# Clustering of Stock Market Data using Square Root Velocity Functions

Aditi Basu Bal
Department of Statistics, Florida State University

**Abstract**

Over the years, analysts have widely contemplated and assessed stock market data to solve different issues pertaining to making informed decisions in the purchasing and selling of stocks. One of the most vital challenges faced in this field is to compute the variability between stock price functions of different products. In this paper, various tools and metrics are studied to precisely measure the distance between functional data. In most instances of real world functional data, there is an issue of alignment which is true for stock market data as well. A layman approach for measuring distance would be to use pointwise $\mathbb{L}^2$ norm in the function space without first adjusting for the lack of alignment. The limitations of this strategy, their solutions and their subsequent limitations are elaborated in this paper. The method of alignment of Square Root Velocity Functions of stock price data overcomes most of the difficulties faced by other techniques. Using this tool to define the metric, a clustering algorithm for functional data is developed. It is found to operate successfully on simulated data. Next, stock price data of some brands from certain sectors of the market are considered and the clustering algorithm is applied on them. The results are compared with the clustering obtained in a layman approach of using Euclidean norm of unaligned functions. Unfortunately, not much improvement in clustering is observed. The possible reasons for such an outcome perhaps being the lack of available information about the underlying variables that control the behavioral biases in this data are discussed.

# 1 Introduction

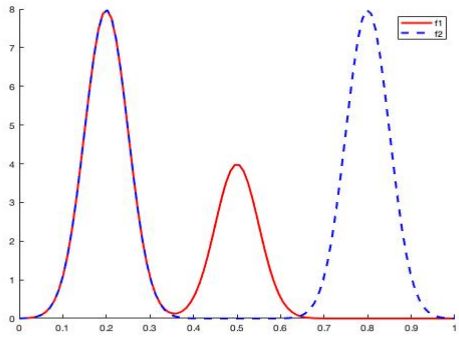
Analysis of stock market data is of great importance to traders as well as investors to arrive at buying and selling decisions. Having an understanding of past information gives one an edge in making educated decisions in stock trade. Researchers have widely studied and evaluated stock market data. It has been observed that a key problem at the base of analysis of this type of data is to measure the distances between the functional data in order to be able to compare stock prices of, say, two equivalent items observed however in two distinctive time ranges or, say, two unique items observed in the same time span. Therefore, in order to proceed to solve any bigger problem in stock market one must ensure that one has the best choice of metric in the right space and the tools to be able to compare data distances meaningfully.
Functional data from stock market are in most cases not aligned and as a result, the usage of a naive $\mathbb{L}^2$ norm to compare distances leads to various problems. This motivated the investigation of the other better alternatives that we have. In section 2.1 various choices of metrics, the challenges they overcome, along with their the limitations, are discussed. Our study is an application of the alignment of Square Root Velocity Functions discussed in section 2.1.4 where a new function called SRVF is generated out of the original function. The group action of warping functions on the SRVF space acts by isometries and is norm-preserving. These characteristics drive out the main challenges that could not be overcome by the previous techniques. We develop an algorithm in section 2.2 that utilizes this tool to perform clustering of functions. Experimental results on simulated data are presented in section 2.3 and that on stock market data are in section 3.

