COMPARISON OF DIMENSIONALITY REDUCTION TECHNIQUES OF SPX500: AN APPROACH TO INDEXING

Nelson Aldana-Martinez

ABSTRACT

Dimensionality reduction techniques are one of the easiest and fastest ways to reach efficiency in data analysis. In this study, the main goal was to find the less number of assets that belong to the Standard and Poor's 500 with which it could be replicated the index behavior. It was started with supervised methodologies such as the high correlation filter between the index and each asset and the variance inflation factor, later, it was used non-supervised techniques as the principal components analysis, principal components analysis with VARIMAX rotation, and sparse principal components analysis. In the last section, it was performed a backtest with a quarter of the data was not used to train any model.

RESUMEN

Las técnicas de reducción de dimensionalidad son tal vez una de las manera más rápidas de lograr eficiencia en el análisis de datos. En el caso de este estudio su uso buscó una forma tal que con el menor número de activos pertenecientes al índice Standard and Poor's 500, se pueda lograr replicar el comportamiento del mismo. Se empezó con metodologías supervisadas como el análisis de la correlación entre cada activo y el índice y el factor de inflación de la varianza, luego se usaron técnicas no supervisadas como el análisis de componentes principales, el analisis de componentes principales con rotación VARIMAX y el analisis de componentes principales con sparse. En la última sección se hizo un backtest con un trimestre de datos no usado en el entrenamiento de los diferentes modelos.

Keywords Data Analysis · Dimensionality Reduction · Principal Components Analysis · Correlation · Passive Investment · Financial Markets · Asset Allocation

1 Introduction

Financial markets generate large amounts of data; it could be secondly, hourly, daily, or the detail level that a study could need. The market that best follows the efficient market theory is the one hosted in the United States, and the index that has the big companies there is the Standard and Poor's 500. It represents USD 40.34 trillion in market capitalization or 47.64% of the total world Gross Domestic Product. And its calculation formula according to Indices (2021) is:

$$Index = \sum_{n=1}^{i} \frac{P_i * Q_i}{Divisor} \tag{1}$$

where:

P_i= Price of the Stock i in the calculation time.

Q_i= Quantity of Stocks i used to calculate the index.

$Divisor$= The level necessary to get the index level in the time t.

Due to the index relevance and after the 2008 financial crisis, the investing strategies migrated to passive ones, they consist in buying each index's component at a time point to replicate the level. Bear in mind this, it would be costly to a small or medium investor, due to the huge amount of stocks. They will have to buy 504 stocks according to the weight they have in the index's portfolio. Therefore, on the starting day June 18th of 2021, investors could not buy the 505 stocks, if they do not hold a USD 1,000,000.

"Investors would be far better off buying and holding an index fund than attempting to buy and sell individual securities or actively managed mutual funds. (...) large trading cost detract substantially from investment returns" Malkiel (2012). Therefore, the principal motivation to explore other ways to approach an indexing strategy is due to the huge amount of money that a small investor will have to spend to buy the index.

"The efficient markets model (...) is the hypothesis that security prices at any point in time "fully reflect" all available information" Fama (1970). Thus, the pricing process given by the interaction among the different stakeholders in the Capital Markets will reflect the value of the financial asset.

Due to the large amount of data generated by the financial markets, it is possible to open possibilities to create other ways to resolve the indexation problem. Bear in mind that the computational ability is today bigger than ever, for that reason, it could be a combination of finance and analytics. Therefore, it could be possible to ask, Do dimensionality reduction techniques allow asset reduction to be indexed to the Standard and Poor's 500?

To answer the business problem, it will explore different dimensionality reduction techniques, (i) in supervised learning, High Correlation Filter and Variance Inflation Factor, and (ii) in unsupervised learning, Principal Components Analysis (PCA), PCA with a VARIMAX rotation and Sparse PCA.

The process will follow ingestion from the financial available data from June 2019 to June 2021 in four different lengths, the first is a data set with 2 years of data since June 18th, 2019 to June 18th, 2021, the second is 1 year of data, beginning on June 18th, 2020, the third is 1 semester of data initiating on December 18th, 2020 and the fourth from March 18th, 2021. Later on, it was performed a backtest with the data from June 18th, 2021 to September 18th, 2021.

Given the above, the main goal will be to compare dimensionality reduction techniques to find which one gets the best tracking error with the least amount of assets to the index to the Standard and Poor's 500. Attached to this, will be three specific goals, which consist in (i) performing a back-test with 2021 3Q's data, to get real conclusions, (ii) finding the optimal number of assets to be purchased to reduce tracking error in the indexing strategy, and (iii) analyze with the different dimensionality reduction techniques and find the least number of assets needed to index. As part of the key success metrics, it is also important to remark that the benchmark will be the Exchanged Traded Funds (ETF) available on market.

It is important to remark that the tracking error was calculated as equation 2, and it was annualized to be comparative always, as the TE$*\sqrt{252}$

$$\text{TE} = \sqrt{\frac{\sum_{i=1}^{n}\left((R_p - R_i) - \overline{(R_p - R_i)}\right)^2}{n}} = \sigma_{R_p - R_i} \tag{2}$$

where R_i is the return of the index and R_p is the return of the portfolio. Thus the tracking error is the standard deviation of the difference between the portfolio returns and the index or benchmark returns. In the case of an indexation strategy, the goal is to have the less possible tracking error.

1.1 Data Description

However, it is important that before performing the study, make a general overview of the data. As main features, it is important to mention: there will be 5 data sets, the back-test one, which is 3 months long (data that was not used to train the models), the 2 years, the 1 year, the 1 semester, and the 1 quarter. This subsection will be exploring the last four.

There are 501 assets on the 1Y data set as they are the ones that have at least 1 year of observations. The Pearson's correlation among the assets and the index ranged from 0.021 to 0.7627, indicating that there are certain data correlated to the index. As it is shown in the figure 1, the correlations are concentrated on 0.455, also demonstrating that there is a positive correlation between the assets and the index.

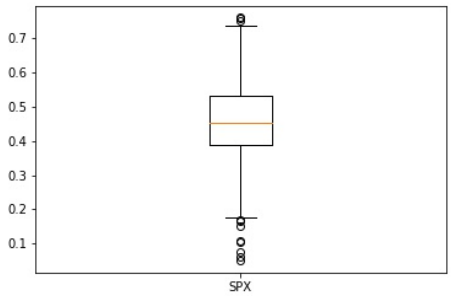

Figure 1: Correlation among the assets and the SPX in 1 Year, data from June 19th 2020 to June 18th 2021

In the 2Y data set are 496 assets, with correlations ranging from 0.054 and 0.869 and the median is 0.66, with a tighter range as it is shown in the figure 2, demonstrating that there is also a positive correlation between the assets and the index in this time frame.

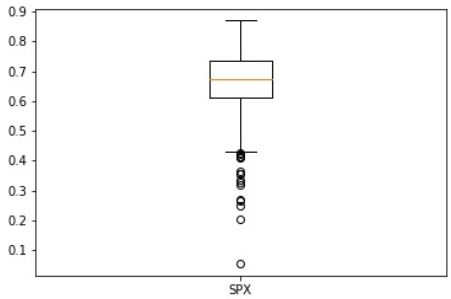

Figure 2: Correlation among the assets and the SPX in two Years, data from June 19th 2019 to June 18th 2021

The semester data set have 503 assets, with correlations ranging from -0.109 and 0.804, with a wider range as it is shown in the figure 3, but it demonstrates also a positive correlation between the assets and the index.

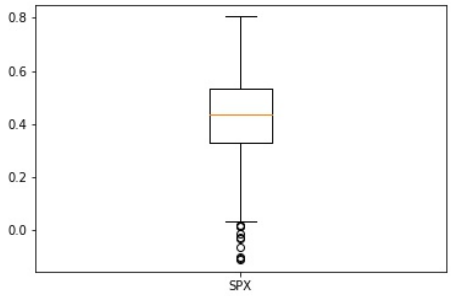

Figure 3: Correlation among the assets and the SPX in one semester, data from December 19th 2020 to June 18th 2021

Finally, the quarter data set have 503 assets, with correlations between -0.153 and 0.810, its range is similar to the correlations in the 1q data as it could be observed in the figure 4. Therefore, it shows a positive correlation between the assets and the index.

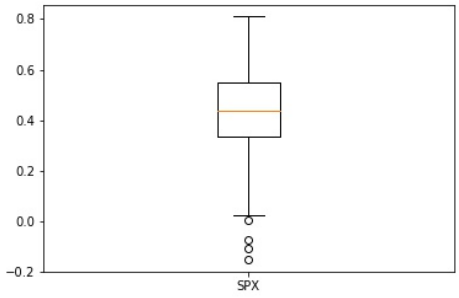

Figure 4: Correlation among the assets and the SPX in one semester, data from March 19th 2021 to June 18th 2021

As there is a correlation across the assets and the index, it guides the study in the right way, giving an incentive to perform it. However, it is important to analyze two other elements the tracking error from the ETFs available in the market and the correlation among the assets, that will be studied in section five.

