MATH410: Majors Project

Thomas Cole

December 2023

Contents

1 Introduction

This paper is being submitted as part of the requirements for MATH410, Majors Project at McGill University. Thank you to both Dr. Russell Steele and Dr. Mehdi Dagdoug for their guidance as supervisor and co-supervisor of this project.

In this paper we follow loosely follow Chapter 8 from Big Data Science in Finance by Aldridge and Avellaneda [1]. To gain further insights into the applications presented in this chapter we reference the thesis "Modelling Systematic Risk in The Options Market" by Doris Dobi [2]. Finally, we utilize the results and methods for retaining principal components as presented in "How many principal components? stopping rules for determining the number of non-trivial axes revisited" [3]. We specifically chose to further peruse the above papers as they provided a clearer guidance on the methods and results introduced in the original chapter.

We start by providing an overview of the prerequisite knowledge as it pertains to the problem at hand. We then move into the first section which discusses the replication of results as they were provided in the thesis. After the initial results we examine the application of random matrix theory through the use of Tracy-Widom and Marchenko-Pastur as it applies to our problem. In the second section, we perform a simulation study using a variety of methods for retaining principal components . We compare these results to the methods presented in the first section under random matrix theory. Finally, we conclude by providing suggestions for where additional work might be done.

2 Principal Component Analysis of Implied Volatility Surfaces

In this section, we analyze the results produced in the thesis "Modelling Systematic Risk in The Options Market" [2]. We first begin by introducing the necessary prerequisite knowledge and then delve into our results.

2.1 Background Knowledge

Our goal in this section is to perform dimensionality reduction of the Implied Volatility Surface via Principal Component Analysis (PCA).

2.1.1 Implied Volatility Surface

To begin, implied volatility represents the expected volatility of an underlying asset over the lifetime of a given option. It is calculated using the Black-Scholes formula. More specifically, given the market price for an option on a given underlying asset we back solve the equation to get the implied volatility.

We note the following notation: **C** Call option price,**S** Current price of underlying asset, **K** Strike price of the option, **r** Risk-free interest rate, σ = Volatility of Underlying Asset, **t** Time to maturity, **N** Normal CDF.

The Black-Scholes equation is given in terms of the following Partial Differential Equation.

$$\frac{\partial \mathrm{C}}{\partial \mathrm{t}} + \frac{1}{2}\sigma^2 \mathrm{S}^2 \frac{\partial^2 \mathrm{C}}{\partial \mathrm{C}^2} + \mathrm{rS}\frac{\partial \mathrm{C}}{\partial \mathrm{S}} = \mathrm{rC} \tag{1}$$

The solution to this equation, which could be derived by the heat equation, is given by

$$\mathrm{C(S,t)} = \mathrm{N(d_1)S} - \mathrm{N(d_2)Ke}^{-rt} \tag{2}$$

$$\mathrm{d_1} = \frac{1}{\sigma\sqrt{\mathrm{t}}}\left[\ln\left(\frac{S}{K}\right) + t\left(r + \frac{\sigma^2}{2}\right)\right] \tag{3}$$

$$\mathrm{d}_2 = \frac{1}{\sigma\sqrt{\mathrm{t}}}\left[\ln\left(\frac{S}{K}\right) + t\left(r - \frac{\sigma^2}{2}\right)\right] \tag{4}$$

$$N(x) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x} \mathrm{e}^{-\frac{1}{2}z^2} dz \tag{5}$$

This implied volatility can then be calculated across a variety of parameters, which we will discuss in a later section.

Once we have the implied volatilities, we can construct an implied volatility surface, which is parameterized by the **t** the time to maturity and **M** the moneyness, which is defined as $\frac{\mathbf{S}}{\mathbf{K}}$. The moneyness of an option describes how the current price of the option contract is positioned relative to its strike price. A call option is said to be in the money, when the market price is above the strike price, at-the-money when they are equal and out of the money when the market price is below the strike price. The opposite is true for put options. Below is an example of the implied volatility surface for Apple (AAPL) on April 8th, 2010.

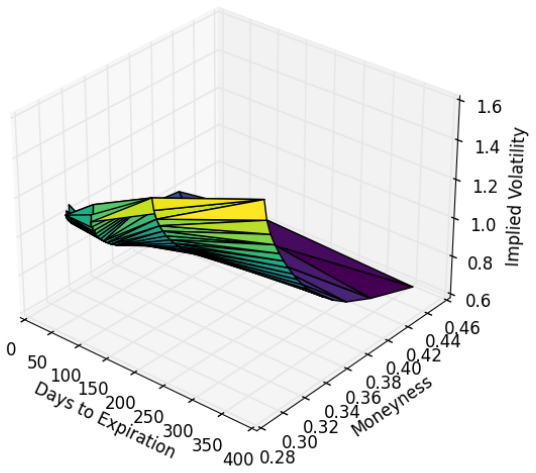


Figure 1: Implied Volatility Surface for AAPL

Lastly, as we mentioned, different implied volatilities can be calculated by varying parameter inputs to our equation. In our application, the two variation of these parameters we make is delta and time to expiration. Delta, commonly referred to as a "greek" in finance, measures the sensitivity of an options price to the change in price of the underling asset. With respect to the mathematical equation given above, delta is the first partial

derivative with respect to S. The second parameter, time to expiration, is fairly straight forward, it is the length of time before the option is set to expire.

2.1.2 Principal Component Analysis

Before moving onto the first application, we provide a brief introduction to Principal Component Analysis, its formulation, solution and assumptions.

Principal Component Analysis (PCA) is a dimensionality reduction procedure for which the goal is to retain as much information from the original data set as possible. PCA does this by constructing new variables which are linear combinations of the original variables. It does this under the following constraints.

- The new variables are linear combinations of the original variables
- The principal components are orthogonal to one another
- They are ordered such that the first accounts for the maximum variation.

More specifically, we want to find principal components (vectors) which maximize the variance of the projected values while minimizing the distance between the original and projected values. It can be shown that these two goals are equivalent.

The formulation is as follows. Given a data set X, we want to find linear combinations $Y_i = w_i^T X$ which maximizes projected variance under the constraint that $w_i^t w_i = 1$.

The solution to this maximization problem can be obtained through La-grange techniques, which is not proved here. We make note of the following values before we move forward as they will be referenced throughout the paper.

- λ_k the eigenvalue of the kth principal component
- Percent of Variance Explained for $\lambda_k = \frac{\lambda_k}{\sum_{i=0}^{n} \lambda_i}$

Assumptions and Limitations:
As with any statistical technique, Principal Component Analysis makes many assumptions which are crucial to understand and make note of when we are analyzing our results. The main assumptions and limitations are as follows:

Assumption 1: The relationships between variables are linear in nature.

Assumption 2: We want orthogonal vectors. These might not always be the best.