# 2 Proposed Framework

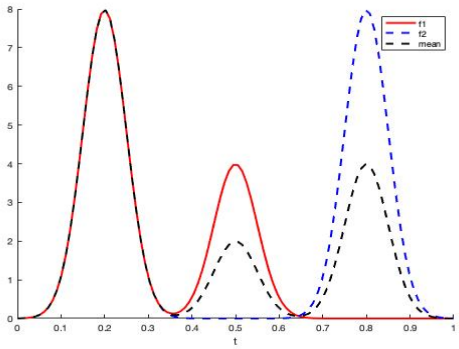
## 2.1 Theoretical Background

### 2.1.1 Euclidean or $\mathbb{L}^2$ metric

Traditionally, in Functional Data Analysis functions have been treated as elements of the $\mathbb{L}^2$ Hilbert space, $\mathbb{L}^2([0,1],\mathbb{R}) = \{f : [0,1] \to \mathbb{R} | \int_0^1 f(t)^2 dt < \infty\}$ and the corresponding $\mathbb{L}^2$ norm has been used to calculate the distance between two functions.
Let $f, g \in \mathbb{L}^2$. The distance between f and g is given by $\|f-g\|^2 = \int_0^1 (f(t)-g(t))^2 dt$.
Functions are infinite dimensional. However, the functional data that we deal with in practice are most often available in the form of a set of discrete observations over a time period. Most of the computational techniques that we use to perform statistical tasks on functional data also produce discrete results. Therefore, the discrete version of the $\mathbb{L}^2$ norm of functional data sampled at n distinct time points should be considered and can be written as $\|f-g\|^2 = \sum_{i=1}^{n} (f(t_i)-g(t_i))^2$.
Using this metric to perform statistical analysis on functional data, however, has some limitations. Under the $\mathbb{L}^2$ norm, the mean of N functions $f_i, i = 1, 2, .., N$ is calculated as $\bar{f}(t) = \frac{1}{N}\sum_{i=1}^{N}(f_i(t))$ i.e., at each t one considers the heights of all the functions and averages them. Averaging functions in this manner is not always sensible. Fig 1a shows two functions $f_1$(red-solid line) and $f_2$(blue-broken line) both having two peaks. The first peaks of both coincide with one another while the second peaks do not. Euclidean mean of such two functions(Fig 1b(black-broken line) generates one that has three peaks. Mean of two double-peaked functions having three peaks makes no sense. One must consider the shapes of the functions and find a way to calculate a mean shape of the functions.




(a) Two double-peaked functions

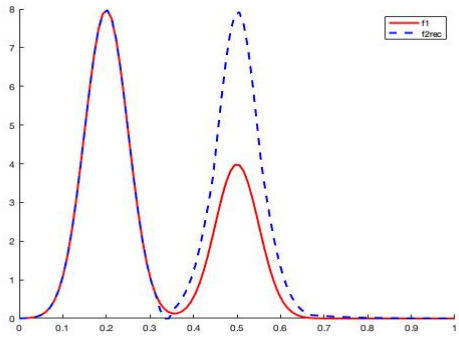



(b) Euclidean mean of two double-peaked functions

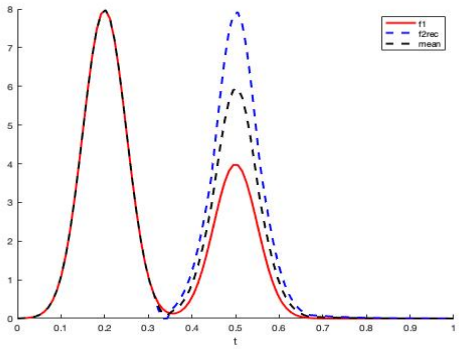
Figure 1: Euclideam mean

### 2.1.2 Alignment of functions using warping

One way to deal with the above-mentioned problem is to separate the phase variability from the amplitude variability between the two functions. We use a warping function $\gamma$ which constitutes the phase variability to align the peaks and valleys of one function to another. This is done by fixing one function and time-warping the other one so that the distance between the two is minimised. Let $f_1$ and $f_2$ be two elements of $\mathcal{F}$=$\{\ f : [0,1] \rightarrow\ \mathbb{R} | f$ is absolutely continuous$\}$.
Then, $\hat{\gamma} = \underset{\gamma \in \Gamma}{argmin} \|f_1 - f_2 \circ \gamma\|_{L^2}$ aligns $f_2$ to $f_1$. Here $\gamma$ is an element of $\Gamma$=$\{\gamma : [0,1] \rightarrow [0,1] | \gamma$ is diffeomorphism, $\gamma(0) = 0, \gamma(1) = 1\}$. $\Gamma$ acts on $\mathcal{F}$ by the following group action:
$\mathcal{F} \times \Gamma \rightarrow \mathcal{F}$
$(f, \gamma) \rightarrow\ \ f \circ \gamma$
1. $((f, \gamma_1), \gamma_2) = (f, \gamma_1 \circ \gamma_2)$
2. $f \circ \gamma_{id} = f$
At this point what remains in the picture is only the amplitude variability. The mean of these aligned functions $f_1$ and $f_2 \circ \gamma$ will preserve the shape(peaks and valleys) of the two functions(Fig 2).




