Decision Support Systems xxx (2014) xxx–xxx

Contents lists available at ScienceDirect

Decision Support Systems

journal homepage: www.elsevier.com/locate/dss

# A kernel entropy manifold learning approach for financial data analysis ☆

Yan Huang [a], Gang Kou [b,c,*]

[a] School of Management and Economics, University of Electronic Science and Technology of China, Chengdu 610054, China
[b] School of Business Administration, Southwestern University of Finance and Economics, Chengdu, China
[c] Collaborative Innovation Center of Financial Security, Southwestern University of Finance and Economics, Chengdu, China

ARTICLE INFO

*Article history:*
Received 23 August 2013
Received in revised form 6 April 2014
Accepted 18 April 2014
Available online xxxx

*Keywords:*
Manifold learning
Financial analysis
Low-dimensional embedding
Information metric

ABSTRACT

Identification of intrinsic characteristics and structure of high-dimensional data is an important task for financial analysis. This paper presents a kernel entropy manifold learning algorithm, which employs the information metric to measure the relationships between two financial data points and yields a reasonable low-dimensional representation of high-dimensional financial data. The proposed algorithm can also be used to describe the characteristics of a financial system by deriving the dynamical properties of the original data space. The experiment shows that the proposed algorithm cannot only improve the accuracy of financial early warning, but also provide objective criteria for explaining and predicting the stock market volatility.



## 1. Introduction

Traditional financial analysis methodologies include quantitative model and textual analysis. The quantitative model is the analysis about financial data by the use of statistical analysis tools or artificial intelligence technologies, which relies on the selection about basic important factors, such as financial ratios, technical indexes, and macroeconomic indexes [1]. The textual analysis utilizes text mining techniques to analyze the context of financial reports, which are dependent on the identification of a predefined set of keywords [2]. Since different factors or keywords are selected for different studies, the results are often subjective.

The real financial indicators are numerous while the complex high-dimensional data tends to obscure the essential feature of data [4]. Identifying intrinsic characteristics and structure of high-dimensional data is important for financial analysis. Inspired by the Quantitative Structure–Property Relationship (QSPR) method [3], whose core idea is that the microscopic structure of a material determines its macroscopic properties, this paper tries to find the inherent relationships between data points of financial dataset, and further derive the overall characteristics of the financial system.

Manifold learning, which explores the inherent low-dimensional manifold structure of high-dimensional data, is a valid choice for this task. In the field of financial analysis, data information characteristics, i.e. probability distributions, are important. However, many existing manifold learning algorithms concern about space geometric characteristics [5–8]. When Probability Density Functions (PDFs) are constrained to form a sub-manifold of interest, the straight-shot distance is no longer an accurate description of the manifold distance [50]. For financial data sets, each data point represents a listed company, while the distance between the data points indicates the degree of difference between the financial positions of listed companies. If the difference was characterized only by the geometric space distance between data points, it may not only unfit the practical significance of financial analysis, but also cause problems in the subsequent analysis. Therefore, this study employs the information metric to measure the relationships between listed companies and obtains the relationship metric model.

Real-world financial data is often nonlinear [10] and linear mapping manifold learning cannot fully capture the data information. Though Qiao et al. proposed a nonlinear mapping [11], the method is too complicated for the current problem. Kernel is often used to discover nonlinear structure in data [12,13]. The objective of this paper is to propose a kernel entropy manifold learning (KEML) algorithm to obtain the low-dimensional representation of high-dimensional financial data from the perspective of manifold learning. The KEML algorithm is extended to a kernel feature space so that the low-dimensional embedding can reflect the characteristics of the original financial data set. Experiments using small and medium-sized companies from China A-share Stock Market are designed to validate the proposed algorithm.

The rest of the paper is organized as follows: Section 2 reviews related works. Section 3 describes the modeling of financial data manifold and the proposed algorithm. Section 4 reports the experimental study and the last section concludes the paper.

☆ Two authors are alphabetically ordered by their last name.

* Corresponding author at: School of Business Administration, Southwestern University of Finance and Economics, Chengdu, China.
*E-mail address:* kougang@yahoo.com (G. Kou).

http://dx.doi.org/10.1016/j.dss.2014.04.004

## 2. Literature review and preliminaries

### 2.1. Machine learning in financial analysis

Over the past few decades, machine learning algorithms have been widely used in the financial field and have been reported to be quite effective in some cases [14]. Machine learning quantitative models include single algorithms, such as ANN [15–17], SVM [18–20] and SOM [21,22], and hybrid techniques, which combine two or more algorithms. Many studies have been conducted to develop hybrid techniques for financial analysis. Serrano-Cinca and Gutiérrez-Nieto [23] combined partial least square (PLS) regression model and principal component analysis (PCA) and multiple linear regression (MLR) for bankruptcy prediction. Yolcu et al. [24] used a hybrid artificial neural network containing linear and nonlinear components for time series forecasting. Kao et al. [25] combined multivariate adaptive regression splines (MARS) and support vector regression (SVR) for stock index forecasting. Lu et al. [26] used independent component analysis (ICA) and support vector regression (SVR) in financial time series forecasting.

Context-based text analysis had been used to analyze unstructured data in financial reporting. Groth and Muntermann [27] and Chan and Franklin [2] and Humpherys et al. [28] adopted text mining technology to analyze the unstructured data of financial reports to improve prediction accuracy of financial risk. Schumaker and Chen [29] used textual representations of financial news articles to estimate the discrete stock price. Olson et al. [30] compared data mining methods for bankruptcy prediction.

