Mining Latent Sources of Causal Time Series Using Nonlinear State Space Modeling

Wei-Shing Chen and Fong-Jung Yu

Department of Industrial Engineering and Technology Management,
Da-Yeh University, 168, University Rd. Dacun, Changhua, 51591, Taiwan, R.O.C.
weishing@mail.dyu.edu.tw

Abstract. Data mining refers to use of new methods for the intelligent analysis of large data sets. This paper applies one of nonlinear state space modeling (NSSM) techniques named nonlinear dynamical factor analysis (NDFA) to mine the latent factors which are the original sources for producing the observations of causal time series. The purpose of mining indirect sources rather than the time series observation is that much better results can be obtained from the latent sources, for example, economics data driven by an "explanatory variables" like inflation, unobserved trends and fluctuations. The effectiveness of NDFA is evaluated by a simulated time series data set. Our empirical study indicates the performance of NDFA is better than the independent component analysis in exploring the latent sources of Taiwan unemployment rate time series.

Keywords: Data mining, latent sources, time series, nonlinear state space modeling, nonlinear dynamical factor analysis.

1 Introduction

Starting from 1980s, modeling time series has become a popular theme [1] for understanding the behaviors of the dynamical process by observing responses. A time series is sequence of regularly sampled quantities from an observed process which is frequently driven by a rather small, typically unobservable set of influences. For example, when there are a multivariate time series that represents the quantity sold and price associated with demand for a product or service, it is often desirable to decide the influences of price and the demand. Such the latent time series describe the underlying data-generating process of a dependent time series using some latent sources (explanatory variables). The developments of this paper were motivated by the need for exploring the driven processes for the given observations. These driven processes contain quantities that cannot be measured direct. Instead, only a portion of noise-corrupted observation is available. Such an effort is needed because the measurement process is either more expensive or more impossible for the latent time series [2]. The present work is therefore interested in the problem of reverse engineering of a time series where one uses a set of given time series x_t to get another hidden set of time series s_t.

In this paper, we used the state space model (SSM) as the data mining technique to explore the latent sources. SSM is a model comprised of two parts: states and

N.T. Nguyen, C.-G. Kim, and A. Janiak (Eds.): ACIIDS 2011, LNAI 6591, pp. 137–148, 2011.

observations. The SSM is one of the most powerful methods for modeling a dynamic system and has been widely employed for engineering control systems[3]. The remaining of the paper is organized as follows. In Section 2 we briefly introduce how an SSM can be used to learn dynamical input-output mappings. In Section 3 we discuss the issue of mining hidden sources from multivariate time series by using nonlinear dynamical factor analysis (NDFA). The methods for deciding the parameters used in NDFA are given in Section 4. In Section 5, an application study is carried out and the results are compared with the FastICA method. The paper is concluded in Section 6.

2 State-Space Models

An SSM describes evolving two time series running in parallel, one referred to as the state process (s_t) and the other as the observation process (x_t). An SSM assumes that an observation vector x_t is generated by its latent sources s_t through an unknown transformation matrix Z and additive observation disturbance n_t:

$$x_t = Zs_t + n_t. \tag{1}$$

In (1), the transformation matrix $\boldsymbol{Z}$ is a parameterized mapping from one state space to an observation space. In a dynamical process, a current source s_t can also be generated through another transformation matrix $\boldsymbol{T}$ from the sources s_{t-1} at the previous time instant as follows

$$s_t = Ts_{t-1} + m_t . \tag{2}$$

The system noise m_t and the observational noise n_t noise can be Gaussian or non-Gaussian. The observation equation (1) and state equation (2) form a linear state-space representation for the dynamic behavior of $\boldsymbol{x}$. An SSM provides an important body of techniques for analyzing time series, only the observations $\boldsymbol{x}$ are known earlier, and both the states $\boldsymbol{s}$ and the mappings $\boldsymbol{Z}$ and $\boldsymbol{T}$ are learned from the data.

