Introduction to Machine Learning, Spring 2016

Spectral Clustering of the Stock Market

1 Introduction

As evidenced by the economic impact of the dot com boom and the more recent real-estate bubble, the stock prices of firms fluctuate with public/investor perceptions, particularly for industry-wide trends. But do these movements occur randomly for individual companies or is there a larger pattern across similar groups? With the underlying assumption that competing companies of the same industry, such as real estate or software, would also experience the same political and technological conditions, it can be seen that the stock prices of related businesses would move together.

However, the primary obstacle to this intuition is the idea that similar companies may react to these trends at different speeds, due to external factors of scale, investor preferences, or even location. This meant that while company stock prices may have had similar up and down reactions, their curves might have differed with an external delay. Such an effect would obviously increase a chance for error, which could manifest as incorrectly correlating companies of different industries that reacted to separate but closely timed events, and vice versa.

But with the influence of these lag structures in mind; we believe we could enrich spectral clustering experiments of time series by evaluating a new type of similarity kernel: the Gaussian Edit Distance with Real Penalty (GERP) kernel, which has seen promise in research for more accurately measuring the similarity between time warped time series.

Overall, this paper demonstrates that while natural clusters of stock price movement may not perfectly correspond to respective industries, consideration for lag may encode more meaningful movement information with prediction potential.

2 Related Work

Our work is in fact an alternate exploration of the Edit Distance with Real Penalty (ERP) metric introduced in another paper written in 2004. The original distance function was introduced as a version of metric L_p-norm category of functions which could counter the lag between time series through local time warping. It was inspired by edit distance scores made more popular in spell-check and natural language processing applications. By introducing a finer metric to evaluate the similarity of time-series, authors Chen and Ng evaluated their metric among other similarity functions through heuristic pruning strategies in search time performance for large historical stock price time series. While their paper succeeded in developing a powerful pruning algorithm with their ERP heuristic score, our paper uses the ERP distance metric to not only test our hypothesis without lag error but also develop a simple trading strategy from the resulting leaders found. While our data may be similar, our paper's objectives are more specific within the Economic field; using ERP's time warping feature to enhance our spectral clustering's performance.

3 Problem Definition and Algorithm

3.1 Task

Using the data collected from the S&P 500, we will attempt to cluster a sample of 72 stock price curves into 24 distinct clusters. We will calculate returns (i.e. percent changes) on the stock prices to normalize them into samples with an approximate mean of 0 and variance independent of time. Building on top of our industry-wide movement intuition, we will be testing the accuracy of the GERP kernel with spectral clustering by cross-referencing our results with the corresponding Global Industry Classification Standard (GICS) industry groups. After optimizing our results, we'll compare the spectral clustering groupings of the GERP against those produced from two baselines: k-means with Euclidean distance and spectral clustering with a rigid Pearson Correlation kernel. We will formally evaluate our results with the both rand index and adjusted rand index metric.

3.2 Algorithm

The primary clustering algorithm we attempt to use is spectral clustering, a clustering method which uses the popular k-means algorithm on the eigenvectors of the data to provide better results on clusters composed of points that are related but not necessarily compact within the $\mathbb{R}^n$ space.

For our primary distance metric, we applied an algorithm in *literature*[1] that describes the success of a metric called the Edit Distance with Real Penalty (ERP) in signal processing to historical stock price returns. This family of elastic measures basically sums the differences in value between "corresponding dates" in the two time series plus any "penalties", where "corresponding dates" refers to the optimal lag structure of the two time series and measure-specific "penalties" are added for each change in lag i.e. when $v_i \neq v_{i-1}$. If there are no lags or leads, then $d(\boldsymbol{a}, \boldsymbol{b}) = \|\boldsymbol{a} - \boldsymbol{b}\|_1$ i.e. the elastic measure is exactly the L_1distance.