2 Supervised Learning

As financial data is correlated among the assets part of the SPX, it will be tested in two ways, the first where the correlation between variables is greater than a defined threes-hold and where the variance inflation factor (VIF) is greater than 5.

2.1 High Correlation

Pearson's correlation coefficient takes values between -1 and 1, where 1 indicates a perfect linear positive relationship and -1 a perfect linear negative relationship; 0 indicates not a relationship at all. It is vastly used in the data analytics world to illustrate relationships among variables and as the first approach to data analysis. It is calculated as the division among the covariance between two variables and the product of standard deviation of both. The formula 3 shows the calculation. However, it is possible to state that positive correlations are above 0.5 and negative correlations below -0.5.

$$\rho_{xy} = \frac{S_{xy}}{S_x S_y} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x}) \cdot (y_i - \overline{y})}{\sqrt{\left(\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2\right) \cdot \left(\frac{1}{n}\sum_{i=1}^{n}(y_i - \overline{y})^2\right)}} \quad \text{with} \quad -1 < \rho < 1 \tag{3}$$

where:

S_{xy}: Covariance between x and y; S_x: Standard deviation of x; S_y: Standard deviation of y

Due to the four data sets having different ranges and characteristics, it was performed three sub-datasets to test the performance of the method. The high correlation filter takes into account the correlation above the 3rd quartile, the 90th percentile, and the 95th percentile for each case. Thus, there will be 3 different models in each sub-sub section and will be compared the performance of each and the index.

2.1.1 High Correlation with 2 Years of Data

As the values of the correlation among the assets and the index in two years, time framework has the following results illustrated in the table 1, the breakpoints to take into account in the study were the following Pearson's correlation coefficients: 0.733517, 0.786423 and 0.812659. There were performed three models were and tested with the historical data.

Percentile	Value
min	0.054411
25%	0.610897
50%	0.675347
75%	0.733517
90%	0.786423
95%	0.812659
max	0.869431

Table 1: Descriptive Percentiles of two years data set

After doing the three asset selections, the results give a difference in money of USD 261,530.43 with the basket confirmed with assets above 3rd quartile, USD 367,572.72 with the basket with assets above 90th percentile and USD 155,555.02; as it is illustrated in figure 5. Despite the last one being less than the two before, the tracking errors were 4.78%, 6.80%, and 7.17%.

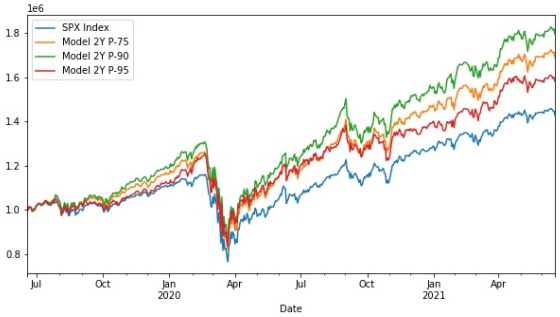

Figure 5: Results of invest USD 1,000,000 in the index and in the other three baskets with two years data

As it could be deduced, more assets guide to less tracking error. However, it was not demonstrated the effectiveness of each model until it will be back-tested with the data from June 18th, 2021 to September 18th, 2021 in section 5.

2.1.2 High Correlation with 1 Year of Data

As it was said before, the following models were calculated with 1 year of data, in table 2 is described the principal percentile to take into account in the study, to the high correlation filter it was the following Pearson's correlation coefficients: 0.532729, 0.625852 and 0.670198. The assets above each of those numbers were the ones that conform to each basket.

Percentile	Value
min	0.050277
25%	0.388260
50%	0.453782
75%	0.532729
90%	0.625852
95%	0.670198
max	0.760994

Table 2: Descriptive Percentiles of one year data set

Even in money, the difference is less than in the previous section, the tracking error with less data is greater. In money, the differences are USD 87,567.84 for the third quartile, USD 124,036.79 for the ninetieth percentile, and USD 141,660 for the ninety-fifth percentile. On the other hand, the tracking errors were: 7.74%, 10.91% and 11.57% respectively. Figure 6 shows the divergence among the models and the index.

Figure 6: Results of invest USD 1,000,000 in the index and in the other three baskets with one year data

2.1.3 High Correlation with 1 Semester of Data

In the 1 semester data set, the descriptive statistics are wider than in the two precedents. The table 3 shows a larger range among the correlations. However, the same method was followed to get the assets check.

Percentile	Value
min	-0.109529
25%	0.331359
50%	0.435980
75%	0.531575
90%	0.616319
95%	0.675882
max	0.803665

Table 3: Descriptive Percentiles of one semester data set

In line with the observed in the last section, even the money differences are tighter, the tracking errors are wider. The figure 7 shows the results of the study. The money gaps are USD 25,025.41, USD 44,627.99, and USD 139,558.14, and the tracking errors are 8.42%, 9.95%, and 12.50% for the third quartile, the ninetieth percentile, and the ninety-fifth percentile.

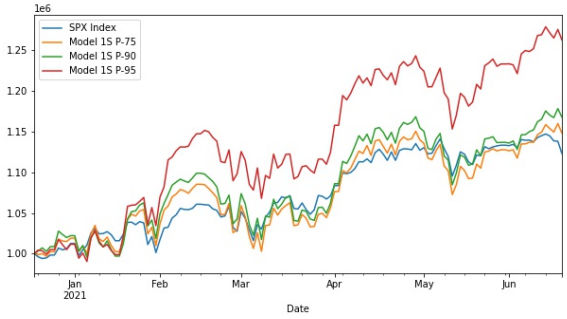

Figure 7: Results of invest USD 1,000,000 in the index and in the other three baskets with one semester data

2.1.4 High Correlation with 1 Quarter of Data

As the last study for this method it was taken into account the data of 1 quarter. The table 4 shows the values of the Pearson's correlation coefficient in the time framework, it is wider than in the others data sets. However, the study was developed equally.

Percentile	Value
min	-0.153522
25%	0.336022
50%	0.438096
75%	0.547976
90%	0.629702
95%	0.705567
max	0.809896

Table 4: Descriptive Percentiles of one quarter data set

As a result of the performed model, the results in money are USD 47,429.30, USD 56,444.91, and USD 63,329.17 and in tracking errors 5.93%, 7.80% and 9.49% for the third quartile, the ninetieth percentile, and the ninety-fifth percentile. Figure 8 shows the performs in historical terms.

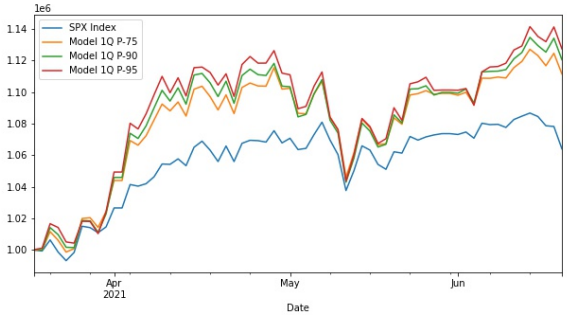

Figure 8: Results of invest USD 1,000,000 in the index and in the other three baskets with one quarter data

To summarize the results obtained in the High Correlation Filter, a greater time framework came with better tracking errors. However, none of the 12 models could be considered as an indexation one due to all of them having tracking errors above 2% as shown in table 2.1.4. Another main conclusion to make a priori is the greater the time framework the difference in money increases. Finally, the number of assets it takes into account the less the tracking error will be.

	Tracking Error				Money Difference			
	2 Years	1 Year	1 Semester	1 Quarter	2 Years	1 Year	1 Semester	1 Quarter
3rd Quartile	4.78%	7.75%	8.43%	5.93%	$ 261,530	$ 87,568	$ 25,025	$ 47,429
90th Percentile	6.81%	10.91%	9.96%	7.80%	$ 367,573	$ 124,037	$ 44,628	$ 56,445
95th Percentile	7.18%	11.58%	12.50%	9.49%	$ 153,555	$ 141,660	$ 139,568	$ 63,329

Table 5: Summary Historical Performance of High Correlation Filters

2.2 Variance Inflation Factor

As the goal of the study is to find a way to index with the less number of assets, sometimes the correlation is not enough as it only shows the collinearity with another variable and not with others. A better way to find multicollinearity is to calculate the variance inflation factor (VIF). As a rule of thumb in literature it could be inferred that a VIF greater than 5 shows a collinearity problem. Equation 4 according to James et al. (2021) shows how it is calculated the coefficient:

$$\text{VIF}(\widehat{\beta}_j) = \frac{1}{1 - R^2_{X_j|X-j}} \tag{4}$$

where $R^2_{X_j|X-j}$ is the R^2 from a regression of X_j onto all the other predictors. If the value of $R^2_{X_j|X-j}$ is close to one, then there is collinearity present, and the VIF will be large.

In this subsection, there will be selected the assets that have a VIF greater than 5. After it, there will be calculated the tracking error among the historical data and the financial index. In section five, it will be developed a back-test with the resulting at least two models.

