On Linear MMSE Approximations of Stationary Time Series

by

Syamantak Datta Gupta

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Doctor of Philosophy
in
Electrical and Computer Engineering

Waterloo, Ontario, Canada, 2013

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.

Abstract

In a large number of applications arising in various fields of study, time series are approximated using linear MMSE estimates. Essentially, the time series of interest is estimated as a linear combination of the past values of the series itself, or other time series that influence the former. Such approximations include finite order moving average and autoregressive approximations as well as the causal Wiener filter. In this dissertation, we study two topics related to the estimation of wide sense stationary (WSS) time series using linear MMSE estimates. First, we study the convergence properties of finite order linear MMSE estimates of a WSS time series. Next, we study the problem of detecting causal connections within a family of WSS time series using linear MMSE estimates.

In the first part of this dissertation, we study the asymptotic behaviour of autoregressive (AR) and moving average (MA) approximations. Our objective is to investigate how faithfully such approximations replicate the original sequence, as the model order as well as the number of samples approach infinity. We consider two aspects: convergence of spectral density of MA and AR approximations when the covariances are known and when they are estimated. Under certain mild conditions on the spectral density and the covariance sequence, it is shown that the spectral densities of both approximations converge in L_2 as the order of approximation increases. It is also shown that the spectral density of AR approximations converges at the origin under the same conditions. Under additional regularity assumptions, we show that similar results hold for approximations from empirical covariance estimates.

In the second part of this dissertation, we address the problem of detecting interdependence relations within a group of WSS time series. The objective is to understand the interaction of different time series and to determine whether one series is causally influenced by others. We use Granger-causality as a tool to identify and measure causal connections. Ideally, in order to infer the complete interdependence structure of a complex system, dynamic behaviour of all the processes involved should be considered simultaneously. However, for large systems, use of such a method may be infeasible and computationally intensive, and *pairwise* estimation techniques may be used to obtain sub-optimal results. In this dissertation, we investigate the problem of determining Granger-causality in an interdependent group of jointly WSS time series by using pairwise causal Wiener filters. Analytical results are presented, along with simulations that compare the performance of a method based on finite impulse response (FIR) Wiener filters to another using directed information, a tool widely used in literature. The problem is studied in the context of cyclostationary (CS) processes as well. Finally, a new technique is proposed that allows the determination of causal connections under certain sparsity conditions.

Acknowledgements

First and foremost, I would like to express my sincere gratitude to my supervisor Professor Ravi R. Mazumdar for his invaluable guidance, unwavering support and encouragement throughout the years I spent as a graduate student at the University of Waterloo. Working under his supervision has been a great learning experience for me. Interactions with him has not only helped me expand my academic horizons and grow as a researcher, but has helped me at a personal level as well.

I would like to sincerely thank my committee members, Professor Andrew Heunis, Professor Patrick Mitran, Professor Yulia Gel and Professor Vikram Krishnamurthy for serving on my PhD advisory committee and for their helpful comments and suggestions, which helped me improve my thesis. In addition, the courses taught by Professor Heunis and Professor Mitran have provided me with excellent insight that has been greatly beneficial to my research.

I would also like to thank the University of Waterloo, and in particular, the Department of Electrical and Computer Engineering for providing me with an excellent academic environment that fosters innovative thinking. I am grateful to the staff of the University and the Department for the help I have received as an International Graduate student.

I am thankful to all my friends and colleagues at Waterloo, whose pleasant company made my life at Waterloo a memorable experience and helped keep my spirits high through difficult times. I would cherish their friendship for the rest of my life.

Finally, I would like to thank my parents, who have encouraged and motivated me in the greatest possible way, who have been my role models and whose constant love and support have sustained me throughout my life. I owe all my accomplishments to them.

Dedication

This dissertation is lovingly dedicated to my parents Mrs. Sudeshna Datta Gupta and Dr. Upal Datta Gupta. They have been my deepest source of inspiration.

Table of Contents

List of Tables

List of Figures

List of Symbols

$\mathbb{R}$	Set of real numbers
$\mathbb{Z}$	Set of integers
$\mathbb{N}$	Set of positive integers
$\mathbb{C}$	Set of complex numbers
ℓ_1	Space of all real sequence $\{\xi_i\}$ such that $\sum_{i\in\mathbb{Z}} \lvert\xi_i\rvert < \infty$
ℓ_2	Space of all real sequence $\{\xi_i\}$ such that $\sum_{i\in\mathbb{Z}} \xi_i^2 < \infty$
$\mathbb{E}[X]$	Expected Value of X
$\mathrm{Var}[X]$	Variance of X
$\lvert\lvert F\rvert\rvert$	$\left\lvert\int_{-\frac{1}{2}}^{\frac{1}{2}} \lvert F(\lambda)\rvert^2 d\lambda\right\rvert^{\frac{1}{2}}$, for all functions $F : \left(-\frac{1}{2}, \frac{1}{2}\right] \to \mathbb{C}$ such that $\int_{-\frac{1}{2}}^{\frac{1}{2}} \lvert F(\lambda)\rvert^2 d\lambda < \infty$,
$k = o\{Z\}$	$\lim_{Z\to\infty} \frac{\lvert k(Z)\rvert}{\lvert Z\rvert} = 0$
$k = O\{Z\}$	$\lim_{Z\to\infty} \frac{\lvert k(Z)\rvert}{\lvert Z\rvert} \infty$
$\lVert\cdot\rVert_2$	Euclidean norm for vectors, spectral norm for matrices
$\lVert\cdot\rVert_F$	Frobenius norm
$\{X(n)\}$	A discrete parameter stochastic process where the index n takes values in the set of integers
$\{R_X(k)\}$	$R_X(k) = \mathbb{E}[(X(n) - \mathbb{E}[X(n)])(X(n-k) - \mathbb{E}[X(n-k)])]$, $k \in \mathbb{Z}$, covariance sequence of $\{X(n)\}$, the suffix $_X$ is dropped when the context is clear
$S_X(\lambda)$	$\sum_{k\in\mathbb{Z}} R(k) e^{-2\pi\imath\lambda k}$, $\lambda \in \left(-\frac{1}{2}, \frac{1}{2}\right]$, Spectral density of $\{X(n)\}$ the suffix $_X$ is dropped when the context is clear
$\overline{\mathbb{E}}[\cdot\lvert\cdot]$	Orthogonal projection operator
$H(n-1)$	Linear span of $\{X(n-1), X(n-2), \ldots\}$
$\{a(k)\}$	Wold decomposition parameters of $\{X(n)\}$
$\{b(k)\}$	infinite order autoregressive (AR) parameters of $\{X(n)\}$
$\overline{\mathbb{E}}[X(n)\lvert H(n-1)]$	infinite order AR estimate

$\nu(n)$	$X(n) - \overline{\mathbb{E}}[X(n)\|H(n-1)]$, innovation sequence
$\overline{R}(k)$	Covariance sequence of $\{\overline{\mathbb{E}}[X(n)\|H(n-1)]\}$
$S_{\overline{X}}(\lambda)$	Spectral density of $\{\overline{\mathbb{E}}[X(n)\|H(n-1)]\}$
$\boldsymbol{R_p}$	Covariance matrix of order p
$\boldsymbol{r_p}$	$[R(1)\ R(2)\ \ldots\ R(p)]^T$
$H^p(n-1)$	Linear span of $\{X(n-1), X(n-2), \ldots, X(n-p)\}$
$\overline{X}_p(n)$	AR estimate of order p
$\{b_p(k)\}$	$k = 1, \ldots, p$, p-th order theoretical AR parameters
$\boldsymbol{B_p}$	$[b_p(1)\ \ldots\ b_p(p)]^T$
$\nu_p(n)$	Estimation error corresponding to AR-p estimate
$\overline{R}_p(k)$	Covariance sequence of $\overline{X}_p(n)$
$S_{\overline{X}_p}(\lambda)$	Spectral density of $\overline{X}_p(n)$
$\widetilde{X}_p(n)$	Moving average of order p
$S_{\widetilde{X}_p}(\lambda)$	Spectral density of $\{\widetilde{X}_p(n)\}$
$R_p(k)$	Covariance sequence of $\sum_{j=1}^{p} b(j)X(n-j)$
$\hat{R}_N(k)$	$\frac{1}{N}\sum_{n=\|k\|+1}^{N} X(n)X(n-k)$, Estimate of $R(k)$ computed from a sample of size N
$\hat{\boldsymbol{R}}_{\boldsymbol{p,N}}$	Covariance matrix of order p of the estimated covariances $\hat{R}_N(k)$
$\hat{\boldsymbol{r}}_{\boldsymbol{p,N}}$	$[\hat{R}_N(1)\ \hat{R}_N(2)\ \ldots\ \hat{R}_N(p)]^T$
$\{\hat{b}_{p,N}(k)\}$	$k = 1, \ldots, p$, p-th order empirical AR parameters
$\hat{X}_{p,N}(n)$	Empirical AR-p estimate
$\hat{\boldsymbol{B}}_{\boldsymbol{p,N}}$	$[\hat{b}_{p,N}(1)\ \ldots\ \hat{b}_{p,N}(p)]^T$
$\tilde{R}_{p,N}(k)$	$\mathbb{E}[\hat{X}_{p,N}(n)\hat{X}_{p,N}(n-k)\|\{\hat{\boldsymbol{B}}_{\boldsymbol{p,N}}\}]$, Covariance sequence of $\{\hat{X}_{p,N}(n)\}$, conditioned on the empirical AR parameters
$S_{\hat{X}_{p,N}}(\lambda)$	$\sum_{k\in\mathbb{Z}} \tilde{R}_{p,N}(k)e^{-2\pi\imath\lambda k}$
$\boldsymbol{X}(n)$	$[X_1(n)\ \ldots\ X_N(n)]^T$, $\mathbb{R}^N$-valued discrete time WSS process
$\boldsymbol{\nu}(n)$	$[\nu_1(n)\ \ldots\ \nu_N(n)]^T$, $\mathbb{R}^N$-valued innovation sequence
$H_j(n-1)$	Linear span of $X_j(n-1)$, $X_j(n-2)$, $\ldots$.
$B(k)$	$[b_{i,j}(k)]$, infinite order AR parameters of the $\mathbb{R}^N$-valued process $\{\boldsymbol{X}(n)\}$
$A(k)$	$[a_{i,j}(k)]$, Wold decomposition parameters of the $\mathbb{R}^N$-valued process $\{\boldsymbol{X}(n)\}$
$H_j^p(n-1)$	Linear span of $X_j(n-1)$, $X_j(n-2)$, $\ldots$, $X_j(n-p)$.
$\overline{\mathbb{E}}[X_i(n)\|H_j^p(n-1)]$	Projection of $X_i(n)$ on the subspace $H_j^p(n-1)$

$\xi[X_i(n)\|H_j^p(n-1)]$	Estimation error in estimating $X_i(n)$ from $H_j(n-1)$
$\hat{\mathbb{E}}[(\xi[X_i(n)\|H_j^p(n-1)])^2]$	Estimated mean squared error when $X_i(n)$ is estimated as a linear combination of $X_j(n-1), \ldots, X_p(n-1)$
T_0	Period of a cyclostationary (CS) process $\{X(n)\}$
$R(n,k)$	Periodically varying covariance of a CS process $\{X(n)\}$
$\beta_p(n,k)$	$k = 1, \ldots, p$, periodically varying AR parameters of a CS process $\{X(n)\}$
$\hat{X}_p(n)$	$\sum_{k=1}^{p} \beta_p(n,k) X(n-k)$
$\widetilde{R}(k)$	$\frac{1}{T_0} \sum_{i=1}^{T_0} \mathbb{E}[X(i)X(i-k)]$
$\widetilde{\boldsymbol{B}}_{\boldsymbol{p}}$	$[\widetilde{b}_p(1) \ \ldots \ \widetilde{b}_p(T_0)]^T$, time-invariant AR parameters for a CS process
$\breve{X}_i^p(n)$	$\sum_{k=1}^{p} \sum_{j=1}^{N} \breve{b}_{i,j}(k) X_j(n-k)$, sparse estimate of $X_i(n)$
$\breve{\boldsymbol{b}}_{i,j}^p$	$[\breve{b}_{i,j}^p(1) \ldots \breve{b}_{i,j}^p(p)]^T$
$g_p(\mathbf{c})$	$\int_{-\frac{1}{2}}^{\frac{1}{2}} \left\lvert \sum_{k=1}^{p} c(k) e^{-2\pi\imath\lambda k} \right\rvert d\lambda$ where $\mathbf{c} = [c(1) \ \ldots \ c(p)]^T$
	the suffix $_p$ is dropped when the context is clear
G_i	$\sum_{j=1}^{N} g(\breve{\mathbf{b}}_{i,j}^p)$

Chapter 1

Introduction

A time series is a sequence of data measured at successive instants of time. It can be represented by an indexed set of observations $\{X(n)\}_{n\in\mathbb{T}}$, $\mathbb{T} \subset \mathbb{R}$; where typically, the indexing set $\mathbb{T}$ is the set of integers $\mathbb{Z}$ or the set of positive integers $\mathbb{N}$. Such sequences are frequently encountered in a wide variety of applications, notably in the fields of statistics, econometrics, statistical signal processing and mathematical finance. Examples of time series include the amounts of rainfall at a certain town recorded daily, the daily closing price of a stock, or the population of a country recorded annually.

When the observations are a set of random variables defined over a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and an index set $\mathbb{T} \subset \mathbb{Z}$, a time series is essentially a discrete time stochastic process or a random sequence. If the statistical properties of the process do not change as a function of the index n, it is called a stationary process.

The main purpose of time series analysis is to formulate efficient mathematical models that facilitate characterization of the stochastic processes being studied. Quite often, information related to the statistical behaviour of the process is contained in the past values of the process itself, or in the past values of some other process that influences the original process. Such a system is known as a **causal** system. The objective of modeling is to accurately analyze and quantify the nature of this information flow from past to present. This enables one to extract meaningful information from the set of observations, which can be used to estimate the properties of the processes and to analyze the inter-relations of multiple processes. These, in turn, facilitate the prediction of future values based on known values from the past, i.e., forecast data before they are observed.

A popular approach in time series analysis is to use an estimation technique that minimizes the mean squared error (MSE), which is a common measure of the quality

of estimation. Such an estimator is called a **minimum mean squared error (MMSE)** estimator. When this estimator is a linear function of the values of one or more time series, we call it a **linear MMSE** estimator. Autoregressive (AR) and moving average (MA) estimates, as well as the causal Wiener filter; all belong to this class of estimators.

1.1 Motivation

Analyzing and estimating time series through linear MMSE approximations is a vast topic and finds application in a wide variety of research areas. In this dissertation, two problems related to the estimation of time series using causal linear MMSE estimators are addressed. In the first part, we study the asymptotic properties of the autoregressive and moving average approximations as the model order approaches infinity. In the second part, we analyze the causal interplay among a number of time series through linear MMSE approximations.

1.1.1 Convergence of the spectral density of finite order approximations of stationary time series

Mathematical models for time series often involve a weighted sum of terms representing values sampled in the past, either of the original process itself or of some other process that carries information of the original process. Due to the limitation of computational power, for all practical purpose such an infinite sequence has to be truncated up to a finite number of terms, thereby reducing the exact model of infinite order to an approximate model of finite order. It is of interest to explore how close such a finite order approximation is to the original infinite order process, and how faithfully it replicates the properties of the latter. In particular, one would like to inquire whether the properties of the approximated version converge asymptotically to those of the original sequence as the order of approximation is increased.

While some results are available on the convergence of the approximating finite order autoregressive process in the time domain, there are not many on the spectral properties of the same as the order of approximation approaches infinity. The spectral density of a wide sense stationary (WSS) stochastic process is defined as the discrete time Fourier transform of its covariance sequence and it represents the distribution of power over the frequency domain. Furthermore, the value of the spectral density at the origin (i.e., at frequency $\lambda = 0$) is of special significance for stationary ergodic sequences because of the following invariance principle. Let $\bar{X}_n = \frac{1}{n}\sum_{k=1}^{n} X(k)$ where $\{X(k)\}$ is a WSS ergodic process. Let the value of the spectral density be finite at the origin and denote it by Γ^2. This quantity

is called the time average variance constant (TAVC) of the process ([1, 2]) and according to the following version of the central limit theorem

$$\sqrt{n}(\bar{X}_n - \mu) \Longrightarrow \mathcal{N}(0, \Gamma^2)$$

where $\Longrightarrow$ denotes convergence in distribution.

As a consequence of the above result, Γ^2 plays a major role in the steady-state simulation problem, where the objective is to compute the limit $\lim_{n\to\infty} \bar{X}_n$ where it exists ([3]).

Exact computation of autoregressive (AR) estimates require the knowledge of the covariance sequence of the original time series. However, in many practical applications, the true sequence is unknown and has to be estimated from a finite sample of observations. The limiting behaviour of the spectral density of the AR approximation, when computed empirically from sampled data, is also an issue of interest. In this case, one has to find a relation between the model order p and the sample size N that can guarantee convergence. While there are some results available that study the asymptotic behaviour of the AR estimate computed empirically, many of these assume the associated innovation sequence to be martingale difference. However, in most signal processing applications, as well as in the simulation of Markov processes, the assumption that the driving sequence is a martingale difference is too strong since all that can be guaranteed is stationarity of the underlying stochastic process and the use of resulting L_2 theory.

1.1.2 Inferring underlying causal structures in a family of stationary time series

Inferring dependence relations in a family of random processes from a finite set of observations is a problem encountered in many applications that arise in a diverse variety of fields. Given a family of time series, the objective is to determine whether one process is affected by the other, and, if possible, to quantify this influence. Granger-causality can be used as a tool to measure such causal connections. The objective is to represent the causal interconnections in the form of a connected graph, like the one in figure 1.1, where nodes indicate individual processes and directed edges indicate causal influences.

Ideally, in order to infer the complete interdependence structure of a complex system, one should simultaneously consider the dynamic behaviour of all the processes involved. However, for a large system, use of such a method may be infeasible due to computational burden. An alternative approach is to consider each pair of processes separately, detect whether one causes the other, and finally use this information to infer about causal links within the entire group. Examples of such *pairwise* techniques include the Wiener filter and

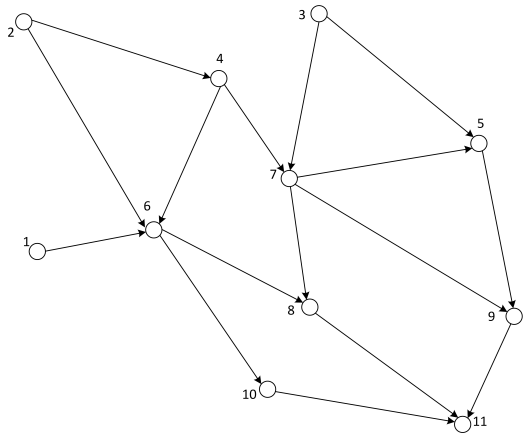

Figure 1.1: An Example of a graph representing causal interdependence within a family of random processes

directed information. It is of interest to know to what extent such pairwise methods may reveal the original interdependence relations of the system. Not many analytical results, however, are available on the topic.

Furthermore, when causal interactions among time series are inferred and represented as a graph, it is preferable to have a depiction which identifies and preserves only the edges corresponding to the *strongest* dependences. Such a representation is devoid of weak or spurious links and is much easier to interpret and use, compared to a complicated mesh of many interconnections. There is therefore a need to incorporate the notion of parsimony in the estimation technique through a sparsity constraint of some form, so that only the most significant edges are preserved.

1.2 Contributions

In the first part of this dissertation, we analyze the asymptotic behaviour of the spectral density of linear MMSE approximations of wide sense stationary time series. We consider convergence with respect to an L_2 norm defined over the frequency domain. Under a mild regularity condition, we show that, as the model order approaches infinity,

- The spectral density of the moving average (MA) estimate converges to that of the original process in L_2.

- The spectral density of the autoregressive (AR) estimate converges to that of an infinite order AR approximation in L_2.
- The time average variance constant (TAVC) of the AR estimate converges to that of the infinite order AR approximation.

Next, we consider the case when AR parameters are computed empirically using the *estimated* covariance sequence. Under additional conditions, we show that as the model order p and the number of observations N both approach infinity, as long as $p = o(N^{1/3})$,

- The AR parameters converge to the parameters corresponding to the infinite order AR approximation in mean square.
- The spectral density of the AR approximation converges to that of an infinite order AR approximation in mean over L_2.
- The time average variance constant (TAVC) of the AR approximation converges to that of the infinite order AR approximation in mean.

In the second phase of this dissertation, we investigate the utility of the pairwise causal Wiener filter in detecting Granger-causality within a group of jointly wide sense stationary real-valued time series.

- We present analytical results on the efficacy of the causal Wiener filter in detecting Granger-causality.
- We compare its performance with that of another popular pairwise estimation technique, namely, directed information, under a Gaussian framework and show that the results are comparable.
- We derive a technique to estimate cyclostationary processes through *time-invariant* AR estimates and extend the method to detect Granger-causality within a family of cyclostationary processes.
- Finally, we present a technique that infers interdependence relations by eliminating weaker connections and preserving only the strongest ones.

In the remainder of this chapter, we review some of the theoretical preliminaries that are relevant to this research.

1.3 Wide sense stationary processes

Definition 1.3.1. ***Wide Sense Stationary (WSS) Process****: Let $\{X(t)\}_{t\in\mathbb{T}}$ be a real-valued stochastic process defined over a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ (with $\mathbb{T} \subset \mathbb{R}$). Let $\mathbb{E}[\ \cdot\ ]$ denote mathematical expectation and let $\rho_X(\cdot,\cdot)$ denote the covariance of the process $\{X(t)\}$, i.e.,*

$$\rho_X(t_1,t_2) = \mathbb{E}\Big[\Big(X(t_1) - \mathbb{E}[X(t_1)]\Big)\Big(X(t_2) - \mathbb{E}[X(t_2)]\Big)\Big]$$

Then $\{X(t)\}$ is said to be ***wide sense stationary (WSS)*** *if*

(i) It is second order, i.e., $\mathbb{E}[X(t)^2] < \infty$ for all $t \in \mathbb{T}$

and

for all t_1, t_2, τ such that $t_1, t_2, t_1+\tau, t_2+\tau \in \mathbb{T}$,

(ii) $\mathbb{E}[X(t_1)] = \mathbb{E}[X(t_1+\tau)]$

(iii) $\rho_X(t_1, t_1+\tau) = \rho_X(t_2, t_2+\tau)$

For a WSS process, covariance is only a function of the separation between the two time instants. As such, it is written as a function with a single argument, denoted by $R_X(\tau)$:

$$\rho_X(t_1, t_1-\tau) = R_X(\tau) \tag{1.3.1}$$

When the context is clear, $R_X(\tau)$ is denoted by $R(\tau)$. In particular, for discrete time WSS processes, τ can only take integer values. In such cases, the covariance sequence is defined as the sequence $\{R(k)\}_{k\in\mathbb{Z}}$. This notation is used in the remainder of this dissertation while dealing with discrete time WSS stochastic processes.

When for a WSS process $\{X(n)\}$, $\mathbb{E}[X(n)] = 0$, the covariance sequence $\{R(k)\}$ is given by

$$R(k) = \mathbb{E}[X(n)X(n-k)]$$

Through the remainder of this dissertation, all processes are considered zero-mean, unless mentioned otherwise.

Definition 1.3.2. ***Jointly Wide Sense Stationary (WSS) Processes****: Let* $\{X(n)\}$, $\{Y(n)\}$ *be two WSS processes defined over the probability space* $(\Omega, \mathcal{F}, \mathbb{P})$ *. The two processes are said to be* ***jointly WSS*** *if*

$$\mathbb{E}[X(n)Y(n-k)] = \mathbb{E}[X(n+\tau)Y(n+\tau-k)] \text{ for all } \tau, k \in \mathbb{Z}$$

The quantity $\mathbb{E}[X(n)Y(n-k)]$ is denoted by $R_{X,Y}(k)$ and is termed as the cross-covariance of $\{X(n)\}$ and $\{Y(n)\}$. Note that while the covariance sequence of a real-valued WSS process is symmetric (i.e., $R(k) = R(-k)$), the same does not hold for the cross-covariance of two real-valued jointly WSS processes. In general, $R_{X,Y}(k) \neq R_{X,Y}(-k)$.

WSS processes are characterized by a quantity known as the power spectral density (often simply referred to as spectral density), which provides a frequency domain representation of the covariance.

Definition 1.3.3. ***Spectral Density****: Let* $\{X(t)\}$ *be a WSS process with covariance* $R_X(\tau)$*. The spectral density (or power spectral density) of* $\{X(t)\}$ *is defined as the Fourier transform of the covariance* $R_X(\tau)$ *([4], P-208). The spectral density of a real-valued discrete time WSS process with covariance sequence* $\{R_X(k)\}$ *is defined as the discrete time Fourier Transform of* $\{R_X(k)\}$ *and is given by*

$$S_X(\lambda) = \sum_{k\in\mathbb{Z}} R_X(k)e^{-2\pi\imath\lambda k}, \ \lambda \in \left(-\frac{1}{2}, \frac{1}{2}\right]$$

When the context is clear, the suffix $_X$ is removed.

The covariance sequence can be represented in terms of the spectral density as follows:

$$R(k) = \int_{-1/2}^{1/2} S(\lambda)e^{2\pi\imath\lambda k}d\lambda$$

The Fourier Transform of a function $f : \mathbb{R} \to \mathbb{C}$ is defined when $f(t)$ is integrable or square-integrable, i.e., when $f \in L_1$ or $f \in L_2$. When $f \in L_1$, i.e., when

$$\int_{\mathbb{R}} |f(t)|dt < \infty$$

the Fourier Transform of f is defined as ([5], P-9):

$$\hat{f}(\lambda) = \int_{\mathbb{R}} f(t)e^{-2\pi\imath\lambda t}dt \tag{1.3.2}$$

Under the additional condition that

$$\int_{\mathbb{R}} |\hat{f}(\lambda)| d\lambda < \infty$$

the inversion formula

$$f(t) = \int_{\mathbb{R}} \hat{f}(\lambda) e^{2\pi\imath\lambda t} d\lambda$$

exists for almost all t. On the other hand, when a function f is square-integrable, i.e., when $f \in L_2$ with

$$\int_{\mathbb{R}} |f(t)|^2 dt < \infty$$

the Fourier Transform is defined in the following way. First, one defines the same when $f \in L_1 \cap L_2$, i.e., when it is both integrable and square-integrable, according to equation (1.3.2). The definition is then extended to any square-integrable function using the fact that $L_1 \cap L_2$ is dense in L_2 ([5], P-157). It follows from Parseval's identity that the Fourier transform of a square-integrable function is square-integrable itself.

It follows that the spectral density of a WSS process exists when $R(\tau)$ is an integrable or square-integrable function of τ. For a discrete time stochastic process, integrability and square-integrability translate to summability and square-summability respectively. Moreover, a summable sequence is also square-summable.