Many statistical methods [4] have been developed to find latent sources from the observation data. Three types of inference, including filtering, smoothing and prediction, for the state and the model from the observations $\boldsymbol{x}$ are commonly discussed in the literature [5]. For linear SSMs with Gaussian process and observation noise, the Kalman filter [6] and Kalman smoothing [7] are the well-known methods of choice for the consistent estimation of the indirectly observed or unobserved states. For many real-world data, the affect of the desired sources to the observed data is, however, not linear. Therefore, an SSM supports nonlinear state space models (NSSM) in the form (3) and (4) are proposed [8] where $\boldsymbol{\theta}$ is a vector containing the model parameters and time $\mathbf{t}$ is discrete.

$$x_t = Z(s_t, \theta_z) + n_t. \tag{3}$$

$$s_t = T(s_{t-1}, \theta_t) + m_t . \tag{4}$$

An SSM is a model to describe how $\boldsymbol{s_t}$ generates or "causes" $\boldsymbol{x_t}$ and $\boldsymbol{s_{t+1}}$. The goal of mining in this paper is to invert this mapping, that is, to mine $\boldsymbol{s_{1:t}}$ from the given $\boldsymbol{x_{1:t}}$ .Methods have been proposed for learning a linear SSM [9], but estimating a nonlinear SSM is inherently more difficult. Indeed, it is almost impossible to find any simple

closed form solution for the distribution of the sources in (3) and (4). Because in an NSSM, the statistical learning problem is no longer solved in the closed form as opposed to the linear SSM and this raises several computational difficulties. Quach et al. [10] used an unscented Kalman filtering (UKF) technique to tackle nonlinearities by deriving NSSMs from ordinary differential equations (ODEs). The nonlinear mapping in (3) and (4) can be modeled by a multilayer perceptron (MLP) networks or a radial basis function (RBF) network which suit well to modeling both strong and mild nonlinearities [11]. In this paper, we apply a method called nonlinear dynamical factor analysis (NDFA), which is based on variational Bayesian learning [12, 13], for learning form of (3) and (4) in an NSSM of a dynamic process. Variational Bayesian methods, also called ensemble learning[14], are a group of methods to approximate intractable integrals arising in Bayesian inference and machine learning.

3 Nonlinear Dynamical Factor Analysis

This section briefly introduces the NDFA model and its learning algorithm. A comprehensive discussion can be found in [12, 13]. The NDFA model is a dynamical extension of nonlinear factor analysis (NFA) [11]. The function of NDFA is to find dynamic sources which explain nonlinearity of the observed data[8]. In NDFA, the mapping from sources to observations is modeled by a multilayer perceptrons (MLP) to represent the nonlinearity. Like in figure 1, the sources are on the top layer and observations in the bottom layer. The middle layer consists of hidden neurons of sigmoidal nonlinearities each of which estimates a nonlinear function of the inputs as

$$x_t = Z(s_t, \theta_Z)+n_t = B\phi(As+a)+b+n_t = B\tanh(As+a)+b+ n_t . \quad (5)$$

where the matrices A and B are the weights of the hidden and output layer and a and b are the corresponding biases. The *tanh* is a common nonlinear activation function ϕ being used in the MLPs. The function T has a similar mapping structure except the MLP network is used to model only the change in the state values.

$$T(s)=s+D\phi(Cs+c)+d=s+D\tanh(Cs+c)+d . \quad (6)$$

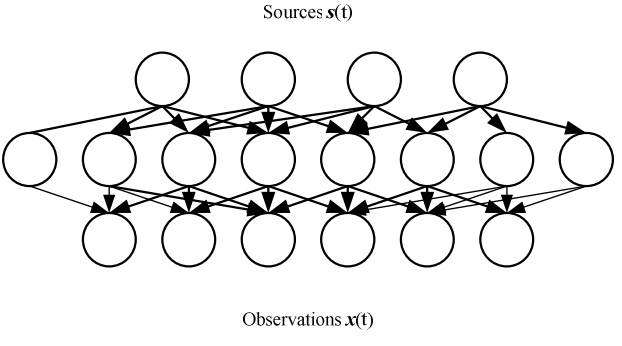

Fig. 1. Mapping structure from sources to observations using a MLP network

All the assumptions of NDFA are expressed in the form of the density model $\boldsymbol{p}(\boldsymbol{X}, \boldsymbol{S}, \boldsymbol{\theta})$. The parameters $\boldsymbol{\theta}$ consist of the parameters of the MLPs in (5) and (6), the variance parameters of the noise terms $\boldsymbol{n}_t$ and $\boldsymbol{m}_t$ as well as hyperparameters. To estimate both the unknown nonlinear function $\boldsymbol{Z}$ and $\boldsymbol{T}$ and the unknown sources $\boldsymbol{s}_t$, the general approach is to estimate the posterior probability distribution of the unknown parts of the model by using the Bayesian approach where all the assumptions made in the model are expressed in the form of the joint distribution of the observations $\boldsymbol{X}$, states $\boldsymbol{S}$ and parameters $\boldsymbol{\theta}$ of the model

