An Ensemble De-noising Method for High Frequency Financial Data

Chaoyong Wang
School of Applied Science
Jilin Teacher's Institute of Engineering and Technology
Changchun, China
dynasty1188@126.com

Yanfeng Sun
College of Computer Science and Technology
Jilin University
Changchun, China
yfsun@jlu.edu.cn

Abstract—**High-frequency financial data are characterized by unbalanced, non-linear and low signal-noise ratio, which often represents a challenge on the study of financial market micro-structure. There has been little research on the de-noising method for high-frequency financial data, with the wavelet analysis as the current major method. Considering that the effect of wavelet analysis is restricted by the signal-noise ratio, we introduced phase space reconstruction and independent component analysis method for analyzing high-frequency financial data. The qualitative and quantitative analyses have shown that high-frequency financial data is chaotic in the time series and suitable to use the phase space reconstruction method. Furthermore, we propose the ensemble de-noising method for the high-frequency financial data. The numerical experiments results show that the de-noising effectiveness of our proposed methods is better than that of wavelet analysis. The improvement is about 2 times and more from the view of prediction precision based on the support vector machine. Our proposed ensemble de-noising method may also become a basis for general studies of financial market microstructure.**

Keywords- Ensemble method, wavelet analysis, phase space reconstruction, independent component analysis, high-frequency financial data

I. INTRODUCTION

High-frequency financial data is not only the kernel of the research on financial market microstructure, but also the prime focus of the industry and academic circles. Using the financial high-frequency data, we can analyze the investors' behaviors; discover the price forming mechanism in the process of trading, explore the relation between information and changing price, learn the whole trading process, evaluate whether the trading process is reasonable, judge whether the information is symmetric, distinguish whether there is market manipulation and so on. Consequently, these are beneficial to reduce the market information asymmetry, increase the strength of market competitiveness, and promote the formation of scientific and reasonable trading mechanism. And then, the profit of market participants can be protected, market risk can be prevented effectively by the market regulator.

At present, the study of high-frequency financial data mainly involves the statistic characteristic of high frequency data, the volatility, the interval between trades, price change, and econometric model etc.. These studies are all dependent on high quality high-frequency data. However, under the open-outcry trading system, there are many errors to be found in these trading records, which are from the trading time, price and volumes, even the missing value can also be found. In addition, there are many other disturbing factors, such as microstructure noise etc., so that the quality of high frequency financial data is not high. Therefore, de-noising work is very important before studying the high frequency financial data.

Fang Wang[1] has proposed threshold pre-average realized volatility, which combines the average method with threshold thinking, where pre-average method is used to reduce the effectiveness of market microstructure noise and threshold method is used to counteract the impact of price jumps on the volatility estimation. Xuguo Ye and Xueqiao Du[2] have proposed several kinds of the error estimation method of market microstructure noise under different assumptions and evaluated the performance of the proposed method through simulation experiments. Jie Zhao[3] has given auto-covariance estimation of market microstructure noise error and using the realized volatility to estimate market microstructure noise error under high frequency data, and obtained the corresponding conclusion through the Monte-Carlo simulation test. Edward W. Sun[4] et. al. have proposed the local linear scaling approximation algorithm based on the linear maximal overlap discrete wavelet transform to decompose the systematic pattern and noise. Qiujun Lan[5] et. al. have analyzed the disadvantages of traditional de-noising method for financial data, proposed a de-noising method based on wavelet analysis, verified the advantage of wavelet de-noising method according to the non-linear threshold theory suggested by Donoho. Cai Peng[6] et. al. have analyzed the characters of earthquake data and random noise signals, exploited the fast independent component analysis based on the minimum mutual information theory to realize the earthquake data de-noising.

