A Study of Russell 3000 Dimensionality Using non-linear Dimensionality Reduction Techniques

Nitish Bahadur, Kelum Gajamannage, and Randy Paffenroth

Abstract—Financial markets are high-dimensional, complex, and constantly changing. Under stressed market conditions the changes are amplified. Financial market can be represented by an underlying manifold in low-dimension that captures the inherent characteristics of the high-dimensional data. Using Russell 3000 constituents and both geodesic and informational geometric schemes, we determine the temporal dimensionality of US market. Further, we use rate of change in US market dimensionality over 30 years to detect early warning system. Additionally, using intra-day prices we zoom into temporal dimensionality around large market movements to detect early perturbation in financial system. We not only study the benefit of using non-linear techniques such as Isomap, over linear technique such as PCA or Multidimensional Scaling but also compare and contrast the use of geodesic distance and informational geometric distances.

Index Terms—Manifold, non-linear dimensionality reduction, Russell 3000, PCA, MDS, Isomap, geodesic, information metric.

1 Introduction

Automation, algorithmic trading, and globalization has not only made financial markets more integrated but also reduced the lag between information diffusion between diverse market centers such as Japan, Hong Kong, London, and New York. However, has automation changed the temporal dimensionality of financial market? Furthermore, for investors who trade frequently, social media has increased the number of analytics an investor needs to analyze. We conjecture if these additonal factors have significantly changed the dimensionality of financial markets? Consequently, the number of stocks traded have increased and the frequency of buying and selling in portfolios have increased. Buying and selling public stocks is predominatly systematic.

To empirically determine the instrinsic dimensionality of the financial market both during normal market conditions and stressed market conditions, we use an index that encompasses the vast majority of financial market. Dimensionality of Russell 3000[1] Index (aka proxy for financial market) is the least number of factors required to explain the market behavior. Given the large number of factors, it is important to distinguish between intrinsic latent factors and noise factors. Moreover, the magnitude of velocity of change in dimensionality indicates change in financial market conditions. This will help build and early warning system that can alert investors earlier.

Investors use linear techniques such as Principal Component Analysis (PCA) [1], where new orthogonal features are created by linearly combining observed factors and projecting them along direction of maximum variability. While PCA reduces dimensionality by preserving the correlation structure of data, Multidimensional scaling (MDS) [2] reduces dimension by preserving the Euclidean distance between data points. The distance matrix arising from Euclidean metric relies on straight line distances, which limits MDS's applicability for non-linear data.

Isometric mapping (Isomap) [3] [4], a non-linear dimension reduction technique, preserves pairwise geodesic distance between data points in original high-dimensional space and successfully addresses the important limitation in MDS [5]. Although Isomap has been used successfully to analyze data from several instances such as collective motion, face recognition, and hand-writing classification, Isomap's usage to reduce dimensionality of Russell 3000 index constituents is fairly limited.

Usage of Isometric mapping, which is based on geodesic distances, is not very intuitive for financial instruments[2]. Financial instruments prices are stochastic and the relationship between the prices are more intuitively explained using information metric in probabilistic space.

1.1 Contributions

While our approach is inspired by Huang, Kau & Peng in August 2016 [6]our research makes the following contributions to the literature:

1) We use daily end of day prices of Russell 3000 index, which is the 3000 largest US traded stocks, consituents over 30 years instead of using phase space reconstruction method to create prices from an index.

- *N. Bahadur is with the Department of Data Science, Worcester Polytechnic Institute, Worcester, MA, 01609. E-mail: nbahadur@wpi.edu*
- *K. Gajamannage is with the Department of Mathematical Sciences, Worcester Polytechnic Institute, Worcester, MA, 01609. E-mail: kdgajamannage@wpi.edu*
- *R. Paffenroth is with with the Department of Mathematical Sciences, Worcester Polytechnic Institute, Worcester, MA, 01609.*

Manuscript revised October 30, 2017.

1. A market capitalization weighted equity index maintained by the Russell Investment Group that seeks to be a benchmark of the entire U.S. stock market. More specifically, this index encompasses the 3,000 largest U.S. traded stocks, in which the underlying companies are all incorporated in the U.S.

2. Financial instruments are stocks, bonds, commodities, derivatives etc.

2) We carefully compare and contrast both geodesic distance approaches and information metric distance approaches over 30 years of data.
3) Further, when intra-day data is available, we zoom into the window around known crashes to carefully analyze changes in dimensionality.

1.2 BACKGROUND

Using daily closing price of CSI (China Stock Index) 800 and S&P 500 during 2005-2015 De Angelis & Dias in 2014 [7] show how data points fit a probabilistic space better than Euclidean distance matrix. The paper proposes an information-metric based manifold learning method to extract the attractor manifold embedded in the reconstructed phase space. The authors use this technique because the adjacency relationship between financial data points are not entirely dependent on geometric relationships. To extract the underlying manifold in dynamic financial systems, the authors used Phase Space Reconstruction (PSR), a high-dimensional phase space is reconstructed from the observed financial time series. Then, Information Metric Manifold Learning (IMML) method is used to extract the manifold embedded in the reconstructed phase space.

Huang & Kau in 2014 [8] study 2006-2010 annual financial data of 205 small and medium-sized companies from China using Information metric distances. The authors find kernel entropy manifold learning technique based on information metric improves the accuracy of financial early warning but also provided objective criteria for explaining and predicting the Chinese stock market volatility. Huang, Kau & Peng in August 2016 [6] uses daily closing price of CSI 800 and the S&P 500 Index during 2005-2015 to build non-linear manifold learning technique for early warnings in financial market. The authors use Kullback-Leibler Divergence [9] as a measure of dissimlaritly to find the manifold.

Zhong and Enke in 2016 [10] used the closing price of the SPDR S&P 500 ETF[3], along with 60 financial and economic factors as the potential features to study daily direction (up or down) of SPY. These daily data were collected from 2518 trading days between June 1, 2003 and May 31, 2013. The most important and influential principal components among all the linear combinations of the 60 factors determined using PCA, fuzzy robust principal component analysis (FRPCA) [10], and kernel-based principal component analysis (KPCA) [10] is input into the classifiers to predict the direction of the SPY for the next day. All classification models based trading strategies generated higher returns than the benchmark one month treasury bill strategy. The paper concludes preprocessing is critical and can help improve the performance of many techniques, such as PCA and artificial neural networks (ANN) , while decreasing the complexity of the mining procedure and achieving reasonable accuracy and high risk-adjusted profits.

Erriksson [11] in 2011 estimates intrinsic dimensionality via clusters. The paper claims clustering exploits structure of data to efficiently estimate intrinsic dimension accurately and efficiently, even when the data does not conform to an obvious clustering structure. Moreover, clustering-based estimation allows for a natural partitioning of the data points that lie on separate manifolds of different intrinsic dimension. Because the point data cloud, predominantly, tends to fall into linear clusters and linear separation techniques are successful. The paper conjectures that lack of linearity in point cloud data might limit the ability to effectively separate growth and values stocks. The techniques used to arrive at these conclusions include the use of ggobi (a grand tour data visualization system), isomap (a non-linear data reduction tool), model-based clustering and multiresolution bootstrap resampling.