2.2.1 VIF with 2 Years of Data

As in earlier models, the goal is to do at least four models, however, after performing the selection of features with the set the VIF was infinite to all the assets. In order to solve the problem, it was tested two different data sets, the one with Pearson's correlation coefficient above the second quartile (251 assets) and the one with the third quartile (126 assets). In all cases, the process was to discard the VIF above the second quartile until it is less than 5, where it is considered that the multicollinearity problem is solved.

After obtaining the assets, it was calculated the performance of investing 1 million dollars in the index and in the new basket of assets, to compare historically how the model performed. Figure 9 shows the performance of the baskets and the index. The allocation that started from the second quartile shows a tracking error of 4.75% and a money difference of USD 46,806.26; however, the one begun with the third quartile through a tracking error of 7.59% and a dollar gap of USD 307,807.33.

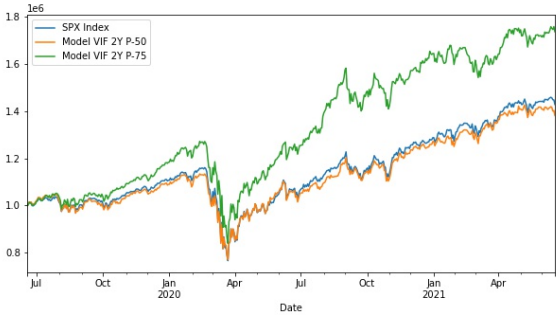

Figure 9: Results of invest USD 1,000,000 in the index and in the assets after drop those with a VIF greater than 5

2.2.2 VIF with 1 Year of Data

As in the earlier method, other time frameworks were tested. In this section, the VIF analysis was performed in the second and third quartile of assets with the highest Pearson's correlation coefficient with the index. After selecting those, the selection was made based on the VIF coefficient.

The results show a mixed conclusion as the VIF starting in the second quartile has a tracking error of 9.26% and a dollar gap of USD 23,548.94. On the other hand, the one beginning with the third quartile results in a tracking error of 5.68% and a difference in money of USD 464.19. Figure 10 shows the behavior of the baskets and the index.

Figure 10: Results of invest USD 1,000,000 in the index and in the assets after drop those with a VIF greater than 5

2.2.3 VIF with 1 Semester of Data

As all cases are different, the 1 semester of data has a higher threes-hold. It is 0.535, where the problem of infinite VIF is resolved, and it is above the third quartile. Bearing this in mind the method for this section was modified to reach the goal. The data sets to be evaluated will be the one above a Pearson's correlation coefficient of 0.535 and the one above the ninetieth percentile.

In line with the previous results, there is no defined tendency in the models' performance, as the results are tracking errors of 10.43% and 10.53% and dollar gaps of USD 26,787.90 and USD 16,693.63. Figure 11 shows the behavior of the baskets and the index.

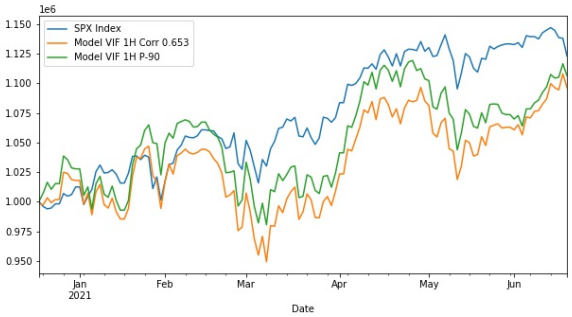

Figure 11: Results of invest USD 1,000,000 in the index and in the assets after drop those with a VIF greater than 5

2.2.4 VIF with 1 Quarter of Data

As all cases are different, the 1 quarter of data has a higher threes-hold. It is 0.62, where the problem of infinite VIF is resolved, and it is above the third quartile. Bearing this in mind the method for this section was modified to reach the goal. The data sets to be evaluated will be the one above a Pearson's correlation coefficient of 0.62 and the one above the ninetieth percentile.

In line with the previous results, there is no defined tendency in the models' performance, as the results are tracking errors of 10.56% and 9.54% and dollar gaps of USD 10,892.23 and USD 19,214.28. Figure 12 shows the behavior of the baskets and the index.

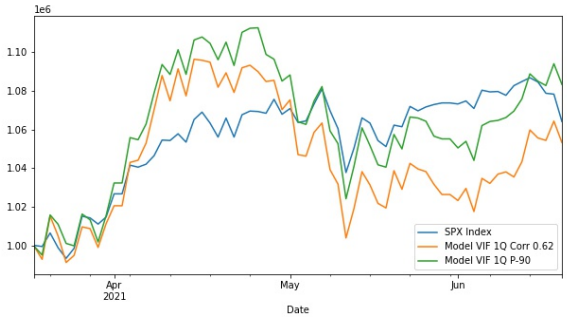

Figure 12: Results of invest USD 1,000,000 in the index and in the assets after drop those with a VIF greater than 5

After doing the study and calculating the models with all the time frameworks, the conclusion is not clear. As there is no relationship between the number of assets and the tracking error or the time framework and the tracking error. Until the model will be tested on the back-test, it will not be possible to conclude.

3 Unsupervised Learning

The main difference across the two great sections in this study is the existence of a "seek" variable, as in the previous one the study search to select based on the Pearson's correlation coefficient and in the case of this section the goal is to find relationship across the variables that aim to find patterns and reduce the number of them that explain the most variance.

3.1 Principal Components Analysis

Given that the High Correlation Filter does not reach the historical tracking error and the VIF does not give a clear conclusion. It was necessary to take another dimensionality reduction technique, Principal Components Analysis (PCA). According to Ruppert and Matteson (2015) "PCA finds structure in the covariance or correlation matrix and used this structure to find low-dimensional subspaces containing most of the variation in the data". The goal in this section will be evaluated with each time framework data set a PCA with 100, 50, 25, and 10 principal components.

PCA is an unsupervised approach because there is no independent variable. It is a form to understand the data, the main goal is to find a low-dimensional representation of data that contains as much as possible of variation. According to James et al. (2021) "PCA seeks a small number of dimensions that are as interesting as possible, where the concept of interesting is measured by the amount that the observations vary along each dimension".

According to Agudelo-Jaramillo et al. (2016) it considers a set of variables $(x_1, x_2, ..., x_n)$ and based on them a new set of variables $(y_1, y_2, ..., y_n)$ is calculated, with the characteristic of being uncorrelated among them and with variance decreasing. Each y_j is a linear combination of the original set of variables $(x_1, x_2, .., x_n)$ as states equation 5.

$$y_j = a_{j1}x_1 + a_{j2}x_2 + ... + a_{jp}x_p = a_j'X \tag{5}$$

where $a_j' = (a_{1j}, a_{2j}, ..., a_{pj})$ is a vector of constants and $X = [x_1, ..., x_p]^T$. The goal is to maximize the variance. And to keep up the transformation orthogonality is required that a_j' module be 1.

$$a_j'A_j = \sum_{k=1}^{p} a_{kj}^2 = 1 \tag{6}$$

The first principal component is calculated by choosing a_1' which maximize the variance of y_1 subjected to $a_1'a_1 = 1$. The second principal component is calculated by an $a_2{}'$ that makes y_2 uncorrelated to y_1. The same procedure is applied to select the rest of the component from y_3 to y_p. The data to process the PCA has to be scaled to correct the differences in variances and means.

3.1.1 PCA 2 Years of Data

The first data set to be examined is the one with two years of data. The main goal is to calculate four models with 100, 50, 25, and 10 components and in each one select the assets that are most important to select the asset allocation.

After do the four asset allocations, the results in Dollar gap are USD 279,375.26, USD 401,044.34, USD 22,864.74 and USD 2,494.96 and the tracking errors of 8.33%, 11.16%, 10.27% and 16.55% with the 100, 50, 25 and 10 components. The figure 13 shows the perform of each basket.

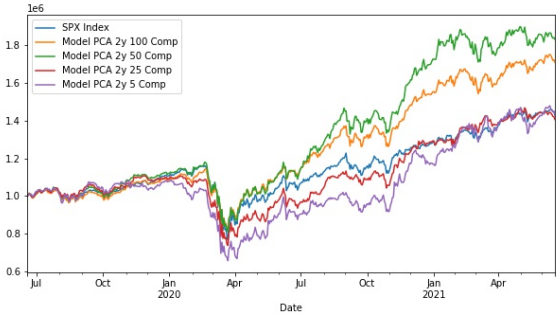

Figure 13: Results of invest USD 1,000,000 in the index and in the four baskets with 100, 50, 25 and 10 components

Unless the do is better than in the earlier models, the historical tracking error is greater than the desired. However, some assets are the most relevant in the different components, thus, there are some assets with two or more proportions in the basket. In order to reach the goal of reduction, it was evaluated without the duplicate assets and the figure 14 shows the new performance with the 71, 44, 25, and 10 unique assets in each model.

The dollar gap is USD 234,070.76 and USD 399,058.41 versus USD 279,375.26 and USD 401,044.34, showing a decrease in both cases and in tracking error the model throw 8.34% and 12.26% versus 8.33% and 11.16%. Although there is not a huge difference between the models in tracking error, the debug of assets allows a reduction in the dollar gap.