High-frequency financial data are characterized by non-stationary, non-linear and low signal-noise ratio, which are usually different from regular frequency time series. In the aspect of solving the non-stationary and non-linear problem, Wavelet analysis performs better than other traditional data preprocessing method (such as all kinds of filtering algorithm), so that it has become a popular preprocessing method in many research field. However, the performance of wavelet-based de-noising closely depends on the signal-noise ratio, so that the satisfying results can't be acquired many times, especially facing the de-noising problem of high-frequency financial data. Therefore, by combining the ideas of wavelet analysis, phase space reconstruction and

independent component analysis, we propose the ensemble de-noising method based on high-frequency financial data in this paper, and the experiments results show that our proposed method performs better than wavelet-based de-noising method.

This paper is organized as follows. In Section 2, we briefly introduce the basic concepts and algorithmic idea of wavelet analysis, phase space reconstruction and independent component analysis. Then we propose the ensemble de-noising method based on high-frequency financial data in Section 3. We show the performance of our proposed method with experiments in Section 4. We summarize in Section 5.

II. RELATED ALGORITHMS AND ALGORITHMIC THINKING

Wavelet is meaning "small wave", this small means that a wave is localized in time domain, so its energy is finite. The basic idea of wavelet analysis is to decompose the original signal into a series of wavelet basis functions, which are obtained by translation and scale dilation of one mother wavelet function [7]. In these years, the multi-resolution wavelet analysis is the most widely used and effective in application field. Its' thinking is that function $f(x) \in L^2(R)$ will be decomposed into a series of subspace sequence with different resolutions, which can be described as a series of approximation functions. Each of which has one corresponding projection of function $f(x)$ at different resolutions. Through these projections, the feature of function $f(x)$ can be analyzed at different resolutions. As shown in Figure 1, the original signal S can be decomposed two parts, one of them is approximation part cA_j denoting useful information and another one is detailed part cD_j delegating noise, where j is the number of decomposition level. Through this kind of multi-resolutions decomposition, useful information and noise part usually have different performance, thus the aim of signal-noise separation can be reached.

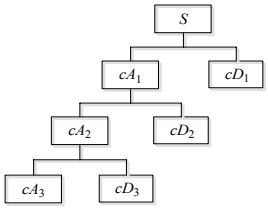

Figure 1. Wavelet multi-resolution decomposition

Phase space reconstruction is one of the best methods to cope with non-linear problem, which has been proposed by Takens based on chaotic theory. The basis idea of which is to reconstruct the chaotic attractors from high-dimensions space, so as to separate a mixture of independent oscillatory sources perfectly, i.e., through estimating the embedding dimension and delay time for a given time series, the original series can be extended to the high dimensional feature space using these parameters, so that the information hidden in the original time series can be exposed. Takens' theory implies that an appropriate time delay and a good embedding dimension play an important role in reconstruction R^n phase space, among which the trajectories may maintain the diffeomorphism with original dynamic system, i.e., in the context of topological equivalence, the dynamic system can be analyzed through phase space reconstruction from certain a time series, and whose chaotic feature can be restored and maintained. Generally, a time series $x = \{x_1, x_2, \cdots, x_N\}$, is most often used to reconstruct phase space, namely the state space:

$$X = \{X_i = (x_i, x_{i+\tau} \cdots, x_{i+(d-1)\tau}),\ i = 1,2,\cdots M = N-(d-1)\tau\} \quad (1)$$

where x_i is a point in state space, d denotes embedding dimension and τ denotes time delay[8].

In the process of phase space reconstruction, it is not only very important but also difficulty to choose embedding dimension and time delay. Only if an appropriate time delay τ and good embedding dimension d are selected, phase-space reconstructed can fully reveal the movement features of system. At present, the usual methods used to select a good delay time include autocorrelation function, multiple autocorrelation, mutual information, etc., while G-P algorithm and False Nearest Neighbor are usually used to choose the appropriate embedding dimension. In addition, there is a kind of method that can estimate the embedding dimension and delay time simultaneously, such as C-C method.

