Dimension Estimation of Equity Markets

Nitish Bahadur, Randy Paffenroth, and Kelum Gajamannage

Abstract—Financial markets are comprised of many instruments, are complex, and are constantly changing. However, it is interesting to consider what, if any, commonalities persist in markets over long time horizons. In particular, one can study financial markets under nominal and stressed market conditions, and attempt to discern which market parameters remain invariant and which market parameters change. Herein, we study financial markets from the perspective of low-dimensional manifolds that capture the inherent characteristics of the high-dimensional data that represent the market performance. Using Russell 3000 constituents, we estimate intrinsic dimension of the US equity market over 30 years (1986-2016) and analyze those times where the change in dimension is abnormal. In particular, our focus in on novel applications of nonlinear techniques such as Isomap and autoencoders, as opposed to linear technique such as principal component analysis (PCA). Such ideas have many applications, including portfolio diversification and market crash analysis.

Index Terms—Russell 3000, principal component analysis, multidimensional scaling, Isomap, autoencoder.

1 Introduction

Automation, algorithmic trading, and globalization have not only made financial markets more integrated but also reduced the information diffusion lag among diverse financial market centers that are located in Japan, Hong Kong, London, and New York. Over the past 30 years, financial markets have seen several technological innovations such as high speed trading [1], alternate news disseminating platforms such as Twitter [2] and Facebook [3], tax efficient financial instruments such as Exchange Traded Funds (ETF) [4] [5] and cognitive computing [6] that is fueled by big data and machine learning. However, it is interesting to consider if there are any ways in which financial markets have *not* changed over the past 30 years. In particular, herein we are interested in studying the *intrinsic dimension* of equity markets over both short-term and long-term time horizons.

One of the key concepts that we leverage in our work is *intrinsic dimension* [7], [8]. By way of foreshadowing, we observe that financial instruments in markets, such as the Russell 3000, constituents are not *independent*. In particular, under normal market conditions, many financial instruments move in lock step, all moving up or down in unison. For example, it is perfectly natural for two highly related healthcare stock such as CI (CIGNA Corp) and UNH (UnitedHealth Group Inc.) to move up and down in unison, and such *correlations* are the basis of many important ideas in finance [9] [10], such as modern portfolio theory [11]. Accordingly, *intrinsic dimension* can be thought of as *the number of degrees of freedom* in a market, and is often much lower than the number of instruments in the market.

Imagine a simple financial market with only two random instruments from either the financial sector, e.g. Bank of America and Wells Fargo as shown in Fig. 1a, or the technology sector, e.g. IBM and Apple as shown in Fig. 1b. The intrinsic dimension of the simple market can be easily estimated by visualizing the relationship between daily returns of two instruments in the market. If the plots of the two instruments lay on a line, then the intrinsic dimensionality is 1 and having both instruments does not provide any diversification. However, if the two instruments do not lay on a line, then the intrinsic dimenionsality is 2 and the two instruments are linearly independent. However, estimating dimension of real financial market with thousands of instruments is a more challenging problem. To empirically estimate the dimension of the financial market under normal market conditions and stressed market conditions, we use the Russell 3000[1] index that encompasses the vast majority of the US financial market. In particular, as we will detail in the sequel, we estimate dimensions using overlapping windows of 60 days.

Of course, we are not the first to consider the intrinsic dimension of financial markets. Enke and Zhong [12] reduce 60 financial and economic features into 36 features to improve forecasting accuracy of the daily direction of the S&P 500 Index ETF (SPY). Jurczyk et. al [13] use eigenvalue decomposition on the similarity between mean-variance portfolios at different times to find critical transition points within a financial market. Zheng et al. [14] use Principal Component Analysis (PCA) on 10 different Dow Jones economic sector indexes and show that the rate of increase in principle components with short 12-month time windows can be effectively used as an indicator of systemic risk. Billio et. al [15] use PCA and Granger-causality on the monthly returns of hedge funds, banks, broker/dealers, and insurance companies to show how interconnectedness in financial sectors have increased. In fact, Modern Portfolio Theory [16] itself is, in many ways, a statement about market dimensionality through its use of the covariance of the market.

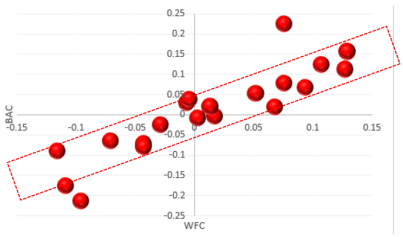
To be precise, we note that dimension estimation (DE) is classically built on top of the closely related idea of di-

- *N. Bahadur is with the Department of Data Science, Worcester Polytechnic Institute, Worcester, MA, 01609. E-mail: nbahadur@wpi.edu*
- *K. Gajamannage is an Assistant Professor at the Department of Mathematics and Statistics, Texas A&M University-Corpus Christi, Corpus Christi, TX 78412. E-mail: kelum.gajamannage@tamucc.edu*
- *R. Paffenroth is the Associate Professor of Mathematical Sciences, Associate Professor of Computer Science, and Associate Professor of Data Science at Worcester Polytechnic Institute, Worcester, MA, 01609. E-mail: rcpaffenroth@wpi.edu*

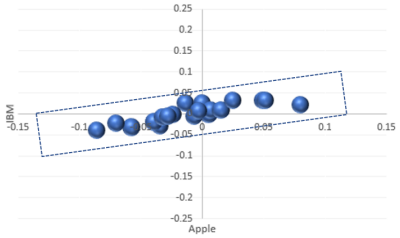
Manuscript revised Sep 25, 2019.

1. A market capitalization weighted equity index maintained by the Russell Investment Group that seeks to be a benchmark of 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.

978-1-7281-0858-2/19/$31.00 ©2019 IEEE
978-1-7281-0858-2/19/$31.00 ©2019 IEEE



(a) Scatter plot of stock prices of a simple financial market comprised of two bank stocks, Bank of America and Wells Fargo. Note these stocks are not well correlated and have higher intrinsic dimensionality.



