2014

MANIFOLD LEARNING-BASED PHASE SPACE RECONSTRUCTION FOR FINANCIAL TIME SERIES

Yan Huang
University of Electronic Science and Technology of China, 1090449942@qq.com

Gang Kou
Southwestern University of Finance and Economics, kougang@yahoo.com

Follow this and additional works at: http://aisel.aisnet.org/pacis2014

Recommended Citation
http://aisel.aisnet.org/pacis2014/364

This material is brought to you by the Pacific Asia Conference on Information Systems (PACIS) at AIS Electronic Library (AISeL). It has been accepted for inclusion in PACIS 2014 Proceedings by an authorized administrator of AIS Electronic Library (AISeL). For more information, please contact elibrary@aisnet.org.
MANIFOLD LEARNING-BASED PHASE SPACE RECONSTRUCTION FOR FINANCIAL TIME SERIES

Yan Huang, School of Management and Economics, University of Electronic Science and Technology of China, Chengdu 610054, 1090449942@qq.com

Gang Kou, School of Business Administration, Southwestern University of Finance and Economics, Chengdu, China, kougang@yahoo.com

Abstract

Phase space reconstruction of financial time series has become an effective approach for financial system analysis. Since the phase space obtained by the traditional method had the large embedding dimension and was susceptible to noise, this paper proposed a novel phase space reconstruction approach based on manifold learning, called manifold learning-based phase space reconstruction (MLPSR), which was applied in financial time series. In our approach, the traditional correlation dimension estimation was employed to reconstruct a rough phase space of financial time series; then manifold learning method was adopted to discover the embedding natural structure of the rough phase space, and thus the intrinsic embedding dimension was obtained. In the empirical study, our method was compared with the traditional correlation dimension method, and showed the better performance. Our MLPSR would reduce the phase space reconstruction error of financial time series, and provide more accurate data support for the subsequent studies about financial system.

Keywords: Manifold learning, Phase space reconstruction, Financial time series.
1 INTRODUCTION

With the rapid development of financial liberalization, globalization and financial innovation, the financial system has become an open, nonlinear complex system. Due to the inherent nonlinearity and complexity of the financial system, it is very difficult and unrealistic to describe the full operating conditions by constructing a mathematical model. In reality, people often only obtained financial time series, which described some state components of financial system. Therefore, to study the complex nonlinear financial system, a useful tool, namely phase space reconstruction technique, can be employed to recover all the characteristics and natures of the real financial system in phase space. The basic idea of Phase Space Reconstruction is: a system state component and its observations in delaying time point are the new dimensions for the system, thus the phase space can be obtained by 'embedding' method which is equal to the original system, and all the features of the original system can be recovered in the phase space. Phase space reconstruction need two key parameters: phase space embedding dimension $m$ and time delay $\tau$. When the two key parameters are selected appropriately, the reconstructed phase space has the same nature and evolution information with the original system. For the selection about embedding dimension and time delay, the most common method is the correlation dimension estimation (Grassberger & Procaccia 1983), which uses the correlation integral to simultaneously estimate the time delay and embedding dimension. Due to the limited data amount and the noise, the calculation results of correlation dimension estimation will be amplified, which will result in a very large phase space reconstruction error, and even the wrong conclusions. (Eckmann & Ruelle 1992).

Recently, manifold learning has emerged in nonlinear feature extraction. It can identify a low-dimensional nonlinear structure hidden in high-dimensional data through several techniques, including Locally Linear Embedding (LLE) (Roweis & Saul 2000), Isometric Feature Mapping (ISOMAP) (Tenenbaum et al. 2000), and Local Tangent Space Alignment (LTSA) (Zhang & Zha 2005), etc. Therefore, we can adopt this merit to extract the intrinsic structure from the reconstructed high-dimensional phase space of financial time series, to reduce noise and improve the quality of the reconstruction phase space.

2 BACKGROUND AND PRELIMINARIES

2.1 Phase Space Reconstruction

The basis of Phase Space Reconstruction was that given access to the state structure of a system, a classification of such systems can be developed. We started by presenting a theoretical construct of the problem. Given a finite-dimensional system state space $M$ and $\varphi : M \rightarrow M$, the dynamics of the system, a system was described by the pair $< M, \varphi >$. We then defined a set $\Phi$ of all possible dynamics on $M$ with a topology $\zeta$. Without loss of generality, we assumed $M$ to be $d$-dimensional, because given any $M' \subset M$, $M$ could be replaced by $M \cup M'$. The system classification then became one of partitioning $\Phi$ according to the requirements of the classification problem with a particular dynamics $\varphi$ identified with a particular partition $P_i$ such that $\Phi = \cup P_i$, where $P_i \cap P_j = \emptyset, i \neq j$.

