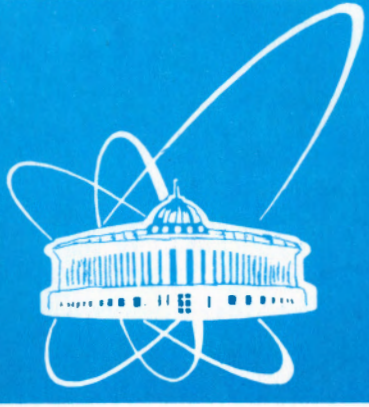

СООБЩЕНИЯ
ОБЪЕДИНЕННОГО
ИНСТИТУТА
ЯДЕРНЫХ
ИССЛЕДОВАНИЙ

Дубна

E5-99-221

Cs.Török[1], H.P.Bernhard[2]

WAVELET SHRINKAGE AND MUTUAL INFORMATION

[1]Technical University, Kosice, Slovakia
[2]Technical University, Wien, Austria

1999

Торок Ч., Бернхард Х.-П. E5-99-221
Вевлет-сжатие и взаимная информация

В работе рассматривается ключевой вопрос прогнозирования, сколько информации о будущих значениях процесса можно получить из прошлых измерений. Введен новый критерий для наилучшей оценки с помощью вевлет-сжатия. Целью работы было нахождение новых компонент для предсказания процессов, создание инструмента для описания отфильтрованных с помощью вевлет-сжатия сигналов и применение теории взаимной информации для вевлетов.

Работа выполнена в Лаборатории вычислительной техники и автоматизации ОИЯИ.

Сообщение Объединенного института ядерных исследований. Дубна, 1999

Török Cs., Bernhard H.P. E5-99-221
Wavelet Shrinkage and Mutual Information

The paper deals with the key question for the prediction gain analysis, how much information about the future values of a process can be obtained from the past. A new criterion is introduced for the best wavelet shrinkage estimator. The paper was motivated by three goals: find new valuable components for predictors, have a tool to characterize the de-noised signals received by wavelet shrinkage, apply the mutual information function to wavelets.

The investigation has been performed at the Laboratory of Computing Techniques and Automation, JINR.

Communication of the Joint Institute for Nuclear Research. Dubna, 1999

1. Introduction

One of the striking problems in the time series analysis and signal processing is the question of prediction and its quality. If there is no a priori knowledge on the optimal predictor and the prediction gain is based on the error of the predicted signal, the achieved prediction gain will depend strongly on the prediction model chosen. There is another approach to assess the prediction gain, which is independent of a special predictor implementation. The paper [1] assesses the achievable maximum of the prediction gain using an information theoretic quantity, known as the mutual information. It proves that for stationary processes the upper bound of the achievable maximum of the prediction gain depends linearly on the mutual information function.

In the eighties the main prediction tool was the Box-Jenkins methodology. Then one began to use the neural networks and afterwards came the wavelets. The wavelets are used for prediction very often in combination with neural networks. This work is on wavelets and mutual information function, it shows how one can assess the information gain of the de-noised signals and so select new components with high degree of information for predictor.

The wavelet shrinkage, the theory and methodology for nonparametric regression and smoothing [4], [5], [7] refers to signal (function, curve) reconstruction from noisy data obtained by wavelet transformation, followed by shrinking the empirical wavelet coefficients towards zero, finished by inverse wavelet transformation. Shrinking noisy wavelet coefficients via thresholding offers very attractive alternatives to existing noise reduction methods, such as splines, linear filters, or kernel smoothers. Unlike these traditional methods of recovering signals from noisy data, which either suppress noise, but erase certain features or leave features sharp, but does not really suppress the noise, wavelets are able to achieve noise suppression by removing the noise from signal while preserving the features (spikes, discontinuities). The questions connected with the degree of smoothing by wavelet shrinkage are not completely answered by now. We offer to apply to this problem the mutual information function.

The sections 2, 3 give the basic definitions and concepts of wavelet analysis and wavelet shrinkage. Section 4 is devoted to the short description of the mutual information function. Section 5 contains results of the application of the mutual information function to wavelet de-noising.

