Regularized Autoregressive Approximation in Time Series

by

Bei Chen

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Master of Mathematics
in
Statistics

Waterloo, Ontario, Canada, 2008

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Bei Chen

Abstract

In applications, the true underlying model of an observed time series is typically unknown or has a complicated structure. A common approach is to approximate the true model by autoregressive (AR) equation whose orders are chosen by information criterions such as AIC, BIC and Parsen's CAT and whose parameters are estimated by the least square (LS), the Yule Walker (YW) or other methods. However, as sample size increases, it often implies that the model order has to be refined and the parameters need to be recalculated. In order to avoid such shortcomings, we propose the Regularized AR (RAR) approximation and illustrate its applications in frequency detection and long memory process forecasting. The idea of the RAR approximation is to utilize a "long" AR model whose order significantly exceeds the model order suggested by information criterions, and to estimate AR parameters by Regularized LS (RLS) method, which enables to estimate AR parameters with different level of accuracy and the number of estimated parameters can grow linearly with the sample size. Therefore, the repeated model selection and parameter estimation are avoided as the observed sample increases.

We apply the RAR approach to estimate the unknown frequencies in periodic processes by approximating their generalized spectral densities, which significantly reduces the computational burden and improves accuracy of estimates. Our theoretical findings indicate that the RAR estimates of unknown frequency are strongly consistent and normally distributed. In practice, we may encounter spurious frequency estimates due to the high model order. Therefore, we further propose the robust trimming algorithm (RTA) of RAR frequency estimation. Our simulation studies indicate that the RTA can effectively eliminate the spurious roots and outliers, and therefore noticeably increase the accuracy. Another application we discuss in this thesis is modeling and forecasting of long memory processes using the RAR approximation. We demonstration that the RAR is useful in long-range prediction of general ARFIMA(p, d, q) processes with $p \geq 1$ and $q \geq 1$ via simulation studies.

Acknowledgements

First of all, I would like to express my sincere gratitude and appreciation to my supervisor, Dr. Yulia R. Gel, for her encouragement, guidance, and unlimited support in every aspect of my study, research and personal life. Her rigorous scholarship and enthusiasm in statistics have inspired and enriched my growth as a student and a researcher. I am indebted to her more than I can express in words. Many thanks in particular to her family for their constant help.

My gratitude also extends to Dr. Bovas Abraham and Dr. Mary Thompson for their precious time to read my thesis and invaluable revising advices. Moreover, I would like to thank all my instructors because of whom I developed my foundation and interest in statistics in the first place.

I am thankful to the graduate studies coordinator Mary Lou Dufton for her patience and valuable advises, to my colleague Li Wang for the Latex help, and to my friend Arpit Kumar and his family for the encouragement and support along the way.

Finally, special thanks to my parents. Without you, nothing is possible.

To the memory of my grandmother

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Wikipedia defines *time series* as "a sequence of data points, measured typically at successive time, spaced at (often uniform) time intervals". *Time series* may be *continuous* or *discrete*. *Continuous time series* are recorded instantaneously and steadily; while *discrete time series* are taken at regular time intervals. In this thesis, we focus on *discrete time series* observed at equal intervals.

1.1 Examples of Time Series

Time series is widely applied in many fields, such as:

- Economics (e.g., Unemployment Rate, GDP, NNP, Interest rate, inflation rate)
- Business (e.g., Inventory, Sales, Quality Indices, Stock Price)
- Ecology (e.g., Air Pollution, Water Pollution, Wildlife Population)
- Astronomy (e.g., Solar Activity, Sun Spots, Star Brightness)
- Meteorology (e.g., Rainfall, Humidity, Wind Speed)
- Sociology (e.g., Crime Rates, Divorce Rates)

In fact, examples of time series can be found everywhere in our university life: a weekly sequence of seminar attendance in the Math Faculty, a monthly series of blood donors in the Student Life Center, a yearly set of employment rates of the co-op program at the University of Waterloo, etc. Figure 1.1 shows the time series plot of real daily wages in pounds in England from 1260 to 1994.

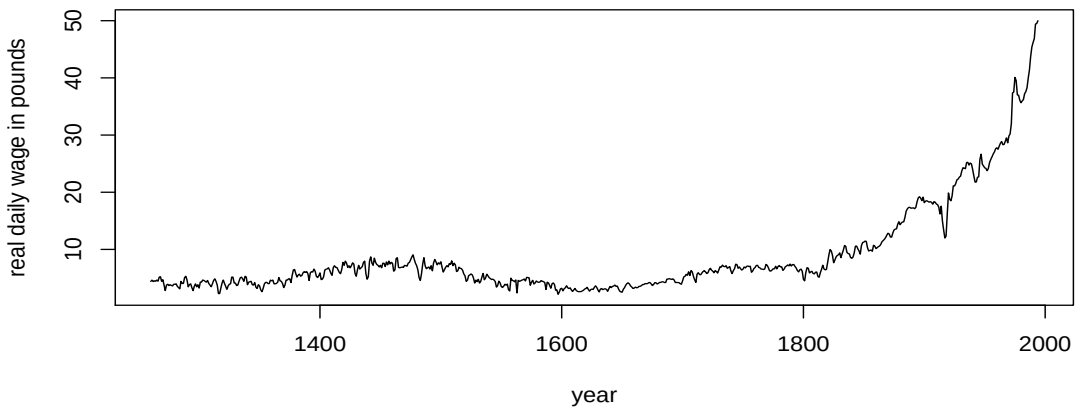

Figure 1.1: Real daily wages in pounds, England. 1260 - 1994

1.2 Thesis Introduction

In applications, the true underlying model of an observed time series is typically unknown or has a complicated structure. A common approach is to approximate the true model of the observed process by an autoregressive (AR) equation. The AR model has a simple polynomial form while its parameter estimation techniques and corresponding asymptotic properties are well-investigated in statistical literature (for example, see Anderson, 1971; Brockwell and Davis, 1987; Box et al., 1994). Therefore, AR models are very attractive for various applications that involve adaptive (online) estimation and forecasting. The main focus of this thesis is to study the statistical properties of Regularized version of AR

approximation and its applications to frequency estimation and modeling of long memory processes.

A common procedure of AR approximation is to select the model order by information criterions such as AIC, BIC and Parzen's CAT, and then to estimate model coefficients by the Least Square (LS), the Yule-Walker (YW) or other methods. However, in practice, the length of observed sample is frequently unknown a priori and may indefinitely increase while the estimation is performed in real time (or online). Hence, the model order has to be refined upon the arrival of every new observation and, thus, all the earlier estimated model parameters need to be recalculated from scratch, which eventually increases the computational costs.

In order to avoid such shortcomings, we consider an alternative approach, called regularized AR (RAR) approximation (Gel and Fomin, 2001; Gel and Barabanov, 2007). The idea is to utilize a "long" AR model whose order significantly exceeds the model order suggested by information criterions, and to estimate AR parameters using Regularized LS (RLS) method. The RLS estimation has a recursive form of the stationary Kalman filter with a regularizer added to the sample information matrix, which enables to estimate AR parameters with different level of accuracy, i.e. the first few model parameters are estimated more precisely than the tail ones and the number of estimated parameters can grow linearly with the sample size. Therefore, the repeated model selection and parameter estimation are avoided as the observed sample increases. The regularizer may be viewed as the smoothing operator applied to the number of the AR coefficients being estimated, and constitutes a link to the model selection criterion. We choose the regularizer by cross validation, and our procedure is similar to the approach of thresholding selection proposed by Bickel and Levina (2007), but in a time series context. The theoretical results indicate that the RLS estimates of an AR(∞) model with exponentially decaying coefficients converge almost surely at a power law rate (Gel and Barabanov, 2007). In this thesis we extend application of the RAR to periodic processes and long memory time series as discussed below.

Many real life time series can be modeled as a sum of sinusoids and noise, and the practical problem of interest lies in detecting the frequency hidden in the sinusoid. One of the most popular approaches is to approximate the generalized spectral density of a

periodic process by an AR model, due to its simple structure and well-studied estimation properties (Tuft and Kumaresan, 1982; Mackisack and Poskitt, 1989, 1990; Li et al., 1994; Hannan and Quinn, 2001). Such an AR approximation can be conducted in three steps. In first, the order of the AR model is selected by information criterias such as AIC or BIC. In second, the model parameters are obtained by the Least Squares (LS), the Yule-Walker or other methods. Finally, the hidden frequency is estimated by determining the zeros of the auxiliary polynomial of the AR model, and then taking the argument of the zero with modulus closest to the unit circle. However, in many electrical engineering, astronomical and biomedical applications, the length of the observations is unknown a priori and therefore frequency detection has to be performed online. In order to avoid the repeated model order selection and parameter estimation, we extend the existing AR approximations and apply the RAR method to estimate the frequency, which significantly reduce the computational burden and improve accuracy of estimates. Our theoretical findings indicate that the RAR estimates of unknown frequency are strongly consistent, i.e. converge almost surely, and asymptotically normally distributed.

In practice, we may encounter spurious frequency estimates when the model order is high (Stoica et al., 1987). Some of the spurious roots of the auxiliary polynomial may have sufficiently large moduli to be mistakenly considered as the true estimates. In order to increase the accuracy of the estimates, we further propose the robust trimming algorithm of RAR frequency estimation, which can be conducted as follows. Firstly, we take a subsample from the observed data sample as a training set (usually the first 1/3 of the sample) and apply the RAR frequency estimation to the training set using different regularizing parameters. Then, we construct a confidence interval (CI) of estimated frequencies based on different regularizers, as well as perform cross validation for selecting an optimal regularizer. Finally, we apply the RAR frequency estimation to the entire observed data sample using the optimal regularizing parameters but only taking into account the frequency estimates falling within the pre-determined CI. Our simulation studies indicate that such a robust trimming of frequency estimates can effectively eliminate the spurious roots and outliers, and thus noticeably increase the accuracy of frequency estimates.

Another application we discuss in this thesis is modeling and forecasting of long memory processes using RAR. Long memory time series are characterized by the property that

dependence between distant observations is small but non-negligible and decays polynomially. Hence, although the long memory process is still weakly stationary, its sum of serial correlations diverges, and a more sophisticated modeling techniques are hence required. Many different approaches are proposed to modeling of long range dependent (LRD) time series (Beran, 1994; Bailie, 1996; Engle, 1995; Nelson, 1991). However, since this thesis is mainly devoted to linear stochastic models, we particularly focus on a linear class of models for LRD, namely autoregressive fractionally integrated moving average (ARFIMA). Forecasting from an ARFIMA model includes parameter estimation based on non-linear optimization and a subsequent representation as a truncated AR model, which leads to high computational burden. Alternatively one can approximate the ARFIMA model by an AR model from the very beginning and, thus, skip the non-linear optimization parameter estimation step. The LRD property indicates that even far apart observations are somewhat correlated, hence, the approximating AR model is usually to be of high order. Most existing literature address modeling the subcases of ARFIMA(p, d, q), i.e. FI$(0, d, 0)$, ARFI$(1, d, 0)$ and FIMA$(0, d, 1)$, by "long" AR models (Ray, 1990; Ray, 1993; Ray and Crato; 1996, Poskitt, 2006). We extend these results to the general ARFIMA(p, d, q) processes with $p \geq 1$ and $q \geq 1$, and our simulation study prove the RAR approximation is useful in long-range prediction.

1.3 Main Contributions

The main contributions of this thesis include:

1. application of the Regularized AR approximation to estimation of unknown frequencies in periodic processes and, hence, extension of the results of Mackisack and Poskitt (1991), i.e. in terms of the increasing maximum possible order for AR approximation, and the results of Gel and Barabanov (2007), i.e. in terms of weakening the restriction on the decay of AR coefficients;

2. derivation of the almost sure convergence of RAR frequencies estimates to the true unknown frequencies;

3. investigation of asymptotic distributional properties of RAR frequency estimates, i.e. show that RAR estimates are asymptotically normally distributed and the corresponding variance-covariance matrix is obtained.

4. a new robust trimming algorithm to eliminate spurious roots and outliers, which noticeably increase the accuracy of the frequency estimates for processes with a low signal-to-noise ratio.

5. application of the Regularized AR approximation to general ARFIMA(p, d, q) processes with $p \geq 1$ and $q \geq 1$ and demonstration that RAR is useful in long-range prediction via simulation studies.

1.4 Thesis Outline

This thesis is organized as follows. In Chapter 2, we introduce some fundamental definitions and concepts in time series analysis that are essential for later discussions, such as autocovariance, stationarity, invertibility and spectral density. Also, we take an overview of the basic linear time series models, including the AR, the MA, and the ARMA, etc. At the end of chapter, we discuss some of the most commonly applied model order selection criterions and parameter estimation techniques.

In Chapter 3, we introduce the RAR approximation which fits an "long" AR model to the process and estimates the AR parameters using the RLS method (Gel and Barabanov, 2007). We demonstrate the properties of such RAR approximation by its application in ARMA model identification. Moreover, we discuss the almost sure convergence of the RLS estimates of an AR(∞) model with exponentially decaying coefficients. The theoretical results of this chapter serve as the motivation and foundation for the later chapters.

In Chapter 4, we apply the RAR method to approximate the generalized spectral density of a sinusoidal process in order to detect the hidden frequency. The strong consistency and asymptotic normality are proved for the RAR frequency estimates. Moreover, we propose a robust trimming algorithm of RAR frequency estimation due to the spurious roots and outliers that we encountered in the simulation studies. We show that the new robust trimming algorithm noticeably increase the accuracy of the estimates, especially for cases

with a low signal-to-noise ratio.

In Chapter 5, we briefly review some of the existing methods for the ARFIMA estimation and discuss an alternative technique, in particular, modeling the ARFIMA process by "long" AR models to approximate its long range dependence structure. We demonstrate that such AR approximation is useful in long-range prediction by simulation studies.

We summarized the main results in Chapter 6 and conclude the thesis with a a outline of future work.

Chapter 2

Linear Stochastic Models

In this chapter, we introduce some fundamental concepts in linear time series analysis that are essential for our later discussions. We begin with an excursus to stochastic processes and concepts of stationarity. After defining autocovariance, autocorrelation and partial autocorrelation functions, we discuss elements of spectral analysis and take an overview of the most commonly used linear models, including AR, MA, ARMA and ARIMA. Then, we proceed with discussion of model order selection and parameter estimation techniques.

2.1 Stochastic Process and Stationarity

2.1.1 Stochastic Process

A stochastic process is a statistical phenomenon that evolves in time according to probabilistic law, and the time series to be analyzed can be considered as a realization of such a stochastic process. Formally, we define the two concepts as follows.

Definition 2.1.1 *(Stochastic Process). Suppose T is an index set. A stochastic process is a family of random variables $\{X_t(\omega), t \in T, \omega \in \Omega\}$ defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$.*

In time series analysis, the index set T is a set of time points. Usually, $T = \mathbb{R}$, $\mathbb{R}^+$, $\mathbb{Z}$ or $\mathbb{Z}^+$. In this thesis, we consider $T = \mathbb{Z}$, where $\mathbb{Z} = \{0, \pm 1, \pm 2, ...\}$. Since X_t is a random vari-

able, for each fixed $t \in T$, X_t is a function $X_t(\cdot)$ on Ω; while for each fixed $\omega \in \Omega$, $X_\cdot(\omega)$ is a function on T.

Definition 2.1.2 *(Realization of a Stochastic Process). The function $X_t(\cdot)$ for fixed $\omega \in \Omega$ and $t \in T$ is called a realization of the stochastic process $\{X_t(\omega), t \in T, \omega \in \Omega\}$.*

A realization $\{X_t, t \in T_0\}$ for $T_0 \subseteq T$ can be regarded as a sample from an infinite population $\{X_t, t \in T\}$. In practice, we make inference about $\{X_t, t \in T\}$ based only on a single observed sample.

2.1.2 Stationarity

In time series modeling, a concept of stationarity plays an important practical and theoretical role. A (weakly) stationary process remains in equilibrium about a constant mean level, while non-stationary processes may have nonconstant mean, time varying variability or both of these properties. Thus, before deciding which model approach to take, one needs to check stationarity of the observed process. In this section, we define concepts of strict and weak stationarity.

Definition 2.1.3 *(Strict Stationarity). The time series $\{X_t, t \in \mathbb{Z}\}$ is said to be strictly stationary if $(X_{t_1}, ..., X_{t_k})$ and $(X_{t_1+h}, ..., X_{t_k+h})$ have the same joint distribution for all positive integers k and for all $t_1, ..., t_k, h \in \mathbb{Z}$.*

Strict Stationarity requires the joint distributions of any subset of $\{X_t, t \in \mathbb{Z}\}$ do not change with time. Intuitively, the plots over two equal-length time intervals of a observed time series need to display similar statistical features.

A very important quantity in time series analysis is the autocovariance function (ACVF), which measures the dependence between different observations of process $\{X_t, t \in \mathbb{Z}\}$.

Definition 2.1.4 *(ACVF). If $\{X_t, t \in \mathbb{Z}\}$ is a process such that $Var(X_t) < \infty$ for each $t \in \mathbb{Z}$, then the ACVF $\gamma(\cdot, \cdot)$ of $\{X_t\}$ is defined by*

$$\gamma(r, s) = Cov(X_r, X_s) = E[(X_r - EX_r)(X_s - EX_s)], r, s \in \mathbb{Z}. \tag{2.1}$$

The definition of weak stationarity of a time series is based on the mean and ACVF of a stochastic process.

Definition 2.1.5 *(Weak Stationarity). The time series* $\{X_t, t \in \mathbb{Z}\}$ *is said to be weakly stationary if*

1. $EX_t^2 < \infty$, *for all* $t \in \mathbb{Z}$,
2. $EX_t = \mu < \infty$, *for all* $t \in \mathbb{Z}$, *and*
3. $\gamma_X(r, s) = \gamma_X(r + t, s + t)$, *for all* $r, s, t \in \mathbb{Z}$.

Strict stationarity, together with the assumption on finite first and second moments, implies weak stationarity. However, the converse is not true in general, except for Gaussian process. All the joint distributions of a Gaussian process are multivariate Gaussian, so weak stationarity is equivalent to strict stationarity. In this thesis, we assume weak stationarity for all stationary processes, unless specified otherwise.

If $\{X_t, t \in \mathbb{Z}\}$ is stationary, then $\gamma_X(r, s) = \gamma_X(r-s, 0)$, for $r, s \in \mathbb{Z}$. Usually, we reduce the ACVF to a one-variable form if $\{X_t\}$ is stationary. For all $t, k \in \mathbb{Z}$, the ACVF at lag k is defined as

$$\gamma_k = Cov(X_{t+k}, X_t) \tag{2.2}$$

and the autocorrelation (ACF) at lag k is defined as

$$\rho_k = Corr(X_t, X_{t+k}) = \frac{Cov(X_t, X_{t+k})}{\sqrt{Var(X_t)Var(X_{t+k})}} = \frac{\gamma_k}{\gamma_0} \tag{2.3}$$

In addition to the autocorrelation between X_t and X_{t+k}, it is equally important to investigate the correlation between X_t and X_{t+k} after removing their mutual linear dependency on the intervening variables $X_{t+1}, X_{t+2}, ..., X_{t+k-1}$, which is indicated by the partial ACF (PACF): $Corr(X_t, X_{t+k}|X_{t+1}, X_{t+2}, ..., X_{t+k-1})$. The idea is presented precisely in the following definition.

Definition 2.1.6 *(PACF). The PACF* $\alpha_\cdot$ *of a stationary time series* $\{X_t, t \in \mathbb{Z}\}$ *is defined by*

$$\alpha_1 = Corr(X_1, X_2) = \rho_1 \tag{2.4}$$

and

$$\alpha_k = Corr(X_{k+1} - P_{\overline{sp}\{1,X_2,\ldots,X_k\}}X_{k+1}, X_1 - P_{\overline{sp}\{1,X_2,\ldots,X_k\}}X_1),\ k \geq 2, \tag{2.5}$$

where

- $\overline{sp}\{1, X_2, \ldots, X_k\}$ *is the closed span of any subset* $\{1, X_2, \ldots, X_k\}$ *of a Hilbert space,*
- *define* $S = \{0, 2, \ldots, k\}$, $X_0 = 1$. $P_{\overline{SP}\{1,X_2,\ldots,X_k\}}X_{k+1} = \sum_{i\in S} a_i X_i$, *where* $a_0, a_2, \ldots, a_k$ *satisfy* $\langle \sum_{i\in S} a_i X_i, X_j\rangle = \langle X_{k+1}, X_j\rangle, j \in S$, *and* $P_{\overline{SP}\{1,X_2,\ldots,X_k\}}X_1 = \sum_{i\in S} b_i X_i$, *where* $b_0, b_2, \ldots, b_k$ *satisfy* $\langle \sum_{i\in S} b_i X_i, X_j\rangle = \langle X_1, X_j\rangle, j \in S$.

The theoretical values of ACVF, ACF and PACF can be calculated exactly if all possible realizations of a stochastic process are known. Otherwise, they can be estimated if multiple independent realizations are available. However, in most applications, we only have one realization of the process to make inference from. Therefore, we analyze their sample quantities in practice. The formulae used to calculate sample ACVF, ACF and PACF will be stated in later chapters.

2.1.3 White Noise Processes

White noise is a basic building block of more complicated time series models. A process $\{\epsilon_t\}$ is defined as white noise if it is stationary, and the ϵ_t's are mutually uncorrelated, for all $t \in \mathbb{Z}$. We usually assume $\{\epsilon_t\}$ has zero mean.

Definition 2.1.7 *(White Noise). The process* $\{\epsilon_t\}$ *is said to be white noise with mean 0 and variance* σ^2, *written*

$$\{\epsilon_t\} \sim WN(0, \sigma^2) \tag{2.6}$$

if and only if $\{\epsilon_t\}$ *has zero mean and ACF*

$$\gamma_k = \begin{cases} \sigma^2 & k = 0 \\ 0 & k \neq 0 \end{cases} \tag{2.7}$$

A white noise is a Gaussian process if its joint distribution is normal. White noise is not linearly forecastable in the sense that the best linear forecast of ϵ_{t+1} based on $\epsilon_t, \epsilon_{t-1}, \ldots$ is zero and does not depend on the present and past observations.

2.2 Spectral Density

The analysis of stationary time series $\{X_t, t \in \mathbb{Z}\}$ by means of its spectral representation is often referred as frequency domain analysis. Frequency or spectral methods for time series are especially popular in physics and engineering communities. Frequency domain can be considered as a dual space to time domain, and all concepts from one domain have a counterpart in another domain. Correspondingly, there is a spectral representation of ACVF of $\{X_t\}$, in terms of spectral density. Therefore, the frequency domain analysis of $\{X_t\}$ based on the spectral density is equivalent to the time domain analysis based on ACVF. Thus, typically a choice of frequency or time domain modeling techniques is subjective to a data analyst and a particular application.

Suppose that $\{\gamma_h\}$ is the ACVF sequence of the stationary process $\{X_t\}$, then $\{\gamma_h\}$ is nonnegative definite due to the following argument. Without loss of generality, assume $\{X_t\}$ has constant mean zero. Then for any sequence of constants $a_1, ..., a_n$

$$0 \leq Var(\sum_{h=1}^{n} a_h X_{t-h}) = E[\sum_{h=1}^{n} a_h X_{t-h}]^2 = E[\sum_{h=1}^{n} a_h X_{t-h} \sum_{s=1}^{n} a_s X_{t-s}] = \sum_{h,s=1}^{n} a_h \gamma_{h-s} a_s.$$

Since $\{\gamma_h\}$ is nonnegative definite, the Herglotz Theorem guarantees the existence of a right-continuous, non-decreasing spectral distribution function $F(\omega)$ defined on $[-\pi, \pi]$, such that $\{\gamma_h\}$ is the Fourier transform of the measure corresponding to F:

$$\gamma_k = \int_{-\pi}^{\pi} e^{ik\omega} \, dF(\omega). \tag{2.8}$$

Definition 2.2.1 *(Spectral Density). If $F(\omega)$ is everywhere continuous and differentiable, with $F'(\omega) = f(\omega)$, then f is called spectral density and*

$$\gamma_k = \int_{-\pi}^{\pi} e^{ik\omega} f(\omega) \, d\omega. \tag{2.9}$$

If $\sum_{h=-\infty}^{\infty} |\gamma_h| < \infty$, the above Fourier transform can be inverted to give an explicit form for spectral density:

$$f(\lambda) = \frac{1}{2\pi} \sum_{k=-\infty}^{\infty} \gamma_k e^{-ik\omega}. \tag{2.10}$$

2.3 Linear Models

In this section, we discuss a set of most commonly utilized linear models such as AR, MA, ARMA and ARIMA. We begin with introducing the backward shift operator B and the backward difference operator ∇, which appear extensively in the model definitions.

Definition 2.3.1 *(Backward Shift Operator). The backward shift operator B is defined as*

$$B^k X_t = X_{t-k},\ k \in \mathbb{Z}. \tag{2.11}$$

In particular, $IX_t = B^0 X_t = X_t$.

Definition 2.3.2 *(Backward Difference Operator). The backward difference operator* ∇ *is defined as*

$$\nabla^k X_t = (1-B)^k X_t,\ k \in \mathbb{Z}. \tag{2.12}$$

In particular, $\nabla^0 X_t = X_t$.

2.3.1 Stationary Models

Stationary models assume that the process remains in equilibrium about a constant mean level. The AR and MA models are the two most fundamental stationary models in time series analysis. The AR model is simply a linear regression of the current value of the process against the previous values of the process, i.e, the variable X_t is regressed on the past values of itself.

Definition 2.3.3 *(AR Processes). An AR(p) process,* $p \in \mathbb{Z}$ *, is defined as*

$$\phi(B) X_t = \epsilon_t \tag{2.13}$$

where

- $\{\epsilon_t\} \sim WN(0, \sigma^2)$,

- $\phi(\lambda) = 1 - \phi_1\lambda - \ldots - \phi_p\lambda^p$ *is a polynomial in* B *of degree* p.

The ACF of a stationary AR(p) process tails off as a mixture of exponential decays, and the PACF vanishes after lag p. Figure 2.1 shows a time series of length 1000 generated by an AR(2) process $X_t = 0.3X_{t-1} + 0.2X_{t-2} + \epsilon_t$, and corresponding sample ACF and PACF. Due to its simplicity, the AR models are very attractive in various applications. In particular, the main focus of this thesis is to use the AR model to approximate different types of processes.

Theoretically, the MA model is a linear regression of current value of the series against the random noise of the past values of the series.

Definition 2.3.4 *(MA Processes). A MA(q) process,* $q \in \mathbb{Z}$*, is defined as*

$$X_t = \theta(B)\epsilon_t, \tag{2.14}$$

where

- $\{\epsilon_t\} \sim WN(0, \sigma^2)$,
- $\theta(\lambda) = 1 - \theta_1\lambda - \ldots - \theta_p\lambda^p$ *is a polynomial in* B *of degree* q.

The ACF of a MA(q) process cuts off after lag p, and the PACF tails off as a mixture of exponential decays. Figure 2.2 shows a process of length 1000 generated by a MA(2) process $X_t = \epsilon_t + 0.3\epsilon_{t-1} + 0.2\epsilon_{t-2}$, and corresponding sample ACF and PACF. In general, the MA(q) process has a finite memory, in the sense that observations spaced more than q time units apart are uncorrelated.

Wold (1938) proved a fundamental result in time series: any zero-mean purely nondeterministic stationary process $\{X_t, t \in \mathbb{Z}\}$ possesses an infinite MA representation (MA(∞)). Such processes are called causal AR processes.

Definition 2.3.5 *(Causality). A time series* X_t *is causal if and only if it has an MA(*∞*) representation of the form*

$$X_t = \epsilon_t + \psi_1\epsilon_{t-1} + \psi_2\epsilon_{t-2} + \ldots = \sum_{j=0}^{\infty} \psi_j\epsilon_{t-j} \tag{2.15}$$

where $\psi_0 = 1$, $\{\epsilon_t\} \sim WN(0, \sigma^2)$, *and* $\sum_{j=0}^{\infty} \psi_j^2 < \infty$.

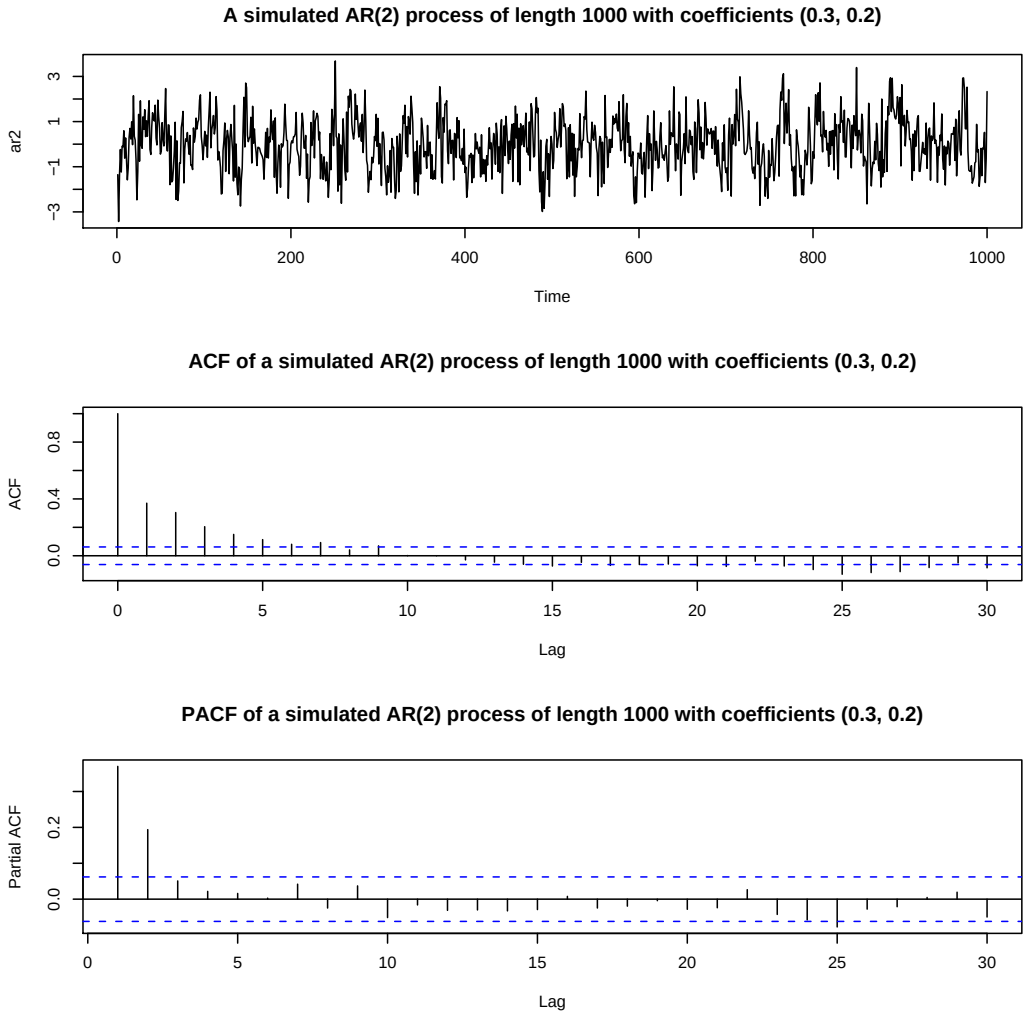

Figure 2.1: An AR(2) process and its ACF, PACF plots

By Definition 2.3.5, an AR(p) process is causal if it can be represented in an MA(∞) form, i.e.,

$$X_t = \frac{1}{\phi(B)}\epsilon_t = \psi(B)\epsilon_t = \sum_{j=0}^{\infty} \psi_j \epsilon_{t-j}, \tag{2.16}$$

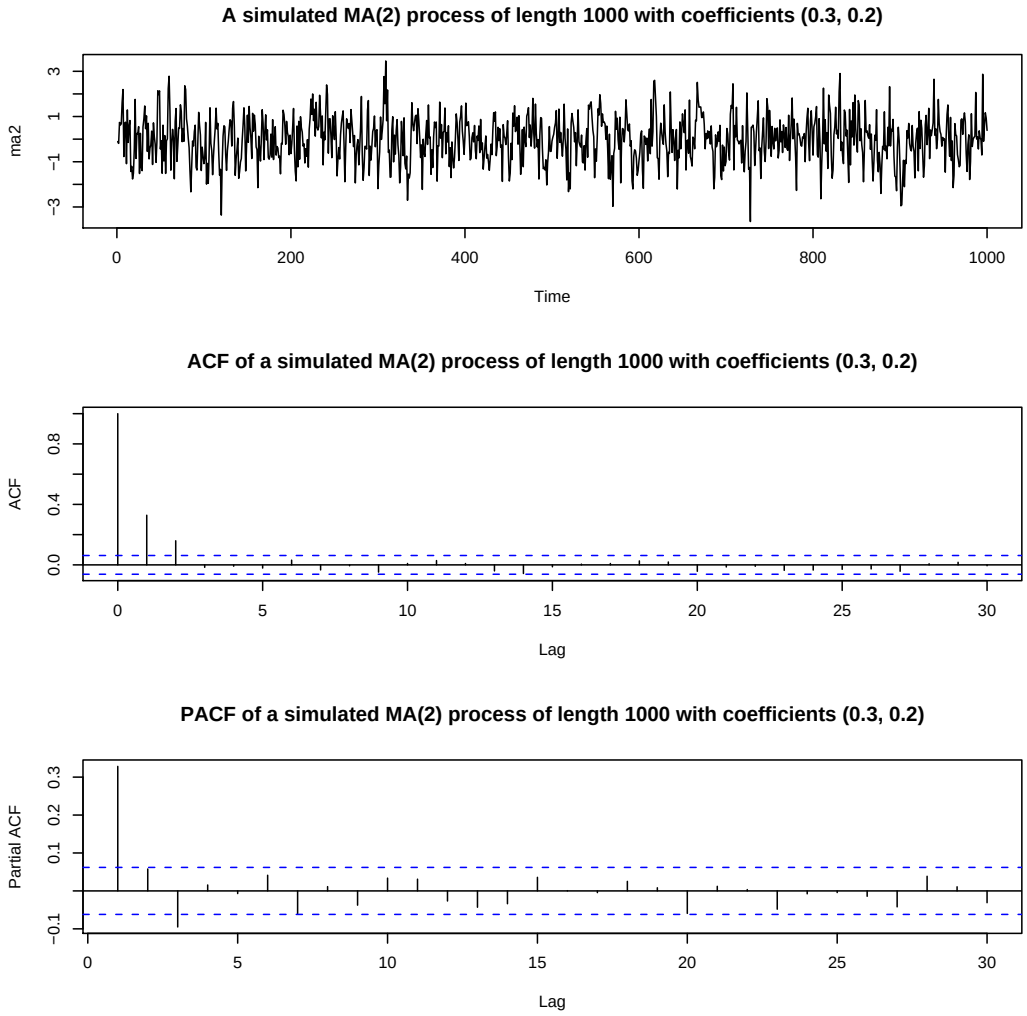

Figure 2.2: A MA(2) process and its ACF, PACF plots

such that $\sum_{j=0}^{\infty} \psi_j^2 < \infty$. To achieve this, all roots of $\phi(\lambda) = 0$ must lie outside the unit circle. Thus, a finite order causal AR process is equivalent to an infinite MA process.

