SOME APPLICATIONS OF WAVELETS TO TIME SERIES DATA

A Dissertation

by

JAE SIK JEONG

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

August 2008

Major Subject: Statistics

SOME APPLICATIONS OF WAVELETS TO TIME SERIES DATA

A Dissertation

by

JAE SIK JEONG

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Approved by:

Co-Chairs of Committee,	Marina Vannucci
	David B. Dahl
Committee Members,	Faming Liang
	Jerry Tsai
Head of Department,	Simon J. Sheather

August 2008

Major Subject: Statistics

ABSTRACT

Some Applications of Wavelets to Time Series Data. (August 2008)

Jae Sik Jeong, B.S., University of Seoul;

M.S., Seoul National University

Co-Chairs of Advisory Committee: Dr. Marina Vannucci
Dr. David B. Dahl

The objective of this dissertation is to develop a suitable statistical methodology for parameter estimation in long memory process. Time series data with complex covariance structure are shown up in various fields such as finance, computer network, and econometrics. Many researchers suggested the various methodologies defined in different domains: frequency domain and time domain. However, many traditional statistical methods are not working well in complicated case, for example, nonstationary process. The development of the robust methodologies against nonstationarity is the main focus of my dissertation. We suggest a wavelet-based Bayesian method which shares good properties coming from both wavelet-based method and Bayesian approach. To check the robustness of the method, we consider ARFIMA$(0, d, 0)$ with linear trend. Also, we compare the result of the method with that of several existing methods, which are defined in different domains, i.e. time domain estimators, frequency domain estimators. Also, we apply the method to functional magnetic resonance imaging (fMRI) data to find some connection between brain activity and long memory parameter.

Another objective of this dissertation is to develop a wavelet-based denoising technique when there is heterogeneous variance noise in high throughput data, especially protein mass spectrometry data. Since denoising technique pretty much depends on threshold value, it is very important to get a proper threshold value which involves estimate of stan-

dard deviation. To this end, we detect variance change point first and get suitable threshold values in each segment. After that, we apply local wavelet thresholding to each segment, respectively. For comparison, we consider several existing global thresholding methods.

To my family,

Thanks for the unlimited help and the strength to finish this work.

ACKNOWLEDGEMENTS

I would like to express thanks to my advisors, Dr. Marina Vannucci and Dr. David B. Dahl. Special thanks go to Dr. Marina Vannucci. Without her guidance and support this dissertation would not have been possible. I also wish to thank my committee members, Dr. Faming Liang and Dr. Jerry Tsai.

I also thank my mom. She always supports me mentally and spiritually in my whole life, no matter what I decide. I am grateful to my brother Jaejun, sister Eunmi, and their families for strong support. I am also thankful to my lovely nephew Seonwoo and pretty nieces Ahjin, and Jinseo.

TABLE OF CONTENTS

Page

CHAPTER Page

LIST OF FIGURES

LIST OF TABLES

TABLE Page

CHAPTER I

INTRODUCTION

With the improvement of computer technology, Bayesian approach has been flourished over the past decades. In many fields, Bayesian approach is being widely used now: finance, microarray study and so on. Various kinds of data which exist in real world have very complex structure. In order to deal with the data easily and apply statistical methods efficiently, we need to simplify such data with complex structure. To this end, wavelet method has been successfully used because it has many good features including decorrelation property. Since the covariance with complex structure can be approximately diagonalized through wavelet transform, we can use the simplified covariance for efficient estimation in the wavelet domain. Basically, there are two kinds of wavelets such as continuous wavelet transform (CWT) and discrete wavelet transform (DWT). DWT has many variants: Maximal overlap DWT (MODWT), Discrete Wavelet Packet Transform (DWPT), and Maximal overlap DWPT (MODWPT). Since each wavelet transform has advantages and disadvantages, we have to carefully decide which wavelet to use based on the situation we face with.

There are three types of memory in time series: short memory, no memory, and long memory. Time series which shows long range dependence has been considered in many fields including hydrology, econometrics, physics, and computer network. Estimation of model parameters in the presence of long range dependence has been major interest of many researchers. Especially, the estimation of long memory parameter d, transforming

---

The format and style follow that of *Journal of the American Statistical Association*.

nonstationary process to stationary one, has been the main issue of many research. Additionally innovation variance σ^2 is another crucial parameter to estimate as well.

Long memory is formally defined in an asymptotic sense by the asymptotic decay of the autocorrelations or the ultimate behavior of spectral density around zero. We will introduce two definitions of a stationary process with long memory which are defined in two different domains: spectral domain and frequency domain. A stationary time series $\{X_t\}$ is said to have long memory when long memory parameter d is between 0 and 0.5. Many methods of estimating long memory parameter based on its decay rate of autocorrelation or behavior of spectral density around zero have been studied by many researchers. However, it is not easy to do that in the presence of long range dependence due to complicated covariance structure. Thus, we propose a wavelet-based Bayesian method which shares good properties from both wavelet-based method and Bayesian approach, i.e. which uses Bayesian modeling on wavelet domain.

Estimation methods in time domain were suggested by many authors. Several heuristic approaches were based on autocorrelation plot or variance plot. More systematic method is R/S statistic, proposed by Hurst (1951). R/S statistic has several variants. For robustness, Beran (1994) and Taqqu, Teverovsky, and Willinger (1995) considered overlapped block when calculating the statistic. Peters (1994) used disjoint block and averaged the statistic. Instead of modifying numerator in R/S, Lo (1991) modified the denominator. Mielniczuk and Wojdyllo (2007) suggested bias corrected R/S estimator and presented simulation result for several variants of R/S.

Various semiparametric estimators of d were proposed, which are based on the behavior of spectral density around zero. The methods are called semiparametric since it is assumed that the spectral density is asymptotically equivalent to $|\lambda|^{1-2H}$. Geweke and Porter-Hudak (1983) proposed an estimator based on periodogram, which is an estimator of spectral density. They also suggest a cut-off for frequencies which are used for efficient

estimation. In the frequency domain, the cut-off value is very important. Under Gaussian assumption, the asymptotic distribution of estimator has been derived by Fox and Taqqu (1985). The Gaussian MLE has been studied by several authors. Yajima (1985) and Dalhaus (1989) presented the asymptotic distribution of exact MLE of long memory parameter. Unfortunately, calculation of the exact MLE need a lot of computational work. To reduce such computation work, approximation of the likelihood function could be an alternative. Fox and Taqqu (1986) used the Whittle method for the estimation of long range dependence parameter. Also, local Whittle method has been studied by Robinson (1995). Taqqu and Teverovsky (1997) studied the robustness of Whittle-type methods through empirical study. Asymptotic properties of the local Whittle estimator of ARFIMA$(0, d, 0)$ is given in Shimotsu and Phillips (2006). Since those methods are parametric or semi-parametric the performance pretty much depends on how well the parametric assumption fits the data.

Other than those methods, development of estimation methods in spatial process was studied by Frias, Alonso, Ruiz-Medina, and Angulo (2008). Hall, Hardle, Kleinow, and Schmidt (2000) suggested new semiparametric bootstrap approach for confidence intervals for the long range dependence parameter.

In this research, we focus on the process which displays long range dependence: AutoRegressive Fractionally Integrated Moving Average (ARFIMA) model, ARFIMA with linear trend and functional magnetic resonance imaging (fMRI).

As another application, wavelet-based approach for denoising is widely used in proteomics. Since mass spectrometry data require complex pre-processing and poorly preprocessed data have a bad effect on the result of statistical analysis, pre-processing is very important. Especially, removal of noise has been the main focus of pre-processing steps. Wavelet denoising techniques have become standard for such task. We suggest local wavelet thresholding method for efficient removal of noise when there is heterogeneous variance noise. In this research, we will focus on the real ovarian cancer mass spectrometry

data.

CHAPTER II

WAVELET TRANSFORM

2.1 Wavelets and wavelet transform

It is very well known that any L^2 function can be represented by functions belongings to a proper basis. Wavelets are those functions which are consisted of orthonormal basis.

Wavelets satisfy some conditions such as admissibility condition, integration to 0 and square-integrable to 1. Following wavelet transforms, we get two kinds of coefficients such as scaling coefficients and wavelet coefficients. Generally speaking, scaling coefficients showing global feature are averages of the original data over corresponding scale while wavelet coefficients showing local nature are differences of weighted averages. Basically, there are two main waves of wavelets: continuous wavelet transform (CWT) and Discrete wavelet transform (DWT). Also, there are three different types of DWT: standard DWT (DWT), maximal overlap DWT (MODWT), discrete wavelet packet transform (DWPT). Here we focus on discrete type wavelets.

2.1.1 Porperties of wavelets

Discrete wavelet transform has various good properties: parsimonious representation, energy decomposition, effective decorrelation properties, and perfect reconstruction. The beauty of wavelet transform is decorrelation property. For instance, many phenomena in real world have very complex structure itself. However, since many statistical methods assume that data have very ideal and simple structure such as independence it doesn't fit the problem. Often, we are faced with this gap between statistical methodologies and real data. Due to the decorrelation property of wavelet transform, this gap can be removed. The covariance with very complex structure can be simplified into wavelet domain in which

transformed data are almost uncorrelated, so that we can apply statistical methods to the data with complex structure.

$Parsimonious\ representation$: we can reconstruct the original signal by using a few coefficients in the wavelet domain via inverse wavelet transform. For illustrative example, the doppler signal is used. Haar wavelet was used as a decomposing wavelet.

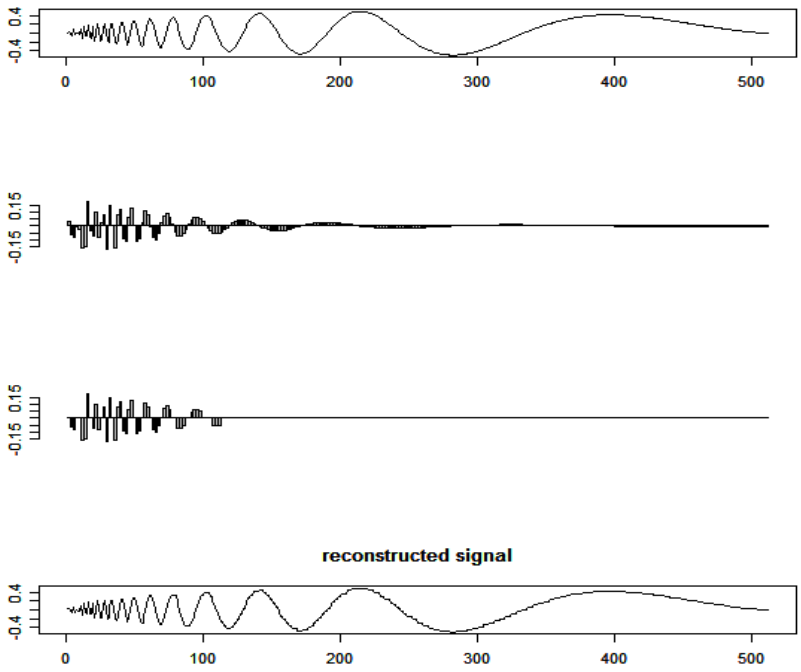

Figure 1: Parsimonious representation.

The mathematical formula of this signal is given

$$f(x) = \sqrt{x(1-x)}\sin\left(\frac{2.1\pi}{x+0.05}\right),\ 0 \le x \le 1.$$

In figure 1, with a few coefficients the signal can be constructed.

$Decorrelation\ property$: The data with very complicated structure can be simplified via wavelet transform. For illustrative example, simulated ARFIMA(0,d,0) data were used. Here long memory parameter d is 0.4 and sample size is 512. As a decomposing wavelet

the least asymmetric wavelet with vanishing moments of 8 was used in figure 2.

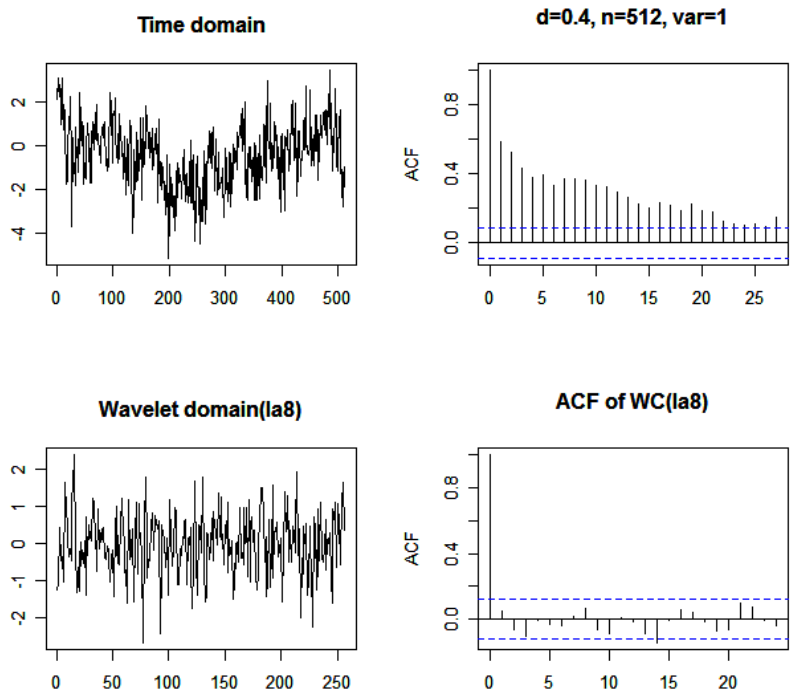

Figure 2: Decorrelation property.

All estimates of the autocorrelation for each lag in the wavelet domain exist in the 95 $\%$ confidence interval which is made under the assumption that the coefficients are uncorrelated.

2.1.2 Discrete wavelet transform

Other than many good properties mentioned in the previous section, DWT has an additional attractive aspect that makes us prefer the method: easy and fast computation. With the help of pyramid algorithm, which was introduced by Mallat (1989), the output through DWT can be computed by using $O(n)$ multiplications. Also, it has some relationship with other wavelet transform. In other words, it can be thought of as a subsampling of the continuous wavelet transforms(CWT) at dyadic scales or wise subsampling and normalizing of

maximal overlap discrete wavelet transforms(MODWT), see Percival and Walden (1999).

If we let X be a time series realizations, then we can write $W = \mathcal{W}X$, where N is a length of X and $\mathcal{W}$ is an $N \times N$ real valued matrix defining the DWT. For example, each row of $\mathcal{W}$ consists of wavelet filter or its shift version of wavelet filter. Let $\{h_l, l = 0, \cdots, L-1\}$ and $\{g_l, l = 0, \cdots, L-1\}$ be wavelet filter and scaling filter, respectively. L is the size of filter. Sometimes wavelet filter is called the mother wavelet. W consists of wavelet coefficients, $w_{i,t}$ and scaling coefficients, $v_{i,t}$ obtained by wavelet transform where i is the resolution level and t is time point of interest.