$$p(x,s,\theta)=p(x|s,\theta)p(s|\theta)p(\theta). \tag{7}$$

Gaussian distributions are used to describe the weights of the MLPs for computational tractability. Based on the Bayesian approach, once the joint distribution (7) is defined and the set of observations is obtained, all the relevant information about the unknown parameters is contained in the posterior

$$p(s,\theta|x)=p(s,\theta,x)/p(x). \tag{8}$$

In general, one would find point estimates of the unknowns by maximizing the posterior (8), which yields the well-known maximum a posteriori (MAP) estimate. However, this approach would easily cause over-fitting problems especially in this ill-posed estimation problem. So, Giannakopoulos and Valpola [15] proposed an ensemble learning technique to approximate the actual posterior probability $p(s,\theta|x)$ (8) by an approximating distribution $q(s,\theta)$ over $\boldsymbol{S}$ and $\boldsymbol{\theta}$. The structure of the model and the learning scheme were optimized based on the value of a cost function which is measured by the Kullback-Leibler divergence:

$$C_{KL}=\int q(s,\theta)log(q(s,\theta)/p(s,q|x)d\theta ds. \tag{9}$$

A more comprehensive explanation of the algorithm is given in [12, 16], and an MATLAB software implementation for the NDFA technique is available at [17].

4 Setting the Parameters for NDFA Algorithm

To apply the NDFA algorithm, a number of parameters including the number of sources, the number of hidden neurons, type of activation function, and an approximation method need to be determined. For the practicable applying the NDFA algorithm in data mining of latent sources, we outline the guidelines for setting up the parameters step-by-step as follows:

Step 1: Convert a one-dimensional time series to a state space with embedded dimensions for modeling the dynamics of the causal time series.

Nonlinear factor analysis is used to extract the undying state from a sequence of observations. Phase space embedding techniques are the standard methods to analyze nonlinear dynamical systems. For a one-dimensional time series data, we need to convert it to an embedded data vector $x(t)=[x^T(t)x^T(t-\tau)x^T(t-2\tau)\ldots x^T(t-m\tau)]$ by using the Takens' embedding theory [18]. The delay parameter (τ) can be determined as the first minimum of the mutual information (MI) function [19] or the first zero of the autocorrelation function. An embedding dimension (m) can be decided by using the

mutual information. To determine the proper embedding dimension m, the false nearest neighbors (FNN) method [20] can be used.

Step 2: Determine number of latent sources and set up the initial values for the sources.

According to the Fig 1, mining latent sources using NDFA requires prior knowledge of the number of sources before the ensemble learning can be performed. Here, we applied ICA approach and used a measure of the number of sources based on the normalized determinant value of the global matrix (G) to determine the number of sources in the given observations [21]. The global matrix G is the product of estimated the mixing matrix and unmixing matrix. It can be stated that if |G| is near to zero, then it points out there are some dependent sources. So, the number of sources should be fewer than the number of variables of observations. Otherwise, if |G| is near to one, the number of sources can be assumed to be equal to the number of observational variables. FastICA [22] can be applied to obtain the mixing matrix and unmixing matrix. Another alternative to determine the number of latent sources is based on the eigenvalue spectrum of the data covariance matrix. Everson & Roberts [23] proposed a method of inferring the true eigenvalue spectrum from the sample spectrum.

Because of the flexibility of the MLP network and the gradient based learning algorithm, the NDFA is sensitive to the initialization. Honkela and Valpola suggested to use linear PCA for initialization of the means of the sources S [16].

Step 3: Determine number of hidden neurons in the mappings Z and T in (3) and (4) which are modeled by an MLP network.