Formally, the optimization problem is recursively defined as:

$$|\boldsymbol{a} - \boldsymbol{b}|_{erp} = d_{ERP}(\boldsymbol{a}, \boldsymbol{b}) = \begin{cases} \sum_{i=1}^{n} |g - b_i| & \text{if } m = 0 \\ \sum_{i=1}^{m} |a_i - g| & \text{if } n = 0 \\ \min \begin{cases} d_{ERP}\big(\text{tail}(\boldsymbol{a}), \text{tail}(\boldsymbol{b})\big) + |\text{head}(\boldsymbol{a}) - \text{head}(\boldsymbol{b})| \\ d_{ERP}(\text{tail}(\boldsymbol{a}), \boldsymbol{b}) + |\text{head}(\boldsymbol{a}) - g| \\ d_{ERP}\big(\boldsymbol{a}, \text{tail}(\boldsymbol{b})\big) + |g - \text{head}(\boldsymbol{b})| \end{cases} & \text{otherwise} \end{cases}$$

Where $\boldsymbol{a} \in \mathbb{R}^T$ and $\boldsymbol{b} \in \mathbb{R}^T$ are two time series, a_i and b_i are the numeric values at time i for time series $\boldsymbol{a}$ and $\boldsymbol{b}$ respectively, $\text{head}(\boldsymbol{a}) = a_1$, $\text{head}(\boldsymbol{b}) = b_1$, $\text{tail}(\boldsymbol{a}) = \langle a_2, \dots, a_m \rangle \in \mathbb{R}^{m-1}$, $\text{tail}(\boldsymbol{b}) = \langle a_2, \dots, a_n \rangle \in \mathbb{R}^{n-1}$, and g is a constant gap cost (usually 0).

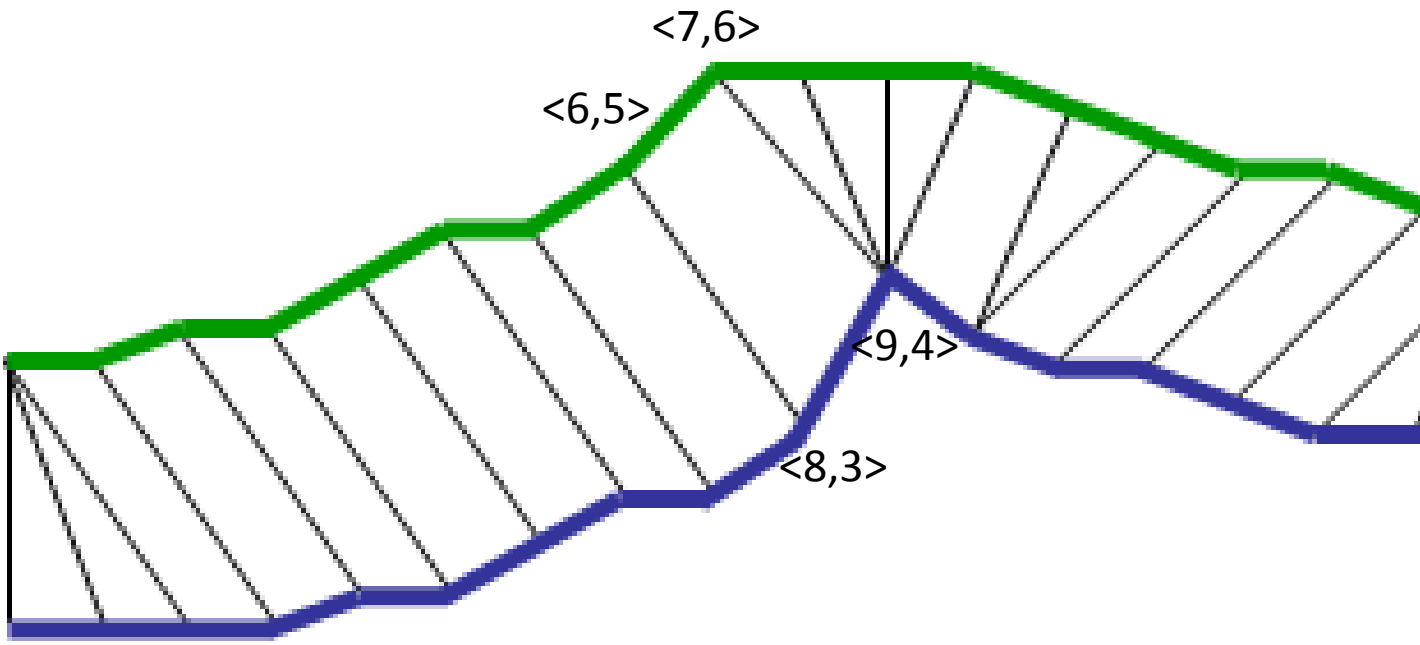
In edit distance parlance, the first alternative in the optimization is known as substitution, the second is insertion, and the third is deletion (from the perspective of transforming time series $\boldsymbol{a}$ into time series $\boldsymbol{b}$).