The problem for real world systems was how to gain access to and represent $\varphi$ for a particular system. The approach used here was phase space reconstruction, also known as phase space embedding, and was first proposed by Takens (1980). The methods for representing $\varphi$, which were the contributions of this work, were presented in the following section.
The central premise was that a space and its associated dynamics, which were topologically equivalent to the original system space $M$ and its dynamics $\varphi$, could be recovered or unfolded from a time series of observations of a single state variable for the original system $\langle M, \varphi \rangle$.

Whitney showed that a $d$-dimensional topological space could be embedded in $\mathbb{R}^{2d+1}$ (Whitney 1936), where an embedding was a homeomorphisms mapping from and topological space to another. Takens (1980) showed that it was a generic property that the map $\Phi_{(\varphi, x)} : M \to \mathbb{R}^{2d+1}$ was defined by an embedding, where $M$ was a $d$-dimensional state space, $\varphi : M \to M$ was a twice continuously differentiable diffeomorphism which described the dynamics of the system, and $x : M \to \mathbb{R}$ was a twice continuously differentiable function representing the observation of a single state variable.

Working from these original theorems, Sauer et al. (1991) extended Takens’ work by showing that almost every time-delay map $\Phi_{(\varphi, x)}$ was an embedding, indicating that except for a set of degenerate cases with measure zero, and the topological equivalence property was guaranteed. Therefore, these theorems guaranteed that for almost every time delay embedding, the reconstructed dynamics of such a map were topologically identical to the true dynamics of the system (Sauer et al. 1991). In addition, they found a tighter bound on the required dimension as $d > 2d_0$, where $d_0$ was the box-counting dimension of the attractor of the underlying system.

In other words, we had a mechanism for obtaining a continuous, one-to-one, and onto transformation from $\langle M, \varphi \rangle$ to $\langle \mathbb{R}^d, X \rangle$, where $d > 2d_0$ and $X$ were the trajectory matrix defined as follows. Given a time series $x = x_n, n = 1, \ldots, N$, a sequence of state variable observations, a trajectory matrix $X$ of dimension $d$ and time lag $\tau$ were defined as

$$X = \begin{bmatrix} X_{1+(d-1)\tau} & \cdots & x_{1+\tau} & x_1 \\ X_{2+(d-1)\tau} & \cdots & x_{2+\tau} & x_2 \\ \vdots & & \ddots & \vdots \\ X_{N} & \cdots & x_{N-(d-2)\tau} & x_{N-(d-1)\tau} \end{bmatrix}$$

Where each row vector in the matrix represents a single point in the space:

$$X_n = [x_{n-(d-1)\tau} \cdots x_{n-\tau} x_n]$$

Where $n = (1 + (d-1)\tau) \ldots N$. A row vector $X_n$ was a point in the RPS. The pattern traced out by $X$ in $\mathbb{R}^d$ was typically referred to as an attractor, even when the technical definition of an attractor was not formally met. We adopt this terminology.

Recall, that the system/signal classification problem in this work was addressed by transforming a signal from a particular system into a RPS, which had a mathematical correspondence with the underlying system. Therefore, given $\mathcal{S}$, the collection of all possible RPS $X$, the system/signal classification problem was to define a partition of $\mathcal{S}$ such that $\mathcal{S} = \cup P_i$ and a mechanism for identifying a particular $X$ with a particular $P_i$.

As discussed next, a statistical machine learning approach was taken for defining the partition and for identifying an $X$ with a $P_i$. In general, the classification accuracy would depend on how well the model of a partition $P_i$ of $\mathcal{S}$ characterizes the signals that were labelled as belonging to that partition and how different the model of $P_i$ was from the models of other partitions.
2.2 Correlation dimension estimation technique

The correlation dimension estimation technique was commonly used to estimate the dimension of attractors of dynamical systems. The correlation dimension was defined as follows: let \( \Omega = x_1, x_2, \ldots, x_N \) be a set of points in \( \mathbb{R}^n \) of cardinality \( N \). If the correlation integral \( C_m(r) \) was defined as:

\[
C_m(r) = \lim_{N \to \infty} \frac{2}{N(N-1)} \sum_{i=1}^{N} \sum_{j=i+1}^{N} I(\|x_j - x_i\| \leq r),
\]

where \( I \) was an indicator function, then the correlation dimension \( D \) of \( \Omega \) is:

\[
D = \lim_{r \to 0} \frac{\ln(C_m(r))}{\ln(r)}.
\]

In our work, we used the correlation dimension method to get an initial value of the phase space dimension.