At the first level, both coefficients are represented

$$w_{1,t} = \sum_{l=0}^{L-1} h_l X_{2t+1-l \text{ mod } N}, \; v_{1,t} = \sum_{l=0}^{L-1} g_l X_{2t+1-l \text{ mod } N} \tag{2.1}$$

where $t = 0, \cdots, N/2$. At the jth level, the scaling coefficients obtained in the $j-1$ level can be used.

At the j-th level, both coefficients are represented

$$w_{j,t} = \sum_{l=0}^{L-1} h_l v_{j-1,2t+1-l \text{ mod } N/2^j}, \; v_{j,t} = \sum_{l=0}^{L-1} g_l v_{j-1,2t+1-l \text{ mod } N/2^j} \tag{2.2}$$

where $t = 0, \cdots, N/2^j$. For consistency, some people consider the original data X_t as $v_{0,t}$. Due to decimating property, we have $N/2^j$ wavelet and scaling coefficients at level j. The constraint on sample size, $N = 2^J$ can be relaxed by considering partial discrete wavelet transform.

2.1.3 Maximal overlap discrete wavelet transform

Maximal overlap discrete wavelet transform (MODWT), a variation on the DWT, has been widely used for some reasons. Unlike the DWT, MODWT does not have constraint on sample size and does not decimate the coefficients as well, which produce the same size of wavelet and scaling coefficients as data at any level. Furthermore, since it does not

depend on starting point, MODWT is called shift-invariant or translation-invariant DWT. Actually, DWT can be considered as a kind of MODWT because we can get all the DWT coefficients through wise subsampling and renormalizing MODWT coefficients, Percival and Walden (1999). Let $\{\tilde{h_l}, l = 0, \cdots, L-1\}$ and $\{\tilde{g_l}, l = 0, \cdots, L-1\}$ be wavelet filter and scaling filter, respectively. The wavelet coefficients and scaling coefficients at level j can be represented

$$w_{j,t} = \sum_{l=0}^{L-1} \tilde{h_l} v_{j-1,2t+1-l \bmod N}, \; v_{1,t} = \sum_{l=0}^{L-1} \tilde{g_l} v_{j-1,2t+1-l \bmod N} \tag{2.3}$$

where $t = 0, \cdots, N$. There is a relationship between filters in DWT and filters in MODWT:

$$\tilde{h_j} = h_j / 2^{j/2}, \; \tilde{g_j} = g_j / 2^{j/2} \tag{2.4}$$

where $t = 0, \cdots, N$.

2.1.4 Discrete wavelet packet transform

Wavelet packets are introduced by Coifman and Meyer and are extended to more general case by Wickerhauser (1994). In general, wavelet packets are regarded as linear combinations of wavelet functions, and form an orthonormal basis of $L^2(R)$, see Vidakovic (1999).

One major difference between standard DWT and DWPT is the frequency band which is decomposed by the methods. For example, the j-th level DWPT decomposes the frequency interval $[0, 1/2]$ into 2^j equal intervals. On the other hand, in the case of the standard DWT, wavelet coefficients at level j describe the frequency band $[1/2^{j+1}, 1/2^j]$.

Let $\{h_l, l = 0, \cdots, L-1\}$ and $\{g_l, l = 0, \cdots, L-1\}$ be wavelet filter and scaling filter, respectively. The elements at level j can be represented

$$w_{j,n,k} = \sum_{l=0}^{L_j - 1} \nu_{j,n,l} X_{2^j[t+1]-1-l \bmod N}, \; t = 0, \cdots, N_j - 1$$

where $\nu_{j,n,l} = \sum_{k=0}^{L-1} \nu_{n,k} \nu_{j-1,n/2,l-2^{j-1}k}, \; l = 0, \cdots, L_j - 1$. Let

$$\nu_{n,k} = \begin{cases} g_l, & \text{if } n \bmod 4 = 0, \text{ or } 3; \\ h_l, & \text{if } n \bmod 4 = 1, \text{ or } 2; \end{cases}$$

For more details about DWPT and MODWPT, see Percival and Walden (1999) and Vidakovic (1999).

2.2 Thresholding policy

According to the way how we process the wavelet coefficients, the thresholding rules are determined. There are so many well-known shrinkage rules: soft, hard, semisoft, firm, nonnegative garrote, n-degree garrote, and hyperbole shrinkage and so on. The mathematical expressions for the hard, soft, semisoft and nonnegative garrote, n degree garrote, and hyperbole thresholding rules are

$$\delta^h(d,\lambda) = d1(|d| > \lambda) \tag{2.5}$$

$$\delta^s(d,\lambda) = (d - sign(d)\lambda)1(|d| > \lambda) \tag{2.6}$$

$$\delta^{ss}(d,\lambda_1,\lambda_2) = sign(d)\frac{\lambda_2(|d|-\lambda_1)}{\lambda_2-\lambda_1}1(\lambda_1 < |d| \le \lambda_2) + d1(|d| > \lambda_2) \tag{2.7}$$

$$\delta^{nng}(d,\lambda) = (d - \frac{\lambda^2}{d})1(|d| > \lambda) \tag{2.8}$$

$$\delta^{g}(d,\lambda) = \frac{d^{2n+1}}{\lambda^{2n}+d^{2n}} \tag{2.9}$$

$$\delta^{hy}(d,\lambda) = sign(d)\sqrt{d^2-\lambda^2}1(|d| > \lambda). \tag{2.10}$$

The plots for these thresholding rules are given in figure 3.

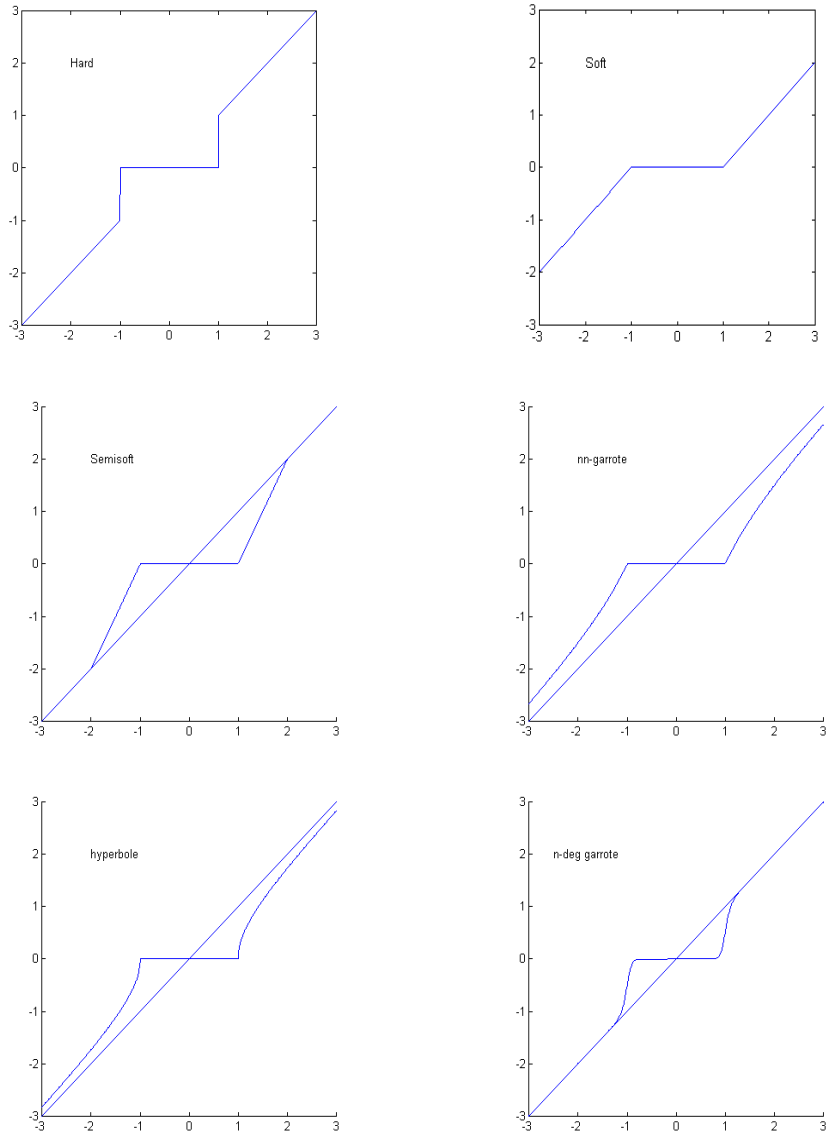

Figure 3: First row: Hard (left) and Soft (right). Second row: Semisoft (left) and nonnegetive-garrote (right). Third row: Hyperbole (left) and n degree garrote (right).

As we can see, the semisoft thresholding rule is getting close to soft rule as λ_2 goes to infinity. Also, it is getting closer to hard thresholding rule as λ_2 goes to λ_1.

Bickel (1983) showed that hard thresholding uniformly better than soft one in terms of the maximum of the MSE. For any λ, Gao and Bruce (1996) demonstrated that one can select appropriate λ_1 and λ_2 such that risk of semisoft rule is smaller than that of hard one.

2.2.1 Ways to select threshold value

We have to decide how to select a threshold value before we apply thresholding rules to the data. To the end, there are so many existing methods: universal, sure, block thresholds, percentile threshold and so on.

$universal\ threshold$: Donoho and Johnstone (1994) call the threshold, $\lambda = \sqrt{2\log n}\sigma$ universal threshold. It has asymptotic minimax properties. For the estimator of σ, two estimators are widely used. One is sample standard deviation estimator obtained by using the finest detail coefficients. The other one is MAD estimator, median absolute deviation from the median. i.e. $\hat{\sigma} = 1.4826MEDIAN[|d^{(J-1)} - MEDIAN(d^{(J-1)})|]$ where $n = 2^J$. Pickands (1967) proved that under some conditions noise is removed from the data after thresholding with threshold value given above.

$SURE\ threshold$: SureShrink threshold is selected by minimizing Stein's unbiased estimator of risk. It is adaptive denoising procedure because it is done by specifying thresholds level-wise.

$$\lambda^{sure} = \underset{\lambda}{\operatorname{argmin}}\, SURE(d, \lambda), \tag{2.11}$$

where $SURE(d,\lambda) = k - 2\sum_{i=1}^{k} 1(|d_i| \leq \lambda) + \sum_{i=1}^{k}(|d_i \wedge \lambda|)^2$. The core part of SureShrink is the combination of λ^{sure} and soft thresholding.

$Block\ thresholding$: Block thresholding rule shrinks wavelet coefficients in groups while other methods mentioned above shrink coefficients individually. Since this rule considers dependence of neighboring coefficients, it has better performance than other rules: small bias. The rule, however, is more sensitive to selection of threshold.

Other than these rules, there are many other existing rules. Nason (1996) addressed twofold cross-validation procedure. He showed that optimal threshold can be almost always found. Also, he said that the rule does not perform well in the case of heavy-tail noise. Bruce (1994) addressed a heavy-tail noise problem with a robust smoother-cleaner

wavelet transformation. In addition to that, threshold selection from the minimum description length (MDL) point of view was introduced by several researchers. According to Saito's approximate MDL (AMDL), the biggest k wavelet coefficients are considered.
Other than equal spaced design, nonequispaced (NES) designs are also studied by many researchers. The simplest way is coercion to equal spacing. i.e. we carry out standard wavelet transform as if the data are equally spaced. Another method is interpolation and averaging. This method has basically two steps: interpolate values at equally spaced points and carry out standard wavelet transform.

2.3 Wavelet denoising

Following the seminal work of Donoho and Johnstone, wavelet thresholding has successfully been used in various applications to remove noise and recover the true signal. This can be done by applying wavelet transform, applying thresholding rule to wavelet coefficients, and coming back to the wavelet domain through inverse wavelet transform. In the course of this process, we have to decide several things: soft or hard thresholding, global or adaptive thresholding, and decomposing wavelet. Soft thresholding maps wavelet coefficients less than a threshold to 0 while hard one shrinks all coefficients by threshold. Global thresholding applies the same cut-off value to all coefficients whereas level-dependent or adaptive thresholding uses different thresholds from level to level, which depend on the resolution level of the wavelet transform. In addition to that, local wavelet thresholding is used for special problems under the assumption that there is heteroscedastic variance. Note that for all rules this work is applied to wavelet coefficients only, not to scaling coefficients. As an illustrative example for denoising, We consider doppler signal and added normal noise to the signal such that signal to noise ratio (SNR) is about 3. The original signal and contaminated signal are given in figure 4.

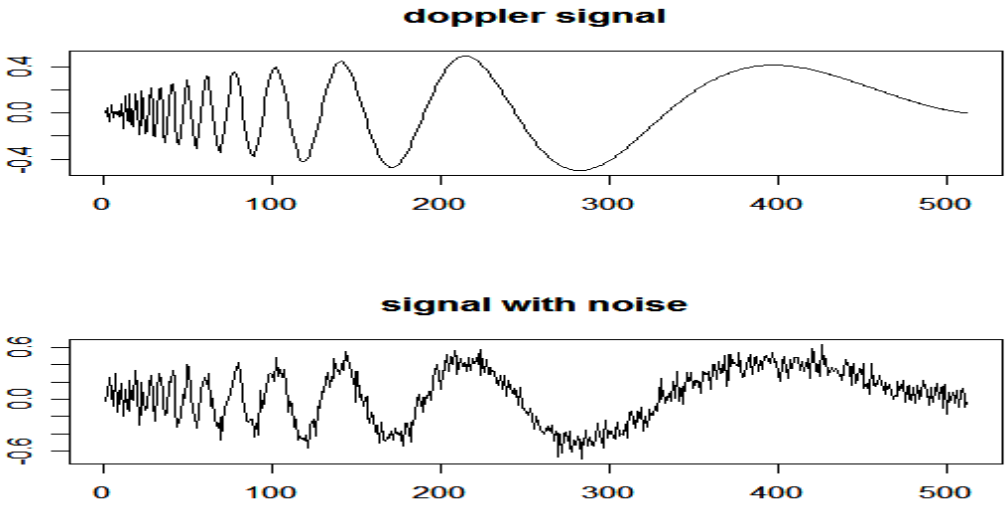

Figure 4: First row: Original signal. Second row: Contaminated signal.