For every insertion, $\boldsymbol{a}$'s cumulative lead over $\boldsymbol{b}$ decreases by 1 day because later values in $\boldsymbol{a}$ are matched to earlier values in $\boldsymbol{b}$. For every deletion, $\boldsymbol{a}$'s cumulative lead over $\boldsymbol{b}$ increases by 1 day because later values in $\boldsymbol{b}$ are matched to earlier values in $\boldsymbol{a}$.

The ERP distance differs from naïve elastic measures (e.g. Dynamic Time Warping) by using the constant g in place of $\text{head}(\boldsymbol{b})$ or $\text{head}(\boldsymbol{a})$ in insertions and deletions (but not substitutions) so that the metric is symmetric and satisfies the triangle inequality.

Visually, what the family of elastic distances does is stretch and transform the lagging time series, choosing an alignment so that the L_1 distance between both time series are minimized. Below is a subset of two example time series ($\boldsymbol{a}$ is blue and $\boldsymbol{b}$ is green) where the optimal operations are two substitutions.

$$|\boldsymbol{a} - \boldsymbol{b}|_{erp} = (\ldots + |3 - 5| + |4 - 6| + \cdots)$$
$$\boldsymbol{a} \text{ cumulative leads over } \boldsymbol{b} = \boldsymbol{v} = \langle \ldots, 8 - 6, 9 - 7, \ldots \rangle$$



In the elastic distance measure family in general, if the lags $\boldsymbol{v}$ between time series $\boldsymbol{a}$ and $\boldsymbol{b}$ vary over time such that $\boldsymbol{a}$ moves together with $\boldsymbol{b}$ on day 1, lags $\boldsymbol{b}$ by 1 day on day 2, leads $\boldsymbol{b}$ by 1 day on day 3, and leads by 3 days on day 4 (i.e. $v_1 = 0, v_2 = -1, v_3 = 1, v_4 = 3$), then,

$$d(\boldsymbol{a}, \boldsymbol{b}) = |a_1 - b_{1+0}| + |a_2 - b_{2-1}| + |a_3 - b_{3+1}| + |a_4 - b_{4+3}| + \cdots + \left|a_n - b_{n+v_n}\right| + \text{penalties}$$

In the case of ERP, the penalties would be the sum of $\left|a_i - b_{i+v_i}\right| - |a_i - g|$ for each insertion of a_i and $\left|a_i - b_{i+v_i}\right| - \left|g - b_{i+v_i}\right|$ for each deletion of a_i.

It is important to note that in our Gaussian kernel, we choose a σ that scales $|\boldsymbol{a} - \boldsymbol{b}|_{erp}$ with the time series length T, which reduces a bias towards shorter time series with an overall lower accumulation of comparisons. This is simply because with longer time series, there are more $|a_i - b_j|$ terms being summed.

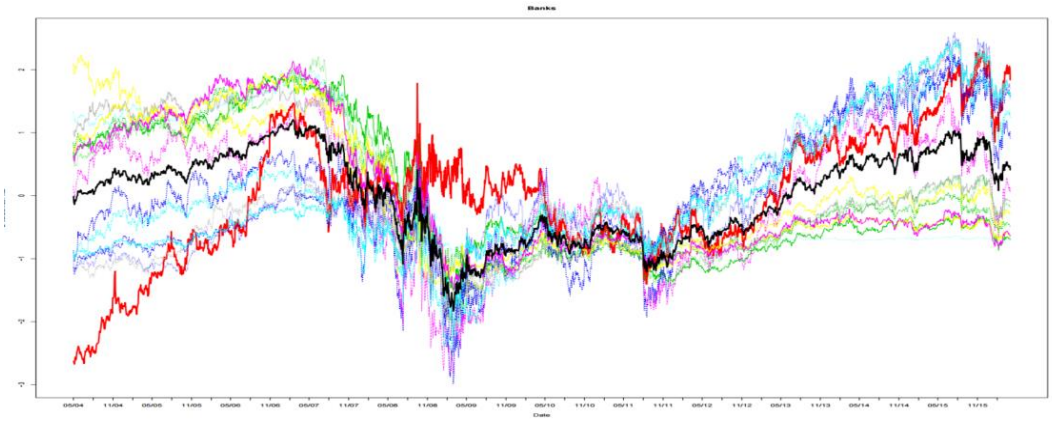
In contrast to other local time warping distance functions, ERP satisfies the four conditions that define a metric space [2]. Therefore, Euclidean distance can be substituted with ERP in the radial basis function (RBF) kernel, creating a pseudo-kernel metric we refer to as GERP. While it can be proved that GERP is still indefinite, the indefiniteness of the kernel theoretically only has a big impact when used in SVM and is inconsequential in the context of spectral clustering. Spectral clustering in particular merely requires an affinity matrix to describe how similar each pair of time series is and performs no convex optimization on the matrix. Thus, it is valid to use the GERP kernel as the basis of our affinity matrices to cluster our time series.