(a) Two double-peaked aligned functions

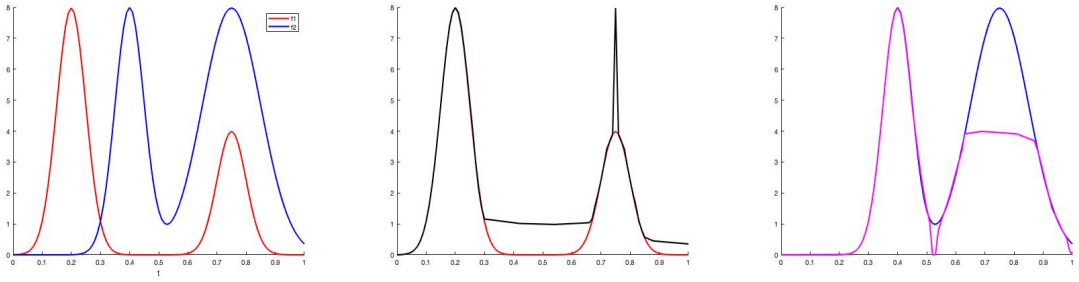



(b) Mean of two double-peaked aligned functions

Figure 2: Mean after Alignment of functions

However, this technique too has some limitations.

- **Pinching Effect** It is the phenomenon that arises when a large part of the function $f_2$ that is being aligned to another function $f_1$ using the warping function $\hat{\gamma}$ gets pinched into a small time interval(Fig 3b), or, alternatively, a small part gets stretched out into a large time interval(Fig 3c).



(a) Two functions (b) Pinching of $f_2$ (c) Stretching of $f_1$

Figure 3: Pinching Effect

- **Lack of Isometry** This property says that the orbits of $f_1$ and $f_2$ are not parallel before and after warping, i.e., $\|f_1 - f_2\| \neq \|f_1 \circ \gamma - f_2 \circ \gamma\|$. The group action does not preserve distance except in the trivial case of $\gamma = \gamma_{id} = \gamma(t) = 1$.

### 2.1.3 Penalization

One solution to these problems is to reduce the warping space by putting a restriction on $\gamma$. An additional regularization term is introduced to the minimization function where a penalty of $\lambda > 0$ takes care of the roughness of $\gamma$. For example, in the following first order penalty setting:
$\hat{\gamma}=\underset{\gamma\in\Gamma}{argmin}(\|f_1 - f_2 \circ \gamma\| + \lambda\mathscr{R}(\gamma))$, $\mathscr{R}(\gamma) = \int_0^1 |\dot{\gamma}(t)^2|dt)$
However, this is not a good solution either as increasing $\lambda$ decreases the importance of the first term or the data term and decreasing $\lambda$ leads to the pinching effect.

### 2.1.4 Alignment of Square Root Velocity Functions (SRVF)

In this technique a new function, the Square Root Velocity Function[1] is created out of the original function in the following manner: For $f \in \mathcal{F}$, where $\mathcal{F}$={ $f : [0,1] \to \mathbb{R} | f$ is absolutely continuous},
SRVF of $f$, $q(t) = sign(\dot{f}(t))\sqrt{|\dot{f}(t)|}$, $\quad q : [0,1] \to \mathbb{R}$.
These SRVFs are square integrable and therefore $q \in \mathbb{L}^2$. A useful fact about the SRVF transformation is that it is invertible. For every $q \in \mathbb{L}^2$, there exists a unique function $f$ that can be obtained back, if $f(0)$ is known, using the equation: $f(t) = f(0) + \int_0^t q(s)|q(s)ds$. Fig 4 shows a sine functions(blue) and its SRVF representation(red).
For a warped function $\tilde{f}(t) = f(\gamma(t))$, one can compute the SRVF using the following steps:
$\dot{\tilde{f}}(t) = \dot{f}(\gamma(t))\dot{\gamma}(t)$
$\tilde{q}(t) = sign(\dot{\tilde{f}}(t))\sqrt{|\dot{\tilde{f}}(t)|} = sign(\dot{f}(\gamma(t))\sqrt{|\dot{f}(\gamma(t))\dot{\gamma}(t)|} = sign(\dot{f}(\gamma(t))\sqrt{|\dot{f}(\gamma(t))|}\sqrt{\dot{\gamma}(t)} = q(\gamma(t))\sqrt{\dot{\gamma}(t)}$,
which is nothing but the right group action of $\Gamma$ on $\mathbb{L}^2$: $\mathbb{L}^2 \times \Gamma \to \mathbb{L}^2$, given by $(q, \gamma) = (q \circ \gamma)\sqrt{\dot{\gamma}}$.
Unlike the group action of $\Gamma$ on the set of functions, $\mathcal{F}$, the group action of $\Gamma$ on the set of SRVFs, $\mathbb{L}^2$ acts by isometries, i,e., it can be shown that $\|q_1 - q_2\| = \|(q_1 \circ \gamma)\sqrt{\dot{\gamma}} - (q_2 \circ \gamma)\sqrt{\dot{\gamma}}\|$. Therefore, $\|q\| = \|(q, \gamma)\|$, i.e., this group action is a norm-preserving transformation. Thus we also get rid of