There is no magic formula for selecting the best number of hidden neurons. Santos et al. [24] proposed a novel technique to estimate the number of hidden neurons of an MLP. The proposed approach consists in the post-training application of SVD/PCA to the back-propagated error and local gradient matrices associated with the hidden neurons. The number of hidden neurons is then set to the number of relevant singular values or eigenvalues of the involved matrices. Some thumb rules are also available for calculating number of hidden neurons such as the following:

1. The number of hidden neurons should be between the size of the input layer and the size of the output layer.
2. The number of hidden neurons should be 2/3 the size of the input layer, plus the size of the output layer.
3. The number of hidden neurons should be fewer than twice the size of the input layer.

Step 4: Decide nonlinear activation function to use in the MLP. Typical choices for ϕ include *tanh* or the logistic *sigmoid* function.

In case of the inputs of the hidden neurons are not Gaussian, use the standard logistic sigmoid activation function in the form of *tanh* which is more common and faster to evaluate numerically.

Step 5: Decide the approximation method for the nonlinearity.

Since it is impossible to find any closed form for the distribution of the sources, we should use a fixed form approximation where the form is fixed and only the parameters are optimized by minimizing the cost function. Two natural choices for the functional form of the approximation for the nonlinearity are Taylor approximation for low

variance inputs and Gauss-Hermite quadrature approximation for high variance inputs and adaptive approach based on the combination of Gauss-Hermite quadrature formula [25]. The Taylor approximation method fails in case of high input variance because it relies on information of the activation function at a single point.

5 Application to Unemployment Time Series

This section takes advantage the NDFA technique to detect presence of latent sources for characterizing the evolution of the unemployment transition in Taiwan. Unemployment and labor market variability typically shows a complex behavior and it is difficult to identify specific patterns in the long run economic cycle. The time series is the monthly registered UR in Taiwan, as published by the Directorate General of Budget, Accounting and Statistics (DGBAS) of Executive Yuan. The series covers the period between January 1978 and June 2010, for 390 values. Figure 2 provides a visual representation of the time series with the y-axis defined as the unemployment rate and the x-axis as the time index. From an eyeball inspection of the plotted series, it seems obvious that this series is nonstationary and seasonal. Table 1 shows the descriptive statistics of the data. The Jarque-Bera test is a goodness-of-fit measure of departure from normality, based on the sample kurtosis and skewness. The wide vertical spaces in the 2000s until the 2010s show high frustration rates throughout the period. A detail discussion of the nonstationarity of the Taiwan unemployment rate can be found in [26].

Table 1. Basic descriptive statistics of the data

Mean	Std Dev	Skew	Kurtosis	C(6)	Max value	Min value	Jarque-Bera
2.79	1.35	0.674	-0.667	-1.330	6.13	0.860	36.6860

The data has enough values for us to apply the techniques developed while it also covers a period of time that is sufficiently homogeneous for us to be able to adequately analyze and characterize the evolution in Taiwan unemployment. According to statistic in Table 1, if the UR random variable sequences obey independent same normal distribution, its Skewness should equal zero and Kurtosis is three. If the random variable sequences obey normal distribution, the Jarque-Bera statistic should obey a chi-square distribution with freedom 2, standard value of which is 9.21 and 5.99 under 1% and 5% significance level respectively. From Table 1, we see the distribution of UR far deviates from normal distribution and the shape is heavy-tailed. The result of remarkable deviation suggests that UR time series may have nonlinear dynamic structure.

The analysis is conducted in the Matlab environment using the NDFA package[17]. Before conducting the analysis, several parameters needed to be set up. First, we convert the time series into an embedded data vector with proper dimensions and delays to model the dynamics of the time series. As shown in Fig 3, the MI function $I(\tau)$ exhibits a local minimum at $\tau=6$ time steps under four cases of different number of boxes for partition. Thus, we should consider $\tau=6$ to be the best delay time in this study. Fig 4 shows the application of FNN method by maximum norm distance yields an embedding dimension m = 5. The embedded data vector was $x(t)=[x^T(t)x^T(t-6)x^T(t-12)x^T(t-18)x^T(t-24)]$.