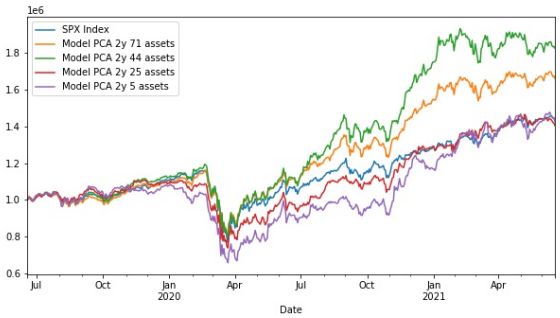

Figure 14: Results of invest USD 1,000,000 in the index and in the four baskets with 74, 44, 25 and 10 assets

3.1.2 PCA 1 Year of Data

After analyzing the 2 years data, the next step is to analyze the 1-year data set. The data were normalized and were broken down into 100, 50, 25, and 10 components.

The exercise gave the following results in tracking error of 15.28%, 6.70%, 9.41% and 11.83% and in dollar gap USD 322,607.63, USD 133,078.13, USD 52,311.08 and USD 7,652.74 respectively. The figure 15 shows the performance. However, as in past models, it is not enough to reach the 5% or less in tracking error.

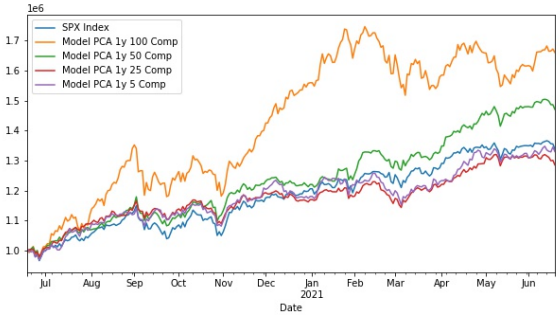

Figure 15: Results of invest USD 1,000,000 in the index and in the four baskets with 100, 50, 25 and 10 components

In the case of the 1 year of data time framework, the reduction in the 100 components model went to 64 assets and the 50 components went to 44 assets.

After doing the new analysis, the figure 16 shows the new performance. Even though in the 2 years of a data model, the tracking error improved with the asset debug, in the case of the 1 year of a data model, it was not the result. As the tracking error went from 15.28% and 6.80% to 14.18% and 6.71%, did not show the same conclusion and did not give enough data to conclude it is a better way to do the study.

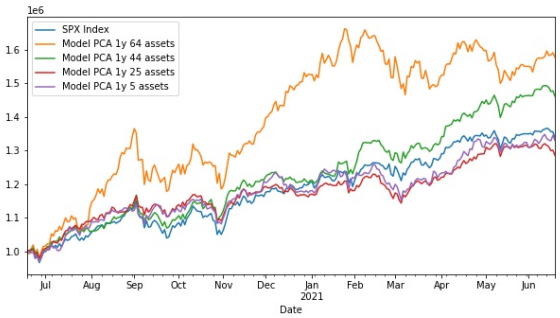

Figure 16: Results of invest USD 1,000,000 in the index and in the four baskets with 64, 44, 25 and 10 assets

3.1.3 PCA 1 Semester of Data

It continues with the analysis of 1 semester of data with 100, 50, 25, and 10 components. After the normalization of the data, the asset that gave more importance to each component was selected to compose the basket.

The 100, 50, 25 and 10 components analysis have results in dollar gap of USD 34,667.28, USD 15,823.10, USD 32,302.33 and USD 89,306.54 and in tracking error of 5.98%, 6.86%, 7.55% and 9.37% respectively. Figure 17 shows the performance of the index and the baskets.

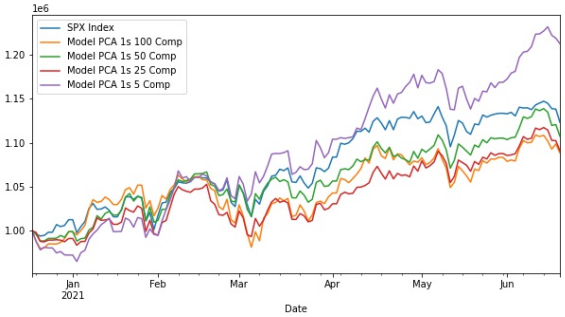

Figure 17: Results of invest USD 1,000,000 in the index and in the four baskets with 100, 50, 25 and 10 components

Even the model of a year of data did not give better results after debugging the basket of assets, it was performed the same process in this data set, throwing new baskets of 86, 48, and 23 assets, the figure 18 shows the performance.

In line with the observed data set of 1 year, this data has a similar conclusion after the debug assets. The tracking error did not improve as they passed from 5.98%, 6.86%, and 7.55% to 6.14%, 5.85%, 7.55%

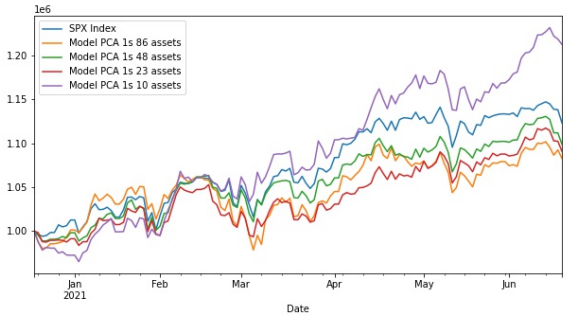

Figure 18: Results of invest USD 1,000,000 in the index and in the four baskets with 86, 48, 23 and 10 assets

3.1.4 PCA 1 Quarter of data

Finally, the analysis of 1 quarter of data has a different length in components due to the number of samples. As this is 64, the greatest number of components to be analyzed is this. Bearing in mind this, the models to analyze were 50, 25, and 10 components.

This data have the following results dollar differences of USD 10,682.74, USD 48,926.42, and USD 75.396,88, and tracking errors of 17.65%, 8.29%, and 13.98%. Figure 19 shows the difference between the index and the asset baskets.

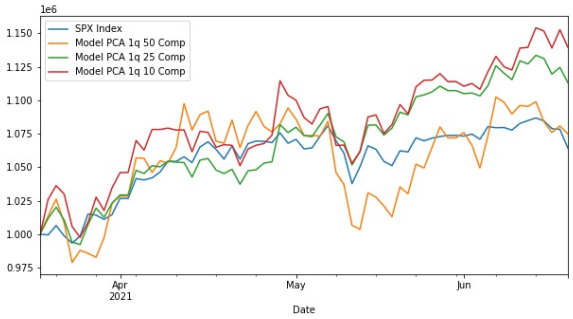

Figure 19: Results of invest USD 1,000,000 in the index and in the four baskets with 50, 25 and 10 components

Finally, in the quarterly data, only the 50 components model has duplicates assets, the number of unique ones are 37. Figure 20 shows the difference. However, the dollar gap is greater after the debugging as it is USD 19,336.88 versus USD 10,682.74 and tracking error goes from 17.65% to 13.81%.

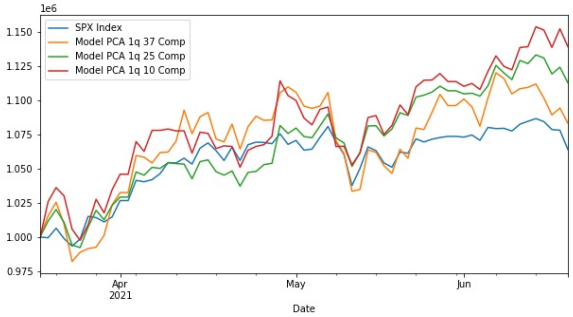

Figure 20: Results of invest USD 1,000,000 in the index and in the four baskets with 37, 25 and 10 assets

Table 6 summarizes the tracking errors of the models which have repeated assets. As the main deductive reasoning conclusion, the largest time framework gave more stable results, however, they are not the best. The shortest time framework has more volatile tracking errors, however, in the 1 semester data, it throws the best results.

Components	2 Years	1 Year	1 Semester	1 Quarter
100	8.33%	15.28%	5.98%	N/A
50	11.16%	6.70%	6.86%	17.65%
25	10.28%	9.41%	7.55%	8.29%
10	16.55%	11.82%	9.37%	13.98%

Table 6: Tracking errors across the different time framework and components

As the number of samples is less, the variance explained is more with fewer components as it is shown in figure 21. However, as the main goal of this study is to do the best way to find an indexing strategy with no known data, the back-test in section 5 will be the real way to conclude.