(b) Scatter plot of stock prices of a simple financial market with two technology stocks, Apple and IBM. Note these stocks are well correlated and have lower intrinsic dimensionality, though their intrinsic dimensionality is not exactly 1.

Fig. 1: Scatter plot of stock prices in September 2008 with different dimensionalities.

mension reduction (DR). To reduce dimension, either linear or nonlinear techniques can be used and linear techniques such as PCA [17] have been widely used in this domain. The key idea of PCA is to construct low-dimensional sub-spaces that preserve as much of the variance in the data as possible and thereby preserve the data's correlation structure. There exist many other principals for determining the intrinsic dimension of data [18] [19] and our ideas here are inspired by that literature. For financial markets, the features are classically log returns, and we take the same approach here. However, as financial markets are quite complicated, we propose to consider more advanced *nonlinear* techniques for dimension estimation.

Over the past 20 years, there has been substantial progress made in nonlinear dimension reduction. Such techniques have already seen applications in studying financial markets [20] [21] and two popular nonlinear dimension reduction techniques that inspire our work are are Isomap [22] [23] [24] and autoencoders [25] [26] [27]. The Isometric Mapping (Isomap) algorithm preserves pairwise *geodesic distances* between data points in the original high-dimensional space and successfully addresses using the limitations of using Euclidean distances in algorithms such as Multi-Dimensional Scaling [28]. Isomap performs nonlinear dimension reduction by considering a graph based distance between data points. For financial markets this can be visualized in Fig. 2b. Isomap chooses the path whose length is shortest which are known in the literature as *geodesics* in the neighborhood graph of the points [28]. An autoencoder [29] [8] is a type of artificial neural network that we will detail in the sequel. However, by way of foreshadowing, we note that autoencoders are a state-of-the-art nonlinear method for computing low-dimensional non-linear latent representation of high dimensional input data using neural networks.

1.1 Contributions

Herein, our approach is inspired by the work in [30] and especially there use of Information Metric Manifold Learning (IMML) [30] to extract the underlying manifold in dynamic financial systems applicable to DE of financial markets. We extend that work in several directions as outlined here:

1) We have preformed a novel analysis of large scale real-world finance data using state of the art nonlinear dimensionality detection techniques.
2) We have developed new techniques for analyzing financial markets that lay at the intersection of neural networks (e.g., autoencoders) and information theory (by way of the Kullback-Leibler divergence).
3) We use end of day prices of Russell 3000 index constituents, which is the 3000 largest US traded stocks, over 30 years instead of using synthetic data, as in [30], to create prices from an index.
4) We compare and contrast Euclidean distance, geodesic distance, and information metric distance based approaches over 30 years of data, and demonstrate an surprising consistency in dimensionalty, apart for market crashes, over this entire time period.
5) In particular, our methodologies can provide a launching pad for advanced portfolio optimization approaches that leverage non-linear dependencies between financial instruments.

1.2 Background

There are two segments of the extent literature that play key roles in our work. First, we have *portfolio diversification*, which is the practice of spreading investments to reduce concentration in a few instruments and reduce investment risk. Of course, two instruments that move the same in the market do not provide any diversification, so as we will discuss in Section 1.2.1 analyzing groups of instruments and their mutual dependencies is important to reduce risk. Second, the use of *dimension estimation* in finance and its relevance in deciding how many instruments to use in building financial portfolios has been an important inspiration for this work. Our survey of dimension estimation literature in finance is summarized in Section 1.2.2 and we note that, with important exceptions, the use of nonlinear techniques such as Isomap and autoencoders for DE in financial markets is less well studied than linear techniques such as PCA.

1.2.1 Portfolio Diversification

The existing portfolio diversification literature can be organized into 3 periods. To begin, in the *pre 2000* period, many simple strategies were developed. For example, [31] suggests that a well-diversified portfolio of randomly chosen stocks must include at least 30 stocks for a borrowing investor and 40 stocks for a lending investor. On the other hand, [32] states this number is between 20 and 50 (Table 8 in [32] - Effect of diversification). Also, [33] argues there is always some diversifiable risk left in the portfolio. Hence, the benefit of holding 50 or more stocks, based on Thaler's

required excess return from imperfect diversification (Table 2 in [33]), is low. *Between 2000 and the financial crisis in 2008*, some authors [34] used the terminal wealth shortfall argument of 20 stock portfolios to conclude that a higher number of stocks are required for diversification. Also, [35] suggested that a portfolio size of 20 is required to eliminate 95% of the diversifiable risk on average. *In the post financial crisis, after 2009*, [36] suggests investors concerned with extreme risk can achieve diversification benefits with a relatively small number of stocks.

However, our motivation is more fundamental: how many independent dimensions are there in the market or a portfolio, and can the number of independent dimensions be effectively estimated? In particular, building diversified portfolios would seem to require as many independent dimensions as possible, in both the linear and non-linear sense. For example a seemingly diversified portfolio that conglomerates such as Amazon, Goldman Sachs, Google, GE, Nike, and Microsoft may appear to be well diversified, but have hidden latent commonalities that increase risk.

1.2.2 Dimension Estimation in Finance

The authors in [24] use Phase Space Reconstruction (PSR) to extract the underlying manifold in dynamic financial systems and leverage Information Metric Manifold Learning (IMML) to build early warning ranges for critical transitions of financial markets. Previously in 2014, Huang & Kou [37] studied 2006-2010 annual financial data of 205 small and medium-sized companies from China using information metric distances. Huang and Kou used kernel entropy manifold learning techniques based on information metric to improve the accuracy of financial early warnings. While use of IMML and the Kullback-Leibler Divergence [38] is inspiring, our research is different on several fronts. First, we use real time series of *Russell 3000* index constituents instead of daily closing prices of the CSI (China Stock Index) 800 and the S&P 500 during 2005-2015. Second, our goal is to estimate the dimension of the market over time using novel non-linear techniques, and track the changes in dimensionality.