Limitation 1: We are using only large variance to search for suitable PCs'

Limitation 2: PCA is scale variant.

We will reference these limitations as we move throughout the next sections and take note of how these assumptions and limitations play into our results.

2.2 PCA Methodologies and Results

Motivation:

Traditional finance literature states that these options volatilities can be characterized by about three factors [1]. One such example can be drawn from the Black-Scholes formula that we described above. As such, the motivation for applying PCA to the implied volatility surface is two fold, first we would like to examine whether traditional finance literature of selecting only three components is sufficient and further we would like to reduce the dimensionality of the data so it can be used elsewhere. The latter half of the thesis also mentions how the results we obtain might be used to classify securities that have more or less systematic vs unsystematic risk.

Now that we have covered the necessary background knowledge on the techniques we are applying, and also given an overview of the data on which we will be applying it we begin by introducing the methods for applying PCA.

2.2.1 Data and Pre-processing

We have two main data sources. We first gather options data, specifically implied volatilities from OptionsMetrics. We also use data from CRSP to get prices on underlying securities. For this analysis, we focus on the constituents of the S&P500. More specifically, within the time frame of August 31, 2004 to August 31, 2013. We estimated the constituents of the S&P500 over this time period by using the constituents on the date of August 31,2013. We gather options data for all options with an underlying asset that is in this set from OptionsMetrics. Importantly, we filter for options with a time to expiry of 30, 91, 182 and 365 days, moving forward we refer to these as days. We also filter for implied volatilities on these options between 20 and 85 delta. Finally, from the CRSP database we obtain prices for all of these underlying assets across the same time period.

2.2.2 Method 1: Non-Coupled

The thesis utilizes a few different ways to construct their data on which they apply PCA. In this section, we examine the first method which we refer to as non-coupled. We will first begin by detailing how to set up our data with an example data set. For explanatory purposes, we will use Apple as we move through the section.

First, we begin with our data set of the implied volatility surface for a specific underlying asset. This is obtained from OptionsMetrics.

Ticker	Date	Days	Delta	Implied Volatility
AAPL	2004-08-31	30	10	0.498111
AAPL	2004-08-31	30	15	0.485301
AAPL	2004-08-31	30	20	0.475574
AAPL	2004-08-31	30	25	0.469565
AAPL	2004-08-31	30	30	0.464794

Next, we transpose this matrix so that we have 52 columns which are the combinations of 13 deltas and 4 days. A sample of this matrix, with 30 delta is given below.

Date	30days30delta	91days30delta	182days30delta	365days30delta
2004-08-31	0.464794	0.428240	0.420453	0.384732
2004-09-01	0.464317	0.427654	0.421653	0.390115
2004-09-02	0.444347	0.423240	0.419102	0.388462
2004-09-03	0.440997	0.415151	0.414734	0.386089
2004-09-07	0.431515	0.423615	0.419042	0.386139

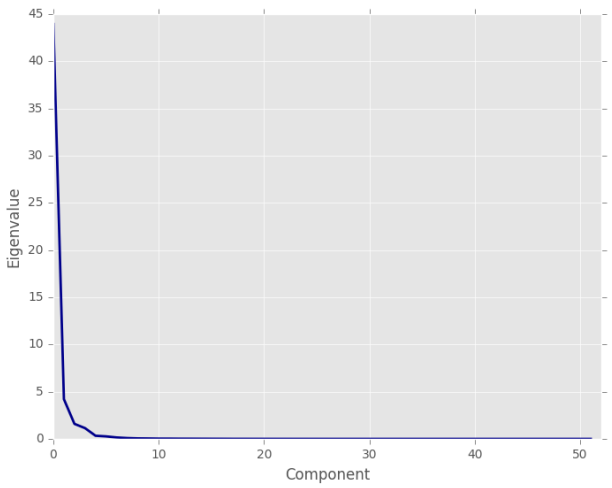
After we have this we calculate row-wise log returns for each column. Our final matrix on which we apply PCA is given below.

Date	30days30delta	91days30delta	182days30delta	365days30delta
2004-08-31	NaN	NaN	NaN	NaN
2004-09-01	-0.001027	-0.001369	0.002850	0.013895
2004-09-02	-0.043962	-0.010375	-0.006068	-0.004246
2004-09-03	-0.007568	-0.019297	-0.010477	-0.006127
2004-09-07	-0.021736	0.020183	0.010334	0.000129

The final dimension of our matrix, when we consider all deltas and maturities is 2266x52. Note that, for certain securities our data set might be smaller than that used in the original study. To account for this, we apply the methods in proportion to our own dimensions. Finally, it is important to reiterate that we repeat this process for each of the individual constituents of the S&P500.

For completeness, before we proceed to the results of this section, we provide the results for Apple.

Component	% Var. Explained
1	0.844714
2	0.080915
3	0.030613
4	0.021951
5	0.006427
6	0.005153
7	0.002938
8	0.001720
9	0.001060
10	0.000870



First, we can see that the first eigenvalue dominates the others in terms of size. Moreover, the first principal component accounts for 84% of the variance within this data set. Naively, if one were to only care about retaining factors that explain more than 80% of our data, then we would be done. That is, we would be able to reduce our original 52 variables to 1 and still retain 84% of the variance within our data set. Though the

result is not entirely intuitive, we will see a similar result for other constituents.

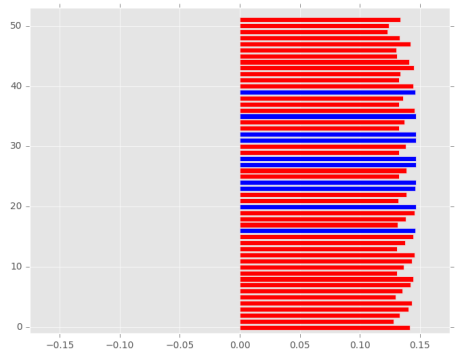


Figure 2: Weights of First Principal Component

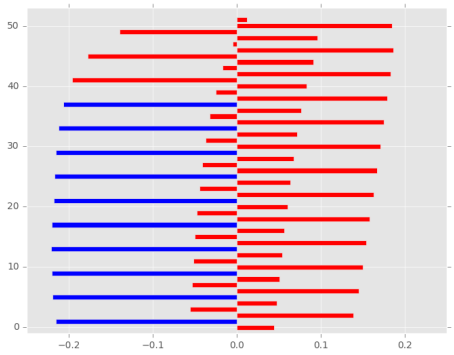


Figure 3: Weights of Second Principal Component

Before we move forward, we examine the weightings on the first and second principal components. We first note that the first principal components weighs positively and nearly equally across all of our factors. That said, those with the highest weights are at 45, 50, 55, 60 delta at 91 and 182 days. As for the second component, we see it contrast the original factors more significantly. Specifically, this component has large negative weights across all deltas for 30 days. It contrasts these factors with all deltas for 365 days. This might have been a fairly intuitive given innate difference between shorter term (30 days) and longer term (365 days) options.