Here we consider only hard and soft thresholding, universal and sure threshold. Daubechies wavelet with vanishing moment of 4 is used. Results obtained by four combinations mentioned above are given in figure 5.

As we can see, we get slightly different reconstructed signals according to what decomposing wavelet we use and what thresholding policy we choose. The selection of decomposing wavelet and decision of thresholding rule should be made with caution according to the situation we face.

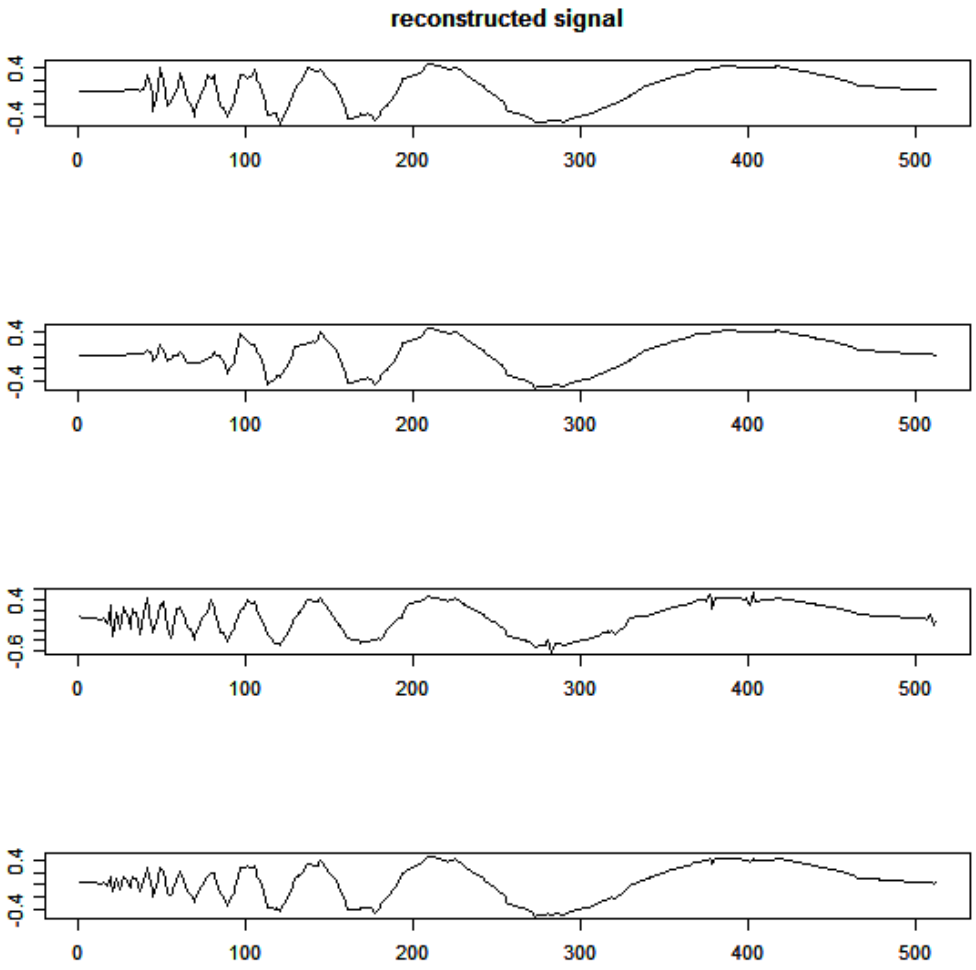

Figure 5: First row: Universal and Hard. Second row: Universal and Soft. Third row: Sure and Hard. Fourth row: Sure and Soft.

CHAPTER III

APPLICATION OF WAVELET THRESHOLDING TO THE CASE OF HETEROGENEOUS ERROR VARIANCE

3.1 Application to proteomics

Over the past decades there has been an increased interest in using high throughput data in cancer studies: mass spectrometry data and microarray gene expression data. Protein mass spectrometry data require quite complex preprocessing techniques. Clearly, poorly preprocessed data have a bad effect on the results of subsequent statistical analysis. Various preprocessing techniques were investigated by many authors. Basically, preprocessing consists of several steps: baseline subtraction, normalization, denoising, peak identification, and peak alignment. Here we focus on denoising step only. In principle, there are two types of noises: the chemical noise and the electrical noise. To remove such noise, wavelet thresholding technique is considered as a standard tool. Existing methods assume that there is homogeneous error variance. We, however, notice that there exists heterogeneous error variance in the mass spectrometry (MS) data. For efficient removal of noise, we apply our local wavelet thresholding to the MS data after variance change point detection.

3.1.1 Variance change point detection

Suppose that we have baseline subtracted MS data. First of all, we have to detect variance change points in the data for local wavelet thresholding. To this end, the iterated cumulative sums of squares algorithm (ICSS), proposed by Inclán and Tiao (1994) is used. The procedure is based on the assumption that the independent observations x_t have mean 0 and variance ${\sigma_t}^2,\ t = 1, \cdots, n$. Null and alternative hypothesis are given:

$$H_0 : {\sigma_1}^2 = \cdots = {\sigma_n}^2 \ v.s.\ H_a : \text{ not } H_0$$

Cumulative sums of squares is defined: $C_k = \sum_{i=1}^{k} x_i$. The test statistic D is given $D = \max(D^+, D^-)$ where

$$D^+ = \max_{1 \le k \le n-1} \left(\frac{k+1}{n} - P_k \right),$$
$$D^- = \min_{1 \le k \le n-1} \left(P_k - \frac{k}{n} \right),$$
$$P_k = \frac{C_k}{C_n},\ k = 1, \cdots, n.$$

If the maximum absolute value of D exceeds a certain predetermined value, then we estimate a change point at point $k^* = \operatorname{argmax}_k D$. For more details about this algorithm, see Inclán and Tiao (1994).

Whitcher, Guttorp, and Percival (2000) apply the ICSS algorithm to coefficients of DWT of long memory data. They also obtained empirical predetermined value of D under the null hypothesis by using Monte Carlo simulation. Gabbanini, Vannucci, Bartoli, and Moro (2004) extended the ICSS algorithm to DWPT and MODWPT.

3.1.2 Binary segmentation procedure

The procedure mentioned above is designed for the detection of single change point. This method, however, can be easily extended to the case of multiple change points. At the first stage of the procedure we test the null hypothesis for the whole data. If we do not reject H_0 we declare that there is no change point in the data, otherwise we divide the data into two subseries by the detected change point. We repeat this process until there is no change point. If this process is done we check those change points detected again to locate more reliable change points. This confirmatory step is to merge neighboring two subseries which were divided by change points. Then we test if the change point is detected again. If we still reject H_0 we keep the point as a change point, otherwise we remove it from the set of change points. This extra step helps to reduce masking effect and to get more reliable change point estimates.

3.1.3 Local wavelet thresholding

As mentioned in section 2, we have several things to decide before applying wavelet thresholding. Commonly used universal threshold and sure threshold are used here and median absolute deviation is used as a noise variance estimator. After variance change point detection, we apply the local wavelet thresholding. The key steps are

1. Compute the wavelet transforms of the data (DWT, MODWT, DWPT, and MODWPT).

2. Use the ICSS algorithm by using DWPT and MODWPT coefficients to locate the variance change points.

3. Divide the MODWT coefficients into segments.

4. Compute local threshold values for each segment.

5. Apply standard wavelet thresholdong to each segment.

6. Reconstruct the signal by using denoised coefficients.

3.2 Application to real ovarian cancer data

We briefly describe the MS data. Serum samples collected at the Mayo Clinic between 1980 and 1989 were analyzed by surface-enhanced laser desorption and ionization time-of-flight (SELDI-TOF) mass spectrometry using the CM10 chip type. The ProteinChip Biomarker System was used for protein expression profiling. More detailed description of the samples and exclusion criteria can be found in Moore, Fung, McGuire, Rabkin, Molinaro, Wang, Zhang, Wang, Yip, Meng, and Pfeiffer (2006).

We focus on the 50 samples obtained after 1986 whose serum was freeze-thawed a single time and applied a Bayesian variable selection approach for classification. In the

analysis we discarded small m/z values less than 2000 due to large noise, and large m/z values greater than 15,000 due to low intensities. For the remaining data, we interpolated the mass spectra on a grid of equally spaced m/z values with 50,000 equi-spaced points using piecewise cubic splines. One MS data is given in figure 6.

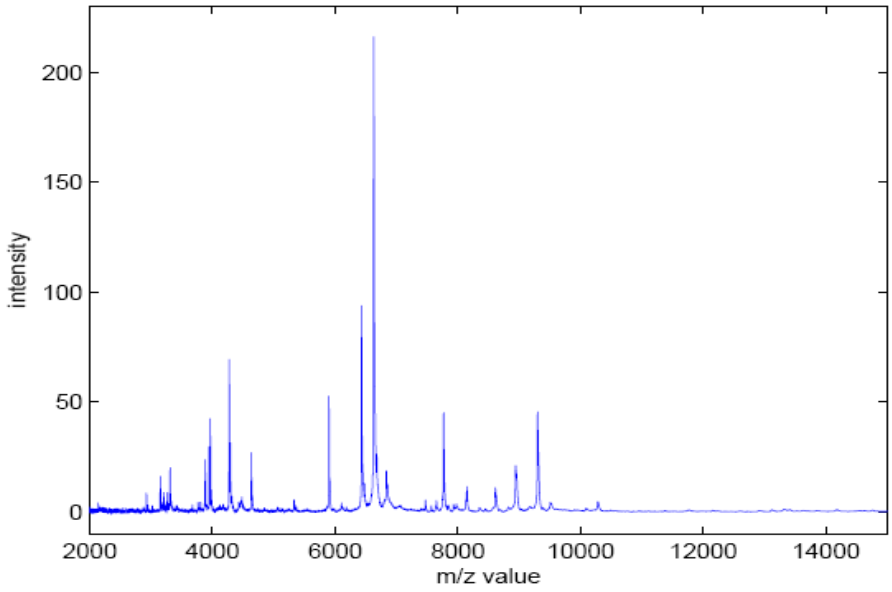

Figure 6: Ovarian cancer: One sample spectrum.

3.2.1 Wavelet denoising

We use the maximal overlap discrete wavelet transforms (MODWT) with Daub(4) along with an adaptive soft thresholding rule. Figure 7 shows estimated standard deviations of the noise for four randomly chosen spectra, which were obtained by running an MAD estimator with window size 1,500 on the finest MODWT coefficients. This shows explicit monotonic decrease as the m/z values increase.

For comparison, we consider global thresholding with two different threshold values: $4\hat{\sigma}^2$ and $40\hat{\sigma}^2$, i.e. C is 4,40.

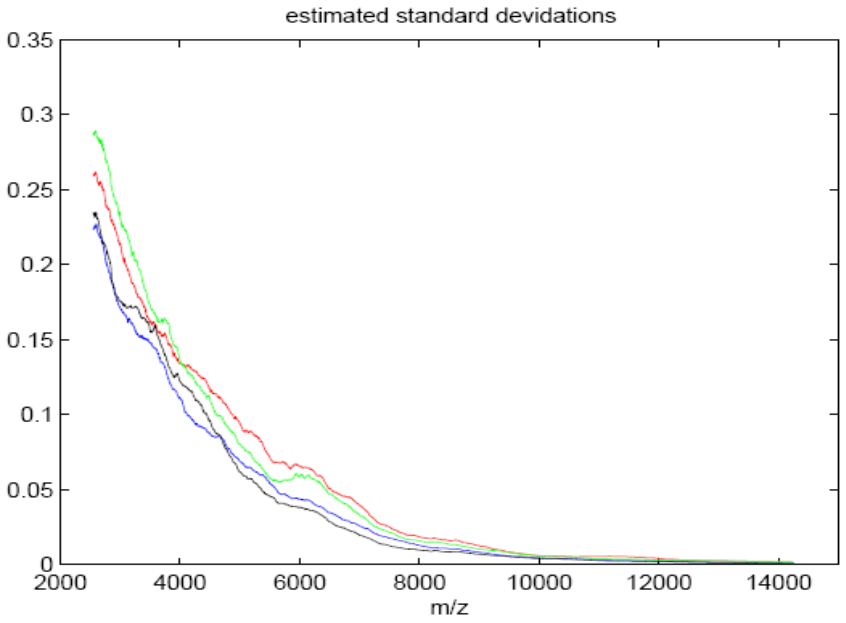

Figure 7: Estimated standard deviations for 4 MS spectra randomly chosen.

3.2.2 Comparison

We applied the local thresholding method to each of 50 mass spectra. For each spectrum, the variance change point detection method was applied. The number of change points detected varies from 7 to 36. Then for peak detection, the SpecAlign software of Wong, Cagney, and Cartwright (2005) was used. The approach based on mean spectrum, proposed by Morris, Coombes, Kooman, Baggerly, and Kobayashi (2005) was used. The local thresholding method detect 58 peaks while the global thresholding detect 53 ($C = 4$) and 48 ($C = 40$) peaks, respectively on the entire m/z range. Let's focus on low range of m/z values, 2000 to 5000. In this interval, the local thresholding detect 21 peaks while global thresholding detect 18, 17 peaks, respectively. Figure 8 presents the detected change point in this part of range.

Then, denoised spectra obtained by using each method are given in figure 9.

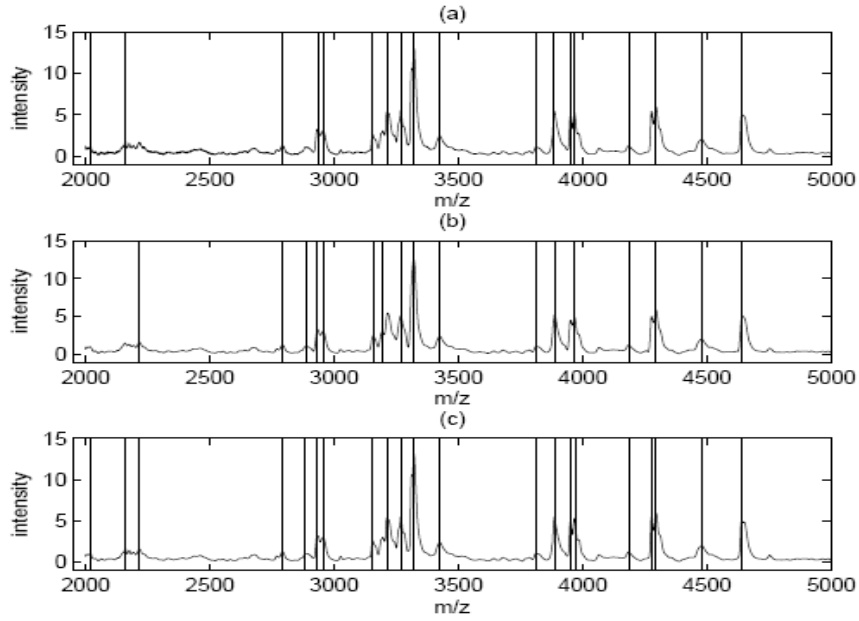

Figure 8: Plot of the location of detected change points by each method.