Independent Component Analysis (ICA) is a kind of blind sources separation method [10]. The aim of ICA is to find a linear decomposition of observed data into statistically independent components, namely to extract the basic source signals (also known as manifold) from mixed signals, and these source signals are independent. In these years, ICA has been widely applied to mobile phone communications, nature language processing, biomedicine, seismic signals and economic analysis etc., while ICA is usually to fulfill the assignments of blind sources separation, features extraction and information filter and so on. The procedure of independent components analysis is shown in Figure 2. Supposing that the subcomponents of signal sources $S(t)$ are all statistically independent from each other, which can be separated from the observed data $x(t)$ through the separation and mixture system B, so that the output signal $y(t)$ approximates the input signal $S(t)$.

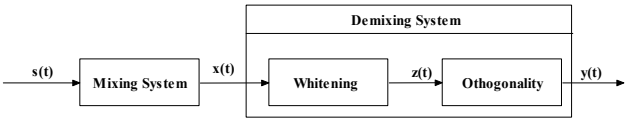

Figure. 2. The procedure of ICA

III. AN ENSEMBLE DE-NOISING METHOD

In view of the inherent nature of high-frequency financial time series, based on algorithmic thinking of wavelet analysis, phase-space reconstruction and independent

components analysis, we proposed an ensemble de-noising method, whose basic idea is as follows. Firstly, using wavelet analysis, one dimension high-frequency financial time series will be decomposed into approximation and noise part at level 3, the next step is to perform initial de-noising by adopting default threshold value, and to perform a multilevel one-dimensional wavelet reconstruction. Secondly, adopting phase-space reconstruction technique, one dimension time series obtained by the first stage is mapped into high dimension space, so as to found chaotic attractors. Thirdly, information manifold (valuable information) will be identified by independent component analysis and support vector machine. Lastly, through reconstructing the high dimension data proposed by the last step method, one dimension time series can be obtained. The concrete procedure and detailed description of our proposed method are as follows.

First stage, let $X = \{x_t \,|\, x_t = f(t) \in L^2(R)\}$ be a high frequency financial time series, $\varphi(t)$ a wavelet scale function and $\psi(t)$ wavelet function. According to the multi-resolutions analysis theory of $L^2(R)$ space, then

$$\sum_{k=-\infty}^{\infty} C_{j,k}\phi_{j,k}(t) = \sum_{k=-\infty}^{\infty} C_{j+1,k}\phi_{j+1,k}(t) + \sum_{k=-\infty}^{\infty} D_{j+1,k}\psi_{j+1,k}(t) \qquad (2)$$

holds. Based on the orthogonality of scale function and wavelet function, the following formulas (3) and (4) should hold.

$$h(k-2m) = <\varphi_{j+1,m}, \varphi_{j,k}> \qquad (3)$$

$$g(k-2m) = <\psi_{j+1,m}, \varphi_{j,k}> \qquad (4)$$

According to the formulas (2), (3) and (4), one can obtain the following result.

$$c_{j,k} = \sum_m h(k-2m)c_{j-1,m}\ ,\ d_{j,k} = \sum_m g(k-2m)c_{j-1,m} \qquad (5)$$

Rewriting (5) into matrix form, one obtains

$$C_j = HC_{j-1}\ ,\ D_j = GC_{j-1} \qquad (6)$$

Then, the reconstruction algorithm can be described as follows.

$$C_j = H^*C_{j+1} + G^*D_{j+1}\ ,\ j = J-1, J-2, \cdots, 0 \qquad (7)$$

where H^* is the dual operator of H, G^* is the dual operator of G. Through reconstructing $d_1, d_2, \cdots, d_J$ and c_J into $D_1, D_2, \cdots, D_J$ and C_J respectively, one obtains a new time series W whose number of samples is same as that of the original series X.

$$W = D_1 + D_2 + \cdots + D_J + C_J\ , \qquad (8)$$

Second stage, given time series $W = \{w_i\}_{i=1}^N$ obtained from the first stage, let delay time be τ and embedding dimension be m. The correlation integral is defined as the following function.