This paper is structured through five sections. In Section 1, we provide a brief introduction of the literature and the problem definition. We explain MDS, Isomap, and information metric distances in Section 2. Section 3 talks about the datasets. Section 4 presents our experiments with data set. Analysis of financial market crashes are analyzed in Section 5. Finally, Section 6 provides a summary and pointers to future work.

2 PCA, MDS, ISOMAP, AND INFORMATION METRIC DISTANCES

In this section, we provide an overview of PCA, MDS, Isomap and information metric distances. While PCA and MDS are linear dimensionality reduction techniques, Isomap is a non-linear dimensionality reduction technique. In contrast to all aforesaid techniques those using the Eucliedean distances, information metric uses the distance between probability distributions.

2.1 Principal Component Analysis

PCA is used in segregating noise and signals in trading models. Using correlations between features, PCA finds the direction of maximum variance in high dimensional data and projects data onto a new subspace of fewer dimension. Using PCA for dimenionsality reduction, we construct $\boldsymbol{W}^{d\times k}$ that allows us to transform input vector $\boldsymbol{x}$ onto a fewer $\boldsymbol{k}$ dimensional subspace. Let

$$\boldsymbol{x} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_i, \ldots, \boldsymbol{x}_d], \boldsymbol{x} \in \mathbb{R}^{d\times k} \tag{1}$$

where $\boldsymbol{x}_i = [x_{1i}, \ldots, x_{di}]^T \in \mathbb{R}^{d\times k}$ collectively represents all the points in the input dataset. PCA assumes that observed variables $\boldsymbol{y}$ is the result from a linear transformation $\boldsymbol{W}$ of p latent variables

$$\boldsymbol{y} = \boldsymbol{x}\boldsymbol{W}, \boldsymbol{W} \in \mathbb{R}^{d\times k} \tag{2}$$

The new features in the $\boldsymbol{k}$ dimensional subspace, where $(k \ll d)$, is

$$\boldsymbol{z} = [\boldsymbol{z}_1, \ldots, \boldsymbol{z}_i, \ldots, \boldsymbol{z}_k], \boldsymbol{z} \in \mathbb{R}^{k} \tag{3}$$

1) Standardize the input data $\boldsymbol{X}$.
2) Use singular value decomposition to decompose $\boldsymbol{X}$ such that $\boldsymbol{X} = \boldsymbol{V}\boldsymbol{\Sigma}\boldsymbol{U}^T$, where $\boldsymbol{V}$, $\boldsymbol{U}$ are unitary matrices ($V^T = V^{-1}$ and $U^T = U^{-1}$), and $\boldsymbol{\Sigma}$ is a matrix with the same size as $\boldsymbol{X}$.
3) Sort the singular values in descending order and select the k largest singular vectors, where k ¡= d.

3. The SPDR S&P 500 trust is an exchange-traded fund which trades on the NYSE Arca under the symbol SPY. SPDR is an acronym for the Standard & Poor's Depositary Receipts, the former name of the ETF. It is designed to track the S&P 500 stock market index.

4) Using the k largest singular vectors in descending order, construct the projection matrix W.
5) Transform input data set $\boldsymbol{X}$ using projection matrix $\boldsymbol{W}^{d\times k}$.

As the linear nature of PCA, its applicability for non-linear data is limited.

2.2 Multidimensional scaling

Multidimensional scaling is a classic approach that can be efficiently used to compute the rank of the distance matrix of the data. Let

$$\boldsymbol{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_i, \ldots, \boldsymbol{x}_n] \tag{4}$$

where $\boldsymbol{x}_i = [x_{1i}, \ldots, x_{di}]^T \in \mathbb{R}^{d\times 1}$ collectively represents all the points in the input dataset. Then MDS computes eigenvalue decomposition of the scaler product matrix,

$$\boldsymbol{S} = \boldsymbol{X}_c^T \boldsymbol{X}_c, \tag{5}$$

which is also known as Gram matrix, of the centered $\boldsymbol{X}_c$ of $\boldsymbol{X}$.

Here we present MDS by assuming that the pairwise Euclidean distance matrix is given. We transform this distance matrix into Gram matrix, $\boldsymbol{S} = [S_{ij}]_{n\times n}$, in two steps. First, squaring the matrix $\boldsymbol{D}$ and then performing double centering of $\boldsymbol{D}$ using

$$S_{ij} = -\frac{1}{2}\big[d_{ij}^2 - \mu_i(d^2) - \mu_j(d^2) + \mu_{ij}(d^2)\big]. \tag{6}$$

Here, while $\mu_i(d^2)$ and $\mu_j(d^2)$ are mean of i-th row and j-th column of the squared distance matrix, respectively, $\mu_{ij}(d^2)$ is the mean of the entire squared distance matrix. Then, we compute the eigenvalue decomposition of the Gram matrix as

$$\boldsymbol{S} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{U}^T. \tag{7}$$

We rearrange $\boldsymbol{\Sigma}$ and $\boldsymbol{U}$ such that the diagonal of $\boldsymbol{\Sigma}$ represents the descending order of magnitudes of eigenvalues and columns of $\boldsymbol{U}$ represent the corresponding eigenvectors in the same order as eigenvalues in rearranged $\boldsymbol{\Sigma}$. We estimate p dimensional latent variables as

$$\hat{\boldsymbol{X}} = \boldsymbol{I}_{p\times n}\boldsymbol{\Sigma}^{1/2}\boldsymbol{U}^T \tag{8}$$

Here $\hat{\boldsymbol{X}}$ is the d-dimensional embedding of the input data $\boldsymbol{Y}$.

This is a linear method, thus it limits the applicability for non-linear data such as financial data. The MDS dimension is PCA dimension + 2 [12]. Isomap overcomes this problem by employing geodesic distance instead of the Euclidean distance.

2.3 Isomap

Isomap ([13]) creates a graph structure over the input data and utilizes that to create geodesics. Isomap inputs one parameter in two forms k or ϵ. Parameter k represents number of nearest neighbors and search k nearest neighbors for each point, while parameter ϵ searches all the nearest neighbors withing an ϵ distance. Nearest neighbor search is converted into a graph structure by treating points as nodes and connecting each pair of nearest neighbors by an edge having the length equal to the Euclidean distance between them.

The geodesic between two given points in the data is the shortest distance between corresponding nodes measured using the Floyd's algorithm [[14]] [[15]]. We compute the shortest path between all pairs of points. Then, we feed the geodesic distances into the distance matrix $\boldsymbol{D}$.

Algorithm 1 *Isomap algorithm.*
Inputs: Data ($\boldsymbol{X}$), number of nearest neighbors (k).
Outputs: List of p largest singular values ($\lambda_l; l = 1, \ldots, p$) and p-dimensional embedding ($\hat{\boldsymbol{X}}$).