2. Wavelet decomposition

The main result of the one dimensional *wavelet analysis* is the (*wavelet*) *decomposition* of the signal y

$$y(t) \approx D_1 + D_2 + \cdots + D_J + S_J \qquad (1)$$

at (multi-resolution) *level J* on the base of the *wavelet transformation*

$$c_{n\times 1} = W_{n\times n} y_{n\times 1}, \qquad (2)$$

where

$$S_j(t) = \sum_k s_{jk}\phi_j(t), \quad D_j(t) = \sum_k d_{jk}\psi_j(t), j = \overline{1,J},$$

the (wavelet) *basis* functions $\phi_{jk}(t)$ and $\psi_{jk}(t)$ are the *scaled* and *translated* versions of the *father* and *mother wavelets* $\phi(t)$ and $\psi(t)$ respectively

$$\varphi_{jk}(t) = \frac{1}{\sqrt{2^j}}\varphi(\frac{t}{2^j} - k), \quad \psi_{jk}(t) = \frac{1}{\sqrt{2^j}}\psi(\frac{t}{2^j} - k), j = \overline{1,J},$$

and the coefficients $c = [\, s_J, d_J, d_{J-1}, \ldots, d_1]$ are given approximately by integrals (if $n=2^{J+1}$, then in d_{jk} the shift index k changes from *one to* $n/2^j$)

$$s_{Jk} \approx \int y(t)\varphi_{Jk}(t)\,dt, \quad d_{jk} \approx \int y(t)\psi_{jk}(t)\,dt, j = \overline{1,J}.$$

The wavelet decomposition (1) is achieved by two steps. In the first step one computes the *wavelet coefficients* c by the discrete wavelet transformation (DWT). In the second one the *detail* and *smooth approximations* $D_1, D_2, \ldots, D_J$ and S_J are evaluated by the inverse discrete wavelet transformation (IDWT). The DWT (2) and IDWT are realized not by matrix multiplication, but by the fast *forward* and *backward pyramidal algorithms*, which use low-pass (for s) and high-pass (for d) filter convolutions along with down-sampling and up-sampling, respectively. The scheme of the forward pyramidal algorithm is shown in figure 1.

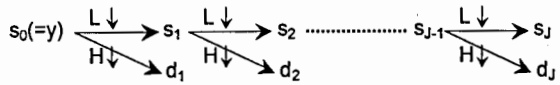

Fig. 1 Forward pyramidal algorithm

The result of the first step of wavelet analysis, the wavelet coefficients, may be visualized by stem plots, see figure 2, or as a time–scaled plot (see [3]). The coefficients $d_1, d_2, \ldots, d_6$ in figure 2 are plotted on the same vertical scale. As we see, the DWT can compact the energy of a signal into a relatively small number of wavelet coefficients.

For the approximations $D_1, D_2, \ldots, D_6$ and S_6 continuous plots are used, see figure 3. Mention must be made that in according with the pyramidal algorithm the smooth approximations $S_1, S_2, \ldots, S_5$ may be viewed too, where $S_{j-1} = D_j + S_j$ $(=D_j+\ldots D_J+S_J)$. From the analysis of the wavelet coefficients and approximations by different wavelet families (or wavelet packet families) one can conclude which wavelet and multi-resolution level give the most appropriate decomposition for a given signal. From figure 3 we can conclude, that for the given wavelet the multi-resolution level 4 is sufficient, for the detail approximations begin to behave from level five as a smooth one with negligible amount of noise.