$$C(m, N, r, \tau) = \frac{2}{M(M-1)} \sum_{1 \le i \le j \le M} \theta(r - d_{ij}) \qquad (9)$$

where $r > 0$, $d_{ij} = \left\| w^{(i)} - w^{(j)} \right\|_\infty$ and $\theta(z) = \begin{cases} 0 & z < 0 \\ 1 & z \ge 0 \end{cases}$

The measurement statistic is defined as follows.

$$S_1(m, N, r, \tau) = C(m, N, r, \tau) - C^m(1, N, r, \tau) \qquad (10)$$

In fact, in calculating the above Equation (10), we must subdivide the time series $W = \{w_i\}_{i=1}^N$ into t disjoint time series as follows.

$$\begin{cases} w^{(1)} = \{w_1, w_{t+1}, \cdots, w_{N-t+1}\} \\ w^{(2)} = \{w_2, w_{t+2}, \cdots, w_{N-t+2}\} \\ \quad \vdots \\ w^{(t)} = \{w_t, w_{2t}, \cdots, w_N\} \end{cases} \qquad (11)$$

Where the length of sub-time series is $N_s = N/t$. By adopting the blocking averaging strategy for Equation (10), $S(m, N, r, t)$ of each sub-time series is calculated as follows.

$$S_2(m, N, r, t) = \frac{1}{t}\sum_{s=1}^{t} C_s(m, \frac{N}{t}, r, t) - C_s^m(1, \frac{N}{t}, r, t) \qquad (12)$$

If $N \to \infty$, then

$$S_2(m, r, t) = \frac{1}{t}\sum_{s=1}^{t} C_s(m, r, t) - C_s^m(1, r, t)\ , (\mathrm{m} = 2, 3, \cdots)\ .$$

One defines the following increment

$$\Delta S_2(m, t) = \max\{S_2(m, r_j, t)\} - \min\{S_2(m, r_j, t)\} \qquad (13)$$

Corresponding to $S_2(m, r, t) \sim t$, the above equation (13) can be used to measure the maximum bias of r. Therefore, one look for the first zero crossing point of $\overline{S}_2(t)$ or the first minimum point of $\Delta\overline{S}_2(t)$ as the optimal time delay τ. Based on $\overline{S}_2(t)$ and $\Delta\overline{S}_2(t)$, the global minimum point of $S_{2or}(t)$ can be viewed as the embedding dimension $\tau_w = (m-1)\tau_d$. The corresponding formulas are as follows.

$$\begin{cases} \overline{S}_2(t) = \dfrac{1}{16}\sum_{m=2}^{5}\sum_{j=1}^{4} S_2(m, r_j, t) \\ \Delta\overline{S}_2(t) = \dfrac{1}{4}\sum_{m=2}^{5} \Delta S_2(m,t) \\ S_{2cor}(t) = \Delta\overline{S}_2(t) + \left|\overline{S}_2(t)\right| \end{cases} \tag{14}$$

After found the time delay τ and embedding dimension m, one can obtain the reconstructed phase space $P = \{P_i\}_{i=1}^{d}$ by the formula (11).

Third stage, what one would like to do is to find the original signals from the mixtures $P(t)$. There exists a linear transformation matrix H_0, such that $P_i(t)$ can be projected into the new subspace as the white vector, i.e.

$$Z_i(t) = H_0 P_i(t) \tag{15}$$

Where H_0 is a white matrix, $Z_i(t)$ is white vector. Through calculating the samples vector, one can obtain a transformation $H_0 = \Lambda^{-1/2} U^T$ by the principle component analysis method, where U is the eigenvalues of the covariance matrix C_P, Λ is the eigenvectors of the covariance matrix C_P. Using orthogonality transformation, one can draw a conclusion that the equation $U^T U = U U^T = I$ holds. Therefore, one can obtain the following equation (16).