Definition 2.3.6 *(Invertibility). A time series X_t is invertible if and only if it has an*

infinite-order AR (AR(∞)) representation of the form

$$X_t = \pi_1 X_{t-1} + \pi_2 X_{t-2} + ... + \epsilon_t = \sum_{j=1}^{\infty} \pi_j X_{t-j} + \epsilon_t, \tag{2.17}$$

where π_j are constants such that $\sum_{j=1}^{\infty} \pi_j^2 < \infty$.

An MA(q) process is always causal, but invertible only if it can be re-written in an AR(∞) representation, i.e.,

$$\frac{1}{\theta(B)} X_t = \pi(B) X_t = \sum_{j=0}^{\infty} \pi_j X_{t-j} = \epsilon_t \tag{2.18}$$

such that $\sum_{j=1}^{\infty} \pi_j^2 < \infty$ is satisfied. To achieve this, all roots of θ must lie outside the unit circle. Thus, a finite order invertible MA process is equivalent to an infinite AR process.

In summary, there is a dual relationship between AR(p) and MA(q) processes. A finite order causal AR(p) process corresponds to an infinite order MA process, and a finite order invertible MA(q) process corresponds to an infinite order AR process. The duality also exits in ACF and PACF. The AR(p) process has its ACF tailing off and PACF cutting off, but the MA(q) process has its ACF cutting off and PACF tailing off.

Although a causal and invertible process can be represented in either AR or MA forms, the number of model parameters may become large, which could possibly reduce the efficiency in estimation. Another alternative is to combine both AR and MA parts in one model, in order to achieve great flexibility in fitting actual time series.

Definition 2.3.7 *(ARMA Processes). An ARMA(p, q) process, $p, q \in \mathbb{Z}$, is defined as*

$$\phi(B) X_t = \theta(B) \epsilon_t, \tag{2.19}$$

where

- $\{\epsilon_t\} \sim WN(0, \sigma^2)$,
- *$\phi(\lambda)$ and $\theta(\lambda)$ are polynomials in B of degree p and q respectively;*

- *$\phi(\lambda)$ and $\theta(\lambda)$ have no common factors.*

ACF of an ARMA(p, q) process tails off as a mixture of exponential decays after first $q - p$ lags. In particular, ACF has $q - p + 1$ initial values if $p \leq q$. The PACF also tails off after first $p - q$ lags. It eventually behaves like the PACF of pure MA process. Figure 2.3 shows a time series of length 1000 generated by an ARMA(2,1) process $X_t = 0.2X_{t-1} + 0.1X_{t-2} + 0.3\epsilon_{t-1} + \epsilon_t$, and corresponding sample ACF and PACF.

Since AR(p) and MA(q) processes are the two basic building blocks of an ARMA(p, q) process, their causality and invertibility conditions still hold in this general case. An ARMA (p, q) process is causal if all roots of $\phi(\lambda) = 0$ lie outside the unit circle, and is invertible if all the roots of $\theta(\lambda) = 0$ lie outside the unit circle. Note that the causality and invertibility are properties not of the process $\{X_t\}$ alone but rather of the relationship between the two processes $\{X_t\}$ and $\{\epsilon_t\}$ appearing in the ARMA equations. If an ARMA(p, q) process is causal, then it can be represented in MA(∞) form as

$$X_t = \frac{\theta(B)}{\phi(B)}\epsilon_t = \psi(B)\epsilon_t = \sum_{j=0}^{\infty} \psi_j \epsilon_{t-j}, \tag{2.20}$$

where $\sum_{j=0}^{\infty} \psi_j^2 < \infty$. If an ARMA$(p, q)$ process is invertible, then it can be represented by AR(∞) form

$$\frac{\phi(B)}{\theta(B)}X_t = \pi(B)X_t = \sum_{j=0}^{\infty} \pi_j X_{t-j} = \epsilon_t, \tag{2.21}$$

where $\sum_{j=0}^{\infty} \pi_j^2 < \infty$. The properties of AR, MA and ARMA processes are summarized in Table 2.1.

2.3.2 Nonstationary Models

In reality, most time series exhibit non-stationary behavior and typically do not vary about a constant mean. In such cases, we should consider non-stationary models. An ARIMA model, a generalization of the ARMA model, incorporates a wide range of non-stationary processes. The key idea of an ARIMA process is to difference a non-stationary process finitely many times in order to achieve stationarity.

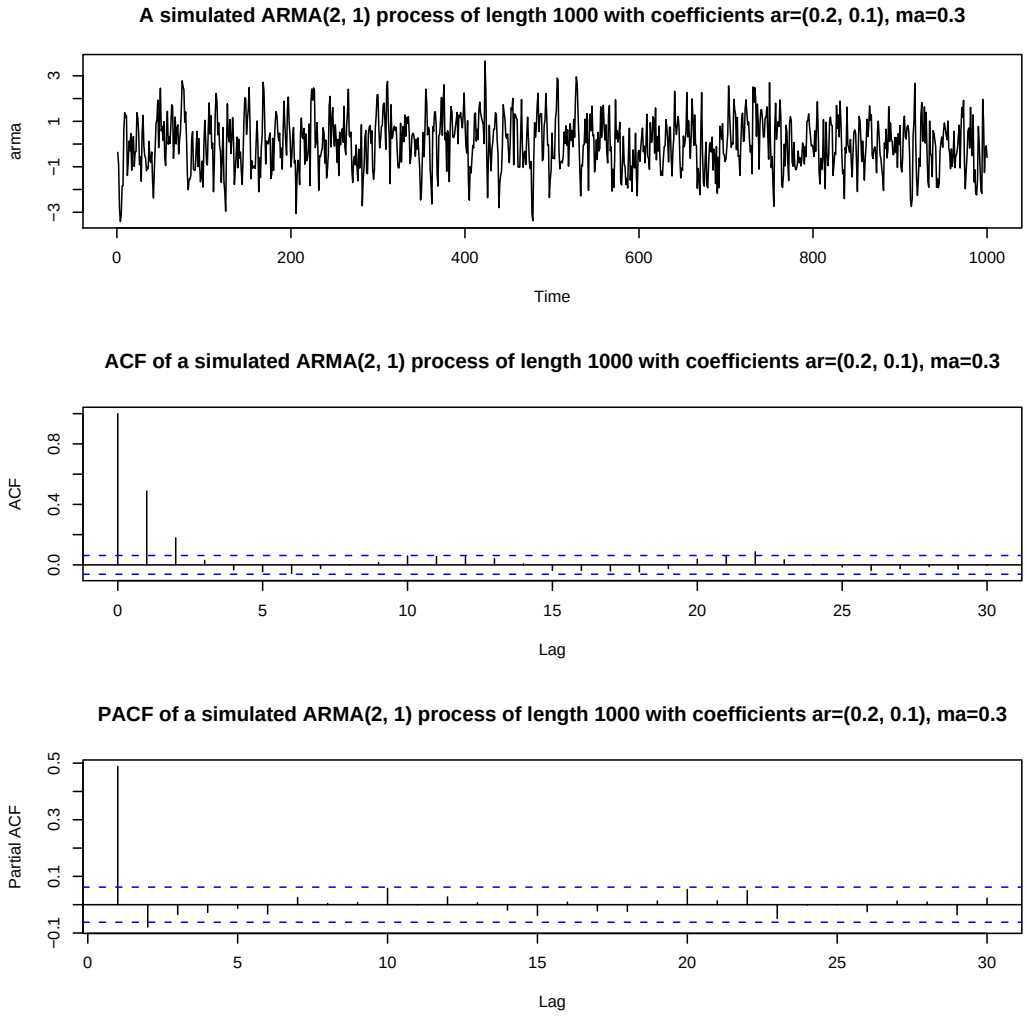

Figure 2.3: An ARMA(2,1) process and its ACF, PACF plots

Definition 2.3.8 *(ARIMA Models). An ARIMA(p, d, q) process, $p, d, q \in \mathbb{Z}$, is defined as*

$$\phi(B)\nabla^d X_t = \theta(B)\epsilon_t, \tag{2.22}$$

where

- $\{\epsilon_t\} \sim WN(0, \sigma^2)$,

	$AR(p)$	$MA(q)$	$ARMA(p,q)$
Causality Condition	Roots of $\phi(\lambda)=0$ lie outside the unit circle	Always causal	Roots of $\phi(\lambda)=0$ lie outside the unit circle
Invertibility Condition	Always invertible	Roots of $\theta(\lambda)=0$ lie outside the unit circle	Roots of $\theta(\lambda)=0$ lie outside the unit circle
AR representation	$\phi(B)X_t=\epsilon_t$, finite	$\theta(B)^{-1}X_t=\epsilon_t$, infinite	$\theta(B)^{-1}\phi(B)X_t=\epsilon_t$, infinite
MA representation	$X_t=\phi(B)^{-1}\epsilon_t$, infinite	$X_t=\theta(B)\epsilon_t$, finite	$X_t=\phi(B)^{-1}\theta(B)\epsilon_t$, infinite
ACF	tails off as a mixture of exponential decays	cuts off after lag q	tails off as a mixture of exponential decays after first $q-p$ lags
PACF	cuts off after lag p	tails off as a mixture of exponential decays	tails off as a mixture of exponential decays after first $p-q$ lags

Table 2.1: Summary of Properties: AR, MA and ARMA processes

- *d is differencing parameter,*
- *$\phi(\lambda)$ and $\theta(\lambda)$ are polynomials in B of degree p and q respectively,*
- *$\phi(\lambda)$ and $\theta(\lambda)$ have no common factor,*
- *$\phi(\lambda)\neq 0$ for $|\lambda|\leq 1$.*

The ARIMA (p,d,q) process X_t is (weakly) stationary if and only if $d=0$. In such case X_t reduces to an ARMA(p,q) process. Figure 2.4 shows a time series of length 1000 generated by an ARIMA(2,1,1) process defined by $(1-B)(X_t-0.2X_{t-1}-0.1X_{t-2})=0.3\epsilon_{t-1}+\epsilon_t$, and the corresponding series after single differencing. The original series

exhibits non-stationary behavior, but becomes (weakly) stationary after being differenced once.

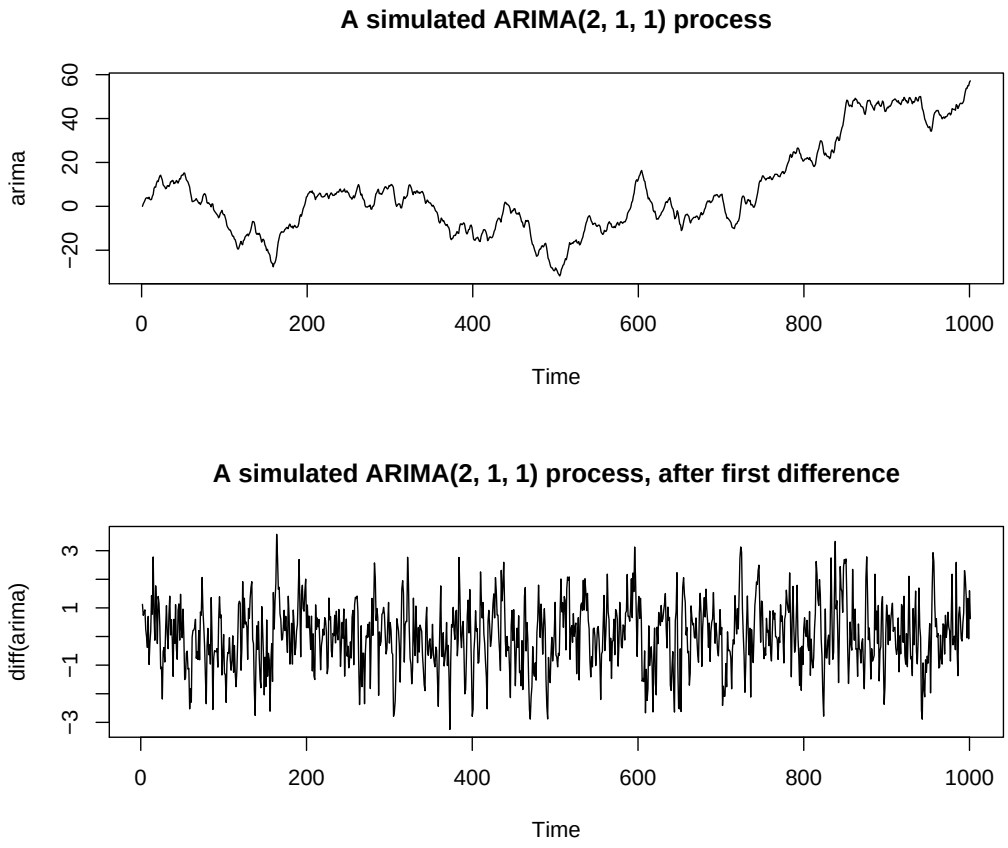

Figure 2.4: An ARIMA(2,1,1) process and its first differencing

2.4 Model Selection

A time series analysis should always begin with a preliminary plot of the data, as an indication of the statistical features that may guide our analysis. Through a careful examination

of the plot, we can often obtain a good idea about whether the series contains a trend, seasonality, nonconstant variance and other non-stationary phenomena. The mains tools for preprocessing data are variance-stablizing transformations and differencing. If the variance of the data is not constant, we may apply logarithmic or Box-Cox's power transformations. If the data displays trend and seasonality, which also can be indicated by slow decaying and periodic ACF, we should consider differencing. In practice, we do not difference data more than twice.

Once the transformed series can potentially be fitted to a causal invertible ARMA model, we need to select appropriate model orders p and q. Using ACF and PACF, we can follow the guidelines from section 2.3,

- An AR(p) process is identified from the property that all values of the PACF after the p-th are negligible.
- A MA(q) process is identified from the property that all values of the ACF after the q-th are negligible.

The sample ACF and PACF approximately have standard deviation $1/\sqrt{n}$, where n is the number of observations in the sample. If the sample ACF and PACF lie between $[-2/\sqrt{n}, 2/\sqrt{n}]$, they can be regarded as negligible. Intuitively, the PACF plot of an AR(p) process has p initial spikes and then fall between the bounds $[-2/\sqrt{n}, 2/\sqrt{n}]$; while its ACF plot tails off as a mixed exponential decay. Conversely, the ACF plot of an MA(q) process has q initial spikes and then fall between the bounds $[-2/\sqrt{n}, 2/\sqrt{n}]$; while its PACF plot tails off as a mixed exponential decay. We may identify the model order depending on the characteristics of ACF and PACF plots. Table 2.1 summaries the important results for selecting p and q using the ACF and the PACF.

However, ACF and PACF plots might suggest multiple adequate models. Based on those candidate models, we need to take further into account the trade-off between bias and variance in order to choose the best suited model. Thus, numerous criteria for model comparison have been developed in the literature for model selection. Below we discuss some of the most popular techniques.

Akaike (1974) introduced Akaike Information Criterion (AIC), which is an asymptotically unbiased estimator of the Kullback-Leibler distance. AIC is defined as

$$AIC = -2\log(L) + 2k \tag{2.23}$$

where L is the maximum likelihood function for the estimated model and k is the number of parameters in the model. AIC attempts to find the model that best explains the data with a minimum parameters. The two terms in AIC equation represent the goodness of fit and number of parameters. AIC not only rewards goodness of fit, but also includes a penalty for increasing model parameters. The optimal model is the one with minimum AIC.

Shibata (1976) shows that AIC is asymptotically efficient in large samples, but it might drastically overfit in small samples. Hurvich and Tsai (1989) prove that AIC can be very biased as an estimate of the Kullback-Leibler distance when a sample size is small, and propose a corrected version of AIC, namely AICc, defined as follows

$$AICc = AIC + \frac{2k(k+1)}{T-k-1} \tag{2.24}$$

where T is the sample size. AICc converges to AIC as $T \to \infty$, so it inherits the asymptotically efficiency property of AIC in large samples but significantly outperforms the AIC in small samples, in the sense that AICc is an almost unbiased estimate of the Kullback-Leibler distance.

Schwartz (1978) suggested a Bayesian Information Criterion (BIC), defined as

$$BIC = -2\log(L) + k\log T \tag{2.25}$$

where L is the maximum likelihood function for the estimated model, k is the number of parameters in the model, and T is a sample size. BIC penalizes the increase of parameters more strongly than AIC. The optimal model is the one with a minimal BIC.

Parzen (1977) proposed a model selection criterion CAT for AR(p) models, defined as

$$\text{CAT (p)} = \begin{cases} -(1+\frac{1}{T}), & p = 0, \\ \frac{1}{T}\sum_{j=1}^{p}\frac{1}{\hat{\sigma}_j^2} - \frac{1}{\hat{\sigma}_p^2}, & p = 1, 2, ... \end{cases} \tag{2.26}$$

where T is a number of observations, $\hat{\sigma}_j^2$ is the unbiased estimate of σ^2 when an AR(j) is fitted to the data. The optimal order p is chosen so that CAT(p) achieves its minimum. Parzen's CAT is asymptotically efficient.

We have discussed above only several most commonly used model selection criteria. There are many other criteria in the literature, based on either residuals or forecasting error, such as final prediction error (FPE) (Akaike, 1969), Mallows CP (Mallows, 1973), Hannan and Quinn criterion (HQ) (Hannan and Quinn, 1979), PLS (Rissanen, 1984), KIC (Cavanaugh, 1999) and others. See Wei (1992), McQuarrie and Tsai (1998), Burnham and Anderson (2002) for more overview and detailed discussion.

2.5 AR Parameter Estimation

After identifying a tentative model, the next step is to estimate the model parameters. The main focus of this thesis is on AR models, hence, so we discuss several most popular AR parameter estimation approaches.

Suppose $\{X_t\}$ is an AR(p) process: $\sum_{k=0}^{p} \phi_k X_{t-k} = \epsilon_t$, and assume $\phi_0 = 1$. Our goal is to estimate $\tilde{\phi} = (\phi_1, ..., \phi_p)'$.

Maximum Likelihood Estimation (MLE)

Assume that ϵ_t are independently and normally distributed random variables with $E[\epsilon_t] = 0$ and $E[\epsilon_t^2] = \sigma^2$. Let $\tilde{X}_t = (X_t, ..., X_{t-p+1})'$. Then the ML estimates of $\tilde{\phi}$ are obtained by maximizing the likelihood function

$$L(\tilde{\phi}, \sigma) = \frac{1}{(2\pi\sigma^2)^{T/2}} \exp\left\{ -\frac{1}{2\sigma^2} \sum_{t=1}^{T} (X_t + \tilde{\phi}' \tilde{X}_{t-1})^2 \right\}. \tag{2.27}$$

The MLE offers the best prospect of giving efficient parameter estimates, whose variances achieves Cramér-Rao lower bounds. The ML estimators are asymptotically consistent and normally distributed. However, there are certain drawbacks. The ML method is based on a non-linear optimization that requires good initial values. Such estimation is computationally very expensive, and poor choices of initial values might result in meaningless

estimates. Also, the distribution of ϵ_t is often unknown a priori which implies that a functional form of L is also unavailable. In practice, most existing methods do not use exact maximum likelihood but various approximations thereto, such as quasi-MLE and Whittle's approximation. A detailed discussion can be found in Brockwell and Davis (1987), Anderson (1971).

Yule-Walker (YW) Estimation

Assume that ϵ_t is WN$(0, \sigma^2)$. Let $\gamma_p = (\gamma_0, \gamma_1, ..., \gamma_p)'$ be the autocovariances of the process X_t. Then the Yule-Walker (YW) estimates of $\tilde{\phi}$ are obtained by solving the YW system of linear equations:

$$\begin{aligned} R_p\tilde{\phi} &= \gamma_p \\ \sigma^2 &= \gamma_0 - \tilde{\phi}'\gamma_p \end{aligned} \tag{2.28}$$

where R_p is the $(p+1) \times (p+1)$-Toeplitz covariance matrix $[\gamma_{i-j}]_{i,j=1}^{p}$. In practice, R_p and γ_p are replaced by sample estimates.

The YW estimation is distributional-free and has a simple computational procedure since the model parameters are obtained by only solving p linear equations. Also, the parameters estimated by the YW method provide the best linear forecast of X_t based on $X_{t-1}, ..., X_{t-p}$, which yields the minimum mean square error of prediction. However, the YW method, based on sample moments, is typically far less efficient than ML. In asymptotics, the YW estimates possess approximately the same distribution properties as the corresponding ML estimates.

Least Square (LS) Estimation

Assume that ϵ_t is WN$(0, \sigma^2)$. The Least Squares (LS) estimates of $\tilde{\phi}$ are obtained by minimizing a sum of square errors. The LS estimates can be obtained from the following recursive equations and hence, can be viewed as a stationary Kalman filter (Abraham and Ledolter, 2005):

$$\begin{aligned} \phi_{t+1} &= \phi_t + \gamma_t\Phi_t(X_{t+1} - \Phi_t'\phi_t) \\ \gamma_{t+1} &= \gamma_t - \gamma_t\Phi_{t+1}(1 + \Phi_{t+1}'\gamma_t\Phi_{t+1})^{-1}\Phi_{t+1}'\gamma_t \end{aligned} \tag{2.29}$$

where $\Phi_t = (X_t, X_{t-1}, \ldots, X_1)'$.

A modification of the LS equations (2.29) for estimation of ARMA models is known as the Extended Least Squares (ELS) method (Panuska, 1969; Young et al, 1970; Ljung, 1977) method. The iterative procedure significantly reduces the computational burden and enables to utilize such recursive estimates for on-line modeling and forecasting, i.e., when the size of observations is unknown a priori. Generally, the LS estimates have larger variances than the ML estimates but possess similar asymptotic properties, i.e., the LS estimates are consistent and normally distributed when sample size is large. We will discuss the regularized version of the LS estimation in the next chapter.

Chapter 3

Regularized AR approximation for ARMA process

The assumption that an observed process follows a finite AR, MA or ARMA model is rarely justified in practice. A common approach is to approximate the true model by finite AR models whose orders are selected by information criterions such as AIC, BIC and whose parameters are estimated by LS, YW or other methods (Akaike, 1969, 1970; Parzen, 1974). However, as sample size increases, it often implies that the model order should be refined and hence all the parameters need to be recalculated. In order to avoid such shortcomings, we introduce an alternative approach: fitting a "long" AR model to the process and estimating its parameters using the Regularized LS (RLS) method (Gel and Fomin, 2001, Gel and Barabanov, 2007). Regularizer in the LS method enables to estimate AR parameters with different level of accuracy. In particular, the first few model parameters are estimated more precisely than the tail ones, and the number of estimated coefficients grows with the sample size. Therefore, the repeated model selection and parameter estimation are avoided as the observed sample increases. In this chapter we illustrate such an approach in the context of identifying ARMA processes and sketch the proofs on almost sure convergence of the RLS estimates for AR(∞) model with exponentially decaying coefficients, as presented by the Gel and Barabanov (2007). The main contribution of this thesis is an extension of the RLS method to a more general class of time series, i.e. periodic and long memory processes, and thus the theoretical results of this chapter serve as motivation and

foundation for main findings of the following two chapters.

3.1 Problem Statement

Consider an ARMA(p, q) process, $p, q \in \mathbb{Z}$,

$$\phi(B)y_t = \theta(B)\epsilon_t, \quad t \in T, \tag{3.1}$$

where

- $\{\epsilon_t\}$ is the martingale difference $E(\epsilon_t|\mathcal{F}_{t-1}) \equiv 0$, and $E(\epsilon_t^2|\mathcal{F}_{t-1}) = \sigma^2$ a.s., where $\mathcal{F}_{t-1}$ is the σ-algebra generated by r.v. $(\epsilon_1, \epsilon_2, \ldots, \epsilon_{t-1})$, $\sup_t E(\epsilon_t^4) < \infty$;
- $\phi(\lambda)$ and $\theta(\lambda)$ are polynomials in B of degree p and q respectively;
- $\phi(\lambda)$ and $\theta(\lambda)$ have no common factor.

There exist various methods for parameter estimation of ϕ and θ that can be generally classified to likelihood-based and sum-of-square-based techniques. The most commonly used method is the Maximum Likelihood Estimation (MLE). As we discussed in the previous chapter, the ML estimates achieve the Cramér-Rao lower bound and are asymptotically consistent and normally distributed (Brockwell and Davis, 1987). However, ML assumes that $\{\epsilon_t\}$ are i.i.d. normal random variables, which is rarely justified in practice. Also, maximizing the likelihood function is a non-linear optimization that requires good initial values. Such estimation procedure is computationally very expensive and poor choices of initial values might result in meaningless estimates. In the case of on-line estimation, i.e., when a size of observations is unknown a priori, the MLE re-estimates all model parameters non-recursively upon arrival of every new observation, which is not feasible in practice due to high computational routine and possible processing delays.

The Extended LS (ELS) method in a form of a stationary Kalman filter, discussed in the previous chapter, has the advantages of computational efficiency, reduced storage requirements and minimum processing delays, which make such estimation ideal for on-line modeling. Agafanov et al.(1982), Fomin (1985), Ljung and Söderström (1983) show the almost sure convergence of the ELS method for an ARMA equation with exponentially

decaying AR coefficients and the positive realness condition on the transfer function, i.e., $|\theta(\lambda) - 1| < 1$ for $|\lambda| = 1$. Fomin (1995) extended the results to the case of weakly stable ARMA equation, i.e. with roots of an AR polynomial along the unit circle, under the same assumptions. However, the positive realness condition of the transfer function is difficult to verify in practice, and hence there exists no assurance that ELS converge with probability 1, which can be important in many applications, e.g. electrical engineering.

Since causal invertible ARMA process can be transformed into an AR(∞) model, an alternative approach for ARMA estimation is to obtain coefficients of the approximating AR(p) model with sufficiently large p and then use the Pade method or stochastically balanced truncation (SBT) to convert the AR parameters into ARMA parameters.

Assume the ARMA(p, q) process is causal invertible. Then ARMA(p, q) can be transformed into an AR(∞) process as follows:

$$a(B)y_t = \epsilon_t, \quad \text{where } a(B) = \frac{\phi(B)}{\theta(B)} = \sum_{j=0}^{\infty} a_j B^j. \tag{3.2}$$

Such a class of models includes (but not limited to) all causal invertible ARMA(p, q) models. In practice, we usually truncate AR(∞) and consider a finite AR(p) approximation, where p is chosen by information criterions such as AIC (Akaike, 1974), BIC (Schwartz, 1978), CAT (Parzen, 1977) or others. Typically the model order p is selected as $p \sim \ln T$. After p is determined, the parameters of the AR(p) model are estimated by LS, YW, the method of stochastic approximation (MSA), MLE, etc. The asymptotic properties of approximating AR parameter estimates are well investigated and date back to Mann and Wald (1943). (For detailed discussion and overview see, for example, Anderson (1971), Said and Dickey (1984), Gel and Fomin (1998), Brockwell and Davis (1987), Wei (1990) and Mari et al.(2000) and references therein.) However, as sample size increases, it often implies that the model order is to be refined and, hence, all the parameters need to be recalculated, which dramatically increases the computational burden.

In this chapter, we investigate the limiting case for approximating by finite AR(p) model, namely an AR(∞) model. In the infinite dimensional case, the usual LS estimator is not consistent in the sense that the loss function does not have a unique minimum. Therefore, the Regularized LS (RLS) (Gel and Fomin, 2001, Gel and Barabanov, 2007) method is applied to identifying AR(∞) models. The utilization of the regularizer in the

LS estimation procedure guarantees the uniqueness of the minimum of the loss function in ℓ_2 norm and ensures that the obtained vector of parameter estimate lies in ℓ_2 space. The RLS estimates for AR(∞) models converges almost surely at a power law rate.

Based on the theoretical results of the RLS estimates of AR(∞) model, we propose a regularized AR (RAR) approximation to identify causal invertible ARMA(p, q) models, which includes the following two steps:

- Step 1: in practice, we approximate the AR(∞) model by a "long" AR model whose order significantly exceeds the order suggested by information criterions, and may be potentially equal to the number of observations. The parameters of such a "long" AR model are estimated by the RLS method. The idea behind this approach is that the model parameters are actually estimated with different level of accuracy which is controlled by a specially chosen regularizer, and the selection of regularizer is equivalent to the selection of model order using information criterions.

- Step 2: estimate the underlying ARMA parameters by the Pade method, which approximates a continuous function by ratio of two polynomials and determines the numerator and denominator coefficients (Baker and Graves-Morris, 1996), or SBT, which transforms a system to a balanced form and then applies the principal component analysis to the partial correlation coefficients of the obtained balanced system (Desai and Pal, 1984; Mari et al., 2000).

