Building Time Series Forecasting Model By Independent Component Analysis Mechanism

Jin-Cherng Lin[1], Yung-Hsin Li[1] and Cheng-Hsiung Liu[2]
[1]Dept. of Computer Science & Engineering Tatung University
Taipei 10451, Taiwan, Email: jclin@ttu.edu.tw, acc_andrew@msn.com
[2]Dept. of Management Information System Ta Hwa Institute of Technology
Hsinchu 307, Taiwan, Email: lhwaliu@yahoo.com.tw

Abstract—**Building a time series forecasting model by independent component analysis mechanism presents in the paper. Different from using the time series directly with the traditional ARIMA forecasting model, the underlying factors extracted from time series is the forecasting base in our model. Within component ambiguity, correlation approximation and mean difference problems, independent component analysis mechanism has intrinsic limitations for time series forecasting. Solutions for those limitations were purposed in this paper. Under the linear time complexity, the component ambiguity and mean difference problem was solved by our proposed evaluation to improve the forecasting reward. The empirical data show that our model exactly reveals the flexibility and accuracy in time series forecasting domain.**

Keywords: Independent component analysis(ICA), Autoregressive(AR), ambiguity, correlation.

1 Introduction

Forecasting has been a necessary technique for economists, scientists and government leaderships nowadays. Combing the independent component analysis (ICA) [1] and autoregressive (AR) [2] concepts, we propose an automatic time series forecasting model to improve the prediction accuracy.

Gross national product(GNP), consumer price index(CPI) and people satisfaction of government implementations ... et al. affect the volatility of financial time series. Making them as underlying factors, we hope to retrieve them from financial time series by independent component analysis mechanism. Independent component analysis (ICA) is a mature technique [1] in signal processing domain for finding underlying factors in mixed signals. We adapt this concept into forecasting the time series data [2], and use it to separate the underlying factors of time series. Assume some factors affect lots time series simultaneously, so those time series change when these factors move. We separate those factors from observed time series, forecasting their behaviors, and retrieve predicted results.

Simple, inexpensive and effective, for the simplicity nature, autoregressive (AR) model is a famous and popular model for predicting time series [2] [4]. In our model, we merge it into the ICA forecasting mechanism and call it ICA-AR forecasting phase.

Independent component analysis (ICA) is a branch of factor analysis. Making the factors as independent as possible [1], ICA owns a different process for component processing with conventional factor analysis concepts. Independent component analysis technique, in practical, is ineffectiveness applied on building financial forecasting model. Back A.D. and Weigend A. S.[5] use threshold and convoluting mixing techniques to modeling the stock prices. But the convolution time period and threshold value of their research are undecidable. Mălăroiu et al.[3] adopted ICA, smoothing and autoregressive(AR) techniques to form a financial forecasting model. They claimed the long-term trend of forecasted series is similar with the original series. However, the forecasted series loss the extreme values of the original ones. Mok et al.[6] used linear regression or artificial neural network(ANN) to solve the component ambiguity problem. Those techniques, especially in ANN, cost lots amount of time complexity. Below we describe basic ICA process and its limitations for building financial forecasting model.

2 Forecasting Model Architecture

Generally speaking, ICA preprocessing and ICA-AR forecasting are two major phases in our model. Figure 1 illustrates the framework of our model.

2.1 ICA Preprocessing Phase

Different from the general ICA procedure [1], ICA is just a preprocessing step in our forecasting model. For retrieving affecting time series factors, we adopt ICA technique. Before using ICA technique, we assume three topics for ICA feasibility [1]. First, We state the number of the time series is the same as the original mixing signals. Then,

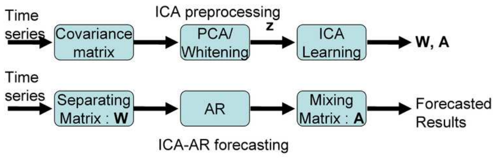

Figure 1: Our model architecture. The upper is the ICA preprocessing stage, and the lower is ICA-AR forecasting stage

the mixing process is a linear matrix and full-rank. Third, we assume the original signals are non-Gaussian and independent with each others. Let the observed time series denote $\mathbf{x}$, linearly mixing procedure $\mathbf{A}$, and underlying factors $\mathbf{s}$. The linearly mixing procedure describes in the form:

$$\mathbf{x} = \mathbf{A}\mathbf{s} \tag{1}$$

The ICA goal is to find another linear transformation process $\mathbf{W}$ to retrieve the underlying factors $\mathbf{s}$.