4 Experimental Evaluation

4.1 Data

Following our intuition that industry leaders would affect the stock market movements, we believed the high speculation and volatility of large firms in the S&P 500 was the perfect place to source our time series. So, we accessed the stock price time-series from the Bloomberg Terminal, downloading historical stock price data from the S&P 500 tracing back from April 2016 to 1969. The following is a graphical sample of one industry over time.

Banks:



To clean our data, we first chose companies that have been public for at least 3000 trading days (i.e. companies with price data between May 5, 2004 and April 5, 2016), which may introduce some survivorship bias because of the failure to include newer companies and failed companies. We further reduced our dataset by removing outliers (highlighted in red) whose correlation to the simple contemporaneous industry average (highlighted in black) was BOTH below 0.5 AND within the bottom 10% of such correlations for the industry group. We also standardized our prices into percentage daily returns (i.e. the percentage difference between today's and yesterday's price) to normalize them into samples with an approximate mean of 0 and variance independent of time. After filtering, a sample of our data set is graphically depicted as follows:

Filtered Banks:

4.2 Methodology

Our initial assumption was that companies in the same industry will be affected by the same macroeconomic factors and so their stock prices will move in tandem. We scaled, cleaned, and sampled our data so to prevent overfitting on idiosyncratic price shocks and to improve the run-time of our experiment. The expectation is that with enough samples, the aggregate cross validation performance should converge to the population performance.

Our hypothesis is that clusters of the GERP kernel would more accurately correlate stock price movements to their respective industries. We compared our results to a 2 other standard metrics: the Euclidean distance with k-means and the Pearson Correlation statistic with spectral clustering.

For our procedure, we investigated the application of clustering of historical stock prices by representing them as time series. A time series data set represents the varying value of a variable over time. A time series can be plotted where the x values are dates and the y values are the value of the variable at that date. A time series itself is straightforward enough to be represented as a feature vector in n-dimensional space, where n is the number of dates and each cell in the vector is a price. However, such a simple representation of time-series data suffers from the curse of dimensionality, where such a large quantity of dimensions does not encode much information on its own but instead co-varies with the relationships between cells. If we naively cluster our time series as these enormous vectors, our high dimensionality might either overfit or produce poor clusters. We turn to custom kernels to prevent naïve clustering algorithms from overemphasizing inconsequential cell-by-cell differences in the parallel feature vectors.

Even when countering the codependence of cells through enrichment with kernels, it does not eliminate time axis distortion. Due to their inflexible nature, the Euclidean distance metric and the related cosine similarity and Pearson correlation statistics generally compare two parallel time series quickly but poorly. This is because when the dates on the two time series are matched up, the movements of the two time series can be offset from each other, either by a constant or variable offset, e.g. time series $\boldsymbol{a}$ may make a move on day 1 and a related time series $\boldsymbol{b}$ may, due to a delayed reaction, make a move on day 2 in response to the same event that moved $\boldsymbol{a}$ earlier. We hereinafter refer to the set of offsets as the lag structure of the two time series, which can be expressed as a vector $\boldsymbol{v}$ where v_i refers to the number of days

that $\boldsymbol{a}$ leads in movement over $\boldsymbol{b}$ on day i. For clarity, a constant offset of k means that $v_i = k \; \forall 1 \leq i \leq n$. We will define a kernel that takes into account the lag structure of two series.

We also evaluated our clusters to our Global Industry Classification Standard industry groups through the popular Rand Index metric and the improved Adjusted Rand Index [3].

Rand Index:

$$\frac{\sum_i \sum_j \binom{N_{i,j}}{2}}{\frac{1}{2}\left[\sum_i \binom{\sum_j N_{i,j}}{2} + \sum_j \binom{\sum_i N_{i,j}}{2}\right]}$$

Adjusted Rand Index:

$$ARI = \frac{\sum_i \sum_j \binom{N_{i,j}}{2} - \left[\sum_i \binom{\sum_j N_{i,j}}{2} \sum_j \binom{\sum_i N_{i,j}}{2}\right] / \binom{n}{2}}{\frac{1}{2}\left[\sum_i \binom{\sum_j N_{i,j}}{2} + \sum_j \binom{\sum_i N_{i,j}}{2}\right] - \left[\sum_i \binom{\sum_j N_{i,j}}{2} \sum_j \binom{\sum_i N_{i,j}}{2}\right] / \binom{n}{2}}$$

We also calculate the inertias of the resulting clusters to get a sense of the cluster densities in the original time series feature space.

