1.6% expected return-

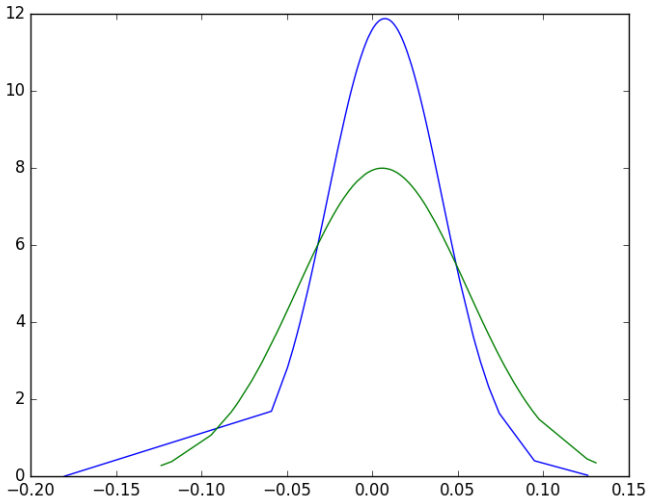


5.2. Normal distribution of the returns for Reconstructed and Random portfolios

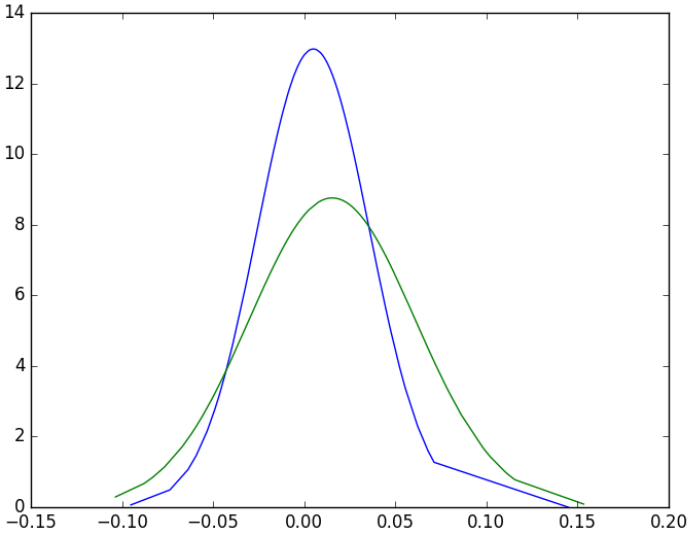
Blue – Reconstructed portfolio

Green – Random portfolio

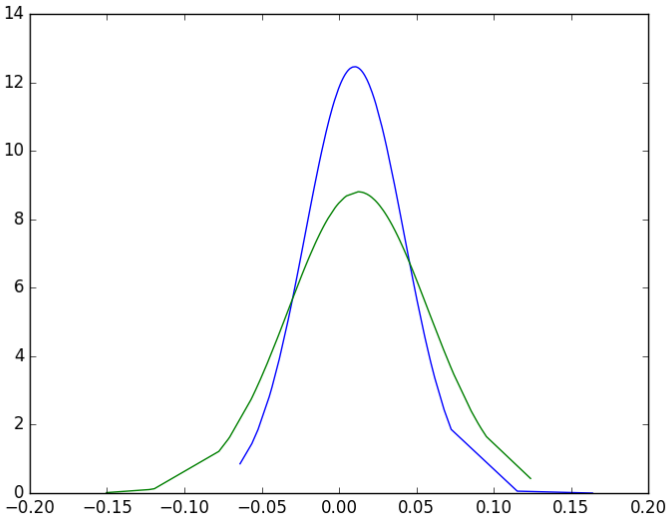
1.1% expected return-



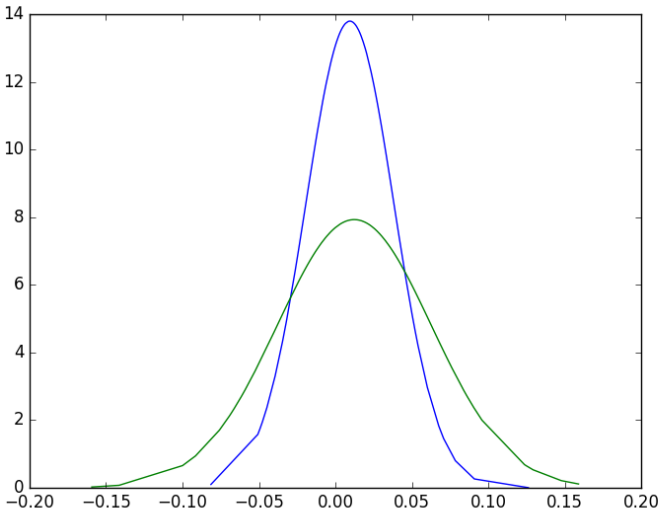
1.2% expected return-



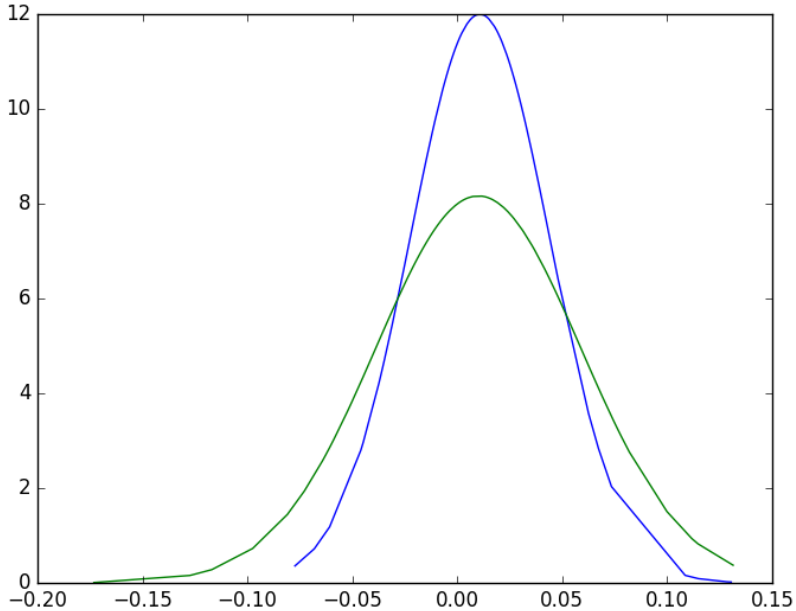
1.3% expected return-



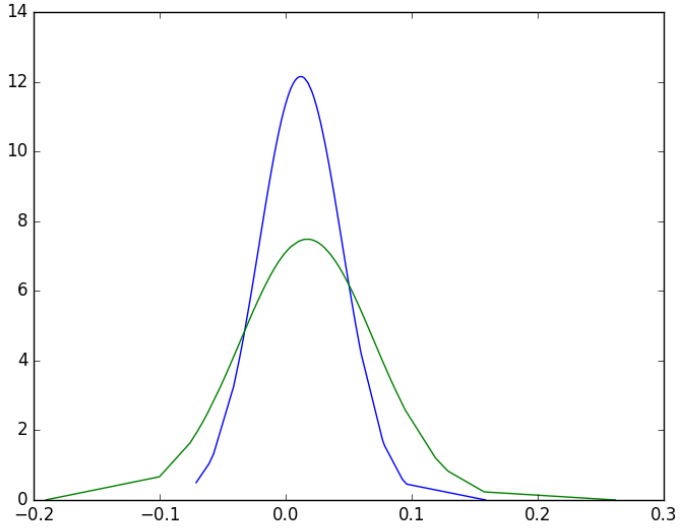
1.4% expected return-



1.5% expected return-



1.6% expected return-



6. Conclusion

From the results obtained we can conclude that the reconstructed portfolio exhibits small variability than the random portfolio. This is consistent with the hypothesis. Additionally, the fact that market capitalization of the stocks in the reconstructed portfolio is higher than random, is consistent with existing literature that states that larger stocks exhibit lower variability than smaller stocks.