$$\mathbf{s} = \mathbf{W}\mathbf{x} \tag{2}$$

For retrieving the separating process matrix $\mathbf{W}$, general ICA process includes three major procedures : Principle component analysis (PCA), whitening and ICA learning procedure. Beside the last step, the front two steps are closed-form formulation procedures, so their operations cost few computation time [1]. For reducing the computation cost in the third learning procedure, we choose a fixed-point ICA learning algorithm : FastICA [1][3]. Below three subsections present details of these three major procedures.

2.1.1 Principle Component Analysis (PCA)

The covariance matrix presents the relationships among of the multivariate data [1]. PCA is the first step to find the principle components from the covariance matrix of the observed time series.

$$Cov_{\mathbf{x}} = \sum (\mathbf{x} - \mathbf{m_x})(\mathbf{x} - \mathbf{m_x})^{\mathbf{T}}, \mathbf{m_x} = \frac{\mathbf{1}}{\mathbf{n}} \sum_{\mathbf{i=1}}^{\mathbf{n}} \mathbf{x_i} \tag{3}$$

The covariance matrix is semi-definite, so its eigenvalues exist and are larger than zero [1]. By the eigenvalue decomposition (EVD) in the linear algebra, we gain the eigenvalue and eigenvector matrix from the covariance matrix of the observed time series $\mathbf{x}$.

$$Cov_{\mathbf{x}} = EDE^T \tag{4}$$

Each row vector under the $\mathbf{E}$ is eigenvector and orthogonal with other row vector, so $\mathbf{E}$ denotes eigenvector matrix. $\mathbf{D}$ is diagonal eigenvalue matrix, and each value on the diagonal line is an eigenvalue of the covariance matrix of the observed time series $\mathbf{x}$. We name the procedure as principle component analysis (PCA) because we can get the principle components from the covariance matrix of the observed time series $\mathbf{x}$ after the observed time series $\mathbf{x}$ performs the linear transformation of the EDE^T.

2.1.2 Whitening

PCA procedure reveals the principle components the the covariance matrix of the observed time series $\mathbf{x}$, however, magnitudes with each components differ from with each others. For the independent requirement [1], we reuse the PCA result and perform another linear transformation process. Reusing the eigenvector matrix $\mathbf{E}$ and eigenvalue matrix $\mathbf{D}$, we perform the whitening process $\mathbf{V}$ to equalize every components magnitude.

$$\mathbf{V} = \mathbf{D}^{-\mathbf{1/2}}\mathbf{E}^{\mathbf{T}} \tag{5}$$

Then making linearly transformation of the time series, we retrieve the uncorrelated (PCA process) and equal variance (whitening process) time series components $\mathbf{z}$.

$$\mathbf{z} = V\mathbf{x} \tag{6}$$

2.1.3 ICA Learning Procedure

The ICA objective is making each component as independent as with each others possible. We propose a linear transformation to achieve the goal.

$$\mathbf{y} = \hat{s} = \mathbf{W}\mathbf{z} \tag{7}$$

To find the independent components, we set the component $\mathbf{y}$ as the estimated value of the whitened components $\mathbf{z}$. We hope to find a linear transformation from the whitened components $\mathbf{z}$ to independent components. Different from above procedures, translating components to independent ones require learning process instead of closed form procedure [1]. The ICA algorithm combines two goals : unsupervised learning and optimization problem. For the inexactly target output value, unsupervised learning algorithm fits the learning mechanism for the independent requirement(8). For achieving the independent output component under the constrains that each norm of $\mathbf{W}$ column vector satisfies one(9). In the view of two dimension, the length of each $\mathbf{W}$ locates on unit circle boundary. We adopt the FastICA learning algorithm [1].