2.3 Manifold Learning Approach

A manifold is a topological space which is locally Euclidean. High-dimensional data observed in the real world are often the consequences of a small number of factors (Law 2006). Manifold learning algorithms assume that the input data resides on or close to a low-dimensional manifold embedded in the ambient space. 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. Researchers have proposed many manifold learning algorithms, such as Isometric Feature Mapping (ISOMAP), Locally Linear embedding (LLE), and Local tangent space alignment (LTSA). This paper employed the Local Tangent Space Alignment (LTSA), which was to construct an approximation for the tangent space at each data point, and then align those tangent spaces to obtain the global coordinates of the data points with respect to the underlying manifold. Computational details and derivation of the algorithm can be found in (Zhang 2005). Due to the nonlinear nature, geometric intuition, and computational feasibility, manifold learning algorithms have attracted extensive attention recently.

3 MANIFOLD LEARNING-BASED THE PHASE SPACE RECONSTRUCTION

Based on Takens theorem, when \( m \geq 2d + 1 \), an attractor embedding can be obtained. Wherein \( m \) is the dimension of the phase space reconstruction, \( d \) is the dimension of the system, but this is only a sufficient condition, which can not help select \( m \) value. Ding et al. (1993) had proved that for noise-free, unlimited data, \( m \) could be taken the smallest integer which was greater than the correlation dimension; but for the noise and finite data, \( m \) was much larger than the dimension of the real data system.

In this paper, a financial time series was finite length and had noise, thus we proposed the manifold learning-based the phase space reconstruction of financial time series (MLPSR). In MLRPS algorithm, by correlation dimension method, we could obtain the Phase space dimension \( m \), which was far from the dimension of real system; Then through the phase space reconstruction technology, financial time series was extended to high-dimensional space, by which the contained information was fully revealed; Finally by use of manifold learning method, the embedding manifold and the intrinsic dimension of system were obtained. The detailed calculation process was as follows:

Given a financial time series \( x_1, \ldots, x_N \), the embedding dimension \( m \) and time delay \( \tau \) could be obtained by correlation dimension estimation. By use of Takens’ theorem, one-dimensional point sequence of the financial time series \( x_1, \ldots, x_N \) were embedded into the \( m \)-dimensional phase
space \( Y = [Y_1, Y_2, \ldots, Y_N] = \begin{bmatrix} x_1 & x_{1+r} & \cdots & x_{1+(m-1)r} \\ x_2 & x_{2+r} & \cdots & x_{2+(m-1)r} \\ \vdots & \vdots & \ddots & \vdots \\ x_N & x_{N+r} & \cdots & x_{N+(m-1)r} \end{bmatrix} \). Thus, we got a rough representation \( Y \) of the reconstructed phase space. Then we employed the LTSA algorithm to extract the intrinsic features of the rough phase space. LTSA was to construct an approximation for the tangent space at each data point, and then aligned those tangent spaces to obtain the global coordinates of the data points with respect to the underlying manifold. The following gave a simple description of the algorithm for the manifold learning about phase space.

Given the phase space \( Y = [Y_1, Y_2, \ldots, Y_N] \), the manifold \( Y' \in \mathbb{R}^{N \times d} \) (\( d \ll m \)) would be obtained by the following steps:

**Local neighbourhood construction:** For each \( Y_i, i = 1, \ldots, N \), determine its \( k \) nearest neighbors \( Y_{i_k}, j = 1, \ldots, k \) and form a neighborhood set \( Y_i' = [Y_{i_1}, \ldots, Y_{i_k}] \).

**Local linear fitting:** Compute the orthonormal basis matrix \( Q_i \) for the \( d \)-dimensional tangent space at \( Y_i \) based on the local neighbourhood \( Y_i' \). It can be taken as the matrix of \( d \) left singular vectors of \( Y_i'(I - ee^T / k) \) corresponding to its \( d \) largest singular values through the singular value decomposition as \( Y_i'(I - ee^T / k) = Q_i \sum_d V_i^T \), where \( e \) is a vector of all ones. Each data point \( Y_{i_j} \) in the neighbourhoods of \( Y_i \) is then projected to the computed tangent space as \( \theta_j(i) = Q_i^T (Y_{i_j} - \bar{Y}_i) \), where \( \bar{Y}_i \) is the mean of \( Y_{i_j} \)’s. Then, we can get \((N \times d)\) local coordinates \( \Theta_i = [\theta_1(i), \ldots, \theta_k(i)], i = 1, \ldots, (N \times d) \).