7. Future Work

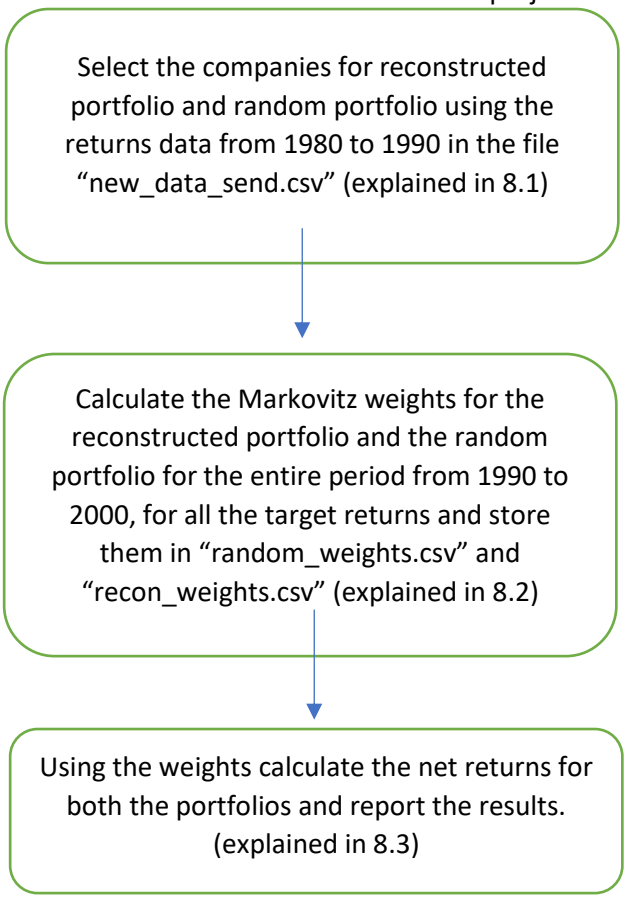
To further improve the results, an autoencoder instead of a feed forward network could be implemented. The same methodology could be implemented with PCA (Principle Component Analysis) instead of an autoencoder for choosing the stocks. Additionally, instead of keeping the companies in the portfolios constant, the composition could be changed every window, to see whether results can improve.

8. References

1. Raeside, D. E. (1976). Monte Carlo Principles and Applications. *Physics in Medicine & Biology*, 181.
2. G. E. Hinton (2006). Reducing the Dimensionality of Data with Neural Networks. *Science*,504

9. Appendix

Flowchart 1. Code flow for the entire project



9.1. Selection of Companies

a. Reading the file "New_data_send.csv" where each row is the financial data of a company and each company represents monthly returns of a particular stock.
b. Run a neural network with 3-layer over the data (from 1980 to 1990) and get the reconstructed values for all the companies. (the neural network will return the original returns but with some error because of the autoencoding)

c. Calculate the sum of square of errors of the original and reconstructed data for each company.
d. Rearrange companies in an ascending order of their reconstruction error.
e. Choose 30 companies with least error satisfying the condition that they have returns from 1980 to 2000. This constitutes the reconstructed portfolio.
f. Choose 30 random companies that have returns from 1980 to 2000. This constitutes random portfolio.