We now move forward to the results of PCA applied to all constituents of the S&P500.

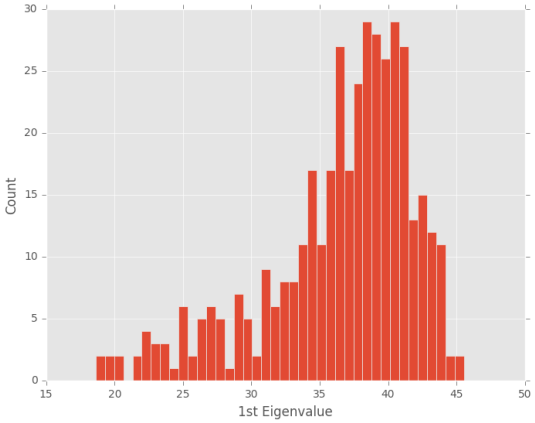


Figure 4: Distribution of Leading Eigenvalue

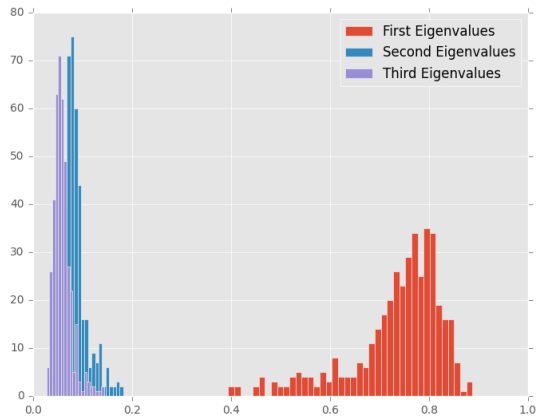


Figure 5: Distribution of % Var. Explained

The graphs provided in Figure 4 and Figure 5 reiterate the same results that we saw

individually when we looked at our example for Apple. First, noting that the sum of our eigenvalues for each matrix should be equal to the number of columns, we can see how heavily dominant the first eigenvalue is across all of our samples. The average first eigenvalue in this sample is 36.5. The drop that we previously saw between the first and following eigenvalues is well visualized in Figure 5.On average the first eigenvalue accounts for 73 percent of the total variance explained. This is in stark contrast to the second and third eigenvalues which on average explain 8 and 5 percent of the total variance. Thus, if we were to look at this naively again we might say that if we wanted only 80% of the variance explained, choosing 1 or even the top 2 components is usually sufficient.

Furthermore, the average percent of variance explained by the top 3 eigenvalues is 88 percent and is also negatively skewed. It is interesting to note the skew of these eigenvalue distributions, nearly all of them are positively skewed, with the first eigenvalue being the exception. The results that we find under this method might suggest that traditional finance literature which suggests to retain 2 or 3 components might be accurate in this sense.

That said, an alternative goal within the thesis is to classify these securities by the amount of systematic risk that they carry. Though we do not delve deeply in to this application, in summary they show that one can classify these constituents using the amount of variance explained by their first principal component. For illustration, we provide in Table 1 and Table 2 the securities on our list which have the highest and lowest leading eigenvalues according to their percentage of variance explained.

Table 1: Highest Leading EV

Ticker	EV1
JPM	0.888106
GS	0.880670
BAC	0.879477
MS	0.863596
ICE	0.858891
XOM	0.858861
NOV	0.858279
MA	0.854355
AXP	0.851582
PSA	0.851488

Table 2: Lowest Leading EV

Ticker	EV1
WEC	0.394646
PNW	0.404667
NLSN	0.409494
ES	0.417082
KMI	0.444843
XEL	0.449996
MKC	0.456654
RSG	0.465517
QEP	0.467710
XYL	0.468343

For reasons mentioned in the thesis, those with high leading eigenvalues might carry more systematic risk. As a result of this, it is fairly intuitive that the securities with the highest leading eigenvalues are also those that are among the largest players in the S&P500.

2.2.3 Method 2: Coupled

We now move on to the second way in which we apply PCA, we will refer to this method as coupled. The only difference here is that we now also include returns on the underlying asset, which we obtain from CRSP.

In reference to step 2 of the Method 1, our data set is now:

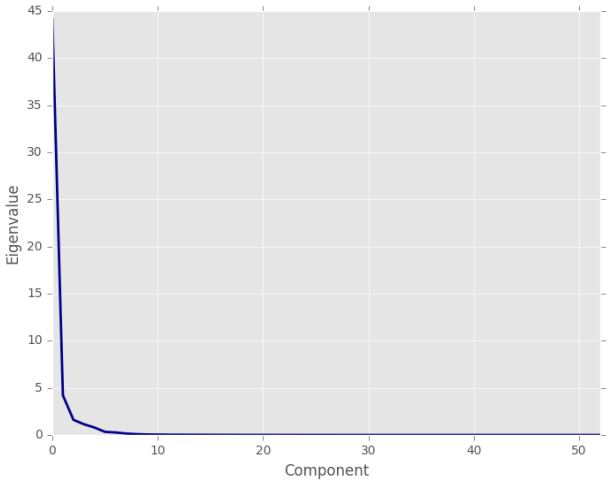
Date	30days30delta	91days30delta	182days30delta	365days30delta	Price
2004-08-31	0.464794	0.428240	0.420453	0.384732	34.490000
2004-09-01	0.464317	0.427654	0.421653	0.390115	35.860000
2004-09-02	0.444347	0.423240	0.419102	0.388462	35.660000
2004-09-03	0.440997	0.415151	0.414734	0.386089	35.230000
2004-09-07	0.431515	0.423615	0.419042	0.386139	35.760000

Finally a sample of the matrix on which we apply PCA is:

Date	30days30delta	91days30delta	182days30delta	365days30delta	Price
2004-08-31	NaN	NaN	NaN	NaN	NaN
2004-09-01	-0.001027	-0.001369	0.002850	0.013895	0.038953
2004-09-02	-0.043962	-0.010375	-0.006068	-0.004246	-0.005593
2004-09-03	-0.007568	-0.019297	-0.010477	-0.006127	-0.012132
2004-09-07	-0.021736	0.020183	0.010334	0.000129	0.014932

The dimension of our complete matrix under this method is now 2266x53. Again, for completeness we provide a sample of the results for Apple under this method.

Component	% Var. Explained
1	0.832461
2	0.079399
3	0.030368
4	0.021539
5	0.014917
6	0.006255
7	0.005039
8	0.002881
9	0.001688
10	0.001039



We note that under this method, not much has changed with respect to the impact of each of our principal components. We find that the leading component explains slightly less of the overall variance, where as most of the others are unchanged. With that said, if we were to naively choose principal components our results would not differ from the previous method. We might again choose only to retain either 1, 2 or 3 components.