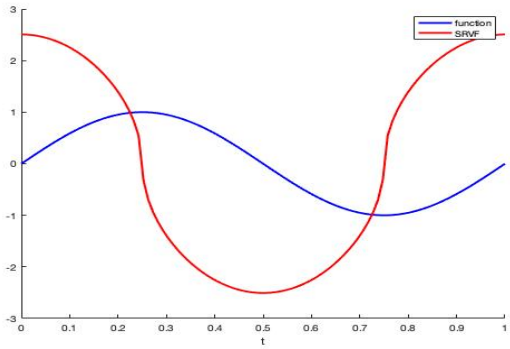
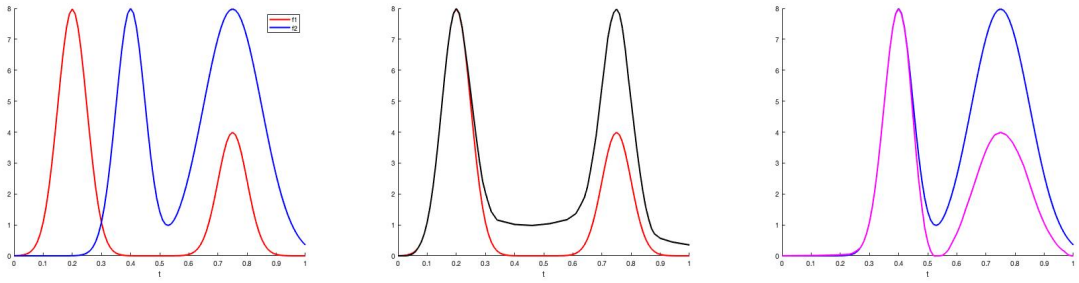



Figure 4: Function and its SRVF

the problem of pinching effect(Fig 5).

In this paper the calculation of distances between functions is done in this SRVF space. Functions $f_1$ and $f_2$ are first converted into their SRVFs $q_1$ and $q_2$. Then one of the SRVFs, $q_1$ is fixed and the other, $q_2$ is acted upon by the warping function $\gamma^*$, the solution to the minimization problem, $\gamma^* = \underset{\gamma \in \Gamma}{argmin} \|q_1 - (q_2 \circ \gamma)\sqrt{\dot{\gamma}}\|^2$. Now that we are in the SRVF space which is just the $\mathbb{L}^2$ space, the distance between $q_1$ and $(q_2, \gamma^*)$ is simply the $\mathbb{L}^2$ norm of their difference. However, in practice, the steps that are followed are not exactly as was just described. The main reason for this is that the calculation of the group action of $\Gamma$ on $\mathbb{L}^2$ is somewhat complicated. Instead, we take a roundabout route to reach the same result. After solving for $\gamma^*$ in the SRVF space we use it to act on $f_2$ thus getting $\tilde{f}_2 = (f_2 \circ \gamma^*)$. $f_1$ and $\tilde{f}_2$ are now aligned(Fig 5). One cannot



(a) Two functions (b) Alignment of $f_2$ to $f_1$ (c) Alignment of $f_1$ to $f_2$

Figure 5: Alignment of functions using SRVFs

just calculate the norm of these two to find the distance between them because of the limitations discussed in section 2.1.2. So now we convert $f_1$ and $\tilde{f}_2$ back to the SRVF space to get $q_1$ and $\tilde{q_2}$. Here, the distance between $q_1$ and $\tilde{q_2}$ is simply the $\mathbb{L}^2$ norm of their difference. Finally, we have the amplitude distance between functions that we use in this paper to perform clustering.