Using PCA, fuzzy robust PCA (FRPCA), and kernel-based PCA (KPCA), Zhong and Enke in 2017 [12] show how to predict next day direction of S&P 500 exchange-traded funds. However, our research differs from their work in that we are interested in estimating the dimension of the S&P 500, not predicting S&P 500 movements.

Deep learning has been widely used in Finance [39] [40] [41]. Recently, Heaton, Polson, and Witte [26] used a 4-step process, namely *a*uto-encode, calibrate, validate, and verify, to find optimal weights of the Biotech iShares Nasdaq Biotechnology exchange-traded fund (IBB) constituents to beat IBB returns. In particular, the 5 nodes in the hidden layer in their 2 layer autoencoder is used for ranking stocks. However, as mentioned previously, while the use of an autoencoder to build portfolios is motivating for our work, we focus on the use of autoencoders to estimate intrinsic dimensionalty.

2 Building Blocks

In this section we introduce many of the algorithmic building blocks of our approach. Additional details for these methods can be found in [8], and herein we will focus on the application of these methods in the financial context. Our key object of study are high-dimensional data points arising from the daily log returns of end of day prices of the financial instruments. In particular, let p_{j,t_i} be today's price and $p_{j,t_{i-1}}$ be yesterday's price for an instrument j, then the log return, denoted by R_{j,t_i} is

$$R_{j,t_i} = log\frac{p_{j,t_i}}{p_{j,t_{i-1}}} \tag{1}$$

We can then form a data matrix X with $X_{i,j} = R_{j,t_i}$. In particular, give d days and k instruments we have that

$$\boldsymbol{X} = \begin{bmatrix} \boldsymbol{x}_1 \\ \vdots \\ \boldsymbol{x}_j \\ \vdots \\ \boldsymbol{x}_d \end{bmatrix} \in \mathbb{R}^{d\times k}, \tag{2}$$

where each row of X, denoted by $\boldsymbol{x}_j$, are the log returns across a single day of all instruments.

2.1 Distance Preservation

Classically, the returns in high-dimension space can be mapped into low-dimension space by preserving the Euclidean distance (Fig. 2a), geodesic distance (Fig. 2b), or information metric distance [24] (Fig. 2c) between two instruments.

2.1.1 Euclidean Distance

The straight line distance between two days, as defined by,

$$\|\boldsymbol{x}_i - \boldsymbol{x}_j\|_2 = \sqrt{(\boldsymbol{x}_i - \boldsymbol{x}_j)\cdot(\boldsymbol{x}_i - \boldsymbol{x}_j)} \tag{3}$$

is called the Euclidean distance between those days. It is classically used in linear dimension reduction techniques and we will describe an application of this distance when we discuss MDS in Section 2.2.2, which works by preserving linear spatial distance.

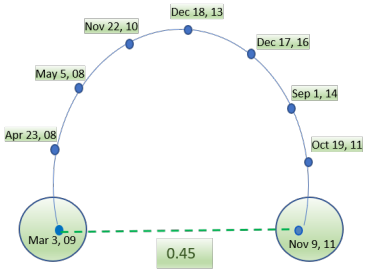
2.1.2 Geodesic Distance

More interestingly, one can also define geodesic distances for financial instruments. In particular, as shown in Fig. 2b, such distances provide an approximation of the distance between instruments *by following a manifold*. Preserving such geodesic distances is paramount to nonlinear dimension reduction technique such as Isomap which we will describe in Section 2.2.3.

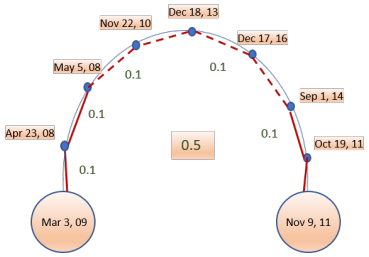
2.1.3 Information Metric Distance

Most importantly for our work, one can also define distance in information theoretic term. In particular, each day $\boldsymbol{x}_i$ can be thought of as a random variable, and the distance between the distributions of two such random variables can be computed. For example, the information metric distances [24] is estimated by using the Kullback-Leibler (KL) divergence [42] [38]. The KL divergence is a measure of dissimilarity between the probability density functions of two random variables, as shown in Fig. 2c, and the KL divergence has been used to find the low dimensional embedding of high dimensional data in financial instruments [24]. In particular, for discrete probability distributions P and Q with m different states, the Kullback-Leibler divergence from Q to P is defined to be

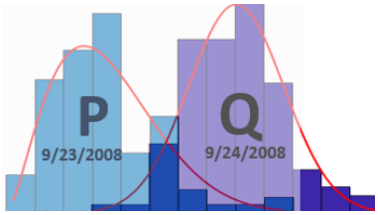
$$D_{KL}(P||Q) = \sum_{i=1}^{m} P(z=i)log\frac{P(z=i)}{Q(z=i)}. \tag{4}$$



(a) Euclidean distance between Mar 3, 09 and Nov 9, 11 is the straight line distance between log returns of the two days in stock market.



(b) Geodesic distance between Mar 3, 09 and Nov 9, 11 is the distance along the curve.



(c) Information metric distance between Sep 23, 08 (P) and Sep 24, 08 (Q) is the distance between the probability distributions of log returns between P and Q.

Fig. 2: Different types of distances.

Note, the KL divergence metric is not symmetric because, in general, $D_{KL}(P||Q) \neq D_{KL}(Q||P)$. Hence we use the a symmetric distance measure based on the KL divergence, $h(P,Q)$, [24] that captures the divergence between two probability distributions P and Q:

$$h(P,Q) = D_{KL}(P||Q) + D_{KL}(Q||P). \tag{5}$$

to measure the distance between the distribution of two instruments.

2.2 Dimension Reduction

Dimension reduction (DR) is the process of reducing the number of features in high-dimensional datasets by obtaining a smaller set of latent variables. DR can be segmented into two categories: linear, such as PCA, and nonlinear, such as Isomap and autoencoders, and we provide a brief overview of these techniques here.

2.2.1 Principal Component Analysis

PCA [43] is quite often used 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 dimensions. Using PCA for dimensionsality reduction of our data matrix X can be accomplished by using the following procedure.