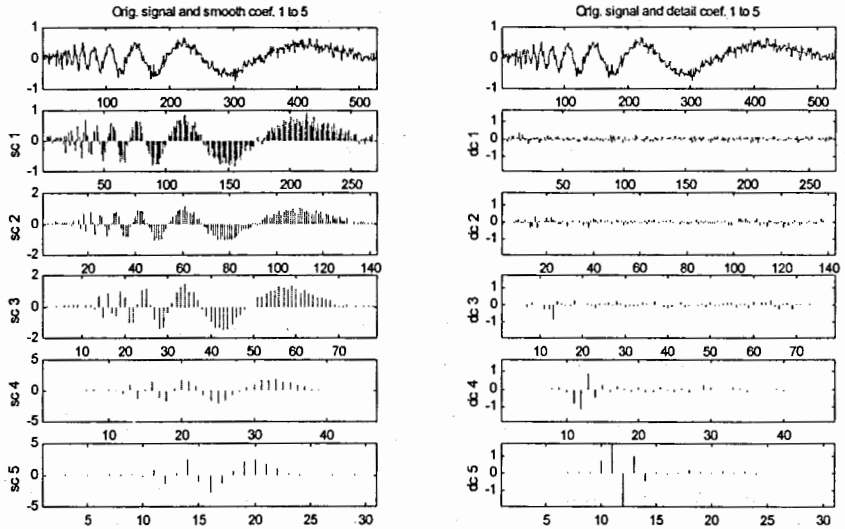

Fig. 2 Smooth and detail coefficients

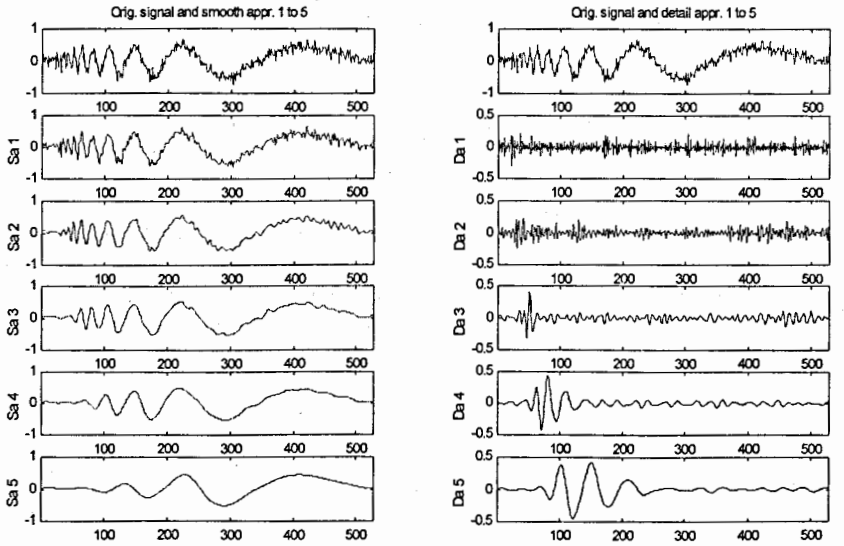

Fig. 3 Smooth and detail approximations

3. Wavelet de-noising

For de-noising orthogonal wavelets are used, so from the orthonormal wavelet transformation

$$c = Wy = W(f + \varepsilon) = Wf + W\varepsilon,$$

it follows that the white noise effects all wavelet coefficients.

It was D. Donoho's idea ([4], [5]) to begin the de-noising process not with the detail approximations D_i, but with the detail coefficients d_i and then compute the wavelet approximations by IDWT on the base of the new, de-noised coefficients. Wavelet de-noising is performed by shrinking the wavelet coefficients towards zero. To carry out the wavelet de-noising one has to chose the *shrinking rule* (from soft, hard ... shrinkage functions), the values of *threshold* λ and *scale* of noise σ (they are needed for the shrinking function).

It is shown (see [4], [7]), that for appropriate choices of λ_i the wavelet shrinkage gives nearly the best possible linear or non-linear estimate of $f(t)$ for a board class of functions from the Besov space B in the sense

$$\inf_{\hat{f}} \sup_{f \in B} \| \hat{f}_w - f \|_2^2,$$

i.e. we want the worst error to be as small as possible when σ tends to zero.

Figures 4, 6 show several reconstructed de-noised signals computed by *symlet* wavelets at different levels for noisy (signal to noise ratio SNR=1) *sin x* and generated process with three periods (see fig.5). Wavelet de-noising can be tuned using different wavelets, levels and shrinking parameters. So far there is no clear-cut rule to say which one of the de-noised signals is the appropriate. We will characterize them by the mutual information function.

4. Mutual information function

Mutual information is a fundamental concept of information theory. It is defined as the relative entropy between the joint probability density and the product probability density of random vectors $\vec{x}, \vec{y}$ by ([1], [6])

$$I(\vec{x}, \vec{y}) = \int \dots \int p(\vec{x}, \vec{y}) \ln \frac{p(\vec{x}, \vec{y})}{p(\vec{x}) p(\vec{y})} d\vec{x}\, d\vec{y},$$

or for discrete random vectors with probability mass functions $p(.)$ by

$$I^{\Delta}(\vec{x}, \vec{y}) = \sum_{\vec{x}} \sum_{\vec{y}} p(\vec{x}, \vec{y}) \ln \frac{p(\vec{x}, \vec{y})}{p(\vec{x}) p(\vec{y})}.$$

The mutual information measures the information that one random vector contains about another one. The basic properties of the mutual information are the following. The mutual information is ([1]):

- symmetric $I(\vec{x},\vec{y}) = I(\vec{y},\vec{x})$;
- nonnegative $I(\vec{x},\vec{y}) \geq 0$;
- $I(\vec{x},\vec{y}) = 0$ if and only if the vectors $\vec{x}$ and $\vec{y}$ are statistically independent;
- $\lim_{\Delta\to 0} I^{\Delta}(\vec{x},\vec{y}) = I(\vec{x},\vec{y})$ if the continuous probability density functions are Rieman-integrable.