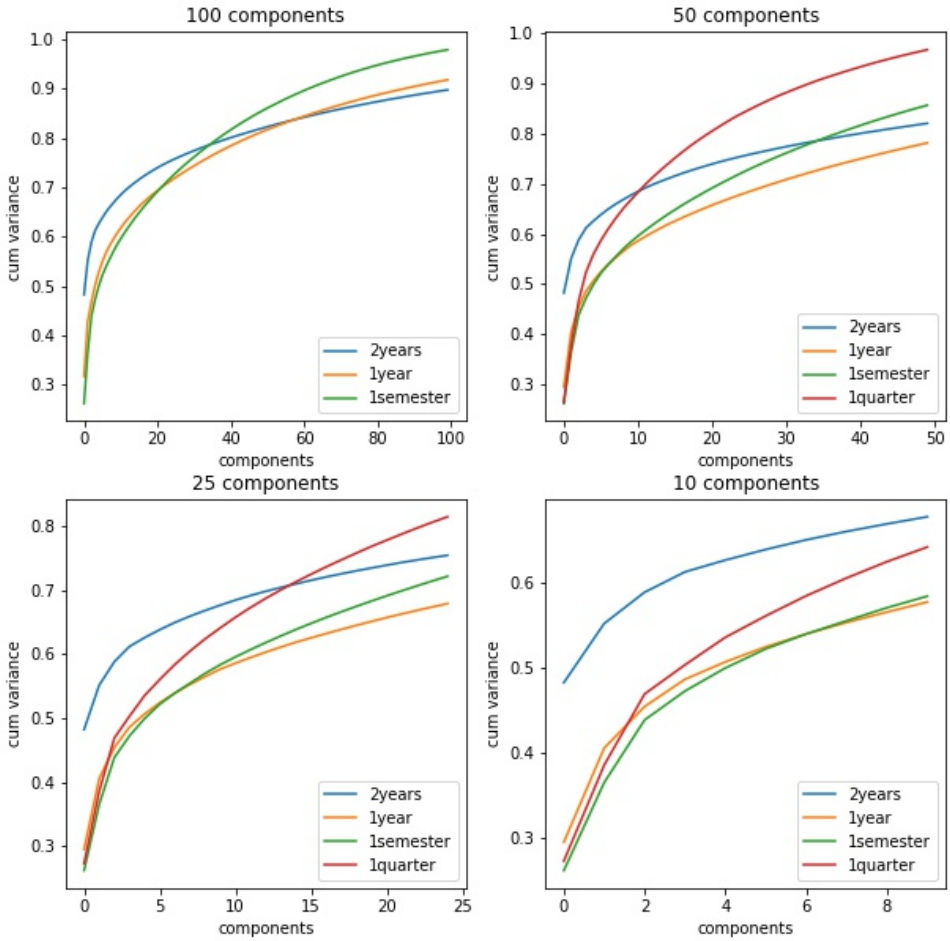

Figure 21: Cumulative variance in each time framework and number of components

Due to the poor results on the historical tracking error, an alternative to PCA was tested. After doing the PCA a VARIMAX rotation was included with the same number of components for each time framework. "Varimax orthogonal rotation tries to maximize the variance of the squared loadings in each factor so that each factor has only a few variables with large loadings and many other variables with low loadings" Lee (2018)

Given said that, the next section analyzes the PCA with a VARIMAX rotation to select the basket of assets in each case. The method is similar to the one used before, with the difference that this time it was not the highest coefficient in the component but the highest coefficient in the rotate one.

3.2 PCA - VARIMAX

"The aim of rotation of the matrix of factor loadings is to facilitate the interpretation so that each factor is associated with a small block of observed variables" Acal et al. (2020). The way to perform the rotation is expected by the following equation stated by Kaiser (1958):

$$R_{\text{VARIMAX}} = \max_{R} \left(\frac{1}{p} \sum_{j=1}^{k} \sum_{i=1}^{p} (\Lambda R)_{ij}^{4} - \sum_{j=1}^{k} \left(\frac{1}{p} \sum_{i=1}^{p} (\Lambda R)_{ij}^{2} \right)^{2} \right) \tag{7}$$

Given said that, the results are presented on the following sections.

3.2.1 PCA - Rotated VARIMAX 2Y of Data

After do a VARIMAX rotation in the models with 2Y of Data, the results throw dollar gap of USD 7,848.66, USD 80,881.17, USD 79,992.30 and USD 183,047.56 and tracking error of 7.13%, 9.33%, 11.93% and 13.64%. The results are show in figure 22.

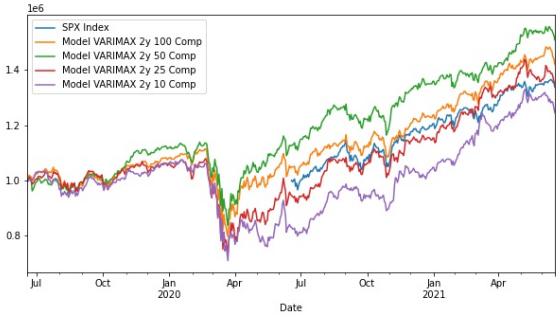

Figure 22: Results of invest USD 1,000,000 in the index and in the four baskets with 50, 25 and 10 components after a VARIMAX rotation

At first, a priori thought it seemed like a VARIMAX rotation loaded better historical results in tracking error terms.

3.2.2 PCA - Rotated VARIMAX 1Y of Data

For this data set, the results of each basket and the index are shown in figure 23. There is a reduction in tracking error across the models compared to the analysis without rotation. The tracking errors were 5.76%, 6.61%, 9.89%, and 5.76%.

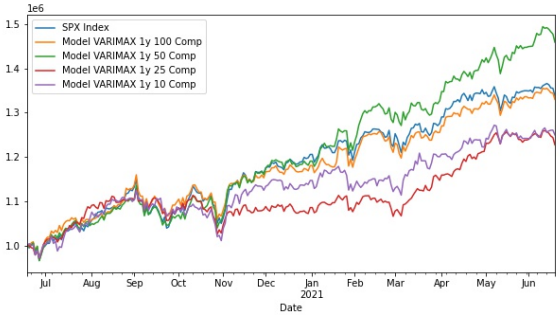

Figure 23: Results of invest USD 1,000,000 in the index and in the four baskets with 50, 25 and 10 components after a VARIMAX rotation

3.2.3 PCA - Rotated VARIMAX 1S of Data

The tracking errors were 8.49%, 6.12%, 7.95% and 11.09%, showing a mixed behaviour with respect to the model without rotation.

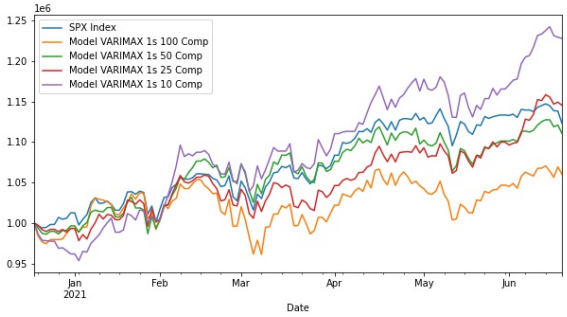

Figure 24: Results of invest USD 1,000,000 in the index and in the four baskets with 50, 25 and 10 components after a VARIMAX rotation

3.2.4 PCA - Rotated VARIMAX 1Q of Data

Finally, the tracking errors of this dataset were 6.35%, 8.10%, and 14.37% as in the previous study, the number of samples is 64 and it has to be the top, as the model requires a matrix multiplication.

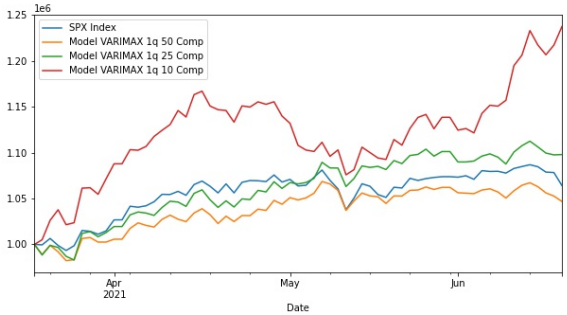

Figure 25: Results of invest USD 1,000,000 in the index and in the four baskets with 50, 25 and 10 components after a VARIMAX rotation

To summarize this section, table 3.2.4 shows that in none of the cases it was reached the goal to have a tracking error below 2%. However, there is an improvement after doing a VARIMAX rotation reaching more stable and lower tracking errors across the different combinations of time frameworks and the number of components.

Method	PCA				PCA - VARIMAX			
Components	2 years	1 year	1 semester	1 quarter	2 years	1 year	1 semester	1 quarter
100	8.33%	15.28%	5.98%	NA	7.13%	5.76%	8.49%	NA
50	11.16%	6.70%	6.86%	17.25%	9.35%	6.61%	6.12%	6.53%
25	10.28%	9.41%	7.55%	8.29%	11.93%	9.89%	7.95%	8.10%
10	16.55%	11.82%	9.37%	13.98%	13.64%	5.76%	11.09%	14.78%

Table 7: Comparative tracking error between PCA and Rotated-PCA across the different combinations of time framework and number of components

3.3 Sparse PCA

The main objective to Sparse PCA is to just take into account the main variables in each component, presenting an advantage above the PCA. It increases the interpretability and the variables selection. It was introduced by Zou et al. (2006) and there he states the following algorithm to perform the study:

1. Let A start at $V[, 1 : k]$, the loadings of the first k ordinary principal components.
2. Given a fixed $A = [\alpha_1, ..., \alpha_k]$, solve the following elastic net problem for $j = 1, 2, ..., k$.

$$\beta_j = \text{argmax}_\beta (\alpha_j - \beta)^T X^T X (\alpha_j - \beta) + \lambda ||\beta||^2 + \lambda_{i,j} ||\beta||_1 \tag{8}$$

3. For a fixed $B = [\beta_1, , \beta_k]$, compute the SVD of $\mathbf{X}^T\mathbf{XB} = \mathbf{UDV}^T$, then update $\mathbf{A} = \mathbf{UDV}^T$
4. Repeat Steps 2–3, until convergence.
5. Normalization: $\tilde{V}_j = \frac{\beta}{||\beta_j||}, j = 1, ..., k$

Given said that, it was performed the study with 100, 50, 25 and 10 components for the four datasets.