Likewise, for jointly WSS processes, $\{X(n)\}$ and $\{Y(n)\}$ the cross-spectral density $S_{X,Y}(\lambda)$ is given by the discrete time Fourier transform of the cross-covariance sequence $\{R_{X,Y}(k)\}$.

$$S_{X,Y}(\lambda) = \sum_{k=-\infty}^{\infty} R_{X,Y}(k) e^{-2\pi\imath\lambda k}$$

1.4 Causal linear MMSE estimators of WSS processes

Let $\{X(n)\}$ and $\{Y(n)\}$ be two real-valued discrete time WSS process defined on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and the former is to be estimated using information from the latter. Let $\mathcal{F}_Y$ be the σ-field generated by $\{Y(n)\}$ ($\mathcal{F}_Y \subset \mathcal{F}$). An estimator of $X(n)$, given the process $\{Y(n)\}$, is a $\mathcal{F}_Y$-measurable function $\hat{X}(n)$.

Definition 1.4.1. *The minimum mean squared error (MMSE) estimate of $X(n)$, given the process $\{Y(n)\}$ is defined as the estimator $\hat{X}^*(n)$ such that the mean squared error $\mathbb{E}[(X(n) - \hat{X}^*(n))^2]$ is a minimum among all $\mathcal{F}_Y$-measurable functions $\hat{X}(n)$.*

The MMSE estimator, in essence, is the conditional expectation of $X(n)$ given $\mathcal{F}_Y$, written as $\mathbb{E}[X(n)|\mathcal{F}_Y]$. It is the "best" estimator of $X(n)$ among all *measurable functions* of $\{\dots, Y(-1), Y(0), Y(1), \dots\}$.

While $\mathbb{E}[X(n)|\mathcal{F}_Y]$ is the "best" estimator of $X(n)$ given the complete information of $\{Y(n)\}$, in general, depending on the statistical properties of the processes involved, it may not be tractable. In such cases, it is often useful to use the *linear* MMSE estimator.

Definition 1.4.2. *The linear MMSE estimator of $X(n)$, given the process $\{Y(n)\}$, is the estimator $\hat{X}^*(n)$ where*

$$\hat{X}^*(n) = \sum_{k=-\infty}^{\infty} a^*(k) Y(n-k)$$

and the parameters are chosen so that the mean squared error

$$\mathbb{E}\left[\left(X(n) - \sum_{k=-\infty}^{\infty} a(k) Y(n-k)\right)^2\right]$$

is minimized when $a(k) = a^*(k)$ for each k.

Essentially, $\hat{X}^*(n)$ is the "best" estimator of $X(n)$ among all the elements in the *closure* of the set of all linear combinations of $\{\dots, Y(-1), Y(0), Y(1), \dots\}$.

In the context of time varying processes, an estimator is termed as causal if it does not use values from the future to estimate the present. Define $\mathcal{F}_Y(n)$ as the σ-field generated by $\{Y(n), Y(n-1), \dots\}$. Then the causal MMSE estimate of $X(n)$ given $\{Y(n)\}$ is $\mathbb{E}[X(n)|\mathcal{F}_Y(n)]$.

Finally, the causal linear MMSE estimate is defined as follows.

Definition 1.4.3. *The causal linear MMSE estimate of $X(n)$ given $\{Y(n)\}$ is an element in the closure of the set of linear combinations of the present and past values of $Y(n)$ of the form $\hat{X}^*(n) = \sum_{k=0}^{\infty} a^*(k) Y(n-k)$ such that the mean squared error $\mathbb{E}[(X(n) - \hat{X}^*(n))^2]$ is a minimum for $\hat{X}^*(n)$ among all linear combinations $\sum_{k=0}^{\infty} a(k) Y(n-k)$.*

When the above estimate includes a finite number of terms from the past of $\{Y(n)\}$, we call it a finite-order linear estimate. In the general form, this is also tantamount to the **causal Wiener filter** (when *all* the past values of $\{Y(n)\}$ are used in estimation) and the **finite impulse response (FIR) Wiener filter** (when only a finite number of past values of $\{Y(n)\}$ are used).

Let $\{X(n)\}$ and $\{Y(n)\}$ be two jointly WSS processes. The FIR Wiener filter of order p that estimates $X(n)$ using $\{Y(n)\}$ is given by

$$\hat{X}(n) = \sum_{k=0}^{p} w_p(k) Y(n-k)$$

where the parameters $\{w(n)\}$, also known as the filter coefficients, minimize the mean squared error. They can be computed by solving the **Wiener-Hopf equations**, given by

$$\sum_{k=0}^{p} w(j) R_Y(k-j) = R_{X,Y}(k) \text{ for } k = 0, \ldots, p$$

where $\{R_Y(k)\}$ is the covariance sequence of $\{Y(n)\}$ and $\{R_{X,Y}(k)\}$ is the cross-covariance sequence of $\{X(n)\}$ and $\{Y(n)\}$. In this dissertation, we would be interested in estimating $X(n)$ using only the *past* values of $\{Y(n)\}$ (starting from $Y(n-1)$). The Wiener filter estimate in that case is a linear combination of $\{Y(n-1), \ldots, Y(n-p)\}$ that minimizes the mean squared estimation error.

In the remainder of this section, two more examples belonging to this class of estimators are introduced. These are the **moving average (MA)** estimate and the **autoregressive (AR)** estimate. We also discuss the **autoregressive moving average (ARMA)** model, which is a combination of the MA and AR models.

Definition 1.4.4. ***Moving Average (MA) Estimate****: A moving average estimate of order p is the linear MMSE estimate of $X(n)$ of the form*

$$\widetilde{X}_p(n) = \sum_{k=0}^{p} a_p(k) \nu(n-k)$$

where $\{\nu(n)\}$ is a white noise sequence, i.e.,

$$\mathbb{E}[\nu(n)\nu(n-k)] = \begin{cases} \sigma_\nu^2 & \text{if } k = 0 \\ 0 & \text{otherwise} \end{cases}$$

It is thus a representation of $X(n)$ in terms of a finite number of terms from a sequence of white noise where the parameters $\{a_p(k)\}$ minimize the mean squared error.

Definition 1.4.5. ***Autoregressive (AR) Estimate****: An autoregressive estimate of order p for a real-valued WSS process $\{X(n)\}$ is the linear MMSE estimate of $X(n)$ of the form*

$$\overline{X}_p(n) = \sum_{k=1}^{p} b_p(k) X(n-k)$$

Thus it represents $X(n)$ in terms of its p most recent values, where the parameters $\{b_p(k)\}$ minimize the mean squared error..

A more general estimation approach is the ARMA model, of which both the MA and AR models are special cases. An ARMA model is defined as follows.

Definition 1.4.6. ***Autoregressive Moving Average (ARMA) Model****: An autoregressive moving average model of order p, q (written as ARMA(p,q)) for a real-valued WSS process $\{X(n)\}$ is the linear MMSE estimate of $X(n)$ of the form*

$$\check{X}_{p,q}(n) = \sum_{j=0}^{q} c_q(j)\nu(n-j) + \sum_{k=1}^{p} d_p(k)X(n-k)$$

where $\{\nu(n)\}$ is a white noise sequence. The parameters $\{c_q(k)\}$ and $\{d_p(k)\}$ minimize the mean squared error.

The **Yule-Walker equations** ([4, 6, 7]) enable one to find the "best" ARMA model of a given order. The equations are based on the principle of least square estimation and are obtained by minimizing the mean squared error in estimation. Let $h(n)$ be the causal impulse response of a WSS process $\{X(n)\}$ with covariance sequence $\{R(k)\}$. The Yule-Walker equations for fitting an ARMA(p,q) model to $\{X(n)\}$ with parameters $\{a(j)\}$ and $\{b(k)\}$ are given by:

$$R(k) + \sum_{k=1}^{p} b(j)R(k-j) = \begin{cases} \sigma_\nu^2 c(k) & \text{if } 0 \le k \le q \\ 0 & \text{if } q < k \le p+q \end{cases}$$

where

$$c(k) = \sum_{j=0}^{q-k} a(j+k)h^*(j)$$

and $\sigma_\nu^2 = \mathbb{E}[\nu^2(n)]$. These equations can be used to estimate a zero-mean WSS discrete time stochastic process whose covariance sequence is known. If the original covariance sequence $\{R(k)\}$ is not available, $R(k)$s may be replaced with their empirical values, computed through sampled observations.

$$\hat{R}_N(k) = \begin{cases} \frac{1}{N}\sum_{n=|k|+1}^{N} X(n)X(n-|k|), & |k| \le N-1 \\ 0, & |k| \ge N \end{cases}$$

where N is sufficiently large.

1.5 Granger-causality

Granger-causality is a mathematical tool that is widely used to quantify causal relations between WSS processes [8, 9]. A process $\{X(n)\}$ is said to Granger-cause another process $\{Y(n)\}$ if the mean squared error in estimating $Y(n)$ from the past values of $Y(n)$ (i.e., $Y(n-1), Y(n-2), \dots$) is greater than that in estimating $Y(n)$ from the past observations of both $\{X(n)\}$ and $\{Y(n)\}$ combined $\Big($ i.e.,$Y(n-1), Y(n-2), \dots$ and $X(n-1), X(n-2), \dots\Big)$. In other words, the past values of $\{X(n)\}$ carry additional information on $Y(n)$ that is not available in the past values of $Y(n)$ itself and therefore an inclusion of these values reduce the error in estimation.

Definition 1.5.1. *Consider a system of two WSS processes $\{X(n)\}$, $\{Y(n)\}$. Let the system have a model order p. $\{Y(n)\}$ is first modeled as an univariate autoregressive process of order p with error θ_Y, i.e.,*

$$Y(n) = \sum_{i=1}^{p} a(i)Y(n-i) + \theta_Y(n)$$

where the parameters $\{a(i)\}$ minimize the mean squared error, and then modeled as an autoregression that also includes past observations of $\{X(n)\}$ with error $\theta_{Y,X}$:-

$$Y(n) = \sum_{i=1}^{p} b(i)Y(n-i) + \sum_{i=1}^{p} c(i)X(n-i) + \theta_{Y,X}(n)$$

where the parameters $\{b(i)\}$, $\{c(i)\}$ minimize the mean squared error. $\{X(n)\}$ is said to Granger-cause $\{Y(n)\}$ if

$$\mathbb{E}[\theta_Y^2] > \mathbb{E}[\theta_{Y,X}^2]$$

Following Geweke [10, 11], the *extent* of Granger-causality can be measured through the quantity known as Wiener-Granger causality from $\{X(n)\}$ to $\{Y(n)\}$, given by

$$F_{X\to Y} = \ln\left(\frac{\mathbb{E}[\theta_Y^2]}{\mathbb{E}[\theta_{Y,X}^2]}\right)$$

Granger-causality attempts to detect causality between time series by comparing mean squared estimation errors. It may be noted that Granger-causality is only a tool to analyze how a number of processes are inter-related, whether one process is truly *caused* by another, is a deeper problem and of a more abstract nature.

1.6 A Hilbert space of square-integrable random variables

The idea of a Hilbert space, i.e., a complete, inner product space, plays a fundamental role in the theory of functional analysis. In this section, we briefly discuss how a Hilbert space can be defined in the context of random variables ([12], P-15).

We begin by considering the space of all real-valued square integrable random variables on $(\Omega, \mathcal{F}, \mathbb{P})$, i.e., random variables $X(\omega)$ with $\mathbb{E}[X^2] < \infty$ and call this space $L_2(\mathbb{P})$. Without loss of generality, let X, $Y \in L_2(\mathbb{P})$ be two zero-mean random variables and define

$$\langle X, Y \rangle = \mathbb{E}[XY]$$

It is easily verified that for X, Y, $Z \in L_2(\mathbb{P})$ and α, $\beta \in \mathbb{R}$,

(i) $\langle X, Y \rangle = \langle Y, X \rangle$

(ii) $\langle \alpha X + \beta Y, Z \rangle = \alpha \langle X, Z \rangle + \beta \langle Y, Z \rangle$
and

(iii) $\langle X, X \rangle = \mathbb{E}[X^2] \geq 0$

Moreover, $\mathbb{E}[X^2] = 0 \Leftrightarrow X = 0$ almost surely.

Therefore, $\langle \cdot, \cdot \rangle$ satisfy the required properties of an inner product on $L_2(\mathbb{P})$. The naturally defined norm of a random variable X, based on this inner product, is $\sqrt{\mathbb{E}[X^2]}$. When X has zero mean, this is equal to the standard deviation of X. It can be further shown that this space is complete with respect to the corresponding norm ([13], [12]). An immediate consequence is the following theorem ([12], P-21, Theorem 2.4).

Theorem 1.6.1. *$L_2(\mathbb{P})$ is a Hilbert space.*

Recall that two elements a, b in a Hilbert space are said to be orthogonal to each other if their inner product $\langle a, b \rangle$ is zero. In this case, $a, b \in L_2(\mathbb{P})$ are orthogonal to each other when $\mathbb{E}[ab] = 0$.

A very important result from the theory of Hilbert spaces is the projection theorem ([5], P-139).

Theorem 1.6.2. ***Projection Theorem****: Let G be a Hilbert subspace of the Hilbert space H and let $G^{\perp}$ be defined as*

$$G^{\perp} = \{z \in H : \langle z, x \rangle = 0 \text{ for all } x \in G\}$$

Let $x \in H$. Then there exists a unique element $y^* \in G$ such that $x - y^* \in G^{\perp}$, and

$$||x - y^*|| = \inf_{y \in G} ||x - y||$$

The element y^* is called the ***orthogonal projection*** *of x on G.*

Let H be a Hilbert space of real-valued second order random variables and G be a Hilbert subspace of H. Let $x \in H$. Then, the orthogonal projection of x on G is the (almost surely) unique element y^* in G that minimizes $\xi = ||y - x||^2 = \mathbb{E}[(y - x)^2]$ for $y \in G$. If one considers y as an estimate of x, given the subspace G, then ξ is the mean squared error in estimation. Therefore, the orthogonal projection of x on G gives the minimum mean squared error (MMSE) estimator of x, given G. This theorem is thus of great importance to estimation theory.

The theory in the next section is largely developed by exploiting properties of Hilbert spaces in the context of second order processes.

1.7 Wold decomposition theorem

The Wold decomposition theorem enables a decomposition of any WSS process into an infinite sum of orthogonal elements. In this section, we discuss the theorem for a real-valued stochastic process. Note that the theorem is valid when $\{\boldsymbol{X}(n)\}$ is an $\mathbb{R}^N$-valued process.

Given a real-valued second order stochastic process $\{X(t), t \in \mathbb{T}\}$ defined on $(\Omega, \mathcal{F}, \mathbb{P})$ we can construct the space of all finite linear combinations of the form

$$\sum_{i=0}^{n} \alpha(i) X(t_i), \ \alpha(i) \in \mathbb{R}, \ t_i \in \mathbb{T}$$

and also include their limits in the mean square when they exist. The space H_X, the closure of all such linear combinations, is a Hilbert space with respect to the inner product $\langle \cdot, \cdot \rangle$ defined above. It is called the Hilbert space generated by (the linear combinations of) the process $\{X(t)\}$, or (the closure of) the linear span of $\{X(t)\}$ ([12], P-21).

For a second order discrete time stochastic process $\{X(n)\}_{n\in\mathbb{Z}}$, we define $H_X(n)$ as the linear span of $\{X(n), X(n-1), X(n-2), \ldots\}$, i.e., the closure of all linear combinations of $X(n)$ and all its past values. When there is no cause of ambiguity, the subscript X is dropped and it is simply written as $H(n)$.

$H(n)$ is a Hilbert space with the inner product $\mathbb{E}[\cdot,\cdot]$. $X(n)$ is an element of $H(n)$ but is (in general) not an element of the subspace $H(n-1)$. Define $\overline{\mathbb{E}}[Y|H(n)]$ as the projection of the random variable Y onto the space $H(n)$. By the projection theorem, $X(n)$ can be written as

$$X(n) = \overline{\mathbb{E}}[X(n)|H(n-1)] + \eta_X(n)$$

where the quantity $\eta_X(n)$ is an element of $H(n)$ but is orthogonal to the subspace $H(n-1)$.

In the same way, $\overline{\mathbb{E}}[X(n)|H(n-1)]$ (which is an element of $H(n-1)$) can further be decomposed as a sum of two orthogonal components, one being its projection on the subspace $H(n-2)$ of $H(n-1)$ and the other being the part orthogonal to $H(n-2)$. It is easy to see that the second component, being an element of $H(n-1)$, is also orthogonal to $\eta_X(n)$. If the procedure is repeated indefinitely, a representation of $X(n)$ is obtained in terms of the orthogonal elements $\eta_X(n)$, $\eta_X(n-1)$, ... and the space $\cap_{n\in\mathbb{Z}} H(n)$.

The process $\{\eta_X(n)\}_{n\in\mathbb{Z}}$ is called the innovation process of X.

Definition 1.7.1. ***Innovation Process****: Given a discrete time WSS process $\{X(n)\}$, the innovation process $\{\eta_X(n)\}$ is defined as*

$$\eta_X(n) = X(n) - \overline{\mathbb{E}}[X(n)|H_X(n-1)]$$

Again, the subscript X is dropped when the context is clear. By construction, $\eta(i)$ and $\eta(j)$ are orthogonal to each other, i.e., $\mathbb{E}[\eta(i)\eta(j)] = 0$, whenever $i \neq j$. It can also be shown that when $\{X(n)\}_{n\in\mathbb{Z}}$ is WSS, $\{\eta(n)\}_{n\in\mathbb{Z}}$ is also WSS ([12], P-27). Without loss of generality, it is assumed that $\mathbb{E}[\eta^2(n)] = 1$.

Let the linear span of $\{\eta(n)\}_{n\in\mathbb{Z}}$ be denoted by $N(n)$. The space $\cap_{n\in\mathbb{Z}} H(n)$ is denoted by $H_{-\infty}$ and corresponds to the "remote past" of $\{X(n)\}$. It can be shown that $N(n) \oplus H_{-\infty} = H(n)$. The statement of the Wold decomposition theorem is as follows ([12], P-28):

Theorem 1.7.1. ***Wold decomposition theorem****: Let $\{X(n)\}_{n\in\mathbb{Z}}$ be a discrete time real-valued WSS process and let $\{\eta(n)\}_{n\in\mathbb{Z}}$ be the corresponding innovation process. Then, $X(n)$ can be written as*

$$X(n) = U(n) + V(n) \tag{1.7.3}$$

where

(i) $U(n)$ is the projection of $X(n)$ on $N(n)$, i.e.,

$$U(n) = \overline{\mathbb{E}}[X(n)|N(n)]$$

and has the representation

$$U(n) = \sum_{k=0}^{\infty} a(k)\eta(n-k)$$

(ii) $V(n)$ is a (linearly) deterministic process. It is the projection of $X(n)$ on $H_{-\infty}$, i.e.,

$$V(n) = \overline{\mathbb{E}}[X(n)|H_{-\infty}]$$

A detailed derivation of the Wold decomposition theorem can be found in [12]. The proof uses the same construction described in this section.

The above theorem allows us to write $X(n)$ as

$$X(n) = \sum_{k=0}^{\infty} a(k)\eta(n-k) + V(n)$$

It is clear from the preceding discussion that $a(0) = 1$. When $\{X(n)\}$ is a (linearly) deterministic process, it is completely determined from $H(n-1)$; i.e., $X(n) \in H(n-1)$ for all n, and the innovation process is the zero-sequence. In that case,

$$X(n) = V(n)$$

When, $\{X(n)\}$ is purely non-deterministic or **regular**, $H_{-\infty} = \{0\}$, so that

$$X(n) = \sum_{k=0}^{\infty} a(k)\eta(n-k)$$

When $\{X(n)\}$ is a zero-mean, regular WSS process, the innovation process $\{\eta(n)\}$ is WSS and zero-mean as well. Moreover, by construction, $\langle \eta(i), \eta(j) \rangle = 0$ for $i \neq j$ and $||\eta(n)|| = 1$ for all n. Thus the above construction corresponds to a moving average-type representation of $\{X(n)\}$ of infinite order.

On the other hand, by the projection theorem, $\overline{\mathbb{E}}[X(n)|H(n-1)]$ gives the minimum mean squared error estimate of $\{X(n)\}$ given the subspace $H(n-1)$. As $H(n-1)$ is the

closure of the vector subspace spanned by the linear combinations of the past values of $\{X(n)\}$, $\overline{\mathbb{E}}[X(n)|H(n-1)]$ can be seen as an infinite order autoregressive approximation of $X(n)$.

If we define $H^p(n-1)$ as the linear span of $\{X(n-1), X(n-2), \ldots, X(n-p)\}$, then $\overline{\mathbb{E}}[X(n)|H^p(n-1)]$ gives the pth order autoregressive approximation of $X(n)$. Likewise, the p-th order FIR Wiener filter estimate of $X(n)$ from the past values of $\{Y(n)\}$ can be seen as the projection of $X(n)$ on the subspace $H_Y^p(n-1)$, the linear span of $\{Y(n-1), Y(n-2), \ldots, Y(n-p)\}$. The infinite impulse response (IIR) causal Wiener filter is a projection of $X(n)$ on $H_Y(n-1)$.

1.8 Organization of the thesis

In this chapter we have presented the motivation behind this research and have listed the key contributions of this dissertation. We have also introduced some of the theoretical concepts that would be used throughout this dissertation. The rest of this dissertation is organized as follows. In chapter 2, a short survey of existing research literature on related topics is presented. Results on the convergence of the spectral densities of the MA and AR estimates using true covariances are discussed in chapter 3. In chapter 4, results on the convergence of the spectral density of the AR estimate using empirical AR parameters are presented, along with some simulation examples. This is followed by some results on the performance of the pairwise causal Wiener filter in inferring Granger-causality in large families of WSS time series in chapter 5. In chapter 6, we discuss a time-invariant AR estimate for cyclostationary processes and extend this method to analyze Granger-causality between two cyclostationary processes. In chapter 7, we present a technique that infers Granger-causality under certain sparsity constraints on the interdependence relations. Finally, the main contributions of this research are reviewed in chapter 8 and possible directions of future research are discussed.

Chapter 2

Literature Review

Analysis and estimation of time series through linear MMSE approximations have a long and rich history of research. In this chapter, we discuss some of the major contributions already available in literature. We begin with a brief introduction on the history of the development of the estimation problem. Due to the vastness of the subject, in the subsequent sections, our focus will be restricted to some specific aspects of the problem that are relevant to this dissertation.

2.1 Introduction

A major breakthrough in the area of MMSE estimation was achieved in the late 1920's and early 1930's due to the contributions of Udny Yule and Gilbert Walker, who developed the well-known Yule-Walker equations. In 1927, Yule [14] suggested the use of autoregressive models to analyze time series data related to sunspots, in lieu of the then established periodogram method. These methods were developed further by Walker in 1931 [15]. About the same time, Eugene Slutsky is attributed to have first demonstrated the method of moving averages in the context of real data [16]. Shortly after, Norbert Wiener published his work [17] on an optimal predictor that minimizes the mean squared estimation error in case of a stationary signal mixed with additive noise.

The problem of estimating an infinite order process and its spectral density using a finite order model is encountered in many applications, and has a long history of research [18]. A detailed survey of the key papers and results in this area is available in [19]. In general, there are two approaches for estimating the spectral density of a WSS process: the **parametric** and the **non-parametric**. In the parametric methods one first solves the

Yule-Walker equations to fit an ARMA model to the available data and then obtains the approximate power spectral density using directly the empirically computed parameters. On the other hand, in the non-parametric approach, the spectral density is estimated as an approximation of the discrete time Fourier Transform of the empirical covariance sequence, given by:

$$\hat{S}_X(\omega) = \sum_{k=-N}^{N} \hat{R}_{k,N} e^{-2\pi\imath k\omega}$$

2.2 AR approximations and their asymptotic behaviour

The autoregressive (AR) estimation problem becomes more complicated when there is no a priori knowledge on the order of the process being estimated. In such cases one has to estimate an order for the model, based on the available information so that it can provide a "best fit" to the observed data. There are a number of problems that have been studied related to the choice of an optimal order and much research has been devoted for such choices.

The basic principle behind Akaike's "Final Prediction Error" Criterion (FPE) [20], [21] is a minimization of mean squared error for the one-step-ahead predictor developed using the proposed model in comparison with an independent sequence of the original process. The AR model order for which this error is the least, is chosen as the optimal order.

Another approach suggested by the same author is the famous Akaike Information Criterion (AIC) [22]. Essentially, this method chooses a model order for which the Kullback-Leibler distance between the model and the original process is the least. Several candidate models for a given process can be ranked according to their AIC, with the model having the least AIC providing the best estimate. Among a wide variety of applications, this method can also be used to obtain the "best" order for an AR model of a discrete time WSS process.

Related issues are the estimation of the power spectral density in the work of Parzen [23, 24, 25]. In [23], the author attempted to compare the time domain approach of the innovation process to the spectral approach in the context of estimating stationary time series. In [24] and [25], the author reviewed the problem of estimating an infinite order AR process from a finite order model and developed a criterion for selecting the order of the AR model. This approach involves the introduction of the "Criterion AR Transfer Function" or CAT function. For each model order M, this criterion computes the integrated relative mean squared error of the spectral estimate of an AR(M) model and chooses that M for which this error is minimized. A generalization of this approach was presented in [26].

Other methods for model order selection include the Hannan-Quinn Criterion (HQC) method [27] and the Bayesian Information Criterion (BIC) or the Schwarz Criterion [28, 29]. These approaches are also based on information theoretic considerations and can be used as alternatives to the AIC.

In [30], the AR and the non-parametric "window" spectral estimation were compared on the basis of an extensive simulation of series constructed from a variety of models. The authors also commented on the relative merits of the FPE, AIC and CAT; and compared the performances of the Burg method and the Yule-Walker Equations with respect to spectral estimation.

Optimal order selection for AR models and various application still remains an active research area. Recent publications include [31], [32] and [33]. Among these, [31] did a comparative study of the AIC, HQC and BIC on the basis of simulation results while the other two papers suggested slight modifications of the existing methods.

There have been a few works that have explored the nature and rate of convergence of finite order estimates of WSS time series in the time and frequency domain. In [34], asymptotic properties of finite-order AR approximations of an infinite-order AR process were considered. It was shown that the model order selected through Akaike's information criterion (AIC) and its variants are asymptotically efficient. The asymptotic properties of the AR approximation when computed from the same realization of a time series were discussed in the works of Ing and Wei [35, 36]. As an extension of the result in [34], it was shown in [36] that the asymptotic efficiency of the AIC is valid even when the predictors are computed from the same realization.

A result on the pointwise convergence of the AR parameters was presented in [37]. A very useful result in this context is Baxter's inequality [38]. This result and its various corollaries have been used in the literature to obtain convergence results that are similar to the ones obtained in this research. In [39], it was shown that under certain assumptions on the relative asymptotic rates of the model order and the number of observations, estimates of the spectral density obtained from finite-order AR estimates are consistent. In the works of Pourahmadi [40, 41, 42], the nature and rates of convergence of the AR parameters were discussed for univariate and multivariate stochastic processes. Similar results on the rate of convergence have been published in [43] and [44].