Inertia:

$$\sum_{k=1}^{K} \sum_{i \in S_k} |\boldsymbol{x_i} - \overline{\boldsymbol{x}}_{\boldsymbol{k}}|_2^2 = \sum_{k=1}^{K} \sum_{i \in S_k} \sum_{j=1}^{T} \left(x_{i,j} - \bar{x}_{k,j}\right)^2$$

Where S_k is the set of time series in the kth cluster, $\overline{\boldsymbol{x}}_{\boldsymbol{k}} \in \mathbb{R}^T$ is the center of the kth cluster, and $\bar{x}_{k,j}$ is the time j value of the cluster center for the kth cluster. This simply sums the Euclidean distance between time series in a cluster to the contemporaneous average of the cluster members and is also known as the k-means objective function.

Even if the clusters do not correspond with industry groups but all time series in a cluster subjectively are similar to each other, we may be able to justify the discrepancy by demonstrating some latent factor common to the varied companies that is not taken into account by the GICS industry groups.

4.3 Results

As a short recap, we framed our experiments to determine whether the proposed GERP kernel would provide significant classifying improvements over similar clustering algorithms. We validated how accurate our resulting natural clusters against the formal Global Industry Classification Standard, which identified firms by their fields and industries.

After conducting our initial experiments, we found that while the GERP kernel provided significant improvements over the standard k-means clustering algorithm, it did not show improved accuracy against the rigid metrics that did not account for time warp.

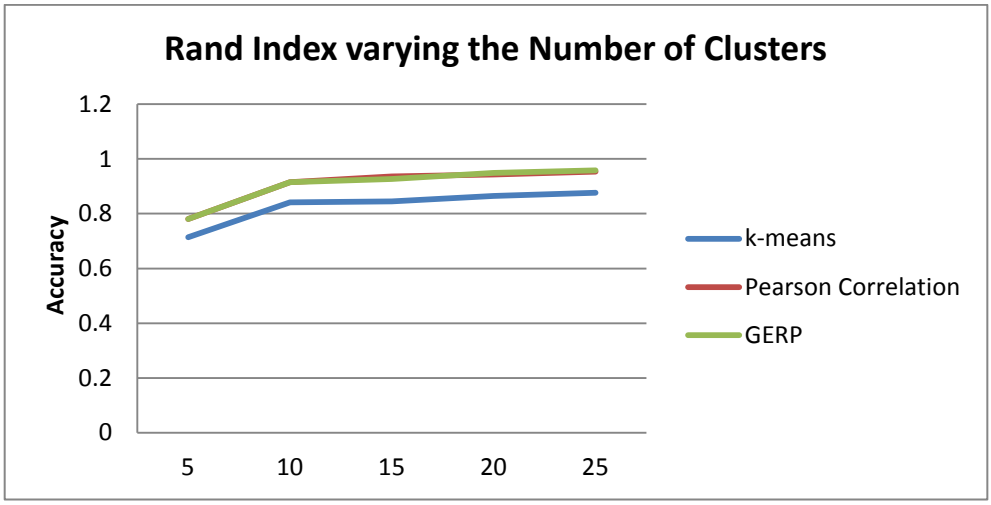
The following figure is a table comparing our three approaches along with 2 evaluation metrics and inertia for a sample of 72 companies across 24 industries, between the dates May 5th, 2004 and July 14th, 2005.