1) Standardize each column of the the input data $\boldsymbol{X}$ by subtracting the mean of each column for every entry of the column dividing by the standard deviation of the column (called the Z-transform [44]).
2) Use singular value decomposition [17] 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 of singular values with the same size as $\boldsymbol{X}$.
3) Sort the singular values in descending order removing all but the k largest, giving $\hat{\boldsymbol{\Sigma}}$, and select the columns of $\boldsymbol{U}$ corresponding to the same k largest singular values, giving $\hat{\boldsymbol{U}}$.
4) The transformed low-dimensional data set can then be written as $\hat{\boldsymbol{X}} = \hat{\boldsymbol{U}}\hat{\boldsymbol{\Sigma}}$.

2.2.2 Multidimensional scaling

Multidimensional scaling [45] [7] is a classic approach that can be efficiently used to compute the rank of the distance matrix of the data. Let

$$\boldsymbol{X_c} = \boldsymbol{X} - \frac{1}{d}\sum_{i=1}^{d} \boldsymbol{x}_i, \tag{6}$$

Then MDS computes the eigenvalue decomposition of the inner product matrix,

$$\boldsymbol{S} = \boldsymbol{X}_c^T \boldsymbol{X}_c, \tag{7}$$

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

Let $\boldsymbol{D}$ be the pairwise Euclidean distance matrix defined by $D_{i,j} = d(\boldsymbol{x}_i, \boldsymbol{x}_j)$. We transform this distance matrix into the Gram matrix, $\boldsymbol{S} = [S_{ij}]_{n\times n}$, in two steps. First, squaring each entry of the matrix $\boldsymbol{D}$, denoted $\boldsymbol{D}^2$, and then performing double centering of $\boldsymbol{D}^2$ using

$$S_{ij} = -\frac{1}{2}\left[D_{ij}^2 - \mu_i(\boldsymbol{D}^2) - \mu_j(\boldsymbol{D}^2) + \mu_{ij}(\boldsymbol{D}^2)\right]. \tag{8}$$

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

$$\boldsymbol{S} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{U}^T. \tag{9}$$

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

$$\hat{\boldsymbol{X}} = \hat{\boldsymbol{U}}\hat{\boldsymbol{\Sigma}}^{1/2} \tag{10}$$

Here $\hat{\boldsymbol{X}}$ is the d-dimensional embedding of the input data $\boldsymbol{X}$, and $\hat{\boldsymbol{U}}$ and $\hat{\boldsymbol{\Sigma}}$ are as defined in Section 2.2.1.

Note, MDS is also a linear method, similar to PCA, hence its applicability for nonlinear data, such as financial markets, is limited. Isomap overcomes this problem by employing geodesic distance instead of the Euclidean distance.

2.2.3 Isomap

Isomap [7] creates a graph structure over the input data and utilizes that to create geodesics. Isomap required a neighborhood of each data point that can be defined in two forms, k or ϵ. The parameter k represents the number of nearest neighbors, while the parameter ϵ searches all the nearest neighbors withing an ϵ distance. The 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 [46] [47]. We compute the shortest path between all pairs of points. Then, we use the geodesic distances into the distance matrix $\boldsymbol{D}$. The full can algorithm can be found in [28], and we paraphrase it here.

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.
2: Consider all the point in $\boldsymbol{X}$ as nodes and if any two nodes are chosen to be neighbors in 1, calculate squared Euclidean distance between them $\boldsymbol{D}^2 = [D_{ij}^2]_{n\times n}$; where $D_{ij} = \|\boldsymbol{x}_i - \boldsymbol{x}_j\|_2^2$ 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 [46] and assign it to $\boldsymbol{D}$.
4: Convert the matrix of distances $\boldsymbol{D}$ into a Gram matrix $\boldsymbol{S}$ by double centering [8], as in MDS, using $S_{ij} = -\frac{1}{2}\left[D_{ij}^2 - \mu_i(\boldsymbol{D}^2) - \mu_j(\boldsymbol{D}^2) + \mu_{ij}(\boldsymbol{D}^2)\right]$.
5: Compute the spectral decomposition S using $\boldsymbol{S} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{U}^T$.
6: Finally, estimate p dimensional latent variables using $\hat{\boldsymbol{X}} = \hat{\boldsymbol{U}}\hat{\boldsymbol{\Sigma}}^{1/2}$ as in MDS.

As in MDS, first we formulate the Gram matrix $\boldsymbol{S}$ from $\boldsymbol{D}$ using Eq. (8) followed by computing the eigenvalue decomposition of $\boldsymbol{S}$ using Eq. (9). The latent variables of the input data are revealed by Eq. (10) and provide a nonlinear projection of the data into a low-dimensional space.

2.2.4 Autoencoder

A detailed discussion of autoencoders can be found in [29]. However, in the interest of completeness, we observe that a simple autoencoder, with one layer hidden, can be defined by way of data $\boldsymbol{x}_i \in \mathbb{R}^{n\times 1}$ passing through the hidden layer, which outputs $\boldsymbol{y}_i \in \mathbb{R}^{m\times 1}$, according to the mapping

$$\boldsymbol{y}_i = f_1(\boldsymbol{W}_1\boldsymbol{x_i} + \boldsymbol{b_1}), \tag{11}$$

where $\boldsymbol{W}_1$ is the weight matrix of the first layer, and, generally, $m < n$. The function f_1 is typically a non-linear activation function such as sigmoid or rectified linear unit (ReLU) [48]. The second layer maps $\boldsymbol{y_i}$ to $\hat{\boldsymbol{x_i}} \in \mathbb{R}^{n\times 1}$ according to

$$\hat{\boldsymbol{x_i}} = f_2(\boldsymbol{W}_2\boldsymbol{y_i} + \boldsymbol{b_2}) = f_2(\boldsymbol{W}_2 f_1(\boldsymbol{W}_1\boldsymbol{x_i} + \boldsymbol{b_1}) + \boldsymbol{b_2}), \tag{12}$$

where $\boldsymbol{W}_2 \in \mathbb{R}^{n\times m}$ and $\boldsymbol{b}_2 \in \mathbb{R}^{n\times 1}$ are the weight matrix and bias vector of the second layer, respectively. The parameters $\boldsymbol{W}_1$, $\boldsymbol{b_1}$, $\boldsymbol{W}_2$, and $\boldsymbol{b_2}$ are found by minimizing a cost function (13) that quantifies the difference between the output $\hat{\boldsymbol{x_i}}$ and input $\boldsymbol{x_i}$ as

$$J(\boldsymbol{W}, \boldsymbol{b}; \boldsymbol{x}) = \frac{1}{2}\|x - \hat{x}\|_2. \tag{13}$$

A *autoencoder* is defined similarly, except that many layers such as (11) and (12) are used.