3.3.1 Sparse PCA with 2 years of data

After calculate the historical tracking error and dollar gap, the models throw USD 166,370.80, USD 20,911.54, USD 51,256.51 and USD 262,329.57 and 5.77%, 8.43%, 8.61% and 12.53%. It shows better results than the same data set with a PCA or a PCA-VARIMAX.

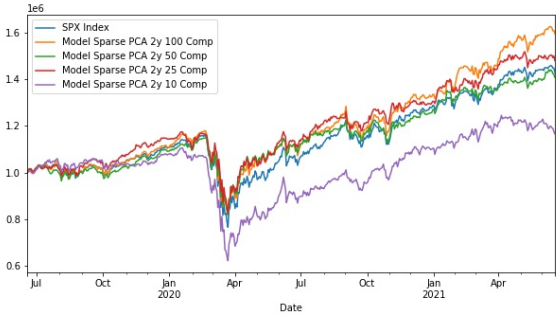

Figure 26: Results of invest USD 1,000,000 in the index and in the four baskets with 100, 50, 25 and 10 components after Sparse PCA

3.3.2 Sparse PCA with 1 year of data

For this dataset, the historical tracking error and dollar gap, the models throw USD 41,498.69, USD 100,967.92, USD 19,089.05 and USD 6,593.52 and 4.90%, 6.66%, 10.07% and 12.79%, in the same way that the last dataset there is an improvement in the results.

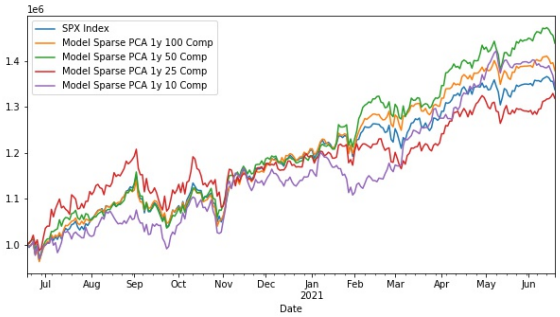

Figure 27: Results of invest USD 1,000,000 in the index and in the four baskets with 100, 50, 25 and 10 components after a Sparse PCA

3.3.3 Sparse PCA with 1 semester of data

In the same line as before the results improved as they throw, 4.27%, 5.49%, 8.60% and 11.38% and USD 26,348.12, USD 11,303.05, USD 16,844.05 and USD 3,002.14.

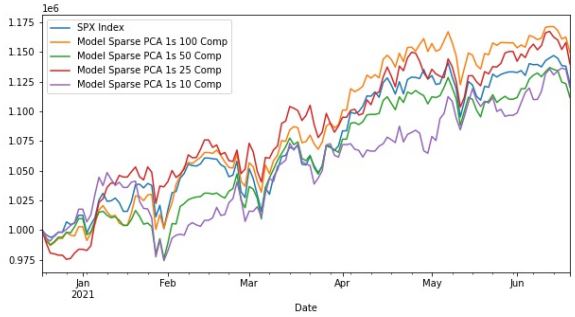

Figure 28: Results of invest USD 1,000,000 in the index and in the four baskets with 100, 50, 25 and 10 components after a Sparse PCA

3.3.4 Sparse PCA with 1 quarter of data

For this methodology it is possible to perform it with a number of components greater than the number of samples. For that reason the model throw results with 100 components. The tracking error continued improving as it was 4.15%, 4.01%, 7.44% and 9.70%. Showing the better results for the different methodologies.

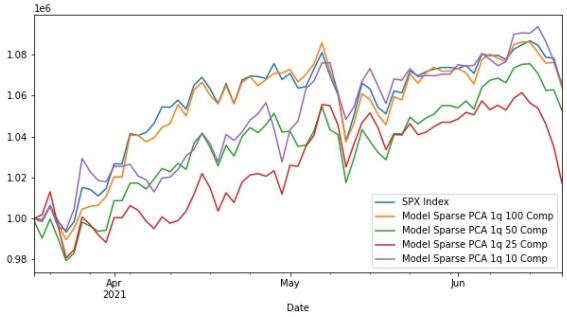

Figure 29: Results of invest USD 1,000,000 in the index and in the four baskets with 100, 50, 25 and 10 components after a Sparse PCA

The results shows a general improvement across all the components and time frameworks. However, there is no reach the 2% where a strategy could be considered as an indexation. Table shows this with the comparative tracking errors. It is important to mention that the Sparse PCA have the better coefficients, however as there is not a clear conclusion due to outliers the comparative backtest will be perform with all the models.

Method	PCA				PCA - VARIMAX				Sparse PCA			
Components	2 years	1 year	1 semester	1 quarter	2 years	1 year	1 semester	1 quarter	2 years	1 year	1 semester	1 quarter
100	8.33%	15.28%	5.98%	NA	7.13%	5.76%	8.49%	NA	5.77%	4.90%	4.27%	4.15%
50	11.16%	6.70%	6.86%	17.25%	9.35%	6.61%	6.12%	6.53%	8.43%	6.66%	5.49%	4.01%
25	10.28%	9.41%	7.55%	8.29%	11.93%	9.89%	7.95%	8.10%	8.61%	10.07%	8.60%	7.44%
10	16.55%	11.82%	9.37%	13.98%	13.64%	5.76%	11.09%	14.78%	12.53%	12.79%	11.38%	9.70%

Table 8: Comparative tracking error among PCA, Rotated-PCA and Sparse PCA across the different combinations of time framework and number of components

4 Comparative Back-test

This section have the main objective to prove the model where the tracking error was less for the objective period from June 18th, 2021 to September 18th, 2021. As it is shown in figure 30 the behaviour to invest in one of the most popular ETFs Spyder and Vanguard fund is to be as good as be on the index.

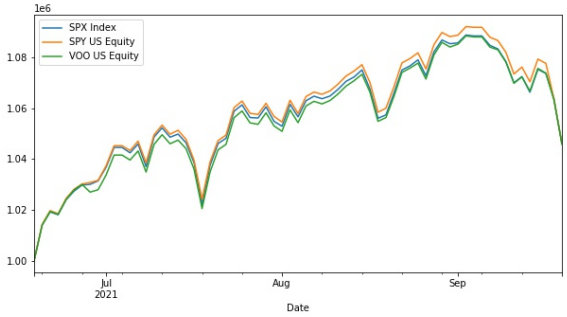

Figure 30: Results of invest USD 1,000,000 in the index and most popular ETF

However, as there are historical behaves above 5% there is interesting to see the results of have the assets in one of those models and observe the tracking errors. In the supervise learning models the figure 31 shows the behaviour of the high correlation filter ones trained with two years of data. Also, it is important to mention that the tracking error is more stable on short periods of time, as it is less possible to have more noise. Table 4 shows that.

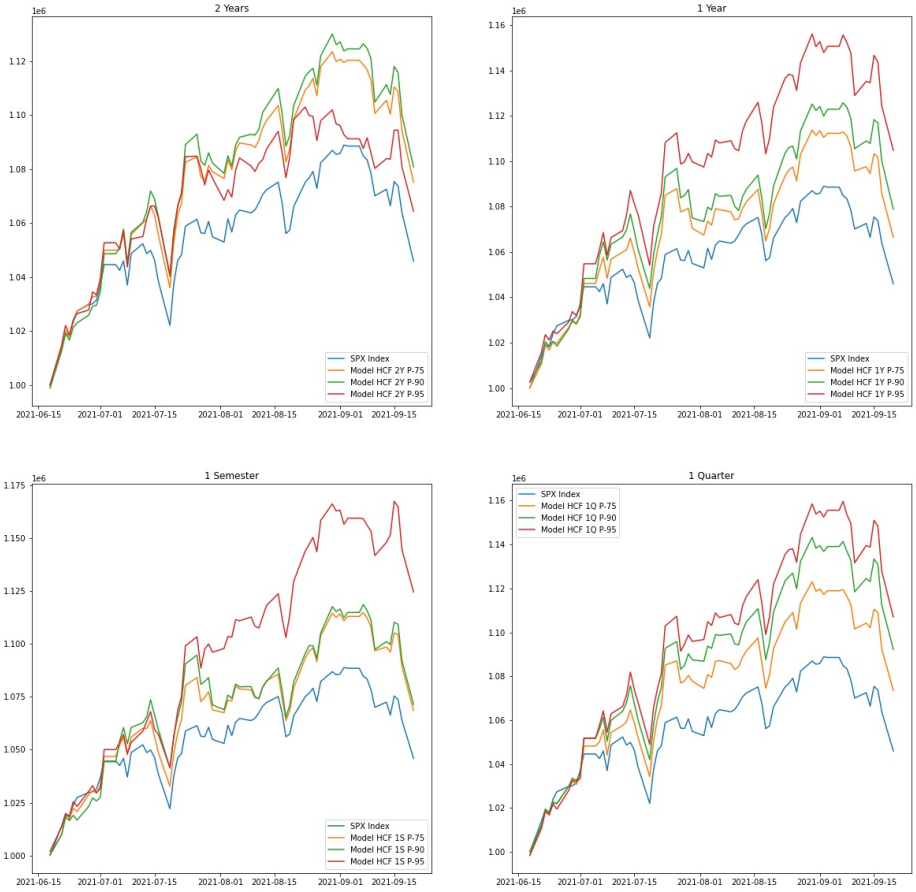

Figure 31: Results of invest USD 1,000,000 in the basket of assets selected by the models with High Correlation Filter