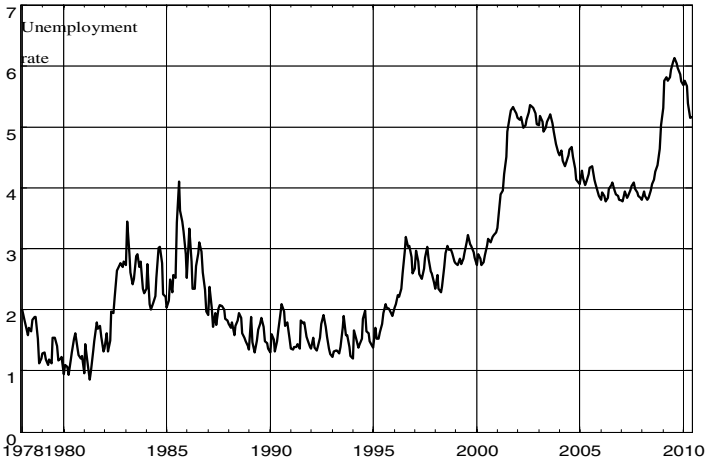

Fig. 2. The time-series representing the historical monthly movement of The UR: Jan 1978 - June 2010

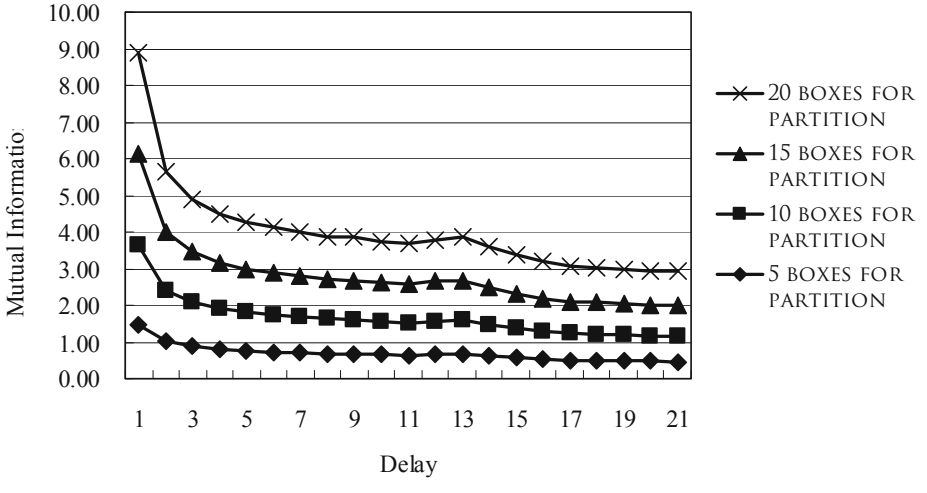

Fig. 3. Mutual information versus delay time

We then estimate the number of sources for searching. Based on the Naik and Kumar [21] method, the determinant of global matrix parameters of ICA is used to estimate the number of search sources. We calculated $|G|=2.6557e\text{-}035\approx 0$ which shows several dependent sources are presented. Observing the plot of eigenvalues of covariance matrix in Figure 6, we eliminated three signals from the five observations and set two latent sources for searching.