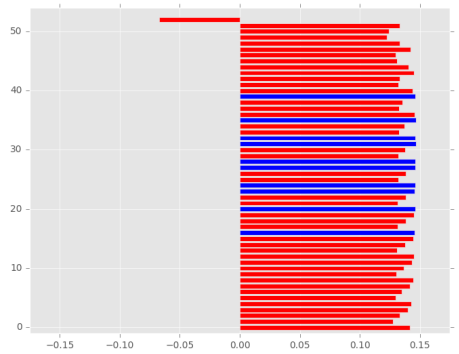


Figure 6: Weights of First Principal Component

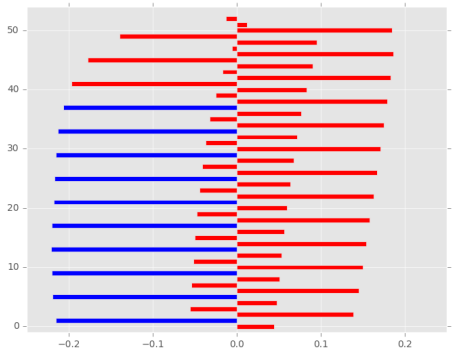


Figure 7: Weights of Second Principal Component

We provide again the two leading principal components for Apple. We find again that the first eigenvector weights most heavily on deltas of 40, 45, 50, 55, 60 for 91 and 182 days. That said we notice one particular change, that is our first principal component contrasts the price of the underlying securities with these weights. As for the second component, we again see a similar large negative weighting across all deltas for 30 days. These are again contrasted with all deltas at 365 days. Our conclusion here, is that largely the composition of the first and second principal component remain relatively unaffected by the addition of this new factor.

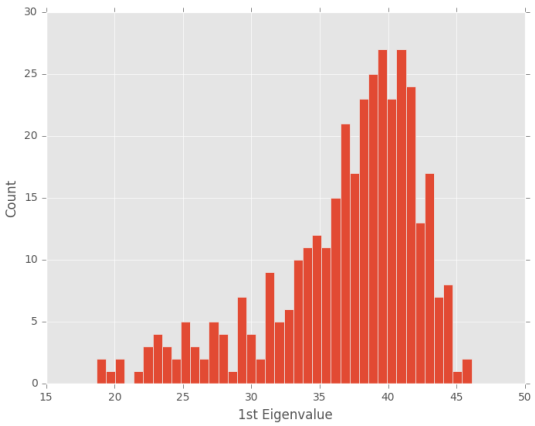


Figure 8: Distribution of Leading Eigenvalue

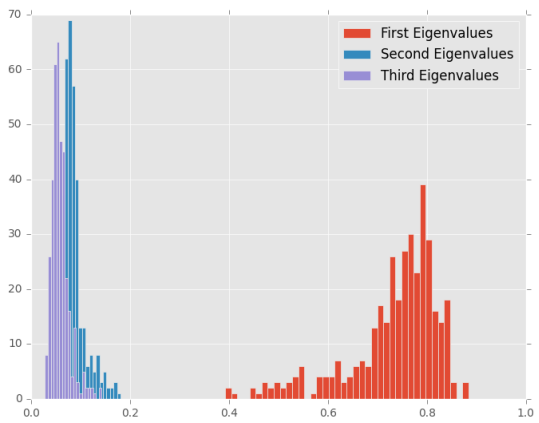


Figure 9: Distribution of % Var. Explained

In Figure 8 and Figure 9 we again provide the similar graphs that we had for the first non-coupled method. Comparing these graphs to the previous method, we see that they are nearly identical. This was apparent within our example of Apple. For comparison

the average of the first eigenvalue under this method is now 33. The average for the second and third eigenvalues are now 3.8 and 2.6 respectively. The positive skew of the other eigenvalues remains the same across this method, while the first eigenvalue remains negatively skewed. Moreover, the insight we get from this method is that it tells us that there is very little extra explanatory power that our additional variable is providing. This is perhaps fairly obvious, as the implied volatility surface of options on the underlying accounts for exactly this- changes in price of the underlying asset. Finally, we provide again the constituents with the highest and lowest leading eigenvalues.

Table 3: Highest Leading EV

Ticker	EV1
JPM	0.883877
GS	0.875682
BAC	0.874023
XOM	0.855339
ICE	0.851265
AXP	0.847453
HIG	0.846647
WFC	0.845386
APA	0.844924
BLK	0.844755

Table 4: Lowest Leading EV

Ticker	EV1
WEC	0.392591
PNW	0.404465
NLSN	0.410093
KMI	0.445132
XEL	0.452784
MKC	0.464288
RSG	0.466964
XYL	0.468762
QEP	0.470567
TSS	0.483616

With respect to the previous top 10 and bottom 10 that we looked at, we see that not much has changed here. We do see a few new additions, but the overall magnitude of these top 10 and bottom 10 leading eigenvalues is roughly unchanged. This provides us with some further stability to the previous analysis we provided across the entire data set.

2.2.4 Method 3.1: Application to Entire S&P500 Implied Volatility Surface

We now move forward from the previous methods and seek to answer a different question, that is, how many principal components can we use to represent the entire market? We will again take two approaches to this. For reference purposes we will refer to this method

as Method 3.1. Our goal here is to apply Principal Component Analysis on the entire S&P500. To do this, under this method, we use the prices for these assets as columns.

To create this data set we gather security prices for all the constituents of the S&P500 from CRSP over the same time period. We now arrange our matrix such that we have the prices for each security in a column. A sample matrix with prices for A, AA, AAPL, ABBV, ABC is given below.

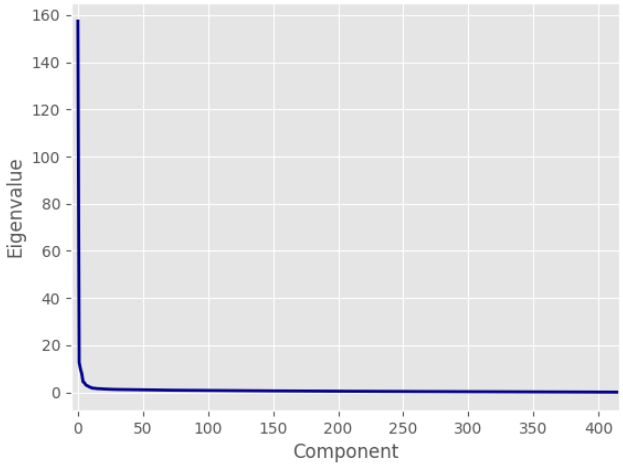
Date	A	AA	AAPL	ABBV	ABC
2004-08-31	20.500000	32.380000	34.490000	NaN	54.100000
2004-09-01	20.690000	32.420000	35.860000	NaN	55.190000
2004-09-02	21.200000	32.720000	35.660000	NaN	56.550000
2004-09-03	20.370000	32.850000	35.230000	NaN	56.270000
2004-09-07	20.580000	33.070000	35.760000	NaN	56.580000