	Historical TE				Backtest TE			
	2 Years	1 Year	1 Semester	1 Quarter	2 Years	1 Year	1 Semester	1 Quarter
P-75	4.78%	7.75%	8.43%	5.93%	2.86%	4.55%	4.68%	4.29%
P-90	6.81%	10.91%	9.96%	7.80%	4.36%	6.21%	6.22%	5.48%
P-95	7.81%	11.58%	12.30%	9.49%	5.09%	6.77%	6.85%	6.86%

Table 9: Tracking error of High Correlation Filter methods

As it is shown in figure 32 the supervise learning methods do not have a great performance none in the historical data or in the backtest.

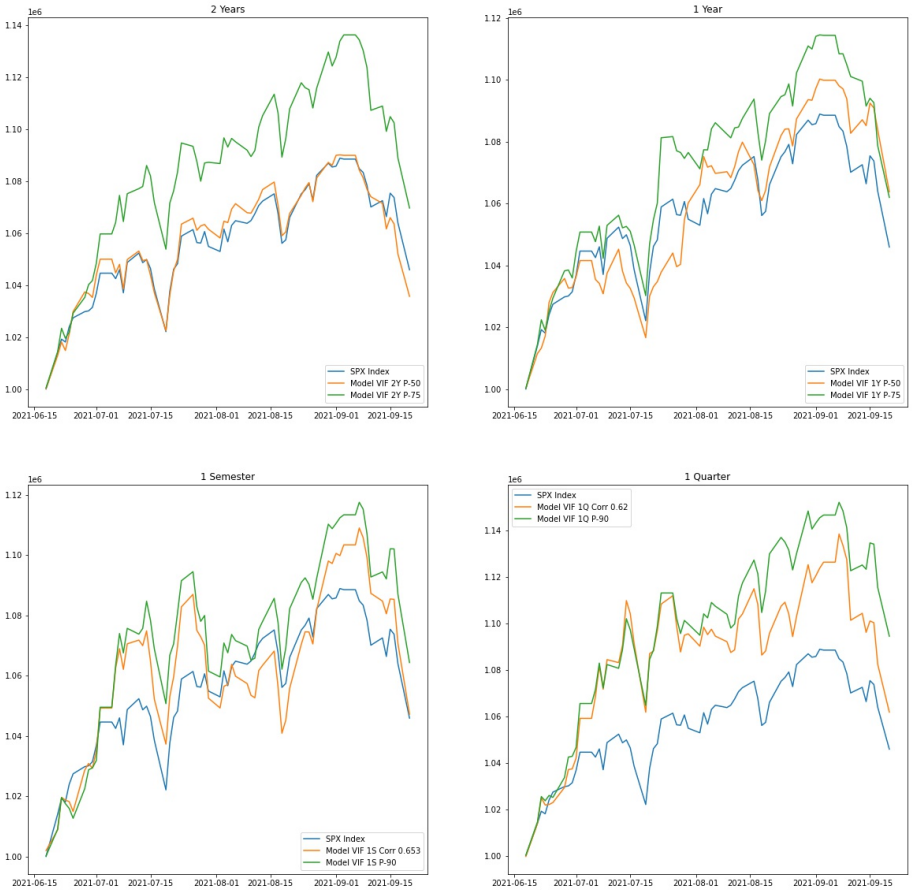

Figure 32: Results of invest USD 1,000,000 in the basket of assets selected by the models with Variance Inflation Filter

However as it was observed on the last model, it is more stable on the short data. For that reason the tracking error is less in the backtest.

	Historical TE				Backtest TE			
	2 Years	1 Year	1 Semester	1 Quarter	2 Years	1 Year	1 Semester	1 Quarter
P-50 / Corr 0.63	4.75%	9.26%	7.54%	9.81%	3.39%	5.89%	7.54%	9.81%
P-75 / P-90	7.59%	5.68%	10.54%	9.54%	5.30%	4.53%	7.63%	7.86%

Table 10: Tracking error of Variance Inflation Factor methods

On the other hand, the unsupervised learning models throw the following results shown in figure 33, the principal components analysis did not have a results as expected with a perfect correlation. However, this go in line with the observed in the historical data.

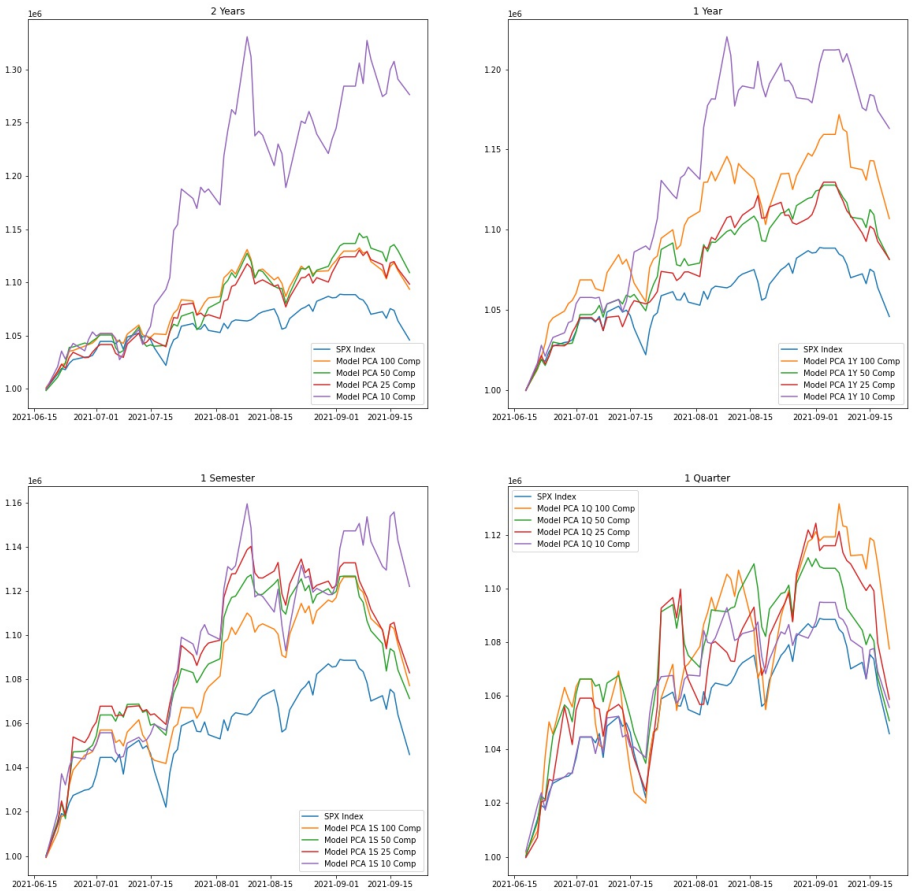

Figure 33: Results of invest USD 1,000,000 in the basket of assets selected by the models with Principal Components Analysis

Instead of the occurred with the supervised learning approach, the PCA gives worse tracking errors numbers with the backtest data.

In line with the happened with PCA, the PCA with rotation VARIMAX did not give better results, there are inconclusive behaviour among the different components and time frameworks.