1: For each point in $\boldsymbol{X}$, choose k nearest points as neighbors [16].
2: Consider all the point in $\boldsymbol{X}$ as nodes and if any two nodes are chosen to be neighbors in 1, calculate Euclidean distance between them $\boldsymbol{D} = [d_{ij}^2]_{n\times n}$; where $d_{ij} = \|\boldsymbol{x}_i - \boldsymbol{x}_j\|_2$ and n is the order of the high-dimensional space. This step converts the dataset into a graph.
3: For each pair of nodes in the graph, find the points $\mathcal{G} = \{\boldsymbol{x}_i | i = 1 \ldots, k\}$ in the shortest path using Floyd's algorithm [14] and assign it to D.
4: Convert the matrix of distances D into a Gram matrix $\boldsymbol{S}$ by double centering [6] using $S_{ij} = -\frac{1}{2}\big[d_{ij}^2 - \mu_i(d^2) - \mu_j(d^2) + \mu_{ij}(d^2)\big]$.
5: Compute its spectral decomposition S using $\boldsymbol{S} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{U}^T$.
6: Finally, estimate p dimensional latent variables as $\hat{\boldsymbol{X}} = \boldsymbol{I}_{p\times n}\boldsymbol{\Sigma}^{1/2}\boldsymbol{U}^T$.

As in MDS, first we formulate the Gram matrix $\boldsymbol{S}$ from $\boldsymbol{D}$ using Eq. (6) followed by computing the eigenvalue decomposition of $\boldsymbol{S}$ using Eq. (7). The latent variables of the input data are revealed by Eq. (8). Isomap ensures non-linear features of the manifold.

2.4 Information Metric Distances

While Isomap classically uses Euclidean distance to create a graph structure where the weights of the edges are Euclidean distances, we can use information metric distances in Isomap too. Now the weight of the edges will be information metric distances. In probabilistic space, information metric distances is estimated by Kullback-Leibler (KL) divergence [9]. Kullback-Leibler (KL) divergence is a measure of dissimilarity in probabilistic space. KL divergence is used to find the low dimensional embedding from high dimensional data. KL divergence captures the change in information beween two stochastic vectors. For discrete probability distribution P and Q, the Kullback-Leibler divergence from Q to P is defined to be

$$\boldsymbol{D_{KL}(P||Q)} = \sum_{\boldsymbol{i=1}} \boldsymbol{P(i) log \frac{P(i)}{Q(i)}}. \tag{9}$$

The KL divergence metric is not symmetrical because $D_{KL}(P||Q) \neq D_{KL}(Q||P)$. Hence we use a transformation that captures the divergence between two probability distirbutions P and Q.

$$h(P,Q) = D_{KL}(P||Q) + D_{KL}(Q||P) \quad (10)$$

$$h(P,Q) = \sum_{i=1} P(i) log \frac{P(i)}{Q(i)} + \sum_{i=1} Q(i) log \frac{Q(i)}{P(i)} \quad (11)$$

P and Q probability distribution are the returns distribution of stocks on two different days.

3 Data

We use the end of day prices of individual stocks that are part of Russell 3000 index. Russell 3000 index is a market capitalization weighted equity index maintained by the Russell Investment Group that consists of 3000 largest U.S. - traded stocks. The following steps were taken to assemble the data for the study:

1) Find out all stock symbol that were part of Russell 3000 index as of October, 2016, the month we started our study.
2) We downloaded the end of day dividend adjusted prices for these tickers from Janurary 2 1986 to September 30 2016 from http://finance.yahoo.com/.
3) The input file for the experiment was created by having dates on the row indexes and stock (ticker) names as the column names.
4) If historical prices are unavailable for a ticker, the ticker is removed from processing.

For intra-day data, we used Wharton Research Data Services (WRDS). The intra-day set was assembled using the following steps:

1) Using the ticker list from above, we downloaded tick[4] data (executed trades) for July and August of 2011. August 8th and 9th of 2011 had sudden change in dimensionality.
2) The input file for the experiment was created by using the last trade executed price at every minute interval between 09:32:00 and 16:00:00, the hours the U.S. stock markets are open. The first 2 minutes of trading alleviates the trading catalysts created by overnight news and supply, demand imbalance before the stock market officially open.
3) If historical prices were unavailable for a ticker, the ticker is removed from processing.
4) The input file for the experiment had date and time stamps in hours:minute:seconds. The time stamps are row indexes and stock symbols are column headers.

4 Experiments

We use Russell 3000 constituents as a proxy of financial market. Using End of Day (EOD) prices for Russell 3000 index constituents from Janurary 2 1986 to September 30 2016 as input, we run various linear and non-linear algorithms to determine the dimensionality of Russell 3000. For all experiments, we use a 60 days moving window of daily log returns over 30 years. We run PCA, MDS, and Isomap algorithm on this 60 days log returns data set for all Russell 3000 index constituents and determine the dimensionality time series of financial market.

Since real data is not low dimensional the singular values are never exactly 0. To address this we use 2 different thresholds: first we use largest singular values till we get 90% variance tand second only singular values that are greater than 1% of the cumulative singular values.

4.1 90% Variance

A large number of instruments explain very little variance in Russell 3000. We believe these instruments do not contribute strongly towards explaining variance in the financial market. Consequently, we do not only consider instruments that explain the most variance but also use a threshold where we ignore instruments variance when the cumulative variance is a threshold of 90% variance. The dimensionality is the number of instruments that contribute to 90% of index variance. Detailed algorithm is listed in Algorithm 2. The time series of PCA, MDS, and Isomap[5] dimensionality is plotted below.

Algorithm 2 *Dimensionality using 90% variance.*
Inputs: Data ($\boldsymbol{X}$), number of nearest neighbors (k), threshold ($t = 90\%$), and window size ($w = 60$).
Output: p, the number of largest squared singular values that explains 90% of variance in ($\boldsymbol{X}$).

1: Initialize data frame(df) with all returns $X = \{X_1, ..., X_N\}$
2: **for** $i \in \{1, \ldots, N - w\}$ **do**
3: Set $df_i = df[i, i + w]$ ▷ $w = 60$ rows of data.
4: Calculate p_{pca} for dimensionality reduction.
5: Use SVD to decompose $\boldsymbol{df_i}$, where $\boldsymbol{df_i} = \boldsymbol{V\Sigma U}^T$ [1].
6: Sort $diag(\boldsymbol{\Sigma})$ in descending order, where $\boldsymbol{\sigma}_i$ are singular values.
7: Calculate $\boldsymbol{\sigma_{sum}} = \sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} {\sigma_i}^2$
8: Calculate $\sigma_{i\%}$, where $\sigma_{i\%} = \frac{{\sigma_i}^2}{\sigma_{sum}}$
9: Dimensionality p, is the value of l where $\sum_{\boldsymbol{l=1}}^{\boldsymbol{w}} \sigma_{i\%} \geq t(90\%)$
10: Calculate p_{mds} using MDS for dimensionality reduction. Repeat steps 5 to 9.
11: Calculate p_{isomap} using Isomap for dimensionality reduction, where $k = 10$. Repeat steps 5 to 9.
12: **end for**

As illustrated in Figure 1, non-linear dimensionality is much lower than linear dimensionality over the 30 year period. The dimensionality time series captures large drops in S&P 500 time series over 30 years.

4.2 Ignore bottom 1% Variance

Using the same set of linear and non-linear techniques, we run another experiment where instead of cumulative

4. A tick is a measure of the minimum upward or downward movement in the price of a security.

5. 10 nearest neighbors were used for this plot.