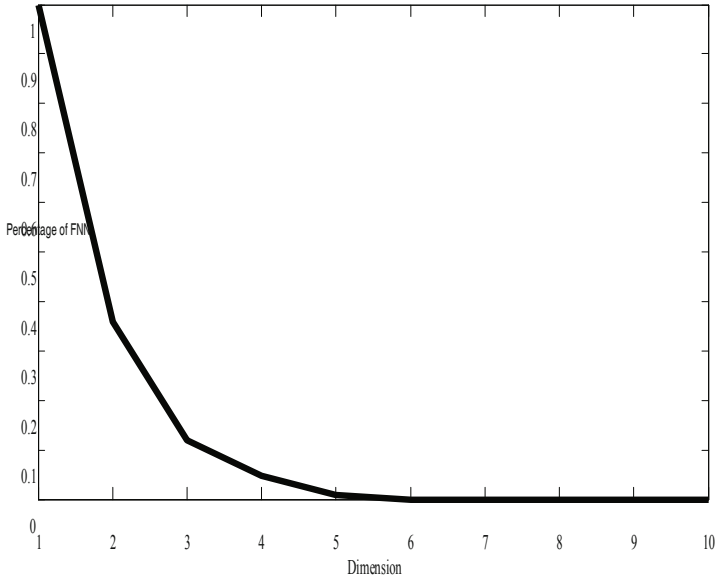

Fig. 4. Percentage of FNN vs m

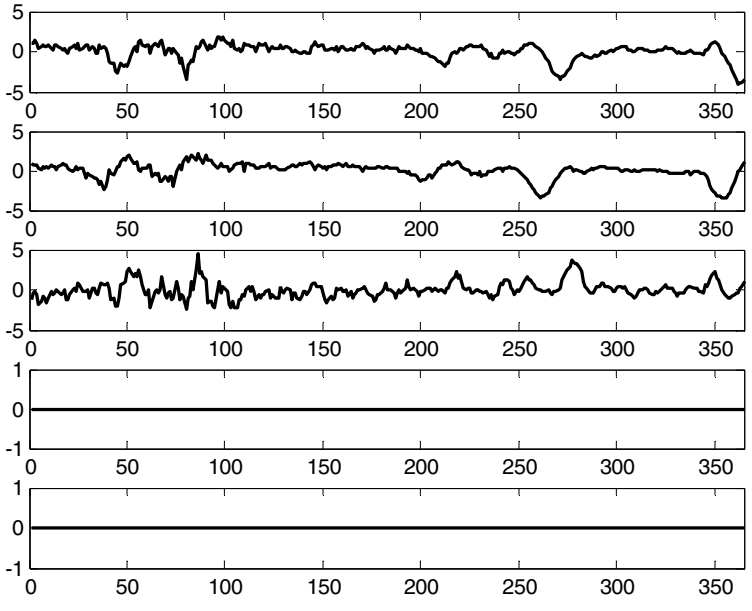

Fig. 5. Sources generated from FastICA

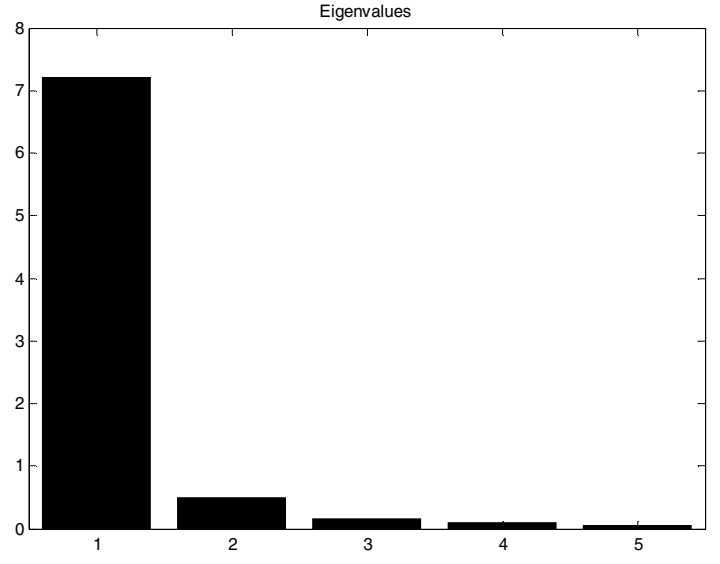

Fig. 6. From the largest to the smallest eigenvalues of covariance matrix

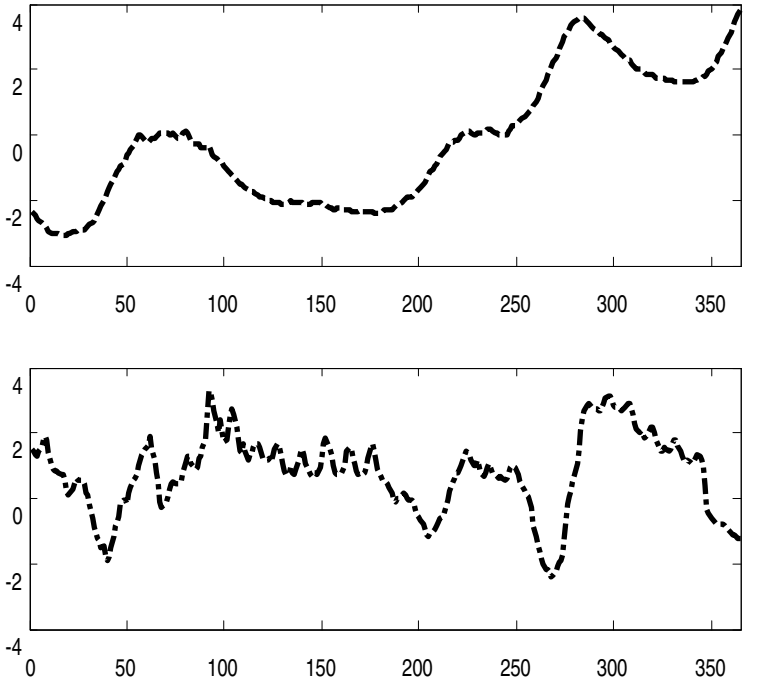

Fig. 7. The two latent sources generated from NDFA

Given two number of search sources, the sources were initialized using PCA. Following the second rule-of-thumb method of the step 3, we set number of both layer hidden neurons to seven. The nonlinear activation is the tanh function. The Gauss-Hermite quadrature approximation is selected as the approximating method for the nonlinearity. We set 500 iterations to run the algorithm. Figure 7 shows the two latent sources of controlling the evolution of Taiwan unemployment rate. The sources have different interpretations. The first latent source drives the unemployment rate in an increasing trend. The second latent source forces the employment market up and down because of some positive and negative forces which include both economic and political issues. Figure 8 shows two latent sources produced from independent component analysis