When the neural network is constrained by a small number of hidden units, the network is forced to learn a compressed representation of the input. However, even when the number of hidden units is large the network can still discover interesting structure by imposing sparsity constraints on the hidden units [49] [8].

$$y_{norm}^i = \frac{y_i}{\sqrt{\sum_{i=1}^k {y_i}^2}} \tag{14}$$

$$J_{sparse}(\boldsymbol{W}, \boldsymbol{b}; \boldsymbol{x}) = J(\boldsymbol{W}, \boldsymbol{b}; \boldsymbol{x}) + \lambda\sum_{i=1}^{k} \|y_{norm}^i\|_1 \tag{15}$$

Using (15) we follow the procedure defined in [8], and use the entries in the lowest dimensional hidden layer y as *singular value proxies*, similar to actual singular values from in the matrix $\hat{\boldsymbol{\Sigma}}$ is PCA, MDS, and Isomap.

3 Dimensionality Estimation Algorithm

Given singular values or singular value proxies [8], we need to define the actual intrinsic dimensionality of the given data X. In particular, for real data, none of the singular values or singular value proxies are exactly 0, so we need to truncate small values based upon some principle. Accordingly, here we use two different analytics:

1) *Greater Than Equal To* 1%: Count the singular values or singular value proxies that are larger than 1% of the sum of all values. The threshold 1% is configurable and the procedure is defined in Algorithm 2)
2) *Up to* 90%: Using the largest singular values or singular value proxies, count the number of values required such that the cumulative value is larger than 90% of the sum of all singular values. Again, 90% threshold is configurable and the procedure is defined in Algorithm 3.

Algorithm 2 *Dimensionality using Greater Than Equal To 1%.*
Inputs: $\{\sigma_1, \sigma_2, ...\sigma_n\}$ - singular values or singular value proxies in case of autoencoder, threshold ($t = 1\%$).
Output: p, the number of singular values greater than equal 1% .

1: Calculate $\sigma_{sum} = \sum_{i=1}^n \sigma_i$
2: Dimensionality $p = \sum I(\sigma_{i\%})$, where

$$I(\sigma_{i\%}) = \begin{cases} 1 & ;\text{if } \frac{\sigma_i}{\sigma_{sum}} \geq 1\%, \\ 0 & ;\text{otherwise.} \end{cases} \tag{16}$$

4 Experiments

Our experiments to estimate dimension are organized in two sections. In Section 4.1 we use log returns of Russell 3000 index constituents as encoded in our data matrix X. Dimension estimated using information metric distance is presented in Section 4.2.

Algorithm 3 *Dimensionality using upto 90%.*
Inputs: $\{\sigma_1, \sigma_2, ...\sigma_n\}$ *- singular values or singular value proxies in case of autoencoder, threshold* ($t = 90\%$).
Output: p*, the number of largest singular values that explains 90% of variance in* ($\hat{\boldsymbol{X}}$).

1: Sort $\{\sigma_1, \sigma_2, ...\sigma_n\}$ in descending order, where $\boldsymbol{\sigma}_i's$ are singular values. Without loss of generality, we assume that the sorted order is $\sigma_1, \sigma_2, ...\sigma_n$.
2: Calculate $\boldsymbol{\sigma_{sum}} = \sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} {\sigma_i}^2$.
3: Calculate $\sigma_{i\%}$, where $\sigma_{i\%} = \frac{{\sigma_i}^2}{\sigma_{sum}}$.
4: Dimensionality p, is the value of l where $\sum_{\boldsymbol{l=1}}^{\boldsymbol{w}} \sigma_{i\%} \geq t(90\%)$.

4.1 Log Returns

We estimate dimension of each 60 day window of daily log returns (Eqn. 1) from 1986 to September 2016 (30 years) moving forward one trading day at a time using PCA, Isomap, and autoencoders for DE. *One important item to note is that not all instruments that we study have existed in the market across this whole time period.* Accordingly, each time window includes instruments that existed at the particular times, and therefore the actual instruments whose dimensionality we compute changes over time. Initial public offerings add new stocks to, merger & acquisitions, bankruptcy and public-to-private market transactions removes stocks from financial markets. For example, Facebook went public on May 2012, Pfizer purchased Wyeth in October 2009 and DELL Technologies became private in October, 2013.

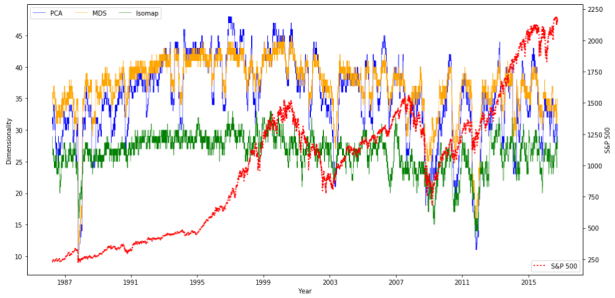


Fig. 3: The dimension is estimated over 30 years using daily log returns of Russell 3000 index constituents. Isomap, MDS and PCA was used for dimension reduction. Dimension estimation presented here used greater than equal to 1% analytic. Note, the intrinsic dimensionality measure by all of these methods is somewhat noisy, but that features, such as crashes, stand out.