	Rand Index	Adjusted Rand Index	Inertia
K-Means	0.88112	0.13743	1.52762
CORR	0.95610	0.30498	2.28468
GERP	0.95282	0.18928	2.92361

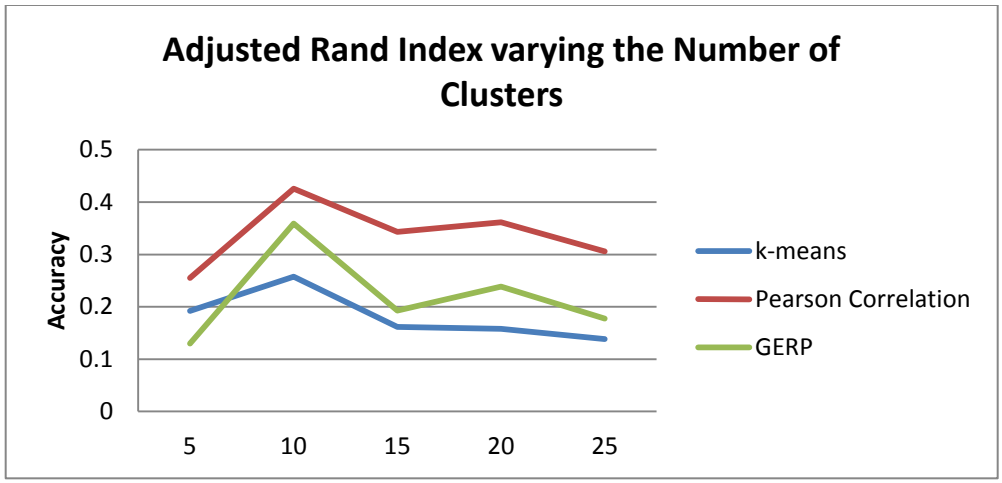
From these results we can conclude the Time Warping GERP kernel does not provide significant improvements over rigid spectral clustering algorithms.

One important caveat we should note is that while we took many measures to avoid overfitting, including cross-validating our results across 5 separate samples, it does not absolutely eliminate a variety of bias which introduce the risk of error in our results. The following measurements were motivated by the desire to further validate the robustness of the previous results.

With the number of clusters greatly outnumbering the individual members, we wondered whether the sparsity of our data was truly affecting our results. To further reduce the impact of overfitting or difficult outlying industries, we re-sampled our data across a variable number of industries, varying from 5 – 24 with exactly 3 members per cluster, to determine if any changes for any of our metrics would occur.

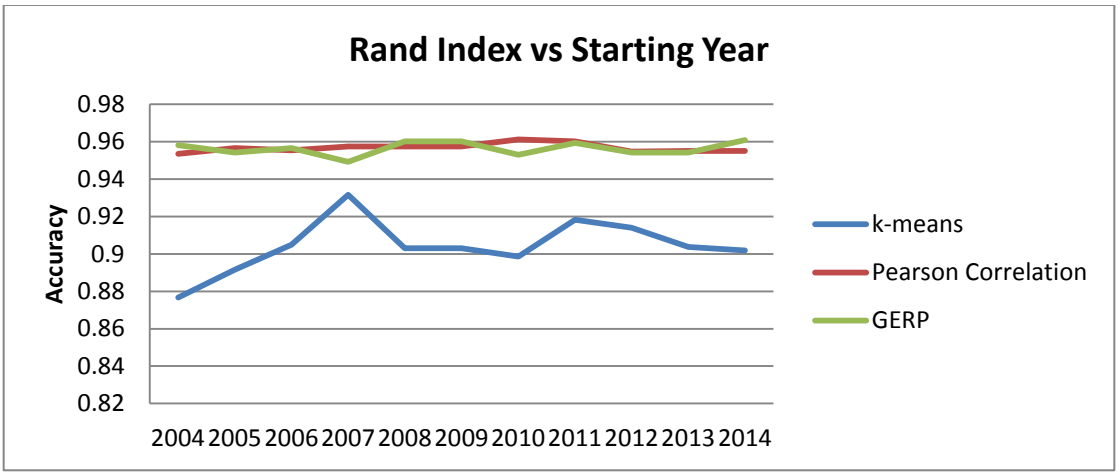


Varying the number of industry clusters has little effect on the overall Rand Index score. The Adjusted Rand Index score saw equally uninteresting results, with a highly fluctuating and seemingly random graph.