As before, we now compute row-wise log returns on this data set. The final matrix on which we apply PCA is:

Date	A	AA	AAPL	ABBV	ABC
2004-08-31	NaN	NaN	NaN	NaN	NaN
2004-09-01	0.009226	0.001235	0.038953	NaN	0.019948
2004-09-02	0.024351	0.009211	-0.005593	NaN	0.024343
2004-09-03	-0.039938	0.003965	-0.012132	NaN	-0.004964
2004-09-07	0.010257	0.006675	0.014932	NaN	0.005494

Note that our full data set is now 2266x415, as we have 415 constituents for which have data over the entire period. Again, we note that we dropped all columns which contained extra NaN values after computing log-returns. Those such as ABBV in the above example would have been dropped, as we do not have price data for the entire sample. As we mentioned before, since this creates subtle differences between our results and those described in the thesis, we will adapt our results moving forward to fit the dimension of the data that we are working with.

Component	% Var. Explained
1	0.379182
2	0.030567
3	0.023711
4	0.018922
5	0.010627
6	0.009824
7	0.007649
8	0.006722
9	0.006248
10	0.005313



When we apply PCA to the entire S&P500 in this manner we still see a large and dominant first component. That said, it is not quite a significant as those which we saw in the previous methods. Furthermore, we have significantly more components which explain a non-trivial amount of variation within our data set. To contrast, we have below the top significant weightings on the first and second component.

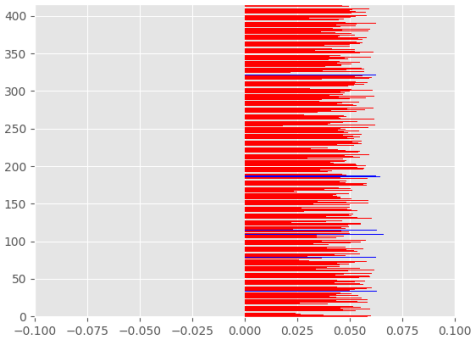


Figure 10: Weights of First Principal Component

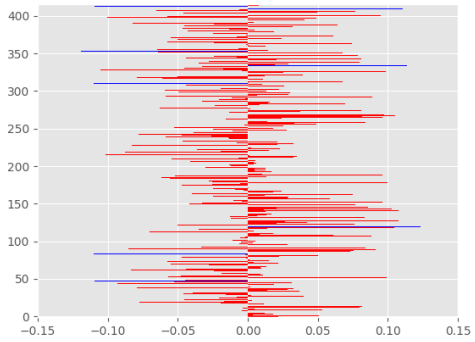


Figure 11: Weights of Second Principal Component

As we saw with the weightings of the previous components, this first component seems to load highly on most variables and thus can be considered to represent the systematic risk within the market. The thesis covers the interpretation and breakdown of these components more thoroughly. In particular, they find that by organizing the data under

this method and controlling for other variables such as grouping these by industry one could classify the amount of systematic risk that might be apparent within an industry by examining its first principal component. We do not go further into this method, but it is an interesting result that relates back to the results we saw under the first two methods when we performed PCA on each of the constituents individually. As a final note, we can see from Figure 11 that the second component draws much more comparison between the individual factors in our data set.

2.2.5 Method 3.2: Application to Entire S&P500 Implied Volatility Surface

We now move on to the final method that we use to apply PCA. This method relates to the previous, in that we are looking to perform PCA on the entire S&P500. As with the other methods, the way in which we do this is slightly different here. In Method 3.1, we only used the prices of the underlying assets in our data set, in this method we would also like to include the implied volatility surface. To do this, the thesis develops an external dimensionality reduction technique known as the pivot method. We will first describe this, as it is central to understanding to how we move forward.

Pivot Method

We provide only a brief summary of the pivot method here as it relates to our application. To begin, in our original data set, we used 52 delta-day combinations. We also found in our original PCA decomposition that the implied volatility surface for each of these assets could be reduced to nearly one principal component. The pivot method seeks to exploit this in a different way. The goal of the pivot method is to determine which combinations of these delta-days can be used to most accurately re-interpolate the entire implied volatility surface by using as little combinations as possible. The thesis does extensive testing on which are most important and their effects on the interpolation. They ultimately find that the 9 pivot model, which uses the 25, 50, and 75 deltas at 30, 182 and 365 days is the most effective. As a result, we use these delta-day combinations in our matrix.

Similar to Method 2, we will now use both the pivot model deltas and the prices to form our data set. More explicitly, for each security in the S&P500 we have 10 columns in total: 9 of these are the delta-day implied volatilities and the last is the price for the underlying asset. Again, we provide an example data set using Apple.

Using only 30 deltas and price for Apple, we get the following:

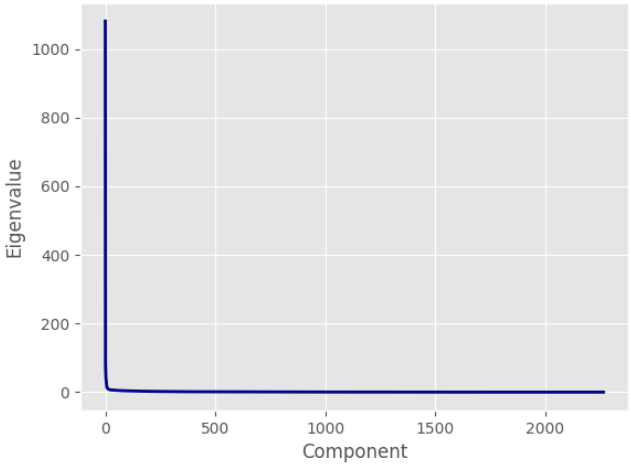
Date	AAPL25delta30days	AAPL25delta182days	AAPL25delta365days	AAPL
2004-08-31	0.469565	0.422278	0.388367	34.490000
2004-09-01	0.470371	0.422594	0.394641	35.860000
2004-09-02	0.443552	0.419531	0.392873	35.660000
2004-09-03	0.444967	0.415459	0.389647	35.230000
2004-09-07	0.431676	0.419090	0.390229	35.760000

Finally, we compute log returns on our data. This gives us the final table on which we will apply PCA.