The financial dataset can be considered as a system, in which each data point is an element. The intrinsic relationships between elements constitute the system structure, which determines the characteristics of the system. Inspired by the idea of QSPR, this study tried to explore the intrinsic structure of the system, and then discover the overall status of the system.

### 2.2. Manifold learning

A manifold is a topological space which is locally Euclidean. High-dimensional data observed in real world are often the consequences of a small number of factors [31]. Manifold learning algorithms assume that the input data resides on or close to a low-dimensional manifold embedded in the ambient space [32]. Thus it is possible to construct a mapping that obeys certain properties of the manifold and obtain low-dimensional representation of high-dimensional data with good preservation of the intrinsic structure in the data [32].

Currently dimension reduction techniques are mainly divided into two categories: linear and nonlinear methods. The most well known linear method is principal component analysis (PCA), which is based on correlation matrices [38]. PCA is a classical feature extraction and data representation technique widely used in pattern recognition and computer vision. Sirovich and Kirby utilized PCA to represent pictures of human faces [54]. Turk and Pentland presented the well-known Eigenfaces method for face recognition in 1991 [54]. Kernel PCA (KPCA), a kernel extension of PCA, is also a very influential method. KPCA performs traditional PCA in a kernel feature space, which is nonlinearly related to the input space [38].

Compared with traditional dimension reduction approaches, manifold learning has advantages such as nonlinear nature, geometric intuition, and computational feasibility. Many manifold learning methods have been developed over the years. Isometric Feature Mapping (ISOMAP) [6] and Locally Linear Embedding (LLE) [7] are the earliest ones. The key idea of ISOMAP algorithm is to preserve the geodesic distance among points on the manifold and embed data into low-dimensional space by multidimensional scaling. LLE computes the reconstruction weights of each point and then minimizes the embedding cost by solving an eigenvalue problem to preserve the proximity relationship among data.

Local tangent space alignment (LTSA) constructs local linear approximations of the manifold in the form of a collection of overlapping approximate tangent spaces at each sample point, and then aligns those tangent spaces to obtain a global parameterization of the manifold [5]. LTSA maps the high dimensional data points on a manifold to points in a lower dimension Euclidean space. This mapping is isometric if the manifold is isometric to its parameter space [5]. Local Multidimensional Scaling (LMDS) is a data embedding method based on the alignment of overlapping locally scaled patches [8] and inputs are local distances. A subset of overlapping patches is chosen by a greedy approximation algorithm of minimum set cover. The patches are aligned to derive global coordinates and minimize a residual measure. LMDS is locally isometric and scales with the number of patches rather than the number of data points. LMDS produces less deformed embedding results than LLE [8].

These manifold learning algorithms use geodesic distance metric or weight measurement to calculate similarities between data points. In many problems of practical interest, however, the manifold geometry is unavailable and the calculation of geodesics must be done in a model-free, nonparametric fashion [34]. In applications like financial analysis, for example, only considering the geometry structure of data space may miss some essential characteristics of data and destroy the proximity relations (topology) of the original data space [9].

### 2.3. Information distance metric

This study adopted an information theory-based metric to measure the difference between data points. Shannon suggested that "information entropy plays a central role in information theory as measures of information, choice, and uncertainty" [35]. Kolmogorov complexity [36] measures information content of an object. Bennett et al. [37] proposed the information distance theory and proved the fundamental universal theorem. Information distance measures the essential relationship between things. Due to its parameter-free, feature-free, and alignment-free characteristics, it can be used to deal with unstructured and incomprehensible data. A distance is a function $D$ with nonnegative real values, defined on the Cartesian product $X \times X$ of a set $X$. It is called a metric on $X$ if for every $x, y, z \in X$:

- $D(x,y) = 0$ iff $x = y$ (the identity axiom);
- $D(x,y) + D(y,z) \geq D(x,z)$ (the triangle inequality);
- $D(x,y) = D(y,x)$ (the symmetry axiom).

A set $X$ provided with a metric is called a metric space. For example, every set $X$ has the trivial discrete metric $D(x, y) = 0$ if $x = y$ and $D(x, y) = 1$ otherwise [37]. The information metric between stochastic sources $X$ and $Y$ is defined as $D(x, y) = H(x|y) + H(y|x)$ [37]. Here $H(x|y)$ is used to measure the difference between probability distributions.

In recent years, entropy-based distance metric has been investigated by the manifold learning field. Costa and Hero [33] proposed geodesic-minimal-spanning-tree (GMST) method that jointly estimates both the intrinsic dimension and intrinsic entropy on the manifold. Jenssen [38] developed kernel entropy component analysis (KECA) for data transformation and dimensionality reduction. KECA reveals structure relating to the Renyi entropy of the input space data set. Carter et al. [34] proposed Fisher Information Nonparametric Embedding (FINE) which utilizes the properties of information geometry and statistical manifolds to define similarities between data sets using Fisher information distance. FINE showed that this metric can be approximated using nonparametric methods. Carter et al. [50] presented methods for

Please cite this article as: Y. Huang, G. Kou, A kernel entropy manifold learning approach for financial data analysis, Decision Support Systems (2014), http://dx.doi.org/10.1016/j.dss.2014.04.004