	Historical TE				Backtest TE			
Components	2 Years	1 Year	1 Semester	1 Quarter	2 Years	1 Year	1 Semester	1 Quarter
100	8.33%	15.28%	5.98%	NA	10.11%	8.85%	6.58%	NA
50	11.16%	6.70%	6.86%	17.25%	11.50%	4.66%	7.30%	12.92%
25	10.28%	9.41%	7.55%	5.29%	9.33%	8.60%	8.60%	7.34%
10	16.55%	11.82%	9.37%	13.98%	28.32%	16.21%	13.72%	12.41%

Table 11: Tracking error of Principal Components Analysis methods

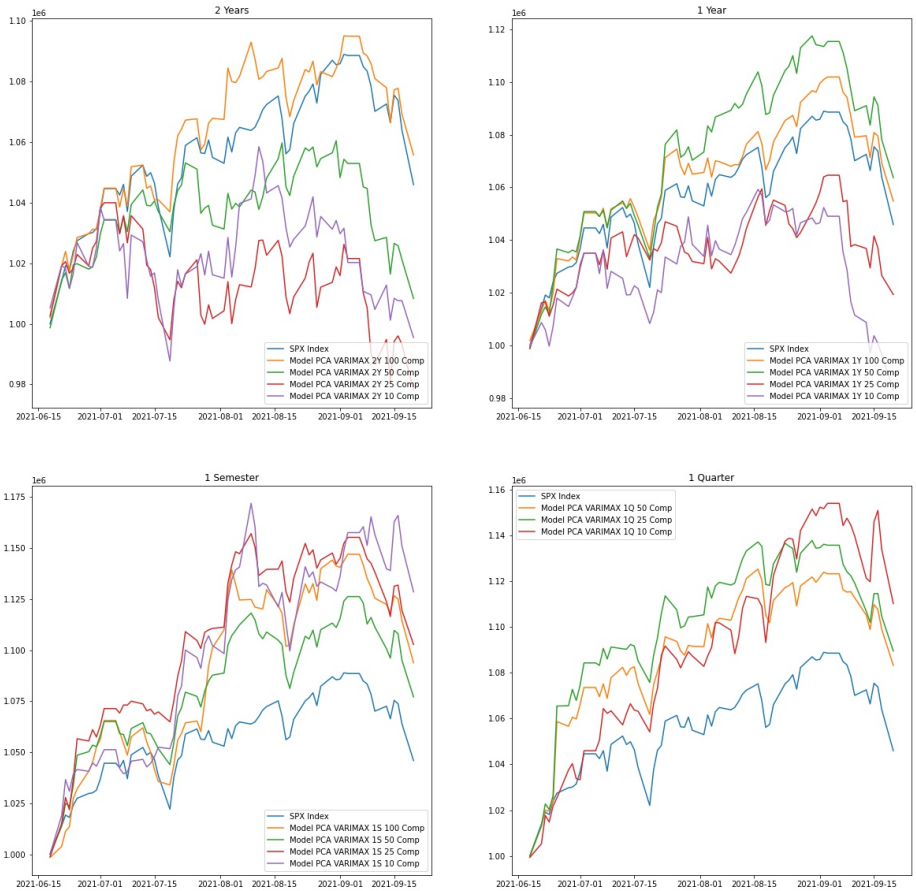

Figure 34: Results of invest USD 1,000,000 in the basket of assets selected by the models with Principal Components Analysis with a VARIMAX rotation

Finally, the sparse PCA throw a result in line with the expected, as the tracking error is less on the shorter data and is better in backtest than in historical.

	Historical TE				Backtest TE			
Components	2 Years	1 Year	1 Semester	1 Quarter	2 Years	1 Year	1 Semester	1 Quarter
100	7.13%	5.76%	8.49%	NA	6.72%	3.85%	9.27%	NA
50	9.35%	6.61%	6.12%	6.53%	6.64%	4.49%	7.16%	7.11%
25	11.93%	9.89%	7.95%	8.10%	7.97%	7.81%	9.41%	8.72%
10	13.64%	5.76%	11.09%	14.78%	9.10%	7.34%	14.35%	9.7%

Table 12: Tracking error of Principal Components Analysis with rotation methods

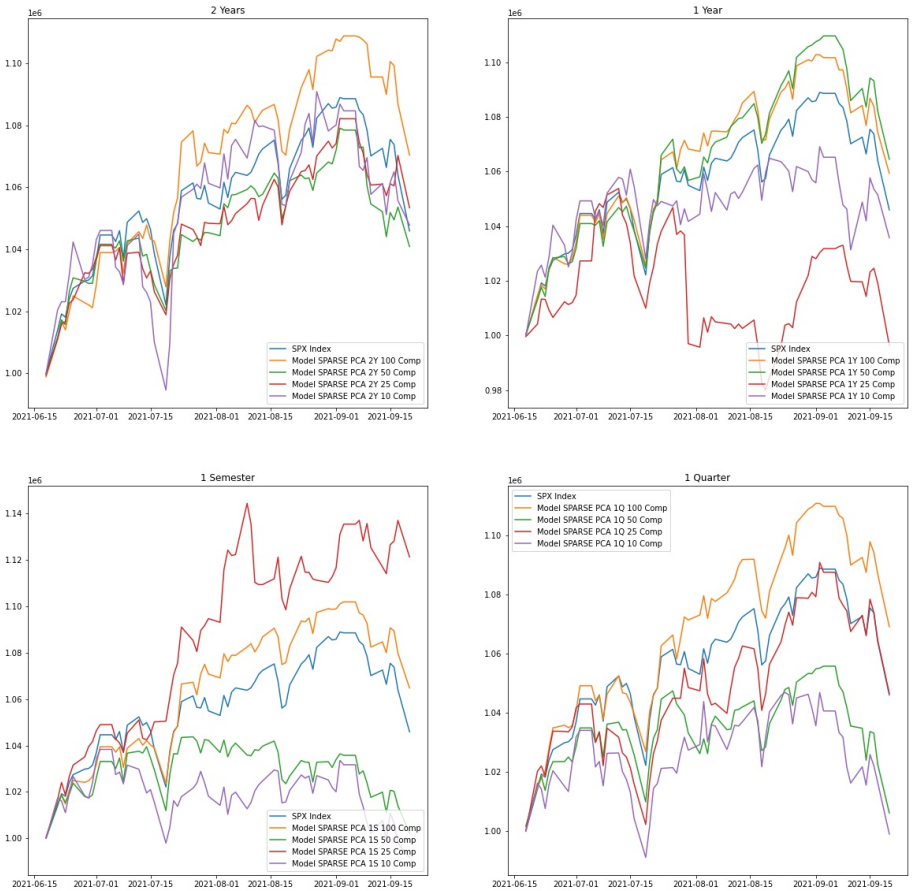

Figure 35: Results of invest USD 1,000,000 in the basket of assets selected by the models with Sparse Principal Components Analysis

	Historical TE				Backtest TE			
Components	2 Years	1 Year	1 Semester	1 Quarter	2 Years	1 Year	1 Semester	1 Quarter
100	5.77%	4.90%	4.27%	4.15%	4.10%	2.75%	3.62%	3.44%
50	8.43%	6.66%	5.49%	4.01%	5.67%	3.25%	3.89%	3.06%
25	8.61%	10.09%	8.60%	7.44%	6.25%	9.89%	11.77%	5.67%
10	12.53%	12.79%	11.38%	9.70%	9.30%	9.14%	7.12%	6.40%

Table 13: Tracking error of Sparse Principal Components Analysis

5 Conclusions

1. The shortest and largest data sets in general give a less tracking error on the historical study. As it is seen on the Sparse PCA, VARIMAX and High Correlation Filter.

2. The supervised learning throw in the both models better results in short periods of time, as when it was evaluated on a three month it collected less noise and have better results.

3. In the unsupervised learning the best model in this study is the Sparse PCA as it takes into account less variables an maximize the variance in less components, putting more 0 and higher scores in other variables.

4. Although the variables could be repeat on some components, it does not give more or less variance.

5. None of the methods tested on this pages could reach the objective of the indexation. However it could be useful in some others studies as a combination with a Black-Litterman model.

References

Acal, C., Aguilera, A. M., and Escabias, M. (2020). New modeling approaches based on varimax rotation of functional principal components. *Mathematics*, 8(11):1–15.

Agudelo-Jaramillo, S., Ochoa-Munoz, M., and Zuluaga-Diaz, F. I. (2016). Principal Component Analysis for Mixed Quantitative and Qualitative Data Proposal Report Research Practise. pages 1–6.

Fama, E. F. (1970). Efficient capital markets: A review of theory and empirical work. *The Journal of Finance*, 25(2):383–417.

Indices, S. D. J. (2021). S&P U.S. Style Indices Methodology. Technical Report November.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2021). *An Introduction to Statistical Learning*. Springer.

Kaiser, H. F. (1958). The varimax criterion for analytic rotation in factor analysis. *Psychometrika*, 23(3):187–200.

Lee, B. (2018). Varimax rotation and thereafter: Tutorial on pca using linear algebra, visualization, and python programming for r and q analysis. *Journal of Research Methodology*, 3:79–130.

Lee, W.-M. (2021). Statistics in Python - Collinearity and Multicollinearity.

Malkiel, B. (2012). *A Random Walk Down Walk Street*. W. W. Norton.

Ruppert, D. and Matteson, D. S. (2015). *Statistics and Data Analysis for Financial Engineering*, volume 2 of *Springer Texts in Statistics*. Springer New York, New York, NY.

Zou, H., Hastie, T., and Tibshirani, R. (2006). Sparse principal component analysis. *Journal of Computational and Graphical Statistics*, 15(2):265–286.