Date	AAPL25delta30days	AAPL25delta182days	AAPL25delta365days	AAPL
2004-08-31	NaN	NaN	NaN	NaN
2004-09-01	0.001715	0.000748	0.016026	0.038953
2004-09-02	-0.058707	-0.007274	-0.004490	-0.005593
2004-09-03	0.003185	-0.009753	-0.008245	-0.012132
2004-09-07	-0.030325	0.008702	0.001493	0.014932

Our final data set is now 2266x3007. Note that, we drop many of the securities from this large data set as we don't have complete data over the entire time frame. This results from either missing data in the implied volatility surface or missing data in the security prices. Again, to keep results consistent we adapt the results presented in the thesis for our own dimension. Before moving forward, we note that we will have maximum of 2266 principal components under this dimension. This is a result of the independence requirement and assumption of PCA.

Component	% Var. Explained
1	0.359395
2	0.027261
3	0.019004
4	0.013702
5	0.010991
6	0.007629
7	0.006928
8	0.005021
9	0.004878
10	0.003953



It is interesting to note that under this method our leading eigenvalue remains relatively unchanged with respect to the amount of variance explained. This is also the case with the second and third eigenvalue.

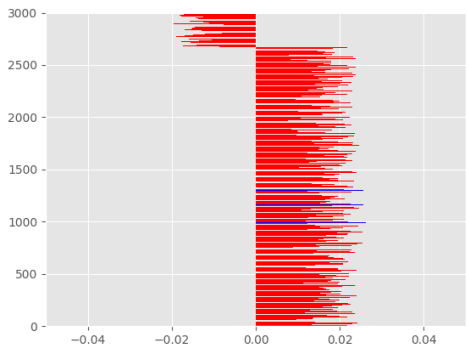


Figure 12: Weights of First Principal Component

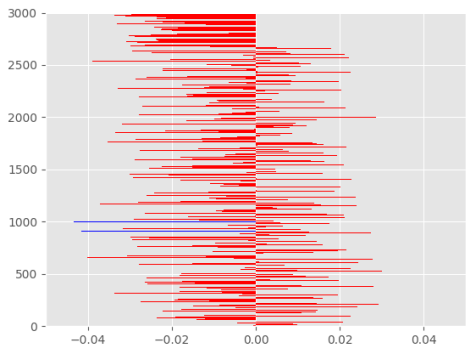


Figure 13: Weights of Second Principal Component

Under this method, we see a change in the construction of the first principal component. Again, it loads highly across most of our original factors, but also contrasts these with other factors. The factors with large negative weights along our first principal component are the stock prices themselves. This might have been foreshadowed by the weightings we saw on the first component under Method 2. For the second component, we see much more variation in the weightings. That said, this component still has large

negative weightings for the factors which represent prices in our original data set.

2.2.6 Conclusion

In this section, we have covered the various methods that were applied to compute PCA on the implied volatility surface, looking at the constituents of the S&P500 both individually and together. This naturally leads us to the next section, whereby we examine how the author of the thesis has chosen which principal components to retain.

2.3 Selection of Principal Components

In natural progression to the previous section on computing principal components, we provide an overview of how the author of the thesis has chosen to retain principal components. In this section, we examine two techniques, Marchenko-Pastur and Tracy-Widom. For each method, we provide an overview of its statement, the assumptions and then finally the results as applied to our data.

2.3.1 Marchenko-Pastur

The first method used in the thesis to determine the number of principal components to retain is the Marchenko-Pastur Law. Its statement is given below.

Statement

The Marchenko-Pastur Law describes the asymptotic behavior of the eigenvalues of large random matrices.

We consider an MxN random matrix X with i.i.d entries with mean 0 and finite variance. Under this law, we describe the limiting distribution of the eigenvalues of the sample covariance matrix given by $Y_N = \frac{1}{N}XX^T$ and the eigenvalues $\{\lambda_1 \dots \lambda_n\}$.

We consider the density of states:

$$\mu_m = \frac{1}{m} \neq \{\lambda_j \in A\}, A \subset R \tag{6}$$

We assume $M, N \to \infty$ such that $\frac{M}{N} \to \Lambda \in (0, \infty)$. Then $\mu_N \to \mu$ in distribution where

$$\mu(A) = \begin{cases} (1 - \frac{1}{\Lambda})1_{0 \in A} + v(A) & \text{if } \Lambda > 1 \\ v(A) & \text{if } 0 \leq \Lambda \leq 1 \end{cases} \tag{7}$$

and

$$dv(x) = \frac{1}{2\pi\sigma^2}\frac{\sqrt{(\lambda_+ - x)(x - \lambda_-)}}{x\Lambda}1_{[\lambda_-,\lambda_+]}dx \tag{8}$$

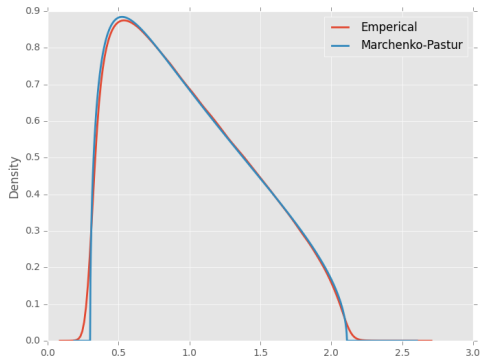
with

$$\lambda_\pm = \sigma^2(1 \pm \sqrt{\Lambda})^2. \tag{9}$$

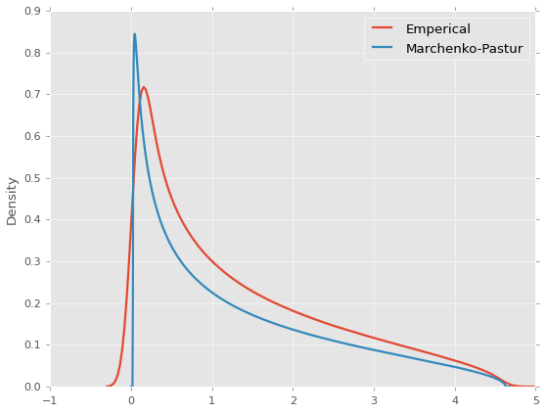
The application of this result for our data is that we would reject all eigenvalues which are below our λ_+ value, and consider them to be not significantly different from noise. The surprising part of this result is that the distribution of our eigenvalues only depends on the dimensions of our matrix.

Application and Results

We first test that the Marchenko Pastur law can be applied to our data before moving forward. We do this by independently permuting the data in our data sets, performing PCA and retaining the largest eigenvalues, 1000 times. We then compose an empirical distribution and compare this to our theoretical distribution.