$$\mathbf{w_{t+1}} \leftarrow E\{\mathbf{z}g(\mathbf{w_t}^T\mathbf{z})\} - E\{g'(\mathbf{w_t}^T\mathbf{z})\}\mathbf{w_t} \tag{8}$$

$$\|\mathbf{w}\| = \sqrt{w_{n1}^2 + w_{n1}^2 + \cdots + w_{nn}^2} = 1 \tag{9}$$

E{} means expectation value in the statistics. In the initial condition, we fill the matrix **W** in orthogonal unit vectors for the orthogonal requirement [1]. For modifying the separating matrix **W**, performing a quantitative value, a contrast function was used here. In the unsupervised learning equation(8), two contrast function g() and g'() observed. The contrast function g() and it derivation g'() are dependent on the Gaussian features of each component. General propose, super-Gaussian and sub-Gaussian are major three types for identifying the data Gaussian feature. Each Gaussian feature requires two contrast function g() and it derivation g'(). Below lists all contrast functions in those the three Gaussian features.

Table 1. Contrast functions list for Gaussian features

	g()	g'()
General purpose	$\frac{1}{a_1} * log(cosh(a_1 * z))$	$tanh(a_1 * z)$
Super Gaussian	$-\exp(-\frac{1}{2}z^2)$	$z * \exp(-\frac{1}{2}z^2)$
Sub Gaussian	$\frac{1}{4}z^4$	z^3

The value a_1 is limited between 1 and 2 [1].

2.2 ICA-Autoregressive (ICA-AR) Forecasting Procedure

After the ICA procedures, we get the separating matrix **W**. Major different from the general ICA approach, we reuse the separating matrix **W** and mixing matrix **A** in our forecasting model, instead of the separated components. And we insert the separating matrix **W** and mixing matrix **A** into the forecasting procedure, so we call the phase : ICA-AR forecasting phase. Separating series to factors, factors forecasting and retrieving predicted time series are major three procedures in the ICA-AR forecasting phase.

2.2.1 Separating time series to factors

For forecasting the time series, we consider the underlying factors of the time series. First step in our forecasting process is to get the factor components from the time series. We reuse the separating matrix **W** from the ICA preprocessing phase [3].

$$\mathbf{f}(t) = \mathbf{W}\mathbf{x}(t) \tag{10}$$

The $\mathbf{f}$(t) denotes underlying factors of the observed time series $\mathbf{x}$(t).

2.2.2 AR forecasting

This is the forecasting mechanism in our model. Focus on the underlying factors instead of time series, we hope to forecast factor time series $\mathbf{f}$(t).

$$\mathbf{f}^p(t+1) = \mathbf{q}[\mathbf{f}(t), \mathbf{f}(t-1), \ldots\ \mathbf{f}(t-k)]\mathbf{x}(t) \tag{11}$$

Using k-order autoregressive (AR) model, we get the predicted factor components $\mathbf{f}^p(t)$.

2.2.3 Forecasted time series

The factor components $\mathbf{f}^p(t+1)$ is forecasted factor components series instead of the predicted time series, but transferring the factor components to time series is our final goal. Inversing the separating matrix **W** in the ICA preprocessing phase, we gain the mixing matrix **A** in ICA mixing procedure. Our model performs a linear transformation by mixing the forecasted factor components $\mathbf{f}^p(t+1)$ to the forecasted time series $\mathbf{T}^p(t+1)$.

$$\mathbf{A} = \mathbf{W}^{-1} \tag{12}$$

$$\mathbf{T}^p(t+1) = \mathbf{A}\mathbf{f}^p(t+1) \tag{13}$$

3 Limitations Solutions And Experiment

In practical, however, ICA approach has some limitations in forecasting time series. The most two significant drawbacks are component ambiguity , time series correlation approximation and mean difference problems. Below we describe those three major problems and propose our solutions for them.

3.1 Component Ambiguity Problem

Intrinsically ICA process generates independent components from the input time series, but it doesn't promise the extracted components ordering is the same as the original ones. The ICA component amplitude means any sign or a scalar multiplier in one of source s_i can be removed by dividing the corresponding column a_i of matrix **A** by some constant c_i. Equation(1) presents the basic ICA mixing model, but it can be presented in another form.

$$\mathbf{x} = \sum_i (\frac{1}{c_i}a_i)(s_i c_i) \tag{14}$$

The ICA component order ambiguity is the order of original source $\mathbf{s}$ is undecidable. Suppose some permutation matrix $\mathbf{P}$ exists, formula(1) can be written in this form.

$$\mathbf{x} = \mathbf{A}\mathbf{P}^{-1}\mathbf{P}\mathbf{s} \tag{15}$$

In formula(15), the elements of $\mathbf{Ps}$ are original source $\mathbf{s}$ within another order. The matrix $\mathbf{AP}^{-1}$ forms another mixing matrix. Below the figure illustrates the situation about the component ambiguity.