Hence, the RAR approximation enables to avoid repeated model selection and parameter estimation when the observed sample increases and hence to reduce the computational cost in online modeling and forecasting. The regularizer may be viewed as the smoothing operator applied to the number of the AR coefficients being estimated, and constitutes a link to the model selection criterion, such as AIC and BIC. We may choose the regularizer by cross validation, which is similar to the approach of thresholding selection proposed by Bickel and Levina (2007), but in a time series context. We discuss the RLS estimation in detail in the next section.

3.2 Regularized LS Estimation

Consider the AR(∞) model in (3.2)

$$a(B)y_t = \epsilon_t,\ t = 1, 2, ...$$

Our goal is to estimate the unknown parameters of the power series $a(z) = 1 + a_1 z + a_2 z^2 + \ldots$, where $a(z)$ is assumed to be analytic and has no zeros in a certain neighborhood of the unit circle. Initial conditions are assumed to be zeros. Equivalently, we may express (3.2) in the form of a linear observation scheme

$$y_t = \boldsymbol{\Phi}'_{t-1}\boldsymbol{\tau}_T + \epsilon_t. \tag{3.3}$$

where $\boldsymbol{\Phi}_{t-1} = (y_{t-1}, y_{t-2}, \ldots, y_1, 0, \ldots)'$ and $\boldsymbol{\tau}_T = -(a_1, a_2, \ldots)'$ are infinite dimensional vectors. Note that $\boldsymbol{\Phi}_t$ has no more than t non-zero elements at time t due to zero initial conditions.

The sequence of estimates $\boldsymbol{\vartheta}$ of parameters $\boldsymbol{\tau}_T$ is determined from the minimum condition for the loss function

$$\begin{aligned} J_T(\boldsymbol{\vartheta}) &= \frac{1}{T}\sum_{t=1}^{T}(y_t - \boldsymbol{\Phi}'_{t-1}\boldsymbol{\vartheta})^2 = \frac{1}{T}\left(\boldsymbol{\vartheta}'\check{\boldsymbol{R}}_T\boldsymbol{\vartheta} - 2\boldsymbol{\vartheta}'\check{\boldsymbol{r}}_T + \sum_{t=1}^{T} y_t^2\right) \\ &= \frac{1}{T}\left((\boldsymbol{\tau}_T - \boldsymbol{\vartheta})'\check{\boldsymbol{R}}_T(\boldsymbol{\tau}_T - \boldsymbol{\vartheta}) + 2(\boldsymbol{\tau}_T - \boldsymbol{\vartheta})'\check{\boldsymbol{r}}^{\epsilon}_T + \sum_{t=1}^{T}\epsilon_t^2\right), \end{aligned} \tag{3.4}$$

where $\check{\boldsymbol{R}}_T = \sum_{t=1}^{T}\boldsymbol{\Phi}_{t-1}\boldsymbol{\Phi}'_{t-1}$, $\check{\boldsymbol{r}}_T = \sum_{t=1}^{T}\boldsymbol{\Phi}_{t-1}y_t$, $\check{\boldsymbol{r}}^{\epsilon}_T = \sum_{t=1}^{T}\boldsymbol{\Phi}_{t-1}\epsilon_t$.

The LS estimates are defined as

$$\hat{\boldsymbol{\tau}}_T = \arg\min_{\boldsymbol{\vartheta}} J_T(\boldsymbol{\vartheta}) = \check{\boldsymbol{R}}_T^{+}\check{\boldsymbol{r}}_T = \boldsymbol{\tau}_T + \check{\boldsymbol{R}}_T^{+}\check{\boldsymbol{r}}^{\epsilon}_T, \tag{3.5}$$

which is equivalent to

$$\inf_{\boldsymbol{\vartheta}} J_T(\boldsymbol{\vartheta}) = \frac{1}{T}\left(-[\check{\boldsymbol{r}}_T]'\check{\boldsymbol{R}}_T^{+}\check{\boldsymbol{r}}_T + \sum_{t=1}^{T} y_t^2\right) = \frac{1}{T}\left(-[\check{\boldsymbol{r}}^{\epsilon}_T]'\check{\boldsymbol{R}}_T^{+}\check{\boldsymbol{r}}^{\epsilon}_T + \sum_{t=1}^{T}\epsilon_t^2\right). \tag{3.6}$$

Here $\check{\boldsymbol{R}}_T^+$ denotes the pseudoinverse of the operator $\check{\boldsymbol{R}}_T : \ell_2(\mathbb{N}) \to \ell_2(\mathbb{N})$,

$$\check{\boldsymbol{R}}_T^+ = \boldsymbol{P}_T \check{\boldsymbol{R}}_T^{-1} \boldsymbol{P}_T + (\boldsymbol{I}_{\ell_2} - \boldsymbol{P}_T), \tag{3.7}$$

where $\boldsymbol{P}_T$ is the orthogonal projector onto the subspace of the range of the operator $\check{\boldsymbol{R}}_T$; $\boldsymbol{P}_T \check{\boldsymbol{R}}_T^{-1} \boldsymbol{P}_T$ is the inversion of the operator $\boldsymbol{P}_T \check{\boldsymbol{R}}_T \boldsymbol{P}_T$ in the invariant subspace $\boldsymbol{P}_T \ell_2(\mathbb{N})$. Note that the operator $\check{\boldsymbol{R}}_T^+$ is unambiguously defined by the relation $\check{\boldsymbol{R}}_T^+ \check{\boldsymbol{R}}_T = \check{\boldsymbol{R}}_T \check{\boldsymbol{R}}_T^+ = \boldsymbol{P}_T$. If the operator $\check{\boldsymbol{R}}_T$ is invertible ($\boldsymbol{P}_T = \boldsymbol{I}_{\ell_2}$), then $\check{\boldsymbol{R}}_T^+ = \check{\boldsymbol{R}}_T^{-1}$.

If $\boldsymbol{\tau}_T$ has a finite number of nonzero elements, i.e, the AR model (3.2) is of finite order, then it can be shown that $\check{\boldsymbol{R}}_T^+ \check{\boldsymbol{r}}_T^\epsilon$ in (3.5) converges to zero with probability 1 as $T \to \infty$. Hence, the LS estimate $\hat{\boldsymbol{\tau}}_T$ is strongly consistent (Kushner and Yin, 2003; Ljung, 1998; Fomin 1999). However, if $\boldsymbol{\tau}_T$ has an infinite number of nonzero elements, i.e., the model (3.2) does not degenerate to AR(p), then $\inf_{\boldsymbol{\vartheta}} J_T(\boldsymbol{\vartheta})$ in (3.7) equals to zero for an arbitrary T. Thus, the LS estimate can not be consistent in such a case. In order to achieve the strong consistency in the infinite dimensional case, we introduce the Regularized LS (RLS) estimates (Gel and Fomin, 2001):

$$\hat{\boldsymbol{\tau}}_T = (\check{\boldsymbol{R}}_T + \varepsilon \boldsymbol{\Lambda})^{-1} \check{\boldsymbol{r}}_T, \tag{3.8}$$

where $\varepsilon > 0$ is a fixed constant, and $\boldsymbol{\Lambda}$ is a regularizer, which is a positive definite operator in the Hilbert space ℓ_2 under the inner product $\langle ., . \rangle$, and $\boldsymbol{\Lambda}^{-\delta}$ is an operator of trace class for a arbitrarily small $\delta > 0$.

The regularizer $\boldsymbol{\Lambda}$ plays an crucial role for the accuracy of the estimate $\hat{\boldsymbol{\tau}}_T$. In the infinite dimensional $\ell_2(\mathbb{N})$, the non-regularized LS loss function may fail to have unique minimum, i.e, the non-regularized LS estimates converge but not necessarily to the vector of unknown parameters $\boldsymbol{\tau}_T$. The utilization of the regularizer guarantees the uniqueness of the minimum, given by $\boldsymbol{\tau}_T$, of the LS loss function in $\ell_2(\mathbb{N})$. Moreover, the elements in the regularizer grow at a certain rate to ensure that the obtained vector of parameter estimates lies in $\ell_2(\mathbb{N})$.

The regularizer may take different forms. In this thesis, we consider the regularizer of

an exponential form (Gel and Fomin, 2001),

$$\mathbf{\Lambda} = \mathrm{diag}\{\mathrm{e}^{\mu\mathrm{k}}\}_{\mathrm{k}=1}^{\infty} = \begin{pmatrix} e^{\mu} & 0 & \dots & 0 & \dots \\ 0 & e^{2\mu} & \dots & 0 & \dots \\ \vdots & \vdots & \ddots & \vdots & \dots \\ 0 & 0 & \dots & e^{\mu k} & \dots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix} \tag{3.9}$$

Barabanov and Gel (2005) proposed the regularizer of a polynomial form ,

$$\mathbf{\Lambda} = diag\{k^p\}_{k=1}^{\infty},\ p > 3, \tag{3.10}$$

which is viewed as a weakening of the regularizing procedure.

For the sake of implementation convenience, we adopt the recursive procedure of the LS estimates (Abraham and Ledolter, 1983; Fomin, 1999). The RLS estimates satisfy the iterative relations represented by the Kalman filter equations:

$$\begin{aligned} \hat{\boldsymbol{\tau}}_{T+1} &= \hat{\boldsymbol{\tau}}_T + \boldsymbol{\gamma}_T^{\varepsilon}\mathbf{\Phi}_T(1 + \mathbf{\Phi}'_{T+1}\boldsymbol{\gamma}_T^{\varepsilon}\mathbf{\Phi}_{T+1})^{-1}(y_{T+1} - \mathbf{\Phi}'_T\hat{\boldsymbol{\tau}}_T) \\ \boldsymbol{\gamma}_{T+1}^{\varepsilon} &= \boldsymbol{\gamma}_T^{\varepsilon} - \boldsymbol{\gamma}_T^{\varepsilon}\mathbf{\Phi}_{T+1}(1 + \mathbf{\Phi}'_{T+1}\boldsymbol{\gamma}_T^{\varepsilon}\mathbf{\Phi}_{T+1})^{-1}\mathbf{\Phi}'_{T+1}\boldsymbol{\gamma}_T^{\varepsilon} \end{aligned} \tag{3.11}$$

with initial conditions $\hat{\boldsymbol{\tau}}_0 = 0$ and $\boldsymbol{\gamma}_0^{\varepsilon} = (\varepsilon\mathbf{\Lambda})^{-1}$. The matrix $\boldsymbol{\gamma}_T^{\varepsilon}$ is inverse to the sample information matrix $\check{\boldsymbol{R}}_T^{\varepsilon}$, i.e. $\boldsymbol{\gamma}_T^{\varepsilon} = (\check{\boldsymbol{R}}_T^{\varepsilon})^{-1}$, where $\check{\boldsymbol{R}}_T^{\varepsilon} = \sum_{t=1}^{T} \mathbf{\Phi}_t\mathbf{\Phi}'_t + \varepsilon\mathbf{\Lambda}$.

The main results of the RLS estimates in the infinite dimensional case are:

- The RLS estimates converges almost surely (a.s.) to the unknown vector $\boldsymbol{\tau}_T$ in terms of ℓ_2 norm, i.e., they are strongly consistent.
- The RLS estimates have a power law rate of a.s. convergence to the unknown vector $\boldsymbol{\tau}_T$ in terms of ℓ_2 norm.

3.3 Brief Overview of Regularization

The concept of regularization was first introduced by Tikhonov (1943), in the context of solving an integral equation in a numerically stable manner. His germinating idea is essentially the method of penalized regression in the statistical framework. After Tikhonov's

introduction of the concept, there have been numerous works dedicated to regularization in statistical inference as diverse as variable selection, covariance estimation, and Efron's bootstrap. In general, regularization is a class of methods used to modify estimation procedures to produce reasonable solutions in unstable situations. In an asymptotic sense, a generic regularization process includes two stages:

- Stage one: construct a sequence of approximating parameters θ_k converge to the target parameter θ, and for each k a sequence of estimators $\hat{\theta}_k$ converges to θ_k,
- Stage two: choose a data dependent value $\hat{k}$ for k.

It is often useful to decompose the difference

$$\hat{\theta}_k - \theta = (\hat{\theta}_k - \theta_k) + (\theta_k - \theta). \tag{3.12}$$

The distance between $\hat{\theta}_k$ and θ_k is called estimation error (variance), and the distance between θ_k and θ is called approximation error (bias). Therefore, the choice of k is essentially the choice of the best balance between bias and variance.

With the advent of information technology age, both size and complexity are the main features of most data sets. The size allows us to analyze the data nonparametrically, but usually in a unstable and discontinuous manner. The complexity often implies high dimensionality and requires more advanced models with a large number of parameters to be fitted to the data, which is inherently unstable (Breiman, 1996). In both cases, regularization is an important tool to extract useful information from the data.

In the contexts of nonparametric regression, most of estimation problems are ill-posed and thus regularization is intensively applied in order to turn ill-posed problems to well-posed ones. In particular, when the number of predictor variables in the regression model is larger than the sample size, i.e, in the case of "overfitting", solutions to the LS equations are not unique, and thus new observations become not uniquely predictable. Hoerl and Kennard (1970) proposed ridge regression to guarantee the uniqueness of the solutions by adding a penalty term to the residual sum of squares. Currently, the counterpart of ridge regression, "Lasso" regression (Tibshirani, 1996; Meinshausen, 2005; Bunea et al., 2005, 2006) is attracting the greatest attention and is extensively investigated. Penalization is definitely not the only form of regularization being used in statistics. In the contexts of

density estimation, besides the oldest method to binning in histogram, Rosenblatt (1956) and Parzen (1962) proposed kernel methods. These methods, in turn, led to the Nadaraya-Watson estimation (Nadaraya, 1964; Watson 1964).

Other than applications in nonparametric regression contexts, regularization is also widely used in estimation of covariance matrix. The most recent methods includes Ledoit and Wolf approach (2003), which considers the Steinian shrinkage toward the identity. Furrer and Bengtsson (2006) proposed the "tapering" the sample covariance matrix. Wu and Pourahmadi (2003) suggested the Cholesky decomposition of the covariance matrix to bound the inverse covariance matrix from below. dAspremont et al. (2007) applied ℓ_1 penalties directly to the entries of the covariance matrix, and Bickel and Levina (2006) considered regularizing the covariance matrix by thresholding. The initial idea of RAR approach comes from applying regularization to the sample information matrix, i.e. (auto)covariance matrix in a time series context.

3.4 Strict Consistency and Rate of a.s. Convergence of RLS in $\ell_2(\mathbb{N})$

The proofs given in this section is a sketch of the convergence analysis of the RLS estimates presented by Gel and Barabanov (2007). The following lemma forms a basis for the subsequent results.

Lemma 3.4.1 *Assume the stochastic variables $\xi_t \geqslant 0$ and ζ_t satisfy*

1. $\xi_0 = 0$, $\forall t \geqslant 0$, $E(\xi_{t+1}|\xi_t, ..., \xi_1) \leqslant \xi_t + \zeta_t$;
2. $\sum_{t=0}^{\infty} E|\zeta_t| = C < \infty$.

Then

$$\forall \alpha > 0 \quad P\left\{\forall T \geqslant 0,\ \xi_T \leqslant \frac{C}{\alpha}\right\} \geqslant 1 - \alpha. \tag{3.13}$$

Corollary 3.4.2 *Let $\psi_T = \mu_T \sum_{t=1}^{T} \epsilon_{t-1}\eta_t$, where μ_T is a decreasing positive function, the vector random process $\{\eta_t\}$ is a martingale difference sequence with respect to σ-algebra*

$\mathcal{F}_t$; ϵ_t is measurable with respect to $\mathcal{F}_t$, and

$$\sum_{t=1}^{\infty} \mu^2 E\Big(\epsilon_{t-1}^2 E\big\{\eta_t^2\big|\mathcal{F}_t\big\}\Big) \leqslant C < \infty.$$

Then

$$\forall \alpha > 0 \quad P\left\{\forall T > 0,\ \psi_T^2 \leqslant \frac{C}{\alpha}\right\} \geqslant 1 - \alpha.$$

Corollary 3.4.2 directly follows from Lemma 3.4.1 with $\xi_t = \psi_t^2$ and $\zeta_t = \mu_t^2 \epsilon_{t-1}^2 E\big\{\eta_t^2\big|\mathcal{F}_t\big\}$. We will use Corollary 3.4.2 of the following form to analyze the rate of a.s convergence. Let $0 < \delta_1 < \delta_2$, $\phi_t = t^{\delta_2}\psi_t^2$ and $\alpha = \beta t_0^{\delta_1}$. Then

$$\forall \beta > 0,\ \forall t_0 > 0, \quad P\left\{\forall t \geqslant t_0,\ \phi_t \leqslant \frac{C}{\beta t^{\delta_2 - \delta_1}}\right\} \geqslant 1 - \frac{\beta}{t_0^{\delta_1}}, \tag{3.14}$$

which implies that ϕ_t converge to zero almost surely with power law rate.

Now, we state the following result (Gel and Barabanov, 2007) on the rate of convergence of the RLS estimates for the AR(∞) model.

Theorem 3.4.3 *For any $0 < \delta < 1$ and $\alpha > 0$, there exist positive constants T_1, C_1 such that*

$$\forall T_0 \geqslant T_1, \quad P\left\{\forall T \geqslant T_0, \quad |\hat{\boldsymbol{\tau}}_T - \boldsymbol{\tau}_T|^2 \leqslant \frac{C_1}{T^{1-\delta}}\right\} \geqslant 1 - \frac{\alpha}{T_0^{\delta}}. \tag{3.15}$$

Theorem 3.4.3 states the power law rate of almost sure convergence to zero of the estimation error for $T_0^{\delta} > \alpha$.

Corollary 3.4.4 *(The power law rate of a.s. convergence). For any $\delta > 0$*

$$\lim_{T \to 0} T^{1-\delta} |\hat{\boldsymbol{\tau}}_T - \boldsymbol{\tau}_T|^2 = 0. \tag{3.16}$$

with probability 1.

The proof of Theorem 3.4.3 follows directly by the following two theorems. Let $\delta > 0$, we define the standard quadratic form associated with the LS algorithm (Barabanov, 1983):

$$V_{T+1} = T^{-\delta}(\hat{\boldsymbol{\tau}}_{T+1} - \boldsymbol{\tau}_{T+1})' \check{\boldsymbol{R}}_T^{\varepsilon} (\hat{\boldsymbol{\tau}}_{T+1} - \boldsymbol{\tau}_{T+1}),\ T = 1, 2, ... \tag{3.17}$$

Assume that all quantities with negative indices are 0, and $N = N(T)$ is a certain deterministic function of T such that $N < T$. The vector $\boldsymbol{\Phi}_t$ may be expressed in a state space form as follows

$$\boldsymbol{\Phi}_t = \boldsymbol{A}^N \boldsymbol{\Phi}_{t-N} + \sum_{j=0}^{N-1} \boldsymbol{A}^j \boldsymbol{B} \epsilon_{t-j}, \tag{3.18}$$

where

$$\boldsymbol{A} = \begin{pmatrix} -a_1 & -a_2 & -a_3 & \dots \\ 1 & 0 & 0 & \dots \\ 0 & 1 & 0 & \dots \\ \vdots & \vdots & \ddots & \end{pmatrix}, \quad \boldsymbol{B} = \begin{pmatrix} 1 \\ 0 \\ 0 \\ \vdots \end{pmatrix}$$

Theorem 3.4.5 *$V_T \to 0$ as $T \to \infty$ with probability 1. Moreover, there exists a positive constant C_2 such that*

$$\forall \alpha > 0, \quad P\left\{ \forall T \geqslant 0, \quad V_{T+1} \leqslant \frac{C_2}{\alpha} \right\} \geqslant 1 - \alpha. \tag{3.19}$$

The proof of Theorem 3.4.5 is essentially based on the inequalities

- $E(V_{T+1}|\mathcal{F}_T) \leqslant V_T + T^{-\delta}\sigma^2 \boldsymbol{\Phi}'_T \boldsymbol{\gamma}^{\varepsilon}_T \boldsymbol{\Phi}_T$;
- $\sum_{t=1}^{\infty} t^{-\delta} E(\boldsymbol{\Phi}'_t \boldsymbol{\gamma}^{\varepsilon}_t \boldsymbol{\Phi}_t) < \infty$.

as well as Lemma 3.4.1 and the auxiliary Lemma 3.4.6.

Lemma 3.4.6 *Let $\boldsymbol{c} = (c_k)_{k=0}^{\infty}$ be a sequence from ℓ_2 and $a(z)$ be an analytical function with no zeros in the neighborhood of the unit circle. Denote*

$$c(z) = \sum_{k=0}^{\infty} c_k z^k, \quad \frac{1}{a(z)} = \sum_{k=0}^{\infty} \gamma_k z^k.$$

Then the sequence $(\boldsymbol{c}' \boldsymbol{A}^k \boldsymbol{B})_{k=0}^{\infty}$ contains the Taylor coefficients of $c(z)/a(z)$:

$$\frac{c(z)}{a(z)} = \sum_{k=0}^{\infty} \boldsymbol{c}' \boldsymbol{A}^k \boldsymbol{B} z^k$$

and

$$\sum_{k=0}^{\infty} |\boldsymbol{c}' \boldsymbol{A}^k \boldsymbol{B}|^2 = \int_{|z|=1} \frac{|c(z)|^2}{|a(z)|^2} dm(z) \leqslant M_a^{-2} \parallel \boldsymbol{c} \parallel^2,$$

where $M_a = \inf_{|z|=1} |a(z)|$ *is positive and* $dm(z) = dz/(2\pi iz)$ *means the normalized Lebesgue measure on the unit circle.*

The result in Theorem 3.4.3 would follow from Theorem 3.4.5 if the matrix $T^{-\delta}\check{\boldsymbol{R}}_T^{\varepsilon}$ is bounded away from 0 for $T > 0$. Below we derive and justify an estimate on the probability that the regularized information matrix $T^{-1}\check{\boldsymbol{R}}_T^{\varepsilon}$ is uniformly bounded from 0, where the regularizer $\boldsymbol{\Lambda}$ plays a crucial role. In fact, $\boldsymbol{\Lambda}$ guarantees the uniqueness of the minimum of the quadratic loss function and ensures the obtained estimates lying in ℓ_2 space. We denote $\hat{\boldsymbol{R}}_T^{\varepsilon} = T^{-1}\check{\boldsymbol{R}}_T^{\varepsilon}$.

Theorem 3.4.7 *For any* $\delta \in (0,1)$*, and* $\alpha > 0$*, there are positive constants* C_4 *and* T_1 *such that*

$$\forall T_0 > T_1,\ P\left\{\forall T \geqslant T_0,\quad \hat{\boldsymbol{R}}_T^{\varepsilon} \geqslant C_4 \boldsymbol{I}\right\} \geqslant 1 - \frac{\alpha}{T_0^{\delta}}. \tag{3.20}$$

The main idea of the proof of Theorem 3.4.7 is outlined as follows. The regularized information matrix $\hat{\boldsymbol{R}}_T^{\varepsilon}$ is divided into three parts such that every part is dominated either by a non-zero observation data set or by the coefficients of the regularizer $\boldsymbol{\Lambda}$. We derive the estimate for every part respectively, based on the state space form of $\boldsymbol{\Phi}_t$ in (3.18).

The infinite-dimensional matrix $\hat{\boldsymbol{R}}_T^{\varepsilon}$ is divided into three terms

$$\hat{\boldsymbol{R}}_T^{\varepsilon} = \frac{1}{T}\sum_{t=1}^{T} \boldsymbol{\Phi}_t \boldsymbol{\Phi}_t' + \frac{\varepsilon}{T}\boldsymbol{\Lambda} = \boldsymbol{Q}_{1,T} + \boldsymbol{Q}_{2,T} + \boldsymbol{Q}_{3,T}, \tag{3.21}$$

and each term is bounded from below as follows:

$$\begin{aligned}
\boldsymbol{Q}_{1,T} &= \frac{1}{T}\sum_{t=1}^{T} \boldsymbol{A}^N \boldsymbol{\Phi}_{t-N} \boldsymbol{\Phi}_{t-N}' \boldsymbol{A}^{*N} + \frac{\varepsilon_1}{T}\boldsymbol{\Lambda} \geqslant \frac{\varepsilon_1}{T}\boldsymbol{\Lambda}, \\
\boldsymbol{Q}_{2,T} &= 2Re \sum_{j=0}^{N-1} \boldsymbol{A}^N \left(\frac{1}{T}\sum_{t=1}^{T} \boldsymbol{\Phi}_{t-N}\epsilon_{t-j}\right) \boldsymbol{B}' \boldsymbol{A}^{*j} + \frac{\varepsilon_2}{T}\boldsymbol{\Lambda} \geqslant -q_{2,T}\boldsymbol{I}, \\
\boldsymbol{Q}_{3,T} &= \sum_{j=0}^{N-1}\sum_{i=0}^{N-1} \boldsymbol{A}^j \boldsymbol{B} \left(\frac{1}{T}\sum_{t=1}^{T} \epsilon_{t-j}\epsilon_{t-i}\right) \boldsymbol{B}' \boldsymbol{A}^{*i} \geqslant \sigma^2 \boldsymbol{U}_N - q_{3,T}\boldsymbol{I}.
\end{aligned} \tag{3.22}$$

where

- $\boldsymbol{U}_N = \sum_{i=0}^{N-1} \boldsymbol{A}^i \boldsymbol{B}\boldsymbol{B}' \boldsymbol{A}^{*i}$;
- $\varepsilon_1, \varepsilon_2 > 0$ are chosen such that $\varepsilon_1 + \varepsilon_2 = \varepsilon$;
- $Re\boldsymbol{X} = (\boldsymbol{X}' + \boldsymbol{X})/2$ for any square matrix $\boldsymbol{X}$;
- $\boldsymbol{I}$ is identity operator, $\boldsymbol{A}^*$ is the complex conjugate transpose of $\boldsymbol{A}$;
- $q_{2,T}$ and $q_{3,T}$ are defined in Lemma 3.4.8 and Lemma 3.4.9.

Lemma 3.4.8 *There exist $C_5, C_6 > 0$ such that for any $\beta > 0$, $\lambda \in (0, 1-\delta)$ and $T_0 > 0$ if $N(T) \leqslant C_6 \beta T^{(1-\delta-\lambda)/2}$ for all $T \geqslant T_0$, then*

$$P\left\{\forall T \geqslant T_0, \quad q_{2,T} \leqslant \frac{C_5 N^2(T)}{\beta T^{(1-\delta-\lambda)/2}}\right\} \geqslant 1 - \frac{\beta^2}{\lambda T_0^{\delta}}. \tag{3.23}$$

Lemma 3.4.9 *1. There exist $C_7 > 0$ such that for any $\beta > 0$, $\lambda \in (0, 1-\delta)$ and $T_0 > 0$*

$$P\left\{\forall T \geqslant T_0, \quad |q_{3,T}| \leqslant \frac{C_7 N^2(T)}{\beta T^{(1-\delta-\lambda)/2}}\right\} \geqslant 1 - \frac{\beta^2}{\lambda T_0^{\delta}}. \tag{3.24}$$

2. There exist an integer K and real $d > 0$ such that for any $N \geqslant K$

$$\boldsymbol{P}_{N-K} \sum_{i=0}^{N-1} \boldsymbol{A}^i \boldsymbol{B}\boldsymbol{B}' \boldsymbol{A}^{*i} P_{N_K} \geqslant d\boldsymbol{P}_{N-K}, \tag{3.25}$$

where $\boldsymbol{P}_k$ is a projector, $\boldsymbol{P}_k \boldsymbol{c} = (c_0, c_1, \ldots, c_{k-1}, 0, 0, \ldots)$ for $\boldsymbol{c} = (c_0, c_1, \ldots)$.

Chapter 4

Regularized AR Frequency Estimation

Many real life periodic time series can be modeled as a sum of sinusoids and noise. The frequencies hidden in such periodic processes can be often detected by approximating the generalized spectral density of the observed process by an autoregressive (AR) model. In this chapter, we apply the Regularized AR (RAR) approximation method introduced in Chapter 3 to the frequency detection problem. Due to the properties of the RAR method, the repeated model selections and parameter estimations are avoided as the observed sample increases. We show that the RAR estimates of frequencies are strictly consistency and asymptotic normally distributed. The presented numerical examples confirm validity of RAR approximation. In addition, we propose a robust trimming algorithm for RAR frequency estimation which aims to minimize the effect of outliers in frequency estimates. Our simulation studies indicate that such robust trimming of frequency estimates can effectively eliminate the spurious and atypical frequency estimates, and therefore noticeably increase the accuracy.

4.1 Introduction

Consider a time series $\{y_t, t \in T\}$ observed from a periodic process

$$y_t = x_t + \epsilon_t, \quad \text{and} \quad x_t = \rho \cos(t\omega_0 + \phi), \tag{4.1}$$

where

- ρ and ω_0 are constants with $\rho > 0$, $\omega_0 \in (0, \pi)$;
- ϕ is a random variable with uniform distribution on $[0, 2\pi)$;
- $\{\epsilon_t\} \sim IID(0, \sigma^2)$ and $E(\epsilon_t^4) = \eta\sigma^4 < \infty$, i.e., $\{\epsilon_t\}$ is a strictly stationary process with continuous spectral density (equal to constant) and finite fourth moment;
- $\{\epsilon_t\}$ is independent of ϕ and hence of $\{x_t\}$.

The structure of y_t is often referred to as the "signal-plus-noise" model, and our interest lies in estimating the frequency ω_0 based on the observed sample $\{y_t, t \in T\}$. The frequency estimation problem is an important subject in both statistical signal processing and time series analysis (Priestley, 1981; Kay, 1988). In the late 18^{th} century, Prony (1795) devised a procedure for fitting exponential models to chemical data to extract the sinusoidal signals; Schuster (1898) was the first one who proposed periodogram in an attempt to find "hidden periodicities" of sun-spot data. After these germinating approaches, substantial studies are conducted on the frequency estimation problem. Most traditional approaches to this problem are based on the Fourier transformation. Walker (1971) introduced periodogram maximization (PM) method, which locates the local maxima in the periodogram as a continuous function of a frequency variable; Hannan (1971, 1973), Rao and Zhao (1993) discussed an approach called nonlinear least square (NLS), which fits a sum of sinusoids to the process by minimizing the sum of squared errors with respect to ρ, ω_0 and ϕ. Both of the PM and NLS frequency estimates are strongly consistent with asymptotical standard deviation $O(T^{-3/2})$. Despite their high accuracy, the utilized non-linear estimation procedures are very computationally intensive and require initial values of accuracy $O(T^{-1})$ in order to converge to the optimal solution (Rice and Roseblatt, 1988).

In application, if the computational burden is not the primary concern, we may apply these classical approaches to acquire satisfactory frequency estimates with high accuracy. However, in real-time (online) frequency estimation, such computational intensive approaches as PM and NLS are not feasible. Our objective is to develop frequency estimation procedures with high accuracy but low computational burden, which is useful in online applications.

The iterative filtering (IF) procedures that utilize AR filters (Kay, 1984; Dragošević and Stanković, 1989; Nehorai, 1985; Stoica and Nehorai, 1988; Tichavský and Händel, 1995) are widely applied in engineering applications due to its effectiveness and simplicity. Kay (1984) proposes an iterative filtering algorithm (IFA) based on an all-pole (AR) filter which is identified by Burg's method. The IFA provides very good estimates even for relatively low signal to noise ratio (SNR). However, since IFA uses a filter whose poles are on the unit circle, the bandwidth is extremely narrow and the iterative procedure requires precise initial values. In order to overcome such disadvantages, Dragošević and Stanković (1989) propose the generalized least square (GLS) approach, which utilizes an all-pole filter with an extra parameter to force the poles to be within the unit circle and applies iterative LS to estimate its parameters. Both the IFA and GLS estimators are asymptotically inconsistent. Kedem (1994) introduces a method of parametric filtering (PF) which unifies and extends ideas of IFA and GLS, and shows that the inconsistency can be eliminated by using an appropriate parameterized filter. Such filter can be chosen on the basis of the bias yielded by the Prony estimator. The IF procedures usually require precise initial values, which is not adequately addressed in current literature.