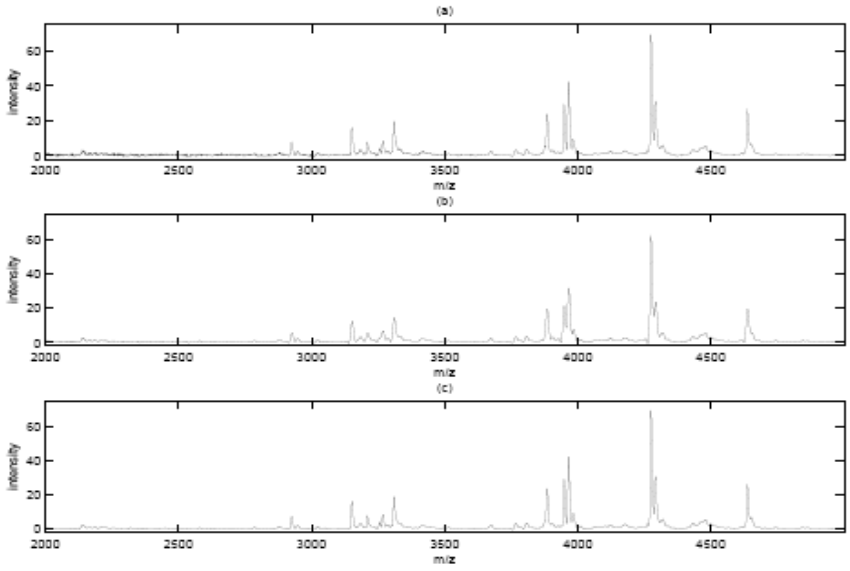

Figure 9: Denoised mass spectra by global and local thresholding: First and second row: global thresholding with $C = 4, 40$, respectively. Third row: local wavelet thresholding.

CHAPTER IV

LONG MEMORY PROCESS

4.1 Long memory process

Long memory process is defined in an asymptotic sense by the asymptotic decay of the autocorrelations or the behavior of spectral density around zero. We here introduce two definitions of a stationary process with long memory or long range dependence which are defined in two different domains: spectral domain and frequency domain.

Definition 4.1.1. *(see J. Beran)* X_t *is called a stationary process with long memory if there exists a real number* $\alpha \in (0, 1)$ *and a constant* $c_\rho > 0$ *such that*

$$\lim_{k\to\infty} \frac{\rho(k)}{c_\rho k^{-\alpha}} = 1,$$

where $\rho(\cdot)$ *is autocorrelation function.*

Definition 4.1.2. X_t *is called a stationary process with long memory if there exists a real number* $\beta \in (0, 1)$ *and a constant* $c_f > 0$ *such that*

$$\lim_{\lambda\to 0} \frac{f(\lambda)}{c_f|\lambda|^{-\beta}} = 1$$

where $f(\cdot)$ *is spectral density function.*

Even though two definitions mentioned above are defined in different domains, Zygmund (1953) proved that these two definitions are equivalent in some sense. For more details about definitions above, see the Beran (1994).

Here we will describe two most widely used models which display long memory or long range dependence: fractional Gaussian noise (fGN) and autoregressive fractionally integrated moving average (ARFIMA).

4.1.1 Fractional Gaussian noise

The fractional Gaussian noise $\{X_t,\ t \geq 1\}$ can be easily obtained through fractional Brownian motion $B_H(t)$ where H is a self-similar parameter or Hurst effect. Here we notice that the relationship between self-similar parameter H and long memory parameter d gives $d = H - \frac{1}{2}$. Even though the relationship between the long memory parameter and self-similar parameter changes in the case of long memory process with symmetric α-stable innovations ($H = d + \frac{1}{\alpha}$), we don't consider the case here. For more information, see Stoev and Taqqu (2005). More specifically, fractional Gaussian noise is the stationary increment of fractional Brownian motion

$$X_t = B_H(t+1) - B_H(t),\ t \geq 1.$$

Time series $\{X_t\}$ is a zero mean stationary process which has autocovariance

$$\gamma(h) \sim H(2H-1)h^{2H-2} \text{ as } h \to \infty$$

and spectral density

$$f(\lambda) \sim C|\lambda|^{1-2H} \text{ as } \lambda \to 0$$

where C is a constant.

4.1.2 ARFIMA(p, d, q) models

Autoregressive fractional integrated moving average(ARFIMA) models are a natural extension of the classic ARIMA models. It was introduced by Granger and Joyeux (1980) and Hosking (1981).

A process $\{X_t,\ t \geq 1\}$ is called ARFIMA(p, d, q) process if it is the stationary solution of

$$\phi(B)(1-B)^d X_t = \psi(B)\epsilon_t \tag{4.1}$$

where B is the backshift operator;

$$\phi(B) = 1 - \sum_{i=1}^{p} \phi_i B^i \tag{4.2}$$

$$\psi(B) = 1 + \sum_{i=1}^{q} \psi_i B^i \tag{4.3}$$

Since we are interested in stationary ($d < \frac{1}{2}$) and invertible ($d > -\frac{1}{2}$) process, the long memory parameter d in (4.1) can be assumed to be any real value in this interval $-\frac{1}{2} < d < \frac{1}{2}$. Furthermore, since we are interested in long memory process, our focus in on $d \in (0, 1/2)$. By binomial expansion, $(1-B)^d$ can be written as

$$(1-B)^d = \sum_{k=0}^{\infty} \binom{d}{k} (-1)^k B^k \tag{4.4}$$

with the binomial coefficients

$$\binom{d}{k} = \frac{\Gamma(d+1)}{\Gamma(k+1)\Gamma(d-k+1)}. \tag{4.5}$$

The simplest case of ARFIMA(p, d, q) models is ARFIMA$(0, d, 0)$ process defined by

$$(1-B)^d X_t = \epsilon_t. \tag{4.6}$$

The autocovariance function of ARFIMA$(0, d, 0)$ process is given by

$$\gamma(k) = \frac{(-1)^k \Gamma(1-2d)}{\Gamma(1-d+k)\Gamma(1-d-k)} \sim C_\gamma k^{2d-1} \tag{4.7}$$

and the autocorrelation function is given by

$$\rho(k) = \frac{\Gamma(1-d)\Gamma(d+k)}{\Gamma(d)\Gamma(1-d+k)} \sim C_\rho |k|^{2d-1} \tag{4.8}$$

These processes are stationary if $d < 0.5$ and possess an invertible moving average representation if $d > -0.5$. More specifically, they have long memory for $0 < d < 0.5$, white noise for $d = 0$ and short memory for $-0.5 < d < 0$.

4.2 Estimation methods: review of existing methods

Many different estimation methods of long memory parameter or long range dependence parameter are introduced by many authors over the past several decades. There are several ways that we can categorize these estimators. For example, each estimator is included in one of these categories: parametric, semi-parametric, and nonparametric estimators. However, we classify each estimator into three different types of methods in terms of the domain in which each estimator is defined: time domain estimators, frequency domain estimators, and wavelet domain estimators.

4.2.1 Time domain estimators

4.2.1.1 Correlogram

As we know, the autocorrelation of the stationary process with long range dependence is proportional to k^{2d-1} where k is lag. The heuristic idea of the correlogram is as follows:

1. calculate sample autocorrelation for each lag k
2. plot $\log|\rho(k)|$ vs $\log k$
3. estimate the slope (β)

We expect the points in the plot to be scattered around a straight line with slope of $2d-1$. Thus we can get long memory parameter estimate of $\hat{d} = \frac{\hat{\beta}+1}{2}$

4.2.1.2 Variance-time estimator

The variance of a stationary process with long range dependence has the following asymptotic order:

$$Var(\bar{X}_n) \sim C_H n^{2H-2}.$$

The variance-time estimator, proposed by Cox and Smith (1953), has the following steps: For an integer k,

1. divide the sequence into m_k non-overlapping subseries of length $k = n/m_k$

2. calculate the sample means and overall mean

$$\bar{X}_j = \text{sample mean of } j \text{ subseries}, \ \bar{X} = {m_k}^{-1} \sum_{i=1}^{m_k} \bar{X}_i$$

3. calculate the sample variance of the sample means

$${S^2}_k = (m_k - 1)^{-1} \sum_{i=1}^{m_k} (\bar{X}_i - \bar{X})^2$$

4. plot $\log {S^2}_k$ vs $\log k$

5. estimate the slope

We can get estimate of long memory parameter, $\hat{d} = \frac{1+\hat{\beta}}{2}$.

4.2.1.3 HR estimator

Yajima (1985) and Dalhaus (1989) proved that, for Gaussian processes with long memory, the MLE is asymptotically efficient in the sense of Fisher. However, the calculation of the log-likelihood function or its derivative requires lots of computational work. Suppose that X_t is a Gaussian process. The log-likelihood function is given by

$$L_n(x;\theta) = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\log|\Sigma(\theta)| - \frac{1}{2}x^t\Sigma^{-1}(\theta)x, \tag{4.9}$$

where $x = (x_1, \cdots, x_n)^t \in R^n$ and $\theta = (\sigma^2, H)$. The first partial derivative of (4.9) is given by

$${L'}_n(x;\theta) = -\frac{1}{2}\frac{\partial}{\partial\theta_j}\log|\Sigma(\theta)| - \frac{1}{2}x^t\left[\frac{\partial}{\partial\theta_j}\Sigma^{-1}(\theta)\right]x \ (j = 1, 2). \tag{4.10}$$

The MLE $\hat{\theta}$ is the solution of the ${L'}_n(x;\hat{\theta}) = 0$. If the dimension of parameters is high or if we have a long time series, the calculation of the exact MLE is not easy and not numerically stable as well because the equation (4.10) involves the calculation of the determinant and

the inverse of $\Sigma(\theta)$. Maximizing an approximation to the likelihood function could be an alternative to solving the exact maximum likelihood equation. There are several approximate MLE methods which are obtained by approximating the likelihood function in many different ways. This HR estimator is one of the approximate MLE's obtained by using the fast and accurate method of Haslett and Raftery. The heuristic idea of approximation of this method is to use the autoregressive approximations. A Gaussian ARFIMA process can be represented by autoregressive process of infinite order. However, since we observe a finite number of samples we have the truncated model:

$$X_t - \theta_1 X_{t-1} - \cdots - \theta_m X_{t-m} = \epsilon_t,\ m < t \leq n$$

where θ_i are the coefficients of $\Phi(B)\Theta^{-1}(B)(1-B)^d$. After more approximations and refinements, a quasi maximum likelihood estimator (QMLE) $\hat{\theta}_n$ is obtained by maximizing

$$L^*{}_n(x;\theta) = K - \frac{n}{2}\log\left(\hat{\sigma}_\epsilon^2(\theta)\right)$$

where $\hat{\sigma}_\epsilon^2(\theta) = \frac{1}{n}\sum_{t=1}^{n}\frac{(X_t-\hat{X}_t)^2}{\nu_t},\ \nu_t = var(X_t - \hat{X}_t),\ \hat{X}_t = \Phi(B)\Theta(B)^{-1}\sum_{i=1}^{t-1}\phi_{ti}X_{t-i}$ and $\phi_{ti} = -\binom{t}{i}\frac{\Gamma(i-d)\Gamma(t-d-i+1)}{\Gamma(-d)\Gamma(t-d+1)}$. For more information about this approximation, see Haslett and Raftery (1989).

4.2.1.4 Beran estimator

Another version of the autoregressive approximation approach is proposed by Beran (1994). The sequence of innovations is given

$$\epsilon_t = X_t - \sum_{i=1}^{\infty}\pi_i X_{t-i}$$

Using observable finite past only, the truncated $\epsilon_t,\ u_t$ is given

$$u_t(\theta) = X_t - \sum_{i=1}^{t-1}\pi_i X_{t-i},\ t = 2, \cdots, n.$$

Let $\theta = (\sigma_\epsilon, \phi_1, \cdots, \phi_p, \theta_1, \cdots, \theta_q, d)$ and $r_t(\theta) = u_t(\theta)/\sigma_\epsilon$. The approximate MLE is obtained by minimizing

$$L_2 = n\log(\theta_1) + \sum_{t=2}^{n} {r_t}^2(\theta).$$

We can see more details about this method in Beran (1994).

4.2.1.5 R/S statistic

This method is based on the rescaled adjusted range statistic, proposed by Hurst (1951). For a time series $\{X_t, t = 1, \cdots, n\}$, let $Y_j = \sum_{i=1}^{j} X_i$ denote the partial sum. The adjusted range $R(t,k)$ is defined by

$$R(t,k) = \max_{0\le i\le k}[Y_{t+i} - Y_t - \frac{i}{k}(Y_{t+k} - Y_t)] - \min_{0\le i\le k}[Y_{t+i} - Y_t - \frac{i}{k}(Y_{t+k} - Y_t)].$$

Also, the function $S^2(t,k)$ is defined by

$$S^2(t,k) = \frac{1}{k}\sum_{i=t+1}^{t+k}(X_i - \bar{X}_{t,k})^2$$

where $\bar{X}_{t,k} = \frac{1}{k}\sum_{i=t+1}^{t+k} X_i$ and k is the block size.

The rescaled adjusted range statistic is defined by the ratio of the two terms above:

$$R/S = \frac{R(t,k)}{S(t,k)}.$$

The idea of the method proposed by Hurst (1951) is to regress the $\log(R/S)$ on $\log(k)$. Since the expectation of the R/S is asymptotically equivalent to $C_H n^H$, the estimate $\hat{H}$ can be obtained in the pox-plot, which is equal to the slope, i.e. $\hat{d} = \hat{\beta} + 1/2$.

There are several variants of R/S method according to the way we calculate the statistic. Peters (1994) considered the disjoint block and averaged $\log(R/S)$ while Beran (1994) and Taqqu et al. (1995) used overlapped block and pox-plot.