As illustrated in Fig. 3, we find that the *intrinsic dimension* of the market over 30 years has been in a narrow range, though it is somewhat noisy. Note, the noise in the dimensionality estimates is not surprising, since the the instruments we study change over time. In fact, given how many features of the market have changed over the 30 years of this study, the stability of the dimensionality is surprising to us. In addition, we find that during crises such as the Black Monday of October 1987, Financial and Banking crisis of 2008, and the Greek Debt crisis of August 2011, the dimension drops but recovers slowly as the market recovers. Although the autoencoder estimated the dimension drops during crisis in financial markets, the estimated dimension time series is much more sensitive compared to that of Isomap.

Unfortunately, in this particular case, the performance of autoencoders is not as good as the competing methods. For example, in Fig. 4 we see that that intrinsic dimension estimated by the autoencoder, using the techniques from [8] are quite noisy.

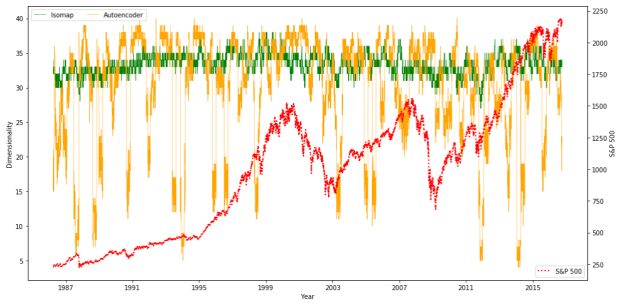


Fig. 4: The dimension is estimated over 30 years using daily log returns of Russell 3000 index constituents. Isomap and autoencoder was used for dimension estimation. Dimension estimation presented here used greater than equal to 1% analytic.

4.2 Information Metric - KL Divergence

In our key result, we estimate the intrinsic dimensionality of the markets using information metric distances [24], specifically the KL divergence, combined with linear and non-linear dimensionality techniques such as autoencoders, Isomap, MDS and PCA.

Consistent with our previous experiment, we find that the *intrinsic dimension* of the market over 30 years has been stable. In fact the dimensionality computed by Isomap is nearly constant, except for dips that exactly coincide with market crashes. During financial crises such as the Black Monday of October 1987, the Financial and Banking crisis of 2008, and the European Sovereign Debt crisis of August 2011, dimension of the market drops and reverts back to its original state as the stress in markets decreases. Our finding is consistent with [50] where the authors study diversification benefits of 5 developed markets.

In particular, we would conjecture that the intrinsic dimensionality of the financial markets, as computed using information metric distances and non-linear methods such as Isomap, can provide indicators of market corrections. *It is interesting to consider whether such techniques are merely indicative of market crashes, or if they are predictive of such events.* We plan to extend these results in the future to answer precisely this type of question.

5 Crash Analysis

Can large dimension changes help investment portfolios avoid large losses? We analyze market dimension from 1986 to 2016 (30 years) to answer the question. On *Black Monday*, October 19 1987, the S&P 500 dropped 20.47% (from 282.7 to 224.84) representing the greatest one-day percentage decline in U.S. stock market history, culminating in a bear market. Dimension of the financial market contemporaneously dropped as illustrated by Fig. 6a.

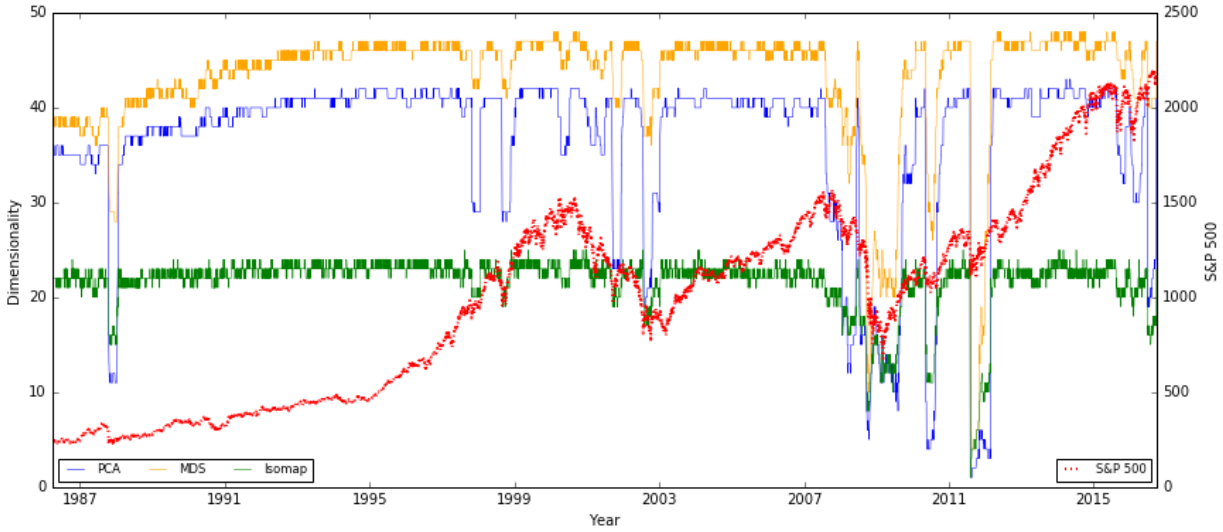


Fig. 5: Dimension is estimated using the information metric distance for 30 years. Bin width of 0.001 is used and PCA, MDS and Isomap is used for dimension reduction. Dimension is calculated using up to 90% analytic. Note that the dimensionality of the market as estimated by Isomap is *almost constant* over a 30 year time-span, except for dips in the dimensionality that exactly coincide with market crashes.

The *Financial and Banking Crisis* of 2007-2009 was the greatest recession after the great depression of 1929. The Subprime mortgage crisis, followed by the liquidity crisis in August 2007 triggered significant financial market dislocation and the S&P 500 index dropped more than 20% over several months. The effect of this was first felt in the gradual drop in dimension, as illustrated in Fig. 6b. In 2007, intrinsic dimension ranged from 22-24 to 19-20, however severe recession caused the intrinsic dimension to crash to the 9-10 range.