**Local coordinates alignment:** Align the \((N \times d)\) local projections \( \Theta_i = [\theta_1(i), \ldots, \theta_k(i)], i = 1, \ldots, (N \times d) \), to obtain the global coordinates \( g_i, i = 1, \ldots, (N \times d) \). Denote \( T_i = [g_{i_1}, \ldots, g_{i_k}] \) with the index set \( \{i_1, \ldots, i_k\} \) determined by the neighbors of each \( Y_i \). Let \( E_i = T_i'(I - ee^T / k) - L_i \Theta_i \) be the local reconstruction matrix. To minimize the local reconstruction error, the optimal alignment matrix \( L_i \) is given by \( L_i = T_i'(I - ee^T / k) \Theta_i^{-1} = T_i \Theta_i^*, \) where \( \Theta_i^* \) is the Moore-Penrose generalized inverse of \( \Theta_i \). Then, \( E_i \) can be written as \( E_i = TSW_i \), where \( S \) is the \( 0-1 \) selection matrix such that \( TS_i = T_i \) and \( W_i = (I - ee^T / k)(I - \Theta_i^* \Theta_i) \). The single data set manifold alignment of LTSA is achieved by minimizing the following global reconstruction error:

\[
\sum_i \|E_i\|^2 = \sum_i \|TSW_i\|^2_F = \|TSW\|^2_F, \quad \text{where} \quad S = [S_1, \ldots, S_{(N \times d)}] \quad \text{and} \quad W = diag(W_1, \ldots, W_{(N \times d)}) .
\]

To uniquely determine \( T \), impose \( TT^T = I_d \). The alignment matrix can be formed as \( B = SWW^T S^T \) to solve the optimal problem.

**Aligning global coordinate:** Compute the \( d + 1 \) smallest eigenvectors of \( B \), pick up the eigenvector matrix \( [u_2, \ldots, u_{d+1}] \) corresponding to the second to the \( d + 1 \) smallest eigenvalues, and set \( T = [u_2, \ldots, u_{d+1}]^T \). The results of \( T(e \in \mathbb{R}^{d \times (n \times L)}) \) then correspond to the global coordinates of the low-dimensional \( Y' \). The \( Y \) matrix can be reorganized to be a \( 3-D \) matrix with the size of \( d \times n \times L \). The \( 3-D \) \( Y' \) structure can be denoted by \( Y'_{d}(t, v) \). Each dimensional \( Y' \) has the same appearance as the \( Y \). Note that the dimension \( d \) is far less than the phase space dimension \( m \).
The $Y'$ reflects the underlying manifold of financial time series by reconstructing phase space, which represents the intrinsic nature of the financial system.

\section{Empirical Studies}

\subsection{Data Selection and Pre-processing}

The stock market has a very important role in the entire financial system. CSI 300, jointly issued by the Shanghai and Shenzhen Stock Exchange, reflects the overall trend of China A-share market, the health of China's securities market overview and stock price movements, which can be used as the evaluation criteria of investment performance. Therefore, we selected the closing price data of CSI 300 Index from January 2006 to December 2012 as the empirical study object, in which the total number was 1703.

As the impact of economic growth and inflation, we could not simply directly use the stock market price index, and need to eliminate the effects of economic growth and inflation on the price series. In this paper, we used Wavelet Analysis for data pre-processing. The wavelet decomposition about the CSI 300 was shown in Figure 1:

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{figure1.png}
\caption{Wavelet Analysis of The CSI 300 Index}
\end{figure}

As shown in Figure 1, through Wavelet Analysis the signal of CSI 300 was decomposed into the approximation coefficients vector and the detail coefficients vector. The noise signal was mainly contained in the detail coefficients vector, which was processed by some threshold, could be reconstructed into the de-noised smooth signals. Figure 2 showed the de-noised smooth signals of CSI300 by wavelet reconstruction:
4.2 Phase Space Reconstruction of Financial Time Series

With the correlation dimension method to estimate the time delay $\tau_d$ and data-dependent maximum time window $\tau_w$. Time delay $\tau_d$ was the first local minimum value of $\Delta S(t)$ corresponding to the time $t$, the maximum time window $\tau_w$ was the global minimum time $t$ corresponding to $S_{cor}(t)$. Time delay, the maximum time window and embedding dimension satisfied the quantitative relationship $\tau_w = (m - 1)\tau_d$, thus the embedding dimension $m$ could be further obtained which was required by the reconstructed phase space of financial system. Here we selected the CSI300 index to reconstruct the phase space of financial system, the value of optimum delay $\tau_d$ was 34, the time window was 280, and therefore the embedding dimension was 10. According Takens’ theorem, we could reconstruct the phase space of the financial system as follows: $X(t_i) = (x_i, x_{i+34}, ..., x_{i+34n}) \in R^{10}, (i = 1, 2, ..., n)$, wherein $n = T - (m - 1)\tau = 1703 - (10 - 1) \times 34 = 1397$.