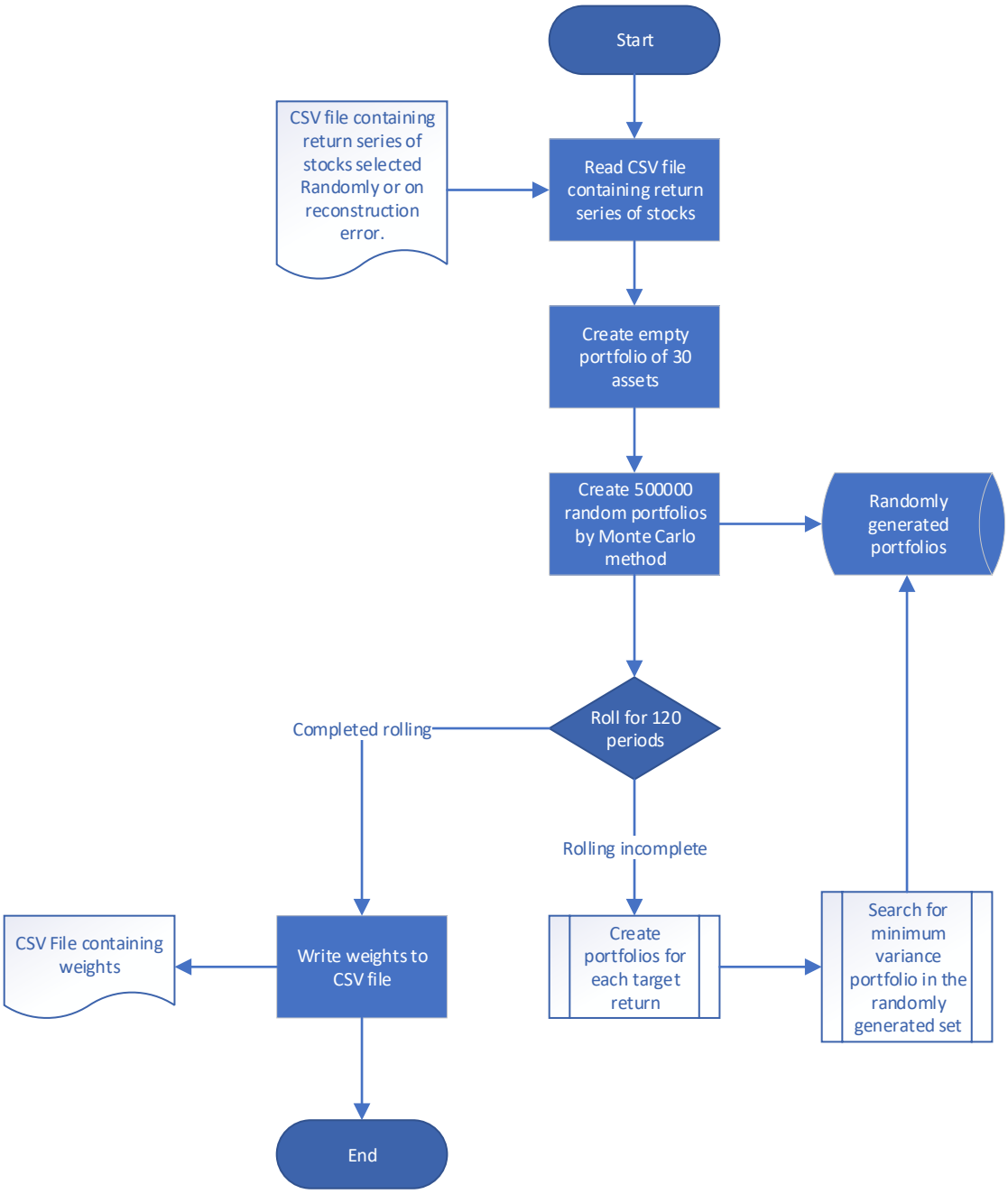
9.2. Calculating Markovitz weights

The Markovitz weights for each of the stocks in each of the time periods is calculated by an R code that does so by the Monte Carlo method. The code creates half a million randomly weighted portfolios by assigning random weights to 30 stocks in each portfolio. Since this method is severely limited by the computational power, its precision is only up to the 3rd decimal, and hence we choose to create only half a million of them, although most solutions can be found in this limited set.