Other commonly used frequency estimation techniques include ARMA-based and eigenanalysis methods. The ARMA based approach utilizes the fact that $\{y_t\}$ satisfies an ARMA$(2, 2)$ equation. Therefore, various existing ARMA estimation techniques can be applied to detecting frequency (Cadzow, 1980, 1982). Among such widely accepted estimation procedures are, for example, the over-determined Yule-Walker (YW) estimator (Friedlander, 1984) and high order YW estimator (Chan and Langford, 1982) which provide advantages of numerical simplicity and asymptotic properties (Stoica and Söderström, 1989). The eigenanalysis methods, such as Pisarenko's harmonic decomposition (Pisarenko, 1972) and the extended Prony method (Kay and Marple, 1981), obtain an eigenvector

with the minimum eigenvalue of a suitably chosen matrix, and then identify the sinusoidal frequency by taking the roots of the polynomial which has the components of the eigenvector as coefficients. However, the performance of the eigenanalysis estimator substantially depends on SNR. A detailed review of frequency estimation methods can be found, for example, in Kay and Marple (1981), Brillinger (1987), Kay(1988) as well as Quinn and Hannan (2001).

In this thesis, we focus on the AR-based methods for frequency estimation. After Yule (1927) proposed to fit an AR(2) model to periodic sunspot data, a considerable literature is developed on using AR models for detection of unknown frequencies (Makhoul, 1975; Kay and Marple, 1981, Stoica and Söderström, 1983, 1987; Truft and Kumaresan, 1982; Mackisack and Poskitt, 1989, 1990; Li et al., 1994; Hannan and Quinn, 2001). Due to its conceptual and numerical simplicity, the AR-based frequency estimation is widely accepted and employed in applied data analysis.

A remarkable feature of $\{y_t\}$ is that the spectral distribution

$$F(\theta) = \sigma^2(\theta + \pi)/(2\pi) + \rho^2 H(\theta - \omega_0)/2, \tag{4.2}$$

where $\theta \in [-\pi, \pi]$ and $H(.)$ denotes the Heaviside function, has jump discontinuities of height $\rho^2/2$ at frequency $\theta = \omega_0$. Therefore its derivative is called "generalized" spectral density. The idea of the AR approach is to estimate the generalized spectral density of $\{y_t\}$, $f(\theta)$, using an AR equation

$$a(z) = 1 + a_1 z + \ldots + a_k z^k, \tag{4.3}$$

The frequency is then obtained by finding the location of the largest peak in the k-th order AR approximation of the generalized spectral density $f_k(z)$, or equivalently, by taking the phase angle of the zero of $a(z)$ closest to the unit circle.

Properties of the AR-based frequency estimates are studied empirically and theoretically. For example, Sakai (1979) empirically shows that the variance of AR-based estimates is inversely proportional to both data length and the square of SNR; Mackisack and Poskitt (1989) illustrate via simulation studies that the standard deviation of the AR-based estimates has a similar order as that of the periodogram estimates and the NLS estimates; Stoica et al. (1989), Mackisack and Poskitt (1989, 1990) as well as Li et al.(1994) prove that the AR-based estimates of frequency are strongly consistent and asymptotically normal.

A common procedure of AR-based frequency estimation includes three steps (see Mackisack and Poskitt, 1991):

- Step 1: select the order k of an AR model (4.3) by information criterions such as AIC, BIC, and PLS;
- Step 2: estimate the AR coefficients by YW, LS or other methods;
- Step 3: the frequency ω_0 is estimated by finding the minimum of the transfer function $\hat{h}_k(\theta) = \left|\hat{a}(e^{i\theta})\right|^2 = \left|\sum_{j=0}^{k} \hat{a}_j(e^{ij\theta})\right|^2$ in $(0, \pi)$, where

$$\hat{a}(z) = 1 + \hat{a}_1 z + \ldots + \hat{a}_k z^k, \tag{4.4}$$

  and $\hat{a}_j$, $j = 1, \ldots, k$, are the sample AR parameter estimates.

The corresponding spectral density estimate is $\hat{f}_k(\theta) = \sigma^2/\{2\pi^2 \hat{h}_k(\theta)\}$. However, in many electrical engineering, astronomical and biomedical applications the length of the observations is not known a priori and may indefinitely increase while frequency detection is performed in real-time (or online). Therefore, the Steps 1 and 2 have to be re-conducted upon arrival of every new observation. In order to avoid the repeated model selection and parameter estimation, we utilize the recursive Regularized Least Squares (RLS) method, introduced in Chapter 3, for estimating AR parameters in (4.3) and name the resulting procedure the Regularized AR-based (RAR) frequency detection. We validate our findings by comparing to the results on AR-based frequency detection, presented by Mackisack and Poskitt (1989).

The key idea of the RAR approach is the same as for the traditional AR-based frequency estimation methods, but the implementation differs. Instead of using the YW method, the RLS method is applied to estimate the AR coefficients and thus Step 1 of the traditional procedure is omitted. As discussed in Chapter 3, the employment of the RLS method enables to estimate the coefficients with different level of accuracy, which is controlled by a regularizer. In particular, the first few model coefficients are estimated more precisely than the tail ones, and the number of estimated parameters grow with the sample size. Hence, the repeated model selection and parameter estimation are avoided as the observed sample increases. In Step 3, we follow an alternative method of minimizing the transfer

function (Stoica et al., 1989; Truft and Kumaresan, 1982). In particular, the estimated frequency $\hat{\omega}_k$ is determined by the argument of zeros of $\hat{a}(z)$ with modulus closest to the unit circle.

The RAR frequency estimation procedure is outlined as follows:

- Step 1: fit a "long" AR(k) model to the process $\{y_t\}$ and then apply the RLS to estimate the AR coefficients. Note that the model order can be adjusted by changing the RLS parameters ε and μ.
- Step 2: find all the roots of $\hat{a}(z)$ (4.4). The estimated frequency $\hat{\omega}_k$ is determined by the argument of the pair of roots with modulus closest to the unit circle.

Our theoretical findings indicate that the RAR estimates of unknown frequency are strongly consistent, i.e., converge almost surely, and asymptotically normally distributed, as shown below.

4.2 Asymptotic Properties of the RAR Frequency Estimate

As discussed in Chapter 3, let us rewrite an AR model in a state-space form

$$y_t = \boldsymbol{\Phi}'_{t-1}\boldsymbol{\tau}_T + \epsilon_t. \tag{4.5}$$

where $\boldsymbol{\Phi}_{t-1} = (y_{t-1}, y_{t-2}, \ldots, y_1, 0, \ldots)'$ and $\boldsymbol{\tau}_T = -(a_1, a_2, \ldots)'$ is an vector of unknown AR parameters and possibly infinite dimensional. We estimate $\boldsymbol{\tau}_T$ by the recursive LS method (Abraham and Ledolter, 1983; Fomin, 1999) which satisfy the Kalman filter equations

$$\begin{aligned}
\hat{\boldsymbol{\tau}}_{T+1} &= \hat{\boldsymbol{\tau}}_T + \boldsymbol{\gamma}^{\varepsilon}_T\boldsymbol{\Phi}_T(1 + \boldsymbol{\Phi}'_{T+1}\boldsymbol{\gamma}^{\varepsilon}_T\boldsymbol{\Phi}_{T+1})^{-1}(y_{T+1} - \boldsymbol{\Phi}'_T\hat{\boldsymbol{\tau}}_T) \\
\boldsymbol{\gamma}^{\varepsilon}_{T+1} &= \boldsymbol{\gamma}^{\varepsilon}_T - \boldsymbol{\gamma}^{\varepsilon}_T\boldsymbol{\Phi}_{T+1}(1 + \boldsymbol{\Phi}'_{T+1}\boldsymbol{\gamma}^{\varepsilon}_T\boldsymbol{\Phi}_{T+1})^{-1}\boldsymbol{\Phi}'_{T+1}\boldsymbol{\gamma}^{\varepsilon}_T
\end{aligned} \tag{4.6}$$

with initial conditions $\hat{\boldsymbol{\tau}}_0 = 0$ and $\boldsymbol{\gamma}^{\varepsilon}_0 = (\varepsilon\boldsymbol{\Lambda})^{-1}$. The matrix $\boldsymbol{\gamma}^{\varepsilon}_T$ is inverse to the sample information matrix $\check{\boldsymbol{R}}^{\varepsilon}_T$, i.e. $\boldsymbol{\gamma}^{\varepsilon}_T = (\check{\boldsymbol{R}}^{\varepsilon}_T)^{-1}$, where $\check{\boldsymbol{R}}^{\varepsilon}_T = \check{\boldsymbol{R}}_T + \varepsilon\boldsymbol{\Lambda}$. Here $\check{\boldsymbol{R}}_T = \sum_{t=1}^{T}\boldsymbol{\Phi}_t\boldsymbol{\Phi}'_t$ and $\boldsymbol{\Lambda} = \mathrm{diag}\{\mathrm{e}^{\mu\mathrm{k}}\}_{\mathrm{k}=1}^{\infty}$ is the regularizer.

To prove strict consistency of the RAR frequency estimates, we follow a similar derivation plan as for the RLS estimates of AR(∞) models with exponentially decaying coefficients (Gel and Barabanov, 2007). (A particular case of such models includes ARMA processes and is discussed in Chapter 3.) However, when the RAR approximation is applied to estimation of ARMA models (see Chapter 3), the power series $a(z) = 1+a_1z+a_2z^2+\ldots$ is assumed to have no zeros in certain neighborhood of the unit circle, which is not valid in frequency estimation problem. In order to analyze asymptotics of the RAR estimates of unknown frequency, we need to update this assumption and revise the proofs from Chapter 3.

We start from an analogue to Theorem 3.4.5. Assume that $a(z)$ may have a finite number of zeros in a certain neighborhood of the unit circle. The quadratic function V_{T+1}, defined as

$$V_{T+1} = T^{-\delta}(\hat{\boldsymbol{\tau}}_{T+1} - \boldsymbol{\tau}_{T+1})'\check{\boldsymbol{R}}^{\varepsilon}_{T}(\hat{\boldsymbol{\tau}}_{T+1} - \boldsymbol{\tau}_{T+1}),\ T = 1, 2, ... \tag{4.7}$$

converges almost surely to zero under the new assumptions that some roots of $a(z)$ are located along the unit circle, as stated in the next theorem.

Theorem 4.2.1 *Let $a(z)$ have finite number of roots along the unit circle, then $V_T \to 0$ as $T \to \infty$ with probability 1. Moreover, there exists a positive constant C_2 such that*

$$\forall \alpha > 0, \quad P\left\{\forall T \geq 0, \quad V_{T+1} \leqslant \frac{C_2}{\alpha}\right\} \geq 1 - \alpha. \tag{4.8}$$

Proof We follow the same approach as the proof of Theorem 3.4.5, which is given in Gel and Barabanov (2007).

Denote the estimation error $\Delta\hat{\boldsymbol{\tau}}_{T+1} = \hat{\boldsymbol{\tau}}_{T+1} - \boldsymbol{\tau}_{T+1}$, then

$$\Delta\hat{\boldsymbol{\tau}}_{T+1} = \Delta\hat{\boldsymbol{\tau}}_{T} + \boldsymbol{\gamma}^{\varepsilon}_{T}\boldsymbol{\Phi}_{T}(\epsilon_{T+1} - \boldsymbol{\Phi}'_{T}\Delta\hat{\boldsymbol{\tau}}_{T}) = \boldsymbol{\gamma}^{\varepsilon}_{T}\check{\boldsymbol{R}}^{\varepsilon}_{T-1}\Delta\hat{\boldsymbol{\tau}}_{T} + \boldsymbol{\gamma}^{\varepsilon}_{T}\boldsymbol{\Phi}_{T}\epsilon_{T+1}, \tag{4.9}$$

which consists of two uncorrelated terms. Since $\boldsymbol{\gamma}^{\varepsilon}_{T}$ decays as T increases, the conditional expectation of $V_{T+1} = T^{-\delta}\Delta\hat{\boldsymbol{\tau}}'_{T+1}\check{\boldsymbol{R}}^{\varepsilon}_{T}\Delta\hat{\boldsymbol{\tau}}_{T+1}$ has the following upper bound

$$E(V_{T+1}|\mathcal{F}_T) = \frac{1}{T^{\delta}}\Delta\hat{\boldsymbol{\tau}}'_{T}\check{\boldsymbol{R}}^{\varepsilon}_{T-1}\boldsymbol{\gamma}^{\varepsilon}_{T}\check{\boldsymbol{R}}^{\varepsilon}_{T-1}\Delta\hat{\boldsymbol{\tau}}_{T} + \frac{\sigma}{T^{\delta}}\boldsymbol{\Phi}'_{T}\boldsymbol{\gamma}^{\varepsilon}_{T}\boldsymbol{\Phi}_{T} \leqslant V_T + \frac{\sigma}{T^{\delta}}\boldsymbol{\Phi}'_{T}\boldsymbol{\gamma}^{\varepsilon}_{T}\boldsymbol{\Phi}_{T}. \tag{4.10}$$

To apply Lemma 3.4.1, we need to show that $\sum_{t=1}^{\infty} t^{-\delta} E\boldsymbol{\Phi}_t' \boldsymbol{\gamma}_T^{\varepsilon} \boldsymbol{\Phi}_t < \infty$. Denote

$$\tilde{\boldsymbol{\Phi}}_t = (\varepsilon \boldsymbol{\Lambda})^{-1/2} \boldsymbol{\Phi}_t, \quad \tilde{\boldsymbol{R}}_T = \sum_{t=1}^{T} \tilde{\boldsymbol{\Phi}}_t \tilde{\boldsymbol{\Phi}}_t' + \boldsymbol{I}, \quad \tilde{\boldsymbol{\gamma}}_T = \tilde{\boldsymbol{R}}_T^{-1}. \tag{4.11}$$

Extend the function $\boldsymbol{\Phi}(t)$ of $\boldsymbol{\Phi}(t) = \boldsymbol{\Phi}_t$ for integer t on the positive semi-axis as a step function on every interval $(t, t+1]$. Consequently, $\tilde{\boldsymbol{R}}(t) = \int_0^t \tilde{\boldsymbol{\Phi}}(s)\tilde{\boldsymbol{\Phi}}(s)' ds$ and $\tilde{\boldsymbol{\gamma}}(s) = \tilde{\boldsymbol{R}}(s)^{-1}$. Then

$$\begin{aligned} \sum_{t=1}^{T} E\boldsymbol{\Phi}_t' \boldsymbol{\gamma}_t^{\varepsilon} \boldsymbol{\Phi}_t = \sum_{t=1}^{T} E\tilde{\boldsymbol{\Phi}}_t' \tilde{\boldsymbol{\gamma}}_t \tilde{\boldsymbol{\Phi}}_t \leqslant \int_0^T E\tilde{\boldsymbol{\Phi}}'(s) \tilde{\boldsymbol{R}}^{-1}(s) \tilde{\boldsymbol{\Phi}}(s) ds \qquad (1) \\ = TrE(\ln \tilde{\boldsymbol{R}}_T - \ln \tilde{\boldsymbol{R}}_0) \leqslant Tr(\ln E\tilde{\boldsymbol{R}}_T - \ln E\tilde{\boldsymbol{R}}_0) \quad (2) \\ = Tr \ln E\tilde{\boldsymbol{R}}_T \end{aligned} \tag{4.12}$$

Here, inequality (1) follows from the condition that $\tilde{\boldsymbol{\gamma}}_s \geqslant \tilde{\boldsymbol{\gamma}}_t, \ s \leqslant t$; inequality (2) follows from concavity of logarithm and the Jensen's inequality. In addition, we take into account that $\ln \tilde{\boldsymbol{R}}_0 = \ln \boldsymbol{I} = 0$.

The matrix $\tilde{\boldsymbol{R}}_T$ may be expressed as $\tilde{\boldsymbol{R}}_T = \mathrm{diag}\{\widehat{\boldsymbol{R}}_{\mathrm{T}}, \boldsymbol{I}\}$ where $\widehat{\boldsymbol{R}}_T = (\widehat{\boldsymbol{R}}_{ij})_1^n$ is a symmetric positive definite $T \times T$ matrix, and therefore $\det \widehat{\boldsymbol{R}}_{\mathrm{T}} \leqslant \prod_{\mathrm{i}=1}^{\mathrm{T}} (\widehat{\boldsymbol{R}})_{\mathrm{ii}}$ holds:

$$\det \widehat{\boldsymbol{R}}_{\mathrm{T}} \leqslant \prod_{\mathrm{i}=1}^{\mathrm{T}} (\widehat{\boldsymbol{R}})_{\mathrm{ii}} \tag{4.13}$$

Substituting this inequality and the identities $\mathrm{Tr} \ln \mathrm{E}\tilde{\boldsymbol{R}}_{\mathrm{T}} = \mathrm{Tr} \ln \mathrm{E}\widehat{\boldsymbol{R}}_{\mathrm{T}}$ and $\mathrm{Tr} \ln \boldsymbol{X} = \ln \det \boldsymbol{X}$ into (4.12), then

$$\begin{aligned} \sum_{t=1}^{\infty} E\tilde{\boldsymbol{\Phi}}_t' \tilde{\boldsymbol{\gamma}}_t \tilde{\boldsymbol{\Phi}}_t \leqslant \ln \det E\tilde{\boldsymbol{R}}_T \leqslant \sum_{i=1}^{T} \ln\{E\tilde{\boldsymbol{R}}_T\}_{ii} \\ = \sum_{i=1}^{T} \ln E\left(1 + \frac{e^{-\mu i}}{\varepsilon} \sum_{k=1}^{T+1-i} y_k^2\right) \leqslant \sum_{i=1}^{T} \ln(1 + \frac{e^{-\mu i}}{\varepsilon} \sum_{k=1}^{T+1-i} Ey_k^2\Big). \end{aligned} \tag{4.14}$$

Recall $\boldsymbol{A}$ and $\boldsymbol{B}$ defined in (3.18),

$$\boldsymbol{A} = \begin{pmatrix} -a_1 & -a_2 & -a_3 & \dots \\ 1 & 0 & 0 & \dots \\ 0 & 1 & 0 & \dots \\ \vdots & \vdots & \ddots & \end{pmatrix}, \quad \boldsymbol{B} = \begin{pmatrix} 1 \\ 0 \\ 0 \\ \vdots \end{pmatrix},$$

then

$$\begin{aligned} Ey_k^2 &\leqslant \sigma^2 C \sum_{i=0}^{k} (\boldsymbol{A}^i \boldsymbol{B})^2 = \sigma^2 C \sum_{i=0}^{k} \|\boldsymbol{A}^i\| \|\boldsymbol{B}\| = \sigma^2 C \sum_{i=0}^{k} i^{\rho-1} \\ &\approx \sigma^2 C \int_0^k x^{\rho-1} dx = \sigma^2 C x^\rho \Big|_0^k = \sigma^2 C k^\rho. \end{aligned} \tag{4.15}$$

Thus

$$\sum_{k=1}^{T} Ey_k^2 \leqslant C\sigma^2 \sum_{k=1}^{T} k^\rho \approx C\sigma^2 \int_1^T x^\rho dx = C\sigma^2 x^{\rho+1} \Big|_1^T \leqslant C\sigma^2 T^{\rho+1} = C_1 T^{\rho+1}. \tag{4.16}$$

Therefore

$$\begin{aligned} \sum_{i=1}^{T} \ln\left(1 + \frac{e^{-\mu i}}{\varepsilon} \sum_{k=1}^{T} Ey_k^2\right) &\leqslant \sum_{i=1}^{T} \ln\left(1 + \frac{e^{-\mu i}}{\varepsilon} C_1 T^{\rho+1}\right) \leqslant \sum_{i=1}^{\infty} \ln\left(1 + \frac{e^{-\mu i}}{\varepsilon} C_1 T^{\rho+1}\right) \\ &\leqslant \int_0^\infty \ln\left(1 + \frac{e^{-\mu x}}{\varepsilon} C_1 T^{\rho+1}\right) dx. \end{aligned} \tag{4.17}$$

Denote $y = \varepsilon^{-1} C_1 T^{\rho+1} e^{-\mu x}$, then

$$\begin{aligned} \int_0^\infty \ln\left(1 + \frac{e^{-\mu x}}{\varepsilon} C_1 T^{\rho+1}\right) dx &= \mu^{-1} \int_0^{\varepsilon^{-1} C_1 T^{\rho+1}} \frac{\ln(1+y)}{y} dy \\ &\sim \mu^{-1} (\ln T^{\rho+1})^2 = \mu^{-1} (\rho+1)^2 \ln^2 T \end{aligned} \tag{4.18}$$

as $T \to \infty$. Hence, there exists C_2 such that

$$\sum_{t=1}^{T} E \boldsymbol{\Phi}_t' \boldsymbol{\gamma}_t^\varepsilon \boldsymbol{\Phi}_t \leqslant C_2 \ln^2 T, \quad T = 1, 2, ... \tag{4.19}$$

Finally, the convergence of the series under consideration is established using the Abel transformation

$$\sum_{t=1}^{T} \frac{1}{t^\delta} E \boldsymbol{\Phi}_t' \boldsymbol{\gamma}_t^\varepsilon \boldsymbol{\Phi}_t = \sum_{t=1}^{\infty} \left(\frac{1}{t^\delta} - \frac{1}{(t+1)^\delta}\right) \sum_{k=1}^{t} E \boldsymbol{\Phi}_k' \boldsymbol{\gamma}_k^\varepsilon \boldsymbol{\Phi}_k \leqslant \sum_{t=1}^{\infty} \frac{C_2 \delta \ln^2 t}{t^{\delta+1}} = C_3 < \infty. \tag{4.20}$$

In the view of inequality (4.10) and the stochastic variable V_T satisfy all the conditions of Lemma 3.4.1. Thus,

$$\forall \alpha > 0 \quad P\left\{\forall T \geq 0, \quad V_{T+1} \leqslant \frac{C_3}{\alpha}\right\} \geq 1 - \alpha.$$

The stochastic variables V_T converges a.s. for all $\delta > 0$. Hence, the limit must be 0. □

By the result of Theorem (4.2.1),

$$\lim_{T\to\infty} T^{-\delta}(\hat{\boldsymbol{\tau}}_T - \boldsymbol{\tau}_T)' \check{\boldsymbol{R}}_T^{\varepsilon}(\hat{\boldsymbol{\tau}}_T - \boldsymbol{\tau}_T) = 0 \quad a.s. \tag{4.21}$$

in order to complete the proof of strong consistency of the RLS estimates $\hat{\boldsymbol{\tau}}_T$, we need to show that with probability 1,

$$T^{-\delta}\check{\boldsymbol{R}}_T^{\varepsilon} > 0. \tag{4.22}$$

The proof for the general case when $\boldsymbol{\tau}_T$ is of infinite dimension ($\boldsymbol{\tau}_T \in \ell^2$) is challenging and we leave (4.23) as a conjecture for now. Instead, we prove an analogous result for the truncated case. In particular, let $\boldsymbol{P}_k$ be an orthogonal projector. Then, the truncated vector of unknown coefficients is given by $\boldsymbol{\tau}_{k,T} = \boldsymbol{P}_k\boldsymbol{\tau}_T$ and the corresponding sample RLS estimates is $\hat{\boldsymbol{\tau}}_{k,T} = \boldsymbol{P}_k\hat{\boldsymbol{\tau}}_T$. Also, the truncated regularized information matrix is denoted as $\check{\boldsymbol{R}}_{k,T}^{\varepsilon} = \boldsymbol{P}_k\check{\boldsymbol{R}}_T^{\varepsilon}\boldsymbol{P}_k'$. Therefore, we need to show that $T^{-1}\check{\boldsymbol{R}}_{k,T}^{\varepsilon}$ is positive definite, i.e.,

$$T^{-1}(\check{\boldsymbol{R}}_{k,T} + \varepsilon\boldsymbol{\Lambda}_k) > 0, \tag{4.23}$$

where

- $\check{\boldsymbol{R}}_{k,T} = \boldsymbol{P}_k\check{\boldsymbol{R}}_T\boldsymbol{P}_k'$,
- $\boldsymbol{\Lambda}_k = \boldsymbol{P}_k\boldsymbol{\Lambda} = \begin{pmatrix} e^{\mu_1} & 0 & \dots & 0 \\ 0 & e^{\mu_2} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & e^{\mu_k} \end{pmatrix}$.

Let us introduce the following notations. Denote the theoretical autocovariance function (ACVF) by

$$r_j = E(y_t y_{t+j}), \quad j = 0, \pm 1, \dots, \tag{4.24}$$

which forms the covariance vectors

$$\begin{aligned} \boldsymbol{r}_k &= (r_1, \ldots, r_k)', \\ \boldsymbol{r}_{k,0} &= (r_0, r_1, \ldots, r_k)', \end{aligned} \tag{4.25}$$

and a $k \times k$-Toeplitz covariance matrix

$$\boldsymbol{R}_{k,T} = \begin{pmatrix} r_0 & r_1 & \ldots & r_{k-1} \\ r_1 & r_0 & \ldots & r_{k-2} \\ \vdots & \vdots & \ddots & \vdots \\ r_{k-1} & r_{k-2} & \ldots & r_0 \end{pmatrix}. \tag{4.26}$$

Denote the sample ACVF by

$$\hat{r}_j = \frac{1}{T} \sum_{t=1}^{T-j} y_t y_{t+j}, \quad j = 0, 1, \ldots, k, \quad \text{where } k \in Z \text{ and } 0 \leq k \leq T-1. \tag{4.27}$$

Correspondingly, sample covariance vectors are given by

$$\begin{aligned} \hat{\boldsymbol{r}}_k &= (\hat{r}_1, \ldots, \hat{r}_k)', \\ \hat{\boldsymbol{r}}_{k,0} &= (\hat{r}_0, \hat{r}_1, \ldots, \hat{r}_k)', \end{aligned} \tag{4.28}$$

and a sample $k \times k$-Toeplitz covariance matrix is

$$\hat{\boldsymbol{R}}_{k,T} = T^{-1} \check{\boldsymbol{R}}_{k,T} = \begin{pmatrix} \hat{r}_0 & \hat{r}_1 & \ldots & \hat{r}_{k-1} \\ \hat{r}_1 & \hat{r}_0 & \ldots & \hat{r}_{k-2} \\ \vdots & \vdots & \ddots & \vdots \\ \hat{r}_{k-1} & \hat{r}_{k-2} & \ldots & \hat{r}_0 \end{pmatrix}. \tag{4.29}$$

Applying the above notations, the inequality (4.23) takes the following form

$$\hat{\boldsymbol{R}}_{k,T} + T^{-1} \varepsilon \boldsymbol{\Lambda}_k > 0, \tag{4.30}$$

which is equivalent to

$$\hat{\boldsymbol{R}}_{k,T} - \boldsymbol{R}_{k,T} + T^{-1} \varepsilon \boldsymbol{\Lambda}_k + \boldsymbol{R}_{k,T} > 0. \tag{4.31}$$

In the rest of this chapter, we suppress the dependence on T in $\boldsymbol{R}_{k,T}$, $\hat{\boldsymbol{R}}_{k,T}$, $\check{\boldsymbol{R}}^{\varepsilon}_{k,T}$, $\boldsymbol{\tau}_{k,T}$ and $\hat{\boldsymbol{\tau}}_{k,T}$, which will be denoted respectively as $\boldsymbol{R}_k$, $\hat{\boldsymbol{R}}_k$, $\check{\boldsymbol{R}}^{\varepsilon}_k$, $\boldsymbol{\tau}_k$ and $\hat{\boldsymbol{\tau}}_k$ for the sake of compactness. Let us state the following result on strong consistency of the truncated RLS estimates $\hat{\boldsymbol{\tau}}_k$.

Theorem 4.2.2 *If $T \to \infty$ and $k \to \infty$ such that $k^{\frac{3}{2}}/T \to 0$, then $\hat{\boldsymbol{\tau}}_k \to \boldsymbol{\tau}_k$ almost surely.*

Proof By definition, for $j = 1, \ldots, k$,

$$\begin{aligned} r_j &= E(y_t y_{t+j}) \\ &= \rho^2 E\bigg(\cos(t\omega_0 + \phi)\cos[(t+j)\omega_0 + \phi]\bigg) + E(\epsilon_t \epsilon_{t+j}) \\ &= \rho^2 E\bigg(\frac{1}{2}[\cos(2t\omega_0 + j\omega_0 + 2\phi) + \cos(j\omega_0)]\bigg) + \delta_{j,0}\sigma^2 \\ &= \frac{\rho^2}{2}\cos(j\omega_0) + \frac{\rho^2}{2}\int_0^{2\pi} \cos(2t\omega_0 + j\omega_0 + 2\phi)\frac{1}{2\pi} d\phi + \delta_{j,0}\sigma^2 \\ &= \frac{\rho^2}{2}\cos(j\omega_0) + \delta_{j,0}\sigma^2. \end{aligned} \tag{4.32}$$

and

$$\begin{aligned} \hat{r}_j &= \frac{1}{T}\sum_{t=1}^{T-j} y_t y_{t+j} \\ &= \frac{1}{T}\sum_{t=1}^{T-j}\bigg\{ \frac{\rho^2}{2}\cos[(j+2t)\omega_0 + 2\phi] + \frac{\rho^2}{2}\cos(j\omega_0) \\ &\quad + \rho\cos[(t+j)\omega_0 + \phi]\epsilon_t + \rho\cos(t\omega_0 + \phi)\epsilon_{t+j} + \epsilon_t\epsilon_{t+j}\bigg\}. \end{aligned} \tag{4.33}$$

Consider an element of $\hat{\boldsymbol{R}}_k - \boldsymbol{R}_k$ for some $j = 0, 1, \ldots, k$:

$$\begin{aligned} \hat{r}_j - r_j &= \frac{j}{T}\frac{\rho^2}{2}\cos(j\omega_0) + \rho^2\cos[(T-1)\omega_0 + 2\phi]\frac{\sin[(T-j)\omega_0]}{2T\sin\omega_0} + \frac{1}{T}\sum_{t=1}^{T-j}\epsilon_t\epsilon_{t+j} \\ &\quad + \frac{1}{T}\sum_{t=1}^{T-j}\rho\cos[(t+j)\omega_0 + \phi]\epsilon_t + \frac{1}{T}\sum_{t=1}^{T-j}\rho\cos(t\omega_0 + \phi)\epsilon_{t+j} - \delta_{j,0}\sigma^2 \\ &= \frac{1}{T}\sum_{t=1}^{T} 2\rho\cos(j\omega_0)\cos(t\omega_0 + \phi)\epsilon_t + \frac{1}{T}\sum_{t=1}^{T}\big\{\epsilon_t\epsilon_{t-j} - \delta_{j,0}\sigma^2\big\} \\ &\quad - \frac{1}{T}\sum_{t=1}^{j}\rho\cos\big\{(t-j)\omega_0 + \phi\big\}\epsilon_t - \frac{1}{T}\sum_{t=T-j+1}^{T}\rho\cos\big\{(t+j)\omega_0 + \phi\big\}\epsilon_t \\ &\quad - \frac{1}{T}\sum_{t=-j+1}^{0}\epsilon_t\epsilon_{t+j} + O\big(\frac{j}{N}\big). \end{aligned} \tag{4.34}$$

Thus, we have

$$\big|(\hat{r}_j - r_j) - S_{j,T}\big| \leqslant \frac{1}{T}\sum_{t=1}^{j}\rho\big|\epsilon_t\big| + \frac{1}{T}\sum_{t=T-j+1}^{T}\rho\big|\epsilon_t\big| + \frac{1}{T}\sum_{t=-j+1}^{0}\rho\big|\epsilon_t\epsilon_{t+j}\big| + O\big(\frac{j}{N}\big), \quad (4.35)$$

where

$$S_{j,T} = \frac{1}{T}\sum_{t=1}^{T}\big\{2\rho\cos(j\omega_0)\cos(t\omega_0+\phi) + \epsilon_{t-j}\big\}\epsilon_t - \delta_{j,0}\sigma^2. \quad (4.36)$$

Since ρ is a constant and $\{\epsilon_t\}$ is assumed to be white noise with finite fourth moment, the four terms on the right-hand side of (4.35) are all $O(j/N)$ a.s. Therefore, for $j = 0, 1, \ldots, k$

$$\hat{r}_j - r_j = S_{j,T} + O(j/N). \quad (4.37)$$

Let $\boldsymbol{S}_T$ be a matrix with elements $S_{l-j,T}$, $j = 1, \ldots, k$ and $l = 1, \ldots, k$, i.e.,

$$\boldsymbol{S}_T = \begin{pmatrix} S_{0,T} & S_{1,T} & \ldots & S_{k-1,T} \\ S_{1,T} & S_{0,T} & \ldots & S_{k-2,T} \\ \vdots & \vdots & \ddots & \vdots \\ S_{k-1,T} & S_{k-2,T} & \ldots & S_{0,T} \end{pmatrix}. \quad (4.38)$$

Also, let $\boldsymbol{E}_T$ be a matrix with elements $O\big((j-l)/T\big)$, $j = 1, \ldots, k$ and $l = 1, \ldots, k$, i.e.,

$$\boldsymbol{E}_T = \frac{1}{T}\begin{pmatrix} 0 & 1 & \ldots & k-1 \\ 1 & 0 & \ldots & k-2 \\ \vdots & \vdots & \ddots & \vdots \\ k-1 & k-2 & \ldots & 0 \end{pmatrix}. \quad (4.39)$$

Thus, we have

$$\hat{\boldsymbol{R}}_k - \boldsymbol{R}_k = \boldsymbol{S}_T + \boldsymbol{E}_T. \quad (4.40)$$

Firstly, let us investigate the asymptotic behavior of $\boldsymbol{S}_T$. By definition, $TS_{i,T}$ can be expressed as

$$TS_{i,T} = \sum_{s=1}^{T} X_{i,s}, \quad (4.41)$$

where $X_{i,s} = \{2\rho\cos(i\omega_0)\cos(s\omega_0+\phi)+\epsilon_{s-i}\}\epsilon_s - \delta_{i,0}\sigma^2$. We can show that $X_{i,s}$'s are uncorrelated and have finite variance.

$$\begin{aligned} E\big[X_{i,s}\big] &= E\big[\{2\rho\cos(i\omega_0)\cos(s\omega_0+\phi)+\epsilon_{s-i}\}\epsilon_s - \delta_{i,0}\sigma^2\big] \\ &= E\big[2\rho\cos(iw_0)\cos(sw_0+\phi)\epsilon_s\big] + E\big[\epsilon_{s-i}\epsilon_s\big] - \delta_{i,0}\sigma^2 \\ &= E\big[2\rho\cos(iw_0)\cos(sw_0+\phi)\big]E\big[\epsilon_s\big] + E\big[\epsilon_{s-i}\epsilon_s\big] - \delta_{i,0}\sigma^2 \\ &= 0 + \delta_{j,0}\sigma^2 - \delta_{j,0}\sigma^2 \\ &= 0. \end{aligned} \tag{4.42}$$

$$\begin{aligned} \mathrm{Var}\big[\mathrm{X_{i,s}}\big] &= E\big[\{[2\rho\cos(iw_0)\cos(sw_0+\phi)+\epsilon_{s-i}]\epsilon_s - \delta_{i,0}\sigma^2\}^2\big] \\ &= 4\rho^2 E\big[\cos^2(iw_0)\cos^2(sw_0+\phi)\epsilon_s^2\big] \\ &+4\rho E\big[\cos(iw_0)\cos(sw_0+\phi)\epsilon_s\epsilon_i\epsilon_{s-i}\big] \\ &-4\rho\delta_{i,0}\sigma^2 E\big[\cos(iw_0)\cos(sw_0+\phi)\epsilon_s\big] \\ &+E\big[\epsilon_{s-i}^2\epsilon_s^2\big] - 2\delta_{i,0}\sigma^2 E\big[\epsilon_{s-i}\epsilon_s\big] + (\delta_{i,0}\sigma^2)^2. \end{aligned} \tag{4.43}$$

Here we consider two cases depending on i.