4.3 Embedding Manifold of the Reconstructed Phase Space

In 4.3, we obtained the matrix $X_{1703 \times 10}$ about the reconstructed phase space of financial system. Here, through LTSA, the embedding manifold $X'$ of $X_{1703 \times 10}$ was got, the embedding dimension $m = 3$. the three-dimensional embedding results as shown in Fig.3:

![De-noised signals of The CSI 300 Index by wavelet reconstruction](image)

**Figure 2.** De-noised signals of The CSI 300 Index by wavelet reconstruction

![The three-dimensional embedding manifold of reconstructed phase space](image)

**Figure 3.** The three-dimensional embedding manifold of reconstructed phase space
4.4 Quality Evaluation of the Phase Space Reconstruction

In this paper, one-step-ahead error was employed to evaluate the quality of the reconstruction phase space. For the times series \( X = [x_1, \ldots, x_N] \), \( N \) was divided into two parts, \( T \) and \( L \), in which \( T \) data were used to construct the model, the \( L \) data were used for prediction. The existing data set were used for one-step-ahead prediction, the prediction results were compared with the already obtained results, and the prediction error was obtained, which the process was known as forward step prediction. Let \( \hat{x}_{1,T+1} \) be the predicted value of \( x_{1,T+1} \), \( \hat{x}_{2,T+1} \) be the predicted value of \( x_{2,T+1} \), \( \ldots \), \( \hat{x}_{M,T+1} \) was the predicted value of \( x_{M,T+1} \), then the predicted values of \( x_{i,T+2}, x_{i,T+3}, \ldots, x_{i,T+L} \) respectively were \( \hat{x}_{i,T+2}, \hat{x}_{i,T+3}, \ldots, \hat{x}_{i,T+L}, i = 1, 2, \ldots, M. \) For example, the prediction of the first variable as shown:

Let the relative prediction error \( NMSE = \frac{RMSE^2}{\sigma^2} \) be the evaluation index, wherein

\[
RMSE(\text{Root Mean Square Error}) = \sqrt{\frac{1}{L} \sum_{l=1}^{L} (x_{1,T+l} - \hat{x}_{1,T+l})^2}, \quad \sigma^2 \text{ was the variance of observations} \{x_{1,T+l}\}_{l=1}^{L}, \text{ the correlation coefficient is}
\]

\[
r = \frac{\sum_{l=1}^{L} x_{1,T+l} \hat{x}_{1,T+l} - \left( \sum_{l=1}^{L} x_{1,T+l} \right) \left( \sum_{l=1}^{L} \hat{x}_{1,T+l} \right)}{\sqrt{\sum_{l=1}^{L} (x_{1,T+l} - \left( \sum_{l=1}^{L} x_{1,T+l} \right))^2} \sqrt{\sum_{l=1}^{L} (\hat{x}_{1,T+l} - \left( \sum_{l=1}^{L} \hat{x}_{1,T+l} \right))^2}},
\]

then we got the prediction errors \( (NMSE_1, NMSE_2, \ldots, NMSE_L) \) of all the prediction variables. Here, the prediction errors about the phase space was \( Error_{mean} = \frac{1}{L} \sum_{l=1}^{L} NMSE_l \), and \( T = 1203, L = 500. \)

Thus, we got the prediction errors \( Error_{ad} = 0.0406 \) by our method, and the \( Error_{Cd} = 0.0936 \) by the traditional correlation dimension method. It was obvious that our method could improve the quality of phase space reconstruction.

5 CONCLUSION

This paper presented a new phase space reconstruction approach for financial time series based on manifold learning, namely manifold learning-based the phase space reconstruction of financial time series (MLPSR). The proposed MLPSR extracted the intrinsic features of the reconstructed phase space obtained by the traditional method, and thus reduced the noise interference. Experimental study showed our MLPSR method decreased the reconstruction error of phase space than the traditional method, and provided more accurate data support for the subsequent studies about financial system.
Reference