The program workflow is as follows:

a) The program first reads the input file of stock returns.
b) Create time series object of stock returns.
c) Initialise empty portfolio of 30 stocks.
d) Generate 0.5 million random portfolio returns subject to constraints such as full investment and max and min investment limits on individual stocks.
e) Find Markovitz portfolio weights by iterating through the collection of random portfolios to find minimum variance portfolio satisfying target return criteria.
f) Repeat e) for all desired target returns.
g) Repeat e) – f) for each rolling window.
h) Write all generated weights to CSV file.

Flowchart 2. Program flow of Markovitz portfolio creation in R



9.3. Calculation of returns for the portfolio

a. Read the file “random_weights” and “recon_weights” and store in matrices.
b. For each cycle/window, read a set of 6 required row from “random_weights”. And “recon_weights”. This will have the weights\ of the companies for that portfolio. (6 rows are taken at a time because each set corresponds to one window -like 1st to 123th

month window- and each row corresponds to one target return, the first row is for 1.1% and the last row of the set corresponds to 1.6%)

c. Calculate the net return for each target return for the portfolio.

$$\text{Net Return} = \Sigma\ [(\text{return of company X})*(\text{weights assigned to company X})]$$

d. Append the returns of the reconstructed and random portfolios to the respective rows of “random_row” and “reconstruct_row” (The first row is for 1.1%, second is for 1.2% and so on)
e. Perform the statistical analysis on the “reconstruct_row” and “random_row” for each target return (example the mean return, median calculation. Standard deviation)
f. Draw the relevant graphs.

Table 1. List of Stocks chosen by the autoencoding process

Company Name **IIII**	**Sector**
Cornerstone Total Return Fund Inc	Financial Services
Western Asset Income Fund	Financial Services
Lincoln National Income Fund inc	Financial Services
Allmerica Securities Trust	Financial Services
Hartford Income Shares Fund Inc	Financial Services
Prospect Street Inc	Financial Services
Fort Dearborn Income Secs Inc	Financial Services
Current Income Shs Inc	Financial Services
Cigna Investment Securities	Financial Services
Vestaur Securities Fund	Financial Services
Pioneer Interest Shares	Financial Services
Hatteras Income Secs Inc	Financial Services
Insight Select Income Fund	Financial Services
Invesco Bond Fund	Financial Services
Boulder Growth & Income Fund Inc	Financial Services
Transamerica Income Shares Inc	Financial Services
Hancock John Investment Trust	Financial Services
Hancock John Income Securities Trust	Financial Services
Montgomery Street Income Securities	Financial Services
Idacorp Inc	Energy
T E C O Energy Inc	Energy
Adams Express Co	Financial Services
Castle Convertible Fund Inc	Financial Services
Avista Corp	Energy
L G & E Energy Corp	Energy
Circle Income Shares Inc	Financial Services
Merchants New York Bancorp Inc	Financial Services
B C E Inc	Telecommunications
Allegheny Energy Inc	Energy
T N P Enterprises Inc	Energy