4.2.1.6 *Modified R/S statistic*

Lo (1991) proposed the modified R/S statistic which is defined by

$$R/S_{lo} = \frac{1}{S_q}[\max_{1\le k\le n}\sum_{j=1}^{k}(X_j - \bar{X}_n) - \min_{1\le k\le n}\sum_{j=1}^{k}(X_j - \bar{X}_n)]$$

where ${S_q}^2 = S^2(0,n) + \frac{2}{n}\sum_{j=1}^{q}(1-\frac{j}{q+1})[\sum_{i=j+1}^{n}(X_i - \bar{X}_n)(X_{i-j} - \bar{X}_n)],\ q < n.$

He applied this method to U.S. market data and noticed that there is a little evidence of long range dependence in the data.

4.2.2 *Frequency domain estimators*

The behavior of periodogram around origin is widely used for estimation of long memory parameter because a stationary process with long range dependence has a spectral density proportional to $|\lambda|^{1-2H}$. For example, spectral densities for the most widely used processes with long range dependence are given as follows:

$$\text{ARFIMA}(0,d,0) \sim |2\sin(\lambda/2)|^{1-2H} \text{ and fGN} \sim |\lambda|^{1-2H}. \tag{4.11}$$

We note that $\lim_{\lambda\to 0}\frac{\sin\lambda}{\lambda} = 1$, i.e. $\sin(\lambda)$ is aysmptotically equivalent to λ around zero. Thus, the choice of cut off for frequencies is very crucial in this case.

4.2.2.1 *GPH estimator*

This estimator is based on the first k periodogram ordinates

$$I_j = \frac{1}{2\pi n}\left|\sum_{t=0}^{n-1} X_t e^{iw_j t}\right|^2,\ j = 1,\cdots,k, \tag{4.12}$$

where $w_j = 2\pi j/n$ and k is a positive integer. In the case of ARFIMA(0,d,0),

$$\log f(\lambda) \sim C_\lambda + (1-2H)\log|2\sin(\lambda/2)|$$

where f is spectral density. Since periodogram $I(\lambda)$ is an asymptotically unbiased estimator of f, we notice the linear relationship between log periodogram and log frequencies.

The estimaotr of d is given by

$$\hat{d}_{GPH} = -\frac{0.5\sum_{j=1}^{k}(x_j - \bar{x})\log I_j}{\sum_{j=0}^{k}(x_j - \bar{x})^2}, \tag{4.13}$$

where $x_j = \log|2\sin(w_j/2)|$ and $\bar{x} = \sum_{j=1}^{k} x_j$.

4.2.2.2 SR estimator

This estimator is based on the regression equation using the smoothed periodogram function in (4.13). The estimaotr of d is given by

$$\hat{d}_{SR} = -\frac{0.5\sum_{j=1}^{k}(x_j - \bar{x})\log {I^S}_j}{\sum_{j=0}^{k}(x_j - \bar{x})^2}, \tag{4.14}$$

where ${I^S}_j$ is the smoothed version of I_j in (4.13).

4.2.2.3 Whittle estimator

This estimator is based on approximate maximum likelihood estimator using spectral density and periodogram. As mentioned above, since exact maximum likelihood method requires lots of computational work Whittle consider approximation of determinant and inverse of covariance matrix. Following Grenander and Szecö (1958), $\frac{1}{n}\log|\Sigma^{-1}|$ can be approximated by

$$\frac{1}{2\pi}\int_{-\pi}^{\pi}\log f(\lambda;\theta)d\lambda.$$

Also, the second term in likelihood equation (4.9) $\frac{1}{2n}x'\Sigma^{-1}x$ is approximated by

$$\frac{1}{4\pi}\int_{-\pi}^{\pi}\frac{I(\lambda)}{f(\lambda)}d\lambda$$

where $I(\lambda) = \frac{1}{n}|\sum_{j=1}^{n}(X_j - \bar{X})e^{(i\lambda j)}|^2$. After appropriate normalization which makes the first approximation term be 0, the Whittle estimator can be regarded as the value of the

parameter which minimizes the function

$$Q(\theta) = \int_{-\pi}^{\pi} \frac{I(\lambda)}{f(\lambda;\theta)} d\lambda. \tag{4.15}$$

For example, the parameter θ is $({\sigma_\epsilon}^2, d)$ in the case of ARFIMA(0,d,0) process. Also, for simplicity, the discrete version of Whittle estimator, proposed by Graf (1983) for fGN, is considered

$$Q^*(\theta) = \sum_{j=1}^{M} \frac{I(\lambda_j)}{f^*(\lambda_j,\theta)}$$

where $\lambda_j = 2\pi j/n,\ M = (n-1)/2$ if n is odd number.

4.2.2.4 Aggregated Whittle estimator

Since the Whittle method is parametric, the Whittle estimator may not be accurate if the parametric assumption is not correct. To reduce this kind of risk, the robust aggregated Whittle estimator was suggested. When a time series is long enough, the estimator is more robust. The idea is to aggregate the data first and get shorter series

$${X_k}^{(m)} = \frac{1}{m} \sum_{i=(k-1)m+1}^{mk} X_i,\ k = 1, \cdots, m/n$$

where m is the aggregation level and n is the length of time series. The choice of m is very crucial because the best value of m is not known. Thus, practical choice of m was used by several authors. The idea is to estimate long range parameter for different levels of aggregation. Then, we find a region in which the plot of estimates is flat.

4.2.2.5 Local Whittle estimator

Robinson (1995) proposed a semiparametric estimator, local Whittle estimator. This method involves two properties from both Whittle estimator and aggregated Whittle estimator: the behavior of the spectral density around zero and minimization of the function Q in equation

(4.15). We get estimate which minimizes the function

$$R(H) = \log\left(\frac{1}{M}\sum_{i=1}^{M}\frac{I(\lambda_i)}{\lambda_i^{1-2H}}\right) - (2H-1)\frac{1}{M}\sum_{i=1}^{M}\log(\lambda_i)$$

where $\lambda_i = \frac{2\pi i}{n}$ and M is the number of periodogram used. Under the some conditions, Robinson (1995) showed that $\hat{H}$ converges in probability to the true value. Also, he showed that $\hat{H}$ had asymptotically normal distribution with mean of H and variance $\frac{1}{4M}$.

4.2.3 Wavelet domain estimators

The parametric methods mentioned above are not accurate if the assumed parametric assumption is not correct. To reduce such risk, various ideas like averaging and aggregation were introduced by many authors. In spite of that, the methods still have some problems if there is deterministic trend, a kind of nonstationarity. However, wavelet-based method is efficient and robust against this kind of nonstationarity when we estimate long memory parameter in the presence of deterministic trend.

4.2.3.1 Abry Veitch estimator

Abry and Veitch (1998) proposed an estimator, which is based on wavelet coefficient obtained by Mallat's pyramidal algorithm. The AV statistic is defined by

$$AV(j,k) = \frac{1}{n_j}\sum_{k}|d(j,k)|^2$$

where $d(j,k)$ is the wavelet coefficient at level j and n_j is the number of wavelet coefficient at level j. This estimator is obtained in the following way:

1. calculate wavelet coefficients, $d(j,k)$ for level $j = j_1, \cdots, j_2$
2. calculate each AV statistic
3. regress $\log_2(AV(j,k))$ on j by weighted least square with weight $(n\ln^2 2)/2^{j+1}$

4. estimate the slope approximately equal to $2H - 1$

Here, the choice of N, vanishing moment is important. Theoretically, the larger N is the better the estimation. In practical situations, however, the estimator is affected by boundary effect because larger N need longer wavelet filter. They suggested $H+1$ as a good practical compromise. Also, we have to consider the choice of $j_1,\ j_2$ and decomposing wavelet as well. They used Daubechies wavelet because it has finite support and it does not result in an excessive extension of the support even though we increase the vanishing moment.

4.2.3.2 Wavelet-based Bayesian estimator

Ko and Vannucci (2006) proposed a wavelet-based Bayesian estimator. They take discrete wavelet transform and compute the exact variances and covariances of the wavelet coefficients via efficient recursive algorithm proposed by Vannucci and Corradi (1999). They used the decorrelation property of wavelet. Twefik and Kim (1992) and Dijkerman and Mazumdar (1994) showed that the correlation of wavelet coefficients decrease exponentially and hyperbolically across scales and along time, respectively. The method is given in the following way:

1. take discrete wavelet transform
2. specify priors on wavelet coefficients
3. calculate posterior
4. get parameter estimates via MCMC

In Ko and Vannucci (2006), they used Daubechies minimum phase wavelet with vanishing moment of 7 and Haar wevelet. They used non-informative prior for unknown parameters and MCMC technique for posterior inference. For comparison, they considered GPH estimator and MLE estimator. By comprehensive simulation study, they noticed that wavelet-

based Bayesian method is better than GPH and MLE in terms of bias and MSE. Also, they apply the method to real US GNP data and Nile river data. In the case of Nile data, they found that their estimate agrees with other estimates obtained by several existing methods. Furthermore, they got empirical credible interval of long memory parameter d.

CHAPTER V

BAYESIAN ESTIMATION VIA WAVELETS

5.1 Linear model

Consider the following usual form of linear regression equation,

$$y = X\beta + \varepsilon, \tag{5.1}$$

where y is an $(N \times 1)$ signal, X is an $(N \times p)$ deterministic design matrix and ε is an $(N \times 1)$ zero-mean Gaussian $1/f$-like noise with ARFIMA$(0, d, 0)$ process with variance-covariance matrix Σ_ε. The first two moments of y are

$$E(y) = X\beta \tag{5.2}$$

and

$$\text{cov}(y) = \Sigma_y = \Sigma_\varepsilon. \tag{5.3}$$

We take discrete wavelet transform (DWT) to the both sides of the model (5.1) in order to get approximately diagonalized variance-covariance matrix of $1/f$-type noise ε. We assume that N is a power of two. The original model (5.1) becomes via DWT

$$y_w = X_w\beta + \varepsilon_w, \tag{5.4}$$

where $y_w = \mathcal{W}y,\ X_w = \mathcal{W}X$ and $\varepsilon_w = \mathcal{W}\varepsilon$ where $\mathcal{W}$ is an $N \times N$ real-valued and orthogonal matrix defining the DWT. The first two moments of y_w are

$$E(y_w) = X_w\beta \tag{5.5}$$

and

$$\text{cov}(y_w) = \text{cov}(\varepsilon_w) = \sigma^2\Sigma_d, \tag{5.6}$$

where Σ_d is an $(N \times N)$ nearly diagonalized matrix such that its diagonal elements are $f_j(d)$ for the approximation level $j = 1, 2, \ldots, J$ of DWT, where $f_j(d)$ is a function of the long memory parameter d. We express it in matrix notation by the product of σ^2 and Σ_d. Since y is multivariate normal distribution and DWT is a linear transformation, y_w are also multivariate normal distribution.

5.2 Bayesian modeling on wavelet domain

The likelihood function is

$$\begin{aligned} L(y_w|\Theta, X_w) &= \frac{|\Sigma_w|^{-1/2}}{(\sqrt{2\pi})^N} \exp\left\{-\frac{1}{2}(y_w - X_w\beta)'{\Sigma_w}^{-1}(y_w - X_w\beta)\right\} \\ &= \frac{(\sigma^2)^{-N/2}|\Sigma_d|^{-1/2}}{(\sqrt{2\pi})^N} \exp\left\{-\frac{1}{2\sigma^2}(y_w - X_w\beta)'{\Sigma_d}^{-1}(y_w - X_w\beta)\right\}, \end{aligned} \tag{5.7}$$

where $\Theta = (\beta, \sigma^2, d)$. Putting priors on parameters of long memory process, we derive posterior distribution for posterior inference.

5.2.1 Prior specification and joint posterior

We use beta distribution as the prior distribution of d which indicate the long range dependent behavior of error term ε

$$\pi(2d) = \frac{\Gamma(\eta + \nu)}{\Gamma(\eta)\Gamma(\nu)}(2d)^{\eta-1}(1 - 2d)^{\nu-1},\ 0 < d < 1/2. \tag{5.8}$$

For the prior distribution of (β, σ^2), we use normal-inverse gamma

$$\begin{aligned} \pi(\beta, \sigma^2) &= \pi(\beta|\sigma^2)\pi(\sigma^2)\ \left(i.e.\ N(\beta_0, \sigma^2 I_2) \times IG\left(\frac{\delta_0}{2}, \frac{\gamma_0}{2}\right)\right) \\ &= (\sigma^2)^{-1} \exp\left\{-\frac{1}{2\sigma^2}(\beta - \beta_0)'(\beta - \beta_0)\right\} \\ &\quad \times \frac{1}{(\sigma^2)^{\delta_0/2+1}} \exp\left\{-\frac{\gamma_0}{2\sigma^2}\right\}. \end{aligned} \tag{5.9}$$

Under the assumption of independence between (β, σ^2) and d, the joint prior distribution is,

$$\pi(\beta, \sigma^2, d) = \pi(\beta|\sigma^2)\pi(\sigma^2)\pi(d). \tag{5.10}$$

The posterior distribution of Θ given (y_w, X_w) is

$$\begin{aligned}\pi(\beta, \sigma^2, d|y_w, X_w) &\propto L(y_w|X_w, \Theta)\pi(\Theta) \\ &\propto (\sigma^2)^{-(\frac{N+\delta_0+2}{2}+1)}|\Sigma_d|^{-\frac{1}{2}}(2d)^{\eta-1}(1-2d)^{\nu-1} \\ &\times \exp\left\{-\frac{1}{2\sigma^2}\left[\gamma_0 + T(\beta, \Sigma_d) + (\beta-\beta_0)'(\beta-\beta_0)\right]\right\}\end{aligned} \tag{5.11}$$

where $T(\beta, \Sigma_d) = (y_w - X_w\beta)'{\Sigma_d}^{-1}(y_w - X_w\beta)$.

5.2.2 Full conditional distribution for posterior inference

We can get easily the full conditional distributions of each parameter. The full conditional distribution of β in our model is

$$\beta|\sigma^2, d, y_w, X_w \sim N\left((X_w{}^{*'}X_w{}^{*} + I_2)^{-1}(X_w{}^{*'}y_w{}^{*} + \beta_0),\ \sigma^2(X_w{}^{*'}X_w{}^{*} + I_2)^{-1}\right), \tag{5.12}$$

where $X_w{}^{*} = {\Sigma_d}^{-1/2}X_w$ and $y_w{}^{*} = {\Sigma_d}^{-1/2}y_w$.

The full conditional distribution of σ^2 is

$$\sigma^2|\beta, d, y_w, X_w \sim IG\left(\frac{N+\delta_0+2}{2},\ \frac{1}{2}\left[\gamma_0 + T(\beta, \Sigma_d) + (\beta-\beta_0)'(\beta-\beta_0)\right]\right) \tag{5.13}$$

where $T(\beta, \Sigma_d) = (y_w - X_w\beta)'{\Sigma_d}^{-1}(y_w - X_w\beta)$.