by using FastICA package [22]. These two latent sources reflect similar variation pattern and provide less knowledge for inferring the causality of the time series. Comparing Figure 7 with Figure 8, we found the NDFA can extract one latent trend source and one variation source while the FastICA would extract both two variation sources.

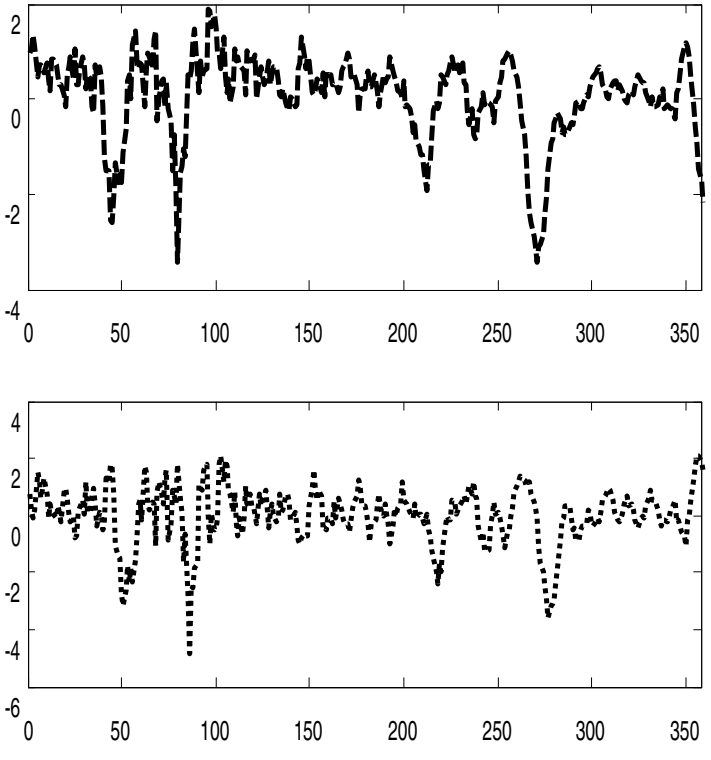

Fig. 8. The two latent sources generated from FastICA

6 Conclusions

It is a tempting alternative to try NDFA on the unemployment time series data. To assume having some underlying independent components and linear transformation in

this specific application may be unrealistic. This paper presents a unified data mining framework for jointly defining process dynamics models and measurements taken on the process. The framework is a state-space model where the process is modeled by the state process and measurements are modeled by the observation process. Parameter estimation and estimation of state process variables can be conducted using MLP with ensemble learning procedures.

Acknowledgements. This work was supported in part by the National Science Council of Republic of China under the NSC-99-2221-E-212-010.