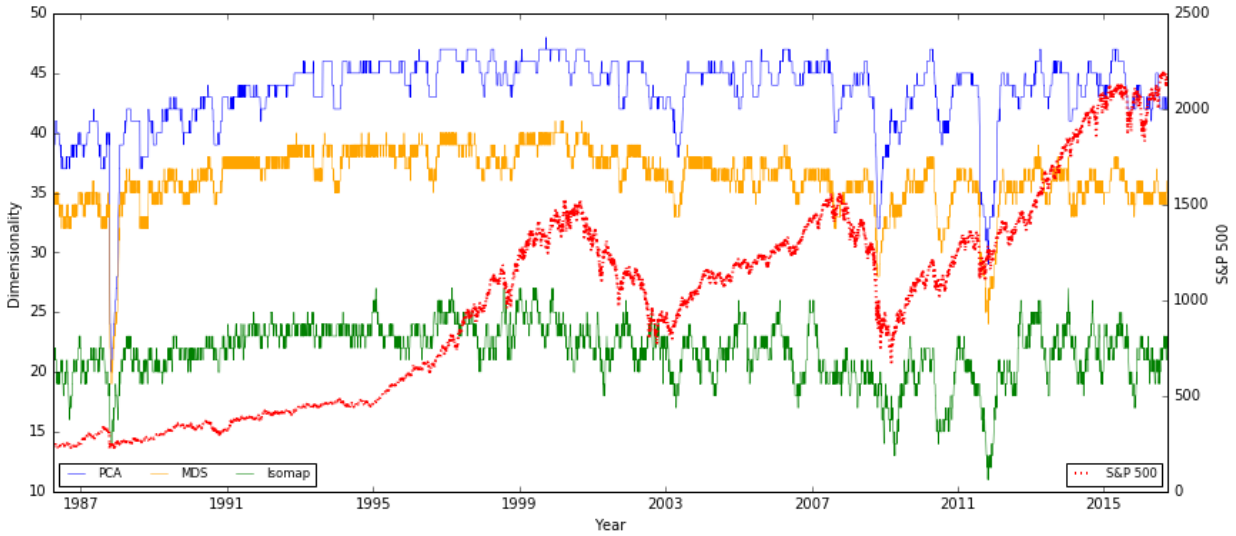


Fig. 1: The dimensionality time series, using Euclidean distances as a dissimilarilty measure, shows the drop in dimensionality when S&P 500 has large drops, when financial markets are under stress. 90% of the cumulative variance is used to calculate dimensionality. Non-linear(Isomap) dimensionality of the market is much lower than linear (MDS/PCA) dimensionality.

variance we consider only index constituent that explained more than 1% of total variance. Detailed algorithm is listed in Algorithm 3.

Algorithm 3 *Dimensionality using large singular values.*
Inputs: Data ($\boldsymbol{X}$), number of nearest neighbors (k), threshold ($t = 1\%$), and window size ($w = 60$).
Output: p, the number of singular values greater than equal 1% .

1: Initialize data frame(df) with all returns $X = \{X_1, ..., X_N\}$
2: **for** $i \in \{1, \ldots, N - w\}$ **do**
3: Set $df_i = df[i, i + w]$ ▷ $w = 60$ rows of data.
4: Calculate p_{pca} for dimensionality reduction.
5: Use SVD to decompose $\boldsymbol{df_i}$, where $\boldsymbol{df_i} = \boldsymbol{V\Sigma U}^T$ [1].
6: Sort $diag(\boldsymbol{\Sigma})$ in descending order, where $\boldsymbol{\sigma}_i$ are singular values.
7: Calculate $\boldsymbol{\sigma_{sum}} = \sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} \sigma_i$
8: Dimensionality $p = \sum \sigma_{i\%}$, where

$$\sigma_{i\%} = \begin{cases} 1 & : \text{if } \frac{\sigma_i}{\sigma_{sum}} \geq 1\% \\ 0 & : \text{otherwise,} \end{cases} \tag{12}$$

9: Calculate p_{mds} using MDS for dimensionality reduction. Repeat steps 5 to 8.
10: Calculate p_{isomap} using Isomap for dimensionality reduction, where $k = 10$. Repeat steps 5 to 8.
11: **end for**

As illustrated in Figure 2, non-linear dimensionality is still lower than linear dimensionality over the 30 year period. Because both PCA and MDS are linear techniques, the change in their dimensionality move in tandem. However, the non-linear dimensionality of the market is significantly lower.

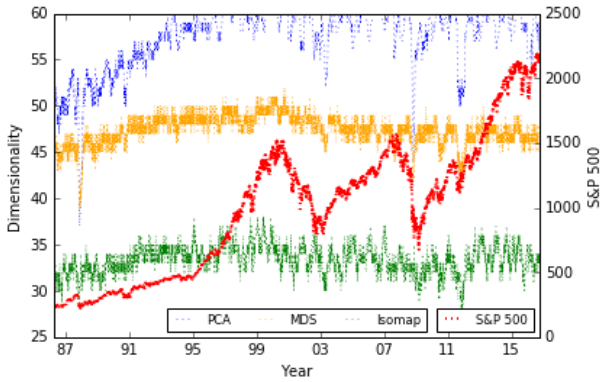


Fig. 2: The dimensionality time series is calculated by using large singular values that contribute at least 1% of cumulative variance show drop in dimensionality when S&P 500 drops significantly, a symptom of stress in financial market. Consistent with Figure 1, Isomap dimensionality is lesser than MDS/PCA.

4.3 Isomap with different k's - 90% variance

While varying the nearest neighbor parameter in Isomap, consider number of instruments 90% of variance is explained.

As illustrated in Figure 3, Isomap dimensionality fluctuates more as k increases. With different k, the wild variations in dimensionality is due to use of Euclidean distance as a measure of dissimilarity.

Moreover, PCA, MDS, and Isomap elbow shapes illustrates that financial market has a well formed structure with approximately 50 dimensions. The steeper drop in MDS is compared to PCA and then that in MDS is compared to Isomap further solidifies our hypothesis that financial market has less than 50 dimensions, and non-linear techniques yield a smaller dimensionality. During stressed mar-

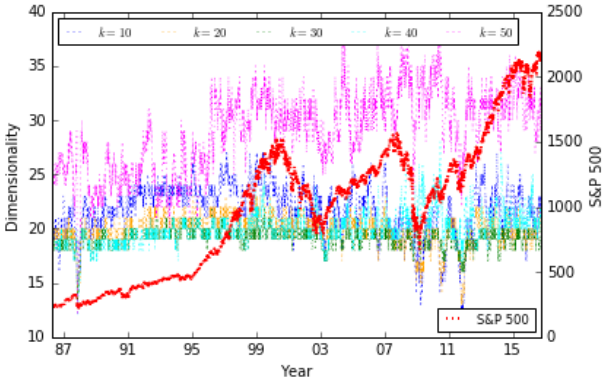


Fig. 3: The non-linear dimensionality time series is sensitive to number of neighbors (k's). We show how dimensionality fluctuates with different k's, where k = 10, 20, 30, 40 and 50. The variance in dimensionality decreases as k increase from 10 to 30 and then variance in dimensionality starts increasign when k increases from 30 to 50.

ket conditions the dimensionality drops drastically reducing the diversification benefit. Our finding is consistent with [17] where the authors study diversification benefits of 5 developed markets and find that for the US, even to be confident of reducing 90% of diversifiable risk 90% of the time, the number of stocks needed on average is about 55. However, in times of distress it can increase to more than 110 stocks.