$$\begin{aligned} &E\{ZZ^T\} = E\{\Lambda^{-1/2} U^T X X^T U \Lambda^{-1/2}\} \\ &= \Lambda^{-1/2} U^T E\{XX^T\} U \Lambda^{-1/2} = \Lambda^{-1/2} \Lambda \Lambda^{-1/2} = I \end{aligned} \tag{16}$$

If Equation $P_i(t) = AS_i(t)$ can be substituted into Equation $Z_i(t) = H_0 P_i(t)$ and let $H_0 A = \widetilde{A}$, then Equation (17) holds.

$$Z_i(t) = H_0 A S_i(t) = \widetilde{A} S_i(t) \tag{17}$$

Where $Z(t)$ is new output signal, which includes the source signal separated from mixture signal, A is mixture matrix.

Fourth stage, the main aim is to find valuable features from the last obtained signal $Z(t)$. Adopting the feature selection algorithm based on SVM, the noise independent component of signal $Z_i(t) \in R^d$ will be identified by the prediction accuracy, and then a new matrix $S_i(t) \in R^s$ can be obtained, where $s \le d$. Denoting $\{I_s^{(S)}\}$, $\{I_i^{(Z)}\}$ and $\{I_j^{(N)}\}$ as the feature set of signal $S(t)$, $Z(t)$ and noise respectively, one can draw the following Equation (18), i.e.,

$$\{I^{(S)}\} = \{I^{(Z)}\} - \{I^{(N)}\}, \tag{18}$$

Then the Equation (19) holds.

$$S(t) = Z_{I(S)}(t) \tag{19}$$

Fifth stage, supposing $\widetilde{A}$ is a mixture matrix, whose columns have been deleted corresponding to noise components, phase space $\widetilde{P}$ will be reconstructed, namely

$$\widetilde{P} = \widetilde{A} S \tag{20}$$

The next step is the inverse procedure of phase space reconstruction, thus new one dimension time series $\hat{x}$ can be reconstructed from phase space $\widetilde{P}$, i.e.,

$$\hat{X}(k) = \widetilde{P}(i, k - (i-1)\tau_t) \tag{21}$$

Thus, the de-noised one dimension time series $\hat{X}$ can be obtained based on our proposed ensemble de-noising method.

IV. Performance Evaluation

In order to verify the effectiveness of our proposed ensemble de-noising method, Data was collected from stocks in the Chinese shanghai composite index from Jan 4 ,2012 to Jan 10, 2013, which is the 5-minute closing price series, the number of samples is 11904.

A. Phase Space Reconstruction Experiments

In fact, not all time series stem from chaotic discrete dynamical system. Thus, before using the dynamical system theory to analyze the time series, one should distinguish between the chaotic and non-chaotic time series firstly. To solve this problem, one adopted two methods which are the qualitative analysis of power spectrum and the quantitative analysis of maximum Lyapunov exponent. The principle of discriminating the chaotic state using the maximum Lyapunov exponent is that, if Lyapunov exponent is positive, and the trajectory can become infinitely distant from the equilibrium state upon small variations, then the chaotic behaviors can be detected. From Table 2, it is clear that the maximum Lyapunov exponent is all positive under the condition of the different embedding dimension and delay time.

TABLE 2. LYAPUNOV EXPONENT STATISTICS

Embedding Dimension	2	3	4	5	6	7
Delay Time	3	3	4	4	5	5
Lyapunov Exponent	0.6614	0.8542	0.5908	0.5392	0.5840	0.6061

The principle of discriminating the chaotic state using the power spectrum is that, if the movement of variable is chaotic, then power spectrum is a consecutive curve rather than horizontal line. Figure 4(b) is the local enlarged graph of Figure 4(a). In view of Figure 4(b), one can observe that it's really a consecutive curve. Thereby, one can conclude

that the 5-minute high frequency data of Chinese shanghai composite index is indeed chaotic time series. Meanwhile, this also indicates that high frequency financial data is suitable for phase space reconstruction theory.