The full conditional distribution of d is

$$\begin{aligned}\pi(d|\beta, \sigma^2, y_w, X_w) \propto\ & |\Sigma_d|^{-1/2}\exp\left\{-\frac{1}{2\sigma^2}\left[(y_w - X_w\beta)'{\Sigma_d}^{-1}(y_w - X_w\beta)\right]\right\} \\ &\times (2d)^{\eta-1}(1-2d)^{\nu-1}.\end{aligned} \tag{5.14}$$

Since the full conditional distribution (5.14) of d is not a closed form of known distribution, the MCMC can be employed by implementing a Metropolis step using beta proposal distribution for d within Gibbs steps for β and σ^2. In the Metropolis step for d the acceptance probability α of a candidate point d_{new} is

$$\alpha = \min\left\{\frac{\pi(d_{new}|\beta, \sigma^2, y_w, X_w)}{\pi(d_{old}|\beta, \sigma^2, y_w, X_w)}, 1\right\}. \tag{5.15}$$

Derivation of full conditional distribution of β: from equation (5.11), we get

$$\begin{aligned}\pi(\beta|\sigma^2, d, y_w, X_w) &\propto (\sigma^2)^{-(\frac{N+\delta_0+2}{2}+1)}|\Sigma_d|^{-\frac{1}{2}}(2d)^{\eta-1}(1-2d)^{\nu-1} \\ &\quad \times \exp\left\{-\frac{1}{2\sigma^2}\left[\gamma_0 + T(\beta,\Sigma_d) + (\beta-\beta_0)'(\beta-\beta_0)\right]\right\} \\ &\propto \exp\left\{-\frac{1}{2\sigma^2}\left[\beta' X_w'{\Sigma_d}^{-1}X_w\beta + \beta'\beta - 2y_w'{\Sigma_d}^{-1}X_w\beta - 2\beta_0'\beta\right]\right\} \\ &= \exp\left\{-\frac{1}{2\sigma^2}\left[\beta'(I + X_w'{\Sigma_d}^{-1}X_w)\beta + -2(y_w'{\Sigma_d}^{-1}X_w - 2\beta_0')\beta\right]\right\} \\ &\propto \exp\left\{-\frac{1}{2\sigma^2}(\beta-K)'[I + X_w'{\Sigma_d}^{-1}X_w](\beta-K)\right\}\end{aligned} \tag{5.16}$$

where $K = [I + X_w'{\Sigma_d}^{-1}X_w]^{-1}(\beta_0 + X_w'{\Sigma_d}^{-1}y_w)$ and $T(\beta,\Sigma_d) = (y_w - X_w\beta)'{\Sigma_d}^{-1}(y_w - X_w\beta)$. K can be represented by $({X_w}^{*'}{X_w}^{*} + I_2)^{-1}({X_w}^{*'}{y_w}^{*} + \beta_0)$ since ${X_w}^{*} = {\Sigma_d}^{-1/2}X_w$ and ${y_w}^{*} = {\Sigma_d}^{-1/2}y_w$.

Since the last row of equation (5.16) is the kernel function of multivariate normal distribution, we get the posterior distribution in equation (5.12) mentioned above.

Derivation of full conditional distribution of σ^2: from equation (5.11), we get directly

$$\begin{aligned}\sigma^2|\beta, d, y_w, X_w \propto \exp\left\{-\frac{1}{2\sigma^2}\left[\gamma_0 + T(\beta,\Sigma_d) + (\beta-\beta_0)'(\beta-\beta_0)\right]\right\} \\ \times (\sigma^2)^{-(\frac{N+\delta_0+2}{2}+1)}.\end{aligned} \tag{5.17}$$

where $T(\beta,\Sigma_d) = (y_w - X_w\beta)'{\Sigma_d}^{-1}(y_w - X_w\beta)$. Since the equation (5.17) is the kernel function of Inverse Gamma distribution we get the posterior distribution in equation (5.13).

Vannucci and Corradi (1999) have proposed a recursive way of computing covariances of wavelet coefficients by using the recursive filters of the DWT and the algorithm has an interesting link to the two-dimensional discrete wavelet transform (DWT2) that makes computations simple. In the context of this paper, the variance-covariance matrix Σ_ε of the wavelet coefficients in equation (5.6) can be computed by first applying the DWT2 to the matrix Σ_ε in equation (5.3). The diagonal blocks of the resulting matrix will provide

the within-scale variances and covariances at the different levels. One can then apply the one-dimensional DWT to the rows of the off-diagonal blocks to obtain the across-scale variances and covariances.

5.3 MCMC

Markov chain Monte Carlo (MCMC) methods are defined in Roberts and Sahu (1997) this way: A MCMC method for the simulation of a distribution f is any method producing an ergodic Markov chain $X^{(t)}$ whose stationary or equilibrium distribution is f. i.e., if the chain is irreducible and aperiodic, then the chain will become stationary at target distribution. Thus, if we run chain long enough, we can consider samples after burn-in period as random samples from a target distribution. Here we introduce two basic MCMC algorithms which are used in application section.

5.3.1 The Metropolis-Hastings Algorithm

Hastings (1970) extended the Metropolis' algorithm to the case when proposal distribution is not necessarily symmetric. Suppose that the distribution q is proposal distribution. Metropolis-Hastings algorithm is given as follows: for given x_t,

1. generate a new value $Y_t \sim q(y|x_t)$
2. take $X_{(t+1)} = \begin{cases} Y_t \text{ with prob. } a(x_t, Y_t), \\ x_t \text{ with prob. } 1 - a(x_t, Y_t), \end{cases}$
   where acceptance probability, $a(x_t, Y_t) = \min\{\frac{f(y)}{f(x)}\frac{q(x|y)}{q(y|x)}, 1\}$.

In the case of random-walk Metropolis, Roberts and Gilks (1994) suggested that the rule of thumb of acceptance rate is to maintain a 25% to 35%.

5.3.2 The Gibbs sampling

The Gibbs sampling algorithm is a special case of Metropolis-Hastings algorithm with acceptance probability of 1. Two types of Gibbs sampling algorithms are given in Liu (2004).

$Randon - Scan\ Gibbs\ Sampler$: For a given $x^{(t)} = ({x_1}^{(t)}, \cdots, {x_p}^{(t)})$,

1. iterate the steps below

   (a) randomly select a index i from $\{1, \cdots, p\}$ according to a given probability vector(e.g. discrete uniform probability)

   (b) generate ${x_i}^{(t+1)} = f_i(x_i|x_{-i})$,

where the density f_i is the full conditional.

$Systematic - Scan\ Gibbs\ Sampler$: For a given $x^{(t)} = ({x_1}^{(t)}, \cdots, {x_p}^{(t)})$,

1. generate ${x_1}^{(t+1)} = f_1(x_1|{x_2}^{(t)}, \cdots, {x_p}^{(t)})$,

2. generate ${x_2}^{(t+1)} = f_2(x_2|{x_1}^{(t+1)}, {x_3}^{(t)}, \cdots, {x_p}^{(t)})$,

   $\vdots$

p. generate ${x_p}^{(t+1)} = f_p(x_p|{x_1}^{(t+1)}, \cdots, {x_{p-1}}^{(t+1)})$,

where the densities $f_1, \cdots, f_p$ are the full conditionals.

Under regularity conditions, it turned out that a Gibbs sampler chain converges geometrically to stationary distribution and its convergence rate varies according to the correlation among variables. Roberts and Sahu (1997) have shown that random scan strategy can outperform systematic scan in terms of convergence rate. Also, for the discrete space problem, Liu (1996) proposed a modified Gibbs sampler, metropolized Gibbs sampler and he proved that this sampler is statistically more efficient than the random scan Gibbs sampler.

In real life example, we can use the combination of the two algorithms above, Metropolis within Gibbs algorithm. Especially, when some full conditionals are known form of distribution and the others are not known form we often use the hybrid algorithm.

More information about Metropolis Hastings sampler and Gibbs sampler is available in various MCMC materials, e.g. Liu (2004) and Robert (1999).

CHAPTER VI

APPLICATION TO ARFIMA AND FMRI

6.1 Simulation design

There are numerous ways to generate a time series that exhibits long memory properties. A computationally simple one was proposed by McLeod and Hipel (1978). It involves the Cholesky decomposition of the correlation matrix. Here we consider two simple processes with long range dependence: ARFIMA and ARFIMA with linear trend. In the application to ARFIMA$(0, d, 0)$, we compare the performance of several methods. Then, we apply the wavelet-based Bayesian method to the simulated ARFIMA data with linear trend and simulated fMRI data. In the second application, we can compare the robustness of each method against nonstationarity such as linear trend.

6.2 Application to simple ARFIMA model

Here we consider the ARFIMA$(0, d, 0)$ model

$$(1 - B)^d X_t = \epsilon_t,\ \epsilon_t \sim (0, \sigma^2) \tag{6.1}$$

where B is the backshift operator.

We generated 100 simulated time series of length 128, 256 and 512 for three different d values (0.1, 0.25, 0.4) respectively. Also, three different σ^2's (0.5,1,3) are considered. The following figure 10 shows trace plot of a time series of length 512 when the long memory parameter is 0.25 and $\sigma^2 = 1$.

The following figure 11 presents the variability of GPH estimator according to sample size n and the figure 12 displays the variability of 4 estimators for sample size of 512.

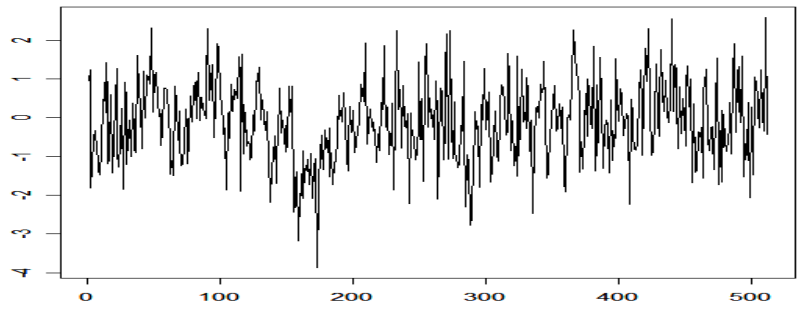

Figure 10: ARFIMA(0,d,0): a time series ($d = 0.25$ and $\sigma^2 = 1$).

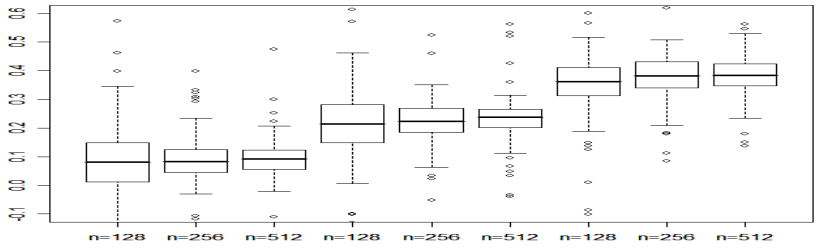

Figure 11: ARFIMA$(0, d, 0)$: Boxplot of $\hat{d}$(GPH).

The following tables (1) and (2) represent the estimates of long memory parameters and the corresponding MSEs obtained by the methods mentioned in the previous section when $\sigma^2 = 1$.

Table 1: ARFIMA$(0, d, 0)$: estimates $\hat{d}$ from each method ($\sigma^2 = 1$)

d	n	GPH	SR	HR	fEXP
0.1	128	0.0780	0.0298	0.0751	0.0954
	256	0.0942	0.0362	0.0806	0.0945
	512	0.1108	0.0688	0.0906	0.0990
0.25	128	0.2229	0.1419	0.1969	0.2296
	256	0.2263	0.1851	0.2230	0.2407
	512	0.2203	0.1943	0.2365	0.2460
0.4	128	0.3783	0.3038	0.3436	0.3936
	256	0.4112	0.3494	0.3700	0.3981
	512	0.4026	0.3653	0.3901	0.4056

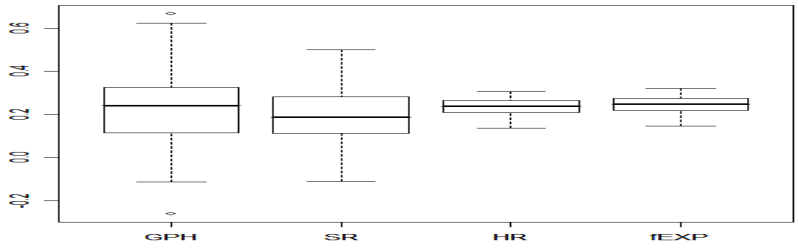

Figure 12: ARFIMA$(0, d, 0)$: Boxplot of $\hat{d}$ of 4 methods ($d = 0.25,\ n = 512$).

Table 2: ARFIMA$(0, d, 0)$: MSE from each method ($\sigma^2 = 1$).

d	n	GPH	SR	HR	fEXP
0.1	128	0.0904	0.0425	0.0043	0.0061
	256	0.0448	0.0338	0.0024	0.0023
	512	0.0222	0.0170	0.0012	0.0011
0.25	128	0.0808	0.0534	0.0086	0.0067
	256	0.0544	0.0310	0.0031	0.0027
	512	0.0313	0.0201	0.0015	0.0014
0.4	128	0.0971	0.0559	0.0071	0.0057
	256	0.0391	0.0250	0.0035	0.0034
	512	0.0204	0.0168	0.0013	0.0015

The results for different values of $\sigma^2 = 0.5,\ 3$ are given in table (3) to table (6), respectively.