The *Europe Sovereign Debt Crisis* of August 2011 started with a debt crisis in Spain and Italy and concerns over France's sovereign bond AAA rating. Slowing economic growth in the United States and it's credit rating downgrade increased volatility of stock market indexes. On August 8th, 2011 the Athens stock market index dropped 1000 points triggering a 6.67% drop in the S&P 500. The effect of the large drop is illustrated by the change in dimension in Fig. 6c. The dimension change estimated by Euclidean distance is more gradual compared to dimension change estimated by information metric distance.

6 Conclusion

Using both linear and nonlinear dimension reduction techniques, and euclidean, geodesic, and information metric distance, we observe that under stressed market conditions, the dimension of financial markets reduces drastically. In fact, the reduction is far more severe when a nonlinear dimension technique is used, as opposed to linear dimensionality reduction techniques. Further, as financial market conditions return to normality the intrinsic dimension of the market returns to its long term historical level depending on the technique used. Surprisingly, in spite of all the innovations and technological advances in trading, we find that the intrinsic dimensionality of the market has remained stable.

Change in dimension is an excellent metric to detect large drops in financial markets, as illustrated by the dimensionality time series during the Black Monday crash in October 1987 (Fig. 6a), the Financial and Banking Crisis during 2007-2009 (Fig. 6b), and the Sovereign Debt Crisis in August 2011 (Fig. 6c).

Additionally, we find in our crash analysis that change in dimension estimated is more sensitive to market movements, as shown in Fig. 6b, when information metric distance measure is used, compared to geodesic distance.

Further, we observe that autoencoders are hard to train in this particular problem domain, especially with limited financial data. While the dimension estimated by autoencoder overlaps dimension estimated by Isomap, as in Fig. 4, the autoencoder estimated dimension is much more sensitive to market movements. We believe this is because autoencoders are data hungry and we have limited end of day data to train a complex autoencoder. In future work, we plan to use intra-day (higher frequency) trade data from NYSE.

Acknowledgement

Results in this paper were obtained in part using a high-performance computing system acquired through NSF MRI grant DMS-1337943 to WPI.

References

[1] C. Borch, K. B. Hansen, and A.-C. Lange, "Markets, bodies, and rhythms: A rhythmanalysis of financial markets from open-outcry trading to high-frequency trading," *Environment and Planning D: Society and Space*, vol. 33, no. 6, pp. 1080–1097, 2015.

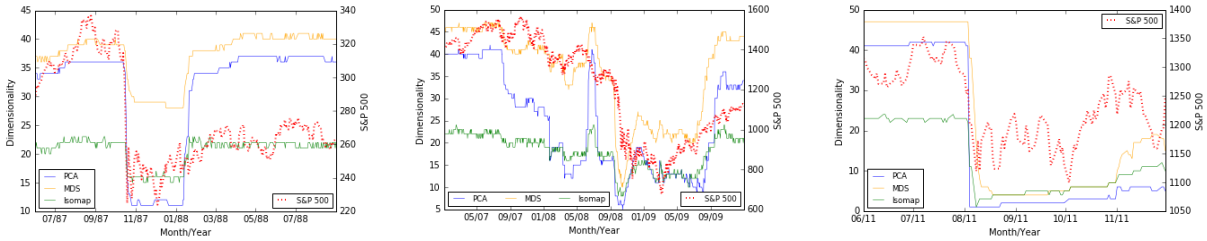
[2] N. Oliveira, P. Cortez, and N. Areal, "The impact of microblogging data for stock market prediction: Using twitter to predict returns, volatility, trading volume and survey sentiment indices," *Expert Systems with Applications*, vol. 73, pp. 125–144, 2017.

[3] S. Kim, Y. Koh, J. Cha, and S. Lee, "Effects of social media on firm value for us restaurant companies," *International Journal of Hospitality Management*, vol. 49, pp. 40–46, 2015.

[4] C. D. Dannhauser, "The impact of innovation: Evidence from corporate bond exchange-traded funds (etfs)," *Journal of Financial Economics*, vol. 125, no. 3, pp. 537–560, 2017.

[5] R. Wermers, "Active investing and the efficiency of security markets," *Available at SSRN*, 2019.

[6] R. Kliman and B. Arinze, "Cognitive computing: Impacts on financial advice in wealth management," in *Aligning business strategies and analytics*. Springer, 2019, pp. 11–23.



(a) Black Monday (Oct 19, 1986) using information metric distance.

(b) Financial and Banking Crisis (2007-2009) using information metric distance.

(c) Europe Soverign Debt Crisis (Aug 2011) using information metric distance.

Fig. 6: How estimated dimension changed during the Black Monday crisis (October 1987), the Financial and Banking crisis (2007-2009) and the Europe Sovereign Debt crisis (August 2011).

[7] J. A. Lee and M. Verleysen, *Nonlinear dimensionality reduction*. Springer Science & Business Media, 2007.

[8] N. Bahadur and R. Paffenroth, "Dimension estimation using autoencoders," *arXiv preprint arXiv:1909.10702*, 2019.

[9] H. K. Pharasi, K. Sharma, R. Chatterjee, A. Chakraborti, F. Leyvraz, and T. H. Seligman, "Identifying long-term precursors of financial market crashes using correlation patterns," *New Journal of Physics*, vol. 20, no. 10, p. 103041, 2018.

[10] S. Mollah, A. S. Quoreshi, and G. Zafirov, "Equity market contagion during global financial and eurozone crises: Evidence from a dynamic correlation analysis," *Journal of International Financial Markets, Institutions and Money*, vol. 41, pp. 151–167, 2016.

[11] H. Markowitz, "Portfolio selection," *Journal of Finance*, vol. 7, no. 1, 1952.

[12] X. Zhong and D. Enke, "Forecasting daily stock market return using dimensionality reduction," *Expert Systems with Applications*, vol. 67, pp. 126–139, 2017.

[13] J. Jurczyk, T. Rehberg, A. Eckrot, and I. Morgenstern, "Measuring critical transitions in financial markets," *Scientific reports*, vol. 7, no. 1, p. 11564, 2017.