- Case a: let $i = 0$.

$$\begin{aligned} \mathrm{Var}\big[\mathrm{X_{i,s}}\big] &= 4\rho^2 E\big[\cos^2(sw_0+\phi)\big]E\big[\epsilon_s^2\big] + 4\rho E\big[\cos(sw_0+\phi)\big]E\big[\epsilon_0\big]E\big[\epsilon_s^2\big] \\ &-4\rho\sigma^2 E\big[\cos(sw_0+\phi)\big]E\big[\epsilon_s\big] + E\big[\epsilon_s^4\big] - 2\sigma^2 E\big[\epsilon_s^2\big] + \sigma^4 \\ &= 4\rho^2 E\big[\cos^2(sw_0+\phi)\big]\sigma^2 + E\big[\epsilon_s^4\big] - \sigma^4 < \infty. \end{aligned} \tag{4.44}$$

- Case b: let $i \neq 0$, i.e., $i = 1, 2, \ldots, k$,

$$\begin{aligned} \mathrm{Var}\big[\mathrm{X_{i,s}}\big] &= 4\rho^2\cos^2(iw_0)E\big[\cos^2(sw_0+\phi)\big]E\big[\epsilon_s^2\big] \\ &+4\rho\cos(iw_0)E\big[\cos(sw_0+\phi)\big]E\big[\epsilon_s\epsilon_i\epsilon_{s-i}\big] + E\big[\epsilon_{s-i}^2\epsilon_s^2\big] \\ &= 4\rho^2\cos^2(iw_0)E\big[\cos^2(sw_0+\phi)\big]\sigma^2 + \sigma^4 < \infty. \end{aligned} \tag{4.45}$$

By Cases (a) and (b), $E[X_{i,s}^2] < \infty$, for any $i = 0, \ldots, k$.

For $p \neq q$ and $p,\ q = 1, \ldots, T$,

$$\begin{aligned}
\mathrm{Cov}(\mathrm{X_{i,p}}, \mathrm{X_{i,q}}) \quad &= E\big[X_{i,p}X_{i,q}\big] \\
&= E\big[\{[2\rho\cos(iw_0)\cos(pw_0+\phi)+\epsilon_{p-i}]\epsilon_p - \delta_{i,0}\sigma^2\} \\
&\{[2\rho\cos(iw_0)\cos(qw_0+\phi)+\epsilon_{q-i}]\epsilon_q - \delta_{i,0}\sigma^2\}] \\
&= 4\rho^2\cos^2(iw_0)E\big[\cos(pw_0+\phi)\cos(qw_0+\phi)\big]E\big[\epsilon_p\big]E\big[\epsilon_q\big] \\
&+2\rho\cos(iw_0)E\big[\cos(qw_0+\phi)\big]E\big[\epsilon_p\epsilon_q\epsilon_{p-i}\big] \\
&-\delta_{i,0}\sigma^2 2\rho\cos(iw_0)E\big[\cos(qw_0+\phi)\big]E\big[\epsilon_q\big] \\
&+2\rho\cos(iw_0)E\big[\cos(pw_0+\phi)\big]E\big[\epsilon_p\epsilon_q\epsilon_{q-i}\big] \\
&-\delta_{i,0}\sigma^2 2\rho\cos(iw_0)E\big[\cos(pw_0+\phi)\big]E\big[\epsilon_p\big] + (\delta_{i,0}\sigma^2)^2 \\
&+E\big[\epsilon_p\epsilon_{p-i}\epsilon_q\epsilon_{q-i}\big] - \delta_{i,0}\sigma^2 E\big[\epsilon_q\epsilon_{q-i}\big] - \delta_{i,0}\sigma^2 E\big[\epsilon_p\epsilon_{p-i}\big] \\
&= E\big[\epsilon_p\epsilon_{p-i}\epsilon_q\epsilon_{q-i}\big] - (\delta_{i,0}\sigma^2)^2 \qquad (4.46)
\end{aligned}$$

Here we consider two cases depending on i.

- Case a': let $i = 0$.

$$E\big[\epsilon_p\epsilon_{p-i}\epsilon_q\epsilon_{q-i}\big] - (\delta_{i,0}\sigma^2)^2 = \sigma^4 - \sigma^4 = 0. \qquad (4.47)$$

- Case b': let $i \neq 0$. If $q = p - i$, then $q - i = p - 2i$, i.e., $p \neq q - i$,

$$E\big[\epsilon_p\epsilon_{p-i}\epsilon_q\epsilon_{q-i}\big] = E\big[\epsilon_q\epsilon_{p-i}\big]E\big[\epsilon_p\big]E\big[\epsilon_{q-i}\big] = 0. \qquad (4.48)$$

  Similar arguments hold if $p = q - i$.

By Cases (a') and (b'), $\mathrm{Cov}(\mathrm{X_{i,p}}, \mathrm{X_{i,q}}) = 0$, for any $i = 0, \ldots, k$. Therefore, $X_{i,s}$'s are uncorrelated and have finite variances.

Let $\upsilon_T^2 = \max\big\{\mathrm{Var}(\mathrm{X_{i,1}}), \ldots, \mathrm{Var}(\mathrm{X_{i,T}})\big\}$ and $\delta \in (0, 1/2)$. By Doob's inequality,

$$P\left\{\left|\frac{1}{T^{\frac{1}{2}+\delta}}\sum_{s=1}^{T} X_{i,s}\right| < \xi\right\} \geqslant 1 - \big\{T^{-2\delta}\max_{1\leqslant s\leqslant T}\mathrm{Var}(\mathrm{X_{i,s}})\big\}/\xi^2 = 1 - \upsilon_{\mathrm{T}}^2/(\mathrm{T}^{2\delta}\xi^2). \qquad (4.49)$$

Our goal is to find such ξ that $\xi \to 0$ and $\upsilon_T^2/(T^{2\delta}\xi^2) \to 0$ simultaneously when $T \to \infty$. Let $\xi = 1/\log T$, then

$$\xi \to 0 \ \text{ and } \ \frac{\upsilon_T^2}{T^{2\delta}\xi^2} = \frac{\upsilon_T^2}{T^{2\delta}/\log^2 T} = \frac{\upsilon_T^2 \log^2 T}{T^{2\delta}} \to 0, \quad \text{as } T \to \infty. \tag{4.50}$$

Hence, the following statement holds with probability 1 when $T \to \infty$,

$$P\left\{ \left| \frac{1}{T^{\frac{1}{2}+\delta}} \sum_{s=1}^{T} X_{i,s} \right| < \frac{1}{\log T} \right\} \to 1. \tag{4.51}$$

which implies that

$$\left| \frac{1}{T^{\frac{1}{2}+\delta}} \sum_{s=1}^{T} X_{i,s} \right| < \frac{c}{\log T}, \quad a.s. \tag{4.52}$$

Since $TS_{j-l,T} = \sum_{s=1}^{T} X_{j-l,s}$ when $T \to \infty$, we get

$$S_{j-l,T} = \frac{1}{T} \sum_{s=1}^{T} X_{j-l,s} = T^{-\frac{1}{2}+\delta} \frac{1}{T^{\frac{1}{2}+\delta}} \sum_{s=1}^{T} X_{j-l,s} < T^{-\frac{1}{2}+\delta} \frac{c}{\log T} = \frac{c}{T^{\frac{1}{2}-\delta} \log T} \to 0. \tag{4.53}$$

Thus,

$$\|\boldsymbol{S}_T\| = \sqrt{\sum_{l=1}^{k} \sum_{j=1}^{k} S_{j-l,T}^2} \le \sqrt{\left(\frac{k^2 c^2}{T^{1-2\delta} \log^2 T} \right)} = \frac{kc}{T^{\frac{1}{2}-\delta} \log T}, \tag{4.54}$$

which implies that

$$\|\boldsymbol{S}_T\| = O\big\{ k/(T^{\frac{1}{2}-\delta} \log T) \big\}, \tag{4.55}$$

requiring $k = T^{\frac{1}{2}-\delta}$, for $\delta \in (0, 1/2)$.

Secondly, we investigate the asymptotic behavior of $\boldsymbol{E}_T = O\big\{(j-l)/T\big\}$, for $j = 1, \ldots, k$, and $l = 1, \ldots, k$.

$$\boldsymbol{E}_T = \frac{1}{T} \begin{pmatrix} 0 & 1 & \ldots & k-1 \\ 1 & 0 & \ldots & k-2 \\ \vdots & \vdots & \ddots & \vdots \\ k-1 & k-2 & \ldots & 0 \end{pmatrix} = \frac{1}{T}(\boldsymbol{D} + \boldsymbol{D}'), \tag{4.56}$$

where $\boldsymbol{D} = \frac{1}{T}\begin{pmatrix} 0 & 0 & \dots & 0 \\ 1 & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ k-1 & k-2 & \dots & 0 \end{pmatrix}$.

Since

$$\|\boldsymbol{D}\| = \sqrt{\lambda_{max}(\boldsymbol{D}^*\boldsymbol{D})} = \sqrt{\sum_{j=0}^{k}(k-j)^2} = O(k^3/2), \tag{4.57}$$

where $\boldsymbol{D}^*$ is the conjugate transpose of $\boldsymbol{D}$ and λ_{max} is the largest eigenvalue of $\boldsymbol{D}^*\boldsymbol{D}$, we obtain

$$\|\boldsymbol{E}_T\| = \frac{1}{T}\|\boldsymbol{D} + \boldsymbol{D}'\| = O\big(k^{\frac{3}{2}}/T\big), \tag{4.58}$$

requiring $k = T^{\frac{2}{3}-\varrho}$, where $\varrho \in \big(0, 2/3\big)$.

By the Cauchy-Schwartz inequality, (4.40) can be re-written as

$$\|\hat{\boldsymbol{R}}_k - \boldsymbol{R}_k\| \leq \|\boldsymbol{S}_T\| + \|\boldsymbol{E}_T\|. \tag{4.59}$$

By results in (4.55) and (4.58), for $\varrho - 1/6 \leq \delta \leq 1/2$ and $\varrho \in \big(0, 2/3\big)$, the following statement holds

$$\|\hat{\boldsymbol{R}}_k - \boldsymbol{R}_k\| = O\big(k^{\frac{3}{2}}/T\big), \tag{4.60}$$

which implies that $\hat{\boldsymbol{R}}_k - \boldsymbol{R}_k \geqslant ck^{\frac{3}{2}}/T$.

Since $T^{-1}\varepsilon\boldsymbol{\Lambda}_k + \boldsymbol{R}_k > 0$ when $T \to \infty$, $k \to \infty$ and $k^{\frac{3}{2}}/T \to 0$, we obtain that

$$\hat{\boldsymbol{R}}_k - \boldsymbol{R}_k + T^{-1}\varepsilon\boldsymbol{\Lambda}_k + \boldsymbol{R}_k > 0 \quad \text{with probability 1,}$$

or equivalently, $T^{-1}\check{\boldsymbol{R}}_k^{\varepsilon} > 0$ a.s., which implies that $\hat{\boldsymbol{\tau}}_k \to \boldsymbol{\tau}$ a.s. $\square$.

Next, we prove that the RLS estimates $\hat{\boldsymbol{\tau}}_k$ are asymptotically normally distributed. In order to derive this result, we need to show that the regularized sample ACVF follows an asymptotic normal distribution. Denote the regularized sample ACVF by

$$\hat{r}_j^{\varepsilon} = \frac{1}{T}\Big\{\sum_{t=1}^{T-j} y_t y_{t+j} + \delta_{j,0}\varepsilon e^{\mu}\Big\} \tag{4.61}$$

Correspondingly, the regularized sample covariance vectors are given by

$$\begin{aligned} \hat{\boldsymbol{r}}_k^{\varepsilon} &= (\hat{r}_1^{\varepsilon}, \ldots, \hat{r}_k^{\varepsilon})', \\ \hat{\boldsymbol{r}}_{k,0}^{\varepsilon} &= (\hat{r}_0^{\varepsilon}, \hat{r}_1^{\varepsilon}, \ldots, \hat{r}_k^{\varepsilon})', \end{aligned} \tag{4.62}$$

and regularized sample $k \times k$-Toeplitz covariance matrix

$$\hat{\boldsymbol{R}}_k^{\varepsilon} = \frac{1}{T}\check{\boldsymbol{R}}_k^{\varepsilon} = \hat{\boldsymbol{R}}_k + \varepsilon \boldsymbol{\Lambda}_k = \begin{pmatrix} \hat{r}_0^{\varepsilon} & \hat{r}_1^{\varepsilon} & \ldots & \hat{r}_{k-1}^{\varepsilon} \\ \hat{r}_1^{\varepsilon} & \hat{r}_0^{\varepsilon} & \ldots & \hat{r}_{k-2}^{\varepsilon} \\ \vdots & \vdots & \ddots & \vdots \\ \hat{r}_{k-1}^{\varepsilon} & \hat{r}_{k-2}^{\varepsilon} & \ldots & \hat{r}_0^{\varepsilon} \end{pmatrix} \tag{4.63}$$

In fact, the utilization of regularizer only changes the diagonal entries of $\hat{\boldsymbol{R}}_k$ which is $\hat{r}_0$, but the remaining entries are equal. Asymptotically, $\hat{\boldsymbol{R}}_k^{\varepsilon}$ is equivalent to $\hat{\boldsymbol{R}}_k$ and $\hat{\boldsymbol{r}}_{k,0}^{\varepsilon}$ is equivalent to $\hat{\boldsymbol{r}}_{k,0}$ by the following argument: the regularizer vanishes as $T \to \infty$, i.e.,

$$\hat{r}_0^{\varepsilon} = \frac{1}{T}\sum_{t=1}^{T} y_t^2 + \frac{\varepsilon e^{\mu}}{T} \to \frac{1}{T}\sum_{t=1}^{T} y_t^2 = \hat{r}_0 \tag{4.64}$$

and therefore,

$$\hat{\boldsymbol{r}}_{k,0}^{\varepsilon} \to \hat{\boldsymbol{r}}_{k,0}, \quad \text{as } T \to \infty. \tag{4.65}$$

We now state the Central Limit Theorem (CLT) of $\hat{\boldsymbol{r}}_{k,0}^{\varepsilon}$.

Theorem 4.2.3 *Given* $\{y_t, t \in T\}$, *under the assumptions stated above,* $\sqrt{T}(\hat{\boldsymbol{r}}_{k,0}^{\varepsilon} - \boldsymbol{r}_{k,0}) \to N(0, \boldsymbol{\Sigma})$ *in distribution. Here,* $\boldsymbol{\Sigma} = [\sigma_{ij}^{\varepsilon}]_{i,j=0,\ldots,k}$ *and*

$$\sigma_{ij}^{\varepsilon} = \begin{cases} \delta_{i,j}\sigma^4 + 2\rho^2\sigma^2\cos(i\omega_0)\cos(j\omega_0), & i,j \neq 0, \\ (\eta - 1)\sigma^4 + 2\rho^2\sigma^2, & i,j = 0. \end{cases} \tag{4.66}$$

Proof By (4.64) and (4.65), $\hat{\boldsymbol{r}}_{k,0}^{\varepsilon}$ is equivalent to $\hat{\boldsymbol{r}}_{k,0}$ as $T \to \infty$. Also, the assumption applied here on $\{\epsilon_t\}$, which assumes $\{\epsilon_t\} \sim IID(0, \sigma^2)$ and $E(\epsilon_t^4) = \eta\sigma^4 < \infty$, is a special case of the assumptions applied in Li et al (1994), which assumes $\{\epsilon_t\}$ is a linear process of the form $\epsilon_t = \sum_{j=-\infty}^{\infty} \psi_j \xi_{t-j}$ where $\{\xi_t\} \sim IID(0, \sigma^2)$, and $E(\xi_t^4) = \eta\sigma^4 < \infty$. Therefore, we can adopt the results from Li et al (1994).

By Li et al. (1994) Theorem 2 (the CLT for $\hat{\boldsymbol{r}}_{k,0}$), $\sqrt{T}(\hat{\boldsymbol{r}}^{\varepsilon}_{k,0} - \boldsymbol{r}_{k,0})$ is asymptotically normally distributed with mean zero and covariance matrix $\boldsymbol{\Sigma}$, where $\boldsymbol{\Sigma} = [\sigma^{\varepsilon}_{ij}]_{i,j=0,\ldots,k}$ and

$$\sigma^{\varepsilon}_{ij} = \lim_{T\to\infty} E\{T(\hat{r}^{\varepsilon}_i - r_i)(\hat{r}^{\varepsilon}_j - r_j)\},\ i,j = 0,\ldots,k. \tag{4.67}$$

By (4.64), $r_j = \frac{\rho^2}{2}\cos(j\omega_0) + \delta_{j,0}\sigma^2$, and the regularized sample ACVF is:

$$\begin{aligned}
\hat{r}^{\varepsilon}_j \quad &= \frac{1}{T}\{\sum_{t=1}^{T-j} y_t y_{t+j} + \delta_{j,0}\varepsilon e^{\mu}\} \\
&= \frac{1}{T}\sum_{t=1}^{T-j}\Big\{\frac{\rho^2}{2}\cos[(j+2t)\omega_0 + 2\phi] + \frac{\rho^2}{2}\cos(j\omega_0) + x_{t+j}\epsilon_t \\
&\quad + x_t\epsilon_{t+j} + \epsilon_t\epsilon_{t+j}\Big\} + \frac{\delta_{j,0}\varepsilon e^{\mu}}{T}
\end{aligned} \tag{4.68}$$

Thus, the estimation error of regularized sample ACVF estimate is given by:

$$\begin{aligned}
\hat{r}^{\varepsilon}_j - r_j \quad &= \frac{1}{T}\sum_{t=1}^{T-j}\Big\{\frac{\rho^2}{2}\cos[(j+2t)\omega_0 + 2\phi] + \frac{\rho^2}{2}\cos(j\omega_0) + x_{t+j}\epsilon_t \\
&\quad + x_t\epsilon_{t+j} + \epsilon_t\epsilon_{t+j}\Big\} + \frac{\delta_{j,0}\varepsilon e^{\mu}}{T} - \big(\frac{\rho^2}{2}\cos(j\omega_0) + \delta_{j,0}\sigma^2\big) \\
&= \frac{j}{T}\frac{\rho^2}{2}\cos(j\omega_0) + \rho^2\cos[(T-1)\omega_0 + 2\phi]\frac{\sin[(T-j)\omega_0]}{2T\sin\omega_0} \\
&\quad + \frac{1}{T}\sum_{t=1}^{T-j} x_{t+j}\epsilon_t + \frac{1}{T}\sum_{t=1}^{T-j} x_t\epsilon_{t+j} + \frac{1}{T}\sum_{t=1}^{T-j}\epsilon_t\epsilon_{t+j} + \frac{\delta_{j,0}\varepsilon e^{\mu}}{T} - \delta_{j,0}\sigma^2 \\
&= A_{1j} + A_{2j} + A_{3j} + A_{4j} + A_{5j} + \frac{\delta_{j,0}\varepsilon e^{\mu}}{T} - \delta_{j,0}\sigma^2,
\end{aligned} \tag{4.69}$$

where $A_{1j} = \frac{j}{T}\frac{\rho^2}{2}\cos(j\omega_0)$, $A_{2j} = \rho^2\cos[(T-1)\omega_0 + 2\phi]\frac{\sin[(T-j)\omega_0]}{2T\sin\omega_0}$, $A_{3j} = \frac{1}{T}\sum_{t=1}^{T-j} x_{t+j}\epsilon_t$, $A_{4j} = \frac{1}{T}\sum_{t=1}^{T-j} x_t\epsilon_{t+j}$, $A_{5j} = \frac{1}{T}\sum_{t=1}^{T-j}\epsilon_t\epsilon_{t+j}$.

Therefore,

$$\begin{aligned}
\sigma^{\varepsilon}_{ij} &= \lim_{T\to\infty} E\{T(\hat{r}^{\varepsilon}_i - r_i)(\hat{r}^{\varepsilon}_j - r_j)\} \\
&= \lim_{T\to\infty} E\big[T(\sum_{l=1}^{5} A_{li} + \frac{\delta_{i,0}\varepsilon e^{\mu}}{T} - \delta_{i,0}\sigma^2)(\sum_{k=1}^{5} A_{kj} + \frac{\delta_{j,0}\varepsilon e^{\mu}}{T} - \delta_{j,0}\sigma^2)\big]
\end{aligned}$$

Note that the first and second moments of x_t are:

$$E(x_t) = E[\rho\cos(t\omega_0+\phi)] = \rho\int_0^{2\pi}\cos(t\omega_0+\phi)\frac{1}{2\pi}d\phi = 0$$

$$\begin{aligned} E(x_t^2) &= E[\rho^2\cos^2(t\omega_0+\phi)] = \rho^2 E\big\{\frac{1}{2}[1+\cos 2(t\omega_0+\phi)]\big\} \\ &= \frac{\rho^2}{2} + \frac{\rho^2}{2}\int_0^{2\pi}\cos 2(t\omega_0+\phi)\frac{1}{2\pi}d\phi = \frac{\rho^2}{2} \end{aligned} \tag{4.70}$$

Also,

$$E(\epsilon_t\epsilon_s\epsilon_u\epsilon_v) = \begin{cases} \eta\sigma^4 & \text{if } t=s=u=v \\ \sigma^4, & \text{if } t=s\neq u=v \\ 0, & \text{if } t\neq s, t\neq u, \text{and } t\neq v \end{cases} \tag{4.71}$$

Let us consider two cases depending on i and j.

Case 1: Let $i,j\neq 0$. Then $\sigma_{ij}^{\varepsilon} = \lim_{T\to\infty} E\big[T\sum_{l=1}^{5}A_{li}\sum_{k=1}^{5}A_{kj}\big]$, where

$$\lim_{T\to\infty} E\big(TA_{1i}A_{1j}\big) = \lim_{T\to\infty}\frac{ij}{T}\frac{\rho^4}{4}\cos(i\omega_0)\cos(j\omega_0) = 0,$$

$$\begin{aligned} \lim_{T\to\infty} E\big(TA_{2i}A_{2j}\big) &= \lim_{T\to\infty}\frac{1}{T}E\left\{\rho^4\cos^2[(T-1)\omega_0+2\phi]\frac{\sin[(T-i)\omega_0]}{2\sin\omega_0}\frac{\sin[(T-j)\omega_0]}{2\sin\omega_0}\right\} \\ &= \lim_{T\to\infty}\frac{1}{T}\rho^4\frac{\sin[(T-i)\omega_0]}{2\sin\omega_0}\frac{\sin[(T-j)\omega_0]}{2\sin\omega_0}E\big\{\cos^2[(T-1)\omega_0+2\phi]\big\} \\ &= \lim_{T\to\infty}\frac{1}{T}\rho^4\frac{\sin[(T-i)\omega_0]}{2\sin\omega_0}\frac{\sin[(T-j)\omega_0]}{2\sin\omega_0}E\big\{\frac{1}{2}\big(1+\cos 2[(T-1)\omega_0+2\phi]\big)\big\} \\ &= \lim_{T\to\infty}\frac{1}{2T}\rho^4\frac{\sin[(T-i)\omega_0]}{2\sin\omega_0}\frac{\sin[(T-j)\omega_0]}{2\sin\omega_0} = 0, \end{aligned}$$

$$\lim_{T\to\infty} E\big(TA_{3i}A_{3j}\big) = \lim_{T\to\infty}\frac{1}{T}E\left(\sum_{t=1}^{T-i}x_{t+i}\epsilon_t\sum_{s=1}^{T-j}x_{s+j}\epsilon_s\right)$$

$$
\begin{aligned}
&= \lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T-\max(i,j)} E[x_{t+i}x_{t+j}]E[\epsilon_t^2] \\
&= \lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T-\max(i,j)} \rho^2 E\{\cos[(t+i)\omega_0+\phi]\cos[(t+j)\omega_0+\phi]\}\sigma^2 \\
&= \lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T-\max(i,j)} \frac{\rho^2}{2} E\{\cos[(2t+i+j)\omega_0+2\phi]+\cos[(i-j)\omega_0]\}\sigma^2 \\
&= \lim_{T\to\infty} \frac{T-\max(i,j)}{T}\frac{\rho^2\sigma^2}{2}\cos[(i-j)\omega_0] \\
&= \frac{\rho^2\sigma^2}{2}\cos[(i-j)\omega_0],
\end{aligned}
$$

$$
\begin{aligned}
\lim_{T\to\infty} E\big(TA_{4i}A_{4j}\big) &= \lim_{T\to\infty} \frac{1}{T} E\bigg(\sum_{t=1}^{T-i} x_t\epsilon_{t+i}\sum_{s=1}^{T-j} x_s\epsilon_{s+j}\bigg) \\
&= \lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T-\max(i,j)} E[x_t x_{t+i-j}]E[\epsilon_{t+i}^2] \\
&= \lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T-\max(i,j)} \rho^2 E\{\cos[t\omega_0+\phi]\cos[(t+i-j)\omega_0+\phi]\}\sigma^2 \\
&= \lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T-\max(i,j)} \frac{\rho^2}{2} E\{\cos[(2t+i-j)\omega_0+2\phi]+\cos[(i-j)\omega_0]\}\sigma^2 \\
&= \lim_{T\to\infty} \frac{T-\max(i,j)}{T}\frac{\rho^2\sigma^2}{2}\cos[(i-j)\omega_0] \\
&= \frac{\rho^2\sigma^2}{2}\cos[(i-j)\omega_0],
\end{aligned}
$$

$$
\lim_{T\to\infty} E\big(TA_{5i}A_{5j}\big) = \lim_{T\to\infty} \frac{1}{T} E\bigg(\sum_{t=1}^{T-i}\epsilon_t\epsilon_{t+i}\sum_{s=1}^{T-j}\epsilon_s\epsilon_{s+j}\bigg) = \lim_{T\to\infty}\frac{T-i}{T}\delta_{i,j}\sigma^4 = \delta_{i,j}\sigma^4,
$$

$$
\begin{aligned}
\lim_{T\to\infty} E\big(TA_{1i}A_{2j}\big) &= \lim_{T\to\infty} E\big(TA_{1j}A_{2i}\big) \\
&= \lim_{T\to\infty} \frac{i}{T}\frac{\rho^4}{2}\cos(i\omega_0)\frac{\sin[(T-j)\omega_0]}{2\sin\omega_0} E\big\{\cos[(T-1)\omega_0+2\phi]\big\} = 0,
\end{aligned}
$$

$$\lim_{T\to\infty} E\big(TA_{1i}A_{3j}\big) = \lim_{T\to\infty} E\big(TA_{1j}A_{3i}\big)$$
$$= \lim_{T\to\infty} \frac{i}{T}\frac{\rho^2}{2}\cos(i\omega_0)\sum_{t=1}^{T-j} E(x_{t+j})E(\epsilon_t) = 0,$$

$$\lim_{T\to\infty} E\big(TA_{1i}A_{4j}\big) = \lim_{T\to\infty} E\big(TA_{4i}A_{1j}\big)$$
$$= \lim_{T\to\infty} \frac{i}{T}\frac{\rho^2}{2}\cos(i\omega_0)\sum_{t=1}^{T-j} E(x_t)E(\epsilon_{t+j}) = 0,$$

$$\lim_{T\to\infty} E\big(TA_{1i}A_{5j}\big) = \lim_{T\to\infty} E\big(TA_{5i}A_{1j}\big)$$
$$= \lim_{T\to\infty} \frac{i}{T}\frac{\rho^2}{2}\cos(i\omega_0)\sum_{t=1}^{T-j} E\big\{\epsilon_t\epsilon_{t+j}\big\} = 0,$$

$$\lim_{T\to\infty} E\big(TA_{2i}A_{3j}\big) = \lim_{T\to\infty} E\big(TA_{3i}A_{2j}\big)$$
$$= \lim_{T\to\infty} \frac{1}{T}\rho^2\frac{\sin[(T-i)\omega_0]}{2\sin\omega_0}\sum_{t=1}^{T-j} E\big\{\cos[(T-1)\omega_0+2\phi]x_{t+j}\epsilon_t\big\} = 0,$$

$$\lim_{T\to\infty} E\big(TA_{2i}A_{4j}\big) = \lim_{T\to\infty} E\big(TA_{4i}A_{2j}\big)$$
$$= \lim_{T\to\infty} \frac{1}{T}\rho^2\frac{\sin[(T-i)\omega_0]}{2\sin\omega_0}\sum_{t=1}^{T-j} E\big\{\cos[(T-1)\omega_0+2\phi]x_t\epsilon_{t+j}\big\} = 0,$$

$$\lim_{T\to\infty} E\big(TA_{2i}A_{5j}\big) = \lim_{T\to\infty} E\big(TA_{5i}A_{2j}\big)$$
$$= \lim_{T\to\infty} \frac{1}{T}\rho^2\frac{\sin[(T-i)\omega_0]}{2\sin\omega_0}\sum_{t=1}^{T-j} E\big\{\cos[(T-1)\omega_0+2\phi]\epsilon_t\epsilon_{t+j}\big\} = 0,$$

$$\lim_{T\to\infty} E\big(TA_{3i}A_{4j}\big) = \lim_{T\to\infty} E\big(TA_{4i}A_{3j}\big)$$
$$= \lim_{T\to\infty} \frac{1}{T}E\bigg(\sum_{t=1}^{T-i} x_{t+i}\epsilon_t \sum_{s=1}^{T-j} x_s\epsilon_{s+j}\bigg)$$

$$
\begin{aligned}
&= \lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T-i} E[x_{t+i+j} x_t] E[\epsilon_{t+j}^2] \\
&= \lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T-i} \rho^2 E\{\cos[(t+i+j)\omega_0 + \phi] \cos[t\omega_0 + \phi]\}\sigma^2 \\
&= \lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T-i} \frac{\rho^2\sigma^2}{2} E\{\cos[(2t+i+j)\omega_0 + 2\phi] + \cos[(i+j)\omega_0]\} \\
&= \frac{\rho^2\sigma^2}{2} \cos[(i+j)\omega_0],
\end{aligned}
$$

$$
\lim_{T\to\infty} E\big(TA_{3i}A_{5j}\big) = \lim_{T\to\infty} E\big(TA_{5i}A_{3j}\big) = \lim_{T\to\infty} \frac{1}{T} E\bigg(\sum_{t=1}^{T-i} x_{t+i}\epsilon_t \sum_{s=1}^{T-j} \epsilon_s \epsilon_{s+j}\bigg) = 0,
$$

$$
\lim_{T\to\infty} E\big(TA_{4i}A_{5j}\big) = \lim_{T\to\infty} E\big(TA_{5i}A_{4j}\big) = \lim_{T\to\infty} \frac{1}{T} E\bigg(\sum_{t=1}^{T-i} x_t\epsilon_{t+i} \sum_{s=1}^{T-j} \epsilon_s \epsilon_{s+j}\bigg) = 0.
$$