(a) Method 3.1: Empirical vs Marchenko-Pastur



(b) Method 3.2: Empirical vs Marchenko-Pastur

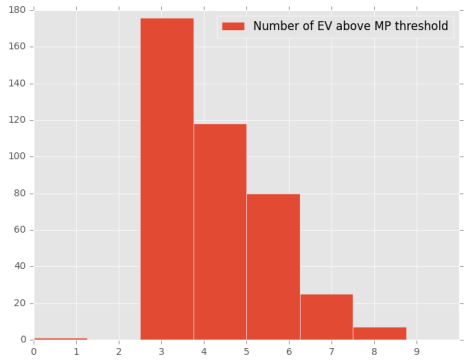
It is clear that our empirical distributions are close to the theoretical distributions guaranteed by Marchenko-Pastur. We perform the Kolmogorov-Smirnov test. Our null hypothesis is that the two distributions are the same. We get results that for both of the above distributions we cannot reject this null hypothesis at the 1 percent significance level. We thus proceed by selecting components using the Marchenko-Pastur Law.

For clarity, our λ_+ value for Apple under Method 1 and Method 2 would be calculated

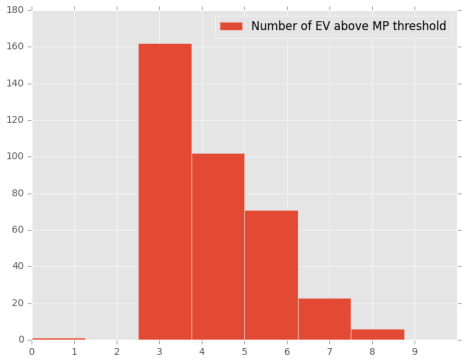
as follows.

$$\lambda_{M1+} = (1 + \sqrt{\frac{52}{2266}})^2 = 1.326 \tag{10}$$

$$\lambda_{M2+} = (1 + \sqrt{\frac{53}{2266}})^2 = 1.329 \tag{11}$$



(c) Method 1: Count of Eigenvalues above MP threshold



(d) Method 2: Count of Eigenvalues above MP threshold

We find that the average number of components above the Marchenko-Pastur threshold is 4.09 and 4.07 for Method 1 and Method 2 respectively. The similarity between these methods was expected under this rule for retaining the number of components, as we saw previously how similar their components were. Again we note that, according to this rule for principal component selection we see the same results that we began with- most are keeping about 3 to 4 components. That said, we do see particular cases in which the number of significant eigenvalues reaches the upper threshold of 8. Intuitively we might assume that if the method retains 8 eigenvalues, it is likely that our leading eigenvalue is small and hence the other components explain more of the remaining variance. Once such example is Ameren Corp (AEE) which, under Method 1, has a leading eigenvalue of 22.2. This would reside on the left tail of our distribution of first eigenvalues in Figure 4.

Under Method 3.1 and Method 3.2 our λ_+ value are:

$$\lambda_{M3.1+} = (1 + \sqrt{\frac{415}{2266}})^2 = 2.039 \tag{12}$$

$$\lambda_{M3.2+} = (1 + \sqrt{\frac{3007}{2266}})^2 = 4.631 \tag{13}$$

By the Marchenko-Pastur Law, under Method 3.1 we would retain 11 eigenvalues, which accounts for 50 percent of the variance explained. As for Method 3.2, we would

retain 68 eigenvalues which accounts for 58 percent of the variance explained. We make note of these values as we will later compare them to the other methods used to retain components.

2.3.2 Tracy-Widom

The second way in which the thesis chooses to retain principal components is through the Tracy-Widom Distribution. We give the statement below.

Statement

Again, the Tracy-Widom distribution describes the distribution of largest eigenvalue of a random correlation matrix. More specifically, Tracy-Widom holds for β-ensembles, that is, Gaussian Ensembles for $\beta = 1, 2, 4$. We use $\beta = 1$, the Gaussian orthogonal ensemble (GOE). GOE model is for random symmetric matrices with independent normally distributed random variables.

The distribution of the largest eigenvalue is given in terms of the Painlevé II differential equation.

$$q''(s) = sq(s) + 2q(s)^3 \tag{14}$$

with boundary condition

$$q(s) \sim Ai(s), s \to \infty \tag{15}$$

where $Ai(s)$ is the Airy Function, not stated here.

The distribution for $\beta = 1$ is then given by

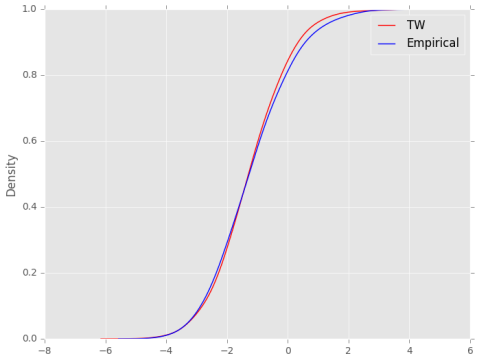
$$F_1(s) = exp(-\frac{1}{2}\int_s^\infty q(x)dx)(F_2(s))^{\frac{1}{2}} \tag{16}$$

$$F_2(s) = exp(-\int_s^\infty (x-s)^2 q(x)^2 dx) \tag{17}$$

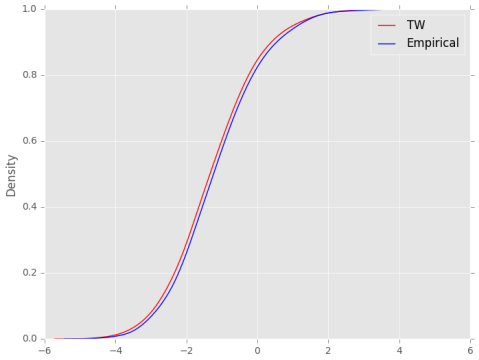
To use this law to retain eigenvalues, we simulate a theoretical distribution and retain eigenvalues which are greater than the maximum of this distribution.

Application and Results

Again, like Marchenko-Pastur we first test that the result can be applied to our data. To do this, we simulate an empirical distribution for our data by permuting our data set, applying PCA and retaining the largest eigenvalues. The theoretical and empirical distributions are given below.



(e) Method 3.1: Empirical vs Tracy-Widom



(f) Method 3.2: Empirical vs Tracy-Widom

It is clear that our empirical distributions are nearly identical to the theoretical Tracy-Widom distributions. To formalize this, we again perform the Kolmogorov-Smirnov test. Our null hypothesis is that the two distributions are the same. We get results that for both of the above distributions we cannot reject this null hypothesis at the 1 percent significance level. We thus proceed by selecting components using the Tracy-Widom distribution.

We find that for Method 3.1, Tracy-Widom retains 11 components which represent 50 percent of the variance explained. Further, for Method 3.2 Tracy-Widom retains 77 components which account for 59 percent of the variance explained.

3 A Simulation Study on Principal Component Retention Techniques

3.1 Introduction

Critical to the application of PCA, as we have mentioned previously, is the requirement of selecting the correct number of principal components to retain. In the latter half of the previous section, we examined two of the methods in which the thesis chose to select the number of principal components. We also saw that both of these methods yielded similar but slightly differing results. In an effort to understand and rationalize the decisions made by the previous theories, in this section we provide extra alternative methods for selecting the number of principal components.