## 2.2 Algorithmic Details

Given a set of n functions $f_1, f_2, ..., f_n$, the goal is to solve a clustering problem that would put the functions that are closer to one another in terms of distance in the same group. The functions must be absolutely continuous. To reiterate, the key challenge here is to align the functions and

calculate the distance between them in the right space for the clustering to be meaningful. Once we have the distances between each pair of functions any generic clustering technique can be applied to reach our final goal. A Dynamic Programming routine DP(h,g), where the arguments h and g are functions in the SRVF space, that solves the minimization problem $\gamma^* = \underset{\gamma\in\Gamma}{argmin}\|h - (g\circ\gamma)\sqrt{\dot{\gamma}}\|^2$ is used here. At a fixed h, DP(h,g) searches over the orbit of g to find the $\gamma$ that minimizes the square of the distance between h and $(g\circ\gamma)\sqrt{\dot{\gamma}}$. This algorithm is described in [1].

Step 1. Convert the functions $f_1, f_2, ..., f_n$ into their SRVFs $q_1, q_2, ..., q_n$ according to $q(t) = sign(\dot{f}(t))\sqrt{|\dot{f}(t)|}$

Step 2. Initialize *count* to 1.

Step 3. For i=1 to n and j=(i+1) to n, repeat steps a to f

a. Find $\gamma^*_{ij}$ by solving the minimization problem, $\gamma^*_{ij} = \underset{\gamma\in\Gamma}{argmin}\|q_i - (q_j\circ\gamma)\sqrt{\dot{\gamma}}\|^2$. This is solved using DP$(q_i, q_j)$.

b. Compute $\tilde{f}_j = (f_j\circ\gamma^*_{ij})$

c. Compute $\tilde{q_j} = \text{SRVF}(\tilde{f}_j) = sign(\dot{\tilde{f}}_j)\sqrt{|\dot{\tilde{f}}_j|}$

d. Compute $AmpDist_{i,j} = \|q_i - \tilde{q_j}\|_{\mathbb{L}^2}$, where $AmpDist_{i,j}$ is the amplitude distance between $q_i$ and $\tilde{q_j}$.

e. Let $Y_{count}$=$AmpDist_{i,j}$

f. Update $count \mapsto count + 1$

Step 4. Use the elements in $Y$ to generate a hierarchical cluster tree, Z, using functions like $linkage(Y)$ in MATLAB.

Step 5. Use Z to display a dendrogram plot of the clustering using functions like $dendrogram(Z)$ in MATLAB.