4.4 Information Metric - KL Divergence

Using the 30 year dataset we determine the dimensionality time series using KL divergence. As in the geodesic case, we use similar threshold of 90%. As illustrated in Figure 4, the change in dimensionality is lot smoother than what we observed using geodesic distance as a measure of dissimilarity.

Unlike Isomap with Euclidean distances [3], Isomap dimensionality with KL divergence [5]fluctuates less k increases from 10 to 60. The Isomap temporal dimensionality is much more stable.

TABLE 1: For bin width 0.009, there are 3853 bins in the data set with >500 returns per bin. For granular bin width of 0.001, there are 558280 bins with $0-10$ returns.

Bin Width	0-10	11-50	51-100	101-500	> 500
0.001	558280	9850	260795	39192	107
0.002	200259	26401	173570	41527	107
0.003	105587	32650	124889	39095	107
0.004	66200	34750	96337	35082	191
0.005	48425	35290	79552	29891	642
0.006	33764	35225	65940	26417	1446
0.007	23583	33938	56174	23499	2342
0.008	18227	32312	49205	21122	3166
0.009	17257	30711	45506	18953	3853

Varying bin widths did not have any noticeable change to dimensionality time series. As illustrated in Figure 6, we also test the stability of our process by using different bin width for discretizing our returns across 30 years. We find no evidence that our process is not stable.

4.5 Intraday Dimensionality

Using Intraday prices between 09:32:00[6] and 16:00:00 we determine the number of singular values required to explain 90% of variance.

Further, analyzing Russell 3000 index constituents intraday prices at 1 minute interval from Janurary 2, 2009 to March 31, 2009 Figure 7 and from July, 2011 to August, 2011 Figure 8, we observe that Isomap, a non-linear dimensionality reduction technique, give us a lower dimensionality than that of linear techniques such as PCA and also capture large drops in financial markets.

Moreover, we also find that Isomap, a non-linear technique, dimensionality of the financial market fluctuates less than PCA, a linear technique, dimensionality.

5 Crash Analysis

We analyze large dimensionality changes over the 30 years to understand if dimenionality change could have predicted the drop in financial market and avoid huge losses.

Black Monday happened on October 19, 1987 when S&P 500 dropped 20.47% (from 282.7 to 224.84). The change in dimensionality in both cases geodesic distances and information metric divergence happened on the next day as illustrated in Figure 1 and Figure 4. The non-linear dimensionality in cases of geodesic distance Figure 9 changed from 23 to 19, whereas information metric dimensionality Figure 10 dropped from 22 to 16. However, in case of information metric divergence the subsequent changes in dimensionality was less.

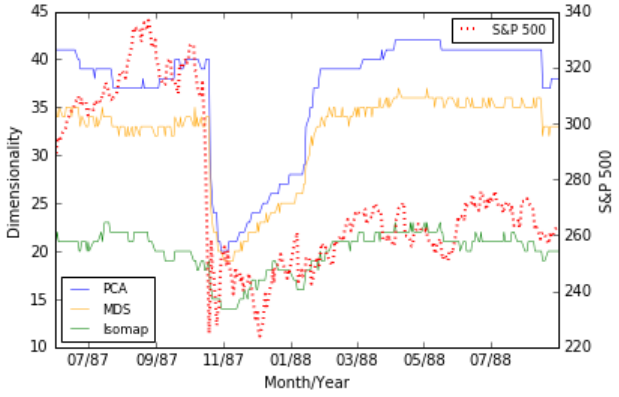


Fig. 9: Black Monday, October 19, 1987 using Euclidean distances

6. We avoid the price fluctuations during opening 2 minutes of trading on exchanges.

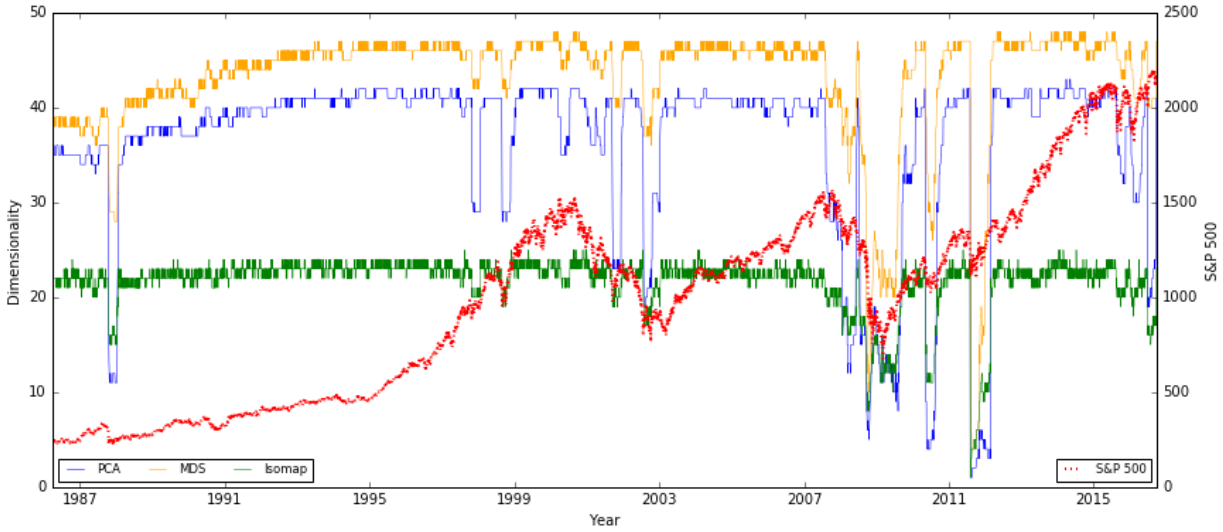


Fig. 4: We use KL divergence, as a dissimilarilty measure, to plot dimensionality time series. Not only large drops in S&P 500 is reflected in PCA, MDS and Isomap dimensionality time series, but KL divergence is also able to capture smaller drops in S&P 500 index. To approximate dimensionality, we use squared values of largest singular values values that cumulatively explain 90% of the variance.

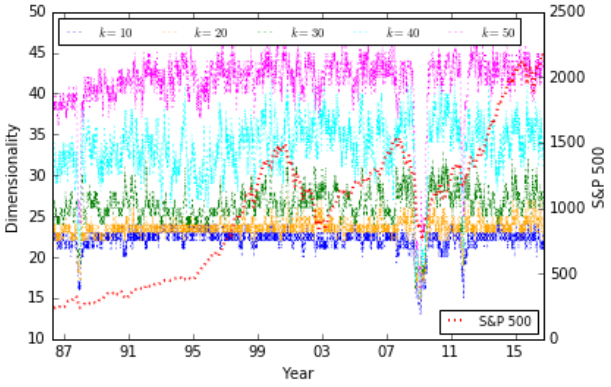


Fig. 5: When KL divergence as a dissimilarity measure with different values of k in Isomap the dimensionality time series varies less comapred to Euclidean distances matrix in Isomap. As k increases from 10 to 20, there is no change in variance. However, as k continues to increase to 30, 40 and 50 the variance of dimensionality time series increases at a faster pace.