Thus, when $i, j \neq 0$,

$$
\begin{aligned}
\sigma_{ij}^{\varepsilon} &= \lim_{T\to\infty} \Big\{ E\big(TA_{3i}A_{3j}\big) + E\big(TA_{4i}A_{4j}\big) + E\big(TA_{3i}A_{4j}\big) + E\big(TA_{4i}A_{3j}\big) + E\big(TA_{5i}A_{5j}\big) \Big\} \\
&= \delta_{i,j}\sigma^4 + \rho^2\sigma^2 \cos[(i-j)\omega_0] + \rho^2\sigma^2 \cos[(i+j)\omega_0] \\
&= \delta_{i,j}\sigma^4 + 2\rho^2\sigma^2 \cos(i\omega_0)\cos(j\omega_0).
\end{aligned}
$$

Case 2: Let $i = j = 0$. Then

$$
\begin{aligned}
\sigma_{00}^{\varepsilon} &= \lim_{T\to\infty} \mathrm{Var}\big\{\mathrm{T}^{1/2} \sum_{\mathrm{l}=1}^{5} \mathrm{A_{lj}}\big\} \\
&= \lim_{T\to\infty} T\big\{\mathrm{Var(A_{30})} + \mathrm{Var(A_{40})} + \mathrm{Var(A_{50})} + 2\mathrm{Cov(A_{30}, A_{40})}\big\},
\end{aligned}
$$

where

$$
\lim_{T\to\infty} T\mathrm{Var(A_{30})} = \lim_{\mathrm{T}\to\infty} \mathrm{TVar(A_{40})} = \lim_{\mathrm{T}\to\infty} \mathrm{TCov(A_{30}, A_{40})} = (\rho^2\sigma^2)/2, \tag{4.72}
$$

$$\lim_{T\to\infty} T\mathrm{Var}(\mathrm{A}_{50}) = \lim_{\mathrm{T}\to\infty} \frac{1}{\mathrm{T}}\left\{\sum_{\mathrm{t}=1}^{\mathrm{T}} \mathrm{E}(\epsilon_{\mathrm{t}}^4) - \sum_{\mathrm{t}=1}^{\mathrm{T}} [\mathrm{E}(\epsilon_{\mathrm{t}}^2)]^2\right\} = (\eta-1)\sigma^4. \tag{4.73}$$

Thus, when $i = j = 0$,

$$\sigma_{00}^{\varepsilon} = (\eta-1)\sigma^4 + 2\rho^2\sigma^2. \tag{4.74}$$

By Case 1 and 2, the result follows. □

Based on the result of Theorem 4.2.3, we now state the CLT for $\hat{\boldsymbol{\tau}}_k$, which mirrors the proof of Lau et al.(2002) on asymptotic normal distribution of the YW estimates.

Theorem 4.2.4 *If $\omega_0 \in (0,\pi)$, $T \to \infty$ and $k \to \infty$ such that $k^{\frac{3}{2}}/T \to 0$, then*

$$\sqrt{T}(\hat{\boldsymbol{\tau}}_k - \boldsymbol{\tau}_k) \to N(0, \boldsymbol{R}_k^{-1}\boldsymbol{M}\boldsymbol{\Sigma}\boldsymbol{M}'\boldsymbol{R}_k^{-1}),$$

where $\hat{\boldsymbol{\tau}}_k = (\hat{a}_1,\ldots,\hat{a}_k)$, $\boldsymbol{\tau}_k = (a_1,\ldots,a_k)$, and

$$\boldsymbol{M} = \begin{pmatrix} a_1 & a_2 & a_3 & \ldots & a_k & 0 \\ a_2 & a_3 & a_4 & \ldots & 0 & 0 \\ a_3 & a_4 & a_5 & \ldots & 0 & 0 \\ \vdots & \vdots & \vdots & \ldots & \vdots & \vdots \\ a_{k-1} & a_k & 0 & \ldots & 0 & 0 \\ a_k & 0 & 0 & \ldots & 0 & 0 \end{pmatrix} - \begin{pmatrix} 0 & a_0 & 0 & \ldots & 0 & 0 \\ 0 & a_1 & a_0 & \ldots & 0 & 0 \\ 0 & a_2 & a_1 & \ldots & 0 & 0 \\ \vdots & \vdots & \vdots & \ldots & \vdots & \vdots \\ 0 & a_{k-2} & a_{k-3} & \ldots & a_0 & 0 \\ 0 & a_{k-1} & a_{k-2} & \ldots & a_1 & a_0 \end{pmatrix}. \tag{4.75}$$

Proof Since $\sqrt{T}(\hat{\boldsymbol{r}}_{k,0}^{\varepsilon} - \boldsymbol{r}_{k,0})$ converges in distribution, as stated by Theorem 4.2.3, it follows the result of Serfling (1980) that $\hat{\boldsymbol{r}}_{k,0}^{\varepsilon} = \boldsymbol{r}_{k,0} + O(1/\sqrt{T})$. Define the following quantities:

- $g(\hat{\boldsymbol{r}}_{k,0}^{\varepsilon}) = (\hat{\boldsymbol{R}}_k^{\varepsilon})^{-1}\hat{\boldsymbol{r}}_k^{\varepsilon} = \hat{\boldsymbol{\tau}}_k$,
- $g(\boldsymbol{r}_{k,0}) = (\boldsymbol{R}_k)^{-1}\boldsymbol{r}_k = \boldsymbol{\tau}_k$,
- $\boldsymbol{\Delta}_{k,i}$= $(p\times p)$-matrix with $\pm i$th off-diagonal elements equal to 1, and 0 otherwise,
- $\boldsymbol{\vartheta}_{k,i}$= $(p\times 1)$-vector with $\pm i$th element equal to 1, and 0 otherwise,

Note in particular that $\boldsymbol{\Delta}_{k,0}$ is the identity matrix and $\boldsymbol{\Delta}_{k,k}$ is a zero matrix. Then, for $i = 0, 1, \ldots, k$

$$\frac{\partial(\hat{\boldsymbol{R}}_k^{\varepsilon})^{-1}}{\partial \hat{r}_i^{\varepsilon}} = -(\hat{\boldsymbol{R}}_k^{\varepsilon})^{-1}\boldsymbol{\Delta}_{k,i}(\hat{\boldsymbol{R}}_k^{\varepsilon})^{-1} \text{ and } \frac{\partial \hat{\boldsymbol{r}}_k^{\varepsilon}}{\partial \hat{r}_i^{\varepsilon}} = \boldsymbol{\vartheta}_{k,i}. \tag{4.76}$$

Thus, by the chain rule,

$$\begin{aligned}
\frac{\partial g(\hat{\boldsymbol{r}}_{k,0}^{\varepsilon})}{\partial \hat{r}_i^{\varepsilon}}\bigg|_{\hat{\boldsymbol{r}}_{k,0}^{\varepsilon}=\boldsymbol{r}_{k,0}} &= \left\{\left[\frac{\partial(\hat{\boldsymbol{R}}_k^{\varepsilon})^{-1}}{\partial \hat{r}_i^{\varepsilon}}\right]\hat{\boldsymbol{r}}_k^{\varepsilon} + (\hat{\boldsymbol{R}}_k^{\varepsilon})^{-1}\left[\frac{\partial \hat{\boldsymbol{r}}_k^{\varepsilon}}{\partial \hat{r}_i^{\varepsilon}}\right]\right\}\bigg|_{\hat{\boldsymbol{r}}_{k,0}^{\varepsilon}=\boldsymbol{r}_{k,0}} \\
&= -(\boldsymbol{R}_k)^{-1}\boldsymbol{\Delta}_{k,i}(\boldsymbol{R}_k)^{-1}\boldsymbol{r}_k + (\boldsymbol{R}_k)^{-1}\boldsymbol{\vartheta}_{k,i} \\
&= -(\boldsymbol{R}_{k,T})^{-1}(\boldsymbol{\Delta}_{k,i}\boldsymbol{\tau}_k - \boldsymbol{\vartheta}_{k,i}).
\end{aligned}$$

Applying Taylor's expansion,

$$\begin{aligned}
\sqrt{T}(\hat{\boldsymbol{\tau}}_k - \boldsymbol{\tau}_k) &= \sqrt{T}\big\{g(\hat{\boldsymbol{r}}_{k,0}^{\varepsilon}) - g(\boldsymbol{r}_{k,0})\big\} \\
&= \sqrt{T}\sum_{i=0}^{k}\frac{\partial g(\hat{\boldsymbol{r}}_{k,0}^{\varepsilon})}{\partial \hat{r}_i^{\varepsilon}}\bigg|_{\hat{\boldsymbol{r}}_{k,0}^{\varepsilon}=\boldsymbol{r}_{k,0}}(\hat{r}_i^{\varepsilon} - r_i) + o(1) \\
&= -\sum_{i=0}^{k}(\boldsymbol{R}_k)^{-1}(\boldsymbol{\Delta}_{k,i}\boldsymbol{\tau}_k - \boldsymbol{\vartheta}_{k,i})\sqrt{T}(\hat{r}_i^{\varepsilon} - r_i) + o(1) \\
&= -(\boldsymbol{R}_k)^{-1}[\boldsymbol{\tau}_k, (\boldsymbol{\Delta}_{k,1}\boldsymbol{\tau}_k - \boldsymbol{\vartheta}_{k,1}), \ldots, (\boldsymbol{\Delta}_{k,k}\boldsymbol{\tau}_k - \boldsymbol{\vartheta}_{k,k})]\sqrt{T}(\hat{r}_i^{\varepsilon} - r_i) + o(1).
\end{aligned}$$

Let $a_i = 0$ for $i < 0$ and $i > k$. Note that

$$\boldsymbol{\Delta}_{k,i}\boldsymbol{\tau}_k - \boldsymbol{\vartheta}_{k,i} = \begin{bmatrix} a_{1+i} \\ a_{2+i} \\ \vdots \\ a_{k+i} \end{bmatrix} - \begin{bmatrix} a_{1-i} \\ a_{2-i} \\ \vdots \\ a_{k-i} \end{bmatrix}. \tag{4.77}$$

Therefore, $[\boldsymbol{\tau}_k, (\boldsymbol{\Delta}_{k,1}\boldsymbol{\tau}_k - \boldsymbol{\vartheta}_{k,1}), \ldots, (\boldsymbol{\Delta}_{k,k}\boldsymbol{\tau}_k - \boldsymbol{\vartheta}_{k,k})] = \boldsymbol{M}$, and the result follows by Theorem 4.2.3. □

Based on the results of Theorem 4.2.3 and Theorem 4.2.4, we now derive the asymptotic properties of estimated frequency $\hat{\omega}_k$. Let us introduce the following polynomial:

$$a^*(z) = 1 + a_1^* z + \ldots + a_k^* z^k, \tag{4.78}$$

and parameter vector $\boldsymbol{\tau}_k^* = (a_1^*, \ldots, a_k^*)$, which is defined to be

$$\boldsymbol{\tau}_k^* = \boldsymbol{R}_k^+ \boldsymbol{r}_k, \tag{4.79}$$

where $\boldsymbol{R}_k^+$ denotes the Moore-Penrose pseudoinverse of $\boldsymbol{R}_k$. Stoica et al. (1989) shows that

$$a^*(z) = B^*(z)A(z), \tag{4.80}$$

where $A(z) = 1 - 2\cos\omega_0 z + z^2$ and $B(z)$ is a monic polynomial of degree $k-2$ uniquely defined by

$$\frac{1}{2\pi}\int_{-\pi}^{\pi} \left|B^*(e^{i\omega})\right|^2 \left|A(e^{i\omega})\right|^2 d\omega = \min_B \frac{1}{2\pi}\int_{-\pi}^{\pi} \left|B(e^{i\omega})\right|^2 \left|A(e^{i\omega})\right|^2 d\omega. \tag{4.81}$$

$A(z)$ has its roots located on the unit circle at $e^{\pm i\omega_0}$, and the remaining roots of $a^*(z)$ are the roots of $B^*(z)$ which located outside the unit circle. As k increases, the roots of $B^*(z)$ tend to the unit circle.

Recall the polynomial $a(z)$ of the AR(k) model corresponding to the unknown coefficient vector $\boldsymbol{\tau}_k$,

$$a(z) = 1 + a_1 z + \ldots + a_k z^k. \tag{4.82}$$

and the polynomial $\hat{a}(z)$ corresponding to the RLS coefficient estimates $\hat{\boldsymbol{\tau}}_k$,

$$\hat{a}(z) = 1 + \hat{a}_1 z + \ldots + \hat{a}_k z^k. \tag{4.83}$$

Denote the complex roots of $\hat{a}(z)$ by $\hat{\beta}_p e^{\pm i\hat{\omega}_p}$, for $p = 1, \ldots, k$. The estimated frequency $\hat{\omega}_k$ is the phrase angle of complex roots $\hat{\beta}_k e^{\pm i\hat{\omega}_k}$ that are closest to the unit circle. Recall that $f_k(\theta) = \sigma^2/\{2\pi|a(e^{i\theta})|\}$ is the k-th order AR approximation to the generalized spectral density of $\{y_t\}$ and $\hat{\omega}_k$ estimates ω_{0k}, the maximum of $f_k(\theta)$. We state the following strong consistency results of the frequency estimate $\hat{\omega}_k$.

Corollary 4.2.5 *If $\omega_0 \in (0, \pi)$, $T \to \infty$ and $k \to \infty$ such that $k^{\frac{3}{2}}/T \to 0$, then for any $\varepsilon > 0$ there exists T' such that*

$$|\hat{\omega}_k - \omega_{0k}| < \varepsilon, \tag{4.84}$$

for all $T > T'$ with probability 1.

The proof of Corollary 4.2.5 follows from the result on a.s. convergence of $\hat{\boldsymbol{\tau}}_k$ using the same derivation steps as that of Theorem 1 in Mackisack and Poskitt (1989).

Corollary 4.2.6 *Under the same conditions as Corollary 4.2.5,* $\hat{\omega}_k \to \omega_0$ *almost surely.*

Proof Note that $\hat{\omega}_k - \omega_0 = (\hat{\omega}_k - \omega_{0k}) + (\omega_{0k} - \omega_0)$. Theorem 2 in Stoica et al. (1987) states that $(\omega_{0k} - \omega_0) = O(1/k^3)$. Hence, by the results of Corollary 4.2.5 and for $k^3 \to \infty$, we get

$$\hat{\omega}_k \to \omega_0, \quad \text{a.s.} \tag{4.85}$$

□.

Finally, we state the CLT of estimated frequency $\hat{\omega}_k$, whose proof applies similar arguments as Stoica et al. (1989) Theorem 5.1.

Theorem 4.2.7 *If* $k^{\frac{3}{2}} \geqslant cT^{1-\delta}$, *for* $0 < \delta < 5/8$ *such that* $k^{\frac{3}{2}}/T \to 0$, *then*

$$\sqrt{T}(\hat{\omega}_0 - \omega_0) \to N(0, \boldsymbol{F}\boldsymbol{G}\boldsymbol{R}_{k,T}^{-1}\boldsymbol{M}\boldsymbol{\Sigma}\boldsymbol{M}'\boldsymbol{R}_{k,T}^{-1}\boldsymbol{G}'\boldsymbol{F}')$$

in distribution, where

- $\boldsymbol{F} = \begin{pmatrix} \psi/(\theta^2+\psi^2), & -\theta/(\theta^2+\psi^2) \end{pmatrix}$, *where* $\theta = [\cos\omega_0, 2\cos 2\omega_0, \ldots, k\cos k\omega_0]\boldsymbol{\tau}_*$ *and* $\psi = [\sin\omega_0, 2\sin 2\omega_0, \ldots, k\sin k\omega_0]\boldsymbol{\tau}^*$;
- $\boldsymbol{G} = \begin{pmatrix} [\cos\omega_0, \cos 2\omega_0, \ldots, \cos k\omega_0]; & [\sin\omega_0, \sin 2\omega_0, \ldots, \sin k\omega_0] \end{pmatrix}'$;

Proof Using the same arguments as in the paper by Stoica et al. (1989) and taking into account the results on asymptotic consistency and normality of $\hat{\boldsymbol{\tau}}_k$ and Corollary 4.2.6, we obtain that for sufficiently large T, $\hat{\omega}_k$ is close to ω_0 and $\hat{\beta}_k$ is close to $\beta_0 = 1$. Hence, the following Taylor expansion holds under regularity conditions:

$$\begin{aligned} 0 &= Re\{\hat{a}(\hat{\beta}_k e^{i\hat{\omega}_k})\} \\ &= Re\{\hat{a}(e^{i\omega_0})\} + \left.\frac{\partial Re\{\hat{a}(\beta e^{i\omega})\}}{\partial\beta}\right|_{\beta=1,\omega=\omega_0}(\hat{\beta}_k - \beta_0) \\ &+ \left.\frac{\partial Re\{\hat{a}(\beta e^{i\omega})\}}{\partial\omega}\right|_{\beta=1,\omega=\omega_0}(\hat{\omega}_k - \omega_0) + O(1/T), \end{aligned} \tag{4.86}$$

$$
\begin{aligned}
0 &= Im\{\hat{a}(\hat{\beta}_k e^{i\hat{\omega}_k})\} \\
&= Im\{\hat{a}(e^{i\omega_0})\} + \left.\frac{\partial Im\{\hat{a}(\beta e^{i\omega})\}}{\partial \beta}\right|_{\beta=1,\omega=\omega_0} (\hat{\beta}_k - \beta_0) \\
&+ \left.\frac{\partial Im\{\hat{a}(\beta e^{i\omega})\}}{\partial \omega}\right|_{\beta=1,\omega=\omega_0} (\hat{\omega}_k - \omega_0) + O(1/T),
\end{aligned} \tag{4.87}
$$

where

$$
\begin{aligned}
\left.\frac{\partial Re\{\hat{a}(\beta e^{i\omega})\}}{\partial \beta}\right|_{\beta=1,\omega=\omega_0} &= [\cos\omega_0, 2\cos 2\omega_0, \ldots, k\cos k\omega_0]\hat{\boldsymbol{\tau}}_k, \\
\left.\frac{\partial Re\{\hat{a}(\beta e^{i\omega})\}}{\partial \omega}\right|_{\beta=1,\omega=\omega_0} &= -[\sin\omega_0, 2\sin 2\omega_0, \ldots, k\sin k\omega_0]\hat{\boldsymbol{\tau}}_k, \\
\left.\frac{\partial Im\{\hat{a}(\beta e^{i\omega})\}}{\partial \beta}\right|_{\beta=1,\omega=\omega_0} &= [\sin\omega_0, 2\sin 2\omega_0, \ldots, k\sin k\omega_0]\hat{\boldsymbol{\tau}}_k, \\
\left.\frac{\partial Im\{\hat{a}(\beta e^{i\omega})\}}{\partial \omega}\right|_{\beta=1,\omega=\omega_0} &= [\cos\omega_0, 2\cos 2\omega_0, \ldots, k\cos k\omega_0]\hat{\boldsymbol{\tau}}_k.
\end{aligned} \tag{4.88}
$$

Let us introduce the following notations:

$$
\begin{aligned}
&\theta = [\cos\omega_0, 2\cos 2\omega_0, \ldots, k\cos k\omega_0]\boldsymbol{\tau}_k^*, \quad \psi = [\sin\omega_0, 2\sin 2\omega_0, \ldots, k\sin k\omega_0]\boldsymbol{\tau}_k^*, \\
&\boldsymbol{h} = [\cos\omega_0, \cos 2\omega_0, \ldots, \cos k\omega_0]', \quad \boldsymbol{g} = [\sin\omega_0, \sin 2\omega_0, \ldots, \sin k\omega_0]', \\
&\boldsymbol{F} = \begin{pmatrix} \psi/(\theta^2+\psi^2) & -\theta/(\theta^2+\psi^2) \end{pmatrix}, \boldsymbol{G} = \begin{pmatrix} \boldsymbol{h}' \\ \boldsymbol{g}' \end{pmatrix}.
\end{aligned} \tag{4.89}
$$

By Theorem 4.2.4, as $k \to \infty$ and $T \to \infty$ such that $k^{\frac{3}{2}}/T \to 0$, $\sqrt{T}(\hat{\boldsymbol{\tau}}_k - \boldsymbol{\tau}_k)$ converges in distribution, and thus it follows the result of Serfling (1980) that $(\hat{\boldsymbol{\tau}}_k - \boldsymbol{\tau}_k) = O(1/\sqrt{T})$. Also, by the result of Theorem 1 in Stoica et al. (1987), $(\boldsymbol{\tau}_k - \boldsymbol{\tau}_k^*) = O(1/k^2)$, thus

$$
\hat{\boldsymbol{\tau}}_k - \boldsymbol{\tau}_k^* = (\hat{\boldsymbol{\tau}}_k - \boldsymbol{\tau}_k) + (\boldsymbol{\tau}_k - \boldsymbol{\tau}_k^*) = O(1/k^2) + O(1/\sqrt{T}). \tag{4.90}
$$

Since $k^{\frac{3}{2}} \geqslant cT^{1-\delta}$, for $0 < \delta < 5/8$, the dominant term in (4.86) is not affected if we replace $\hat{\boldsymbol{\tau}}_k$ by $\boldsymbol{\tau}_k^*$, which is

$$\begin{aligned} 0 &= Re\{\hat{a}(e^{i\omega_0})\} + \theta(\hat{\beta}_k - \beta_0) - \psi(\hat{\omega}_k - \omega_0) + O(1/T), \\ 0 &= Im\{\hat{a}(e^{i\omega_0})\} + \psi(\hat{\beta}_k - \beta_0) + \theta(\hat{\omega}_k - \omega_0) + O(1/T). \end{aligned} \tag{4.91}$$

Since $a^*(e^{i\omega}) = 0$,

$$\begin{aligned} Re\{\hat{a}(e^{i\omega_0})\} &= Re\{\hat{a}(e^{i\omega_0}) - a^*(e^{i\omega_0})\} = \boldsymbol{h}'(\hat{\boldsymbol{\tau}}_k - \boldsymbol{\tau}_k^*), \\ Im\{\hat{a}(e^{i\omega_0})\} &= Im\{\hat{a}(e^{i\omega_0}) - a^*(e^{i\omega_0})\} = \boldsymbol{g}'(\hat{\boldsymbol{\tau}}_k^T - \boldsymbol{\tau}_k^*). \end{aligned} \tag{4.92}$$

Denote $\Delta_{\boldsymbol{\tau}} = (\hat{\boldsymbol{\tau}}_k - \boldsymbol{\tau}_k^*)$, $\Delta_\beta = (\hat{\beta}_k - \beta_0)$, and $\Delta_\omega = (\hat{\omega}_k - \omega_0)$, then we have

$$\begin{aligned} 0 &= \boldsymbol{h}'\Delta_{\boldsymbol{\tau}} + \theta\Delta_\beta - \psi\Delta_\omega + O(1/T), \quad (1) \\ 0 &= \boldsymbol{g}'\Delta_{\boldsymbol{\tau}} + \psi\Delta_\beta + \theta\Delta_\omega + O(1/T). \quad (2) \end{aligned} \tag{4.93}$$

Solve for Δ_ω in terms of $\Delta_{\boldsymbol{\tau}}$, we get

$$\Delta_\omega = \frac{\psi\boldsymbol{h}' - \theta\boldsymbol{g}'}{\psi^2 + \theta^2}\Delta_{\boldsymbol{\tau}} + O(1/T) = \boldsymbol{F}\boldsymbol{G}\Delta_{\boldsymbol{\tau}} + O(1/T),$$

which is

$$(\hat{\omega}_k - \omega_0) = \boldsymbol{F}\boldsymbol{G}(\hat{\boldsymbol{\tau}}_k - \boldsymbol{\tau}_k^*) + O(1/T). \tag{4.94}$$

Equivalently,

$$(\hat{\omega}_k - \omega_0) = \boldsymbol{F}\boldsymbol{G}(\hat{\boldsymbol{\tau}}_k - \boldsymbol{\tau}_k) + \boldsymbol{F}\boldsymbol{G}(\boldsymbol{\tau}_k - \boldsymbol{\tau}_k^*) + O(1/T). \tag{4.95}$$

We now consider $\boldsymbol{F}\boldsymbol{G}(\boldsymbol{\tau}_k - \boldsymbol{\tau}_k^*)$. By the result of Stoica et al. (1987) Theorem 1,

$$(\boldsymbol{\tau}_k - \boldsymbol{\tau}_k^*) = O(1/k^2), \tag{4.96}$$

$$\begin{aligned} \frac{1}{k}\theta &= -\frac{2}{k^2}[\cos\omega_0, 2\cos 2\omega_0, \ldots, k\cos k\omega_0]\boldsymbol{G}' \begin{pmatrix} 1 \\ 0 \end{pmatrix} \\ &\quad + \frac{1}{k}[\cos\omega_0, 2\cos 2\omega_0, \ldots, k\cos k\omega_0]O(1/k^2) \\ &= -\frac{1}{2} + O(1/k), \end{aligned} \tag{4.97}$$

$$\begin{aligned}\frac{1}{k}\psi &= -\frac{2}{k^2}[\sin\omega_0, 2\sin 2\omega_0, \ldots, k\sin k\omega_0]\boldsymbol{G}'\begin{pmatrix}1\\0\end{pmatrix} \\ &\quad + \frac{1}{k}[\sin\omega_0, 2\sin 2\omega_0, \ldots, k\sin k\omega_0]O(1/k^2) \\ &= O(1/k).\end{aligned} \tag{4.98}$$

Substituting (4.96) - (4.98) to $\boldsymbol{FG}(\boldsymbol{\tau}_k - \boldsymbol{\tau}_k^*)$, we get

$$\boldsymbol{FG}(\boldsymbol{\tau}_k - \boldsymbol{\tau}_k^*) = \frac{O(1) + kO(1)}{O(1) + k^2O(1)}O(1/k^2) = O(1/k^3). \tag{4.99}$$

Therefore,

$$(\hat{\omega}_k - \omega_0) = \boldsymbol{FG}(\hat{\boldsymbol{\tau}}_k - \boldsymbol{\tau}_k) + O(1/k^3) + O(1/T). \tag{4.100}$$

Multiplying both sides of (4.100) by $\sqrt{T}$, we get

$$\sqrt{T}(\hat{\omega}_k - \omega_0) = \sqrt{T}\boldsymbol{FG}(\hat{\boldsymbol{\tau}}_k - \boldsymbol{\tau}_k) + O(T^{\frac{1}{2}}/k^3) + O(1/T^{\frac{1}{2}}). \tag{4.101}$$

If $k^{\frac{3}{2}} \geqslant cT^{1-\delta}$, for $0 < \delta < 3/4$, then $T^{\frac{1}{2}}/k^3 \to 0$, and $O(T^{\frac{1}{2}}/k^3) \to 0$. Also, as $T \to \infty$, $O(1/T^{\frac{1}{2}}) \to 0$, so we have

$$\sqrt{T}(\hat{\omega}_k - \omega_0) = \sqrt{T}\boldsymbol{FG}(\hat{\boldsymbol{\tau}}_k - \boldsymbol{\tau}_k). \tag{4.102}$$

By Theorem 4.2.4, $\sqrt{T}(\hat{\boldsymbol{\tau}}_k - \boldsymbol{\tau}_k) \to N(0, \boldsymbol{R}_k^{-1}\boldsymbol{M}\boldsymbol{\Sigma}\boldsymbol{M}'\boldsymbol{R}_k^{-1})$, and thus, if $k^{\frac{3}{2}} \geqslant cT^{1-\delta}$, for $0 < \delta < 5/8$ such that $k^{\frac{3}{2}}/T \to 0$,

$$\sqrt{T}(\hat{\omega}_k - \omega_0) \to N(0, \boldsymbol{FG}\boldsymbol{R}_k^{-1}\boldsymbol{M}\boldsymbol{\Sigma}\boldsymbol{M}'\boldsymbol{R}_k^{-1}\boldsymbol{G}'\boldsymbol{F}'). \quad \square \tag{4.103}$$

In this section, we have proved the strong consistency of the truncated RLS estimates $\hat{\boldsymbol{\tau}}_k$ in the case where a has roots on the unit circle. Moreover, we have showed that the estimated frequency $\hat{\omega}_k$ converges almost surely under the assumption that $T \to \infty$ and $k \to \infty$ such that $k^{\frac{3}{2}}/T \to 0$. Compared to the strong consistency results of the estimated frequency in Mackisack and Poskitt (1989), which assumes that $T \to \infty$ and $k \to \infty$ such that $k^2/T \to 0$, our result extends to higher order of approximating AR equations. In practice, Mackisack and Poskitt (1989) apply AIC to select the model order, while the order of our regularized AR approximation may be much higher than the order suggested by AIC.

4.3 Robust Trimming Algorithm

For sufficiently large T, the zeroes of $\hat{a}(z)$ are close to the zeros of $a^*(z)$, or equivalently, $B^*(z)A(z)$. Stoica (1987) points out that as k increases, the zeros of $B^*(z)$ tend to the unit circle. Therefore, it is possible that the zeros of $B^*(z)$ moves faster towards the unit circle than those corresponding to $A(z)$, which results in false frequency estimates. Such situation was encountered in our simulation studies. In order to increase the accuracy of the estimates, we propose the robust trimming algorithm (RTA) of RAR frequency estimation.