[14] Z. Zheng, B. Podobnik, L. Feng, and B. Li, "Changes in cross-correlations as an indicator for systemic risk," *Scientific reports*, vol. 2, p. 888, 2012.

[15] M. Billio, M. Getmansky, A. W. Lo, and L. Pelizzon, "Econometric measures of connectedness and systemic risk in the finance and insurance sectors," *Journal of financial economics*, vol. 104, no. 3, pp. 535–559, 2012.

[16] E. J. Elton, M. J. Gruber, S. J. Brown, and W. N. Goetzmann, *Modern portfolio theory and investment analysis*. John Wiley & Sons, 2009.

[17] I. Jolliffe, *Principal component analysis*. Wiley Online Library, 2002.

[18] Y. Wang and P. Shang, "Analysis of financial stock markets through multidimensional scaling based on information measures," *Nonlinear Dynamics*, vol. 89, no. 3, pp. 1827–1844, 2017.

[19] K. Bhattacharjee, S. Pal, and S. Pal, "Detection of dissimilarity among different indian banks: An mds approach," *IUP Journal of Bank Management*, vol. 16, no. 1, 2017.

[20] X. Wang, "On the effects of dimension reduction techniques on some high-dimensional problems in finance," *Operations Research*, vol. 54, no. 6, pp. 1063–1078, 2006.

[21] C. Albanese, K. Jackson, and P. Wiberg, "Dimension reduction in the computation of value-at-risk," *The Journal of Risk Finance*, vol. 3, no. 4, pp. 41–53, 2002.

[22] M.-H. Yang, "Extended isomap for pattern classification," in *AAAI/IAAI*, 2002, pp. 224–229.

[23] M. Balasubramanian and E. L. Schwartz, "The isomap algorithm and topological stability," *Science*, vol. 295, no. 5552, pp. 7–7, 2002.

[24] Y. Huang, G. Kou, and Y. Peng, "Nonlinear manifold learning for early warnings in financial markets," *European Journal of Operational Research*, vol. 258, no. 2, pp. 692–702, 2017.

[25] W. Bao, J. Yue, and Y. Rao, "A deep learning framework for financial time series using stacked autoencoders and long-short term memory," *PloS one*, vol. 12, no. 7, p. e0180944, 2017.

[26] J. Heaton, N. Polson, and J. H. Witte, "Deep learning for finance: deep portfolios," *Applied Stochastic Models in Business and Industry*, vol. 33, no. 1, pp. 3–12, 2017.

[27] M. Schreyer, T. Sattarov, D. Borth, A. Dengel, and B. Reimer, "Detection of anomalies in large scale accounting data using deep autoencoder networks," *arXiv preprint arXiv:1709.05254*, 2017.

[28] 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.

[29] Y. Wang, H. Yao, and S. Zhao, "Auto-encoder based dimensionality reduction," *Neurocomputing*, vol. 184, pp. 232–242, 2016.

[30] 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.

[31] M. Statman, "How many stocks make a diversified portfolio?" *Journal of financial and quantitative analysis*, vol. 22, no. 3, pp. 353–363, 1987.

[32] E. J. Elton and M. J. Gruber, "Risk reduction and portfolio size: An analytical solution," *The Journal of Business*, vol. 50, no. 4, pp. 415–437, 1977.

[33] S. Benartzi and R. H. Thaler, "Naive diversification strategies in defined contribution saving plans," *American economic review*, vol. 91, no. 1, pp. 79–98, 2001.

[34] D. L. Domian, D. A. Louton, and M. D. Racine, "Portfolio diversification for long holding periods: how many stocks do investors need?" *Studies in Economics and Finance*, vol. 21, no. 2, pp. 40–64, 2003.

[35] G. Y. Tang, "How efficient is naive portfolio diversification? an educational note," *Omega*, vol. 32, no. 2, pp. 155–160, 2004.

[36] V. Alexeev and F. Tapon, "Equity portfolio diversification: how many stocks are enough? evidence from five developed markets," *Evidence from Five Developed Markets (November 28, 2012). FIRN Research Paper*, 2012.

[37] Y. Huang and G. Kou, "A kernel entropy manifold learning approach for financial data analysis," *Decision Support Systems*, vol. 64, pp. 31–42, 2014.

[38] 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.

[39] Z. Jiang, D. Xu, and J. Liang, "A deep reinforcement learning framework for the financial portfolio management problem," *arXiv preprint arXiv:1706.10059*, 2017.

[40] A. Tsantekidis, N. Passalis, A. Tefas, J. Kanniainen, M. Gabbouj, and A. Iosifidis, "Using deep learning to detect price change indications in financial markets," in *2017 25th European Signal Processing Conference (EUSIPCO)*. IEEE, 2017, pp. 2511–2515.

[41] J. Sirignano, A. Sadhwani, and K. Giesecke, "Deep learning for mortgage risk," *arXiv preprint arXiv:1607.02470*, 2016.

[42] D. J. MacKay and D. J. Mac Kay, *Information theory, inference and learning algorithms*. Cambridge university press, 2003.

[43] H. Hotelling, "Analysis of a complex of statistical variables into principal components." *Journal of educational psychology*, vol. 24, no. 6, p. 417, 1933.

[44] J. D. Petruccelli, B. Nandram, and M. Chen, *Applied statistics for engineers and scientists*. Prentice Hall New Jersey, 1999.

[45] T. F. Cox and M. A. Cox, *Multidimensional scaling*. CRC press, 2000.

[46] R. W. Floyd, "Algorithm 97: shortest path," *Communications of the ACM*, vol. 5, no. 6, p. 345, 1962.

[47] T. H. Cormen, *Introduction to algorithms*. MIT press, 2009.

[48] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *Proceedings of the 27th international conference on machine learning (ICML-10)*, 2010, pp. 807–814.

[49] A. Ng *et al.*, "Sparse autoencoder," *CS294A Lecture notes*, vol. 72, no. 2011, pp. 1–19, 2011.

[50] 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.