### 2.3 Illustrative Examples

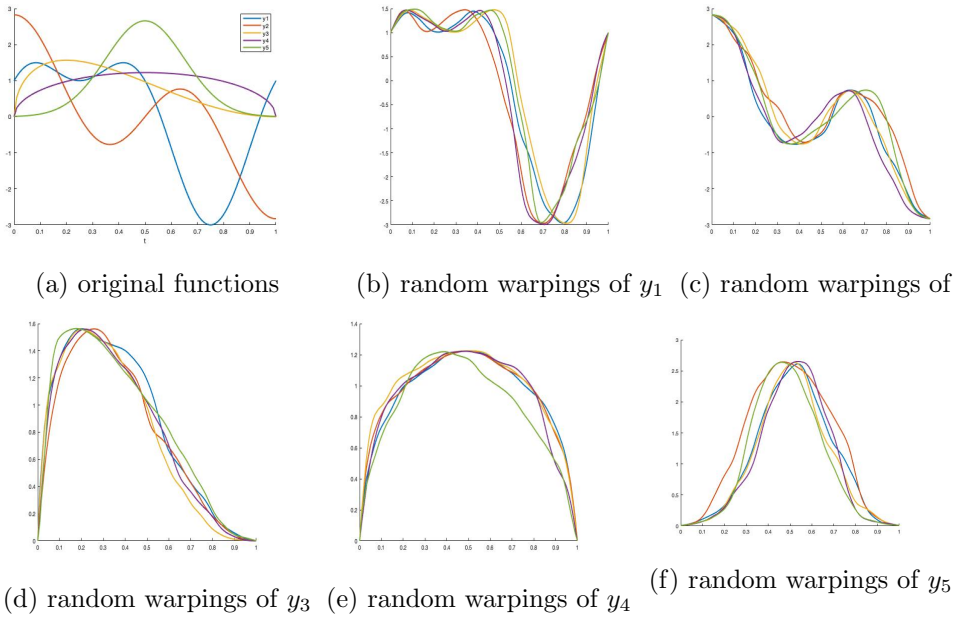
In this section the algorithm described in section 2.2 is implemented on a set of simulated functions. Due to the manner in which these functions are simulated one already knows what kind of clustering outcome to expect. Therefore obtaining the expected outcome at the end illustrates the validity of our algorithm.

Initially, five visibly different and smooth functions of t∈[0,1] are chosen as described below and shown in Fig 6a:

1. $y_1(t) = 2sin(\pi t) + cos(4\pi t)$
2. $y_2(t) = \sqrt{2}cos(\pi t) + \sqrt{2}cos(3\pi t)$
3. $y_3(t) = \sqrt{\frac{t(1-t)^4}{Beta(2,5)}}$
4. $y_4(t) = \sqrt{\frac{t(1-t)}{Beta(2,2)}}$
5. $y_5(t) = \frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}(\frac{t-\mu}{\sigma})^2}, \pi = 0.5, \sigma = 0.15$

Next, for each of these functions, $y_1, y_2, ..., y_5$, five random warping functions $\gamma$ from the set $\Gamma$=$\{\gamma : [0,1]\to[0,1] | \gamma$ is diffeomorphism, $\gamma(0) = 0, \gamma(1) = 1\}$ are generated and acted on the respective functions thus producing five random warpings of each of the five original functions as shown in Fig 6b-6f.

Taking all these together we now have a set of 25 functions $f_1, f_2, ..., f_{25}$ in which $f_1, f_2, ..., f_5$ are random warpings of the original function $y_1$; $f_6, f_7, ..., f_{10}$ of $y_2$; $f_{11}, f_{12}, ..., f_{15}$ of $y_3$ and so on. Our objective is to cluster these functions using the distances between their SRVF representations



(a) original functions (b) random warpings of $y_1$ (c) random warpings of $y_2$

(d) random warpings of $y_3$ (e) random warpings of $y_4$ (f) random warpings of $y_5$

Figure 6: Simulated data

as described in the algorithm in section 2.2. Ideally, we would expect to see $f_1, f_2, ..., f_5$ clustered together, $f_6, f_7, ..., f_{10}$ clustered together, $f_{11}, f_{12}, ..., f_{15}$ clustered together and so on.
We apply Steps 1 through 5 to these twenty five functions ultimately generating a dendrogram plot of the hierarchical cluster tree which is shown in Fig 7. We observe that the results are as expected, i.e., that our algorithm clusters the functions correctly.

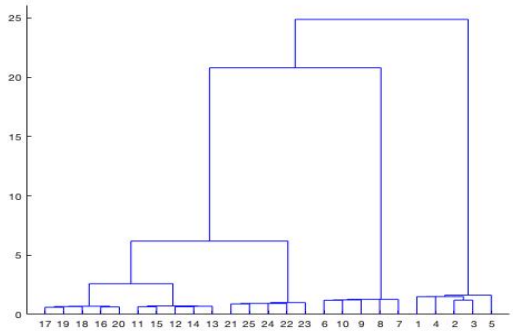



Figure 7: Dendrogram plot of simulated data

## 3 Experimental Results

Here, we align each pair of functional data on stock prices of some companies in the SRVF space and then measure the distances between them as described in 2.1.4 following the algorithm given in 2.2. The validity of our algorithm was demonstrated in section 2.3 and now we apply it to stock

market data.