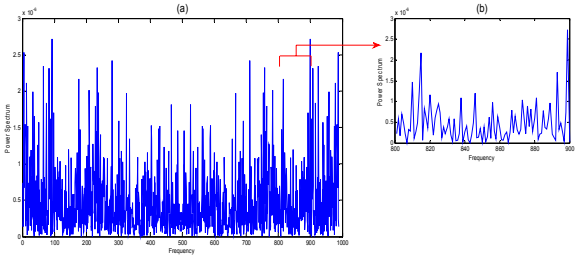

Figure 4. Power spectrum of 5-minute high frequency

Base on the thinking of removing the continuity of high frequency financial time series and transforming one dimension into multi-dimension time series problem which is convenient to analyze, phase space reconstruction method is used to splits original time series into a set of many sub-time series with the same length by the embedding dimension and time delay, which can be mapped into the multi-dimension space. Thereby, the embedding dimension and time delay are very important factors for reconstruction, and which can be usually obtained through coping with the original time series and doing some estimation. In this paper, one adopted the C-C method which is based on the idea of the embedding dimension and time delay estimated simultaneously. Data comes from the composite index in China Shanghai Stocks Exchange from Jan 4 ,2012 to Jan 18, 2012, which is the 5-minute closing price series, the number of samples is 500.

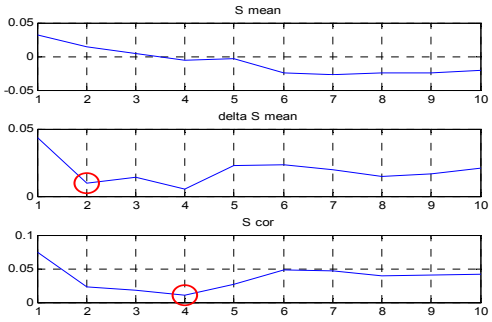

Figure 5. Estimation of the embedding dimension and time delay

According to the C-C algorithm theory, the first local minimum of statistical estimator $\Delta\overline{S}_2(t)$ is the optimal time delay. From the view of the middle one of Figure 5, the first local minimum is obtained at point 2, i.e., the optimal time delay is 2. The global minimum of statistical estimator $S_{2cor}(t)$ is acquired at point 4. According to this method, the optimal embedding dimension and time delay of time series with different quantitative samples have been achieved, and the corresponding results are as the following Table 3. This also illustrates that, when the number of samples used to research is different, the embedding dimension and time delay for reconstruction are usually both different. However, from the view of statistical result, one can observe that the embedding dimension and time delay are usually not too large. Thus, it can ensure that not only the continuity of high frequency financial time series can be broken appropriately, but also the valuable information can be separated effectively. Meanwhile, this also successfully avoids "dimension disaster" in the separation procedure.

TABLE 3. COMPARISON OF HE EMBEDDING DIMENSION AND TIME DELAY

Number of Samples	500	1000	1500	2000	2500	3000
Time Delay	2	4	3	4	3	3
Embedding Dimension	4	4	7	7	6	3

B. *Ensemble De-noising Experiments*

Because wavelet de-noising method has been successfully applied to many fields, it has been regarded as the effective de-noising method. In this paper, in order to testify the effect of our proposed ensemble de-noising method, the result will be compared with wavelet de-noising method. As for the wavelet de-noising experiment, the above mentioned data will still be used for this experiment; db1 is selected as the wavelet basic function; default threshold value will be used to de-noise. As for ensemble de-noising method based on high frequency financial data, experiment will be conducted according to the procedure of our proposed method in Part 3. The below Figure 6 shows the distribution curves of original and de-noised data by different method and.