Table 3: ARFIMA$(0, d, 0)$: estimates $\hat{d}$ from each method ($\sigma^2 = 0.5$)

d	n	GPH	SR	HR	fEXP
0.1	128	0.0915	0.0132	0.0811	0.1035
	256	0.0870	0.0578	0.0802	0.0935
	512	0.0819	0.0602	0.0901	0.0980
0.25	128	0.2668	0.1847	0.2030	0.2376
	256	0.2522	0.2035	0.2360	0.2547
	512	0.2596	0.2195	0.2352	0.2447
0.4	128	0.3827	0.3185	0.3504	0.4033
	256	0.4415	0.3795	0.3742	0.4012
	512	0.4089	0.3621	0.3866	0.4024

Table 4: ARFIMA$(0, d, 0)$: MSE from each method ($\sigma^2 = 0.5$)

d	n	GPH	SR	HR	fEXP
0.1	128	0.0682	0.0424	0.0044	0.0064
	256	0.0488	0.0289	0.0027	0.0026
	512	0.0300	0.0168	0.0011	0.0011
0.25	128	0.0671	0.0453	0.0093	0.0083
	256	0.0450	0.0291	0.0023	0.0024
	512	0.0231	0.0175	0.0017	0.0016
0.4	128	0.0678	0.0490	0.0067	0.0060
	256	0.0571	0.0256	0.0024	0.0021
	512	0.0268	0.0205	0.0011	0.0011

Table 5: ARFIMA$(0, d, 0)$: estimates $\hat{d}$ from each method ($\sigma^2 = 3$)

d	n	GPH	SR	HR	fEXP
0.1	128	0.1390	0.0598	0.0800	0.1007
	256	0.1460	0.0885	0.0850	0.0986
	512	0.0836	0.0474	0.0923	0.1002
0.25	128	0.2481	0.1716	0.2111	0.2474
	256	0.2143	0.1630	0.2307	0.2491
	512	0.2744	0.2085	0.2420	0.2521
0.4	128	0.3813	0.2949	0.3339	0.3800
	256	0.4258	0.3694	0.3762	0.4053
	512	0.4082	0.3520	0.3884	0.4035

Table 6: ARFIMA$(0, d, 0)$: MSE from each method ($\sigma^2 = 3$)

d	n	GPH	SR	HR	fEXP
0.1	128	0.0864	0.0438	0.0048	0.0069
	256	0.0372	0.0233	0.0024	0.0026
	512	0.0300	0.0193	0.0014	0.0014
0.25	128	0.0654	0.0513	0.0075	0.0069
	256	0.0553	0.0420	0.0030	0.0029
	512	0.0308	0.0233	0.0015	0.0015
0.4	128	0.0628	0.0553	0.0093	0.0071
	256	0.0342	0.0158	0.0026	0.0026
	512	0.0329	0.0233	0.0014	0.0015

Table 7: ARFIMA$(0, d, 0)$ with linear trend: The estimates $\hat{d}$ from each method($\sigma^2 = 1, \beta = 0.001$)

d	n	GPH	SR	HR	fEXP	WB
0.1	256	0.0539	-0.0063	0.0730	0.0804	0.1190
	512	0.0359	0.0050	0.0813	0.0851	0.1022
	1024	0.0653	0.0372	0.0905	0.1000	0.1014
0.25	256	0.1657	0.1311	0.2208	0.2209	0.2556
	512	0.2234	0.1756	0.2312	0.2275	0.2439
	1024	0.2063	0.1864	0.2404	0.2348	0.2489
0.4	256	0.3484	0.2723	0.3501	0.3488	0.3665
	512	0.3413	0.3167	0.3831	0.3617	0.3965
	1024	0.3557	0.3329	0.3890	0.3752	0.3968

6.3 Application to ARFIMA model with linear trend

We consider the model with linear trend

$$y = X\beta + \epsilon$$

where ϵ follows AFRIMA$(0, d, 0)$. For comparison, we consider the same estimation methods mentioned in the previous section.

We generated 100 simulated time series of length 256, 512 and 1024 for three different d values (0.1, 0.25, 0.4) respectively. Also, two different σ^2 $(0.5, 1)$ and β $(0.01, 0.001)$ are considered. The following table (7) represents the estimates of long memory parameter and the corresponding MSEs obtained by the methods explored in the previous section when $\sigma^2 = 1$ and $\beta = 0.001$. In order that we compare our method with other methods, we use the idea of Beran (1994). He fit a linear trend $g(x_t) = \beta_0 + \beta_1 t$ to the data and then a ARFIMA$(0, d, 0)$ model to the residuals. Our method estimates both the slope and the long memory parameter simultaneously while other methods estimate d after detrending the data with the help of OLS estimates. We used 5000 MCMC iterations and the first 2500 was used as burn-in.

Table 8: ARFIMA$(0, d, 0)$ with linear trend: MSE from each method$(\sigma^2 = 1, \beta = 0.001)$

d	n	GPH	SR	HR	fEXP	WB
0.1	256	0.0685	0.0505	0.0028	0.0355	0.0023
	512	0.0331	0.0311	0.0019	0.0075	0.0012
	1024	0.0178	0.0155	0.0006	0.0039	0.0006
0.25	256	0.0438	0.0368	0.0043	0.0217	0.0036
	512	0.0340	0.0280	0.0016	0.0075	0.0015
	1024	0.0267	0.0175	0.0007	0.0048	0.0007
0.4	256	0.0608	0.0526	0.0047	0.0240	0.0036
	512	0.0358	0.0276	0.0013	0.0159	0.0012
	1024	0.0293	0.0210	0.0009	0.0088	0.0008

The corresponding MSEs of the estimators of long memory parameter is given in table (8).

The following box-plot (figure 13) gives us more insight into the bias and MSE. As we can see, the estimators of long-memory parameter are getting closer and closer to the true d and the variances of those estimators are getting smaller and smaller as sample size increases.

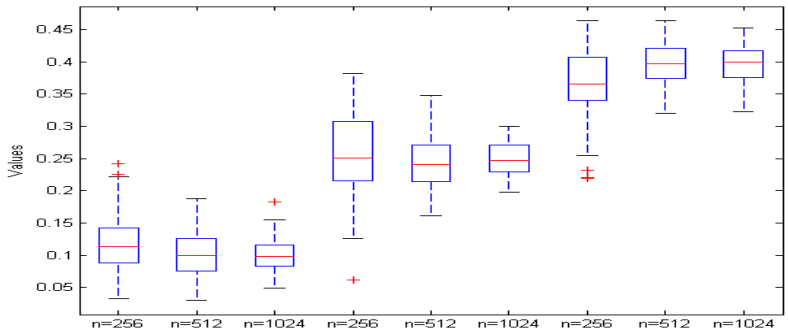

Figure 13: Boxplot of $\hat{d}$.

Three following tables present the estimates $\hat{d}$ and corresponding MSE for different values of σ^2 and β.

The table (9) presents the estimated value when $\sigma^2 = 0.5$ and $\beta = 0.001$.

Table 9: ARFIMA$(0, d, 0)$ with linear trend: The estimates $\hat{d}$ and MSE$(\sigma^2 = 0.5, \beta = 0.001)$

d	n	$\hat{d}/MSE$	$\hat{\sigma}^2/MSE(\sigma^2 = 0.5)$	$\hat{\beta}/MSE(\beta = 0.001)$
0.1	256	0.1266/0.0022	0.5878/0.0091	0.0010/2.9814$\times 10^{-6}$
	512	0.1110/0.0015	0.5433/0.0027	0.0009/3.6058$\times 10^{-7}$
	1024	0.1054/0.0008	0.5240/0.0012	0.0010/4.9228$\times 10^{-8}$
0.25	256	0.2435/0.0030	0.5797/0.0078	0.0058/4.8804$\times 10^{-5}$
	512	0.2504/0.0011	0.5401/0.0023	0.0008/9.2552$\times 10^{-7}$
	1024	0.2526/0.0006	0.5232/0.0009	0.0010/1.5215$\times 10^{-7}$
0.4	256	0.3811/0.0021	0.5780/0.0073	0.0015/1.5322$\times 10^{-5}$
	512	0.3894/0.0016	0.5365/0.0022	0.0008/2.9519$\times 10^{-6}$
	1024	0.3946/0.0007	0.5231/0.0009	0.0009/6.2392$\times 10^{-7}$

Table 10: ARFIMA$(0, d, 0)$ with linear trend: The estimates $\hat{d}$ and MSE$(\sigma^2 = 0.5, \beta = 0.01)$

d	n	$\hat{d}/MSE$	$\hat{\sigma}^2/MSE(\sigma^2 = 0.5)$	$\hat{\beta}/MSE(\beta = 0.01)$
0.1	256	0.1245/0.0027	0.5789/0.0076	0.0099/2.6058$\times 10^{-6}$
	512	0.1125/0.0012	0.5446/0.0030	0.0100/3.2935$\times 10^{-7}$
	1024	0.1067/0.0007	0.5255/0.0011	0.0100/5.2547$\times 10^{-8}$
0.25	256	0.2512/0.0029	0.5843/0.0087	0.0103/4.9333$\times 10^{-6}$
	512	0.2496/0.0013	0.5440/0.0027	0.0100/8.4964$\times 10^{-7}$
	1024	0.2443/0.0006	0.5247/0.0012	0.0100/3.2295$\times 10^{-7}$
0.4	256	0.3839/0.0028	0.5835/0.0084	0.0103/1.2890$\times 10^{-5}$
	512	0.3966/0.0012	0.5397/0.0024	0.0102/2.6690$\times 10^{-6}$
	1024	0.3968/0.0007	0.5218 /0.0010	0.0099 /8.3810$\times 10^{-7}$

The table (10) presents the estimated value when $\sigma^2 = 0.5$ and $\beta = 0.01$.

The table (11) presents the estimated value when $\sigma^2 = 1$ and $\beta = 0.001$.

6.3.1 Inference on other parameters

Since our method estimates innovation variance as well as slope other than long-memory parameter, we can see the results for two other parameters. The following table (12) shows us the estimates ($\hat{\sigma}^2,\ \hat{\beta}$) and the corresponding MSE's. 1 and 0.01 were considered as the true value for innovation variance and slope, respectively.

Table 11: ARFIMA$(0, d, 0)$ with linear trend: The estimates $\hat{d}$ and MSE$(\sigma^2 = 1, \beta = 0.001)$

d	n	$\hat{d}/MSE$	$\hat{\sigma}^2/MSE(\sigma^2 = 1)$	$\hat{\beta}/MSE(\beta = 0.001)$
0.1	256	0.1188/0.0022	1.0081/0.0067	0.0009/4.2312$\times 10^{-6}$
	512	0.1027/0.0015	1.0055/0.0036	0.0009/7.3720$\times 10^{-7}$
	1024	0.1028/0.0007	1.0019/0.0023	0.0011/1.2182$\times 10^{-7}$
0.25	256	0.2409/0.0042	0.9959/0.0071	0.0010/1.0543$\times 10^{-5}$
	512	0.2449/0.0016	1.0031/0.0033	0.0011/1.9395$\times 10^{-6}$
	1024	0.2422/0.0007	1.0035/0.0016	0.0009/3.4809$\times 10^{-7}$
0.4	256	0.3662/0.0041	0.9972/0.0062	0.0013/2.1019$\times 10^{-5}$
	512	0.3893/0.0014	1.0052/0.0036	0.0007/5.8666$\times 10^{-6}$
	1024	0.3968/0.0005	1.0032/0.0019	0.0011/1.4461$\times 10^{-6}$

Table 12: ARFIMA$(0, d, 0)$ with linear trend: The estimates $\hat{\sigma}^2, \hat{\beta}$ and corresponding MSE$(\sigma^2 = 1, \beta = 0.01)$

d	n	$\hat{\sigma}^2/MSE(\sigma^2 = 1)$	$\hat{\beta}/MSE(\beta = 0.01)$
0.1	256	1.0227/0.0082	0.0102/4.7843$\times 10^{-6}$
	512	1.0068/0.0034	0.0101/7.8048$\times 10^{-7}$
	1024	1.0032/0.0014	0.0100/1.4192$\times 10^{-7}$
0.25	256	1.0153/0.0048	0.0103/1.1568$\times 10^{-6}$
	512	1.0122/0.0034	0.0101/2.2790$\times 10^{-6}$
	1024	0.9986/0.0025	0.0100/3.9425$\times 10^{-7}$
0.4	256	1.0129/0.0065	0.0106/3.1462$\times 10^{-5}$
	512	0.9995/0.0033	0.0097/5.8004$\times 10^{-6}$
	1024	0.9966/0.0019	0.0099/9.8694$\times 10^{-7}$

6.4 Application to fMRI

Various applications of long memory process have been extended to many fields such as genetics and psychology. Here we apply our method to simulated functional magnetic resonance imaging (fMRI) data. The model we adopt here can be written as

$$y = X\beta + e, \tag{6.2}$$

where e is a long memory process with an innovation variance ${\sigma_L}^2$.

We generated fMRI data by convolving a square wave signal with a Poisson hemodynamic response function (HRF). A square wave signal can be defined as

$$x(t) = A \sum_{k=-\infty}^{\infty} g(t - kP), \tag{6.3}$$

where A and P are amplitude and fundamental period of the signal, respectively. The function $g(t)$ is defined as

$$g(t) = \begin{cases} 1, & 0 \leq t < P/2 \\ -1, & P/2 \leq t < P \\ 0, & ow \end{cases} . \tag{6.4}$$

Since the typical size of fMRI data is between 300 and 400, we considered three different sample sizes such as 128, 256 and 512 in this section. We consider $\sigma^2 = 1$, $\beta = 0.01$. For other parameters, three different $d(0.1, 0.25, 0.4)$ and SNR (0.5,5,10) are considered, respectively. The figure 14 shows a simulated fMRI signal which is generated by convolving a square wave signal with $N = 512$, $d = 0.1$ and a period of 16 with the poisson HRF with $\lambda = 4$. Parameters $(\beta,\ {\sigma_L}^2)$ is set to (0.01,1). The three different signal-to-noise ratio (SNR) are considered:

$$SNR = 10 \log \left(\frac{A}{{\sigma_L}^2} \right). \tag{6.5}$$

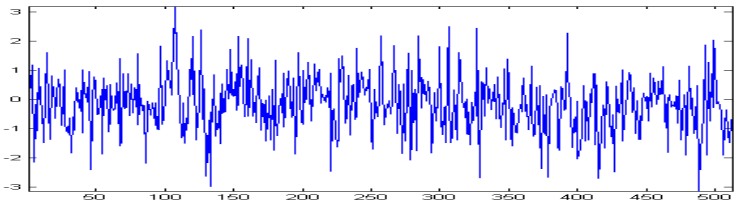

Figure 14: Simulated fMRI signals($d = 0.1$).

The following tables (table (13) to table (15)) represent estimated parameters and MSEs based on 100 replications.

Table 13: fMRI: The estimates $\hat{d}$ and MSE($SNR = 0.5$)

SNR=0.5, $\sigma^2 = 1,\ \beta = 0.01$				
d	n	$\hat{d}$/MSE	$\hat{\sigma^2}$/MSE	$\hat{\beta}$/MSE
0.1	128	0.1380/0.0052	1.0276/0.0107	0.0025/0.0101
	256	0.1151/0.0020	1.0288/0.0064	0.0063/0.0040
	512	0.1052/0.0014	1.0019/0.0031	0.0018/0.0019
0.25	128	0.2387/0.0064	0.9981/0.0087	0.0125/0.0110
	256	0.2582/0.0042	1.0280/0.0062	0.0026/0.0044
	512	0.2544/0.0016	1.0077/0.0037	-0.0032/0.0020
0.4	128	0.3463/0.0076	0.9921/0.0095	-0.0126/0.0100
	256	0.3822/0.0027	1.0074/0.0065	0.0016/0.0034
	512	0.4001/0.0017	1.0051/0.0042	0.0063/0.0020

The results for different parameter values are given in the following tables (table (13) and table (14)).