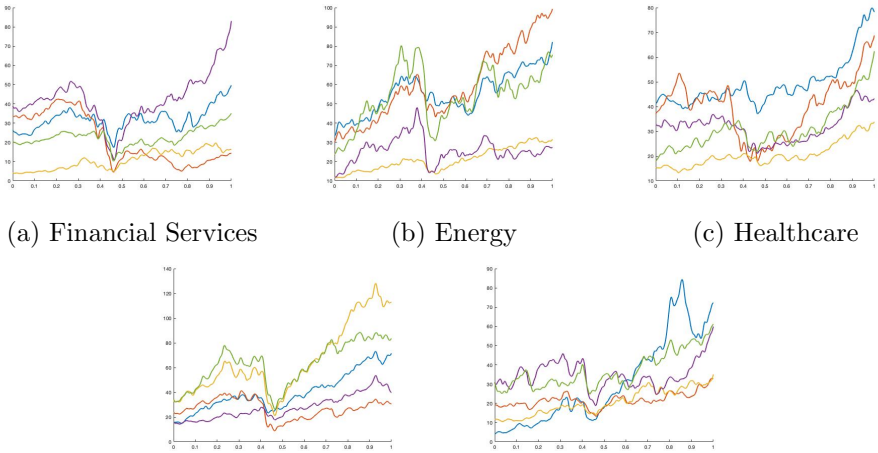
### 3.1 Data and its Preprocessing

Five different sectors of the market are chosen: Financial Services, Energy, Healthcare, Real Estate and Technology. Five brands are selected from each of these five sectors(Table 1) and their historical data for each business day from 1st January, 2005 to 31st December, 2013 are obtained from `https://finance.yahoo.com/`. The choice of brands is made carefully to ensure that there are no

| Financial Services | Energy | Healthcare | Real Estate | Technology |
|---|---|---|---|---|
| JP Morgan | Exxon Mobil | Johnson n Johnson | American Tower | Apple |
| Bank of America | Chevron | United Healthcare | Prologis | Microsoft |
| HDFC | TC Energy | Abbott Laboratories | Simon Property | Oracle |
| American Express | Suncor Energy | Eli Lilly and Company | Welltower Inc | Adobe |
| US Bank | Schlumberger | CVS | Boston Properties | Qualcomm |

Table 1: Table showing names of brands from sectors chosen for this study

impossible data combinations, missing values, etc. Out of all the variables in each of these datasets the variable 'Adjusted Closing Price' is of importance to us. This variable presents the closing stock price for the day adjusted for the dividend. Upon plotting the Adjusted Prices over this 9 year time span it is observed that these functions are very rough and therefore their derivatives are not defined at the sharp edges. Since the calculation of derivatives of the functions is essential to our method, the data was further processed through a smoothening function using the 'moving average' method and the results are shown in Fig 8. It appears that the functions, especially the



(a) Financial Services (b) Energy (c) Healthcare

(d) Real Estate (e) Technology

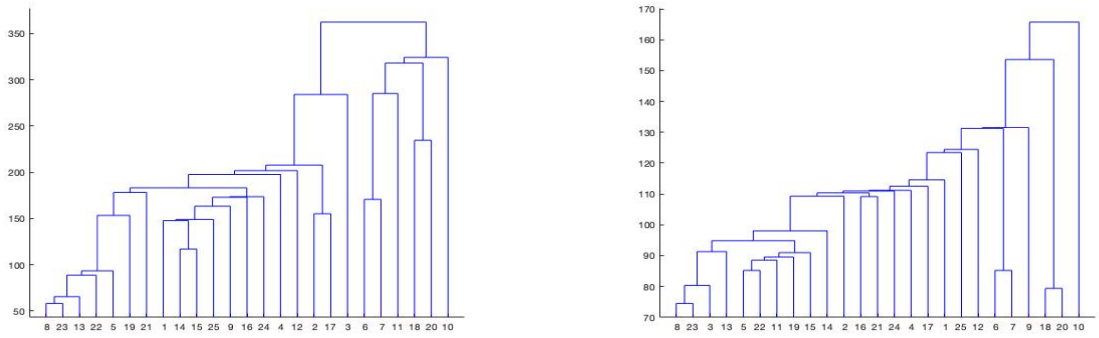
Figure 8: Smoothened stock price functions