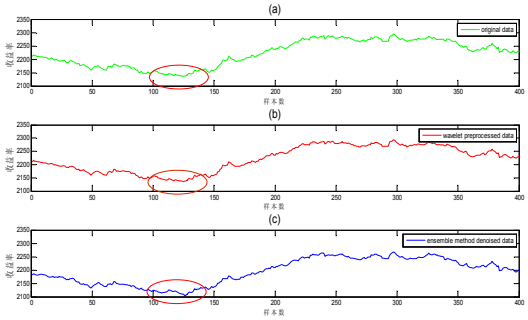

Figure 6. Comparison of de-noising effect

In Figure 6, (a) represents the distribution curves of original data. (b) represents the distribution curves of data de-noised by wavelet. (c) represents the distribution curves of data de-noised by ensemble de-noising method. In order to display the result of data de-noising effectively, one can only scratch the former 400 samples. From the view of Figure 6 (a) and (b), it is clear that the distribution curve of data de-noised by wavelet is not obviously different from that of the original data. This also indicates that the effectiveness of separating noise from the original data is not obvious to the wavelet de-noising method. However, Figure 6 (a) and (c) are obviously different, such as red ellipse marker. This illustrates that when signal-noise ratio

of high frequency financial data is relatively lower, wavelet de-noising method will be badly affected, but ensemble de-noising method can achieve the better result than wavelet de-noising method.

In addition, in order to further verify the effectiveness of ensemble de-noising method based on high frequency financial data and illustrate the effect of noise on research of financial market microstructure, the single and multi-step prediction have been made about the 5-minute high frequency data of China shanghai composite index based on the standard support vector machine (SVM). Root mean square error (RMSE), mean average error (MAE) and mean average percentage error (MAPE) are adopted to evaluate the effectiveness of wavelet and ensemble de-noising method. Because SVM is supervised learning method, 300 samples are selected as training set and 200 samples are used as test set. For rationality, SVM parameters are set as the same value. Penalty parameter is set as 100. RBF (Radial basis function) is selected as kernel function. The parameter of RBF kernel function is chosen as 0.01. Experimental results are shown in the following Table 4.

TABLE 4. COMPARISON OF DE-NOISING EFFECT BASED ON SVM PREDICTION

Prediction Step	RMSE			MAE			MAPE		
	O_data	W_data	E_data	O_data	W_data	E_data	O_data	W_data	E_data
1	0.0052	0.0050	0.0048	8.4732	7.9525	7.1211	0.0038	0.0035	0.0032
2	0.0055	0.0052	0.0050	9.4524	8.6900	7.9234	0.0042	0.0039	0.0035
3	0.0058	0.0056	0.0051	10.1170	9.5414	8.5231	0.0045	0.0042	0.0038
4	0.0062	0.0058	0.0056	10.6240	10.0190	9.6026	0.0047	0.0044	0.0043
5	0.0062	0.0059	0.0054	10.5440	10.0260	9.4600	0.0047	0.0044	0.0042
6	0.0064	0.0063	0.0056	10.9530	10.5330	9.8023	0.0049	0.0047	0.0044
7	0.0064	0.0064	0.0061	11.0100	10.7650	10.6240	0.0049	0.0048	0.0047
8	0.0067	0.0067	0.0060	11.3990	11.1120	10.6400	0.0051	0.0049	0.0047
9	0.0069	0.0067	0.0058	11.9370	11.0600	10.2820	0.0053	0.0049	0.0046
10	0.0072	0.0068	0.0062	12.3320	11.2890	10.6610	0.0055	0.0050	0.0048
12	0.0075	0.0073	0.0075	12.9080	12.3620	12.2920	0.0057	0.0055	0.0055
24	0.0125	0.0123	0.0124	23.5110	23.0640	22.7000	0.0104	0.0103	0.0102
Average	0.0069	0.0067	0.0063	11.9384	11.3678	10.8026	0.0053	0.0050	0.0048
Improvement		0.0002	0.0006		0.5706	1.13576		0.0003	0.0005
Contrast	**3.0**			**1.9906**			**1.8125**		

In Table 4, O_data denotes the original data, W_data denotes the de-noised data by wavelet de-noising method. E_data denotes the de-noised data by ensemble de-noising method. From the view of Table 4, it is clearly shown that, relative to prediction accuracy of original data, the prediction accuracy improvement of data obtained by ensemble de-noising method is about 2 times and more than wavelet de-noising method. Thus, this also indicates that ensemble de-noising method can improve the prediction accuracy and contribute to the financial market microstructure research.