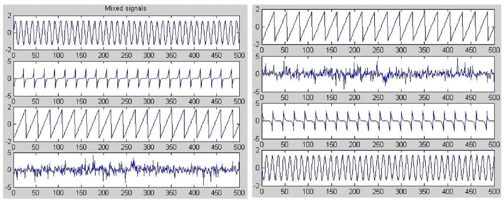

Figure 2: The component ambiguity effect.

Figure 2 is the toy data demonstration. Four artificial time series feed to ICA separation procedure. The ICA component amplitude and permutation ambiguity both present in the demonstration. Series in row three on left side graph, permutation ambiguity makes it appears in the row one on the right side graph after the ICA procedure. The magnitude ambiguity reverses those series values in opposed sign.

Those two ambiguity problems are harmful for ICA forecasting mechanism. Below we propose our solution to solve those two problems for them. ICA component ambiguity problem(14)(15) affects the sign and order of forecasted time series $\mathbf{T}^p(t+1)$. To solve the amplitude and sign ambiguity, the researchers use two procedures to overcome the component ambiguity problem : the absolute value of the forecasted time series $\mathbf{T}^p(t+1)$ and linear matching .

$$\mathbf{T}^p_a(t+1) = |\mathbf{T}^p(t+1)| \tag{16}$$

Equation(16) removes the sign ambiguity of the forecasted time series $\mathbf{T}^p(t+1)$. According to the ICA components order ambiguity problem, the order of forecasted time series $\mathbf{T}^p_a(t+1)$ is undecidable. To matching the approximated time series $\mathbf{T}^p_a(t+1)$, the mean-squared error(MSE) measurement among all forecasted time series $\mathbf{T}^p_a(t+1)$ was used. Assume each series length is L, and the optimum matched series denotes $\mathbf{T}^p_o(t+1)$. The optimum mapping equation proposes in equation(17). With n components, the computation time of the matching procedure costs O(n).

$$\mathbf{T}^p_o(t+1) = min\{\frac{1}{L}\sum_{k=1}^{L}(\mathbf{T}^p_a(k) - x(k))^2\} \tag{17}$$

3.2 Time Series Correlation Approximation Problem

The basic assumption of ICA approach for forecasting model is some underlying factors affect lots time series simultaneously. How can we choose lots time series affected by the same underlying factors? Our purposed method is the statistical correlation coefficient $\rho_{x,y}$. The $\rho_{x,y}$ describes each two component's relationship between time series x and y. As the correlation coefficient approximates positive one, the forecasted time series are more close to the original time series.

$$\mu_k = \frac{1}{L}\sum_{i=1}^{L}k_i \tag{18}$$

$$Cov(X,Y) = \frac{1}{L}\sum_{i=1}^{L}(x_i - \mu_x)(y_i - \mu_y) \tag{19}$$

$$\sigma_k = \sqrt{\frac{1}{L}\sum_{i=1}^{L}(k_i - \mu_k)^2} \tag{20}$$

$$\rho_{X,Y} = \frac{Cov(X,Y)}{\sigma_X \sigma_Y} \tag{21}$$

Equation(18) to (21) list how to retrieve the correlation coefficient between two time series. Below the empirical information reveals how the correlation coefficient affects the forecasting result.

Table 2. Statistical correlation coefficients compared with two application domains

Compared series	2nd	3rd	4th	5th	6th
Temperature	0.993	0.995	0.995	0.997	0.996
Stock	0.876	0.676	0.810	0.877	0.836

In table 2, monthly average temperature and stock time series are compared. Obviously, the correlation coefficients in the monthly temperature are closer to positive one than those coefficients in the stock data. Figure(3) illustrates the forecasting results.