In practice, the AR parameters are often derived using estimates of the covariance sequences and not the original covariance sequence of the process itself, as the latter is not readily available[1]. While studying the asymptotic behavior of such estimates, one has to

[1]We use the term "theoretical AR estimates" to refer to AR approximations based on solving the Yule-Walker equations using the true covariances and the term "empirical AR estimates" to refer to AR

impose conditions on the relationship between the number of samples N used for estimation and the order of estimation p. It was shown by Berk [39] that the AR parameters derived from an estimated covariance sequence converges in probability as long as $p = o\{N^{\frac{1}{3}}\}$ [2]. Mean square convergence of the AR parameters as the number of samples approaches infinity while the model order remains finite was studied in [45]. In [46], the authors proved theorems on the rate of almost sure convergence of the covariance estimates to their true values, and subsequently derived the same for the AR parameters based on AR models driven by martingale difference innovation sequences. Their theorems require $p = O\{(\ln N)^{\alpha}\}$ for some $\alpha < \infty$. Among more recent works, [47, 48] and [49] have studied the convergence of the estimated covariance matrix and AR parameters to their corresponding theoretical values under similar assumptions. This assumption, however, is too restrictive in the context of signal processing applications, where in many cases only wide sense stationarity of the process can be guaranteed.

2.3 Determining causality within a family of WSS time series

Complex systems consisting of several interacting units are encountered extensively in various fields of study [50, 51], ranging from econometrics and finance to neuroscience, climatology and ecology. Detailed discussions on the key issues and challenges in this area have been presented in [52, 53] and [54], along with comprehensive surveys of the available literature. Analyzing the interplay between the dynamic behaviour of the various units involved and their interdependence relations has recently become a key issue in multidisciplinary research [55, 56, 57]. Given a family of dynamic systems, the objective is to analyze the interdependence relations among the individual units, and identify whether one is influenced by the others. Such systems are often modeled as a group of stochastic processes, and the interdependence relations among processes are represented in the form of a graph, where nodes correspond to the processes involved and edges (directed or undirected) indicate dependence relations [53]. The problem arises in a wide variety of

approximations based on empirical estimates of the covariance sequence.

[2] Standard notation:

$$k = o\{Z\} \quad \Rightarrow \quad \lim_{Z\to\infty} \frac{|k(Z)|}{|Z|} = 0$$

$$k = O\{Z\} \quad \Rightarrow \quad \lim_{Z\to\infty} \sup \frac{|k(Z)|}{|Z|} < \infty$$

applications including economics, biology, cognitive sciences and ecology [58, 59]; and as pointed out in [57] and [60], the issue of inferring information on the network topology from a set of measurements is one that still lacks proper understanding.

One of the earliest works in the area was presented in [61], where the authors discussed means to derive phylogenic connections among various organisms. For each of a large number of species, a number of physiological characteristics were identified and quantified in terms of numbers. The "closeness" of two species was estimated by computing the product-moment correlation coefficient for the two; which was used in turn to reconstruct a hierarchial structure. Similar work was done in [62], [63] and [64], among others.

Another well-known approach [59] is found in the field of finance, where one is interested in deriving a network-like structure for a given set of stock prices that evolve over time [65]. A metric is defined on the basis of correlation coefficients (the idea of such a metric was introduced in [66]) and a hierarchical topology is obtained using the minimum spanning tree approach.

The issue of determining interdependence among several time series was also studied by [67], [68] and [69]. These papers suggested modified methods to estimate the correlation matrix, which can then be used to derive network topology following the ideas developed in [65].

Granger-causality [8, 9, 70, 71] has often been used to identify how one time series "influences" another. Essentially, a process $\{X(n)\}$ is said to "Granger-cause" another process $\{Y(n)\}$ if given the past history of $\{Y(n)\}$, the additional knowledge of the past history of $\{X(n)\}$ leads to a better prediction of the present value of $\{Y(n)\}$ (i.e., a reduction in the mean squared error). The theory was further developed by Geweke [10, 11] through an analysis of Granger-causality in the spectral domain. Related is the notion of causal Wiener filters [17]; where one attempts to fit a linear predictor for $Y(n)$ using the past values of $\{X(n)\}$.

A useful summary of some of the interesting works done in this field was presented in [59]. A more recent survey on the research on Granger-causality across various fields of applications, including graphical representations of causality can be found in [72].

Among the recent works that have addressed this problem, in [57], the authors proposed a method in which driving signals were externally applied to the system and the dynamic response was measured. The network structure was inferred by comparing this response with the "undriven" dynamics of the system. A network of coupled phase oscillators was considered and the required information was obtained by performing measurements under different (and independent) driving conditions, under the assumption that all units communicate with each other. In [73], a means of identifying modular structure within a

given network was suggested by using the properties of phase oscillators. The notion of Granger-causality was used in [74] to study the effect of one part of the nervous system upon another. In [60], the authors considered a sparsely connected network and suggested a methodology to reconstruct the topology in a noisy environment and under low availability of data; using an L_1 minimization technique.

Information theoretic tools present a useful framework to investigate such influences. While mutual information [75] is a symmetric quantity that only measures the amount of information *shared* between random vectors without any insight on the direction of information flow, it can be modified to develop asymmetric measures; including directed information; a quantity that was introduced in [76] and subsequently formalized in [77] and transfer entropy [78, 79]. In recent years, directed information has been used in the context of identifying hidden causal links among time series [80, 81, 82].

It was pointed out in [83] that when two processes are jointly Gauss-Markov, there is an equivalence between Granger-causality and directed information. Equivalence relations between the two were derived in [53] for Gaussian linear models and in [54] under fairly general frameworks. In [84], it was shown that transfer entropy and Granger-causality are entirely equivalent, for Gaussian variables. Similar results were presented in [82] in the context of neural spike trains, where the authors developed an estimator of directed information that infers causality.

When a system consists of more than two processes, ideally, observations from each process should be simultaneously taken into account to detect the interdependence relations. For large systems, however, this is computationally demanding. Moreover, in most practical applications, *approximations* of moments and covariances are used in lieu of their true values, which may lead to the detection of false edges between nodes. Alternatively, instead of considering all processes simultaneously, one can look at the interdependence relations between each pair of processes. Even though estimating pairwise causal relations of processes may fail to reveal all the interconnections of a network; it can provide valuable information on the structure at much less computational costs.

In [85], the authors considered the problem of approximating the joint distribution of multiple random processes where each node has at most one parent in the directed information graph. Using the Kullback-Leibler (KL) divergence as the metric to find the best approximation, it was shown that the optimal approximate joint distribution can be obtained by maximizing the sum of pairwise directed information. In [86], the authors used pairwise, infinite impulse response (IIR) Wiener filters to define a coherence-based metric and then reconstruct the network topology as the minimum spanning tree connecting the nodes. This method was shown to outperform the correlation-based methods, especially when the number of available samples was low. In [58], sufficient conditions were derived

to guarantee the exact reconstruction of the link structure of a network while considering the coherence-based metric defined earlier. Similar methods based on the Wiener filter were suggested in [87] and [59] along with several interesting analytical results.

An important consideration in MMSE estimation is often that of parsimonious modeling, a priority on keeping the model simple. When the interaction of several inter-dependent agents are being described, it is particularly useful if the *most significant* dependences are identified, as opposed to having a model which detects *all* dependences. Not only is a model involving a smaller number of non-zero parameters easier to interpret, but it is also more accurate as it eliminates some of the spurious connections detected because of inadequate data. In the context of a graphical representation of the causal connections of a family of time series, it would be useful to restrict the number of edges going into a node, so that only edges corresponding to the strongest influences are retained.

The above issue is often addressed through tools like the lasso ("least absolute shrinkage and selection operator") [88], where the optimal MMSE estimator is derived under a constraint that enforces a bound on the absolute sum (the ℓ_1 norm) of the parameters. While this method does not directly restrict the *number* of non-zero parameters, it can successfully reduce the less significant parameters while retaining the more influential ones, and is relatively easy to implement. In [89], the authors provided a necessary condition for the lasso variable selection to be consistent, and subsequently presented a modified version where different parameters in the ℓ_1 penalty where assigned adaptive weights.

An extension of the lasso method is the group lasso [90, 91], where estimation accuracy is improved by dividing the prediction parameters in groups, and optimizing the estimator through the implementation of a constraint on these groups. Conditions under which the group lasso method consistently estimates sparse structures of causal connections were derived in [92].

A lasso-based technique to measure causal relationships from time-course gene expression data was presented in [93]. In [94], the authors presented a comparative discussion on various graphical Granger methods, including the lasso. A method similar to the group lasso technique that seeks to identify the group structure among various lag variables involved in a temporal model was discussed in [95]. Related is the method proposed in [96] which derived sparsity connections under a Gaussian framework, or that of [87], where a sub-optimal Wiener filter based algorithm was used to restrict the number of edges.

Achieving parsimonious models through the detection of the most significant groups of parameters remains an area of considerable interest at present. In the context of the problem of detecting Granger-causality among a number of time series, this is an area of particular importance.

2.4 Conclusion

In this chapter we have presented a short review of the available literature on finite order approximations of time series and on the topic of inferring causal relations from the dynamic response of a complex system. It is evident from our survey, that while the AR estimation problem has been studied in some detail, and several issues, including that of an optimal AR model order has been addressed quite thoroughly; few results are available on the convergence of the spectral density of the approximating finite order AR process.

The main motivation behind the first part of this dissertation is to study the asymptotic behavior of the spectral density of finite order approximation models, as the order approaches infinity, and to obtain conditions under which the spectral density of the approximation converges to the spectral density of the infinite order AR estimate of the original process. We thereby identify a class of stochastic processes for which the spectral density of the original process may be derived from that of an approximating AR sequence. We look at convergence of the spectral density at the origin, and with respect to an L_2 norm defined over the frequency domain. We also consider the case where the AR parameters are computed empirically, using estimated covariances, and present conditions for convergence of the spectral density of the approximating AR sequence. Some of our preliminary results were presented in [97] and [98]. The main results of this part were published in [99].

Identifying interdependence and causal connections within a given collection of time series, on the other hand, is a topic that has begun to attract significant interest in recent years. While several techniques have been proposed for the purpose, the problem is still not completely understood and many issues remain yet to be addressed. The objective of the second part of this dissertation is to find ways to understand the interplay among a number of time series and devise methods to detect causal dependence. We investigate the efficiency of a simple pairwise estimation technique, i.e., the causal Wiener filter, in detecting Granger-causality in a large system of time series and present some analytical results. Furthermore, we also propose a sub-optimal method using pairwise FIR Wiener filters to detect causal connections among families of WSS time series and compare its performance to that of directed information through simulation. Noting that a large class of time series are cyclostationary, we propose a simple, time-invariant AR estimation technique for such processes and use a similar idea to detect Granger-causality. Finally, we propose a technique that restricts the number of edges in the graphical representation of a system of interdependent time series through the implementation of a novel penalty function. Some of these results were presented at [100].

Part I

Chapter 3

Convergence of Spectral Density of Finite Order Estimates of Stationary Time Series

3.1 Introduction

Linear estimation of a time series from a finite number of past observations is a problem encountered in a diverse variety of applications including econometrics, statistical signal processing and neuroscience. This chapter deals with the spectral properties of finite order linear approximations for a *regular* zero-mean, real-valued wide sense stationary (WSS) sequence with finite second moment. From the Wold decomposition [12, 101], it follows that such a process can be expressed as an infinite weighted sum of unit-variance, uncorrelated random variables called the *innovation process*. This is a moving average (MA) type model for the process. By a one-step application of the projection theorem, on the other hand, one obtains an infinite order autoregressive (AR) model of the process, wherein the process is expressed as a weighted sum of all its past values, plus the current value of the innovation process. The main motivation of this research is to explore how such models behave asymptotically, as the model order approaches infinity, and whether the approximated versions approach the original process in the limit both in the time and frequency domain. Specifically, we are interested in studying the asymptotic behaviour of the spectral density of autoregressive (AR) and moving average (MA) estimates.

We look at convergence of the spectral density at the origin, and with respect to an L_2 norm defined over the frequency domain. The motivation for considering convergence in L_2 comes from an Electrical Engineering perspective, noting that the spectral density of a discrete time stochastic process represents the distribution of power over the frequency domain, and that its square-integral over the domain of frequency is an indicator of the total energy of the process.

Additionally, we also look at the convergence of the spectral density at the origin. Recall that the time average variance constant (TAVC, denoted by Γ^2) of a WSS ergodic process $\{X(n)\}$ is defined as the spectral density of the process at the origin. If $\bar{X}_T = \frac{1}{T}\sum_{n=1}^{T} X(n)$ is the sample mean, then according to the central limit theorem due to Ibragimov and Linnik [102],

$$\sqrt{T}(\bar{X}_T - \mu) \Longrightarrow \mathcal{N}(0, \Gamma^2)$$

where $\Longrightarrow$ denotes convergence in distribution.

In steady state simulations, where the objective is to find the limit of $\bar{X}_T$ as $T \to \infty$ [3], Γ^2 plays an important role. The quantity can be estimated through approximations for the spectral density at the origin, instead of directly estimating the moments.

We begin this chapter with a study of the asymptotic behavior of the spectral density of MA estimates of a stationary time series. We consider a truncated version of Wold's equation, consisting of the first p terms of the sum. It is shown that for any p, this truncated version gives the MMSE moving average estimation for the given process. Next it is shown that this estimate converges in quadratic mean to the original process and when the coefficients of the Wold expansion of the original process are absolutely summable (i.e., the sequence is in ℓ_1 [1]), its spectral density converges to that of the original process in L_2.

Next, we study the asymptotic properties of the spectral density of AR estimates. In this chapter, we consider the case where AR parameters are derived based on knowing the true covariances of the original process. It is shown that when the spectral density of the process is strictly non-vanishing in $\left(-\frac{1}{2}, \frac{1}{2}\right]$ and its covariance sequence is absolutely summable, the spectral density of the approximating AR sequence converges in L_2, and also at the origin, as the order of approximation goes to infinity. In the context of simulating Markov processes, this is a fairly general condition, as a non-vanishing spectral density is guaranteed when a Markov process has a continuous spectral density [103].

[1] Let $\mathbb{Z}$ denote the set of integers.
Then, ℓ_1 denotes the space of all real sequence $\{\xi_j\}$ such that $\sum_{j\in\mathbb{Z}} |\xi_j| < \infty$
ℓ_2 denotes the space of all real sequence $\{\psi_j\}$ such that $\sum_{j\in\mathbb{Z}} |\psi_j|^2 < \infty$

3.2 Preliminaries

Consider a zero-mean, regular, discrete time, real-valued WSS stochastic process $\{X(n)\}_{n\in\mathbb{Z}}$ defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Let $L_2(\mathbb{P})$ denote the Hilbert space of random variables with finite second moment with the inner product $\mathbb{E}[\cdot,\cdot]$ defined thereon. Two zero mean random variables $X, Y \in L_2(\mathbb{P})$ are said to be orthogonal if $\mathbb{E}[XY] = 0$. Let $R(k) = \mathbb{E}[X(n)X(n-k)]$ denote the covariance sequence.

The spectral density $S(\lambda)$ is defined as the discrete time Fourier transform of the covariance sequence.

$$S(\lambda) = \sum_{k\in\mathbb{Z}} R(k)e^{-2\pi\imath\lambda k}, \quad \lambda \in \left(-\frac{1}{2}, \frac{1}{2}\right]$$

For $n \in \mathbb{Z}$ define the space $H(n) =$ linear span of $\{X(n), X(n-1), X(n-2), \ldots\}$.

Let Y be a random variable defined on $L_2(\mathbb{P})$. Define $\overline{\mathbb{E}}[Y|H(n)]$ as the projection of Y onto the space $H(n)$ with respect to the inner-product $\mathbb{E}[\cdot,\cdot]$. Then $\overline{\mathbb{E}}[Y|H(n)]$ is the minimum mean squared error (MMSE) linear estimate of Y given $H(n)$.

Define $\{\nu(n)\}_{n\in\mathbb{Z}}$ as the innovation process associated with $\{X(n)\}_{n\in\mathbb{Z}}$, i.e.,

$$\nu(n) = X(n) - \overline{\mathbb{E}}[X(n)|H(n-1)]$$

with $\mathbb{E}[\nu(n)] = 0$ and $\mathbb{E}[\nu(n)\nu(k)] = \sigma_\nu^2\delta(n-k)$, where $\delta(k)$ denotes the Kronecker delta. Without loss of generality it is assumed that $\sigma_\nu^2 = 1$. By construction, $\nu(n)$ is orthogonal to $H(k-1)$, for all $k \leq n$. The sequence $\{\nu(k)\}_{k\in\mathbb{Z},k\leq n}$ spans the subspace $H(n)$ and constitutes an orthogonal basis in $L_2(\mathbb{P})$ for the latter. Note that $\overline{\mathbb{E}}[X(n)|H(n-1)]$ corresponds to the MMSE linear estimate of $X(n)$ given the space $H(n-1)$ and it can therefore be written as a weighted sum of all the past values of $X(n)$ as follows:

$$\overline{\mathbb{E}}[X(n)|H(n-1)] = \sum_{k=1}^{\infty} b(k)X(n-k) \tag{3.2.1}$$

where the $b(k)$s minimize the mean squared error. The corresponding mean squared error is

$$\begin{aligned}\mathbb{E}\left[\left(X(n) - \sum_{k=1}^{\infty} b(k)X(n-k)\right)^2\right] &= \mathbb{E}\left[\left(X(n) - \overline{\mathbb{E}}[X(n)|H(n-1)]\right)^2\right] \\ &= \mathbb{E}[\nu^2(n)] = 1 \text{ for all } n \end{aligned} \tag{3.2.2}$$

$\overline{\mathbb{E}}[X(n)|H(n-1)]$ is thus an infinite order AR estimate of $X(n)$. On the other hand, being a WSS process, $X(n)$ may be expressed in terms of its Wold decomposition as follows [12, 101].

$$X(n) = \sum_{k=0}^{\infty} a(k)\nu(n-k) \text{ for all } n \in \mathbb{Z} \tag{3.2.3}$$

$$\text{with} \sum_{k=0}^{\infty} |a(k)|^2 < \infty \text{ and } a(0) = 1.$$

This gives an infinite order moving average representation of $X(n)$. The coefficients $\{b(k)\}$ and $\{a(k)\}$ are related to each other as follows. For each k,

$$b(k) = \sum_{j=1}^{k} a(j)b(k-j) \tag{3.2.4}$$

From (3.2.3) and the properties of the innovations sequence it is seen that:

$$R(k) = \sum_{n=0}^{\infty} a(n)a(n-k)$$

Let p be a positive integer, and define the space $H^p(n) =$ linear span of $\{X(n), X(n-1), X(n-2), \ldots X(n-p+1)\}$, i.e., the space of all linear combinations of the p most recent values of the sequence at time n, including $X(n)$; and all their limits in the mean square when they exist. $H^p(n)$ is then a closed Hilbert subspace of $H(n)$, for all p and n. Clearly, $H^{\infty}(n) = H(n)$, by definition.

Define $\overline{X}_p(n)$ as the MMSE AR approximation of $X(n)$ of order p. Then $\overline{X}_p(n)$ is a linear combination of $\{X(n-1), X(n-2), \ldots, X(n-p)\}$ and is given by

$$\overline{X}_p(n) = \overline{\mathbb{E}}[X(n)|H^p(n-1)] = \sum_{k=1}^{p} b_p(k)X(n-k) \tag{3.2.5}$$

where the coefficients $b_p(k)$ minimize the error $\mathbb{E}[(X(n) - \sum_{k=1}^{p} b_p(k)X(n-k))^2]$. The coefficients $b_p(k)$ can be obtained as solutions to the Yule-Walker equations [4, 6] given by:

$$\boldsymbol{B_p} = \boldsymbol{R_p}^{-1}\boldsymbol{r_p} \tag{3.2.6}$$

where

$$\boldsymbol{R_p} = \begin{pmatrix} R(0) & R(1) & \cdots & R(p-1) \\ R(1) & R(0) & \cdots & R(p-2) \\ \vdots & \vdots & \ddots & \vdots \\ R(p-1) & R(p-2) & \cdots & R(0) \end{pmatrix}$$

$$\boldsymbol{r_p} = [R(1)\ R(2)\ \ldots\ R(p)]^T$$

and

$$\boldsymbol{B_p} = [b_p(1)\ \ldots\ b_p(p)]^T$$

Let $\{\overline{R}_p(k)\}_{k\in\mathbb{Z}}$ and $\{\overline{R}(k)\}_{k\in\mathbb{Z}}$ be the covariance sequences of $\overline{X}_p(n)$ and $\overline{\mathbb{E}}[X(n)|H(n-1)]$ respectively. Let $S_{\overline{X}_p}(\lambda)$ and $S_{\overline{X}}(\lambda)$ denote the respective spectral densities.

$$S_{\overline{X}_p}(\lambda) = \sum_{k\in\mathbb{Z}} \overline{R}_p(k) e^{-2\pi\imath\lambda k}$$

$$S_{\overline{X}}(\lambda) = \sum_{k\in\mathbb{Z}} \overline{R}_k e^{-2\pi\imath\lambda k}$$

Finally, let $\widetilde{X}_p(n)$ denote the best p-th order moving average estimate of $\{X(n)\}$ and let $S_{\widetilde{X}_p}(\lambda)$ denote its spectral density.

A sequence of random variables $\{X(n,\omega)\}$ is said to converge to a random variable $\{X(\omega)\}$ in quadratic mean if the convergence is in $L_2(\mathbb{P})$. A sequence of functions $f_n : \left(-\frac{1}{2}, \frac{1}{2}\right] \to \mathbb{R}$ is said to converge in L_2 when the convergence is in $L_2\left(\left(-\frac{1}{2}, \frac{1}{2}\right], \mathbb{R}\right)$.

3.3 Moving average approximations of regular stationary sequences

Consider a moving average approximation of $X(n)$ of order p, constructed using the innovation sequence $\{\nu(n)\}$. Note by assumption, $\mathrm{Var}[\nu(n)] = \sigma_\nu^2 = 1$ for all $n \in \mathbb{N}$. We begin with some results on the moving average estimate.

Proposition 3.3.1. *The best p-th order moving average approximation of $X(n)$ is given by $\widetilde{X}_p(n) = \sum_{k=0}^{p} a(k)\nu(n-k)$.*

Proof. Let $\widetilde{X}_p(n) = \sum_{k=0}^{p} \hat{a}(k)\nu(n-k)$ be the p-th order moving average estimator of $X(n)$ and $\xi(n)$ be the corresponding mean squared estimation error. Then

$$\begin{aligned}\xi(n) &= \mathbb{E}\left[\left(\sum_{k=0}^{p}\hat{a}(k)\nu(n-k) - \sum_{k=0}^{\infty}a(k)\nu(n-k)\right)^2\right] \\ &= \mathbb{E}\left[\left(\sum_{k=0}^{p}(\hat{a}(k)-a(k))\nu(n-k) - \sum_{k=p+1}^{\infty}a(k)\nu(n-k)\right)^2\right]\end{aligned}$$

The MMSE estimate is obtained by setting

$$\frac{\partial\xi(n)}{\partial\hat{a}(k)} = 0 \text{ for } k = 0, 1, \ldots, p \tag{3.3.7}$$

$$\text{This gives } \hat{a}(k) = a(k) \quad k = 0, 1, \ldots, p \tag{3.3.8}$$

Thus, the best p-th order moving average approximation of $X(n)$ is given by a truncation of the infinite sum up to the first p terms. □

The next result is related to the mean square convergence of the p-th order approximation that readily follows from the properties of the innovation process and the Wold decomposition.