The *auto-mutual information function* was introduced for (strictly) stationary stochastic processes in [6] and generalized for any stochastic process in [1]. The *mutual information function* (**MIF**) is defined as

$$M_{\vec{x},\vec{y}}(t,L) = I(\vec{x}(t),\vec{y}(t+L)),$$

where $\vec{x}(t)$ and $\vec{y}(t)$ are (in general) n-dimensional processes, e.g.

$$\vec{y}(t) = \{y_1(t), y_2(t), \ldots, y_n(t)\}.$$

The MIF measures the information contained in one signal about the other for every time lag L. The investigation of signals by MIF is similar to the correlation analysis. The main difference between them is that the MIF reflects both *linear* and *nonlinear* system properties.

The main properties of the mutual information function and the auto-mutual information function $M_{\vec{y},\vec{y}}(L) = I(\vec{y}(0),\vec{y}(L))$ for stationary processes are:

- $M_{\vec{x},\vec{y}}(t,L) \geq 0$;
- $M_{\vec{y},\vec{y}}(L) = M_{\vec{y},\vec{y}}(-L)$;
- for the *prediction gain*

$$G_y(L) = 10\log\frac{E(y^2(t))}{E(\varepsilon^2(L))}$$

it holds

$$G_y(L) \leq 6.02(M_{y,\vec{y}}(L) + \Delta H),$$

where ΔH is the entropy difference between the entropy of a Gaussian random variable with variance equal to $E(y^2(t))$ and the entropy of $y(t)$.

The last property implies that the information gain for stationary processes independently of the generating system (linear or nonlinear) underlying $y(t)$ and the predictor type *can not be greater* than the linear combination of the MIF and the entropy difference.

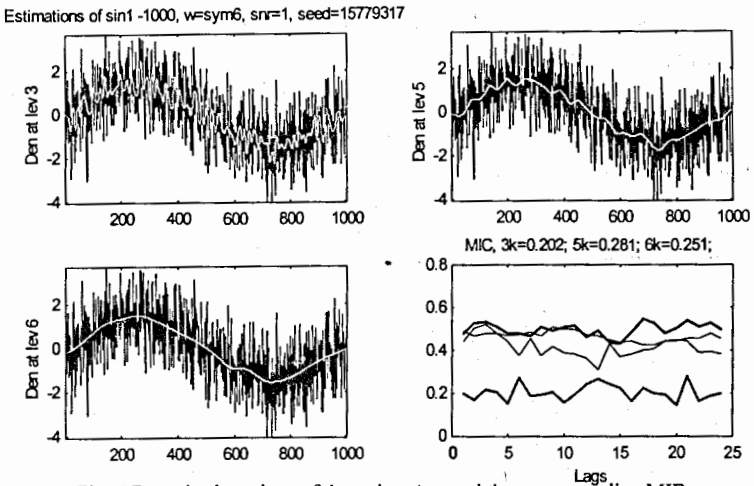

Fig. 4 De-noised versions of the noisy *sin x* and the corresponding MIFs

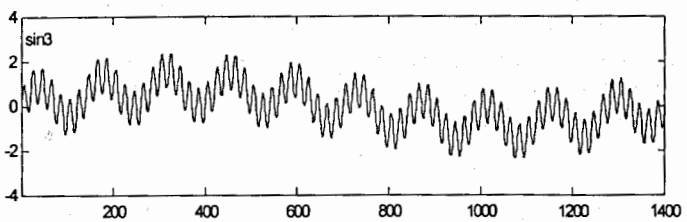

Fig.5 Generated process with three periods

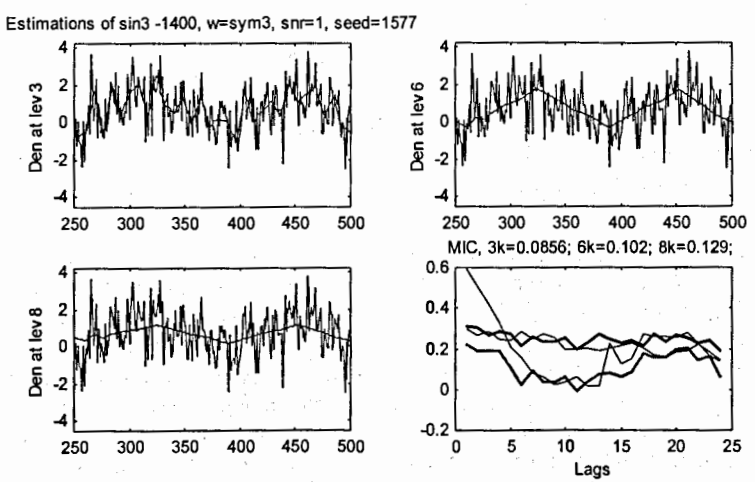

Fig. 6 De-noised versions of the noisy process from fig 5 and the corresponding MIFs