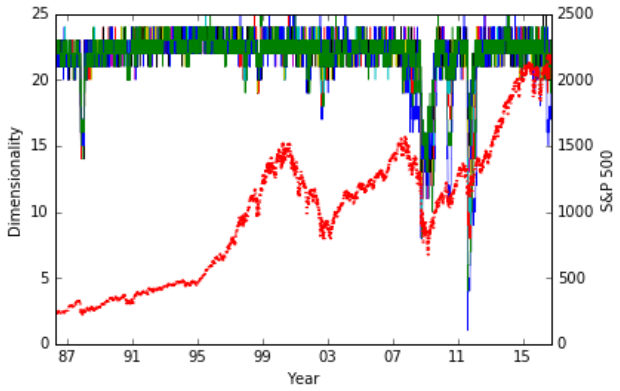


Fig. 6: The plot of varying bin widths from 0.001 to 0.009 to create discretized distribution of returns and then using information metric divergence as a dissimilarity measure with Isomap is stable. This shows that bin widths has insignificant effect on determining non-linear dimensionality time series.

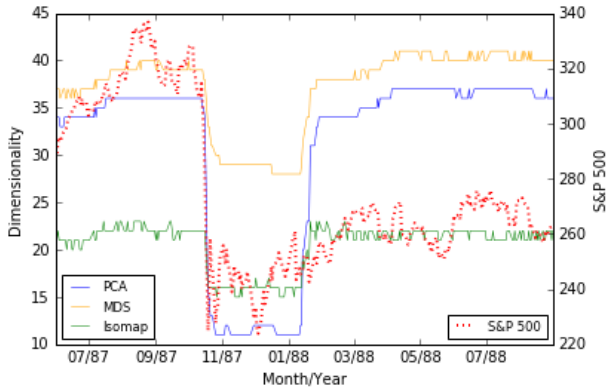


Fig. 10: Black Monday, October 19, 1987 using KL divergence

Early 1990's recession when Iraq invaded Kuwait S&P dropped from 359.54 to 304 from July 1, 1990 to end of October 30, 1990. However, the drop was gradual during the period. Hence, the change in dimensionality in both cases, geodesic distances and information metric divergence, was insignificant. The non-linear dimensionality Figure 11 in cases of geodesic distance oscillated between 26 and 28, whereas information metric dimensionality Figure 12 oscillated between 21 and 24. The recession lasted barely 8 months and the change in dimensionality indicates that there was minor perturbations in the underlying features of the financial market as opposed to significant dislocations.

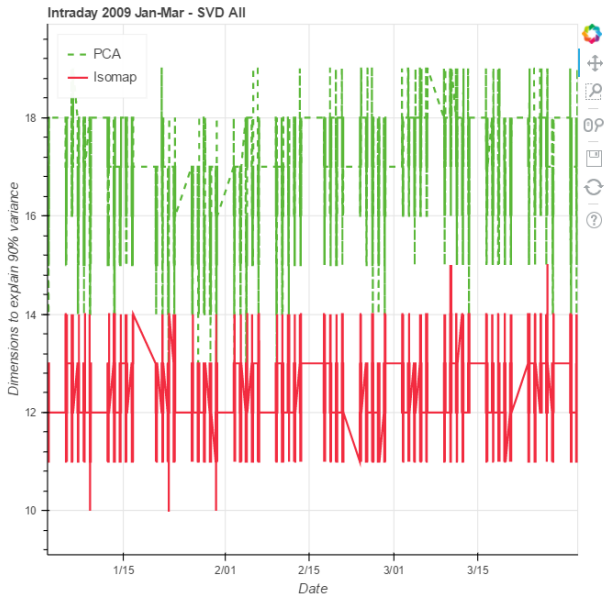


Fig. 7: The intraday dimensionality time series for Russell 3000 index constituents determined by PCA and Isomap during January and February of 2009. Prices at the end of every 1 minute interval during hours when the market was open was used. For Isomap $k = 10$ was used.

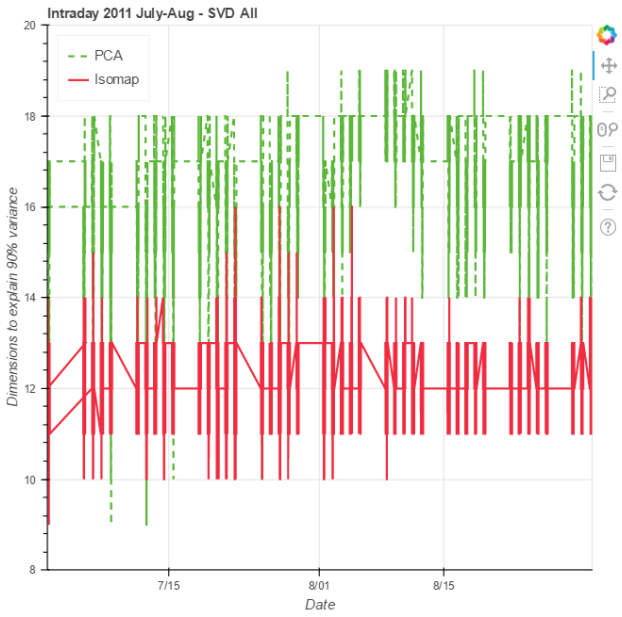


Fig. 8: Prices at the end of every 1 minute interval during hours when the market was open was used. The intraday dimensionality time series for Russell 3000 index constituents determined by PCA and Isomap. For Isomap $k = 10$ was used.

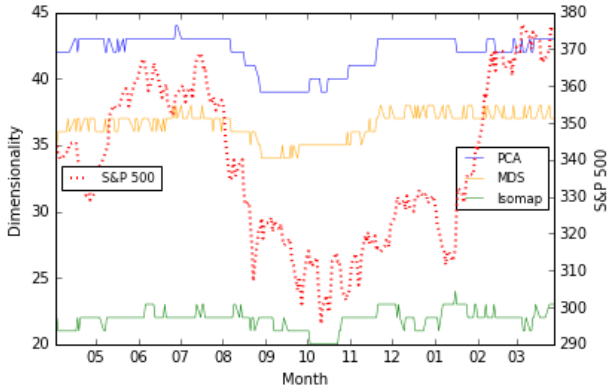


Fig. 11: Iraq Kuwait war, July 1990 using Euclidean distances

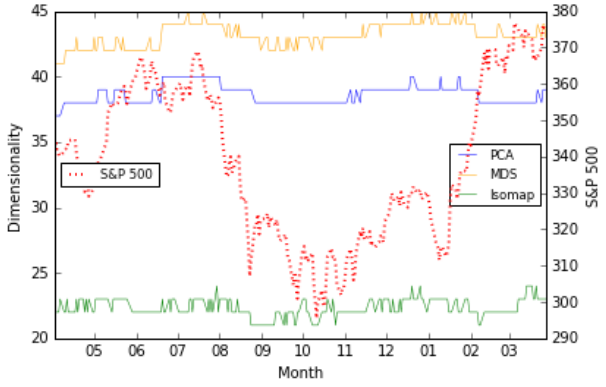
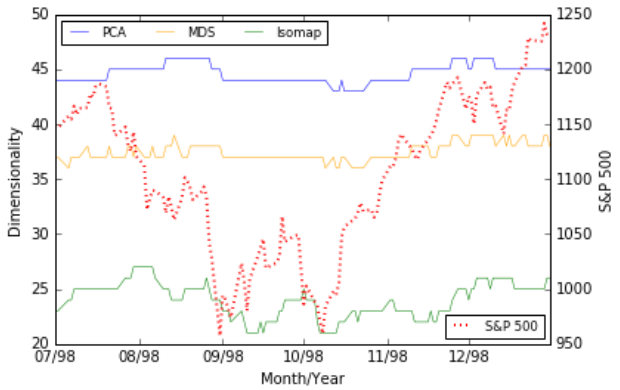