Proposition 3.3.2. *As $p \to \infty$, $\widetilde{X}_p(n)$ converges to $X(n)$ in quadratic mean.*

Proof. The proof is trivial. We start by noting that as the process $\{X(n)\}$ is square-integrable, we have $\sum_{k=0}^{\infty} a^2(k) < \infty$; since $\mathbb{E}[X^2(n)] = \sum_{k=0}^{\infty} a^2(k)$. Then,

$$\begin{aligned}\lim_{p\to\infty}\mathbb{E}[(\widetilde{X}_p(n) - X(n))^2] &= \lim_{p\to\infty}\mathbb{E}\left[\left(\sum_{k=p+1}^{\infty}a(k)\nu(n-k)\right)^2\right] \\ &= \lim_{p\to\infty}\left(\sum_{k=p+1}^{\infty}a^2(k)\right) \\ &= 0 \end{aligned} \tag{3.3.9}$$

where we use the orthogonality of the $\nu(n)$s and the square-summability of the $a(k)$s. □

3.3.1 Convergence of the spectral density in L_2

For all functions $F : \left(-\frac{1}{2}, \frac{1}{2}\right] \to \mathbb{C}$ such that $\int_{-\frac{1}{2}}^{\frac{1}{2}} |F(\lambda)|^2 d\lambda < \infty$, define $||\cdot||$ to be the L_2 norm as follows:

$$||F|| = \left| \int_{-\frac{1}{2}}^{\frac{1}{2}} |F(\lambda)|^2 d\lambda \right|^{\frac{1}{2}}$$

Proposition 3.3.3. *Under the condition $\sum_{k=0}^{\infty} |a(k)| < \infty$ where $\{a(k)\}$ denotes the sequence of coefficients in the Wold decomposition, the spectral density of $\widetilde{X}_p(n)$ converges in L_2 to that of $X(n)$ as $p \to \infty$.*

Proof. Let $\sum_{k=0}^{\infty} |a(k)| = S$. It follows from the properties of the innovation process, that the covariance sequence of $\{X(n)\}$ is $\{R(k)\}$ where

$$R(k) = \sum_{l=0}^{\infty} a(l)a(l-k)$$

The spectral density $S(\lambda)$ is given by

$$\begin{aligned} S(\lambda) &= \sum_{k\in\mathbb{Z}} \sum_{l=0}^{\infty} a(l)a(l-k)e^{-2\pi\imath k\lambda} \\ &= \sum_{l=0}^{\infty} a(l)e^{-2\pi\imath l\lambda} \sum_{l-k=-\infty}^{\infty} a(l-k)e^{2\pi\imath(l-k)\lambda} \\ &= A_0^{\infty}(\lambda)A_0^{\infty}(-\lambda) \end{aligned}$$

where for $M, N \in \mathbb{N}$

$$A_M^N(\lambda) = \sum_{l=M}^{N} a(l)e^{-2\pi\imath l\lambda}$$

and

$$A_M^{\infty}(\lambda) = \sum_{l=M}^{\infty} a(l)e^{-2\pi\imath l\lambda}$$

Similarly, the spectral density of $\widetilde{X}_p(n)$ is given by

$$S_{\widetilde{X}_p}(\lambda) = A_0^p(\lambda)A_0^p(-\lambda)$$

Now consider

$$\begin{aligned}
\left\|S(\lambda) - S_{\widetilde{X}_p}(\lambda)\right\| &= \|A_0^\infty(\lambda)A_0^\infty(-\lambda) - A_0^p(\lambda)A_0^p(-\lambda)\| \\
&= ||(A_0^p(\lambda) + A_{p+1}^\infty(\lambda))(A_0^p(-\lambda) + A_{p+1}^\infty(-\lambda)) - A_0^p(\lambda)A_0^p(-\lambda)|| \\
&= \left\|(A_0^p(\lambda)A_{p+1}^\infty(-\lambda) + A_{p+1}^\infty(\lambda)A_0^\infty(-\lambda)\right\| \\
&\leq \left\|(A_0^p(\lambda)A_{p+1}^\infty(-\lambda)\right\| + \left\|A_{p+1}^\infty(\lambda)A_0^\infty(-\lambda)\right\| \\
&= \left|\int_{-\frac{1}{2}}^{\frac{1}{2}} I_1(p,\lambda)d\lambda\right|^{\frac{1}{2}} + \left|\int_{-\frac{1}{2}}^{\frac{1}{2}} I_2(p,\lambda)d\lambda\right|^{\frac{1}{2}} \qquad (3.3.10)
\end{aligned}$$

where

$$I_1(p,\lambda) = |A_0^p(\lambda)A_{p+1}^\infty(-\lambda)|^2$$
$$I_2(p,\lambda) = |A_{p+1}^\infty(\lambda)A_0^\infty(-\lambda)|^2$$

$$\begin{aligned}
\text{Then, } I_1(p,\lambda) &= |A_0^p(\lambda)|^2|A_{p+1}^\infty(-\lambda)|^2 \leq \left(\sum_{k=0}^{p}|a(k)|\right)^2\left(\sum_{k=p+1}^{\infty}|a(k)|\right)^2 \\
&\leq S^2\left(\sum_{k=p+1}^{\infty}|a(k)|\right)^2
\end{aligned}$$

and similarly,

$$\begin{aligned}
I_2(p,\lambda) &= |A_{p+1}^\infty(\lambda)|^2|A_0^\infty(-\lambda)|^2 \\
&\leq \left(\sum_{k=p+1}^{\infty}|a(k)|\right)^2\left(\sum_{k=0}^{\infty}|a(k)|\right)^2 \\
&= S^2\left(\sum_{k=p+1}^{\infty}|a(k)|\right)^2
\end{aligned}$$

Since the sum $\sum_{k=0}^{\infty}|a(k)|$ is non-decreasing and converges to $S < \infty$, for any given $\epsilon > 0$, there exists a positive integer p such that $\sum_{k=p+1}^{\infty}|a(k)| < \frac{\epsilon}{2S}$, so that

$$I_1(p,\lambda) < \frac{\epsilon^2}{4} \qquad (3.3.11)$$

$$I_2(p,\lambda) < \frac{\epsilon^2}{4} \qquad (3.3.12)$$

Combining (3.3.11) and (3.3.12) with (3.3.10) yields

$$\left\|S(\lambda) - S_{\widetilde{X}_p}(\lambda)\right\| < \left(\left|\int_{-\frac{1}{2}}^{\frac{1}{2}} \frac{\epsilon^2}{4} d\lambda\right|^{\frac{1}{2}} + \left|\int_{-\frac{1}{2}}^{\frac{1}{2}} \frac{\epsilon^2}{4} d\lambda\right|^{\frac{1}{2}}\right) = \epsilon$$

Thus, for any $\epsilon > 0$, there exists a positive integer p such that $||S(\lambda) - S_{\widetilde{X}_p}(\lambda)|| < \epsilon$ for all λ. Therefore, $S_{\widetilde{X}_p}(\lambda)$ converges in L_2 to $S(\lambda)$ as $p \to \infty$. □

3.4 AR approximations of regular stationary sequences based on true covariances

We begin by looking at the asymptotic behavior of the AR approximation given by (3.2.5). Define $\nu_p(n)$ as the error in estimation corresponding to the AR-p approximation of $\{X(n)\}$, i.e.,

$$\nu_p(n) = X(n) - \mathbb{E}[X(n)|H^p(n-1)]$$

Lemma 3.4.1. *As $p \to \infty$, $\overline{\mathbb{E}}[X(n)|H^p(n-1)] \to \overline{\mathbb{E}}[X(n)|H(n-1)]$ and $\nu_p(n) \to \nu(n)$ in quadratic mean.*

Note that for all $p \in \mathbb{N}$

$$\begin{aligned}
\overline{\mathbb{E}}[\nu_p(n)|H(n-1)] &= \overline{\mathbb{E}}[(X(n) - \overline{\mathbb{E}}[X(n)|H^p(n-1)])|H(n-1)] \\
&= \overline{\mathbb{E}}[X(n)|H(n-1)] - \overline{\mathbb{E}}[X(n)|H^p(n-1)] \\
&= (X(n) - \overline{\mathbb{E}}[X(n)|H^p(n-1)]) - (X(n) - \overline{\mathbb{E}}[X(n)|H(n-1)]) \\
&= \nu_p(n) - \nu(n) \qquad (3.4.13)
\end{aligned}$$

Let $\overline{\mathbb{E}}[\nu_p(n)|H(n-1)] = \epsilon_p(n)$. Then $\epsilon_p(n) = \nu_p(n) - \nu(n)$ and $\mathbb{E}[\epsilon_p(n)] = \mathbb{E}[\nu_p(n) - \nu(n)] = 0$. As $X(n)$ is a WSS sequence in n and $\nu_p(n)$ is constructed as a linear combination of $X(n)$, $X(n-1)$, ..., $X(n-p)$ whose coefficients do not depend on n, it follows that $\nu_p(n)$ is also a WSS sequence in n. The variance of $\nu_p(n)$, then, is only a function of p. Let $\text{Var}[\nu_p(n)] = \sigma_p^2$. By (3.4.13), and the fact that $\nu(n)$ is orthogonal to the subspace $H(n-1)$ (and hence to $\epsilon_p(n)$) we obtain:

$$\begin{aligned}
\sigma_p^2 &= \text{Var}[\nu(n)] + \text{Var}[\epsilon_p(n)] \\
&= 1 + \text{Var}[\epsilon_p(n)] \\
&= 1 + \mathbb{E}[\epsilon_p(n)^2] \\
&\geq 1 \text{ for all } p \in \mathbb{N}
\end{aligned}$$

Note that for any q, $p \in \mathbb{N}$ such that $q > p$, $H^p(n-1) \subset H^q(n-1)$. It follows then, that $\overline{\mathbb{E}}[X(n)|H^q(n-1)]$ is at least as good a linear estimate of $X(n)$ as $\overline{\mathbb{E}}[X(n)|H^p(n-1)]$, in terms of mean squared error. Therefore,

$$\mathbb{E}[(X(n) - \overline{\mathbb{E}}[X(n)|H^q(n-1)])^2] \leq \mathbb{E}[(X(n) - \overline{\mathbb{E}}[X(n)|H^p(n-1)])^2]$$

and hence

$$\sigma_q^2 \leq \sigma_p^2$$

Hence, the sequence σ_p^2 is a non-increasing sequence in p, bounded below by 1 and must therefore have a limit as $p \to \infty$ that is bounded from below by 1. It can be shown that the limit is in fact equal to 1. A rigorous proof is available in [12, lemma 3.1(b)].

We now present a few preliminary lemmas that would be used to derive the convergence results on the spectral density of the AR approximation. Throughout the rest of this chapter, it is assumed that the spectral density $S(\lambda)$ is non-vanishing in $\left(-\frac{1}{2}, \frac{1}{2}\right]$.

We start with the following lemma on the pointwise convergence of the coefficients $b_p(k)$ [37, Proposition 3.1].

Lemma 3.4.2. *As $p \to \infty$, $b_p(k) \to b(k)$ for each $k \in \mathbb{N}$.*

Proof. For each $p \in \mathbb{N}$ and $k \in \{1, \ldots, p\}$ define $\tilde{b}_p(0) = 1$ and $\tilde{b}_p(k) = -b_p(k)$. For any p, $\nu_p(n)$ is given by

$$\nu_p(n) = \sum_{k=0}^{p} \tilde{b}_p(k) X(n-k)$$

The above can be written in the matrix form as $\nu = B\mathbf{X}$ where $\nu = [\nu_0(n-p) \cdots \nu_p(n)]^T$, $\mathbf{X} = [X(n-p) \cdots X(n)]^T$ and B is a lower triangular matrix whose first column is $[\tilde{b}_0(0) \cdots \tilde{b}_p(p)]^T$. The matrix B is invertible with inverse A which satisfies $\mathbf{X} = A\nu$. The inverse A is lower triangular with first column $[a_0(0) \cdots a_p(p)]^T$ and the elements $a_p(k)$ satisfy for all $n \in \mathbb{N}$

$$X(n) = \sum_{k=0}^{p} a_p(k) \nu_{p-k}(n-k) \tag{3.4.14}$$

and for $p > k$

$$\tilde{b}_p(k) = -\sum_{j=1}^{k} a_p(j) \tilde{b}_{p-j}(k-j) \tag{3.4.15}$$

By definition, $\nu_n(n)$ and $\nu_m(m)$ are orthogonal to each other for $n \neq m$ and hence (3.4.14) provides an orthogonal decomposition of $X(n)$. Therefore,

$$\mathbb{E}[X(n)\nu_{p-k}(n-k)] = a_p(k)\sigma_{p-k}^2 \tag{3.4.16}$$

From (3.2.3), on the other hand, we have

$$\mathbb{E}[X(n)\nu(n-k)] = a(k) \tag{3.4.17}$$

However, by lemma 3.4.1, $\nu_p(n) \to \nu(n)$ in quadratic mean. It then follows from (3.4.16) and (3.4.17) that for all $k \in \mathbb{N}$

$$\begin{aligned}\lim_{p\to\infty} |a_p(k)\sigma_{p-k}^2 - a(k)| &= \lim_{p\to\infty} |\mathbb{E}[X(n)\nu_{p-k}(n-k) - X(n)\nu(n-k)]| \\ &\leq \lim_{p\to\infty} \left|\mathbb{E}[X^2(n)]\mathbb{E}[(\nu_{p-k}(n-k) - \nu(n-k))^2]\right|^{\frac{1}{2}} \\ &= 0\end{aligned}$$

$$\text{and hence } \lim_{p\to\infty} a_p(k) = a(k) \tag{3.4.18}$$

Finally, to show the pointwise convergence of $\tilde{b}_p(k)$ (and therefore that of $b_p(k)$) first observe that $\tilde{b}_p(0) = b(0) = 1$ holds for all p. Let

$$\lim_{p\to\infty} \tilde{b}_p(j) = -b(j)$$

for all $j \leq k$. Then, using the recursive relation given by (3.4.15) and comparing with (3.2.4), one obtains

$$\lim_{p\to\infty} \tilde{b}_p(k+1) = -b(k+1)$$

Therefore, by the principle of mathematical induction, as $p \to \infty$, $\tilde{b}_p(k) \to -b(k)$; i.e., $b_p(k) \to b(k)$ for each $k \in \mathbb{N}$. □

Next, we present a key result on the summability of the AR coefficients known as Baxter's inequality [38, Theorem 2.2].

Let $\{X(n)\}$ be a WSS process with spectral density function $S(\lambda) > 0$ and let $\overline{X}_p(n)$ be the p-th order MMSE linear predictor of $X(n)$, defined by (3.2.5) and let σ_p^2 be the corresponding mean squared error. Let $\{b(k)\}$ be the limits of the coefficients $\{b_p(k)\}$ and let $\sigma^2 > 0$ be the limit of σ_p^2 as $p \to \infty$ (in our case $\sigma^2 = 1$ by lemma 3.4.1). Define the sequence $\{u_p(k)\}$ as $u_p(k) = -\frac{b_p(k)}{\sigma_p^2}$ and let $\{U(k)\}$ be the limit of $\{u_p(k)\}$. Then, the theorem is stated as follows.

Theorem 3.4.1. ***Baxter's Inequality****: Let $S(\lambda)$ is a positive continuous function whose Fourier coefficients have γ moments for some $\gamma \geq 0$, i.e.,*

$$\sum_{m=0}^{\infty} m^{\gamma}|c_m| < 0$$

where $\{c_m\}$ *are the Fourier coefficients. Then, there exists an integer* $N > 0$ *and a constant* $c > 0$*, both depending only on* $S(\lambda)$ *such that for all* $p \geq N$*,*

$$\sum_{k=1}^{p}(2^{\gamma} + k^{\gamma})|u_p(k) - U(k)| \leq c \sum_{k=p+1}^{\infty}(2^{\gamma} + k^{\gamma})|U(k)|$$

Note that the Fourier coefficients of the spectral density are the elements of the covariance sequence $\{R(k)\}$.

The above theorem can be used to establish the following lemma on the convergence of the coefficients $\{b_p(k)\}$ as $p \to \infty$.

Lemma 3.4.3. *When the spectral density of* $\{X(n)\}$ *is strictly positive in* $\lambda \in \left(-\frac{1}{2}, \frac{1}{2}\right]$*, and the covariance sequence is in* ℓ_1*, i.e.,*

$$\sum_{k\in\mathbb{Z}} |R(k)| < \infty$$

then

$$\lim_{p\to\infty} \sum_{k=1}^{p} |b_p(k) - b(k)| = 0$$

Proof. The proof is a simple application of Baxter's inequality. Note that when the covariance sequence is in ℓ_1, the spectral density is continuous. Pointwise convergence of the $b_p(k)$ to $b(k)$ for each k follows from lemma 3.4.2. Moreover, summability of the covariance sequence also implies that the sequence has a finite 0-th moment ($\gamma = 0$ in Theorem 3.4.1). It then follows from Theorem 3.4.1 that under the assumption that the spectral density of $\{X(n)\}$ is strictly positive in $\lambda \in \left(-\frac{1}{2}, \frac{1}{2}\right]$, there exists a positive integer N and a constant $c > 0$ such that

$$\sum_{k=1}^{p} \left| \frac{b_p(k)}{\sigma_p^2} - b(k) \right| \leq c \sum_{k=p+1}^{\infty} |b(k)| \tag{3.4.19}$$

for all $p > N$. Since the covariance sequence has been assumed to be in ℓ_1, the sequence of AR coefficients of the original process are also in ℓ_1 [26], i.e.,

$$\sum_{k=1}^{\infty} |b(k)| < \infty$$

Then, for any $\epsilon > 0$, there exists an integer N_0 such that

$$\sum_{k=p+1}^{\infty} |b(k)| < \frac{\epsilon}{c}$$

for all $p > N_0$. Define $N^* = \max\{N, N_0\}$. Then for all $p > N^*$,

$$\sum_{k=1}^{p} \left| \frac{b_p(k)}{\sigma_p^2} - b(k) \right| < \epsilon$$

Therefore,

$$\lim_{p\to\infty} \sum_{k=1}^{p} \left| \frac{b_p(k)}{\sigma_p^2} - b(k) \right| = 0 \tag{3.4.20}$$

It follows from the triangle inequality that

$$\begin{aligned} \sum_{k=1}^{p} \left| \frac{b_p(k) - b(k)}{\sigma_p^2} \right| &\leq \sum_{k=1}^{p} \left| \frac{b_p(k)}{\sigma_p^2} - b(k) \right| + \sum_{k=1}^{p} \left| \frac{b(k)}{\sigma_p^2} - b(k) \right| \\ &\leq \sum_{k=1}^{p} \left| \frac{b_p(k)}{\sigma_p^2} - b(k) \right| + \frac{\sigma_p^2 - 1}{\sigma_p^2} \sum_{k=1}^{p} |b(k)| \end{aligned}$$

Therefore, as $p \to \infty$,

$$\lim_{p\to\infty} \sum_{k=1}^{p} \left| \frac{b_p(k) - b(k)}{\sigma_p^2} \right| \leq \lim_{p\to\infty} \sum_{k=1}^{p} \left| \frac{b_p(k)}{\sigma_p^2} - b(k) \right| + \lim_{p\to\infty} \frac{\sigma_p^2 - 1}{\sigma_p^2} \sum_{k=1}^{p} |b(k)|$$

By (3.4.20), the first limit on the right hand side is zero and by lemma 3.4.1,

$$\lim_{p\to\infty} \sigma_p^2 = 1$$

Therefore,

$$\lim_{p\to\infty} \sum_{k=1}^{p} |b_p(k) - b(k)| = 0$$

□

The two main results of the following section are applications of the above lemma.

Lemma 3.4.4. *If the spectral density $S(\lambda) > 0$ for $\lambda \in (-\frac{1}{2}, \frac{1}{2}]$ and the covariance sequence is in ℓ_1, ie.,*

$$\sum_{k\in\mathbb{Z}} |R(k)| < \infty$$

then

$$\sum_{k=0}^{\infty} |a(k)| < \infty$$

i.e., the coefficients of Wold decomposition are also in ℓ_1.

The above is a direct consequence of [104, Theorem 3.8.4, P-78].

3.4.1 Convergence of the spectral density in L_2

Here we present a sufficient condition for the L_2 convergence of the spectral density of the AR approximation as $p \to \infty$.

Proposition 3.4.1. *Let $S(\lambda) > 0$ for $\lambda \in \left(-\frac{1}{2}, \frac{1}{2}\right]$ and let $\sum_{k\in\mathbb{Z}} |R(k)| < \infty$. Then as $p \to \infty$, $S_{\overline{X}_p}(\lambda)$ converges to $S_{\overline{X}}(\lambda)$ in L_2.*

Proof. We begin by noting that when the sequence $\{R(k)\}$ is in ℓ_1 and the spectral density is strictly positive (as stated above), both $\{a(k)\}$ (by lemma 3.4.4) and $\{b(k)\}$ are also in ℓ_1 and the conditions of Lemma 3.4.3 are satisfied.

For some p and k, $\overline{R}_p(k)$ may be obtained from (3.2.5) as

$$\begin{aligned}
\overline{R}_p(k) &= \mathbb{E}[\overline{X}_p(n)\overline{X}_p(n-k)] \\
&= \mathbb{E}\left[\sum_{j=1}^{p} b_p(j)X(n-j)\sum_{l=1}^{p} b_p(l)X(n-k-l)\right] \\
&= \sum_{j=1}^{p} b_p^2(j)R(k) + \sum_{t=1}^{p-1}\sum_{j=1}^{p-t} b_p(j)b_p(j+t)(R(k-t)+R(k+t)) \qquad (3.4.21)
\end{aligned}$$

Consider the WSS process given by $\sum_{j=1}^{p} b(j)X(n-j)$ and let $\{R_p(k)\}$ be its covariance sequence. Proceeding as in the case of (3.4.21), we can obtain a similar expression for $R_p(k)$

as follows.

$$R_p(k) = \sum_{j=1}^{p} b^2(j)R(k) + \sum_{t=1}^{p-1}\sum_{j=1}^{p-t} b(j)b(j+t)(R(k-t)+R(k+t))$$

so that for all $k \in \mathbb{Z}$

$$\begin{aligned}
|R_p(k) - \overline{R}_p(k)| &= \Bigg| \sum_{j=1}^{p} (b^2(j) - b_p^2(j))R(k) \\
&\quad + \sum_{t=1}^{p-1}\sum_{j=1}^{p-t} (b(j)b(j+t) - b_p(j)b_p(j+t))\,(R(k-t)+R(k+t)) \Bigg| \\
&\leq \sum_{j=1}^{p} \left|(b^2(j) - b_p^2(j))\right| |R(k)| \\
&\quad + \sum_{t=1}^{p-1}\sum_{j=1}^{p-t} |(b(j)b(j+t) - b_p(j)b_p(j+t))|\,(|R(k-t)| + |R(k+t)|)
\end{aligned}$$

Summing over all $k \in \mathbb{Z}$

$$\begin{aligned}
\sum_{k\in\mathbb{Z}} |R_p(k) - \overline{R}_p(k)| &\leq \Bigg(\sum_{j=1}^{p} \left|b^2(j) - b_p^2(j)\right| \\
&\quad + 2\sum_{t=1}^{p-1}\sum_{j=1}^{p-t} |b(j)b(j+t) - b_p(j)b_p(j+t)| \Bigg) \sum_{k\in\mathbb{Z}} |R(k)| \\
&= \left(\sum_{j=1}^{p}\sum_{i=1}^{p} |b(i)b(j) - b_p(i)b_p(j)| \right) \sum_{k\in\mathbb{Z}} |R(k)| \\
&\leq \left(\sum_{j=1}^{p}\sum_{i=1}^{p} |b(i)||b(j) - b_p(j)| + \sum_{j=1}^{p}\sum_{i=1}^{p} |b_p(j)||b(i) - b_p(i)| \right) \sum_{k\in\mathbb{Z}} |R(k)| \\
&\leq \left(\sum_{i=1}^{p} |b(i)| + \sum_{i=1}^{p} |b_p(i)| \right) \left(\sum_{i=1}^{p} |b(i) - b_p(i)| \right) \sum_{k\in\mathbb{Z}} |R(k)|
\end{aligned}$$

As the covariance sequence has been assumed to be in ℓ_1, the sequence $\{b(k)\}$ is also in ℓ_1 [26] and by lemma (3.4.3), as $p \to \infty$ the second term of the above product goes to 0. Finally, by the same lemma,

$$\lim_{p\to\infty}\sum_{i=1}^{p}|b_p(i)| \le \lim_{p\to\infty}\sum_{i=1}^{p}|b(i)|$$

Therefore,

$$\lim_{p\to\infty}\sum_{k\in\mathbb{Z}}|R_p(k)-\overline{R}_p(k)| = 0 \tag{3.4.22}$$

Now note that

$$\begin{aligned}
R_p(k) &= \mathbb{E}\Big[\Big(X(n)-\nu(n)-\sum_{i=p+1}^{\infty}b(i)X(n-i)\Big)\Big(X(n+k)-\nu(n+k)\\
&\quad -\sum_{j=p+1}^{\infty}b(j)X(n+k-j)\Big)\Big]\\
&= \overline{R}(k)-\sum_{j=p+1}^{\infty}b(j)R(k-j)-\sum_{i=p+1}^{\infty}b(i)R(k+i)\\
&\quad +\sum_{j=p+1}^{\infty}b(j)\mathbb{E}[X(n+k-j)\nu(n)]+\sum_{i=p+1}^{\infty}b(i)\mathbb{E}[X(n-i)\nu(n+k)]\\
&+ \mathbb{E}\left[\left(\sum_{i=p+1}^{\infty}b(i)X(n-i)\right)\left(\sum_{j=p+1}^{\infty}b(j)X(n+k-j)\right)\right]\\
&= \overline{R}(k)-\sum_{j=p+1}^{\infty}b(j)R(k-j)-\sum_{i=p+1}^{\infty}b(i)R(k+i)+\sum_{j=p+1}^{\infty}b(j)a(|k|-j)\\
&\quad +\left(\sum_{i=p+1}^{\infty}b(i)^2\right)R(k)+\sum_{t=1}^{\infty}\sum_{i=p+1}^{\infty}b(i)b(i+t)(R(k+t)+R(k-t))
\end{aligned}$$

Therefore,

$$\begin{aligned}
|R_p(k)-\overline{R}(k)| &\le \sum_{j=p+1}^{\infty}|b(j)||R(k-j)|+\sum_{i=p+1}^{\infty}|b(i)||R(k+i)|+\sum_{j=p+1}^{\infty}|b(j)||a(|k|-j)|\\
&\quad +\left(\sum_{i=p+1}^{\infty}|b(i)|^2\right)|R(k)|+\sum_{t=1}^{\infty}\sum_{i=p+1}^{\infty}|b(i)||b(i+t)|(|R(k+t)|+|R(k-t)|)
\end{aligned}$$

Summing over all $k\in\mathbb{Z}$,

$$\begin{aligned}
\sum_{k\in\mathbb{Z}} |R_p(k) - \overline{R}(k)| &\leq \left(\sum_{j=p+1}^{\infty} |b(j)|\right) \sum_{k\in\mathbb{Z}} |R(k-j)| + \left(\sum_{i=p+1}^{\infty} |b(i)|\right) \sum_{k\in\mathbb{Z}} |R(k+i)| \\
&\quad + \left(\sum_{j=p+1}^{\infty} |b(j)|\right) \sum_{k\in\mathbb{Z}} |a(|k|-j)| + \left(\sum_{i=p+1}^{\infty} |b(i)|^2\right) \sum_{k\in\mathbb{Z}} |R(k)| \\
&\quad + \left(\sum_{t=1}^{\infty} \sum_{i=p+1}^{\infty} |b(i)||b(i+t)|\right) \left(\sum_{k\in\mathbb{Z}} |R(k+t)| + \sum_{k\in\mathbb{Z}} |R(k-t)|\right) \\
&= \left(\sum_{j=p+1}^{\infty} |b(j)|\right) \left(2\sum_{k\in\mathbb{Z}} |R(k)| + \sum_{k=0}^{\infty} |a(k)|\right) \\
&\quad + \left(\sum_{i=p+1}^{\infty} |b(i)|\right)^2 \sum_{k\in\mathbb{Z}} |R(k)|
\end{aligned}$$

As $p \to \infty$, each term on the right hand side of the above inequality goes to zero, because the covariance sequence and the sequences $\{b(k)\}$ and $\{a(k)\}$ are in ℓ_1. Therefore,

$$\lim_{p\to\infty} \sum_{k\in\mathbb{Z}} |R_p(k) - \overline{R}(k)| = 0 \tag{3.4.23}$$

Combining the results of (3.4.22) and (3.4.23) we obtain

$$\lim_{p\to\infty} \sum_{k\in\mathbb{Z}} |\overline{R}_p(k) - \overline{R}(k)| = 0 \tag{3.4.24}$$

Finally,

$$\begin{aligned}
\lim_{p\to\infty} \left\| S_{\overline{X}_p}(\lambda) - S_{\overline{X}}(\lambda) \right\| &= \lim_{p\to\infty} \left| \int_{-\frac{1}{2}}^{\frac{1}{2}} \left| \sum_{k\in\mathbb{Z}} (\overline{R}_p(k) - \overline{R}(k)) e^{-2\pi\imath\lambda k} \right|^2 d\lambda \right|^{\frac{1}{2}} \\
&\leq \lim_{p\to\infty} \left| \int_{-\frac{1}{2}}^{\frac{1}{2}} \sum_{k\in\mathbb{Z}} |\overline{R}_p(k) - \overline{R}(k)|^2 d\lambda \right|^{\frac{1}{2}} \\
&\leq \lim_{p\to\infty} \left(\sum_{k\in\mathbb{Z}} |\overline{R}_p(k) - \overline{R}(k)| \right) \\
&= 0
\end{aligned}$$

where (3.4.24) is used for the last equality. This completes the proof. □

3.4.2 Convergence of the spectral density at the origin

We now study conditions under which the spectral density of the finite order AR approximation converges at the origin. As mentioned earlier, $S(0)$ is referred to as the Time Average Variance Constant (TAVC) and plays an important role in simulations.