In the next section we will be interested in $M_{\hat{y},y(L)}$, where y, $\hat{y} \equiv \hat{y}_{w,j}$ stand for the one dimensional original signal and the de-noised one (its estimation by wavelet w at level j), respectively. $M_{\hat{y},y(L)}$ represents the amount of information that one can know about the future values of signal y that are separated by the time lag L from $\hat{y}$. On the computation of $M_{\hat{y},y(L)}$ see [2], [6].

5. Wavelet de-noising and the MIF

In this section we apply the signal analysis method based on the MIF to the characterization of the de-noised signals received by wavelet shrinkage.

To perform a wavelet shrinkage one has to choose first of all a wavelet, a multi-resolution level and set the shrinking parameters. Different choices give different estimators. It is hard to conclude which de-noised version is better. Figure 4 shows that de-noising at a higher level results in a smoother estimation, but from the smoothest one some important information may be swept out. The situation in fig.6 is even more intricate. The de-noised signals copy two different periods. Which period carries more valuable information about the original signal, its future? We can conclude observing the corresponding MIF in figure 4 (the auto-MIF is at bottom) that more smoothing must not imply greater information gain, greater values of MIF (from the given three estimators the de-noised signal at level 5 gives the maximal information gain 0.281).

It would be desirable to have an analytical rule to say which wavelet shrinkage estimation is the best (in some sense). We offer to use the criterion (*MI criterion*)

$$\min_{w,j} \frac{1}{L} \sum_{i=1}^{L} (M_{\hat{y}_{w,j},y(L)}(i) - M_{y,y(L)}(i)). \qquad (3)$$

So if we want to use one of the de-noised versions of a signal as a predictor component, we have to choose the estimation, which gives the maximum average difference between $M_{\hat{y},y(L)}$ and $M_{y,y(L)}$.

Figures 7, 8 show the values of MI criterion (3) for noisy processes generated with two different SNR's. The wavelet shrinkage estimators were computed by DWT using default settings of wavelet shrinkage and orthogonal wavelets (5 coiflets, 7 symlets and 9 daublets) at every available level. We see, that the estimators received by the Haar wavelet (daublet 1) give in three from four cases the maximal difference (at levels 8, 8, 10). Figures 9, 10 show the original signals, their *best* MI estimators and the values of the corresponding MIFs. It is well known what role the zero (constant) process plays in estimating the white noise. Figure 10 shows the best MI estimator in presence of large noise *tends* to a constant value.