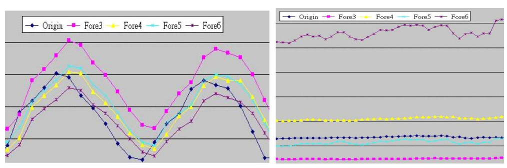

Figure 3: The two kinds forecasted results compare.

Figure(3) presents the forecasting difference between those two time series. The dark blue lines, in both sides,

ISBN:978-988-98671-2-6

are the base original time series. On the left side area, the correlation coefficients among those series approximate positive one, and those forecasted time series $T^P(t+1)$ are more closer to the original time series as the number of time series increases. On the right side, with the lower correlation values, the forecasted time series diverse from the original time series when the number of time series increases. Increasing the number of time series whose correlation coefficients are positive correlation can improve the forecasting accuracy.

3.3 Mean Difference Problem

The predicting behavior of our model can write in this form.

$$T^P(t+1) = A \cdot AR \cdot Wx(t) \tag{22}$$

Due to the AR is a linear model and matrix A is the inverse of matrix W, equation (22) could be written in another form.

$$T^P(t+1) = AR \cdot x(t) \tag{23}$$

AR(1) model expresses in the form: $T^P(t+1) = (1 - \varphi_1)\mu_x + \varphi_1 x(t)$. Since φ_1 is bounded between -1 and 1, the output value of $T^P(t+1)$ begins x(t) to $2\mu_x - x(t)$. The inference reveals that the mean of every time series affects the forecasted results of our model. Below the figure illustrates the mean difference problem and solved result.

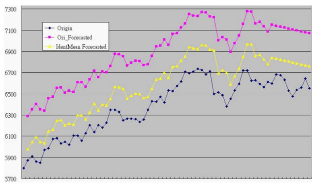

Figure 4: The mean difference problem and solved result.

In the figure(4), the dark blue line is the original time series and the pink one is the forecasted result with the mean difference effect. The diversity significantly exists between the forecasted series to original ones. After unifying the mean of every time series, we plot the forecasted series in yellow lines. Comparing with the pink line and the yellow one, you can easily observe that the yellow line is more close to the original time series than the pink one. The autoregressive mechanism causes the mean difference effect when the model was used in the forecasting usage.

3.4 Stock time series

In our first experiment, stock series is our forecasting target. Using the Nasdaq Industry Index in the United State and Taiwan Weighted Stock Index (TAIPEX) in Taiwan, we verify the performance of our model. For the vacation on those different areas, the common business days remain 97 days in those stock markets. The learning period begins from November 1, 2005 to February 21, 2006. Seventy-seven days are the learning period in both business period. The forecasted result remains 21 days from January 24, 2006 to March 3, 2006. Figure 5 is the original stock data. To simplify the long-term title, TSICA is our model's abbreviation in later article.

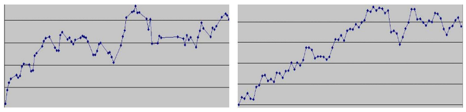

Figure 5: The original stock indices. Left is Nasdaq Indices, and the right is TAIPEX from the Noverber 1, 2005 to March 3, 2006

After PCA and whitening procedures (4)(6), the equal magnitude whitened component **z** released. Figure 6 shows the whitened components.

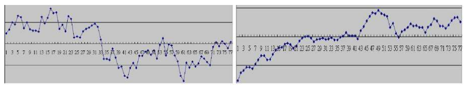

Figure 6: The whitened components **z** after the linear transformation $z = Vx$.

After the ICA preprocessing phase (7)(8), our model separates the underlying factors **y** of these two stock indices. Figure 7 illustrates separated components **y**.

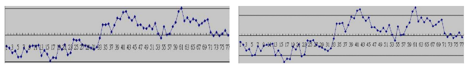

Figure 7: The separated components **y** after the separating matrix **w** retrieved.

In the figure 8, the black points are the original stock time series, the pink ones are time series forecasted by TSICA model and the yellow ones are contrast series generated by AR(1) model.

Table 3. Forecasting mean-squared error(MSE) between autoregressive(AR(1)) and TSICA

	5 days	10 days	15 days	20 days
AR(1)	1044.318	691.874	1346.786	9694.008
TSICA	1083.215	869.842	2319.763	13434.037

In the first view of the mean square error(MSE) compare table 3 above, AR(1) model seems to induce fewer forecasted error value than TSICA model. Observing the figure 8, however, the forecasted series is significant difference between the AR(1) and TSICA model. The fore-

ISBN:978-988-98671-2-6

casted time series of AR(1) smoothly extends the previous time sample, but the series from the TSICA forms a falling trend of the time series.