The RTA procedures are based on the result that the RAR frequency estimates are normally distributed in large samples (see Theorem 4.2.7). We first take a training set from the sample in order to choose an "optimal" combination of model order k and regularizing parameter μ, and to construct a $(1-\alpha)\%$-confidence interval (CI) based on the chosen k and μ. Model order k should be significantly higher than the order suggested by the information criterions. We select (empirically) a set of k and μ, and then apply the RLS estimation with all the combinations of k and μ from this set. We choose such k and μ that yield the minimum mean square error (MSE). This approach of selecting k and μ is similar to the thresholding method suggested by Bickel and Levina (2006), and we will investigate its theoretical justifications in our future research.

Since the sample distribution of frequency estimates is approximately normal, we can construct a $(1-\alpha)\%$-CI based under the assumption of normality. After k, μ and the $(1-\alpha)\%$-CI are obtained, we can apply the RLS estimation with the chosen k and μ to the entire sample, and only take the frequency estimates falling into the $(1-\alpha)\%$ CI. Our simulation studies indicate that such a robust trimming of frequency estimates can effectively eliminate the spurious roots and outliers, and therefore noticeably increase the accuracy.

To demonstrate the performance of the RTA method, we conduct the following simulation study (see Mackisack and Poskitt, 1989). Consider a observed sample of size 1024 generated by the following process:

$$y_t = 20\cos(1.24t + 0.01) + \epsilon_t, \tag{4.104}$$

where $\{\epsilon_t\}$ are iid N(0,1). Notice that the true frequency $\omega_0 = 1.24$.

Mackisack and Poskitt (1989) (MPAR) utilize a three-step procedure:

1. select the order of AR(k) model by AIC; AIC suggests $k = 44$.
2. estimate the AR coefficients by the YW method.
3. the frequency $\hat{\omega}_k$ is estimated by finding the minimum of the transfer function $\hat{h}_k(\theta) = \left|\hat{a}(e^{i\theta})\right|^2 = \left|\sum_{j=0}^{k} \hat{a}_j(e^{ij\theta})\right|^2$ in $(0, \pi)$.

On the other hand, we conducted the RTA approach as follows:

1. From the observed data sample $\{y_t\}_{t=1}^{T}$, take a sub-sample of size T_1, $T_1 << T$.
2. Fit AR(k) models to the sub-sample $\{y_t\}_1^{T_1}$ by the RLS estimation with various regularizer parameter μ_j, $\mu_j \in (0.001, 0.015)$, $j = 1, \ldots, 15$, and model order k_i, $k_i \in (45, 60)$, $i = 1, \ldots, 15$.
3. Find the "optimal" μ^* and corresponding k^* providing the minimum of MSE. In particular, we select $\mu^* = 0.135$ and $k^* = 55$.
4. Construct a 95% CI using the sample distribution of $\hat{\omega}_{T_1}(\mu^*, k^*)$, $j, i = 1, \ldots, 15$.
5. Fit the AR(k) model to $\{y_t\}_{T_1+1}^{T}$ by the RLS estimation with μ^*; take into account only $\hat{\omega}_t$ falling within the $(1 - \alpha)$% CI.

Note that in step (4), the 95% CI is constructed by using the median absolute deviation (MAD).

We denote MSE of the RTA and MPAR methods by MSE_{RTA} and MSE_{MPAR} respectively. In Figure 4.1, both MSE_{RTA} and MSE_{MPAR} decay exponentially when SNR increases. In particular, MSE_{RTA} is equivalent to MSE_{MPAR} when SNR is high ($> 22dB$). However, MSE_{RTA} is significantly lower than MSE_{MPAR} when SNR is low ($< 22dB$). Low SNR means that the signal is hidden in noise, so the frequency is more difficult to detect. Thus, RTA can be a preferred method for detection of unknown frequency in noisy conditions of online modeling.

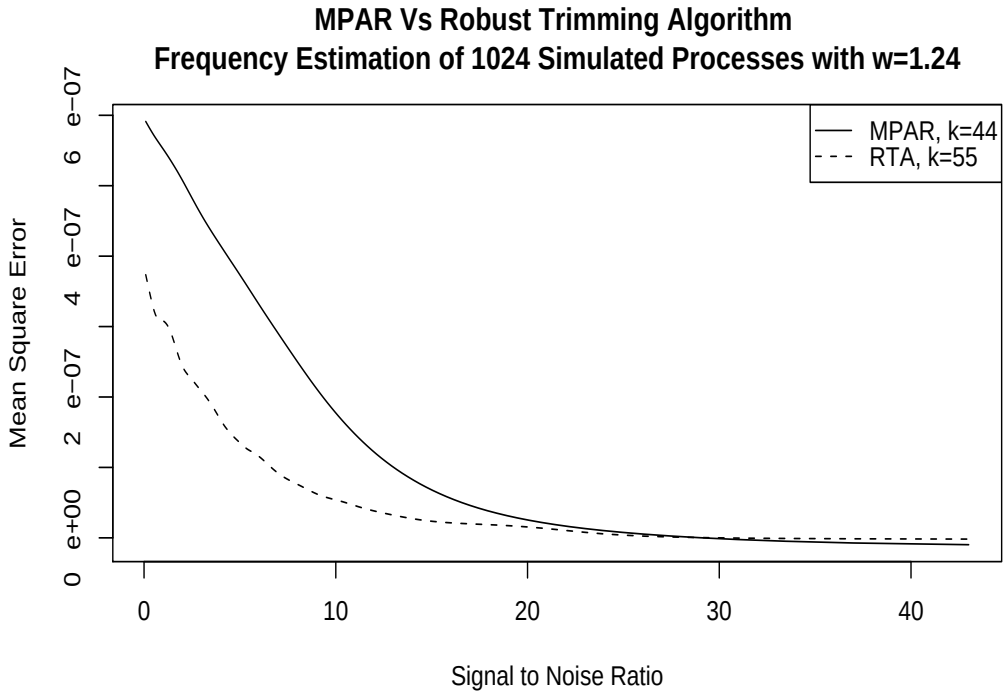

Figure 4.1: Comparison of MPAR and RTA in the plot SNR vs. MSE.

Chapter 5

AR Approximation to ARFIMA Process

According to Hosking (1984), long memory is defined as "the significant dependence between observations a long time span apart". Many time series, particularly from financial and computer science applications, possess such properties. In general, there are two major schools of analyzing long memory processes, namely, continuous models, such as the fractional Gaussian noises (Mandelbrot and van Ness, 1968), and discrete models, such as the Autoregressive Fractional Integrated Moving Average (ARFIMA) models (Granger and Joyeux, 1980; Hosking, 1981). The main focus of this thesis is on discrete time series and linear stochastic models. Hence, we are particularly interested in a class of ARFIMA models that will be discussed in this chapter. The ARFIMA model is a generalization of the ARIMA model but with fractional differencing parameter d and fits into the Box-Jenkins framework. However, estimation of ARFIMA parameters is very computationally expensive and raises issues on robustness to the choice of initial values. Therefore, we discuss an alternative method of approximating the ARFIMA process by "long" AR models (Ray, 1993; Ray and Crato; 1996, Poskitt, 2007). We also apply the Regularized AR (RAR) approximation to make long-range forecasting. We will prove the AR approximation, particularly the RAR approach, is useful in long-range prediction by simulation studies.

5.1 Introduction to the ARFIMA Process

A long memory process can be characterized in different ways (Hosking, 1984):

- the ACVF decays hyperbolically as oppose to exponentially as lag increases;
- the spectral density increases without limit as the frequency tends to zero;
- the rescaled adjusted range (Hurst, 1956) behaves as a function T^h, $h > \frac{1}{2}$ of the sample size T, rather than as $T^{1/2}$ for short memory processes. Here, h is known as Hurst exponent.

Remark. The re-scaled adjusted range is defined as follows. Given a realization of a time series $\{y_t, t \in T\}$, with sample mean $\bar{y}$ and sample variance s^2, the re-scaled adjusted range is $R = s^{-1}\{\max(S_1, ..., S_T) - \min(S_1, ..., S_T)\}$, where $S_t = \sum_{i=1}^{t} y_i - t\bar{y},\ t = 1, ..., T$ are the adjusted partial sums.

The early stage of modeling long memory processes was motivated by the Hurst phenomenon. In particular, Hurst (1951) observed the long range dependence of wet and drought periods in the Nile stream flows. Mandelbrot and Van Ness (1968) show that such dependence is compatible with stationarity by constructing fractional Brownian motion (fBm), which essentially is a weakly stationary stochastic process with a hyperbolically decaying autocorrelation function in continuous time. Mandelbrot and Wallis (1969) indicate that fractional Gaussian noise (fGn), which is a discrete time analogue of fBm, exhibits the long range dependence. Also, Mandelbrot and Wallis derive that the Hurst exponent h is equal to $d + 1/2$, where d is a fractional differencing parameter.

Other studies suggest that the Hurst phenomenon may be represented as a short memory ARMA process (O'Connell, 1971, 1974; Hipel and McLeod, 1978). The reason is that in finite samples the re-scaled adjusted range of an ARMA process does not necessarily behave as $T^{1/2}$. Instead, it can exhibits Hurst exponent larger than 1/2. Therefore, ARMA processes can capture long range dependence structure when a sample size is finite.

Granger and Joyeux (1980) and Hosking (1981) propose the long memory ARFIMA process, which generalizes the ARIMA process (Box and Jenkins, 1976) by permitting the differencing parameter d to take fractional values, in order to represent the hyperbolically decaying correlation structure. Both theoretical and empirical results indicate that

ARFIMA models are useful in modeling long memory processes. Compared with the fGn model, the ARFIMA equation enables modeling of the short memory as well as the long memory processes; while compared to the ARMA approach, the ARFIMA model takes into account the degree of the long range dependence. We define the ARFIMA process as follows.

Definition 5.1.1 *(ARFIMA Process). An ARFIMA (p, d, q) process, $p, q \in \mathbb{Z}$ and $d \in \mathbb{R}$ is defined as*

$$\phi(B)\nabla^d X_t = \theta(B)\epsilon_t, \tag{5.1}$$

where

- $\{\epsilon_t\} \sim WN(0, \sigma^2)$,
- $\phi(\lambda)$ *and* $\theta(\lambda)$ *are polynomials in* B *of degree* p *and* q *respectively,*
- $\phi(\lambda)$ *and* $\theta(\lambda)$ *have no common factor,*
- $\phi(\lambda) \neq 0$ *for* $|\lambda| \leq 1$,
- *the fractional differencing operator* ∇^d *represents*

$$(1-B)^d = \sum_{k=0}^{\infty} \binom{d}{k} (-B)^k = \sum_{k=0}^{\infty} b_k B^k, \qquad \text{where } b_k = \frac{\Gamma(-d+k)}{\Gamma(-d)\Gamma(k+1)}. \tag{5.2}$$

When $-1/2 < d < 1/2$, all roots of $\phi(\lambda)$ and $\theta(\lambda)$ lie outside the unit circle. Hence, the ARFIMA (p, d, q) process is causal and invertible. In particular, if $0 < d < 1/2$, the process exhibits significantly positive dependence between distant observations; with $d = 0$, the ARFIMA process becomes a short memory ARMA process; if $-1/2 < d < 0$, the process exhibits significantly negative dependence between distant observations. When $p = q = 0$ and $d > -\frac{1}{2}$ but not equal to zero, the ARFIMA $(0, d, 0)$ process is called fractional noise process, which can be considered as the result of applying fractional differencing to white noise.

The long term behavior of an ARFIMA(p, d, q) process is similar to that of the fractional noise process with the same value of d. The reason is that the effect of d on distant

observations decays hyperbolically as the lag increases, while the effects of ϕ and θ decay exponentially. For very distant observations, the effects of ϕ and θ are negligible.

Figure 5.1 presents a time series plot as well as ACF and PACF plots of 1000 observations generated from an ARFIMA$(1, d, 1)$ model with $\phi_1 = 0.7$, $d = 0.3$, $\theta_1 = 0.2$. Note that d indicates the degree of long range dependence in the process. As d increases, the correlation between distant observations becomes more and more evident, and vice versa.

5.2 Estimation and Forecasting of the ARFIMA Process

After the introduction of ARFIMA processes, a large number of studies have been conducted on its estimation procedures, both in frequency and time domains. Generally, the existing estimation methods can be divided into three categories: two-step estimation procedure, simultaneous estimation (one-step) procedure and AR approximations.

The two-step estimation procedure is conducted as follows:

1. estimate the differencing parameter d alone;

2. estimate the remaining AR and MA parameters.

There are many different ways to perform the step 1. Since $d = h - 1/2$, where h is the Hurst exponent, we may utilize the existing estimators of h to estimate d, for instance, the Hurst's (1951, 1956) K coefficient $K = \log R/(\log T - \log 2)$, and the Mandelbrot and Wallis estimator (1969) which uses the slope of the regression of $\log R$ on log(series length) with R calculated from subseries of various lengths. However, these estimators are proved to be biased upward if $h < 0.7$ and biased downward if $h > 0.7$, while their sampling variability is large (Wallis and Matalas, 1970).

A number of alternative estimators have been developed for estimating d. Geweke and Porter-Hudak (1983) investigate behavior of spectral density of long memory processes around zero and propose an estimator of d, based on a regression slope of log(periodogram). Chen et al. (1994) proposed the lag window spectral density estimator to improve the estimator in Geweke and Porter-Hudak (1983). In general, the existing methods can be

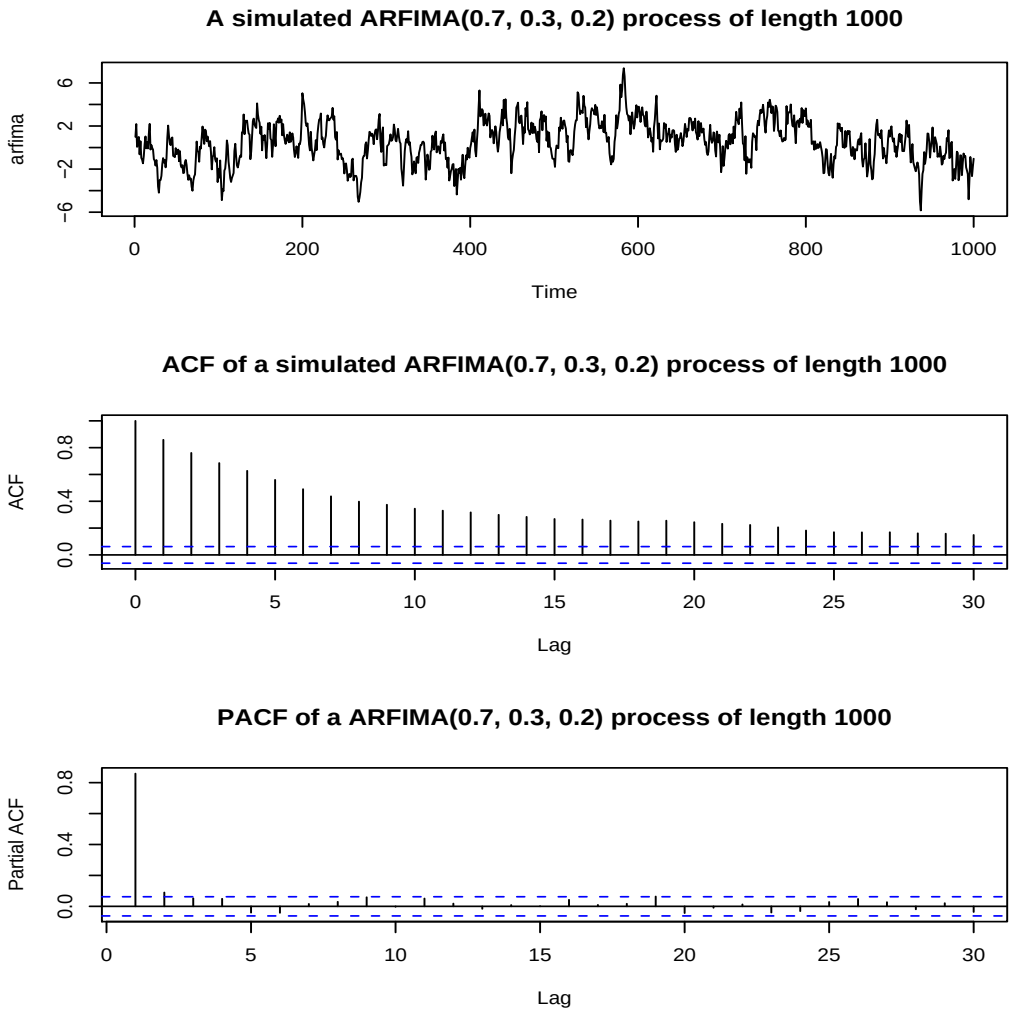

Figure 5.1: An ARFIMA(0.7,0.3,0.2) process and its ACF, PACF

grouped into the parametric and semiparametric methods. The discussions and examples for the parametric method may be found in, for instance, Fox and Taqqu (1986), Dahlhaus (1989) and Ludeña (2000); for the semiparametric method may be found in, for instance, Geweke and Porter-Hudak (1983), Reisen(1993, 1994), Robinson (1995) and references therein. Some recent simulation studies comparing different techniques for estimation d

are presented in Reisen and Lopes (1999), Hurvich and Deo (1999).

In the second step, the estimated differencing parameter is used to transform the observed series into a series that follows a standard ARMA(p, q) model, and then we can identify and estimate the ϕ and θ parameters using Box-Jenkins modelling procedure.

We can also combine the two-step procedure into a simultaneous estimation of all parameters, i.e. estimate d jointly with the ARMA parameters ϕ and θ. Both McLeod and Hipel (1978) and Sowell (1992) apply the MLE method to estimate the ARFIMA parameters d, ϕ and θ simultaneously. Tieslau et al. (1996) propose the minimum-distance estimator based on estimated and theoretical autocorrelations of an ARFIMA (p, d, q) process. Reisen et al. (2001) considered an iterative estimation procedure by Hosking (1981) to estimate parameters. More detailed discussions on the simultaneous estimation procedure can be found in Smith (1997), Lobato and Robinson (1996), and Cheung and Dielbold (1994).

In fact, an ARFIMA process can be transformed into an AR(∞) process. A common approach to make k-step-ahead prediction from an ARFIMA process is based on the truncated AR(∞) model. After some literature research, we find that the exact procedure on how to choose an "optimal" ARFIMA model and the order of truncation especially for k-steps ahead forecasting remains unclear. Also, most current estimation methods are likelihood based. The ML method offers the best prospect of giving efficient parameter estimates, which achieves the Cramér-Rao lower bounds. However, ML is a non-linear optimization that requires good initial values. Therefore, such estimation is computationally very expensive, and poor choices of initial values might result in meaningless estimates.

An alternative estimation approach is to skip the estimation of parameters d, ϕ, θ and fit an "long" AR model to an ARFIMA process from the beginning (Ray, 1993; Ray and Crato; 1996, Poskitt, 2006). AR models provide the advantage of estimation simplicity and well-investigated theoretical properties. We discuss the AR modeling of the ARFIMA process in the following section.

5.3 AR approximation to the ARFIMA process

Theoretically, an ARFIMA (p, d, q) process can be converted to an AR(∞) process. Substituting (5.2) into (5.1), the ARFIMA (p, d, q) process becomes

$$\phi(B) \sum_{k=0}^{\infty} b_k B^k X_t = \theta(B)\epsilon_t \tag{5.3}$$

where $b_0 = 1$, $b_1 = -d$, $b_2 = d(1-d)/2$, $b_k = b_{k-1}(k-1-d)/k$, for $k \geq 3$.

Then, substituting $\phi(B) = \sum_{i=1}^{p} \phi_i B^i$, $\theta(B) = \sum_{j=1}^{q} \theta_j B^j$ into (5.3), we obtain

$$\frac{\sum_{i=1}^{p} \phi_i B^i \sum_{k=0}^{\infty} b_k B^k}{\sum_{j=0}^{q} \theta_j B^j} X_t = \left(1 - \sum_{i=1}^{\infty} \delta_i L^i\right) X_t = \epsilon_t \tag{5.4}$$

where $\delta_i = b_i - \sum_{j=1}^{q} \theta_j \delta_{i-j} + \sum_{i=1}^{p} \phi_j b_{i-j}$, and $\sum \delta_i^2 < \infty$, for $|d| < 1/2$. Note that (5.4) is a AR(∞) process.

The AR model has a simple structure, and its estimation techniques as well as the asymptotic properties are well-established. The estimation of AR parameters requires much less effort compared to that of the ARFIMA parameters. In practice, an AR(∞) model is truncated to finite AR(p) to approximate the ARFIMA process. Although AR(p) does not possess long memory properties, we attempt to choose the model order p large enough so that the tail of the AR(∞) is negligible. Such modeling is equivalent to approximating a hyperbolically decaying correlation function by a sum of exponentials.

There exists a number of theoretical and empirical studies that verify feasibility of using "long" AR models to approximate and forecast ARFIMA processes. Geweke and Porter-Hudak (1983) indicate that the AR(50) model provides forecasts comparable to the fractional noise models, based on three economic series. Ray (1993) show empirically and theoretically that AR(p) models produce accurate long-range forecasts of the fractional noise process. Ray (1993) also addresses the need for considering different model orders for forecasting at different lead times. Moreover, Crato and Ray (1996) indicate that "long"

AR models perform competitively or better for long-range forecasting than the subcases of ARFIMA(p, d, q): FI$(0, d, 0)$, ARFI$(1, d, 0)$ and FIMA$(0, d, 1)$.

Based on our simulation study, we find that

1. "long" AR approximation is also useful for the general ARFIMA(p, d, q) with $p \geq 1$ and $q \geq 1$,
2. the RAR approximation can potentially be a competitive method for producing forecasts of long memory processes,
3. "percentage increase in prediction error" keeps decreasing as model order p of the AR(p) approximation. increases.

We present our numerical example in the following section.

5.4 Numerical Examples

We present two examples on application of the RAR approach to modeling and forecasting of long memory processes.

Example (1). In this example, we illustrate that the AR approximation is useful for prediction of the general ARFIMA(p, d, q) processes with $p \geq 1$ and $q \geq 1$, which extends the results of Ray (1993) and Crato and Ray (1996) who consider only FI$(0, d, 0)$, ARFI$(1, d, 0)$ and FIMA$(0, d, 1)$. The simulation study is conducted as follows:

1. simulate 10000 observations from the long memory ARFIMA$(1, 0.15, 1)$ process with coefficients $ar = 0.2$, $ma = -0.4$.
2. fit {AR(1), AR(2),..., AR(80)} models into the first 9980 observations by the Yule-Walker estimation method and make 20-step-ahead predictions using each AR model.
3. calculate the relative percentage increase (RPI) in mean square prediction errors (MSPE) for each AR(p) approximation, $p = 1, ..., 80$. The RPI for k-step-ahead prediction is define as

$$\text{RPI}(k) = \frac{V^*(k) - V(k)}{V(k)} \times 100 \tag{5.5}$$

where $V^*(k)$ is the k-step-ahead MSPE using AR(p) model, and $V(k)$ is the k-step-ahead MSPE using the true model.

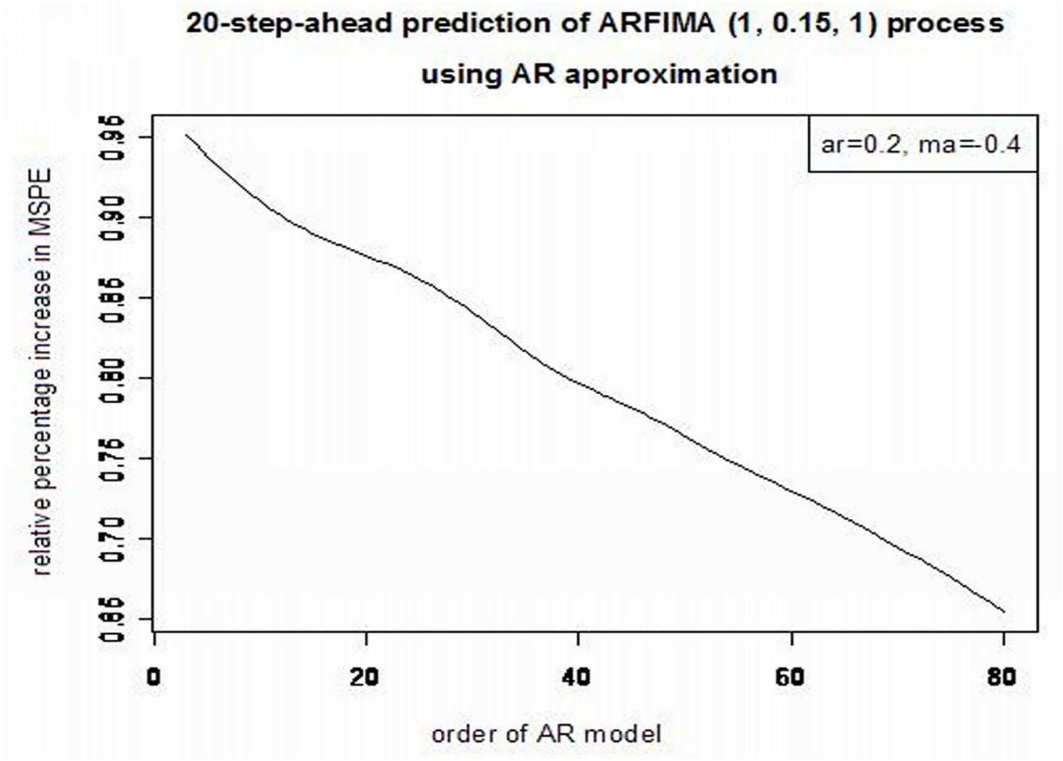

Figure 5.2: An ARFIMA process approximated by "long" AR model

In Figure 5.2, RPI in MSPE decreases monotonically as the order p of the approximating AR model increases, which has same pattern as the a special case of ARFIMA(p, d, q), FI$(0, d, 0)$, discussed in Ray (1993).

Example (2). An application of the RAR approximation in long memory daily temperature data from Portland, Oregon, USA. We have 13140 daily observations that are pre-processed, i.e. de-meaned. We take the first 10000 observations as our training sample, and observations $[10001, \ldots, 10010]$ as the verification sample.

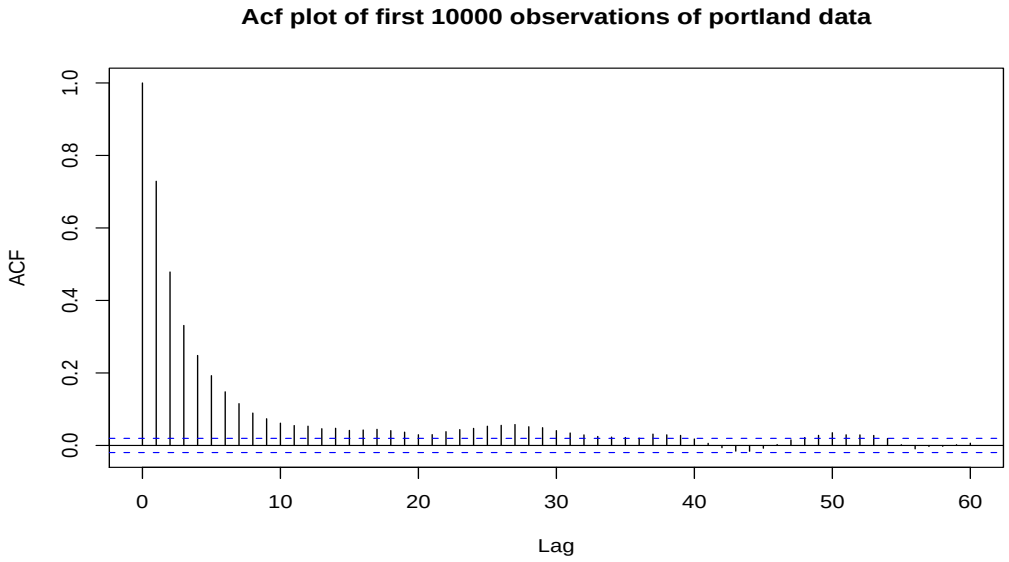

Figure 5.3: ACF plot of first 10000 observations of Portland data

In Figure 5.3, the acf plot of the sample exhibits long memory property, i.e. acf decays hyperbolically. We model such long memory process using the ARFIMA model and the RAR approximation. Our goal is to compare the two approaches in terms of 10-step-ahead MSPE, which is implemented as follows:

- Approach (1). ARFIMA model.
  We apply two-step-estimation procedure to fit an ARFIMA model to the sample. Firstly, the fractional differecing parameter d is estimated by Geweke and Porter-Hudak method, which gives $d = 0.12$. After we fractionally difference the sample data, we compare AIC values of the ARMA models of the differenced sample data, and select an ARMA(10,2) model which achieves lowest AIC. Then, we calculate the corresponding AR coefficients for ARFIMA(10, 0.12, 2) model. Based on the truncated AR(80) model, we produce 10-step-ahead forecasts of observations $[10001, \ldots, 10010]$.

- Approach (2). RAR approximation.

We take a training set of 1000 observations from the sample in order to find an "optimal" model. Firstly, we apply the RAR approximation with various regularizing parameters $\mu \in (0.05, 0.5)$ and model order $p \in (30, 80)$ to the first 990 observations, and then make 10-step-ahead predictions using the AR models with different combinations of regularization parameters. The "optimal" model is selected in such way that achieves minimum MSPE. In particular, we choose $\mu = 0.1$ and $p = 70$. Then, we apply the RAR approximation using the selected "optimal" model to the entire sample, and produce 10-step-ahead forecasts.

We now compare the 10-step-ahead root mean square prediction errors (RMSPE) of the two approaches:

RMSPE of 10-step prediction	
ARFIMA(10, 0.12, 2) using Truncated AR(80)	RAR approximation with $\mu = 0.1$ and $p = 70$
0.8522	0.414

Table 5.1: Comparison of RMSPE: ARFIMA model Vs RAR approximation

In Table 5.1, the 10-step-ahead RMSPE of RAR approximation has about 50% reduction from that of the ARFIMA. In this particular application, we conclude that the RAR approximation performs better than the ARFIMA model. Though a more extended study is needed to validate all properties of RAR, this initial analysis indicates that the RAR approximation can be potentially a competitive method for producing forecasts of long memory processes.

Chapter 6

Conclusion and Future Work

In this thesis, we discuss the Regularized AR (RAR) approximation for modeling and forecasting of time series. The RAR idea is to fit a "long" autoregressive (AR) model to an observed process and estimate AR coefficients by the Regularized LS (RLS) method, which constitutes a version of the iterative stationary Kalman filter. RAR enables to avoid the repeated model order selection and parameter estimation with an increase of a sample size. Two applications of the RAR method are presented in details, i.e. frequency estimation and long memory process forecasting.

In our frequency estimation study, we extend the results of Mackisack and Poskitt (1989) and Gel and Barabanov (2007) and apply the RAR procedure for detection of unknown frequency in periodic signals. Our theoretical findings indicate that the RAR estimates of unknown frequency converge almost surely as the approximating AR model order k increases at the rate $k^{\frac{3}{2}}/T \to 0$ when $T \to \infty$. We also show that RAR estimates of unknown frequencies are asymptotically normally distributed and the corresponding variance-covariance matrix is obtained. Moreover, a new robust trimming algorithm is proposed to eliminate spurious roots and outliers, which noticeably increase the accuracy of the frequency estimates for processes with a low signal-to-noise ratio. We can conclude that the RAR frequency estimation along with the robust trimming algorithm noticeably reduces the computational burden and improves accuracy of estimates.

The results of Ray (1993) and Crato and Ray (1996) indicate that "long" AR models perform competitively or better in forecasting long memory processes when compared

to the traditional ARFIMA-based methods, in particular, for such subcases of ARFIMA as FI$(0, d, 0)$, ARFI$(1, d, 0)$ and FIMA$(0, d, 1)$. We investigate this result via simulation studies for more general ARFIMA(p, d, q) models with $p \geq 1$ and $q \geq 1$. We also apply the Regularized AR (RAR) approximation to forecasting of long memory processes. In the Portland temperature example, we show that the RAR approach yields a lower 10-step-ahead root mean square prediction error (RMSE) than yielded by the ARFIMA model chosen by AIC. Although a more extended validation study is need to assess performance of RAR, we can conclude that the RAR approximation is potentially a competitive method for modeling and forecasting of long memory processes.