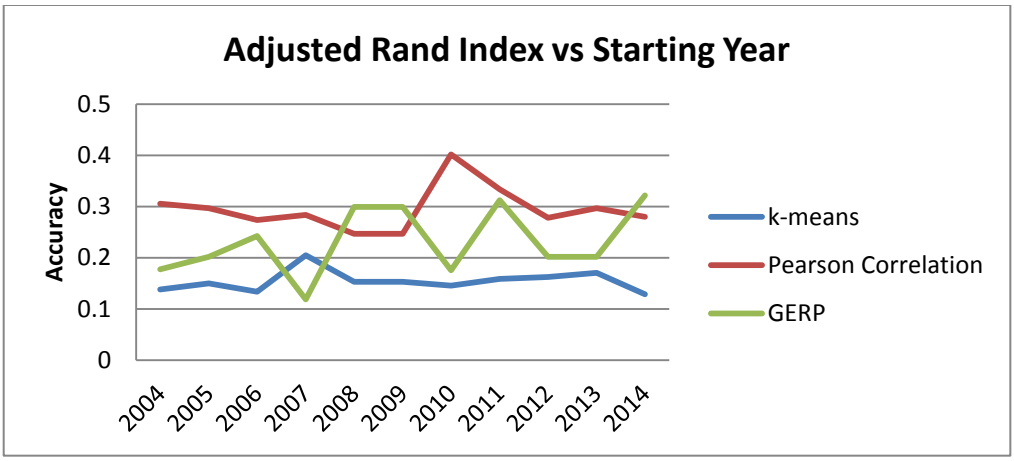


However, even by changing the ratio of clusters to data points, the results are rather consistent with our original observations that the Pearson Correlation algorithm slightly outperforms GERP and k-means. Regardless of the comparative results, the low margin of error for accuracy across different number of clusters demonstrates that our results remain consistent across varying levels of sparsity.

Another variable we also wanted to test that could potential affect our results was the range of dates of our time series, which could present bias toward specific economic conditions. Once again, we re-sampled our 24 industry groups across 10 1-year periods, which resulted in the following accuracy graphs.



Once again, while the Rand Index evaluations showed little difference by varying the range of dates, the Adjusted Rand index provided similar conclusions in a high variance and seemingly random graph.



The narrow Rand Index band – with a width of 0.10 for the entire decade despite high volatility and a dramatic transformation of the economic environment from the 2005 Real Estate Boom to the 2008 Recession – further demonstrates the generalizable and robust results of our clustering methods. While in some cases GERP did outperform, on average the Pearson Correlation was more accurate.

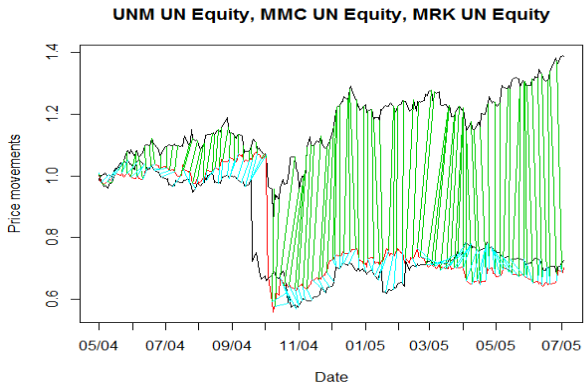
4.4 Discussion

While the GERP kernel did provide strong results compared to k-means, our experiments did not validate our hypothesis, showing instead the GERP kernel did not provide significant accuracy performance over a baseline of non time-warping spectral clustering methods.

From our primary results, the first point of interest we notice is that inertia, or the sum of squared distances function describing the compactness of clusters, is not necessarily correlated to the accuracy of our algorithms. This independent relationship can easily be attributed to the novel features of spectral clustering, which ignore simple convex boundaries of k-means for the finer –grained and more abstract evaluation with subspace connectivity. With a significant performance boost in tradeoff with less compact clusters, these results briefly verify spectral clustering was the correct model to optimize time-series clustering accuracy.

The second observation we find in our results is that while our rigid Pearson Correlation kernel performs better than GERP, the difference is rather miniscule, which may suggest similar performance. This observation however is not found in our findings with the Adjusted Rand Index matrix, which shows the same results but noticeably different scores which may disprove the similar performance case and that the simple Pearson Correlation kernel performs better than GERP in classifying time series.

Upon further investigation, we discovered latent information that the GERP kernel may have taken into consideration, undermining a key assumption behind our hypothesis. The following figure is the lag structure of the most correlated cluster from spectral clustering with GERP. The three members include 2 large insurance providers: the Unum Group and Marsh & McLennan Companies Inc. as well as the pharmaceutical giant Merck & Co.