Proposition 3.4.2. *Let $S(\lambda) > 0$ for $\lambda \in \left(-\frac{1}{2}, \frac{1}{2}\right]$ and let $\sum_{k\in\mathbb{Z}} |R(k)| < \infty$. Then, as $p \to \infty$, the spectral density of $\overline{\mathbb{E}}[X(n)|H^p(n-1)] = \overline{X}_p(n)$ converges to that of $\overline{\mathbb{E}}[X(n)|H(n-1)]$ at the origin.*

Proof. Refer to equation (3.4.21) for an expansion of $\overline{R}_p(k)$ for each k, for a given p:

$$\overline{R}_p(k) = \sum_{j=1}^{p} b_p^2(j)R(k) + \sum_{t=1}^{p-1}\sum_{j=1}^{p-t} b_p(j)b_p(j+t)(R(k-t)+R(k+t))$$

Summing the above over all $k \in \mathbb{Z}$ gives

$$\begin{aligned}\sum_{k\in\mathbb{Z}} \overline{R}_p(k) &= \left(\sum_{j=1}^{p} b_p^2(j) + 2\sum_{t=1}^{p-1}\sum_{j=1}^{p-t} b_p(j)b_p(j+t)\right)\sum_{k\in\mathbb{Z}} R(k) \\ &= \left(\sum_{j=1}^{p} b_p(j)\right)^2 \sum_{k\in\mathbb{Z}} R(k)\end{aligned}$$

From where it follows that

$$\lim_{p\to\infty}\sum_{k\in\mathbb{Z}} \overline{R}_p(k) = \lim_{p\to\infty}\left(\sum_{j=1}^{p} b_p(j)\right)^2 \sum_{k\in\mathbb{Z}} R(k) \tag{3.4.25}$$

Proceeding similarly, using the expression in (3.2.1), one obtains the following expression for $\sum_{k\in\mathbb{Z}} \overline{R}_k$:

$$\sum_{k\in\mathbb{Z}} \overline{R}_k = \left(\sum_{j=1}^{\infty} b(j)\right)^2 \sum_{k\in\mathbb{Z}} R(k) \tag{3.4.26}$$

Recall that by lemma 3.4.3, under the stated conditions,

$$\lim_{p\to\infty}\sum_{k=1}^{p} |b_p(k) - b(k)| = 0$$

Clearly, then,

$$\lim_{p\to\infty}\sum_{k=1}^{p}(b_p(k)-b(k))=0 \tag{3.4.27}$$

i.e.,

$$\lim_{p\to\infty}\left(\sum_{k=1}^{p}b_p(k)\right)^2=\left(\sum_{k=1}^{\infty}b(k)\right)^2 \tag{3.4.28}$$

Combining (3.4.28) with (3.4.25), and comparing with (3.4.26), we obtain

$$\lim_{p\to\infty}\sum_{k\in\mathbb{Z}}\overline{R}_p(k)=\left(\sum_{k=1}^{\infty}b(k)\right)^2\sum_{k\in\mathbb{Z}}R(k)=\sum_{k\in\mathbb{Z}}\overline{R}(k)$$

This completes the proof.

□

3.5 Conclusion

In this chapter, we have considered two kinds of finite order approximation for a real-valued, zero-mean WSS process $\{X(n)\}$: a moving average approximation and an autoregressive approximation. We have shown that both approximations converge in quadratic mean as $p\to\infty$.

We have provided a proof of the convergence of the spectral density of an autoregressive approximation of a WSS process when the spectral density is strictly positive and the covariance sequence is absolutely summable. Thus, under the said conditions, the TAVC exists and is well defined for the limiting AR approximation. Moreover, any unbiased spectral estimator derived from a finite autoregressive approximation will converge to the spectrum of the original process at the origin. This will enable easy approximation of the TAVC of the original process, which plays a significant role in the context of steady-state simulation.

Further, it has been shown that the spectral density of both the moving average and the autoregressive type approximations converge in L_2 when the spectral density is strictly positive and the coefficients of the Wold expansion are absolutely summable. These are fairly general conditions and are satisfied ny a large class of WSS stochastic processes.

A sufficient condition for the spectral density of a WSS process to be strictly positive is given in [105, theorem 11.2]. If $\{R(k)\}$ is such that

$$R(k) + R(k+2) \geq 2R(k+1) \text{ for } k \geq 0$$

and

$$R(0) + R(2) > 2R(1)$$

then the spectral density is strictly positive. For a Markov process, the strict positivity of the spectral density is guaranteed as long as the process has a continuous spectral density [103].

For a zero-mean wide sense stationary process having an infinite order moving average representation, one example where the condition $\sum_{0 \leq k < \infty} |a(k)| < \infty$ is met is the case when $R(k)$ tends to zero at an exponential rate as $k \to \infty$; i.e., there exists constants $C \in \mathbb{R}$, $\alpha \in (0,1)$ such that $R(k) \sim C\alpha^{|k|}$ [106]. A more trivial example is that of a process that has a finite order moving average representation. In such cases, the spectral density of the original process can be approximated over $\lambda \in \left(-\frac{1}{2}, \frac{1}{2}\right]$ from that of the finite order estimate.

Chapter 4

Convergence of Spectral Density of Empirically Computed AR Approximations of Stationary Time Series

4.1 Introduction

In the previous chapter, results were presented on the asymptotic behaviour of the spectral density of finite order MMSE estimates of WSS time series. However, due to the unavailability of the actual covariance sequence of the process under consideration, in most practical applications, AR approximations are computed using estimates of the covariance quantities based on a finite set of observations. In this chapter we study the asymptotic behavior of the spectral density of AR approximations when they are derived empirically. Throughout this chapter, the term "theoretical AR estimates" is used to refer to AR approximations based on solving the Yule-Walker equations using the true covariances and the term "empirical AR estimates" is used to refer to AR approximations based on empirical estimates of the covariance sequence.

Define $\{X(n)\}$ to be a **strong mixing**, *regular*, real-valued, zero-mean, discrete time WSS process with a strictly positive spectral density and an absolutely summable covariance sequence. The fourth moment of the corresponding innovation sequence is assumed to

be finite. We study the asymptotic properties of the spectral density of the AR estimates of such processes when they are derived using estimates of covariance computed using a sample of size N. Under a mild assumption, we show that as long as the model order $p = o\{N^{\frac{1}{3}}\}$, spectral density of the AR estimate converges in mean with respect to an L_2 norm, as both p and N approach infinity. It is further shown that under the same condition on N and p, the spectral density of the approximating AR sequence converges at the origin in mean.

Finally, we study the spectral estimation of two WSS processes using AR parameters through simulation for different sample sizes and model orders. These results complement our theoretical findings.

4.2 Preliminaries

Recall that, when the true covariance sequence of $\{X(n)\}$ is known, the AR parameters are computed using the Yule-Walker equations (3.2.6).

$$\boldsymbol{B_p} = \boldsymbol{R_p}^{-1}\boldsymbol{r_p}$$

where

$$\boldsymbol{R_p} = \begin{pmatrix} R(0) & R(1) & \cdots & R(p-1) \\ R(1) & R(0) & \cdots & R(p-2) \\ \vdots & \vdots & \ddots & \vdots \\ R(p-1) & R(p-2) & \cdots & R(0) \end{pmatrix}$$

$$\boldsymbol{r_p} = [R(1)\ R(2)\ \ldots\ R(p)]^T$$

and

$$\boldsymbol{B_p} = [b_p(1)\ \ldots\ b_p(p)]^T$$

$\boldsymbol{B_p}$ as defined above gives the *theoretical* parameters for the p-th order AR estimate. However, in practice, the covariance sequence $\{R(k)\}$ of $\{X(n)\}$ is often unavailable and has to be estimated from a sequence of observations. Given an ensemble of N observations of $\{X(n)\}$, $R(k)$ is typically estimated by

$$\hat{R}_N(k) = \begin{cases} \frac{1}{N}\sum_{n=|k|+1}^{N} X(n)X(n-|k|), & |k| \leq N-1 \\ 0, & |k| \geq N \end{cases} \tag{4.2.1}$$

The p-th order AR parameters $\{\hat{b}_{p,N}(k)\}$ (for some model order $p < N$) are then estimated by replacing $R(k)$ with $\hat{R}_N(k)$ in the Yule-Walker equations:

$$\hat{\boldsymbol{B}}_{\boldsymbol{p,N}} = \hat{\boldsymbol{R}}_{\boldsymbol{p,N}}^{-1}\hat{\boldsymbol{r}}_{\boldsymbol{p,N}} \tag{4.2.2}$$

The corresponding AR estimate of $X(n)$ is given by

$$\hat{X}_{p,N}(n) = \sum_{k=1}^{p} \hat{b}_{p,N}(k) X(n-k) \tag{4.2.3}$$

Unlike the theoretical AR parameters $\{b_p(k)\}$, which are fixed for a given p, the estimated AR parameters $\{\hat{b}_{p,N}(k)\}$ are random variables and in general, $\hat{X}_{p,N}(n)$ is different from the theoretically derived $\overline{X}_p(n)$.

Let $\{\tilde{R}_{p,N}(k)\}_{k\in\mathbb{Z}}$ denote the covariance sequence of the estimates $\{\hat{X}_{p,N}(n)\}$, conditioned on the set of parameters $\hat{\boldsymbol{B}}_{\boldsymbol{p,N}} = [\hat{b}_{p,N}(1) \ \ldots \ \hat{b}_{p,N}(p)]^T$.

$$\tilde{R}_{p,N}(k) = \mathbb{E}[\hat{X}_{p,N}(n)\hat{X}_{p,N}(n-k)|\{\hat{\boldsymbol{B}}_{\boldsymbol{p,N}}\}] \tag{4.2.4}$$

Then $\{\tilde{R}_{p,N}(k)\}_{k\in\mathbb{Z}}$ is dependent on the parameters $\{\hat{b}_{p,N}(k)\}$ and thus a stochastic sequence. Let $S_{\hat{X}_{p,N}}(\lambda)$ denote the corresponding spectral density.

$$S_{\hat{X}_{p,N}}(\lambda) = \sum_{k\in\mathbb{Z}} \tilde{R}_{p,N}(k) e^{-2\pi\imath\lambda k}$$

For a given N and p, the spectral density of $\{\hat{X}_{p,N}(n)\}$ is not a deterministic function but a stochastic process defined on $\left(-\frac{1}{2}, \frac{1}{2}\right] \times (\Omega, \mathcal{F}, \mathbb{P})$. For a fixed value of λ, $S_{\hat{X}_{p,N}}(\lambda)$ is a random variable.

Let $\|\cdot\|_2$ denote the vector Euclidean norm and let $\|\cdot\|_2$ and $\|\cdot\|_F$ denote the spectral norm and the Frobenius norm of matrices respectively.

Let $\{X(n)\}$ be a real-valued, WSS sequence that is strong mixing and whose innovation sequence has a finite fourth moment. As in [45] (a similar assumption can also be found in [39]) we assume that the following condition holds.

A 4.2.1. *There exists an $N_0 \in \mathbb{Z}$ such that for all $N > N_0$ and for all $p < N$,*

$$\left\| \boldsymbol{R}_{\boldsymbol{p}}^{-1}(\hat{\boldsymbol{R}}_{\boldsymbol{p,N}} - \boldsymbol{R}_{\boldsymbol{p}}) \right\|_2 \leq K_0 < 1$$

Lemma 4.2.1. *At any instant n, for any $l, m \in \{1, \ldots, p\}$, $X(n-l)X(n-m)$ are asymptotically independent of the empirical AR parameter set $\{\hat{\boldsymbol{B}}_{\boldsymbol{p,N}}\}$ as $|(n-p) - N| \to \infty$.*

Proof. The above follows from the fact that $\{X(n)\}$ is a strong mixing process. The difference $|(n-p) - N|$ indicates the separation in time between the computation of the parameters $\hat{b}_{p,N}(k)$ and the finite past of $\{X(n)\}$ used in its autoregressive approximation.

Let $\mathcal{F}_i^j = \sigma\{X(t) : i \leq t \leq j\}$ be the sigma-algebra generated by the random variables $\{X(i), \ldots, X(j)\}$. Let $A \in \mathcal{F}_{-\infty}^i$ and $B \in \mathcal{F}_{i+t}^{\infty}$ with $t > 0$. Define the strong mixing coefficient α_t as follows:

$$\alpha_t = \sup_{A,B} |\mathbb{P}(A \cap B) - \mathbb{P}(A)\mathbb{P}(B)|$$

By the strong mixing property, we have

$$\lim_{t\to\infty} \sup_{A,B} |\mathbb{P}(A \cap B) - \mathbb{P}(A)\mathbb{P}(B)| = 0$$

For some positive integer N, $\hat{R}_N(k)$ is $\mathcal{F}_{-\infty}^N$-measurable (as per its definition in (4.2.1)) for $k \in \{1, \ldots, N\}$. Each $\hat{b}_{p,N}(k)$ is a measurable function of $\hat{R}_N(1), \ldots, \hat{R}_N(N)$; and therefore also $\mathcal{F}_{-\infty}^N$-measurable, for $k \in \{1, \ldots, p\}$. On the other hand, the product terms $X(n-l)X(n-m)$ are $\mathcal{F}_{n-p}^{\infty}$-measurable for $l, m \in \{1, \ldots, p\}$. Let $\sigma\{\hat{\boldsymbol{B}}_{\boldsymbol{p,N}}\}$ and $\sigma\{X(n-l)X(n-m)\}$ denote the sigma-algebras generated by $\{\hat{b}_{p,N}(k)\}$ and $X(n-l)X(n-m)$ with $k, l, m \in \{1, \ldots, p\}$. Then,

$$\sigma\{\hat{\boldsymbol{B}}_{\boldsymbol{p,N}}\} \subset \mathcal{F}_{-\infty}^N \text{ and } \sigma\{X(n-l)X(n-m)\} \subset \mathcal{F}_{n-p}^{\infty}$$

From the strong mixing property, it follows that for any $C \in \sigma\{\hat{\boldsymbol{B}}_{\boldsymbol{p,N}}\}$ and $D \in \sigma\{X(n-l)X(n-m)\}$,

$$\lim_{|(n-p)-N|\to\infty} |\mathbb{P}(C \cap D) - \mathbb{P}(C)\mathbb{P}(D)| = 0$$

and the result follows. □

As a corollary, it also follows that

$$\lim_{|(n-p)-N|\to\infty} \left| \mathbb{E}[\hat{b}_{p,N}(k)\hat{b}_{p,N}(j)X(n-l)X(n-m)] - \mathbb{E}[\hat{b}_{p,N}(k)\hat{b}_{p,N}(j)]\mathbb{E}[X(n-l)X(n-m)] \right| = 0$$

Thus, whenever $(n-p) >> N$, $\hat{b}_{p,N}(k)\hat{b}_{p,N}(j)$ and $X(n-l)X(n-m)$ are uncorrelated.

Lemma 4.2.2. *Let $R(k)$ be the covariance sequence and let $\hat{R}_N(k)$ be its estimate defined by (4.2.1). Let the spectral density $S(\lambda)$ be square-integrable, i.e.,*

$$\int_{-\frac{1}{2}}^{\frac{1}{2}} (S(\lambda))^2 d\lambda < \infty$$

and let $\mathbb{E}[\nu^2(n)] < \infty$. Then, for large N, $\mathbb{E}[(\hat{R}_N(k) - R(k))^2] \leq \frac{K_1}{N}$ where K_1 is a constant.

Define $c_N(k)$ as the least-square estimate of $R(k)$, i.e.,

$$c_N(k) = \frac{1}{N-|k|} \sum_{n=|k|+1}^{N} X(n)X(n-|k|) \text{ for } |k| \leq N-1$$

Hannan[107, P-39] has shown that

$$\lim_{N\to\infty} (N-|k|)\mathbb{E}[(c_N(k) - R(k))]^2 \leq K_1'$$

where K_1' is some positive constant. The result follows readily, observing that $\hat{R}_N(k) = \frac{N-|k|}{N} c_N(k)$.

Next we present a key lemma on the mean square convergence of the AR parameters. This result is stronger than that by Berk [39] where a convergence in probability of the AR parameters was shown.

Lemma 4.2.3. *Let $\{X(n)\}$ be a real-valued zero-mean, strong mixing, wide sense stationary process with covariance sequence $\{R(k)\}$ and spectral density $S(\lambda)$ such that $\{R(k)\}$ is summable and $S(\lambda)$ is strictly positive for $\lambda \in \left(-\frac{1}{2}, \frac{1}{2}\right]$. Let A 4.2.1 hold and let $\mathbb{E}[\nu^4(n)] < \infty$. Then, when model order $p = o\{N^{\frac{1}{3}}\}$*

$$\lim_{N\to\infty} \mathbb{E}\left[\left(\sum_{k=1}^{p} |b_p(k) - \hat{b}_{p,N}(k)|\right)^2\right] = 0$$

Proof. Begin by noting that the conditions of lemma 4.2.2 are satisfied as a square-summable covariance sequence guarantees a square-integrable spectral density by Parseval's theorem. Recall that $\sum_{i=1}^{p}(\hat{b}_{p,N}(k) - b_p(k))^2$ gives the square of the Euclidean norm of the difference between vectors $\boldsymbol{B_p}$ and $\hat{\boldsymbol{B}}_{\boldsymbol{p,N}}$.

It follows from the strict positivity of the spectral density that $\left\|R_p^{-1}\right\|_2$ is bounded for all p [108]. $\|\boldsymbol{r}_p\|_2$ is bounded as the covariance sequence is square-summable. Let $\left(\sup_p \left\|R_p^{-1}\right\|_2 \|\boldsymbol{r}_p\|_2\right)^2 = K_2$. Recall also, that the Frobenius norm of a matrix is always at least as large as its spectral norm.

Let $\boldsymbol{R_p^{-1}}(\hat{\boldsymbol{R}}_{\boldsymbol{p,N}} - \boldsymbol{R_p}) = \boldsymbol{C_{p,N}}$. Observe that

$$\begin{aligned}\left\|\hat{\boldsymbol{R}}_{\boldsymbol{p,N}}^{-1}\right\|_2 &= \left\|R_p^{-1}(\boldsymbol{I} + \boldsymbol{C_{p,N}})^{-1}\right\|_2 \\ &\leq \left\|\boldsymbol{R_p^{-1}}\right\|_2 \left\|(\boldsymbol{I} + \boldsymbol{C_{p,N}})^{-1}\right\|_2\end{aligned}$$

By A 4.2.1, for all $N > N_0$, the quantity $(\boldsymbol{I} + \boldsymbol{C_{p,N}})^{-1}$ can be expanded as a power series and we obtain

$$\begin{aligned}
\left\|\hat{\boldsymbol{R}}_{\boldsymbol{p,N}}^{-1}\right\|_2 &\leq \left\|\boldsymbol{R_p^{-1}}\right\|_2 \left\|\boldsymbol{I} + \sum_{k=1}^{\infty}(-\boldsymbol{C_{p,N}})^k\right\|_2 \\
&\leq \left\|\boldsymbol{R_p^{-1}}\right\|_2 \left(1 + \sum_{k=1}^{\infty}\|-\boldsymbol{C_{p,N}}\|_2^k\right) \\
&\leq \left\|\boldsymbol{R_p^{-1}}\right\|_2 \left(1 + \sum_{k=1}^{\infty}K_0^k\right) \\
&= \frac{\left\|\boldsymbol{R_p^{-1}}\right\|_2}{1 - K_0}
\end{aligned}$$

Therefore, $\left\|\hat{\boldsymbol{R}}_{\boldsymbol{p,N}}^{-1}\right\|_2$ is bounded for all $N > N_0$. Let

$$\sup_{p \leq N, N > N_0} \left\|\hat{\boldsymbol{R}}_{\boldsymbol{p,N}}^{-1}\right\|_2^2 = K_3$$

$$\begin{aligned}
\left\|\boldsymbol{B_p} - \hat{\boldsymbol{B}}_{\boldsymbol{p,N}}\right\|_2 &= \left\|\boldsymbol{R_p^{-1}}\boldsymbol{r_p} - \hat{\boldsymbol{R}}_{\boldsymbol{p,N}}^{\boldsymbol{-1}}\hat{\boldsymbol{r}}_{\boldsymbol{p,N}}\right\|_2 \\
&= \left\|\boldsymbol{R_p}^{-1}(\hat{\boldsymbol{R}}_{\boldsymbol{p,N}} - \boldsymbol{R_p})\hat{\boldsymbol{R}}_{\boldsymbol{p,N}}^{-1}\boldsymbol{r}_p + \hat{\boldsymbol{R}}_{\boldsymbol{p,N}}^{\boldsymbol{-1}}(\boldsymbol{r}_p - \hat{\boldsymbol{r}}_{\boldsymbol{p,N}})\right\|_2 \\
&\leq \left\|\hat{\boldsymbol{R}}_{\boldsymbol{p,N}}^{-1}\right\|_2 \left(\left\|\boldsymbol{R_p^{-1}}\right\|_2 \left\|\hat{\boldsymbol{R}}_{\boldsymbol{p,N}} - \boldsymbol{R_p}\right\|_F \|\boldsymbol{r}_p\|_2 + \|\boldsymbol{r}_p - \hat{\boldsymbol{r}}_{\boldsymbol{p,N}}\|_2\right) \\
&\leq \left\|\hat{\boldsymbol{R}}_{\boldsymbol{p,N}}^{-1}\right\|_2 \left(\sqrt{p\sum_{k=-p}^{p}(R(k) - \hat{R}_N(k))^2}\left\|\boldsymbol{R_p^{-1}}\right\|_2 \|\boldsymbol{r}_p\|_2\right) \\
&\quad + \left\|\hat{\boldsymbol{R}}_{\boldsymbol{p,N}}^{-1}\right\|_2 \left(\sqrt{\sum_{k=-p}^{p}(R(k) - \hat{R}_N(k))^2}\right)
\end{aligned}$$

It follows, then, that

$$\begin{aligned}
\mathbb{E}\left[\left(\left\|\boldsymbol{B_p} - \hat{\boldsymbol{B}}_{\boldsymbol{p,N}}\right\|_2\right)^2\right] &\leq K_3(2p+1)\sup_{-p \leq k \leq p}\mathbb{E}\left[\left(R(k) - \hat{R}_N(k)\right)^2\right]\left(\sqrt{K_2 p} + 1\right)^2 \\
&\leq K_3(2p+1)\frac{K_1}{N}\left(\sqrt{K_2 p} + 1\right)^2
\end{aligned}$$

where lemma 4.2.2 has been used. Therefore,

$$\begin{aligned}\lim_{N\to\infty} \mathbb{E}\left[\left(\sum_{k=1}^{p} |b_p(k) - \hat{b}_{p,N}(k)|\right)^2\right] &\leq \lim_{N\to\infty} p\mathbb{E}\left[\left(\left\|\boldsymbol{B_p} - \hat{\boldsymbol{B}}_{\boldsymbol{p,N}}\right\|_2\right)^2\right] \\ &\leq \lim_{N\to\infty} K_3 p(2p+1)\frac{K_1}{N}(\sqrt{K_2 p}+1)^2\end{aligned}$$

Clearly, when $p = o\{N^{\frac{1}{3}}\}$ [1], the above limit is zero. This completes the proof. □

4.3 Convergence of the spectral density of the empirical AR estimate

4.3.1 Convergence of the spectral density in L_2

In this subsection, we study the asymptotic behavior of the spectral density of the empirically derived AR estimate with respect to an L_2 norm. As before, for all functions $F : \left(-\frac{1}{2}, \frac{1}{2}\right] \to \mathbb{C}$ such that $\int_{-\frac{1}{2}}^{\frac{1}{2}} |F(\lambda)|^2 d\lambda < \infty$, define $||\cdot||$ to be the L_2 norm as follows:

$$\|F\| = \left|\int_{-\frac{1}{2}}^{\frac{1}{2}} |F(\lambda)|^2 d\lambda\right|^{\frac{1}{2}}$$

We consider the limiting properties of $\left\|\mathbb{E}\left[\left|S_{\hat{X}_{p,N}} - S_{\overline{X}}\right|\right]\right\|$ as both p, N go to infinity.