Fig. 12: Iraq Kuwait war, July 1990 using KL divergence

Asian Financial Crisis in July 1997 was localized to South East Asian countries. This was also the time of *dot com boom* in US stock market. While the tremor was felt with the collapse of Long Term Financial Capital (LTCM) but there was no noticeable change in financial market dimensionality. Between June of 1997 and August of 1997, the nonlinear dimensionality Figure 13 in cases of geodesic distance oscillated between 28 and 31, whereas information metric dimensionality Figure 14 that oscillated between 21 and 24. ***Mini crash*** in October 27, 1997 followed the financial crisis where S&P 500 dropped as illustrated in Figure 1 and Figure 4 from 941.64 to 876.99, but recovered the next day. While volatility in S&P 500 index was higher, it was range bound. Consequently, the dimensionality adjusted toward the lower end of the ranges mentioned above and hovered around there before rising higher with the technology boom[7].



7. This was between 1997 and 2000.

Fig. 13: Asian Financial Crisis, July 1997 using Euclidean distances

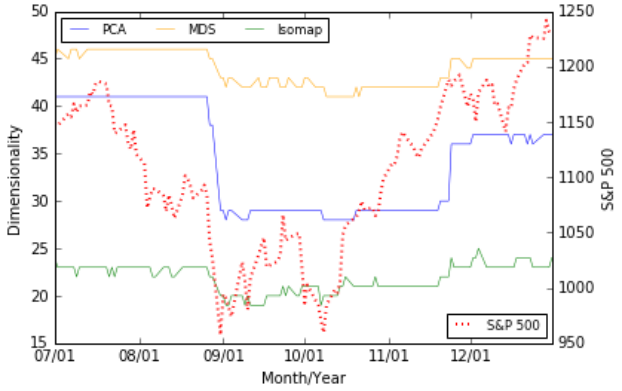


Fig. 14: Asian Financial Crisis, July 1997 using KL divergence

Dot com bubble burst in 2001 Figure 15 and Figure 16. Further, the 9/11 terrorist attack stressed the market. Even though the technology weighted NASDAQ[8] index rose 85.6% in 1999, S&P 500 inde only rose 19.5%. Hence when the dot com era companies collapsed the resultant change in dimensionality of the financial market was muted.

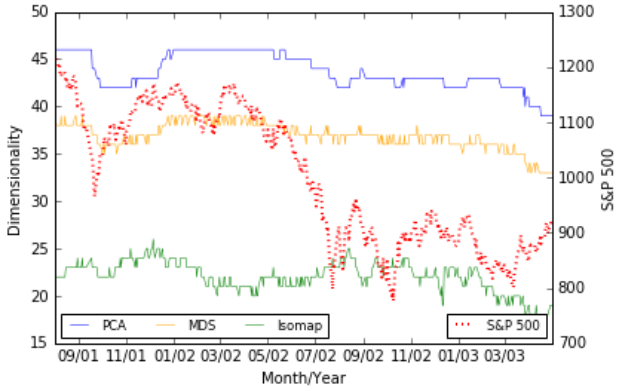
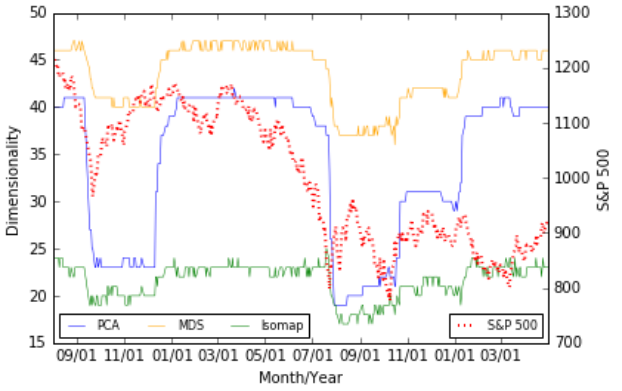


Fig. 15: Dot com crash and 9/11 stressed the market. Using Euclidean distances, we show how the time series of dimensionality changes with S&P 500, indicated in red.



8. The NASDAQ Composite is a stock market index of the common stocks and similar securities (e.g. ADRs, tracking stocks, limited partnership interests) listed on the NASDAQ stock market.

Fig. 16: Shows the effect of Dot com crash and 9/11 on the time series of dimensionality. When we use KL divergence as a dissimilarity measure the drop in dimensionality is large and without lag.

Financial and Banking Crisis of 2007-2008 was the greatest recession after the great depression of 1929. While the subprime started collapsing in 2007, the liquidity crisis started around August 7, 2007 and then subsequently got amplified with Lehman Brother[9] collapse in September 15, 2008. This was a significant market dislocation where all major indexes dropped more than 20%. The effect of this was first felt in the gradual drop in dimensionality as illustrated in Figure 17 and Figure 18 in 2007 from 22-24 range to 19-20 range and then the slide ampified after the Lehman crash to 9-10 range. Apparently, small market dislocations manifest as localized disturbances in financial market manifold without affecting the dimensionality drastically, but large market shifts completely alter the market dimensionality.

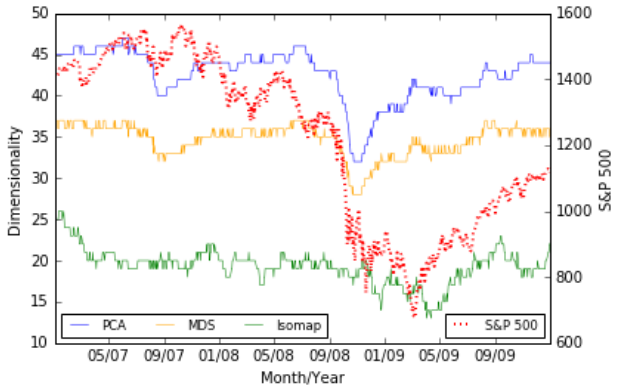


Fig. 17: The effect of Financial and Banking Crisis of 2007-2008, using Euclidean distances as a dissimilarity measure, on dimensionality time series show changes in linear dimensionality is more amplified than change in non-linear dimensionality. The slant in Isomap indicates that the non-linear dimensionality drops before linear dimensionality drops.

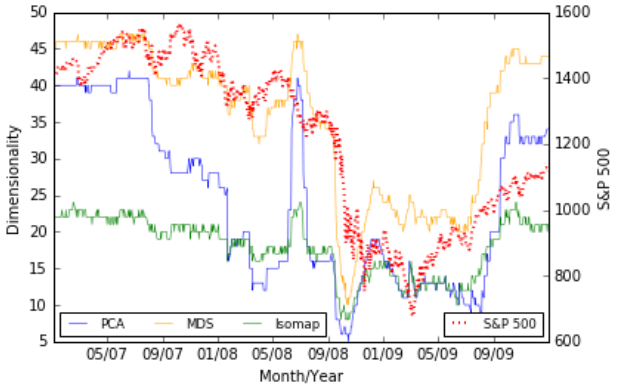


Fig. 18: When KL divergence as a dissimilarity measure, changes in dimensionality time series during Financial and Banking Crisis of 2007-2008 exhibit contemproneous drop in dimensionality along with S&P 500.