Table 2. Statistics related to companies in Reconstructed Portfolio

Sl no.	Company Name	Mean	St. Deviation	Max	Min
1	Cornerstone Total Return Fund Inc	0.010366	0.034443057	0.146789	-0.06897
2	Western Asset Income Fund	0.013363	0.034681428	0.1558442	-0.06476
3	Lincoln National Income Fund inc	0.015124	0.035777942	0.1332766	-0.07258
4	Allmerica Securities Trust	0.012469	0.037833537	0.1230769	-0.08642
5	Hartford Income Shares Fund Inc	0.011204	0.038398218	0.1643077	-0.07333
6	Prospect Street Inc	0.012425	0.038364393	0.1582353	-0.11765
7	Fort Dearborn Income Secs Inc	0.012663	0.039175639	0.1428571	-0.10024
8	Current Income Shs Inc	0.012828	0.039269129	0.1967742	-0.07945
9	Cigna Investment Securities	0.011541	0.039699011	0.1588785	-0.10241
10	Vestaur Securities Fund	0.012485	0.040391246	0.1527778	-0.13043
11	Pioneer Interest Shares	0.011221	0.041116676	0.1354167	-0.07317
12	Hatteras Income Secs Inc	0.011271	0.041680612	0.2021277	-0.125
13	Insight Select Income Fund	0.012959	0.041074257	0.1776577	-0.08584
14	Invesco Bond Fund	0.010383	0.042221113	0.1732284	-0.11043
15	Boulder Growth & Income Fund Inc	0.011344	0.04270438	0.1537931	-0.10769
16	Transamerica Income Shares Inc	0.012478	0.042198761	0.1593443	-0.13572
17	Hancock John Investment Trust	0.012251	0.04357959	0.1544715	-0.10639
18	Hancock John Income Securities Trust	0.011979	0.044283214	0.2483146	-0.11881
19	Montgomery Street Income Securities	0.011013	0.044636588	0.1697479	-0.10222
20	Idacorp Inc	0.016422	0.043725869	0.1434211	-0.10404
21	T E C O Energy Inc	0.017896	0.045172968	0.1339286	-0.104
22	Adams Express Co	0.013541	0.047102234	0.2072956	-0.14205

23	Castle Convertible Fund Inc	0.013543	0.047198338	0.1674312	-0.18408
24	Avista Corp	0.014015	0.047077188	0.1386139	-0.09836
25	L G & E Energy Corp	0.01528	0.047187241	0.2166667	-0.11741
26	Circle Income Shares Inc	0.010995	0.049318739	0.2378667	-0.14787
27	Merchants New York Bancorp Inc	0.017777	0.046341129	0.1781609	-0.20376
28	B C E Inc	0.013671	0.048684715	0.1205442	-0.11618
29	Allegheny Energy Inc	0.017345	0.047551814	0.1677852	-0.09969
30	T N P Enterprises Inc	0.016404	0.048139429	0.1691176	-0.10615

Table 3. Statistics related to companies in Reconstructed Portfolio

Sl no.	Mean	St. Deviation	Min	Max
1	0.017753	0.073790234	-0.25333	0.251163
2	0.017069	0.054534561	-0.11061	0.203008
3	0.020191	0.073916552	-0.18072	0.302817
4	-0.00588	0.155159761	-0.34483	0.56
5	0.013557	0.091240291	-0.2973	0.301947
6	0.014352	0.128064002	-0.36216	0.378531
7	0.01571	0.127860496	-0.35738	0.501362
8	0.019873	0.078819732	-0.23	0.197368
9	0.009212	0.092307726	-0.31317	0.322727
10	0.02241	0.064311442	-0.17625	0.187817
11	0.013399	0.076726924	-0.35349	0.175676
12	0.01206	0.081933224	-0.27442	0.240941
13	0.032155	0.089977492	-0.33962	0.427481
14	0.014031	0.137733282	-0.38955	0.480469
15	0.016589	0.082328993	-0.27027	0.384615
16	0.015611	0.0904395	-0.33333	0.309859
17	0.009579	0.112300552	-0.25161	0.301887
18	0.023448	0.095730237	-0.2967	0.277027
19	0.01713	0.077545479	-0.17829	0.27027
20	0.011891	0.085654209	-0.2	0.625
21	0.009829	0.100457653	-0.27451	0.341772
22	0.01438	0.098469427	-0.32572	0.502416
23	-0.00205	0.152485904	-0.57732	0.45098
24	0.023967	0.166070187	-0.34375	1.162162
25	0.011605	0.089459559	-0.23828	0.337079
26	0.014619	0.141231662	-0.41558	0.379562
27	0.028092	0.080705239	-0.19328	0.25
28	0.019435	0.061504368	-0.1828	0.18552
29	0.012699	0.111605149	-0.36486	0.6875
30	0.011259	0.056205558	-0.15946	0.293204