The following box plot (figure 15) includes results for 9 combinations of sample size and long memory parameter.(SNR=5)

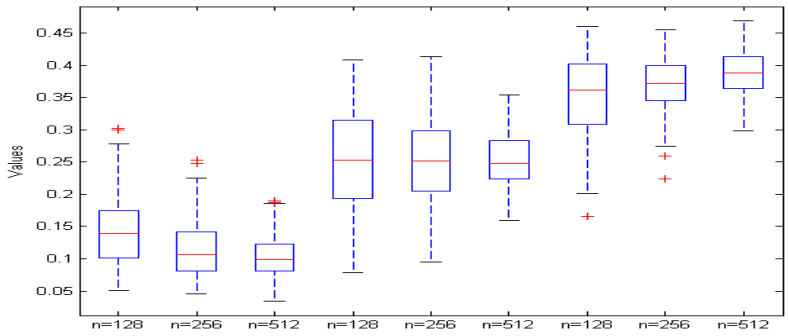

Figure 15: Box plot for SNR=5.

It is not surprising that the estimates are getting closer to the true value as sample size increases and that the variation of estimates is decreasing according to the increase of sample size.

Table 14: fMRI: The estimates $\hat{d}$ and MSE($SNR = 5$)

SNR=5, $\sigma^2 = 1,\ \beta = 0.01$				
d	n	$\hat{d}$/MSE	$\hat{\sigma^2}$/MSE	$\hat{\beta}$/MSE
0.1	128	0.1413/0.0050	1.0406/0.0137	0.0063/0.0098
	256	0.1146/0.0023	1.0208/0.0064	0.0005/9.3771$\times 10^{-5}$
	512	0.0989/0.0012	1.0154/0.0036	0.0085/0.0015
0.25	128	0.2579/0.0058	1.0304/0.0089	-0.0064/0.0087
	256	0.2546/0.0041	1.0211/0.0069	-0.0014/0.0018
	512	0.2552/0.0016	1.0045/0.0027	0.0047/0.0018
0.4	128	0.3574/0.0056	1.0240/0.0092	0.0043/0.0081
	256	0.3782/0.0023	1.0032/0.0048	0.0062/0.0046
	512	0.3923/0.0013	1.0078/0.0027	0.0058/0.0020

Table 15: fMRI: The estimates $\hat{d}$ and MSE($SNR = 10$)

SNR=10, $\sigma^2 = 1,\ \beta = 0.01$				
d	n	$\hat{d}$/MSE	$\hat{\sigma^2}$/MSE	$\hat{\beta}$/MSE
0.1	128	0.1453/0.0066	1.0299/0.0130	-0.0068/0.0061
	256	0.1192/0.0025	1.0143/0.0075	-0.0107/0.0030
	512	0.1054/0.0011	1.0096/0.0028	-0.0047/0.0022
0.25	128	0.2519/0.0068	1.0237/0.0110	-0.0080/0.0068
	256	0.2465/0.0033	1.0002/0.0060	-0.0043/0.0032
	512	0.2554/0.0012	1.0061/0.0035	0.0006/0.0018
0.4	128	0.3546/0.0067	1.0339/0.0117	-0.0115/0.0061
	256	0.3917/0.0026	1.0171/0.0065	-0.0011/0.0045
	512	0.3974/0.0011	0.9918/0.0037	0.0063/0.0018

CHAPTER VII

SUMMARY AND FUTURE STUDY

7.1 Summary

We have investigated a wavelet-based Bayesian parameter estimation method in the process which shows long range dependence. The method uses the decorrelation property of wavelet transform and introduces the Markov chain Monte Carlo method in the Bayesian framework for fast and efficient posterior inference. Simulation studies have shown that the method displays robustness against nonstationarity such as linear trend.

The commonly used models are ARFIMA$(0, d, 0)$ and fractional Gaussian model. For an illustrative example, we considered simple Gaussian ARFIMA$(0, d, 0)$ model to check the performance of several existing methods. In the analysis, four estimation methods are chosen for comparison. The performance of each method is getting better based on bias and MSE as sample size increases. Also, we noticed that maximum likelihood based estimator is stable in terms of MSE.

To check if the estimation methods behave well even in the case of nonstationarity, we considered ARFIMA model with linear trend. As we expected, wavelet-based Bayesian method is the most robust against nonstationarity such as linear trend. We can see that GPH estimator is not stable while the maximum likelihood-based estimator is relatively stable.

Since wavelet has many good properties, it is very widely used in various fields. In the estimation process, it is commonly used because of its robustness against nonstationarity and model misspecification. In addition to that, wavelet has been used as a shrinkage tool. For example, it has been used as a standard tool for noise removal. For this purpose, we used wavelet transform in the high throughput data, mass spectrometry data in which

there is heterogeneous noise variance. First of all, the local thresholding method detects variance change points and uses the points detected to get better estimate of threshold value in each segment. To check the performance of the method, we applied the local thresholding method to the real ovarian cancer data. For comparison, we also considered global thresholding method with different threshold values along with soft thresholding policy. The local thresholding method is quite helpful to detect peak or biomarker which has important biological meaning in proteomics.

7.2 Future study

The application of long memory parameter estimation is extended to various fields including functional magnetic resonance imaging(fMRI). We are going to apply wavelet-based Bayesian method to fMRI data to find some connection between brain activity and long memory parameter. Also, we will consider estimating other parameters as well as long memory parameter simultaneously in the more complicated setup, for example, nonstationary ARFIMA (p, d, q).

REFERENCES

Abry, P. and Veitch, D. (1998). "wavelet analysis of long-range dependent traffic". *IEEE Transactions on Information Theory* **44**, 2–15.

Beran, J. (1994). *Statistics for long-memory processes*. London: Chapman & Hall.

Bickel, P. (1983). *Minimax estimation of the mean of a normal distribution subjectto doing well at the point*. New York: Academic Press.

Bruce, A. G., Donoho, D. L., Gao, H. Y., and Martin, R. D. (1994). Smoothing and robust wavelet analysis.. In COMPSTAT. Proceedings in Computational Statistics, 11th Symposium, pages 531-547.

Dalhaus, R. (1989). "efficient parameter estimation for self-similar processes". *Annals of Statistics* **17**, 1749–1766.

Dijkerman, R. W. and Mazumdar, R. R. (1994). "on the correlation structure of the wavelet coefficients of fractional brownian motion". *IEEE Transactions on Information Theory* **40**, 1609–1612.

Donoho, D. L. and Johnstone, I. M. (1994). "ideal spatial adaptation by wavelet shrinkage". *Biometrika* **81**, 425–455.

Fox, R. and Taqqu, M. S. (1985). "non-central limit theorems for quadratic forms in random variables having long-range dependence". *Ann. Probability* **13**, 428–446.

Fox, R. and Taqqu, M. S. (1986). "large sample properties of parameter estimates for strongly dependent stationary gaussian time series". *Annals of statistics* **14**, 517–532.

Frias, M., Alonso, F., Ruiz-Medina, M., and Angulo, J. (2008). "semiparametric estimation of spatial long-range dependence". *Journal of Statistical Planning and Inference* **138**, 1479–1495.

Gabbanini, F., Vannucci, M., Bartoli, G., and Moro, A. (2004). "wavelet packet methods for the analysis of variance of time-series with application to crack widths on the brunelleschi dome". *Journal of Computational and Graphical Statistics* **13**, 639–658.

Gao, H.-Y. and Bruce, A. G. (1996). "waveshrink with firm shrinkage". *Technical Report* **39**. StatSci Division of MathSoft, Inc.

Geweke, J. and Porter-Hudak, S. (1983). "the estimation and application of long memory time series models". *Journal of Time Series Analysis* **4**, 221–237.

Granger, C. and Joyeux, R. (1980). "an introduction to long-range time series models and fractional differencing". *Journal of Time Series Analysis* **1**, 15–30.

Hall, P., Hardle, W., Kleinow, T., and Schmidt, P. (2000). "semiparametric bootstrap approach to hypothesis tests and confidence intervals for the hurst coefficient". *Statistical Inference for Stochastic Processes* **3**, 263–276.

Haslett, J. and Raftery, A. (1989). "space-time modeling with long-memory dependence: Assessing ireland's wind power resource". *Journal of Applied Statistics* **38**, 1–50.

Hastings, W. K. (1970). "monte carlo sampling methods using markov chains and their applications". *Biometrika* **57**, 97–109.

Hosking, R. (1981). "fractional differencing". *Biometrika* **68**, 165–176.

Hurst, H. (1951). "long-term storage capacity of reservoirs". *Trans American Society Civil Engineers* **116**, 770–799.

Inclán, C. and Tiao, G. (1994). ”use of cumulative sums of squares for retrospective detection of changes of variance”. *Journal of the American Statistical Association* **89**, 913–923.

Ko, K. and Vannucci, M. (2006). ”bayesian wavelet anaysis of autoregressive fractionally integrated moving average processes”. *Journal of Statistical Planning and Inference* **136**, 3415–3434.

Liu, J. S. (1996). ”peskun’s theorem and a modified discrete-state gibbs sampler”. *Biometrika* **83**, 681–682.

Liu, J. S. (2004). *Monte Carlo strategies in scientific computing*. New York: Springer.

Lo, A. W. (1991). ”long-term memory in stock market prices”. *Econometrica* **59**, 1279–1313.

Mallat, S. G. (1989). ”multiresolution approximations and wavelet orthonormal bases of $l^2(r)$”. *Trans. Amer. Math. Society* **315**, 69–87.

McLeod, A. I. and Hipel, K. W. (1978). ”preservation of the rescaled adjusted range”. *Water Resources Res.* **14**, 491–518.

Mielniczuk, J. and Wojdyllo, P. (2007). ”estimation of hurst exponent revisited”. *Computational Statistics and Data Analysis* **51**, 4510–4525.

Moore, L. E., Fung, E. T., McGuire, M., Rabkin, C. C., Molinaro, A., Wang, Z., Zhang, F., Wang, J., Yip, C., Meng, X. Y., and Pfeiffer, R. M. (2006). ”evaluation of apolipoprotein a1 and post-translationally modified forms of transthyretin as biomarkers for ovarian cancer detection inan independent study population”. *Cancer Epidemiology Biomarkers and Prevention* **15**, 1641–1646.

Morris, J. S., Coombes, K. R., Kooman, J., Baggerly, K. A., and Kobayashi, R. (2005). "feature extraction and quantification for mass spectrometry data in biomedical applications using the mean spectrum". *Biometrics* **21**, 1764–1775.

Nason, G. P. (1996). "wavelet shrinkage by cross-validation". *Journal of Royal Statistical Society Series B* **58**, 463–479.

Percival, D. and Walden, A. (1999). *Wavelet methods for time series analysis*. London: Cambridge University Press.

Peters, E. (1994). *Fractal market analysis*. New York: Wiley.

Pickands, J. (1967). "maxima of stationary gaussian process". *Wahrscheinlichkeitstheorie verw. Geb.* **7**, 190–223.

Robert, C. P. and Casella, G. (1999). *Monte Carlo statistical methods*. New York: Springer.

Roberts, G. O. and Gilks, W. R. (1994). "convergence of adaptive direction sampling". *Journal of Multivariate Analysis* **49**, 287–298.

Roberts, G. O. and Sahu, S. K. (1997). "updating scheme; correlation structure; blocking and parameterization for the gibbs sampler". *Journal of the Royal Statistical Society, Series B* **59**, 291–317.

Robinson, P. (1995). "gaussian semiparametric estimation of long range dependence". *The Annals of Statistics* **23**, 1630–1661.

Shimotsu, K. and Phillips, P. C. (2006). "local whittle estimation of fractional integration and some of its variants". *Journal of Econometrics* **130**, 209–233.

Stoev, S. and Taqqu, M. S. (2005). "asymptotic self-similarity and wavelet estimation for long-range dependent fractional autoregressive integrated moving average time series with stable innovations". *Journal of Time Series Analysis* **26**, 211–249.

Taqqu, M. S. and Teverovsky, V. (1997). "robustness of whittle-type estimators for time series with long-range dependence". *Stochastic Models* **13**, 723–757.

Taqqu, M. S., Teverovsky, V., and Willinger, W. (1995). "estimators for long-range dependence: An empirical study". *Fractals* **3**, 785–798.

Twefik, A. H. and Kim, M. (1992). "correlation structure of the discrete wavelet coefficients of fractional brownian motion". *IEEE Transactions on Information Theory* **38**, 904–909.

Vannucci, M. and Corradi, F. (1999). "covariance structure of wavelet coefficients: Theory and models in a bayesian perspective". *Journal of Royal Statistical Society Series B* **61**, 971–986.

Vidakovic, B. (1999). *Statistical modeling by wavelets*. New York: Wiley Series.

Whitcher, B., Guttorp, P., and Percival, D. (2000). "multiscale detection and location of multiple variance changes in the presence of long memory". *Journal of Statistical Computation and Simulation* **68**, 65–88.

Wickerhauser, M. V. (1994). *Adaptive wavelet analysis from theory to software algorithms*. Wellesley, MA: AK Peters.

Wong, J. W., Cagney, G., and Cartwright, H. M. (2005). "specaligh-processing and alignment of mass spectra datasets". *Biometrics* **21**, 2088–2090.

Yajima, Y. (1985). "on estimation of long-memory time series models". *Australia Journal of Statistics* **27**, 303–320.

Zygmund, A. (1953). *Trigonometrics series*. London: Cambridge Unviersity Press.

VITA

Jae Sik Jeong, son of Chang-Gyun Jeong and Jeong-Sim Kang, was born in Haenam, Korea. He received a Bachelor of Science degree in computer science and statistics from University of Seoul in Seoul, Korea in 1999. He received a Master of Science degree in statistics from Seoul National University under the direction of Dr. Byeonguk Park in 2001. He continued his studies in statistics under the direction of Dr. Marina Vannucci and received a Doctor of Philosophy degree in statistics from Texas A&M University in College Station, Texas, in August 2008. His address is Department of Statistics, Texas A& M University, 3143 TAMU, College Station, TX 77843.