In this section, we examine the paper "How many principal components? stopping rules for determining the number of non-trivial axes revisited" [3], in which they provide a variety and analysis of the different methods that could be used to determine the number of significant principal components to retain. We begin by providing an overview of the methods that we have chosen to apply, we then summarize their results and provide possible explanations for the behaviour of some of these methods.

3.2 Methods

In this section, we describe the methods that we have applied to select the number of significant components to retain. Note that these methods have been chosen by both their ease of application to our problem and relevance to the previously identified methods of choosing principal components. We will apply these methods to both Method 3.1 and Method 3.2 to asses their results.

3.2.1 Parallel Analysis

This first method, which we will refer to as Parallel Analysis, takes a Monte Carlo approach to generating a sample of eigenvalues based on the simulated data sets. For comparison purposes, we generate these sample data sets under the same dimension for both Method 3.1 and Method 3.2. The method proceeds as follows:

1. Simulate N(0,1) i.i.d random variables with given dimension.

2. Perform PCA on the generated matrix

3. Repeat steps 1) and 2) 1000 times.

Once we have all this data, we gather the eigenvalues for each axis and calculate percentile intervals. For example, if we want $\alpha = 0.05$ we calculate the 95% interval and accept/reject principal components based on these bounds.

One of the large underlying assumptions that is being made in this method is that our data follows a normal distribution. As a result, this test can also be seen as evaluating the normality assumption of our data set. We will further discuss this implication in the results section.

3.2.2 Randomization Based on Eigenvalues

Under this method we seek to perform permutation tests on our data set to determine the optimal number of principal component to retain. The steps are as follows:

1. Permute the data in the original matrix

2. Perform PCA on the generated matrix

3. Repeat steps 1) and 2) 1000 times.

For each PCA matrix in the above computation we calculate and record the following measures.

1. λ_k: The observed eigenvalue

2. % variance explained by eigenvalue λ_k

3. $\frac{\lambda_k}{\lambda_{k+1}}$: The ratio between the current and next adjacent eigenvalue

4. $\lambda_k - \lambda_{k+1}$: The difference between the current and next adjacent eigenvalue

After this, we calculate a p-value for a given axis across all four statistics gathered which is the number that exceeded the observed value over 1000. We then accept/reject principal components based on these p-values.

3.2.3 Randomization Based on Eigenvectors

This method employs the same permeation tests as above but now analyzes the eigenvectors weights. The steps are the same as above.

1. Permute the data in the original matrix
2. Perform PCA on the generated matrix
3. Repeat steps 1) and 2) 1000 times.

The difference for this method is only in the calculations that we perform for each of the PCA computations. Under this method, we now compare the absolute loading's of the original data with the absolute loading's of the permuted data. We then say that an eigenvector is significant if at least two of these loading's are greater than the observed. We then again calculate a p-value for each axis as we did in the previous method and retain components based on that.

This second method takes note of an implicit assumption in PCA, that is, at least two of the loading's on each principal component are significant.

3.2.4 Kaiser-Guttman

We include this method of principal component selection as it is the one that is most commonly used. The idea here, is since we are applying our PCA methods to correlation matrices, we should retain all components which have an eigenvalue greater than 1.

3.2.5 Average under Parallel Analysis

This method utilizes the first method that we discussed in this section, Parallel Analysis. Under this method, we take the PCA results obtained from the parallel analysis method, average the resulting eigenvalues by axis and retain all the components which have eigenvalues greater than the average.

3.2.6 Average under Permutation

Much like the above method, we again use the results from the Randomization based on Eigenvalues method. Here, we again take an average of all the eigenvectors we got by axis, and then retain those for which the observed eigenvalue exceeds the average.

3.3 Results

We provide in Table 5 the result from the tests above for Method 3.1 and Method 3.2 and additionally the previous tests we performed.

	Method 3.1	Method 3.2
Parallel Analysis	415	2265
Permutation Eigenvalues	11	110
Permutation Eigenvalues % Var.	11	110
Permutation Eigenvalues Ratio	155	807
Permutation Eigenvalues Diff	34	222
Kaiser Guttman	51	465
Random Avg Parallel Analysis	11	110
Random Avg Perm.	11	110
Permutation Eigenvectors	415	2266
Marchenko-Pastur	11	68
Tracy-Widom	11	77

Table 5: Number of Components Retained by Method and Rule

It is clear above that we see varying performance amongst the number of principal components to retain. Interestingly, for Method 3.1 many of the methods suggest that 11 principal components should be retained. More over, Tracy-Widom and Marchenko-Pastur agree with the other methods of selecting principal components under this construction of our PCA. For Method 3.2 the results are more varied but the general theme is that a large number of principal components should be retained. While Tracy-Widom and Marchenko-Pastur suggest 77 and 68 respectively other methods suggest retaining 110 components. Finally, we note that a few of the tests performed very poorly in our sample. More specifically, the retention based on eigenvectors performed poorly under both methods. This is likely due to the fact that given our eigenvectors are very large, the chances of us having 2 weightings higher than the observed becomes fairly likely. Further, we see that Parallel Analysis performed equally as poorly across our sample. This result was likely expected, as under parallel analysis we are also implicitly testing

the normality assumption of our data. This is where the previously explored methods of Marchenko-Pastur and Tracy-Widom are able to surpass these methods. Finally, we note that Kaiser-Guttman performed fairly poorly on Method 3.2, where our data set is much larger, which is a well known disadvantage of this selection rule.

4 Conclusion and Extension

To conclude, we analyzed the PCA decomposition of the implied volatility surface for the constituents of the S&P500 and as well the overall market. We then performed a simulation study assessing different methods of retaining principal components and contrasting these with the methods originally used from random matrix theory. All code used to generate these results is available upon request.

There are various extensions which could be made to the work that we have done here. The most obvious extension is to further pursue the classification methods that were introduced within the original thesis. This would provide some further insight into how the constituents in the S&P500 are differentiated from a view of their systematic risk. Furthermore, due to computational limitations we were unable to compute these methods across the entire database of equities available on OptionsMetrics. It would be interesting to pursue this further and see how an even larger sample might perform under these methods. In addition to this, it would be interesting to include other firm specific variables in our Method 2, rather than just price of the underlying security such as the volume of trades on a specific security over that period.

References

[1] Irene Aldridge and Marco Avellaneda. *Big Data Science in Finance*. John Wiley amp; Sons, Inc., 2021.

[2] Doris Dobi. *Modeling Systemic Risk in The Options Market*. PhD thesis, New York University, May 2014.

[3] Pedro R. Peres-Neto, Donald A. Jackson, and Keith M. Somers. How many principal components? stopping rules for determining the number of non-trivial axes revisited. *Computational Statistics Data Analysis*, 49(4):974–997, 2005.