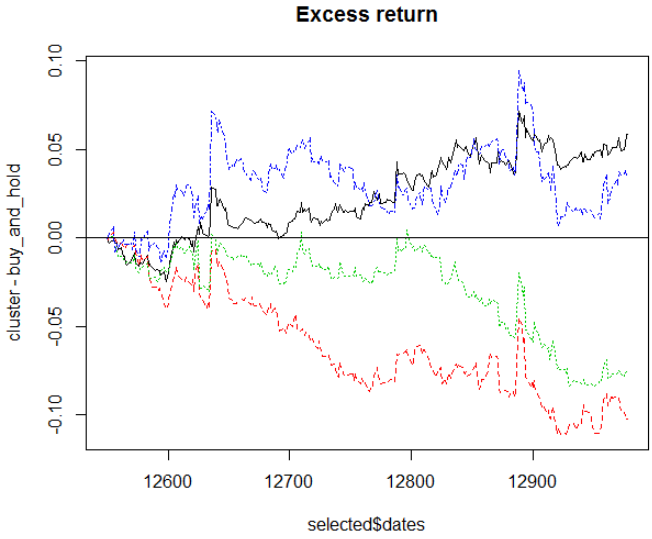
While pharmaceuticals and insurance may not belong to the same industry, their stock prices do betray related movement which we initially believed would exist among related firms and competitors of the same industry. One particular price fluctuation was an enormous drop in the Fall of 2004, which corresponded with the withdrawal of the dangerous Vioxx drug from Merck's offering. With the increasing rise of studies linking life-threatening health defects to the popular painkiller, Merck lost billions of dollars from the fallout of lawsuits and media storms that followed. From these lawsuits, Merck's insurance providers, which some we identify in the cluster above, were also ordered to pay out and lose investor trust as a result of the scandal. While this was only one of many examples, the idea that movements within the stock market may also occur deliberately between industries is a concept that may have challenged the power of our ground truth and overall hypothesis.

One interesting note of GERP that we did notice was the different size distributions between algorithms. In our initial data set, we sampled exactly 3 companies from a total of 24 different industries which we fed into 3 different clustering algorithms. The following are the results for our 2 spectral clustering approaches:

CORR cluster sizes: [2 5 3 3 4 2 2 3 6 3 2 3 3 4 2 2 2 3 5 2 2 2 3 4]
CORR size mean: 3, CORR size standard deviation: 1.14208
GERP cluster sizes: [3 3 4 3 2 3 3 3 3 3 2 3 3 4 4 2 2 4 4 3 3 2 3 3]
GERP size mean: 3, GERP size standard deviation: 0.65938

While both representations average 3 members per cluster exactly, the variation of the Pearson Correlation cluster size is much larger than that of the GERP kernel, a perhaps minor observation which may raise the issue of the natural sparsity of our data, particularly with Rand Index metrics which traditionally stronger results with larger clusters.

The final additional aspect of this project we started to explore was the practical usage of our lag structures in finding industry leaders which could signal early movements within the market. The following figure shows our returns of an active group trading strategy over a simple buy and hold baseline, which did show promising returns.



Blue = k-means, Black = GERP, Red = Correlation, Green = Industry

However, the results in different time periods had extremely high variance, indicating that the trading parameters were overfitted for the sample we trained on. We concluded that the cluster leaders were too inconsistent to make a competitive strategy, so we decided to defer a deeper analysis for future work.

5 Conclusions

In conclusion, while our results remained consistent over variations in sample companies, data sparsity, and time, we found no evidence that the time warping metric: the Gaussian Edit Distance with Real Penalty provides a significant improvement in accuracy for classifying economic time series.

We initially wished to also explore an objective way to evaluate whether the reason the GERP kernel had a lower Rand Index was because the companies in each cluster it produced were more similar to each other than the companies in each industry group. The high Rand Index for the correlation based clusters discourages this exploration however because it suggests that the time series in an industry group are in fact similar enough that a high quality clustering should tend towards industry groups.

We also proposed to evaluate a trading strategy using the ERP lag structures in the hopes that each cluster would have a clearly identifiable price movement leader, but this was quickly found to be an erroneous assumption after we prototyped the strategy and analyzed the poor trading results.

In the future, we would not only like to see the investigation of other time warping metrics for clustering time series to be evaluated, such as Dynamic Time Warp Distance, but more in depth exploration of trading strategies which take advantage of the resulting lag structures. Perhaps we can be more selective about the clusters we trade as well as refine our entry and exit signals.

6 Bibliography

[1] http://www.cs.ust.hk/~leichen/pub/04/vldb04.pdf

[2] http://arxiv.org/ftp/arxiv/papers/0802/0802.3522.pdf

[3] https://pdfs.semanticscholar.org/52d4/8b393f3f838f2370c50af03703eee0bbd669.pdf