European Journal of Operational Research 258 (2017) 692–702

Contents lists available at ScienceDirect

European Journal of Operational Research

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

Innovative Applications of O.R.

Nonlinear manifold learning for early warnings in financial markets

Yan Huang[a,1], Gang Kou[b,1], Yi Peng[c,1,*]

[a] School of Economics, Xihua University, Chengdu 610039, China
[b] School of Business Administration, Southwestern University of Finance and Economics, Chengdu, China
[c] School of Management and Economics, University of Electronic Science and Technology of China, Chengdu 610054, China

ARTICLE INFO

Article history:
Received 16 June 2015
Accepted 20 August 2016
Available online 27 August 2016

Keywords:
Data mining
Manifold learning
Financial markets
Early warning
Dynamic system

ABSTRACT

A financial market is a complex, dynamic system with an underlying governing manifold. This study introduces an early warning method for financial markets based on manifold learning. First, we restructure the phase space of a financial system using financial time series data. Then, we propose an information metric-based manifold learning (IMML) algorithm to extract the intrinsic manifold of a dynamic financial system. Early warning ranges for critical transitions of financial markets can be detected from the underlying manifold. We deduce the intrinsic geometric properties of the manifold to detect impending crises. Experimental results show that our IMML algorithm accurately describes the attractor manifold of the financial dynamic system, and contributes to inform investors about the state of financial markets.

1. Introduction

Scholars and practitioners have developed increasingly elaborate techniques intended to forecast the approach of financial crises. As a complex, dynamic system, financial markets can exhibit tipping points at which abrupt transitions to a contrasting dynamic regime may occur (Scheffer et al., 2009). This shift is called a critical transition in financial markets, and it is exemplified by systemic market crashes or global crises (Ang & Timmermann, 2012). It is difficult to predict reliably when critical thresholds approach because markets might show little change before reaching the critical point (Scheffer et al., 2009). In addition, shifts in financial markets are usually triggered by stochastic and unpredictable externalities (Sugihara et al., 2012; Battiston et al., 2016). However, investors need adequate warning that an impending crisis is highly probable and reduce potential losses.

The intrinsic complexity and nonlinearity of financial markets make it hard to construct an integral mathematical model to characterize the financial system, and thus the corresponding early warning model is unable to be constructed (Christofides, Eicher, & Papageorgiou, 2016; Kou et al., 2014). However, financial time series are comprehensive reflections of market operations and provide a database for market analysis (Ausín, Galeano, & Ghosh, 2014). In practice, observations about the state of dynamic systems are often one-dimensional time series data. Through Phase Space Reconstruction (PSR), time series data can be reconstructed in a space in which the topology is equivalent to the original dynamic system (Richard, Michael, Andrew, & Ye, 2004). PSR describes the trajectory of the dynamic system in the reconstructed high-dimensional space (He, Liu, Long, & Wang, 2012). Takens' embedding theorem shows that the N-dimensional dynamic system has a low-dimensional structure because the system state is confined to an attractor (M) of dimension $d(d < N)$ in the state space (Han & Christopher, 2011).

An underlying manifold governs dynamic systems and reveals their dynamic nature (Sugihara et al., 2012). Therefore, extracting the intrinsic manifold structure is a primary objective of market research. Manifold learning is a hot topic in the fields of data mining and machine learning, which seek to find the intrinsic low-dimensional embedding structures within high-dimensional data. Our study proposes a manifold learning approach to extract the structure of the manifold underlying high-dimensional phase spaces, explore early warning ranges for critical transitions in markets, and discover further intrinsic structural properties.

Numerous manifold learning methods have been developed, including Isometric Feature Mapping (ISOMAP) (Tenenbaum, Sivlar, & Langford, 2000), Locally Linear Embedding (LLE) (Roweis & Saul, 2000), and Local Tangent Space Alignment (LTSA) (Zhang & Zha, 2004). These methods have successfully discovered the embedded low-dimensional manifold. However, classical manifold learning algorithms are concerned with space geometric characteristics. In financial analysis, data information characteristics—i.e., probability distributions—are important (Huang & Kou, 2014). When

* Corresponding author.
E-mail addresses: kougang@yahoo.com, kou.gang@qq.com, kougang@swufe.edu.cn (G. Kou), pengyi@uestc.edu.cn (Y. Peng).
[1] These authors contributed equally to this work.

http://dx.doi.org/10.1016/j.ejor.2016.08.058

probability density functions (PDFs) are restricted to form an intrinsic manifold of high-dimensional data, geodesic distance no longer accurately describes the manifold distance (Carter, Raich, Finn, & Hero, 2011).

We selected stock market composite indices as observed in financial time series data. Each data point represents a financial system state, and the distance between them indicates the degree of difference between system states. If the difference is characterized only by the geometric space between data points, the result may not fit the practical significance of the financial analysis but rather cause errors in subsequent analyses. Therefore, this study proposes an IMML algorithm to discover the structure embedded in the high-dimensional phase space, which is reconstructed by the observed financial time series range.

Our study is conducted in three steps. First, we reconstruct an observed financial time series as a high-dimensional phase space. Second, we propose the IMML algorithm and employ it to extract the manifold embedded in the high-dimensional phase space. Third, we use the underlying manifold to detect early warning ranges for critical transitions in markets. In addition to the crisis diagnosis, we implement market prognosis from the perspective of the inherent geometric properties in the manifold.