References

1. Makridakis, S.: Time series prediction: Forecasting the future and understanding the past. In: Weigend, A.S., Gershenfeld, N.A. (eds.), p. 643. Addison-Wesley Publishing Company, Reading (1993), ISBN 0-201-62; International Journal of Forecasting 10, 463–466 (1994)
2. Hu, X., Xu, P., Wu, S., Asgari, S., Bergsneider, M.: A data mining framework for time series estimation. Journal of Biomedical Informatics 43, 190–199 (2010)
3. Chen, C.T.: Linear System Theory and Design, 3rd edn. Oxford University Press, New York (1999)
4. Everitt, B.S., Dunn, G.: Applied Multivariate Data Analysis. Oxford University Press, New York (1992)
5. West, M., Harrison, J.: Bayesian Forecasting and Dynamic Models. Springer, New York (1990)
6. De Jong, P.: The diffuse Kalman filter Annals of Statistics 19 (1991)
7. Anderson, B.D.D., Moore, J.B.: Optimal filtering. Prentice-Hall, Englewood Cliffs (1979)
8. Ilin, A., Valpola, H., Oja, E.: Nonlinear dynamical factor analysis for state change detection. IEEE Transactions on Neural Networks 15, 559–575 (2004)
9. Overschee, P.v., Moor, B.D.: Subspace Identification for Linear Systems: Theory, Implementation Applications. Springer, Heidelberg (1996)
10. Quach, M., Brunel, N., d'Alché-Buc, F.: Estimating parameters and hidden variables in nonlinear state-space models based on ODEs for biological networks inference. Bioinformatics (2007)
11. Lappalainen, H., Honkela, A.: Bayesian Nonlinear Independent Component Analysis by Multi-Layer Perceptrons. In: Girolami, M. (ed.) Advances in Independent Component Analysis, pp. 93–121. Springer, Heidelberg (2000)
12. Valpola, H., Karhunen, J.: An unsupervised ensemble learning method for nonlinear dynamic state-space models. Neural Comput. 14, 2647–2692 (2002)
13. Giannakopoulos, X., Valpola, H.: Nonlinear dynamical factor analysis. In: Bayesian Inference And Maximum Entropy Methods in Science And Engineering: 20th International Workshop. AIP Conference Proceedings, vol. 568 (2001)
14. Barber, D., Bishop, C. (eds.): Ensemble learning in Bayesian neural networks. Springer, Berlin (1998)
15. Giannakopoulos, X., Valpola, H.: Nonlinear dynamical factor analysis. In: AIP Conference Proceedings, vol. 568, p. 305 (2001)
16. Honkela, A., Valpola, H.: Unsupervised variational Bayesian learning of nonlinear models. In: Saul, L.K., Weis, Y., Bottou, L. (eds.) Advances in Neural Information Processing Systems (NIPS 2004), vol. 17, pp. 593–600 (2005)

17. Valpola, H., Honkela, A., Giannakopoulos, X.: Matlab Codes for the NFA and NDFA Algorithms (2002), http://www.cis.hut.fi/projects/bayes/
18. Takens, F.: Detecting strange attractors in turbulence. LNM, vol. 898, pp. 366–381. Springer, Heidelberg (1981)
19. Fraser, A.M., Swinney, H.L.: Independent coordinates for strange attractors from mutual information. Physical Review A 33, 1134 (1986)
20. Sprott, J.C.: Chaos and Time Series Analysis, vol. 507. Oxford University Press, Oxford (2003)
21. Naik, G.R., Kumar, D.K.: Determining Number of Independent Sources in Undercomplete Mixture. EURASIP Journal on Advances in Signal Processing 5, Article ID 694850 (2009), doi:10.1155/2009/694850
22. Gävert, H., Hurri, J., Särelä, J., Hyvärinen, A.: FastICA Package (2005), http://www.cis.hut.fi/projects/ica/fastica/code/dlcode.shtml
23. Everson, R., Roberts, S.: Inferring the eigenvalues of covariance matrices from limited, noisy data. IEEE Transactions on Signal Processing 48, 2083–2091 (2000)
24. Santos, J.e.D.A., Barreto, G.A., Medeiros, C.a.M.S.: Estimating the Number of Hidden Neurons of the MLP Using Singular Value Decomposition and Principal Components Analysis: A Novel Approach. In: 2010 Eleventh Brazilian Symposium on Neural Networks, pp. 19–24 (2010)
25. Honkela, A.: Approximating Nonlinear Transformations of Probability Distributions for Nonlinear Independent Component Analysis. In: Proceedings of the 2004 IEEE International Joint Conference on Neural Networks (IJCNN 2004), Budapest, Hungary, pp. 2169–2174 (2004)
26. Chen, W.-S.: Use of recurrence plot and recurrence quantification analysis in Taiwan unemployment rate time series. Physica A: Statistical Mechanics and its Applications (in Press, 2011), doi:10.1016/j.physa.2010.12.020