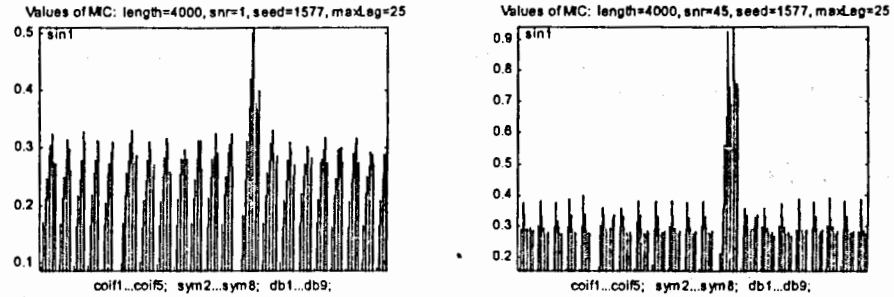

Fig. 7 Values of MIC for estimates of *sin x*

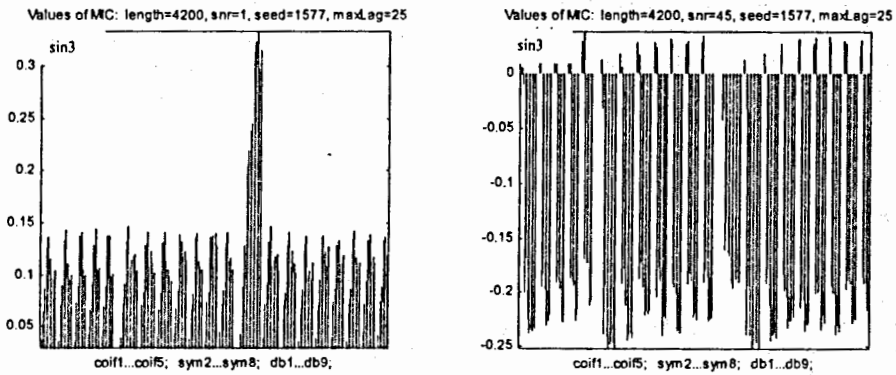

Fig. 8 Values of MIC for estimates of process from fig.5

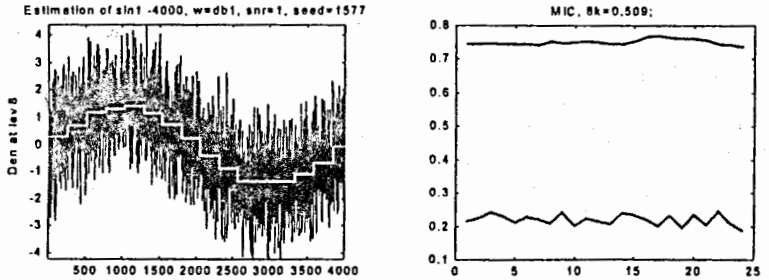

Fig. 9 The best MI estimation of *sin x* and the corresponding MIF

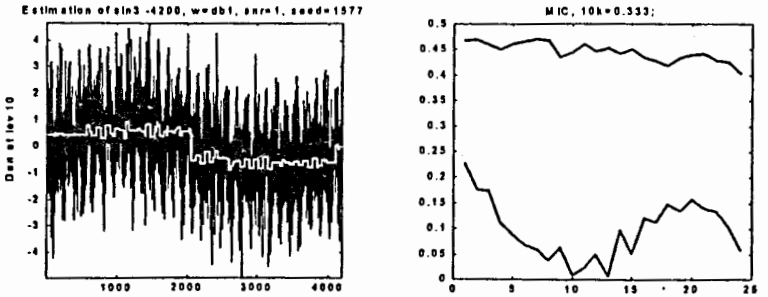

Fig. 10 The best MI estimation of *sin₃* and the corresponding MIF

For non-stationary signals the mutual information function $M_{\hat{y},y(L)} \equiv I(\hat{y}(t), y(t,L))$ depends on t. We analyzed $M_{\hat{y},y(L)}$ with different lengths of y and the *Haar* wavelets gave in general the maximum average difference.

6. Conclusion

The paper analyzed the wavelet de-noising and mutual information. It offers a new way one can look at the wealthy world of de-noised signals obtained by wavelet shrinking due to the introduced criterion. The mutual information function, by which one can characterize the de-noised signals from the viewpoint of prediction gain, seems to be a promising new tool not only for choosing new predictor components, but also for the wavelet analysis itself. The study has to be extended to some other important aspects and techniques in wavelet analysis, as e.g. the use of wavelet packet transformation.

References

[1] H.P.Bernhard, A Tight Upper Bound on the Gain of Linear and Nonlinear Predictors for stationary Stochastic Processes, IEEE (in print)

[2] H.P.Bernhard, G.Kubin, A fast mutual information calculation algorithm, in Signal processing VII: Theories and Applications, M.J.J.Holt, C.F.N. Cowan, P.M.Grant, and W.A.Sandham, Eds. vol.1, pp.50-53, Elsevier Science B.V., Amsterdam, 1994

[3] A.Bruce, H.Y.Gao, Wavelet Analysis with S-Plus, Springer, 1996

[4] D.L.Donoho, I.M.Johnstone, Ideal spatial adaptation via wavelet shrinkage, Biometrika, 81, 425-455, 1994

[5] D.L.Donoho, De-noising by soft threshholding, IEEE Transactions on Information Theory, 41(3), 613-627, May 1995

[6] A.M.Fraser, H.L.Swinney, Independent coordinates for strange attractors from mutual information, Physical Review A, vol.33, no.2, pp. 1134-1140, Feb.1986

[7] Handbook of numerical analysis, vol.V, Techniques of Scientific computing (part 2), Wavelets and Fast Numerical Algorithms (by Y.Meyer), Edited by P.G.Ciarlet and J.L.Lions, 1977, Elsevier NH

Received by Publishing Department
on August 12, 1999.