AR(1) model purely uses the autocovariance coefficients and the last time stamp value to forecast the future series. In different mechanism, TSICA explores the underlying factors of the time series. So the hidden information reveals the future trend of the time series. TSICA forecasted series reflects the inside intent.

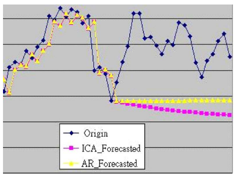

Figure 8: The original, independent component analysis forecasted and autoregressive forecasted time series.

3.5 Weather temperature time series

In the second experiment, the Taiwan central weather bureau released average monthly temperature data feeds in our model. The learning period of temperature information begins January 1998 to December 2004, totally 97 months. Figure 9 is the original monthly average temperature data of two cities.

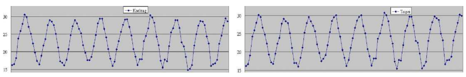

Figure 9: The original monthly average temperature of two cities in Taiwan.

Figure 10 illustrates the forecasted temperature result. We realize the forecasted time series approximates the original temperature data, and the mean square error (MSE) between the original series and the forecasted ones is 2.133 Celsius degree.

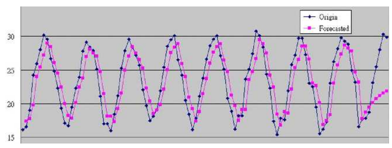

Figure 10: The forecasted result and the original monthly average temperature of two cities in Taiwan.

4 Conclusion

Component ambiguity, time series correlation approximation and mean difference problem are major intrinsic limitations for building a financial time series forecasting model by independent component analysis mechanism. In this paper, we purpose mean-squared error(MSE) measurement, statistical correlation coefficient measurement and mean unification to overcome those drawbacks for building the financial forecasting model.

In the future, the researchers will quantitate the correlation coefficient as a threshold value to determine the approximation degree of two time series. Without a specific threshold value, the approximation degree of each two time series has been relative to other ones. In the forecasting phases, AR model is not the only one choice. Eltoft [7] had purposed both a multi-layer perception (MLP) network or a finite impulse response (FIR) filter as one step predictor. Juan M. Górriz et al. [8] use radial basis function (RBF) artificial neural network (ANN) to forecast the financial time series. The accuracy among those forecasting techniques can be compared in the future papers.

References

[1] Aapo Hyvärinen et al., *INDEPENDENT COMPONENT ANALYSIS*, John Wiley & Sons, 2001.

[2] Maw-Wen Lin, *Time Series Analysis and Forecasting*, Hwa-Tai Books, 1992

[3] Mălăroiu S. K. Kiviluoto and E. Oja, "Time Series prediction with independent component analysis", *In Proc. of Advanced Investment Technology(AIT'99)*, Gold Cost, Australia, pp. 895-898, 1999

[4] B. Wu, *Time Series Analysis*, Hwa-Tai Books, 1995

[5] A.D. Back and A. S. Weigend, "A First Application of Independent Component Analysis to Extracting Structure from Stock Returns", *Int. J. of Neural Systems*, V8, N4, pp. 473-484, 1997

[6] P.Y.Mok et al., "AN ICA DESIGN OF INTRADAY STOCK PREDICTION MODELS WITH AUTOMATIC VARIABLE SELECTION", *Proceedings. of the IEEE International Joint Conference on Neural Networks*, V3, N84, Hong Kong, China, pp. 2135-2140, 2004

[7] Torbjorn Eltoft, "Data augmentation using a combination of independent component analysis and non-linear time-series prediction", *Proceedings of the 2002 International Joint Conference of Neural Networks*, Honolulu, USA, pp. 448 - 453, 2002

[8] J.M.Gorriz, C.G.Puntonet, M. Salmeron, E.W. Lang, "Time series prediction using ICA algorithms", *Proceedings of the Second IEEE International Workshop on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications*, pp. 226-230, 2003

ISBN:978-988-98671-2-6