V. CONCLUSION

In this paper, based on the algorithmic thinking of wavelet analysis, phase space reconstruction and independent component analysis, an ensemble de-noising method has been proposed in this paper. The 5-minute high frequency data of Shanghai composite index From Chinese shanghai stock exchange is used in numerical experiments. The experimental results show that the prediction of SVM based on the de-noised data by an ensemble de-noising method has got more accurate than wavelet de-noising method. From the numerical results, one can observe that the relative improvement based on an ensemble de-noising method is about 2 times and more than wavelet de-noising method. Namely, it also shows that an ensemble de-noising method based on high frequency financial data is capable of improving the prediction accuracy. Therefore, one can conclude that the ensemble de-noising method is a more effective method for de-noising, so that it can provide the higher quality data to further research the microstructure of financial market. Meanwhile, the ensemble de-noising method provides selection for the de-noising problem of other field. How can one further improve the de-noising capacity of our proposed ensemble de-noising method for the high frequency data with different frequency, which is the focus of the future work.

ACKNOWLEDGMENT

The authors are grateful to the support of the Natural Science Foundation of Jilin Province (Grant No. 201215168), Science and Technology research project of Jilin Provincial Department of. Education during the 12th Five-Year plan (Grant No. 2011299).

REFERENCES

[1] Fang Wang. Study on volatility of financial high frequency data based on market microstructure noise and jump[D]. Southwestern university of finance and economics, 2011.

[2] Xuguo Ye and Xueqiao Du. Estimation of market microstructure noise error using high frequency financial data[J]. College Mathematics. 28.5(2012): 62-69.

[3] Jie Zhao. Market microstructure noise error under high frequency finance data[J]. Journal of Hefei University of Technology. 31.3(2008): 380-383.

[4] Sun, Edward W., and Thomas Meinl. A new wavelet-based denoising algorithm for high-frequency financial data mining[J]. European Journal of Operational Research 217.3 (2012): 589-599.

[5] Joo, Jong-Hoon, and Peihua Qiu. Jump detection in a regression curve and its derivative[J]. Technometrics 51.3 (2009): 289-305.

[6] Cai Peng, Jianku Sun, Junhua Chen, Ling xia and Dongshan Huang. Noise elimination with independent component analysis[J]. Progress in Exploration Geophysics. 30.1(2007):30-32.

[7] Nason, Guy P. and Rainer Von Sachs. Wavelets in time-series analysis. Philosophical Transactions of the Royal Society of London[J]. Series A: Mathematical, Physical and Engineering Sciences 357.1760 (1999): 2511-2526.

[8] Tang, Longkun, and Jianli Liang. CC method to phase space reconstruction based on multivariate time series[J]. Intelligent Control and Information Processing (ICICIP), 2011 2nd International Conference on. Vol. 1. IEEE, 2011.

[9] Shaoqing Yang and Chuanying Jia. Two practical methods of phase space reconstruction methods[J]. Acta Physica Sinica. 51.11(2002): 2452-2457.

[10] Kim H. S., Eykhoh R. and Salas J. D.. Nonlinear dynamics delay times and embedding windows [J]. Physica D. 127 (1999): 48~60

[11] [11] Min Zhang, Zhu Mu, and Ma Wenjie. Implementation of FastICA on DSP for Blind Source Separation[J]. Procedia Engineering. 29 (2012): 4228-4233..