9. Lehman was fourth-largest ,158 year, investment bank in the United States doing business in investment banking, equity and fixed-income sales and trading.

Greek Debt Crisis is the sovereign debt crisis faced by Greece following the financial crisis of 200708. Tax increases, spending cuts led to financial losses and social unrest. On August 8th, 2011 Athens stock market index dropped 1000 triggering a 6.67% drop in S&P 500. The effect of the large drop is illustrated by change in time series of dimensionality in Figure 19 and Figure 20. While the change in dimensionality using Euclidean distance approach is insignificant, the change in dimensionality using KL divergence is very prominent.

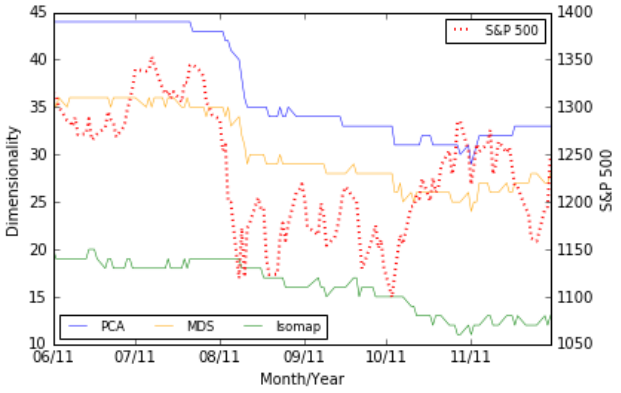


Fig. 19: Although S&P 500 index (red line) drops 1000 points, the dimensionality time series during Greek Debt Crisis of August 2011, using Euclidean distances, changes gradually. Moreover, there appears to be a lag before dimensionality time series starts dropping.

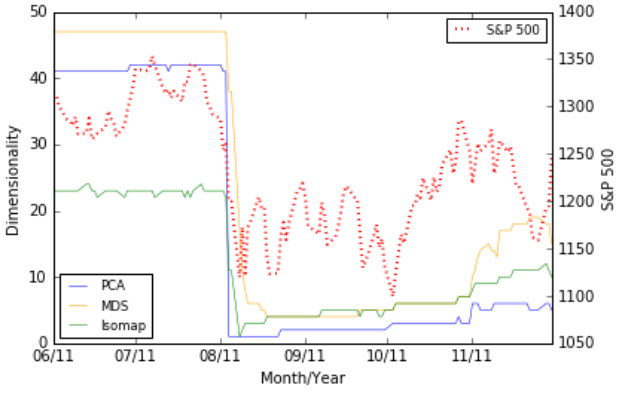


Fig. 20: The drop in dimensionality time series during Greek Debt Crisis of August 2011, when KL divergence is used, is contemporaneous. Moreover, the magnitude of dimensionality change appears to instantenous and apropos to 1000 point drop in S&P 500 index.

6 Conclusion

Using both linear and non-linear dimensionality reduction techniques, and euclidean distance and KL divergence, we observe that under stressed market conditions dimensionality of financial market reduces drastically. In fact, the reduction is far more severe when non-linear dimensionality technique is used, as opposed to linear dimensionality reduction technique. Further, as financial market conditions return to normality the instrinsic dimentionality of the market returns to its long term historical level depending on the technique used. Surprisingly, inspite of all the innovations and technological advances in trading, we find that the instrinsic dimensionality of the market has remained stable.

Change in dimensionality is an excellent metric to detect large drops in financial markets as illustrated by the dimensionality time series during Black Monday crash in October 1987, Financial and Banking Crisis during 2007-2009, and Greek Debt Crisis in August 2011.

Additionally, we find in our crash analysis that when Kullback Liebler divergence Figure 18 measure is used, instead of geodesic Figure 17 distances, in Isomap the dimensionality is more sensitive to drops in S&P 500 and precedes the large drops in financial markets.

In the next iteration we plan to use of intra-day (higher frequency) trades and quotes data from NYSE. Additionally, we will look into different asset classes such as Futures and Option makets to study if the change in dimensionality of these asset classes are contemproneous or lags large drop in financial markets. We want to further analyze if large changes in dimensionality of markets can be detected by change in trading volume of financial instruments.

References

[1] I. Jolliffe, *Principal component analysis*. Wiley Online Library, 2002.
[2] T. F. Cox and M. A. Cox, *Multidimensional scaling*. CRC press, 2000.
[3] M.-H. Yang, "Extended isomap for pattern classification," in *AAAI/IAAI*, 2002, pp. 224–229.
[4] M. Balasubramanian and E. L. Schwartz, "The isomap algorithm and topological stability," *Science*, vol. 295, no. 5552, pp. 7–7, 2002.
[5] J. B. Tenenbaum, V. De Silva, and J. C. Langford, "A global geometric framework for nonlinear dimensionality reduction," *science*, vol. 290, no. 5500, pp. 2319–2323, 2000.
[6] Y. Huang, G. Kou, and Y. Peng, "Nonlinear manifold learning for early warnings in financial markets," *European Journal of Operational Research*, vol. 258, pp. 692–702, 2016.
[7] L. De Angelis and J. G. Dias, "Mining categorical sequences from data using a hybrid clustering method," *European Journal of Operational Research*, vol. 234, pp. 720–730, 2014.
[8] Y. Huang and G. Kou, "A kernel entropy manifold learning approach for financial data analysis," *Decision Support System*, vol. 64, pp. 31–42, 2014.
[9] K. M. Carter, R. Raich, W. G. Finn, and A. O. Hero, "Information-geometric dimensionality reduction," *IEEE Signal Processing Magazine*, vol. 28, pp. 89–99, 2011.
[10] X. Zhong and D. Enke, "Forecasting daily stock market return using dimensionality reduction," *Expert Systems With Applications*, vol. 67, pp. 126–139, 2016.
[11] M. Crovella and B. Eriksson, "Estimation of intrinsic dimension via clustering," *BU/CS Technical Report 2011-12*, 2011.
[12] N. Krislock and H. Wolkowicz, "Euclidean distance matrices and applications," http://www.math.uwaterloo.ca/ hwolkowi/henry/reports/EDMhandbook.pd p. 38, 2010.
[13] J. A. Lee and M. Verleysen, *Nonlinear dimensionality reduction*. Springer Science & Business Media, 2007.
[14] R. W. Floyd, "Algorithm 97: shortest path," *Communications of the ACM*, vol. 5, no. 6, p. 345, 1962.
[15] T. H. Cormen, *Introduction to algorithms*. MIT press, 2009.
[16] J. H. Friedman, J. L. Bentley, and R. A. Finkel, "An algorithm for finding best matches in logarithmic expected time," *ACM Transactions on Mathematical Software (TOMS)*, vol. 3, no. 3, pp. 209–226, 1977.
[17] A. Vitali and T. Francis, "Equity portfolio diversification: how many stocks are enough? evidence from five developed markets." http://www.utas.edu.au/economics-finance/research/, p. 44, 2014.