In the future, we plan to investigate the following topics:

1. Analysis of consistency and distributional properties of the RLS and, hence, the RAR estimates in infinite dimensional case.

2. Extension of the RAR approach to detection of multiple unknown frequencies.

3. Development of systematic procedures for selecting "optimal" regularizing parameters μ and ϵ.

4. Analysis of linkage between the regularizer and AIC, BIC, CAT and other information criterions, as well as with recent advances on regularization of covariance matrices in other fields of statistics.

5. Investigation on how RLS is connected to shrinkage and "Lasso" regressions.

6. Asymptotic properties of RLS for AR(∞) with hyperbolically decaying coefficients, i.e. long memory processes.

Appendix

Chapter 2

##R code##

```
*****************************************************
Simulation of an AR(2) process of length 1000
with coefficient (0.3, 0.2) and its ACF, PACF
*****************************************************
ar2=arima.sim(list(order = c(2, 0, 0), ar=c(0.3, 0.2)), n=1000)
par(mfrow=c(3,1))
ts.plot(ar2, main="A simulated AR(2) process of length 1000
with coefficients (0.3, 0.2)")
acf(ar2, main="ACF of a simulated AR(2) process of length
1000 with coefficients (0.3, 0.2)")
pacf(ar2, main="PACF of a simulated AR(2) process of length
1000 with coefficients (0.3, 0.2)")

*****************************************************
Simulation of an MA(2) process of length 1000
with coefficient (0.3, 0.2) and its ACF, PACF
*****************************************************
ma2=arima.sim(list(order = c(0, 0, 2), ma=c(0.3, 0.2)), n=1000)
```

```
par(mfrow=c(3,1))
ts.plot(ma2, main="A simulated MA(2) process of length 1000
with coefficients (0.3, 0.2)")
acf(ma2, main="ACF of a simulated MA(2) process of length
1000 with coefficients (0.3, 0.2)")
pacf(ma2, main="PACF of a simulated MA(2) process of length
1000 with coefficients (0.3, 0.2)")

*****************************************************************
Simulation of an ARMA(2,1) process of length 1000
with coefficient ar=(0.2, 0.1), ma=0.3 and its ACF, PACF
*****************************************************************
arma=arima.sim(list(order = c(2, 0, 1), ar=c(0.2, 0.1), ma=0.3), n=1000)
par(mfrow=c(3,1))
ts.plot(arma, main="A simulated ARMA(2, 1) process of length 1000
with coefficients ar=(0.2, 0.1), ma=0.3")
acf(arma, main="ACF of a simulated ARMA(2, 1) process of length
1000 with coefficients ar=(0.2, 0.1), ma=0.3")
pacf(arma, main="PACF of a simulated ARMA(2, 1) process of length
1000 with coefficients ar=(0.2, 0.1), ma=0.3")

*************************************************************************
Simulation of an ARIMA(2,1,1) process of length 1000
with coefficient ar=(0.2, 0.1), ma=0.3 and its its first differencing
*************************************************************************
arima=arima.sim(list(order = c(2, 1, 1), ar=c(0.2, 0.1), ma=0.3), n=1000)
par(mfrow=c(2,1))
ts.plot(arima, main="A simulated ARIMA(2, 1, 1) process")
ts.plot(diff(arima), main="A simulated ARIMA(2, 1, 1) process,
after first difference")
```

```
Chapter 4

##Matlab code##

*****************************************
Mackisack and Poskitt example:
*****************************************

noise=(14:-0.1:0.1);
snr=10*log10(0.5*400./noise.^2);

for i=1:140,
        for j=1:1000,
         arw1024(j)=arfreq(noise(i), 1024, 44, 1.24, 0.01);
        end;
        armse1024(i)=var(arw1024);
end;

function w=arfreq(noise,n,k,intial,fi)
y=20*cos(intial*[1:n]'+ fi ) + noise*randn(n,1);
bg=arburg(y, round(k));
n=roots(bg);
m=abs(roots(bg));
[maxval, maxindex]=max(m);
w=angle(n([maxindex]));

*********************************************
RAR example
*********************************************

for i=1:140,
```

```
        for j=1:1000,
         rarw1024(j)=newlar(noise(i),1024,50,1.24, 0.01, 0.005);
        end;
        rarmse1024(i)=var(rarw1024);
end;

function w=newlar(noise,n,k,initial, fi, mu)
y=20*cos(initial*[1:n]'+ fi) + noise*randn(n,1);
t=MNK(y,k,n,mu);
a=[1,-t];
n=roots(a);
m=abs(n);
[minval, minindex]=min(abs(m-1));
w=angle(n([minindex]));

function tau=MNK(y, k, n, mu) ## RLS##
gamma=zeros(k,k);
for i=1:k,
      gamma(i,i)=exp(mu*k);
      tau(i)=0.1;
end;

for t=(k+1):(n-1),
        for i=1:k,
             F1(i)=y(t-i+1);
        end;
        F=F1';
        gamma=gamma-gamma*F*F'*gamma/(1+F'*gamma*F);
        tau=tau+(gamma*F*(y(t+1)-F'*tau')/(1+F'*gamma*F))';
end;
```

```
*********************************************
RAR robust trimming algorithm
*********************************************

data=20*cos(1.24*[1:1024]'+ 0.01) + randn(1024,1); ##training set##
sample=data(1:400);
for s=1:101
        t=MNK(sample,55, 400, 0.005*(s-1));
        a=[1,-t];
        n=roots(a);
        m=abs(n);
        [minval, minindex]=min(abs(m-1));
        o(s)=angle(n([minindex]));
end;

med=median(o);
for k=1:101
   mo=abs(o(k)-median(o));
end;
sd=median(mo)*1.4826;
cl=med-2*sd; ##confidence inteval##
cu=med+2*sd;

for p=1:101
     bias(p)=abs(o(p)-med);
end;
[minval, minindex]=min(bias);
mu=0.005*minindex;

ave=mean(o);
variance=var(o);
```

```
cl=ave-sqrt(variance/100);
cu=ave+sqrt(variance/100);

for i=1:1000,
        counter=0;
        for j=1:1000,
         freq=newlar(noise(i),1024,55,1.24, 0.01, mu);
              if freq>cl & freq<cu
                    counter=counter+1;
               rartrimw(counter)=freq;
end;
         end;
         if counter ==0,
         rartrimmse(i)=0;
         else
         rartrimmse(i)= var(rartrimw);
         end;
end;

##R code##

************************************************
MP VS RAR
************************************************
tt=seq(14,0.1,by=-0.1)
t=10*log10(200/tt^2)

la1024m=read.table("G:/BeiChen/numericalexample/
1000poskitt1024_1.24_44.txt",sep="")
lar1024m=as.numeric(la1024m)
las1024m=smooth.spline(lar1024m)
```

```
robust1024w=read.table("G:/BeiChen/numericalexample/
1000robust1024_1.24_55_0.135.txt", sep="")
robustas1024w=as.numeric(robust1024w)
robostsmooth1024w=smooth.spline(robustas1024w)

plot(t,las1024m$y, type="n", xlab="Signal to Noise Ratio",
ylab="Mean Square Error", main="MP VS Robust Trimming Algorithm
Frequency Estimation of 1024 Simulated Processes with w=1.24")
lines(t,las1024m$y,lty=1)
lines(t,robostsmooth1024w$y,lty=2)
legend("topright", c("MP, k=44", "RTA, k=55"),lty=c(1,2), pch=-1)


Chpater 5

##R code##

**********************************************************************
Simulation of an ARFIMA(1,d,1) process of length 1000
with coefficienta ar=0.7, ma=0.2, d=0.3 and its ACF, PACF
**********************************************************************

arfima=farimaSim(n=1000, model=list(ar=0.7, d=0.3, ma=0.2),
method=c("time") )
ts.plot(arfima, main="A simulated ARFIMA(0.7, 0.3, 0.2)
process of length 1000")
acf(arfima, main="ACF of a simulated ARFIMA(0.7, 0.3, 0.2)
process of length 1000")
pacf(arfima, main="PACF of a ARFIMA(0.7, 0.3, 0.2)
process of length 1000")
```

```
*****************
Example 1
*****************

rr=function(l){
rr=0
for( j in -p:p){
rr=rr+ru[abs(j)+1]*rx[abs(j+l)+1]
}
rr
}

d=0.15
q=rep(0,80)
for(s in 1:80){
vv=rep(0,1000)
for(u in 1:1000){
simdata=fracdiff.sim(10000,ar=c(0.2), ma=c(-0.4), d=0.15)
k=5
p=s
fit=ar.yw(test,FALSE,p)

ru=ARMAacf(ar=0, ma=da,p)
c=ARMAtoMA(ar=arcoef,ma=0,k+p)
rx=rep(0,k+p+1)
rx[1]=gamma(1-2*d)/(gamma(1-d))^2
for (i in 2:(k+p+1)){
rx[i]=(rx[i-1]*(i-1+d))/(i-d)
}
v=0
for (i in 0:(k-1)){
```

```
    for (j in 0:(k-1)){
        v=v+c[i+1]*c[j+1]*rr(i-j)
        }
}
vv[u]=v
}
q[s]=mean(vv)
}

********************
Example 2
********************

##Approach (1) by ARFIMA

##R code##

portland=read.table("G:/BeiChen/my master thesis/portland.txt", sep="")
sample=portland[1:10000]
verification=sample[10001:10010]
mGPH=fdGPH(sample)
mGPH$d
r=diffseries(sample, mGPH$d)
long=arima0(r, order=c(10, 0, 2))


## Matlab code##

load('-ascii','portland.mat');
subset=portland(1:10000);
sample=portland(1:fsize-10);
```

```
fsize=10000;
numpred=10;
forward=sample;
nov=subset(9991:10000);
verification=portland(10001:10010);
pred=[];
d=0.12;
p=10;
q=2;
phi=[0.3132,  -0.1087,  0.1373,  -0.0075,  0.0155,  -0.0030,  0.0023,
-0.0140,  0.0024,  -0.0128]; ##result from R##
theta=[0.3662, 0.2383];
k=80;

for j=1: numpred
   for i=1:numpred
      t=arfimacoef(d,k,theta, phi, p, q);
      train=forward(fsize-k+1:fsize);
      new(i)=1 + t*train';
      forward=[forward(2:9990),new(i)];
   end;
pred(j)=new(numpred);
forward=[sample(1+j : fsize), nov(2:j+1)];
end;

sum=0;
for i=1:numpred
   sum=sum + (pred(i)-verification(i))^2;
end;

mse=sum/10;
```

```
function coef=arfimacoef(d,k,theta,phi,p,q)
%negative
%first coef is not 1, should be added 1 when used

b(1)=-d;
b(2)=1/2*d*(1-d);

if k>=3
for t=3:k
b(t)= b(t-1)*(t-1-d)/t;
end;
end;

for i = 1:k
  aa=0;
  for j=1:q
     if i-j>1
       aa=aa+theta(j)*c(i-j);
     else
       aa=aa;
     end;
  end;

  bb=0;
  for h=1:p
     if i-j>1
       bb=bb+phi(j)*b(i-j);
     else
       bb=bb;
     end;
  end;
```

```
  c(i)=b(i)-aa+bb;
end;
coef = c;

##Approach (2) by RAR ##

********************
training sample
********************

fsize=1000; #####take a training set of 1000########
numpred=10; ####10-step-ahead###########
real=portland(991 :1000);
forward=portland(1:fsize-10);

for k= 1:50
   for mu=1:50
      mse=0;
      for i=1:numpred
          t=MNK(forward, k+30, fsize, mu/100);
            for j=1:k+30
               train(j)=forward(fsize+1-j);
            end;
          new(i)=1+t*train';
          fsize=fsize+1;
          forward=portland(1:fsize);
      end;

      for p=1:10
      mse= mse+(real(p)-new(p))^2;
      end;
```

```
  pmse(mu)=mse/10;
  end;
  cmse(:,k)=pmse;
end;

******************
Entire sample
******************

load('-ascii','portland.mat');
subset=portland(1:10000);
sample=portland(1:fsize-10);
fsize=10000;
numpred=10;
forward=sample;
nov=subset(9991:10000);
verification=portland(10001:10010);
pred=[];
k=70;
mu=0.1;
for i=1:numpred
   for j=1:numpred
      t=MNK(forward, k, fsize, 0.1);
      train=forward(fsize-k+1:fsize);
      new(i)=1 + t*train';
      forward=[forward(2:9990),nov(2:j+1)];
   end;
pred(j)=new(numpred);
forward=[sample(1+j : fsize), nov(2:j+1)];
end;
```

```
sum=0;
for i=1:numpred
   sum=sum + (pred(i)-verification(i))^2;
end;
mse=sum/10;
```

Bibliography

Abraham, B. and Ledolter, J. (2005). *Statistical Methods for Forecasting*. Wiley, New York.

Agafanov, S., Barabanov, A. and Fomin, V. (1982). "Adaptive filtering of stochastic processes". *In book: Problems of Cybernetic;. Actual Adaptive Control Problems, Cybernetics Counsel of USSR*, Acad. Sc., N. 4-30.

Akaike, H. (1969). "Fitting autoregressive models for prediction". *Annals of The Institute of Statistical Mathematics*. **21**, 243-247.

Akaike, H. (1974). "A new look at the statistical model identification". *IEEE Transactions on Automatic Control*. **19** (6), 716-723.

Anderson, T. W. (1971). *The Statistical Analysis of Time Series*. John Wiley and Sons.

Baillie, R. T. (1996). "Long memory processes and fractional integration in econometrics". *Journal of Econometrics*. **73**, 5-59.

Barabanov, A. E. (1983). "On strong convergence of the method of least squares". *Automatics and Telemechanika*. **10**, 119-127.

Barabanov, A. and Gel, Y. R. (2005). "Strong consistency of the Least-Squares method with a polynomial regularizer for infinite AR models". *Automat. Remote Control*. **66**, 92-108.

Baker, Jr., G. A. and Graves-Morris, P. (1996). *Padé Approximants*. Cambridge University Press, New York.

Beran, J. (1994). *Statistics for Long Memory Processes*. Chapman and Hall, New York.

Bickel, P. J. and Levina, E. (2006). "Covariance regularization by thresholding". *Submitted to the Annals of Statistics*.

Bickel, P. J. and Levina, E. (2008). "Regularized estimation of large covariance matrices". *Annals of Statistics*. **36** (1), 199-227.

Bickel, P. J. and Li, B. (2006). "Regularization in Statistics". *Sociedad de Estadística e Investigación Operativa Test*. **15** (2), 271-344.

Box, G. E. P., Jenkins, G. M. and Reinsel, G. C. (1994). *Time Series Analysis: Forecasting and Control*. Prentice-Hall, Inc., New Jersey.

Breiman, L. (1996). "Heuristics of instability and stabilization in model selection". *Annals of Statistics*. **24** (6), 2350-2383.

Brockwell, P. J. and Davis, R. A. (1987). *Time Series: Theory and Methods*. Springer, New York.

Brillinger, D. R. (1987). "Fitting cosines: some procedures and some physical examples". In Applied Probability, Stochastic Processes, and Sampling Theory, Ed. I. B. MacNeill and G. J. Umphrey, 75-100. Boston: Reidel.

Bunea, F., Wegkamp, M. H., and Auguste, A. (2006). "Consistent variable selection in high dimensional regression via multiple testing". *Journal of Statistical Planning and Inference*. **136**(12), 4349-4364.

Burnham, K. P. and D. R. Anderson. (2002). *Model Selection and Multimodel Inference: A Practical-Theoretic Approach*. 2nd ed. Springer-Verlag.

Cadzow, J. A. (1980). "High performance spectral estimation - A new ARMA method". *IEEE Trans. Acoust., Speech, Signal Processing*. Vol. ASSP-28, 524-529.

Cadzow, J. A. (1982). "Spectral estiamtion: An overdetermined rational model equation approach". *Proc. IEEE*. **70**, 907-939.

Caporale, G. M., Pittis, N. and Spagnolo, N. (2003). "IGARCH models and structural breaks". *Journal Applied Economics Letters*. **10** (12), 765-768.

Cavanaugh, J. E. (1999). "A large-sample model selection criterion based on Kullbacks symmetric divergence". *Statistics and Probability Letters*. **42** (4), 333-343.

Chan Y. T. and Langford, R. P. (1982). "Spectral estimation via the high order Yule-Walker equations,". *IEEE Trans. Acoust., Speech, Signal Processing*. Vol. ASSP-30, 689-698.

Chen, G., Abraham, B., and Peires, S. (1994). "Lag window estimation of the degree of differencing in fractionally integrated time series models." *Journal of Time Series Analysis*. **15**, 473-487.

Cheung, Y. and Diebold, F. X. (1994). "On maximum-likelihood estimation of the differencing parameter of fractionally integrated time series models". *Journal of Econometrics*, **62**, 301-316.

Cleveland, W. (1993). *Visualizing Data*. Hobart Press.

dAspremont, A., Banerjee, O., and El Ghaoui, L. (2007). "First-order methods for sparse covariance selection". *SIAM Journal on Matrix Analysis and its Applications*. To appear.

Dahlhaus, R. (1989). "Efficient parameter estimation for self-similar processes". *The Annals of Statistics*. **17**, 1749-1766.

Desai, U. B. and Pal, D. (1984). "A transformation approach to stochastic model reduction". *IEEE Trans. Automat. Control*. **29** (12), 1097-1100.

Dragošević, M. V. and Stanković, S. S. (1989). "A generalized least square method for frequency estimation". *IEEE Trans.Acoust.,Speech, Signal Processing*. **37** (6), 805-819.

Engle, R. F. (1995). *ARCH: Selected Readings*. Oxford University Press.

Fomin, V. (1985). *Recurrent Estimation and Adaptive Filtering*. Nauka, Moscow.

Fomin, V. (1995). "Regression models of nonstationary time series". *In book: Modern Problems of Computer Data Analysis and Modelling (Edited by Kharin, Y.)*, **2**, 269-274. Minsk.

Fomin, V. (1999). *Optimal Filtering. Filtering of Stochastic Processes*, vol. 1. Kluwer Academic Publishers, Dordrecht.

Fox, R. and Taqqu, M. S. (1986). "Large-sample properties of parameter estimates for strongly dependent stationary Gaussian time series". *The Annals of Statistics*. **14**, 517-532.

Friedlander, B. (1984). "The overdetermined recursive instrumental variable method,". *IEEE Trans. Automat. Contr.* Vol. AC-29, 353-356.

Furrer, R. and Bengtsson, T. (2006). "Estimation of high-dimensional prior and posteriori covariance matrices in Kalman filter variants". *Journal of Multivariate Analysis*. To appear.

Galbraith, J. W. and Zinde-Walsh, V. (2001). "Conditional quantiles of volatility in equity index and foreign exchange data". *Working paper, Department of Economics, McGill University.*

Gel, Y. R. and Barabanov, A. (2007). "Strong consistency of the regularized least-squares estimates of infinite autoregressive models".*Journal of Statistical Planning and Inference*, **137**, 1260-1277.

Gel, Y. R. and Fomin, V. N. (2001). "Identification of an unstable ARMA equation". *Math. Problems Eng.* **7**, 97-112.

Gel,Y. R. and Fomin, V.N. (1998). "Linear model approximation of stochastic stationary time series". *Vestnik St.-Petersburg Univ.* **2**, 24-31.

Geweke, J. and S. Porter-Hudak. (1983). "The Estimation and Application of Long Memory Time Series Models". *Journal of Time Series Analysis.* **4**, 221-238.

Granger, C. W. J. and Joyeux R. (1980). "An introduction to long-memory time series models and fractional differencing". *Time Series Analysis.* **1**, 15-29.

Hannan, E. J. (1971). "Nonlinear time series regression". *Journal of Applied Probability.* **8**, 767-780.

Hannan, E. J. (1971). "The estimation of frequency". *Journal of Applied Probability.* **10**, 510-519.

Hannan, E. J. and Quinn, B. G. (1979). "The determination of the order of an autoregression". *Journal of the Royal Statistical Society. Series B (Methodological).* **41** (2), 190-195.

Hipel, K. W. and McLeod, A. I. (1978). "Preservation of the rescaled adjusted range, 2, simulation studies using Box-Jenkins models". *Water Resour.Res.,* **14**, 509-516.

Hoerl, A. E. and Kennard, R. W. (1970). "Ridge regression: Biased estimation for nonorthogonal problems". *Technometrics*, **12** (1), 55-67.

Hosking, J. R. M. (1981). "Fractional differencing". *Biometrika*. 68, 165-176.

Hosking, J. R. M. (1984). "Modeling persistence in hydrological time series using fractional differencing". *Water Resour. res.,* **20** (12), 1898-1908.

Hurvich, C. M., and Tsai, C. L. (1989). "Regression and time series model selection in small samples". *Biometrika*. **76**, 297-307.

Hurvich, C. M., and Deo, R. S. (1999). "Plug-in selection of the number of frequencies in regression estimates of the memory parameter of a long-memory time series". *Journal of Time Series*. **20**, 331-341.

Hurst, H. E. (1951). "Long-term storage capacity of reservoirs". *Trans. Am. Soc. Civ. Eng.* **116**, 770-799.

Hurst, H. E. (1956). "Methods of using long-term storage in reservoirs". *Proc. Inst. Civ. Eng.,* **1**, 519-543.

Kay, S. M. and Marple, S. L. (1981). "Spectrum analysisa modern perspective". *Proc. IEEE*. **69**, 1380-1419.

Kay, S. M. (1984). "Accurate frequency estimation at low signal-to-noise ratio". *IEEE Trans.Acoust.,Speech, Signal Processing*. Vol.ASSP-32, No.3, 540-547.

Kay, S. M. (1988). *Modern Spectral Estimation: Theory and Application*. Prentice-Hall, Englewood Cliffs.

Kedem, B. (1994). *Time series analysis by higher-order zero crossing*. IEEE Press, Pis-

cataway, NJ.

Kushner, H. J. and Yin, G. G. (2003). *Stochastic Approximation Algorithms and Applications, Applications of Mathematics*. 2nd Edition, Springer-Verlag, New York.

Ljung, L. (1977). "On Positive real transfer functions and the convergence of some recursive schemes".*IEEE Trans.Automat.Control*, AC-22(4), 539-551.

Ljung, L and Söderström, T. (1983). *Theory and Practice of Recursive Identijcation*. The MIT press, London.

Li,T. H. and Kedem, B (1994). "Iterative filtering for multiple frequency estimation". *IEEE Trans on Signal Processing*. **42** (5). 1120-1130.

Ledoit, O. and Wolf, M. (2004). "A well-conditioned estimator for large-dimensional covariance matrices". *Journal of Multivariate Analysis*. **88** (2), 365-411.

Lobato, I, and Robinson, P. M. (1996). "Averaged periodogram estimation of long memory". *Journal of Econometrics*. **73**, 303-324.

Ludeña, C. (2000). "Parametric estimation for Gaussian long-range dependent process based on the log-peroidogram". *Bernoulli*. **6**, 709-728.

MacKisack, M. S. and Poskitt, D. S. (1989). "Autoregressive frequency estimation". *Biometrika*. **76** (3), 565-575.

MacKisack, M. S. and Poskitt, D. S. (1990). "Some properties of autoregressive estimates for processes with mixed spectra". *Journal of Time Series Analysis*. **11**(4), 325-337.

Mackisack, M. S., Osborne, M. R. and Smyth, G.K. (1994). "A modified Prony algorithm for estimating sinusoidal frequencies". *Journal of Statistical Computation and Simulation*.

49. 111-124.

Makhoul, J. (1975). "Linear prediction: a tutorial review". *Proc.IEEE*. **63**, 561-580.

Mandelbrot, B. and Van Ness, J. W. (1969). "Fractional Brownian motions, fractional noises and applications". *SIAM Review*. **10**, 422-437.

Mandelbrot, B. and Wallis, J. R. (1968). "Noah, Joseph and operational hydrology". *Water Resources Research*. **4**, 909-918.

Mandelbrot, B. and Wallis, J. R. (1969). "Computer experiments with fractional Gaussian noises". *Water Resour. Res.,* **5**, 228-267.

Mallows, C. L. (1973). "Some comments on c_p". *Technometrics*. **15** (4), 661-675.

Mann, H. and Wald, A. (1943). "On statistical treatment of linear stochastic difference equations". *Econornetrica*. 11.

Mari, J., Dahlen, A. and Lindquist, A. (2000). "A covariance extension approach to identification of time series". *Automatica J.* IFAC **36** (3), 379-398.

Meinshausen, N. (2005). "Lasso with relaxation". Unpublished.

McQuarrie, A. D. R. and Tsai, C. L. (1998). *Regression and Time Series Model Selection*. World Scientific.

Nadaraya, E. A. (1964). "On estimating regression". *Theory of Probability and Its Applications*. **10**, 186-190.

Nehorai, A. (1985). "A minimum parameter adaptive notch filter with constrained poles and zeros". *IEEE Trans.Acoust. Speech Signal Process*. **33**, 983-996.

Nelson, D. B. (1991). "Conditional heteroskedasticity in asset returns: A new approach". *Econometrica*. **59**, 347-370.

Osborne, M. R. (1975). "Some special nonlinear least squares problems". *SIAM Journal of Numerical Analysis*. **12**. 571-592.

O'Connell, P. E. (1971). "A simple stochastic modeling of Hurst's law". *Mathematical Models in Hydrology, Symposium. Warsaw*. **1**, 169-187.

O'Connell, P. E. (1974). "Stochastic modeling of long-term persistence in stream flow sequences", Ph.D. thesis, Eng. Dep., Imperial Coll., London.

Parzen, E. (1962). "On estimation of a probability density function and mode". *The Annals of Mathematical Statistics*. **33**, 1065-1076.

Parzen, E. (1974). "Some recent advances in time series modeling". *Trans. Automat. Control*. AC-**19**, 723-730.

Parzen, E. (1977). "Multiple time series modeling: Determining the order of approximating autoregressive scheme". *Multivariate Analysis IV (Ed.P.Krishnaiah)*. 283-295. North-Holland, Amsterdam.

Pisarenko, V. F.(1973). "The retrieval of harmonics from a covariance function Geophysics". *J. Roy. Astron. Soc.*. **33**, 347-366.

Poskitt, D. S. (2006). "Autoregressive approximation in nonstandard situations: the fractionally integrated and non-invertible cases". *Annals of the Institute of Statistical Mathematics*. **59** (4), 697-725.

Priestley, M. B. (1981). *Spectral Analysis and Time Series*. Academic Press, London.

Prony, R. (1795). "Essai éxperimental et analytique". *J. de L'École Polytechnique*. **2**, 24-76.

Quinn, B. G. and Hannan, E. J. (2001). *The estimation and tracking of frequency*. Cambridge University Press, UK.

Rao, C. R. and Zhao, L. C. (1993). "Asymptotic behavior of maximum likelihood estimates of superimposed exponential signals". *IEEE Trans. Signal Process*. **42**, 1461-1464.

Ray, B. K.(1993). "Modeling long-memory processes for optimal long-range prediction". *Journal of Time Series Analysis*. **14**, 511-525.

Ray, B. and Crato, N. (1996). "Model Selection and Forecasting of Long-range Dependent Processes". *Journal of Forecasting*. **15**, 107-125.

Reisen, V. A. (1993). "Long memory time series models". Ph.D Thesis. Department of Mathematics, UMIST, Manchester, U.K.

Reisen, V. A. (1994). "Estimation of the fractional difference parameter in the ARFIMA(p, d, q) model using the smoothed periodogram". *Journal of Time Series Analysis*. **15**, 335-350.

Reisen, V. A., Abraham, B., and Lopes, S. (2001). "Estimation of parameters in ARFIMA process. A simulation study". *Communications in Statistics: Simulation and Computation*. **30**(4), 787803.

Reisen, V. A. and Lopes, S. (1999). "Some simulation and applications of forecasting long-memory time series models". *Journal of Statistical Planning and Inference*. **80**, 269-287.

Rice, J. A. and Rosenblatt, M. (1988). "On frequency estimation". *Biometrika*. **75**, 477-484.

Rissanen, J. (1978). "Modeling by shortest data description". *Automatica.* **14** (5), 465-471.

Robinson, P. M. (1995). "Log-peroidogram regression of time series with long range dependence". *The Annals of Statistics.* **23**, 1048-1072.

Rosenblatt, M. (1956). "Remarks on some nonparametric estimates of a density function". *The Annals of Mathematical Statistics.* **27**, 832-837.

Said, E. S. and Dickey, D. A. (1984). "Testing for unit roots in autoregressivemoving average models of unknown order". *Biometrika* **71**, 599607.

Sakai, H. (1984). "Statistical analysis of Pisarenkos method for sinusoidal frequency estimation". *IEEE Trans. Acoust., Speech, Signal Processing.* Vol. ASSP-**32**, 95-101.

Scguster, A. (1989). "On the investigation of hidden periodicities with application to a supposed 26 day period of meteorological phenomena". *Terrestrial Magnetism.* **3**, 13-41.

Schwarz, G. (1978). "Estimating the dimension of a model". *Annals of Statistics.* **6** (2), 461-464.

Shibata, R. (1976). "Selection of the order of an autoregressive model by Akaike's information criterion". *Biometrika.* **63**, 117-126.

Smith, J., Taylor, N., and Yadav, S. (1997). "Comparison the bias and mis-specification in ARFIMA models". *Journal of Times Series Analysis.* **18**, 507-527.

Sowell, F. (1992). "Maximum likelihood estimation of stationary univariate fractionally differenced ARMA models". *Journal of Econometrics.* **53**, 165-188.

Stoica, P., Friedlander, B. and Söderström, T. (1987). "Asymptotic bias of the high-order autoregressive estimates of sinusoidal frequencies". *Circuits System Signal Process.* **6** (3),

287-298.

Stoica, P. and Nehorai, A. (1988). "Performance analysis of an adaptive notch filter with constrained poles and zeros". *IEEE Trans.Acoust. Speech Signal Process.* **36**, 911-919.

Stoica, P., Söderström, P. and Ti, F (1989). "Asymptotic properties of the high-order Yule-Walker estimates of sinusoidal frequencies". *IEEE Trans. Acoust., Speech, Signal Processing.* **37** (11), 1721-1734.

Tibshirani, R. (1996). "Regression shrinkage and selection via the lasso". *Journal of the Royal Statistical Society.* Series B, **58** (1), 267-288.

Tichavský, P. and Händel, P. (1995). "Two algorithms for adaptive retrieval of slowly time-varying multiple cisoids in noise". *IEEE Trans. Signal Process.* **43**, 1116-1127.

Tieslau, M. A., P. Schmidt and R. T. Baillie. (1996). "A Minimum-Distance Estimator for Long-Memory Processes". *Journal of Econometrics.* **71**, 249-264.

Tikhonov, A. N. (1943). "On the stability of inverse problems". *C. R. (Doklady) Acad. Sci. URSS (N.S.).* **39**, 176-179.

Tufts, D. W. and Kumaresan, R. (1982). "Estimation of frequencies of multiple sinusoids: Making linear prediction perform like maximum likelihood. *Proc. IEEE.* **70**, 975-989.

Walker, A. M. (1971). "On the estimation of a harmonic component in a time series with stationary independent residuals". *Biometrika.* **58**, 21-36.

Wallis, J. R. and Matalas, N. C. (1970). "Small sample properties of H and K, estimators of the Hurst coefficient h". *Water Resour. Res.,* **6**, 1583-1594.

Watson, G. S. (1964). "Smooth regression analysis". *Sankhyā.* Series A, **26**, 359-372.

Wei, W. W. S. (1990). *Time Series Analysis: Univariate and Multivariate Methods.* Addison-Wesley.

Wei, C. Z. (1992). "On Predictive Least Squares Principles". *Annals of Statistics.* **20** (1), 1-42.

Wu, W. B. and Pourahmadi, M. (2003). "Nonparametric estimation of large covariance matrices of longitudinal data". *Biometrika.* **90** (4), 831-844.

Yule, G. U. (1921). "On the time correlation problem". *Journal of the Royal Statistical Society.* **84**, 497-510.