This study is organized as follows. Section 2 reviews related theory and methods. Section 3 describes our manifold learning method. Section 4 reports the experimental study. Section 5 concludes the paper.

2. Preliminaries

2.1. Phase Space Reconstruction

The theoretical basis of PSR originates in Takens' embedding theorem (Takens, 1981, chap. 21). The theorem shows that complete information about the hidden state of dynamic systems can be preserved in observed time series data. The phase space is a time-delay reconstruction using time-delayed versions of a time series as coordinates for the space. Specifically, given a time series $x = x_n,\ n = 1,\ \dots\ , N$, a reconstructed phase space matrix X of dimension m and time lag τ is defined by its row vectors:

$$x = [x_{n-(m-1)\tau},\ \dots\ , x_{n-\tau},\ x_n], \tag{1}$$

where $n = (1 + (m-1)\tau) \cdots N$ and a row vector x_n is a point in the reconstructed phase space.

A proper time lag can reduce the required RPS dimension. A common heuristic for selecting time lag is to use the first minimum of the automutual information function (Richard et al., 2004). The automutual information function is defined as

$$I_n(X_0, X_1,\ \dots\ , X_n) = \sum_j (H(x_j) - H(X_0, X_1,\ \dots\ , X_n)), \tag{2}$$

where $H(x_j)$ is the entropy and $H(X_0, X_1,\ \dots\ , X_n)$ is the joint entropy of the time series data points. τ is at the first local minimum of mutual information.

Embedding dimension m is another vital parameter for PSR, which is not previously known. Many methods seek to determine the dimension parameter, including the global false nearest-neighbor technique and the Cao method (Cao, 1997). We adopt the Cao method for its robust handling of noise and because it presents no need to set threshold values manually. The related process of calculating m is as follows (Cao, 1997):

$$Y_i(m) = (x_i, x_{i+\tau},\ \dots\ , x_{i+(m-1)\tau}),\ i = 1, 2,\ \dots\ , N - (m-1)\tau, \tag{3}$$

where m is the embedding dimension, τ is the time lag, and $Y_i(m)$ denotes the ith reconstruction vector with embedding dimension m. Moreover, let

$$a(i, m) = \frac{\left\| Y_i(m+1) - Y_{n(i,m)}(m+1) \right\|}{\left\| Y_i(m) - Y_{n(i,m)}(m) \right\|} \quad i = 1, 2, \dots, N - m\tau, \tag{4}$$

where $\|Y_k(m) - Y_l(m)\| = \max_{0 \le j \le m-1} |x_{k+j\tau} - x_{l+j\tau}|$ and $a(i, m)$ $(1 \le a(i, m) \le N - m\tau)$ is an integer such that $Y_{n(i,m)}(m)$ is the nearest neighbor of $Y_i(m)$ in the m-dimensional reconstructed phase space.

2.2. Manifold Learning

A manifold can be viewed as a nonlinear object that is locally linear (Jamshidi, Kirby, & Broomhead, 2011). For high-dimensional real world data, a perceptually meaningful structure has few degrees of freedom. In other words, high-dimensional data points can be mapped into a surrogate low-dimensional space (Seung & Lee, 2000). Hence, it is possible to construct a mapping that obeys specific properties of the manifold and obtains low-dimensional representation of high-dimensional data while preserving the intrinsic structure underlying the data (Lin & Zha, 2008).

Of the many manifold learning methods, ISOMAP and LLE are the earliest. The key idea of the ISOMAP algorithm is to maintain geodesic distance among points on the manifold and embedded data into low-dimensional space through multidimensional scaling. LLE calculates the reconstruction weights of each point and minimizes embedding cost by solving an eigenvalue problem to preserve the proximity relationship among data. LTSA constructs local linear approximations of the manifold by using a collection of overlapping approximate tangent spaces at each data point and aligns these tangent spaces to obtain a global parameterization of the manifold (Zhang & Zha, 2004). 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. Local Multidimensional Scaling (LMDS) is a data embedding method based on the alignment of overlapping locally scaled patches, and its inputs are local distances (Yang, 2008). 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. Also a common nonlinear method for dimension reduction, Kernel Principal Component Analysis (KPCA) is a kernel extension of PCA and a special manifold learning algorithm. KPCA conducts traditional PCA in a kernel feature space, which is nonlinearly related to the input space (Jenssen, 2010).

These manifold learning algorithms use a geodesic distance metric or weight measurement to calculate similarities among data points. In financial practice, considering only the geometric structure of a data space disguises essential characteristics of the data and destroys the proximity relations (topology) of the original data space.

2.3. Information distance metric

The theoretical basis of information distance originates in Shannon information theory and Kolmogorov complexity theory. It is framed as the universal cognitive similarity distance that measures the essential relationship between things (Kolmogorov, 1965). Owing to its parameter-free, feature-free, and alignment-free characteristics, it can be used to manage unstructured and incomprehensible data. The information distance metric (Bennett, Gács, Li, Vitányi, & Zurek, 1998) is the Riemannian metric between PDFs p_1