Proposition 4.3.1. *Let $\{X(n)\}$ be a real-valued zero-mean, strong mixing, wide sense stationary process with covariance sequence $\{R(k)\}$ and spectral density $S(\lambda)$ such that $\{R(k)\}$ is in ℓ_1 and $S(\lambda)$ is strictly positive for $\lambda \in \left(-\frac{1}{2}, \frac{1}{2}\right]$. Let A 4.2.1 hold and let $\mathbb{E}[\nu^4(n)] < \infty$. Then, when model order $p = o\{N^{\frac{1}{3}}\}$; as $p \to \infty$, $\left\|\mathbb{E}\left[\left|S_{\hat{X}_{p,N}} - S_{\overline{X}}\right|\right]\right\|$ converges to 0.*

[1] Standard notation:

$$\begin{aligned}k = o\{Z\} &\Rightarrow \lim_{Z\to\infty} \frac{|k(Z)|}{|Z|} = 0 \\ k = O\{Z\} &\Rightarrow \lim_{Z\to\infty} \sup \frac{|k(Z)|}{|Z|} < \infty\end{aligned}$$

Proof. Consider the p-th order AR estimate $\sum_{k=1}^{p} b_p(k)X(n-k)$ and its covariance sequence $\{\overline{R}_p(k)\}$, given by (3.4.21):

$$\overline{R}_p(k) = \sum_{j=1}^{p} b_p^2(j)R(k) + \sum_{t=1}^{p-1}\sum_{j=1}^{p-t} b_p(j)b_p(j+t)(R(k-t)+R(k+t))$$

It follows from (4.2.4), that

$$\begin{aligned}
\tilde{R}_{p,N}(k) &= \mathbb{E}[\hat{X}_{p,N}(n)\hat{X}_{p,N}(n-k)|\{\hat{\boldsymbol{B}}_{\boldsymbol{p,N}}\}] \\
&= \mathbb{E}\left[\left(\sum_{j=1}^{p}\hat{b}_{p,N}(j)X(n-j)\sum_{l=1}^{p}\hat{b}_{p,N}(l)X(n-k-l)\right)\middle|\{\hat{\boldsymbol{B}}_{\boldsymbol{p,N}}\}\right] \\
&= \sum_{j=1}^{p}\hat{b}_{p,N}(j)^2R^*(k) + \sum_{t=1}^{p-1}\sum_{j=1}^{p-t}\hat{b}_{p,N}(j)\hat{b}_{p,N}(j+t)(R^*(k-t)+R^*(k+t))
\end{aligned}$$

where $R^*(k) = \mathbb{E}[X(n)X(n-k)|\{\hat{\boldsymbol{B}}_{\boldsymbol{p,N}}\}]$. When the parameters $\{\hat{b}_{p,N}(k)\}$ are computed sufficiently ahead of time from the current ensemble of the finite past of $\{X(n)\}$, ie., when $(n-p) >> N$, $R^*(k)$ can be replaced by $R(k)$ by virtue of the asymptotic independence shown in lemma 4.2.1 and we obtain

$$\tilde{R}_{p,N}(k) = \sum_{j=1}^{p}\hat{b}_{p,N}(j)^2R(k) + \sum_{t=1}^{p-1}\sum_{j=1}^{p-t}\hat{b}_{p,N}(j)\hat{b}_{p,N}(j+t)(R(k-t)+R(k+t)) \tag{4.3.5}$$

For all $k \in \mathbb{Z}$

$$\begin{aligned}
&|\tilde{R}_{p,N}(k) - \overline{R}_p(k)| \\
&= \left|\sum_{j=1}^{p}(\hat{b}_{p,N}(j)^2 - b_p^2(j))R(k)\right. \\
&\left.+\sum_{t=1}^{p-1}\sum_{j=1}^{p-t}(\hat{b}_{p,N}(j)\hat{b}_{p,N}(j+t) - b_p(j)b_p(j+t)\,(R(k-t)+R(k+t))\right| \\
&\leq \sum_{j=1}^{p}\left|\hat{b}_{p,N}(j)^2 - b_p^2(j)\right||R(k)| \\
&+\sum_{t=1}^{p-1}\sum_{j=1}^{p-t}\left|\hat{b}_{p,N}(j)\hat{b}_{p,N}(j+t) - b_p(j)b_p(j+t)\right|\Big(|R(k-t)|+|R(k+t)|\Big)
\end{aligned}$$

Summing over all $k \in \mathbb{Z}$

$$\begin{aligned}
&\sum_{k\in\mathbb{Z}} |\tilde{R}_{p,N}(k) - \overline{R}_p(k)| \\
\leq\ & \left(\sum_{j=1}^{p}\sum_{i=1}^{p} \left|\hat{b}_{p,N}(i)\hat{b}_{p,N}(j) - b_p(i)b_p(j)\right|\right) \sum_{k\in\mathbb{Z}} |R(k)| \\
\leq\ & \left(\sum_{j=1}^{p}\sum_{i=1}^{p} |\hat{b}_{p,N}(i)||\hat{b}_{p,N}(j) - b_p(j)| + \sum_{j=1}^{p}\sum_{i=1}^{p} |b_p(j)||\hat{b}_{p,N}(i) - b_p(i)|\right) \sum_{k\in\mathbb{Z}} |R(k)| \\
=\ & \left(\sum_{i=1}^{p} |\hat{b}_{p,N}(i)| + \sum_{i=1}^{p} |b_p(i)|\right) \left(\sum_{i=1}^{p} |\hat{b}_{p,N}(i) - b_p(i)|\right) \sum_{k\in\mathbb{Z}} |R(k)| \qquad (4.3.6)
\end{aligned}$$

Then,

$$\begin{aligned}
\left\|\mathbb{E}\left[\left|S_{\hat{X}_{p,N}} - S_{\overline{X}}\right|\right]\right\| &= \left\|\mathbb{E}\left[\left|\sum_{k\in\mathbb{Z}} (\tilde{R}_{p,N}(k) - \overline{R}(k)) \exp^{-2\pi i\lambda k}\right|\right]\right\| \\
&\leq \left\|\mathbb{E}\left[\sum_{k\in\mathbb{Z}} |\tilde{R}_{p,N}(k) - \overline{R}_p(k)|\right]\right\| + \left\|\sum_{k\in\mathbb{Z}} |\overline{R}_p(k) - \overline{R}(k)|\right\| \qquad (4.3.7)
\end{aligned}$$

It was shown in proposition 3.4.1, that under the given conditions, the second term of the above inequality goes to zero as p (and N) goes to infinity. From (4.3.6), it follows that,

$$
\begin{aligned}
& \lim_{N\to\infty} \mathbb{E}\left[\sum_{k\in\mathbb{Z}} |\tilde{R}_{p,N}(k) - \overline{R}_p(k)|\right] \\
\leq\ & \lim_{N\to\infty} \mathbb{E}\left[\left(\sum_{i=1}^{p} |\hat{b}_{p,N}(i)| + \sum_{i=1}^{p} |b_p(i)|\right)\left(\sum_{i=1}^{p} |\hat{b}_{p,N}(i) - b_p(i)|\right)\right] \\
& \sum_{k\in\mathbb{Z}} |R(k)| \\
\leq\ & \left(\sum_{k\in\mathbb{Z}} |R(k)|\right) \lim_{N\to\infty} \left(\sqrt{\mathbb{E}\left[\left(\sum_{i=1}^{p} |\hat{b}_{p,N}(i)|\right)^2\right]}\right. \\
& \sqrt{\mathbb{E}\left[\left(\sum_{i=1}^{p} |\hat{b}_{p,N}(i) - b_p(i)|\right)^2\right]} \\
& \left. + \left(\sum_{i=1}^{p} |b_p(i)|\right) \mathbb{E}\left[\left(\sum_{i=1}^{p} |\hat{b}_{p,N}(i) - b_p(i)|\right)\right]\right) \qquad (4.3.8)
\end{aligned}
$$

where we use the Cauchy-Schwarz inequality. It follows from Baxter's inequality [38, Theorem 2.2] that $(\sum_{i=1}^{p} |b_p(i)|)$ is bounded for all p. By lemma 4.2.3, $\mathbb{E}\left[\left(\sum_{i=1}^{p} |\hat{b}_{p,N}(i) - b_p(i)|\right)^2\right]$ converges to zero. Consequently, the lower moment

$$
\mathbb{E}\left[\left(\sum_{i=1}^{p} |\hat{b}_{p,N}(i) - b_p(i)|\right)\right]
$$

goes to zero as well. Finally, it can be shown using Baxter's inequality and lemma 4.2.3 that $\lim_{N\to\infty} \mathbb{E}\left[\left(\sum_{i=1}^{p} |\hat{b}_{p,N}(i)|\right)^2\right]$ is bounded. Therefore, the first term in equation (4.3.7) converges to zero and the result follows. □

4.3.2 Convergence of the spectral density at the origin

Finally, we present a result on the convergence of $S_{\hat{X}_{p,N}}(0)$ to $S_{\overline{X}}(0)$.

Proposition 4.3.2. *Let $\{X(n)\}$ be a real-valued zero-mean, strong mixing, wide sense stationary process with covariance sequence $\{R(k)\}$ and spectral density $S(\lambda)$ such that $\{R(k)\}$ is in ℓ_1 and $S(\lambda)$ is strictly positive for $\lambda \in \left(-\frac{1}{2}, \frac{1}{2}\right]$. Let A 4.2.1 hold and let $\mathbb{E}[\nu^4(n)] < \infty$. Then, when model order $p = o\{N^{\frac{1}{3}}\}$, as $p \to \infty$, $S_{\hat{X}_{p,N}}(0)$ converges to $S_{\overline{X}}(0)$ in mean.*

Proof. Refer to (4.3.5) for an expression for the estimated covariance $\tilde{R}_{p,N}(k)$. Summing these terms over all $k \in \mathbb{Z}$ gives

$$\begin{aligned}\sum_{k\in\mathbb{Z}} \tilde{R}_{p,N}(k) &= \left(\sum_{j=1}^{p} \hat{b}^2_{p,N}(j) + 2\sum_{t=1}^{p-1}\sum_{j=1}^{p-t} \hat{b}_{p,N}(j)\hat{b}_{p,N}(j+t)\right)\sum_{k\in\mathbb{Z}} R(k) \\ &= \left(\sum_{j=1}^{p} \hat{b}_{p,N}(j)\right)^2 \sum_{k\in\mathbb{Z}} R(k)\end{aligned}$$

Using the expression for $\sum_{k\in\mathbb{Z}} \overline{R}(k)$ derived in (3.4.26), we obtain

$$\begin{aligned}&\mathbb{E}\left[\left|S_{\hat{X}_{p,N}}(0) - S_{\overline{X}}(0)\right|\right] \\ =\ &\mathbb{E}\left[\left|\sum_{k\in\mathbb{Z}} \tilde{R}_{p,N}(k) - \sum_{k\in\mathbb{Z}} \overline{R}(k)\right|\right] \\ =\ &\mathbb{E}\left[\left|\left(\sum_{j=1}^{p} \hat{b}_{p,N}(j)\right)^2 - \left(\sum_{j=1}^{\infty} b(j)\right)^2\right|\right] \left|\sum_{k\in\mathbb{Z}} R(k)\right| \\ =\ &\mathbb{E}\left[\left(\sum_{j=1}^{p} \hat{b}_{p,N}(j) + \sum_{j=1}^{\infty} b(j)\right)\left(\sum_{j=1}^{p} \hat{b}_{p,N}(j) - \sum_{j=1}^{\infty} b(j)\right)\right] \sum_{k\in\mathbb{Z}} R(k) \\ \leq\ &\sqrt{\mathbb{E}\left[\left(\sum_{j=1}^{p} \hat{b}_{p,N}(j) + \sum_{j=1}^{\infty} b(j)\right)^2\right] \mathbb{E}\left[\left(\sum_{j=1}^{p} \hat{b}_{p,N}(j) - \sum_{j=1}^{\infty} b(j)\right)^2\right]} \left|\sum_{k\in\mathbb{Z}} R(k)\right|\end{aligned}$$

where the Cauchy-Schwarz inequality has been used at the last step. By construction, $\hat{\boldsymbol{R}}_{\boldsymbol{p,N}}$ is a positive semi-definite Toeplitz matrix [45]. Consider the characteristic polynomial given by:

$$P(x) = 1 - \hat{b}_{p,N}(1)x - \hat{b}_{p,N}(2)x^2 - \cdots - \hat{b}_{p,N}(p)x^p$$

Since $\hat{\boldsymbol{R}}_{\boldsymbol{p,N}}$ is positive semi-definite, $P(x)$ must have all its roots outside the unit disk [109, P-540], [110]. It follows, then, that the algebraic sum of the parameters $\hat{b}_{p,N}(k)$ is bounded by 1 [111]. Similarly, the sum of theoretically derived AR parameters, i.e., $\sum_{k\in\mathbb{Z}} b_k$ is also bounded by 1. Therefore,

$$\begin{aligned}
&\mathbb{E}\left[\left|S_{\hat{X}_{p,N}}(0) - S_{\overline{X}}(0)\right|\right] \\
&\leq 2\sqrt{\mathbb{E}\left[\left(\sum_{j=1}^{p}\hat{b}_{p,N}(j) - \sum_{j=1}^{\infty} b(j)\right)^2\right]}\left|\sum_{k\in\mathbb{Z}} R(k)\right| \\
&= 0
\end{aligned}$$

by lemma 4.2.3 and lemma 3.4.3. □

4.4 Simulation results

In this section, we present simulation results for the spectral estimation problem using AR approximations based on a finite number of observations. The WSS processes under consideration include an AR(12) process and an ARMA(4,4) process. Note that the first process has a finite order AR representation while the second has an infinite order AR representation. The AR(12) process is given by

$$\begin{aligned}
X(n) \;=\; & \nu(n) + 0.9X(n-1) - 0.75X(n-2) + 0.8X(n-3) - 0.6X(n-4) + 0.5X(n-5) \\
& -0.45X(n-6) + 0.3X(n-7) - 0.25X(n-8) + 0.15X(n-9) + 0.05X(n-11) \\
& +0.25X(n-12)
\end{aligned}$$

and the ARMA(4,4) process is given by

$$\begin{aligned}
X(n) \;=\; & \nu(n) + 0.9\nu(n-1) - 0.5\nu(n-1) - 0.2\nu(n-3) + 0.1\nu(n-4) \\
& +0.7X(n-1) - 0.6X(n-2) + 0.4X(n-3) - 0.5X(n-4)
\end{aligned}$$

The spectral densities of these two processes are given in figures 4.1 and 4.2 respectively.

Two types of innovation process are considered. First, the innovation sequence $\{\nu(n)\}$ is generated from a standard normal distribution. Next, $\{\nu(n)\}$ is generated from a Gaussian mixture distribution with the following parameters:

$$\boldsymbol{\mu} = [-0.5 \;\; -0.4 \; 0 \; 0.4 \; 0.5]^T;\; \boldsymbol{\sigma^2} = [.25 \;\; .44 \;\; 1 \;\; 1.34 \;\; 1.1]^T;\; \boldsymbol{w} = [0.25 \;\; 0.15 \;\; .2 \;\; .3 \;\; .1]^T$$

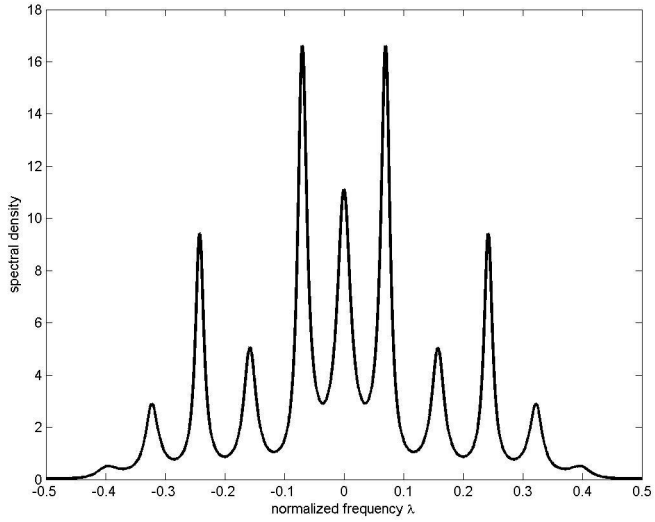

Figure 4.1: Spectral density of AR(12) process under consideration

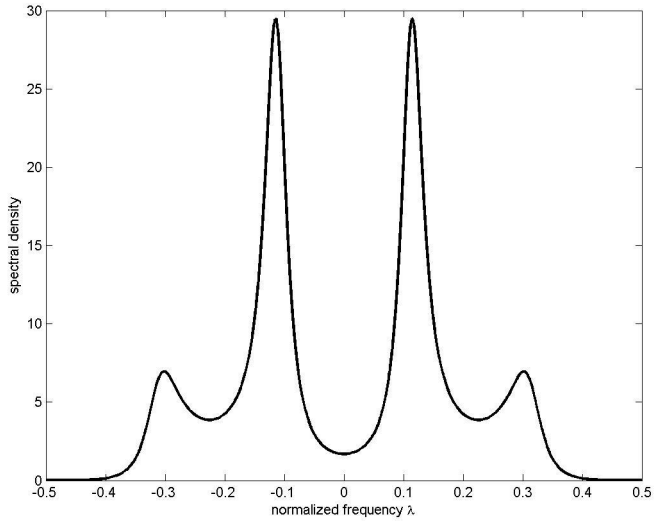

Figure 4.2: Spectral density of ARMA(4,4) process under consideration

where $\boldsymbol{\mu}$, $\boldsymbol{\sigma^2}$ and $\boldsymbol{w}$ denote the expectations, variances and weights of the respective components.

For both types of innovation, samples of sizes 1000, 20,000 and 100,000 are drawn arbitrarily; and for each sample size, various model orders are considered. Using the empirical AR parameters $[\hat{b}_{p,N}(1) \ \ldots \ \hat{b}_{p,N}(p)]^T$, the spectral density is estimated as

$$\hat{S}_{X_{p,N}}(\lambda) = \frac{1}{\left(1 - \sum_{k=1}^{p} \hat{b}_{p,N}(k) e^{-2\pi\imath\lambda k}\right)\left(1 - \sum_{k=1}^{p} \hat{b}_{p,N}(k) e^{2\pi\imath\lambda k}\right)}$$

For each set of observations i, $i = 1, \ldots, 50$; the corresponding error $\zeta_{p,N}(i)$ in estimation is found by computing numerically the L_2 norm of the difference $|\hat{S}_{X_{p,N}}(\lambda) - S(\lambda)|$ over $\left(-\frac{1}{2}, \frac{1}{2}\right]$:

$$\zeta_{p,N}(i) = \sqrt{\int_{-\frac{1}{2}}^{\frac{1}{2}} |\hat{S}_{X_{p,N}}(\lambda) - S(\lambda)|^2 d\lambda}$$

Finally, for each pair of p and N, the expected value of $\zeta_{p,N}$ is estimated from the sample mean calculated over 50 runs of simulation. Semilogarithmic plots of average error versus $\frac{p^3}{N}$ for Gaussian and Gaussian mixture innovation are presented in figures 4.3, 4.4 and figures 4.5 and 4.6 respectively.

It is seen that for both the processes, and for both types of innovation sequence, for a finite number of observations, the estimation error reaches a minimum when p is close to $N^{\frac{1}{3}}$. For lower values of $\frac{p^3}{N}$ the error is high due to underestimation, as the model order p is too low. For large p, the effect of bias in higher order terms of the estimated covariance sequence $\hat{R}_N(k)$ becomes significant, thereby increasing the error. It is also noted that for both processes, a larger sample size corresponds to a lower estimation error for the same $\frac{p^3}{N}$ ratio.

We conclude this section with simulation results that examine the behavior of the estimation error as the sample size N increases while the model order p remains fixed. For the AR(12) process we keep the model order fixed at 20 and calculate the average L_2 error in spectral estimation for samples sizes of 5000, 10,000, 20,000, 50,000, 100,000 and 500,000. For the ARMA(4,4) process the same sample sizes are considered for model orders 20 and 200. The plots for Gaussian and Gaussian mixture innovations are presented in figures 4.7, 4.8 and figures 4.9 and 4.10 respectively.

For a fixed p, as the sample size is increased, the estimated covariance terms $\hat{R}_N(k)$ approach the original covariance quantities $R(k)$ for $k = 1, \ldots, p$; thereby leading to the

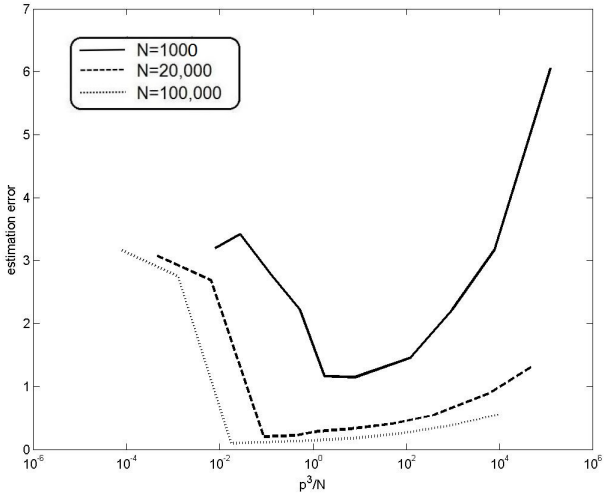

Figure 4.3: Estimation error for different sample size (N) and model order (p) for AR(12) process - Gaussian innovation

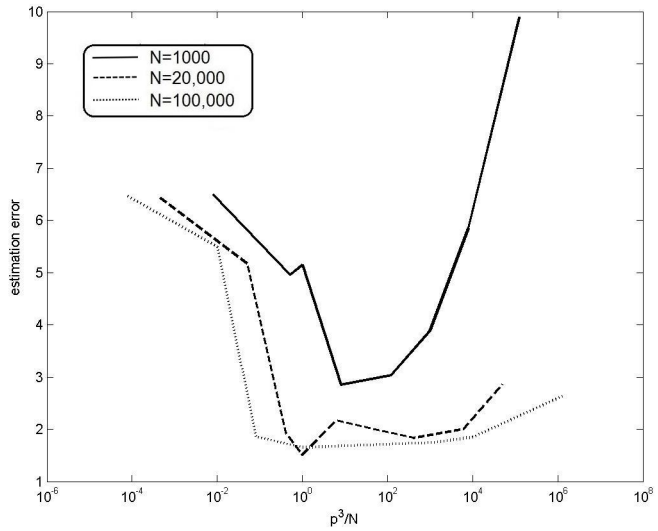

Figure 4.4: Estimation error for different sample size (N) and model order (p) for ARMA(4,4) process - Gaussian innovation

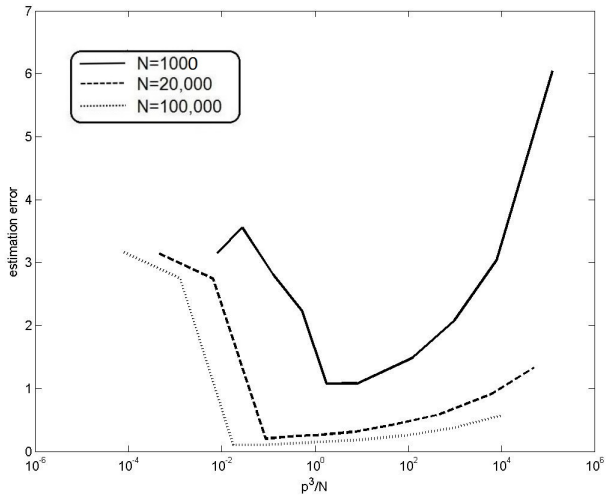

Figure 4.5: Estimation error for different sample size (N) and model order (p) for AR(12) process - Gaussian mixture innovation

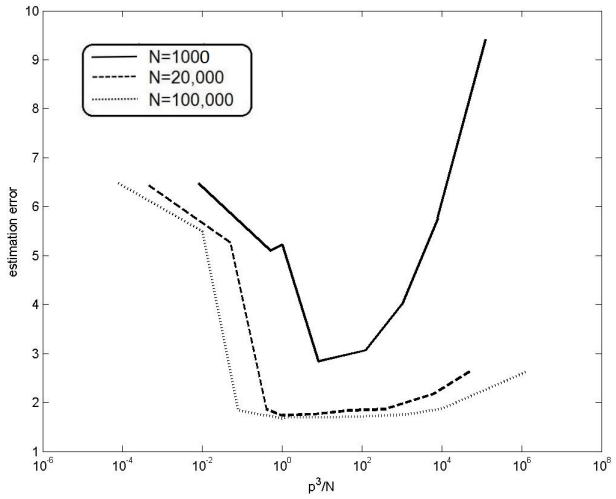

Figure 4.6: Estimation error for different sample size (N) and model order (p) for ARMA(4,4) process - Gaussian mixture innovation

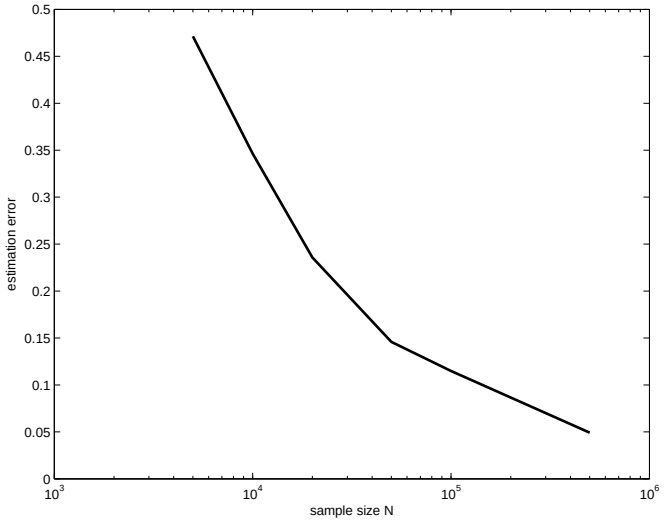

Figure 4.7: Estimation error for different sample size (N) for p=20, AR(12) process - Gaussian innovation

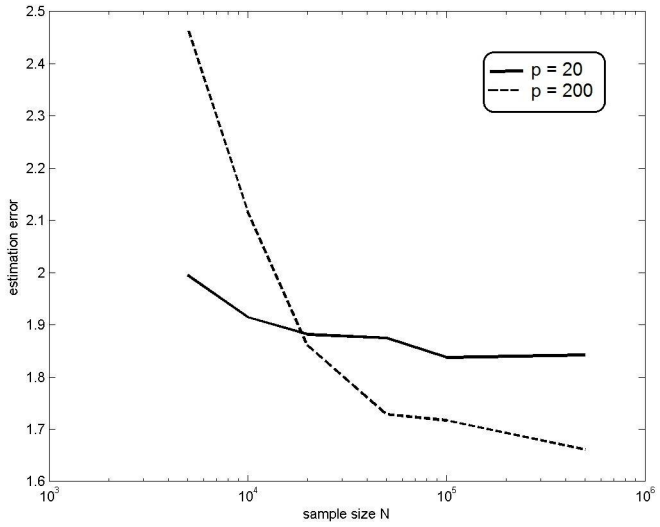

Figure 4.8: Estimation error for different sample size (N) for p=20 and p=200, ARMA(4,4) process - Gaussian innovation

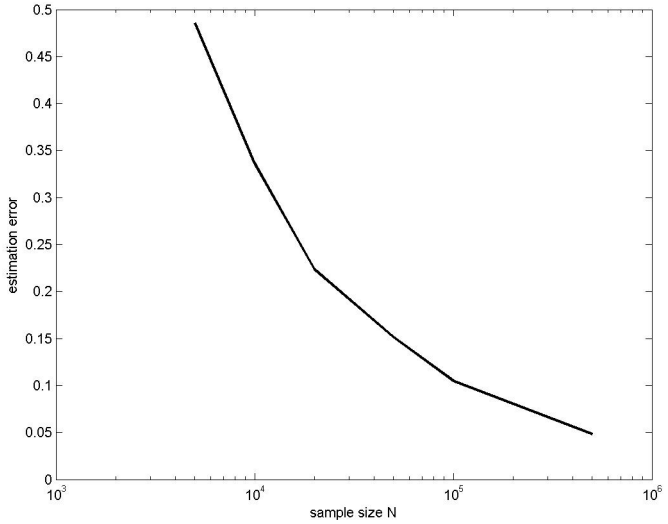

Figure 4.9: Estimation error for different sample size (N) for p=20, AR(12) process - Gaussian mixture innovation

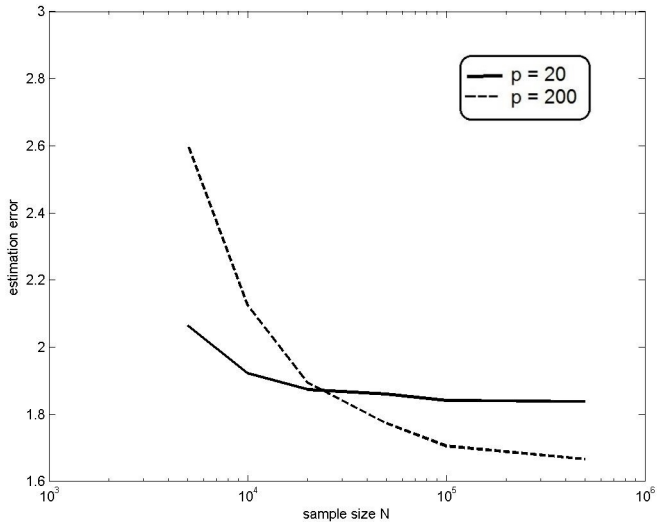

Figure 4.10: Estimation error for different sample size (N) for p=20 and p=200, ARMA(4,4) process - Gaussian mixture innovation

better estimation of the AR-p parameters. In case of the finite-order AR process, since the order of estimation p is greater than the actual order of the process, the estimation error reduces steadily and approaches zero as N is increased.

The ARMA model, on the other hand, has an infinite order AR representation and hence in its case the error can never go down to zero for a finite p; as there would always be an infinite number of higher order AR parameters that would remain undetermined. However, the error would reduce and tend to converge to a non-zero limit as N is increased while p is kept fixed. Moreover, this limiting error would be lower for a higher model order.

Simulation results for the ARMA model for both the model orders show a reduction in estimation error as N is increased. For sample sizes $5,000$ and $10,000$, the error corresponding to model order 200 is higher than that corresponding to a model order of 20. This is because $p = 200$ is too large for these values of N and the effect of bias in the estimated covariance terms is significant. For large N, the limiting error for $p = 200$ is seen to be lower than that for $p = 20$; as expected by theory.