ones from the sectors Financial Services and Real Estate are already quite aligned with respect to a time point between 0.4 and 0.5 where the stock prices for all the brands seem to drop. This point, in fact, refers to the Great Recession of 2008 when a general economic decline was observed

in world markets. Financial Services and Real Estate were among the first few industries to get hit by this recession. This is the time period our study focuses on. So, instead of working with this 9 year long data(2265 time poins), roughly only the middle 1001 time points corresponding to the years 2007 to 2009 were considered for this analysis.

### 3.2 Results

We have 25 stock price functions, $f_1, f_2, ..., f_{25}$ in which $f_1, f_2, ..., f_5$ correspond to brands from the Financial Services sector, $f_6, f_7, ..., f_{10}$ from the Energy sector, $f_{11}, f_{12}, ..., f_{15}$ from the Healthcare sector, $f_{16}, f_{17}, ..., f_{20}$ from Real Estate and $f_{21}, f_{22}, ..., f_{25}$ from Technology. Naurally, we expect $f_1, f_2, ..., f_5$ to be clustered together, $f_6, f_7, ..., f_{10}$ together, $f_{11}, f_{12}, ..., f_{15}$ together and so on. First, we use $\mathbb{L}^2$ norm to calculate and compare the distance between each pair of functions as described in section 2.1.1. The hierarchical clustering obtained using this metric is shown in the dendrogram plot in 9a. The result obtained is not impressive and that was expected. Next, Algorithm 2.2 is applied on the stock price functions and the results are observed. Fig 9b shows the dendrogram plot for the clustering of our stock market data using alignment of their SRVF representations. It appears that with the exception of only a few brands, the stock price functions belonging to




(a) using $\mathbb{L}^2$ norm

(b) using SRVF

Figure 9: Dendrogram plot of stock market data

a particular sector which should have been grouped together are placed quite far apart from one another which is in contrast with the 100% accurate results obtained for simulated data shown in Fig 7. In fact, the SRVF results do not exhibit much improvement in clustering from that obtained under $\mathbb{L}^2$ norm.
For further investigation, we consider the *AmpDist* matrix. $AmpDist[i, j]$ is the amplitude distance between the aligned SRVFs of the ith and jth functions. From this, for each function we find its nearest neighbour and determine whether they belong to the same sector. In the case of $\mathbb{L}^2$ norm, 37.5% of the nearest neighbours are found to be from the same sector and in the case of SRVFs it is 33.33%.

## 4 Discussion and Conclusion

In this paper we discussed the importance of choosing the correct metric and space for analysis of functional data for one's inferences to be meaningful. Starting from the most natural choice,

the Euclidean space, various other options were discussed and their limitations demonstrated. Ultimately, the method of time warping of Square Root Velocity Functions was chosen to develop an algorithm for functional alignment and calculation of distances between them and then clustering of the ones that are closer to one another in the same group. This procedure successfully clustered a set of simulated data where each group of functions was purposely chosen to be distinctly different form the other groups.
Next, we took up the challenge of testing our procedure on stock market data. The results were not very impressive. Infact, clustering done based on the SRVF distances was no better than that done using Euclidean distances between the unaligned data. Futhermore, the percentage of the nearest neighbors of each function that were placed in the same group as the function was lesser for SRVF distances than that for Euclidean distances by a slight bit. There can be a number of reasons why our method was not as successful as it should have been. We simply do not know what underlying variables control the behavioral biases in this data. Perhaps our presumption that commodities from the same market sector should have similar stock price data over the same time period is a gross simplification of a much more complex problem. One reason behind the equally good performance of clustering without any alignment of the functions may be the fact that the data are already quite aligned. Upon observing the data plots(Fig 8) it is apparent that there is some alignment present already. For future research, a better comprehension of the data needs to be achieved and the dominant factors influencing it need to be incorporated in the study or a more robust technique needs to be developed.

# 5 Bibliography

# References

[1] Anuj Srivastava and Eric Klassen. *Functional and Shape Data Analysis*. Springer, New York, 2016.