It is observed that the plots remained almost identical when the innovation sequence was drawn from a Gaussian mixture distribution instead of a Gaussian distribution. While ARMA and AR models with Gaussian noise can be estimated through an explicit computation of the maximum likelihood function, such analysis is not feasible for processes generated from non-Gaussian type noises. The usefulness of the AR estimation method lies in its applicability across a variety of distributions, as long as the innovation sequence is uncorrelated and the resulting process is WSS.

4.5 Conclusion

In this chapter we have considered asymptotic behavior of AR approximations when the covariance sequence of the original process is unavailable and has to be estimated through observed samples. For this case, the following additional conditions are imposed: the process is assumed to be strong mixing and the corresponding innovation sequence is assumed to have a finite fourth moment. Under a mild assumption, a result on the convergence in mean square of the empirical AR parameters is derived when the model order $p = o\{N^{\frac{1}{3}}\}$. Furthermore, under the same conditions, it is shown that the spectral density of the AR approximation converges in mean with respect to an L_2 norm defined over the space of functions on $\lambda \in \left(-\frac{1}{2}, \frac{1}{2}\right]$ and that the spectral density at the origin, i.e., the time average variance constant (TAVC) of the AR approximation converges to that of an infinite order AR expansion of the process in mean.

The condition $p = o\{N^{\frac{1}{3}}\}$ is the same as the one imposed in [39] where a convergence in probability result was established for the AR parameters. However, the results of this paper are stronger than that of [39] since convergence in mean square of the AR parameters implies convergence in probability. It may be noted that Berk's result could have been used to show convergence in mean square, if we could show that the squares of the AR parameters were uniformly integrable. However, proving uniform square integrability would have been much harder, compared to the direct approach used in this thesis.

Finally, through simulation, we have studied the problem of estimating the spectral density of a WSS process through AR estimates based on a finite number of observations. For different sample sizes N and model orders p, spectral estimation of an ARMA(4,4) and an AR(12) process has been simulated and the average of the L_2 norm of the error has been plotted. As expected, for the same model order, the simulation results indicate a general improvement in estimation with an increasing sample size. It is also seen that for both the examples considered, the error reaches a minimum when p^3 is close to N.

Part II

Chapter 5

Detecting Causality in a Family of Stationary Processes: A Wiener Filter Based Approach

5.1 Introduction

Causality is the relation between two events, where the second event is seen to occur as a consequence of the first. The first event is termed as the "cause" and the second is termed as the "effect". The topic of causality has been studied and analyzed extensively in philosophy over millennia and has gained significant interest in science. In the deterministic sense, an event A is said to cause an event B when the occurrence of A is *always* followed by that of B. In contrast, A is said to *probabilistically* cause B, if (informally), the occurrence of A increases the probability of the occurrence of B.

There are various approaches that attempt to detect and quantify causality between events, random variables and stochastic processes. In this thesis, we use "Granger-causality" as a tool to determine and quantify causality among stochastic processes. In this case, mean squared estimation error is interpreted as a measure of causality. It may be noted that Granger-causality is distinct from true causality; "$\{X(t)\}$ Granger-causes $\{Y(t)\}$" does not necessarily imply that $\{X(t)\}$ *causes* $\{Y(t)\}$. Granger-causality is only an instrument used to comprehend the interplay among a number of stochastic processes; the true nature of causality is a much deeper problem.

Inferring causal dependences in a family of dynamic systems from a finite set of observations is a problem encountered in a diverse variety of fields, including economics, biology, neuroscience, meteorology and ecology. Given a set of random processes, the objective is to determine whether one process is influenced by the others, and to investigate the nature of this influence. It is customary to represent causal connections in the form of a connected graph, where the individual processes are depicted by nodes and the interdependence relations are depicted by directed edges.

Ideally, in order to infer the complete interdependence structure of a complex system, dynamic behaviour of all the processes involved should be considered simultaneously. However, for large systems, use of such a method may be infeasible. Alternatively, *pairwise* methods, i.e., methods that evaluate causal interdependence between each pair of processes, can be used to obtain sub-optimal solutions to the problem at lower computational costs. In this chapter, we investigate the problem of determining Granger-causality in an interdependent group of jointly wide sense stationary (WSS) real-valued discrete time stochastic processes by using pairwise causal Wiener filters. Through simulation examples, we compare the performance of this approach with another that uses directed information as a tool to infer causality. It is seen that a pairwise Wiener filter-based method can help obtain useful and reasonably accurate information about the causal structure of the system.

We begin by discussing the problem formally and introduce the notion of Granger-causality. This is followed by several analytical results that relate the pairwise Wiener filter to Granger-causality. Next, we propose a technique that uses pairwise finite impulse response (FIR) Wiener filters to detect causal interdependence relations. The performance of this method is compared to that of a method using directed information through a simulation example. Finally, the efficacy of the pairwise FIR Wiener filter based technique is illustrated through an example that uses real data.

5.2 Preliminaries

Consider a system of N jointly wide sense stationary (WSS) discrete time, real-valued, zero-mean, regular stochastic processes defined over a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Denote each process by $\{X_i(n)\}$ for $i = \{1, \ldots, N\}$. The index n indicates time and takes values in the set of integers. Since these processes are jointly WSS, the column vector $[X_1(n) \ \ldots \ X_N(n)]^T$ denotes an $\mathbb{R}^N$-valued multivariate discrete time WSS process. Denote this process by $\boldsymbol{X}(n)$.

The objective is to determine the inter-dependence relations among these processes and

represent these dependences in the form of a directed or undirected graph; where nodes correspond to the individual processes and edges indicate dependence relations. A directed edge in the graph essentially represents a causal filter; i.e., the existence of a directed edge from node i to node j implies that $X_j(n)$ is affected by $X_i(n-\tau)$ for some $\tau > 0$. An undirected edge, on the other hand, represents a mutual dependence relation; i.e., an undirected edge between nodes i and j implies that $X_i(n)$ and $X_j(n)$ are inter-dependent. No prior information is available on the interconnections and the topology of the system has to be estimated using only a series of observations recorded at these nodes.

In general, while investigating the inter-relations among a collection of processes, one of the four following objectives may be of interest.

1. To determine if a node i is *directly related* to node j; i.e., whether the nodes i and j are connected by an edge.
2. To ascertain the *direction* of this dependence; i.e., to determine if i causes j, if j causes i or if both of them mutually influence each other.
3. To find a *quantitative measure* of the interdependence between processes i and j. This relates to the distance between the corresponding nodes.
4. To determine *how* the processes are related; i.e., to obtain a model that clearly depicts interdependence relations and can be used to predict and estimate one process from the knowledge of the other.

In this chapter, we are primarily interested in questions 1, 2 and 3.

The following terminology is used.

1. If there is a directed edge from j to i, node j is called a "parent" of node i and node i is called a "child" of node j.
2. If there is a directed path from j to i, node j is called an "ancestor" of node i and node i is called a "descendant" of node j.
3. Processes corresponding to nodes that do not have edges entering from any other process are called "driving processes".

A simple example of three nodes is illustrated in figure 5.1, where $B_{3,1}(z)$ and $B_{3,2}(z)$ are the z-domain representation of causal linear filters, i.e.,

$$B_{3,1}(z) = \sum_{k=1}^{\infty} b_{3,1}(k) z^{-k}$$

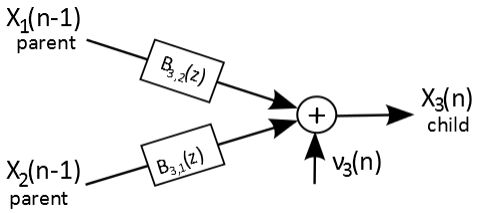

Figure 5.1: A system of three processes

$$B_{3,2}(z) = \sum_{k=1}^{\infty} b_{3,2}(k) z^{-k}$$

and $\{\nu_3(n)\}$ is a white noise sequence. The process corresponding to the child node 3 is given by

$$X_3(n) = \sum_{k=1}^{\infty} b_{3,1}(k) X_1(n-k) + \sum_{k=1}^{\infty} b_{3,2}(k) X_2(n-k) + \nu_3(n)$$

In the system represented in figure 5.2, nodes 1, 2 and 5 are driving processes. For node 7, nodes 4 and 6 are parents, nodes 1 to 6 are ancestors, nodes 8 and 9 are children and nodes 8, 9 and 10 are descendants.

The system can be formally described as follows. Let $L_2^N(\mathbb{P})$ denote the Hilbert space of $\mathbb{R}^N$-valued random vectors with finite second moment, equipped with the inner product $\mathbb{E}[\cdot,\cdot]$ defined thereon. Let $H_{\boldsymbol{X}}(n)$ denote the linear span of $\{\boldsymbol{X}(n), \boldsymbol{X}(n-1), \boldsymbol{X}(n-2), \ldots\}$, i.e., the closure of the linear combinations of $\boldsymbol{X}(n)$ and all its past values at time n. As $\{\boldsymbol{X}(n)\}$ is an $\mathbb{R}^N$-valued WSS stochastic process, $H_{\boldsymbol{X}}(n)$ is a Hilbert space [12, P-21]. For any square integrable $\mathbb{R}^N$-valued random variable $\boldsymbol{Y}$, define $\overline{\mathbb{E}}[\boldsymbol{Y}|H_{\boldsymbol{X}}(n)]$ as the projection of $\boldsymbol{Y}$ onto the space $H_{\boldsymbol{X}}(n)$. Then $\overline{\mathbb{E}}[\boldsymbol{Y}|H_{\boldsymbol{X}}(n)]$ is the minimum mean squared error (MMSE) linear estimate of $\boldsymbol{Y}$ given $H_{\boldsymbol{X}}(n)$.

By Wold decomposition theorem [12, 101], $\{\boldsymbol{X}(n)\}$ may be uniquely represented in the following form:

$$\boldsymbol{X}(n) = \sum_{k=0}^{\infty} A(k) \boldsymbol{\nu}(n-k) \tag{5.2.1}$$

where $\boldsymbol{\nu}(n) = [\nu_1(n) \ \ldots \ \nu_N(n)]^T$ is the corresponding $\mathbb{R}^N$-valued innovation process and

$$A(k) = [a_{i,j}(k)] \in \mathbb{R}^{N \times N}$$

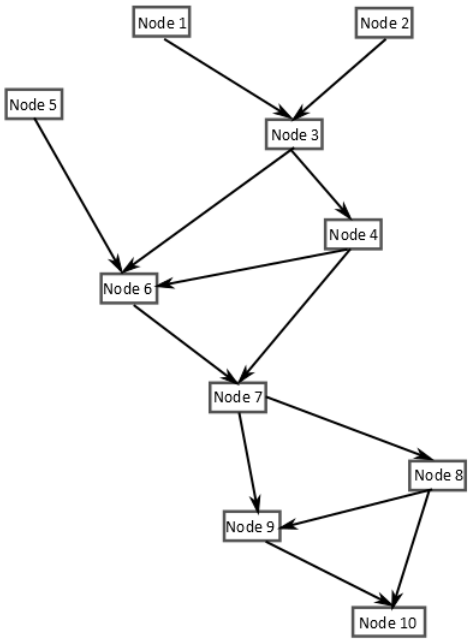

Figure 5.2: Example of a system of interdependent WSS processes

for each k. The innovations are such that

$$\mathbb{E}[\nu_i(n)\nu_i(n-k)] = 0 \text{ for all } k \neq 0$$

$$\mathbb{E}[\nu_i(n)\nu_j(n-k)] = 0 \text{ for all } i \neq j, \text{ and for all } k \in \mathbb{Z}$$

In addition, the process $\{\boldsymbol{X}(n)\}$ can also be expressed as an infinite order autoregression as follows.

$$\boldsymbol{X}(n) = \overline{\mathbb{E}}[\boldsymbol{X}(n)|H_{\boldsymbol{X}}(n-1)] + \boldsymbol{\nu}(n)$$

where $\overline{\mathbb{E}}[\boldsymbol{X}(n)|H_{\boldsymbol{X}}(n-1)]$ is essentially a weighted sum of the past values of $\boldsymbol{X}(n)$; i.e.,

$$\overline{\mathbb{E}}[\boldsymbol{X}(n)|H_{\boldsymbol{X}}(n-1)] = \sum_{k=1}^{\infty} B(k)\boldsymbol{X}(n-k) \tag{5.2.2}$$

with $B(k) = [b_{i,j}(k)] \in \mathbb{R}^{N\times N}$ for each k.

Following (5.2.2), Each scalar innovation process $\{\nu_i(n)\}$ corresponds to the real-valued process $\{X_i(n)\}$ as follows:

$$X_i(n) = \sum_{k=1}^{\infty} \sum_{j=1}^{N} b_{i,j}(k) X_j(n-k) + \nu_i(n) \tag{5.2.3}$$

and following (5.2.1), each process can also be expressed as a linear combination of the past and present innovations:

$$X_i(n) = \sum_{k=1}^{\infty} \sum_{j=1}^{N} a_{i,j}(k) \nu_j(n-k) + \nu_i(n) \tag{5.2.4}$$

The mathematical tool that quantifies causal relations among processes in the context of the proposed model is Granger-causality. Recall that, for a system of two process $\{X_i(n)\}$ and $\{X_j(n)\}$ with a model order p, Granger-causality [8] is defined as follows. Denote by $H_i^p(n-1)$ the linear span of $\{X_i(n-1), \ldots, X_i(n-p)\}$, and let $H_{i,j}^p(n-1)$ be the linear span of $\{X_i(n-1), \ldots, X_i(n-p), X_j(n-1), \ldots, X_j(n-p)\}$. Recall that $\overline{\mathbb{E}}[\cdot|\cdot]$ denotes the projection operator. Let $\xi[\cdot|\cdot]$ be the corresponding estimation error.

$X_i(n)$ is first modeled as an univariate autoregressive process using the past values of itself; i.e., as a projection on $H_i^p(n-1)$

$$X_i(n) = \overline{\mathbb{E}}[X_i(n)|H_i^p(n-1)] + \xi[X_i(n)|H_i^p(n-1)]$$

where

$$\overline{\mathbb{E}}[X_i(n)|H_i^p(n-1)] = \sum_{k=1}^{p} \alpha_{i,i}(k) X_i(n-k)$$

and then modeled as an autoregression that also includes past observations of $\{X_j(n)\}$

$$X_i(n) = \overline{\mathbb{E}}[X_i(n)|H_{i,j}^p(n-1)] + \xi[X_i(n)|H_{i,j}^p(n-1)]$$

where

$$\overline{\mathbb{E}}[X_i(n)|H_{i,j}^p(n-1)] = \sum_{k=1}^{p} \beta_{i,i}(k) X_i(n-k) + \sum_{k=1}^{p} \beta_{i,j}(k) X_j(n-k)$$

$\{X_j(n)\}$ is said to Granger-cause $\{X_i(n)\}$ if the mean squared error in case of the latter is strictly less than that for the former.

$$\mathbb{E}\left[(\xi[X_i(n)|H_i^p(n-1)])^2\right] > \mathbb{E}\left[(\xi[X_i(n)|H_{i,j}^p(n-1)])^2\right]$$

When the system is causally influenced by an infinite past, the model order p is allowed to approach ∞. For the rest of this dissertation, we assume the processes to be dependent on an infinite past. In other words, the system is allowed to have infinite memory.

When there is more than one process involved, one has to take into account all the processes simultaneously to determine Granger-causality. Given a family of N processes $\{X_k(n)\}$, $k \in \{1, \ldots, N\}$, $\{X_i(n)\}$ Granger-causes $\{X_j(n)\}$ $(j \neq i)$ if the mean squared error in estimating $X_j(n)$ from the past values of all the processes excluding those of $\{X_i(n)\}$ is greater than that in estimating $X_j(n)$ from the past values of all the processes, including those of $\{X_i(n)\}$.

The above definition of Granger-causality can be extended beyond linear estimates as follows. Let $\mathcal{F}(n)$ be the sigma-algebra generated by all the processes $\{X_k(n)\}$, $k = 1, \ldots, N$, up to time n, and let $\mathcal{F}_{-j}(n)$ be the sigma-algebra generated by the processes $\{X_k(n)\}_{k \neq j}$ up to time n. Clearly, $\mathcal{F}_{-j}(n) \subset \mathcal{F}(n) \subset \mathcal{F}$. $\{X_j(n)\}$ is said to Granger-cause $\{X_i(n)\}$, if

$$\mathbb{P}(X_i(n) \in A|\mathcal{F}(n-1)) \neq \mathbb{P}(X_i(n) \in A|\mathcal{F}_{-j}(n-1))$$

where A is any Borel subset of $\mathbb{R}$. Essentially, this means that the past of $\{X_j(n)\}$ carries additional information on the present of $\{X_i(n)\}$, not included in the past of the other processes. While the second definition is more general, it is not very easy to use in practical applications, as it requires the knowledge of distributions. Since we are interested in linear MMSE estimates in this thesis, we would be using the first definition, that defines Granger-causality in terms of mean squared errors and autoregressive parameters.

Determining causality in this way will necessitate the determination of the autoregressive parameters through the method of least squares. However, for a family of WSS processes having infinite memory, the problem of determining the parameters is non-trivial and involves a significant level of computation. If the model is simplified by assuming that the system has a finite memory; i.e., instead of an infinite weighted sum of past observations, the estimate depends on the p most recent values; even then for each i, there are $p(N-1)$ and pN parameters to be derived. For large N and p this becomes computationally burdensome.

The same problem may also be addressed from an information theoretic perspective. Recall that for two random variables X and Y, the mutual information is defined as

$$I(X;Y) = \int_Y \int_X p(x,y) log\left(\frac{p(x,y)}{p(x)p(y)}\right) dxdy$$

where $p(x)$, $p(y)$ are the respective marginal probability distributions of X and Y, and $p(x,y)$ is their joint probability distribution. The conditional mutual information of two random variables X and Y, conditioned on another random variable Z, is the expected value of the mutual information of X and Y, given Z. It is defined as

$$I(X;Y|Z) = \int_Z \int_Y \int_X p(x,y,z) log\left(\frac{p(x,y,z)p(z)}{p(x,z)p(y,z)}\right) dxdydz$$

where $p(x,y,z)$ denotes the joint probability distribution function of X, Y and Z; and $p(x,z)$ and $p(y,z)$ are the joint distributions of X, Z and Y, Z respectively. Finally, for a number of random variables $\{X_i\}_{i=1,\ldots,N}$, the mutual information is defined using the following recursion.

$$I(X_1;X_2;\ldots;X_N) = I(X_1;X_2;\ldots;X_{N-1}) - I(X_1;X_2;\ldots;X_{N-1}|X_N)$$

The mutual information of two random variables (or two random vectors) is a symmetric quantity, devoid of any directional aspect. However, while the notion of causality is not inherent in mutual information, it may be introduced through directed information. For two $\mathbb{R}^t$-valued random vectors $\boldsymbol{X^t}$, $\boldsymbol{Y^t}$; $\boldsymbol{X^t} = [X(1)\ \ldots\ X(t)]^T$, $\boldsymbol{Y^t} = [Y(1)\ \ldots\ Y(t)]^T$; the directed information from $\boldsymbol{X^t}$ to $\boldsymbol{Y^t}$ is given by [83, 53]:

$$I(\boldsymbol{X^t} \to \boldsymbol{Y^t}) = \sum_{i=1}^{t} I(\boldsymbol{X}^i; Y(i)|\boldsymbol{Y}^{i-1})$$

where $\boldsymbol{X}^i = [X(1)\ \ldots\ X(i)]$.

For two stochastic processes, $\{X(n)\}$ and $\{Y(n)\}$, the directed information rate is given by

$$I(X \to Y) = \lim_{t\to\infty} \frac{1}{t} I(\boldsymbol{X^t} \to \boldsymbol{Y^t})$$

when the limit exists [81], where $\boldsymbol{X^t}$, $\boldsymbol{Y^t}$ are random vectors as defined above. When both the processes are stationary and ergodic, the limit is well-defined [81, 112].

We conclude this section with a result that relates Granger-causality and multivariate AR models. Let $H_i(n)$ denote the linear span of $\{X_i(n), X_i(n-1), X_i(n-2), \ldots\}$ and let $H(n)$ denote the linear span of the present and past values of all the processes $\{X_k(n)\}$; $k = 1,\ldots,N$. Let $H_{-j}(n)$ denote the same for all the processes except $\{X_j(n)\}$. In each case, we also include the limits in the mean square of the sums when they exist. By construction, $H_i(n)$, $H(n)$ and $H_{-j}(n)$ are Hilbert spaces; equipped with the inner-product $\mathbb{E}[\cdot,\cdot]$. Recall that $\overline{\mathbb{E}}[\cdot|\cdot]$ denotes the projection operator and $\xi[\cdot|\cdot]$ denotes the corresponding estimation error.

For any process $\{X_i(n)\}$, let $P_i \subset \{1, \ldots, N\}$ denote the set of indices corresponding to its parents, and also include the process $\{X_i(n)\}$ itself. Following (5.2.3), for all $l \in P_i$, there exists at least one $k \geq 1$ such that $b_{i,l}(k) \neq 0$. Let $H_{P_i}(n)$ denote the linear span of all the processes $\{X_l(n)\}_{l \in P_i}$ and their past values. Then, (5.2.3) can be expressed as:

$$X_i(n) = \sum_{k=1}^{\infty} \sum_{l \in P_i} b_{i,l}(k) X_l(n-k) + \nu_i(n) \tag{5.2.5}$$

Lemma 5.2.1. *$\{X_j(n)\}$ Granger-causes $\{X_i(n)\}$ if and only if $j \in P_i$.*

Proof. The proof uses the fact that the projection of an element of a Hilbert space on a closed Hilbert subspace is unique. Let $j \in P_i$ and $\{X_j(n)\}$ does not Granger-cause $\{X_i(n)\}$. Then,

$$\mathbb{E}\left[(\xi[X_i(n)|H(n-1)])^2\right] = \mathbb{E}[\nu_i^2(n)]$$

and in this case,

$$\overline{\mathbb{E}}[X_i(n)|H(n-1)] = \overline{\mathbb{E}}[X_i(n)|H_{P_i}(n-1)]$$

We have,

$$X_i(n) = \sum_{k=1}^{\infty} \sum_{l \in P_i} b_{i,l}(k) X_l(n-k) + \nu_i(n)$$

Let $\xi[X_i(n)|H_{-j}(n-1)] = \nu_{i,-j}(n)$. Since $H_{-j}(n-1) \subset H(n-1)$, $\overline{\mathbb{E}}[X_i(n)|H_{-j}(n-1)]$ is an element in $H(n-1)$, and therefore,

$$\mathbb{E}[\nu_{i,-j}^2(n)] \geq \mathbb{E}[\nu_i^2(n)]$$

On the other hand, since by our assumption $\{X_j(n)\}$ does not Granger-cause $\{X_i(n)\}$,

$$\mathbb{E}[\nu_{i,-j}^2(n)] = \mathbb{E}[\nu_i^2(n)] \tag{5.2.6}$$

But if the above holds, then there are *two* elements in $H(n-1)$ that achieve the minimum distance from $X(n)$, namely, $\overline{\mathbb{E}}[X_i(n)|H_{-j}(n-1)]$ and $\overline{\mathbb{E}}[X_i(n)|H(n-1)]$. In other words, $X(n)$ has two projections on $H(n-1)$. From the uniqueness of the projection operation, this is a contradiction and (5.2.6) can only hold when the parameters $\{b_{i,j}(k)\}$ are zero for all k, in which case $j \notin P_i$. Therefore, when $j \in P_i$, (5.2.6) does not hold and $\{X_j(n)\}$ Granger-causes $\{X_i(n)\}$.

To prove the reverse, recall that the process $\{X_j(n)\}$ is said to Granger-cause $\{X_i(n)\}$ if and only if

$$\mathbb{E}\left[(\xi[X_i(n)|H(n-1)])^2\right] < \mathbb{E}\left[(\xi[X_i(n)|H_{-j}(n-1)])^2\right]$$

i.e., the exclusion of information on the past history of $\{X_j(n)\}$ strictly increases the error in estimating $X(n)$.

Clearly then, since the representation in (5.2.5) is unique, it immediately follows that if $\{X_j(n)\}$ Granger-causes $\{X_i(n)\}$; past values of $\{X_j(n)\}$ will appear in the MMSE estimate of $X_i(n)$. In other words, $X_i(n)$ must have an expansion of the form (5.2.5); i.e., $j \in P_i$.

This completes the proof. Note that both results hold when the terms involving $\{X_i(n)\}$ are absent on the right hand side of (5.2.5).

□

Corollary 5.2.1. *For a system of N processes, Granger-causality can be completely determined by implementing an N dimensional Multivariate autoregressive (MVAR) model.*

The result follows directly from the above lemma. An MVAR model solves for each parameter $\{b_{i,j}(k)\}$ in (5.2.2). For the processes j not in P_i, $b_{i,j}(k)$ will be zero for all k.

For a system of many interdependent processes, ideally, causal links should be detected following corollary 5.2.1, where the parameters are computed using covariance and cross-covariance sequences of the processes. However, for a large system, implementation of such a method would be computationally challenging. Moreover, since one has to work with *estimates* of covariances in lieu of the exact quantities; false edge detection is likely.

As an example, we simulate a system of 10 jointly WSS time series having causal connections as depicted in figure 5.2, with arbitrarily chosen parameters and standard normal innovation. MVAR parameters are computed using 1,000,000 sample realizations. The computed parameters below a certain threshold are set to zero. The configuration of the system thus obtained is presented in figure 5.3.

While most of the edges of the original system are identified, some additional edges are falsely detected as well. This results due to the deviation of the estimates of cross-covariance terms from their true values. The results are also sensitive to the selection of the model order p, and closer p is to the actual model order, more accurate are the results.

Instead of considering all the processes together; one may attempt to infer the underlying structure by observing the pairwise dynamics of processes. In other words, it is simply

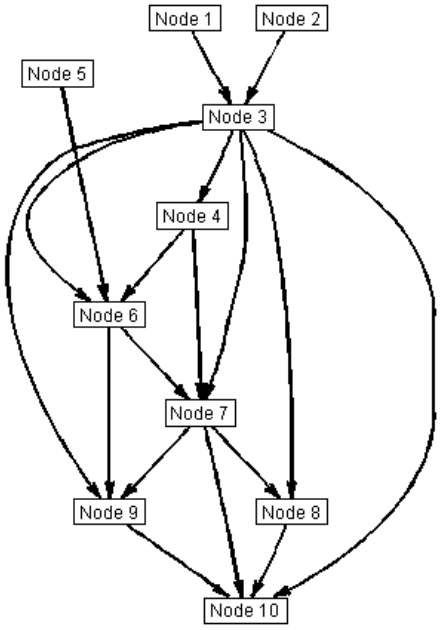

Figure 5.3: Granger-causality inferred through MVAR approach

investigated whether there is a causal link from $\{X_i(n)\}$ to $\{X_j(n)\}$, ignoring the dynamics of the other processes. This is repeated for each pair of processes and tested for both directions of causality, and the results are then used to determine the general structure of the system.

For a system of N processes and a model order p, a pairwise approach requires solving N^2 systems of linear equations, each with p unknowns. Noting that solving each such system requires $O(p^3)$ operations, this method requires $O(N^2p^3)$ operations in all. In comparison, the MVAR approach that directly solves for the Np least square parameters for each of the N individual processes, involves $O(N^4p^3)$ operations. Thus, even though a pairwise method may not reveal all causal branches, it can nonetheless provide a means to obtain a quick and easy insight into the system.

In the following sections, we investigate the efficacy of *pairwise* causal Wiener filters for this purpose. Some theoretical results are presented on the applicability and limitations of such an approach. This is followed by simulation examples where the performance of pairwise FIR Wiener filters in detecting causal connections is compared to that of directed information.

5.3 Results on a pairwise Wiener filter based approach

The following assumptions are made on the system.

1. There is no instantaneous causality; i.e., there is no pair of nodes $\{i, j\}$ such that $X_i(n)$ and $X_j(n)$ affect each other.
2. There are no closed loops or directed cycles in the graph.

The first assumption is inherent in the model formulation of 5.2.2. While the first assumption asserts that there is no undirected edge (representing instantaneous mutual causation) in the system, it does not eliminate the possibility of **bidirectional** edges between two nodes, in the case where present values of both nodes are causally affected by the *past values* of each other. The existence of such edges is precluded by the second assumption. In the rest of this section, we present a few theoretical results on the pairwise causal Wiener filter.

Theorem 5.3.1. *Let $\{X_1(n)\}$ and $\{X_2(n)\}$ be jointly WSS and let the error in estimating $X_2(n)$ from $X_1(n-1)$ through a causal Wiener filter be $\xi[X_2(n)|H_1(n-1)]$. If $\{X_1(n)\}$ Granger-causes $\{X_2(n)\}$, then*

$$\mathbb{E}[(\xi[X_2(n)|H_1(n-1)])^2] < \mathbb{E}[X_2^2(n)]$$

Proof. Note that by definition,

$$\mathbb{E}[(\xi[X_2(n)|H_1(n-1)])^2] \leq \mathbb{E}[X_2^2(n)]$$

We have to show that the above inequality is strict. Now, the equality can hold only if

$$\mathbb{E}[(\overline{\mathbb{E}}[X_2(n)|H_1(n-1)])^2] = 0$$

Since $\overline{\mathbb{E}}[X_2(n)|H_1(n-1)]$ is the projection of $X_2(n)$ on $H_1(n-1)$, this can only happen if $X_2(n)$ is orthogonal to the subspace $H_1(n-1)$, i.e., $\mathbb{E}[X_2(n)X_1(n-l)] = 0$ for all $l > 1$. We shall show that this cannot hold.

Since the system consists of only two processes,

$$\xi[X_2(n)|H_{1,2}(n-1)] = \xi[X_2(n)|H(n-1)] = \nu_2(n)$$

Recall that the causal Wiener filter from $X_1(n-1)$ to $X_2(n)$ is the linear minimum mean squared error (MMSE) estimator of the latter based on $X_1(n-1)$ and its past values.

$$X_2(n) = \sum_{k=1}^{\infty} w_{2,1}(k) X_1(n-k) + \xi[X_2(n)|H_1(n-1)]$$

Since $\{X_1(n)\}$ Granger-cause $\{X_2(n)\}$, due to lemma 5.2.1, $X_2(n)$ can be represented as

$$X_2(n) = \sum_{k=1}^{\infty} b_{2,2}(k) X_2(n-k) + \sum_{k=1}^{\infty} b_{2,1}(k) X_1(n-k) + \nu_2(n)$$

Taking z-transforms on both sides, and rearranging, we obtain

$$\left(1 - \sum_{k=1}^{\infty} b_{2,2}(k) z^{-k}\right) X_2(z) = \left(\sum_{k=1}^{\infty} b_{2,1}(k) z^{-k}\right) X_1(z) + N_2(z)$$

Assuming $\left(1 - \sum_{k=1}^{\infty} b_{2,2}(k) z^{-k}\right)$ to be invertible, $X_2(n)$ can be expressed as an autoregression of $\{X_1(n)\}$ as follows

$$X_2(n) = \sum_{k=1}^{\infty} c_{2,1}(k) X_1(n-k) + \xi_{2|1}(n)$$

where the parameters $\{c_{2,1}(k)\}$ are derived from

$$\sum_{k=1}^{\infty} c_{2,1}(k) z^{-k} = \left(1 - \sum_{k=1}^{\infty} b_{2,2}(k) z^{-k}\right)^{-1} \left(\sum_{k=1}^{\infty} b_{2,1}(k) z^{-k}\right)$$

and the error $\xi_{2|1}(n)$ is a linear combination of $\{\nu_2(n)\}$, with the z-transform

$$\xi_{2|1}(z) = \left(1 - \sum_{k=1}^{\infty} b_{2,2}(k) z^{-k}\right)^{-1} N_2(z)$$

In general, $c_{2,1}(k) \neq 0$, and therefore, for $l > 1$,

$$\mathbb{E}[X_2(n) X_1(n-l)] \neq 0$$

Clearly, then, $X_2(n)$ is not orthogonal to $H_1(n-1)$, i.e.,

$$\mathbb{E}[(\xi[X_2(n)|H_1(n-1))^2] < \mathbb{E}[X_2^2(n)]$$

□

This result relates the pairwise causal Wiener filter and the notion of Granger-causality. It is worth noting here that the reverse is not necessarily true. Consider, for instance, the case where

$$X_1(n) = \alpha X_2(n-1) + \nu_1(n)$$
$$X_2(n) = \beta X_2(n-1) + \nu_2(n)$$

Here, $\mathbb{E}[(\xi[X_2(n)|H_1(n-1))^2] < \mathbb{E}[X_2^2(n)]$, and yet $\{X_2(n)\}$ is not Granger-caused by $\{X_1(n)\}$.

Proposition 5.3.1. *If two processes $\{X_1(n)\}$ and $\{X_2(n)\}$ constitute a system such that $\{X_1(n)\}$ Granger-causes $\{X_2(n)\}$ and $\{X_2(n)\}$ does not Granger-cause $\{X_1(n)\}$ then, in general,*

$$\mathbb{E}\left[(\xi[X_1(n)|H_1(n-1)])^2\right] < \mathbb{E}\left[(\xi[X_1(n)|H_2(n-1)])^2\right]$$

Proof. Let the processes be defined as follow.

$$\begin{bmatrix} X_1(n) \\ X_2(n) \end{bmatrix} = \sum_{k=1}^{\infty} \begin{bmatrix} b_{1,1}(k) & 0 \\ b_{2,1}(k) & b_{2,2}(k) \end{bmatrix} \begin{bmatrix} X_1(n-k) \\ X_2(n-k) \end{bmatrix} + \begin{bmatrix} \nu_1(n) \\ \nu_2(n) \end{bmatrix}$$

It immediately follows that

$$\mathbb{E}\left[(\xi[X_1(n)|H_1(n-1)])^2\right] = \mathbb{E}[\nu_1^2(n)]$$

On the other hand, we have

$$X_2(n) = \sum_{k=1}^{\infty} b_{2,1}(k)X_1(n-k) + \sum_{k=1}^{\infty} b_{2,2}(k)X_2(n-k) + \nu_2(n)$$

Proceeding through the same steps as in theorem 5.3.1, we obtain

$$X_2(n) = \sum_{k=1}^{\infty} c_{2,1}(k)X_1(n-k) + \xi_{2|1}(n) \tag{5.3.7}$$

$X_1(n)$ can be expressed as

$$\begin{aligned} X_1(n) &= \sum_{k=2}^{\infty} b_{1,1}(k) X_1(n-k) + b_{1,1}(1) \left(\sum_{j=1}^{\infty} b_{1,1}(j) X_1(n-j-1) + \nu_1(n-1) \right) + \nu_1(n) \\ &= \sum_{k=2}^{\infty} d_{1,1}(k) X_1(n-k) + b_{1,1}(1) \nu_1(n-1) + \nu_1(n) \end{aligned}$$

where the parameters $\{d_{1,1}(k)\}$ are functions of the parameters $\{b_{1,1}(k)\}$. Note that by construction, both $\nu_1(n)$ and $\nu_1(n-1)$ are orthogonal to $H_2(n-1)$ and hence

$$\mathbb{E}\left[(\xi[X_1(n)|H_2(n-1)])^2\right] \geq \mathbb{E}\left[(\xi[X_1(n)|H_1(n-1)])^2\right] \tag{5.3.8}$$

where the equality holds if $b_{1,1}(1) = 0$ and

$$\overline{\mathbb{E}}[X_1(n)|H_2(n-1)] = \sum_{k=2}^{\infty} d_{1,1}(k) X_1(n-k)$$

For the above to hold there must exist parameters $\{f_{1,2}\}$ such that

$$\sum_{k=1}^{\infty} f_{1,2}(k) X_2(n-k) = \sum_{k=2}^{\infty} d_{1,1}(k) X_1(n-k)$$

or, using (5.3.7),

$$\sum_{k=1}^{\infty} f_{1,2}(k) \left(\sum_{j=1}^{\infty} c_{2,1}(j) X_1(n-k-j) + \xi_{2|1}(n-k) \right) = \sum_{k=2}^{\infty} d_{1,1}(k) X_1(n-k)$$

The sequence $\{\xi_{2|1}(n)\}$ is by construction, a linear combination of the innovation $\{\nu_2(n)\}$ and hence orthogonal to the process $\{X_1(n)\}$. Therefore, the above equality can only hold if

$$\sum_{k=1}^{\infty} f_{1,2}(k) \xi_{2|1}(n-k) = 0$$

However, each of the terms $\xi_{2|1}(n-k)$ is a linear combination of $\{\nu_2(n)\}$, an orthogonal sequence, and therefore the above, in general, will not be zero. Therefore, the inequality in (5.3.8) is strict and the result is proved.

□

In order to infer interdependence between two processes using a causal Wiener filter approach, $X_i(n)$ is estimated as $\overline{\mathbb{E}}[X_i(n)|H_j(n-1)]$. As long as $X_i(n)$ is not orthogonal to $H_j(n-1)$, this estimate will be non-zero and the mean squared error $\mathbb{E}\left[(\xi[X_i(n)|H_j(n-1)])^2\right]$ will be strictly less than $\mathbb{E}[X_i^2(n)]$. Pairwise causality can then be inferred through an analysis of the mean squared errors $\mathbb{E}\left[(\xi[X_i(n)|H_j(n-1)])^2\right]$. Here, we present three results on a causal Wiener filter approach to a system consisting of more than two jointly WSS processes.

Proposition 5.3.2. *If* $\mathbb{E}\left[(\xi[X_i(n)|H_j(n-1)])^2\right] < \mathbb{E}[X_i(n)^2]$, *then at least one of the following must be true:*

1. $\{X_i(n)\}$ *is an ancestor of* $\{X_j(n)\}$.
2. $\{X_j(n)\}$ *is an ancestor of* $\{X_i(n)\}$.
3. $\{X_i(n)\}$, $\{X_j(n)\}$ *have a common ancestor.*

Proof. For any two processes $\{X_i(n)\}$ and $\{X_j(n)\}$, the following always holds.

$$\mathbb{E}\left[(\xi[X_i(n)|H_j(n-1)])^2\right] \leq \mathbb{E}[X_i(n)^2]$$

Let the processes at nodes i, j be such that they do not satisfy any of the three criteria. Let A_i be the set of indices corresponding to the ancestor processes associated with $\{X_i(n)\}$ and include the index i. Let A_j be the set of indices corresponding to the ancestor processes associated with $\{X_j(n)\}$ and include the index j. The processes $\{X_i(n)\}$ and $\{X_j(n)\}$ can then be expressed as:

$$X_i(n) = \sum_{l \in A_i} \sum_{k=1}^{\infty} c_l(k)\nu_l(n-k) + \nu_i(n)$$

$$X_j(n) = \sum_{l \in A_j} \sum_{k=1}^{\infty} c'_l(k)\nu_l(n-k) + \nu_j(n)$$

Such an expression can be arrived at by a step-by-step expansion of the ancestor processes associated with $\{X_i(n)\}$ and $\{X_j(n)\}$. Since nodes i and j have neither a common ancestor nor an ancestor-descendant relation; A_i and A_j are mutually exclusive. Since the individual $\{\nu_l(n)\}$ are innovations, they are orthogonal to each other. Therefore, the terms $\{\nu_l(n-k)\}_{l \in A_i}$ and $\{\nu_l(n-k)\}_{l \in A_j}$ are uncorrelated.

Therefore, for any $\tau \in \mathbb{N}$,

$$\mathbb{E}[X_i(n)X_j(n-\tau)] = 0$$

Thus, $X_i(n)$ is orthogonal to $H_j(n-1)$ (likewise, $X_j(n)$ is orthogonal to $H_i(n-1)$) and hence

$$\mathbb{E}\left[(\xi[X_i(n)|H_j(n-1)])^2\right] = \mathbb{E}[X_i^2(n)]$$

Therefore for $\mathbb{E}\left[(\xi[X_i(n)|H_j(n-1)])^2\right] < \mathbb{E}[X_i(n)^2]$ to hold, at least one of the three criteria must be satisfied.

□

Proposition 5.3.3. *If node j is an ancestor of node i, then*

$$\mathbb{E}\left[(\xi[X_i(n)|H_j(n-1)])^2\right] < \mathbb{E}[X_i^2(n)]$$

Proof. Let j be an ancestor of i. Then, $X_i(n)$ can be expressed as

$$X_i(n) = \psi_j(n-\tau) + \phi_{i,j}(n)$$

where $\psi_j(n-\tau) \in H_j(n-\tau)$ for some $\tau \in \mathbb{N}$ and $\phi_{i,j}(n)$ is orthogonal to $H_j(n-\tau)$. Noting that $H_j(n-\tau) \subset H_j(n-1)$, it is easy to see that

$$\mathbb{E}\left[(\overline{\mathbb{E}}[X_i(n)|H_j(n-1)])^2\right] \geq \mathbb{E}[\psi_j^2(n-\tau)] > 0$$

and therefore,

$$\begin{aligned}\mathbb{E}\left[(\xi[X_i(n)|H_j(n-1)])^2\right] &= \mathbb{E}[X_i^2(n)] - \mathbb{E}\left[(\overline{\mathbb{E}}[X_i(n)|H_j(n-1)])^2\right] \\ &< \mathbb{E}[X_i^2(n)]\end{aligned}$$

□

Proposition 5.3.4. *Let the processes $\{X_i(n)\}$ and $\{X_j(n)\}$ be such that*

1. *There is no directed path between the two and the process $\{X_k(n)\}$ is their only common ancestor.*

   Or,

2. *The two have no common ancestor and the process* $\{X_k(n)\}$ *is the only intermediate node in a directed path between the two that extends (without loss of generality) from* j *to* i*.*

Then,

$$\mathbb{E}\left[(\xi[X_i(n)|H_j(n-1)])^2\right] \geq \mathbb{E}\left[(\xi[X_i(n)|H_k(n-1)])^2\right]$$
$$\mathbb{E}\left[(\xi[X_j(n)|H_i(n-1)])^2\right] \geq \mathbb{E}\left[(\xi[X_j(n)|H_k(n-1)])^2\right]$$

The two causal structures mentioned are depicted in figure 5.4.

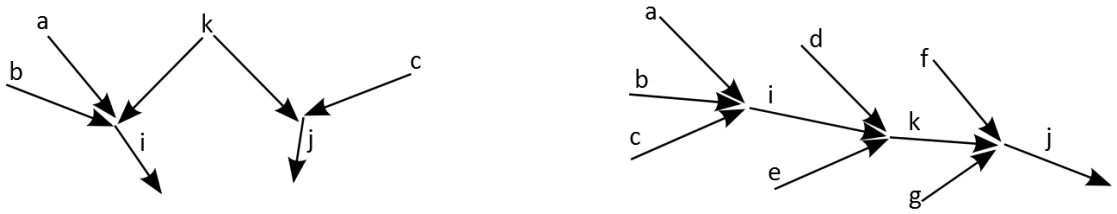

Figure 5.4: Causal structures corresponding to proposition 5.3.4

Proof. Consider the case when $\{X_k(n)\}$ is a common ancestor. Note that

$$X_i(n) = \overline{\mathbb{E}}[X_i(n)|H_k(n-1)] + \xi[X_i(n)|H_k(n-1)]$$

The error $\xi[X_i(n)|H_k(n-1)]$ is the part in $X_i(n)$ that is orthogonal to $H_k(n-1)$. This can be expressed as a weighted sum of innovations $\{\nu_i(n)\}_{i\in A_{i/k}}$ corresponding to the processes $\{X_i(n)\}_{i\in A_{i/k}}$. The set of indices $A_{i/k}$ includes the process i itself and all ancestors of $\{X_i(n)\}$, excluding k.

Since there is no directed path going from $\{X_j(n)\}$ to $\{X_i(n)\}$, the corresponding innovation $\{\nu_j(n)\}$ is absent in $\xi[X_i(n)|H_k(n-1)]$. Since there is no directed path from $\{X_i(n)\}$ to $\{X_j(n)\}$, the innovation $\{\nu_i(n)\}$ and $\{X_j(n)\}$ are orthogonal to each other. Finally, since $\{X_j(n)\}$ has no common ancestor with $\{X_i(n)\}$ apart from $\{X_k(n)\}$, the innovations $\{\nu_i(n)\}_{i\in A_{i/k}}$ are not present in $X_j(n)$ and are therefore orthogonal to $\{X_j(n)\}$. Combining, it is seen that each of the individual innovations that constitute $\xi[X_i(n)|H_k(n-1)]$ are orthogonal to the process $\{X_j(n)\}$ and its past values, and consequently to the subspace $H_j(n-1)$. Therefore, $\xi[X_i(n)|H_k(n-1)]$ is orthogonal to $H_j(n-1)$ too. The

error in estimating $X_i(n)$ from $H_j(n-1)$ will include the quantity $\xi[X_i(n)|H_k(n-1)]$ (and possibly some additional terms consisting of innovations not included in $A_{i/k}$. Therefore,

$$\mathbb{E}\left[(\xi[X_i(n)|H_j(n-1)])^2\right] \geq \mathbb{E}\left[(\xi[X_i(n)|H_k(n-1)])^2\right]$$

Now consider the case when, without loss of generality, there is a directed path from $\{X_j(n)\}$ to $\{X_i(n)\}$ and $\{X_k(n)\}$ is the only intermediate node. As the nodes have no common ancestor, it immediately follows that the innovations that constitute $\xi[X_i(n)|H_k(n-1)]$ come from processes other than $\{X_j(n)\}$. None of these processes is an ancestor to $\{X_j(n)\}$, neither is $\{X_j(n)\}$ an ancestor to any of them. Therefore, these processes are all orthogonal to the process $\{X_j(n)\}$ and hence to $H_j(n-1)$. Therefore,

$$\mathbb{E}\left[(\xi[X_i(n)|H_j(n-1)])^2\right] \geq \mathbb{E}\left[(\xi[X_i(n)|H_k(n-1)])^2\right]$$

□

While a causal Wiener filter has the notion of direction inherent in itself, that is not the case with the non-causal Wiener filter. However, the latter, too, may be used in conjunction with the former to infer information on causal links. The following result is related to this.

Proposition 5.3.5. *Let $G_{i|j}(\lambda)$ be the frequency response of the non-causal Wiener filter that estimates process $\{X_i(n)\}$ from $\{X_j(n)\}$. If $G_{i|j}(\lambda)$ is causal (i.e., devoid of anti-causal terms), then $\{X_i(n)\}$ does not cause $\{X_j(n)\}$.*

Proof. It is known that the frequency response of the non-causal Wiener filter that estimates process $\{X_i(n)\}$ from $\{X_j(n)\}$ is given by

$$G_{i|j}(\lambda) = \frac{S_{i,j}(\lambda)}{S_j(\lambda)}$$

where $S_j(\lambda)$ is the power spectral density of $\{X_j(n)\}$ and $S_{i,j}(\lambda)$ is the cross power spectral density of $\{X_i(n)\}$ and $\{X_j(n)\}$.

$$S_j(\lambda) = \sum_{k=-\infty}^{\infty} \mathbb{E}[X_j(n)X_j(n-k)]e^{-2\pi\imath\lambda k}$$

$$S_{i,j}(\lambda) = \sum_{k=-\infty}^{\infty} \mathbb{E}[X_i(n)X_j(n-k)]e^{-2\pi\imath\lambda k}$$

Consider processes $\{X_i(n)\}$ and $\{X_j(n)\}$ that are related as follows.

$$X_i(n) = \sum_{k=1}^{\infty} \beta(k) X_j(n-k) + \xi_{i|j}(n)$$

where $\sum_{k=1}^{\infty} \beta(k) X_j(n-k) = \overline{\mathbb{E}}[X_i(n)|H_j(n-1)]$ and the error $\xi_{i|j}(n)$ is such that $\mathbb{E}[X_j(n-k)\xi_{i|j}(n)] = 0$ for all $k > 0$. Observe that, by construction, $\xi_{i|j}(n)$ is a weighted sum of innovations that includes $\nu_i(n)$, since $\nu_i(n) \notin H_j(n-1)$.

The cross power spectral density $S_{i,j}(\lambda)$, then, is

$$\begin{aligned} S_{i,j}(\lambda) &= \sum_{t=-\infty}^{\infty} \mathbb{E}[X_i(n)X_j(n-t)]e^{-2\pi\imath\lambda t} \\ &= \sum_{t=-\infty}^{\infty} \left(\sum_{k=1}^{\infty} \beta(k) \mathbb{E}\left[X_j(n-k)X_j(n-t)\right] + \mathbb{E}\left[X_j(n-t)\xi_{i|j}(n)\right] \right) e^{-2\pi\imath\lambda t} \\ &= B(\lambda)S_j(\lambda) + S_{\xi,j}(\lambda) \end{aligned}$$

where $B(\lambda)$ is given by

$$B(\lambda) = \sum_{k=1}^{\infty} \beta(k) e^{-2\pi\imath\lambda k}$$

and $S_{\xi,j}(\lambda)$ is the cross power spectral density of $\{\xi_{i,j}(n)\}$ and $\{X_j(n)\}$, given by

$$S_{\xi,j}(\lambda) = \sum_{t=-\infty}^{\infty} \mathbb{E}[X_j(n-t)\xi_{i|j}(n)]e^{-2\pi\imath\lambda t}$$

Then,

$$G_{i|j}(\lambda) = B(\lambda) + \frac{S_{\xi,j}(\lambda)}{S_j(\lambda)}$$

$G_{i|j}(\lambda)$ represents a causal filter if it does not contain negative powers of $e^{-2\pi\imath\lambda}$. By definition, $B(\lambda)$ is a causal filter. $S_j(\lambda)$, the spectral density of $\{X_j(n)\}$, has both causal and anti-causal terms. Finally, by construction, $\mathbb{E}[X_j(n-t)\xi_{i|j}(n)]$ can be non-zero only for $t \leq 0$; and consequently, if $S_{\xi,j}(\lambda)$ is not zero it must be anti-causal. Therefore, if $G_{i|j}(\lambda)$ is a causal filter, then

1. $\mathbb{E}[X_j(n+k)\xi_{i|j}(n)] = 0$ for all $k > 0$, i.e., $X_j(n)$ is orthogonal to $\xi_{i|j}(n-k)$ for all $k > 0$.

2. $G_{i|j}(\lambda) = B(\lambda)$

$\xi_{i|j}(n)$ is a linear combination of the innovations, and it includes the quantity $\nu_i(n)$. For $X_j(n)$ to be orthogonal to $\xi_{i|j}(n-k)$ for all $k > 0$, it must be orthogonal to each of the individual innovations that constitute $\xi_{i|j}(n-k)$. Therefore

$$\mathbb{E}[X_j(n)\nu_i(n-k)] = 0$$

for all $k > 0$. But this can only happen if $\{X_i(n)\}$ is not an ancestor of $\{X_j(n)\}$, in which case $\{X_i(n)\}$ does not Granger-cause $\{X_j(n)\}$, by lemma 5.2.1. This completes the proof. □

5.4 Efficacy of Wiener filters in detecting causality: simulation and real data

Motivated by the results obtained in section 5.3, in this section, the effectiveness of FIR Wiener filters in determining the causal structure of a system is investigated through simulation. The processes are first normalized to make them zero-mean and of equal variance σ^2. Recall that $H_j^p(n-1)$ denotes the linear span of $\{X_j(n-1), \ldots, X_j(n-p)\}$ for each $j \in \{1, \ldots, N\}$. The objective is to fit a one-step FIR Wiener filter-based predictor of order p for each pair of processes. Essentially, this is equivalent to projecting $X_i(n)$ onto the space $H_j^p(n-1)$; for each i and j, $i \neq j$; i.e., $X_i(n)$ is estimated as

$$\hat{X}_i(n) = \sum_{k=1}^{p} \hat{b}_{i,j}^p(k) X_j(n-k)$$

The parameters $[\hat{b}_{i,j}^p(1) \ \ldots \ \hat{b}_{i,j}^p(p)]^T = \boldsymbol{\hat{B}_{i,j}^p}$ can be estimated as

$$\boldsymbol{\hat{B}_{i,j}^p} = \boldsymbol{hatR}(X_j^p)^{-1}\hat{\boldsymbol{r}}(X_{i,j}^p)$$

where

$$\hat{\boldsymbol{R}}(X_j^p) = \begin{pmatrix} \hat{R}_j(0) & \hat{R}_j(1) & \cdots & \hat{R}_j(p-1) \\ \hat{R}_j(1) & \hat{R}_j(0) & \cdots & \hat{R}_j(p-2) \\ \vdots & \vdots & \ddots & \vdots \\ \hat{R}_j(p-1) & \hat{R}_j(p-2) & \cdots & \hat{R}_j(0) \end{pmatrix}$$

is the estimated p-th order covariance matrix for $\{X_i(n)\}$ and $\hat{\boldsymbol{r}}(X_{i,j}^p) = [\hat{R}_{i,j}(1) \dots \hat{R}_{i,j}(p)]^T$ is a vector containing the estimated cross-covariance terms of $\{X_i(n)\}$ and $\{X_j(n)\}$. Let the corresponding estimated mean squared error be denoted by $\hat{\mathbb{E}}[(\xi[X_i(n)|H_j^p(n-1)])^2]$. If this error is significantly close to the variance of the processes, i.e., if for some pre-defined threshold of significance ϵ_1,

$\left|\sigma^2 - \hat{\mathbb{E}}[(\xi[X_i(n)|H_j^p(n-1)])^2]\right| < \epsilon_1$ it is concluded that there is no edge from j to i. Otherwise, if $\hat{\mathbb{E}}[(\xi[X_i(n)|H_j^p(n-1)])^2] < \hat{\mathbb{E}}[(\xi[X_j(n)|H_i^p(n-1)])^2]$, it is concluded that there is an edge from j to i. This ensures that there are no closed loops in the connected graph that represents causal relations.

Once all edges are detected, an additional step is employed to eliminate some of the false edges. Within the graph, we detect all triangles; i.e., all sets of three nodes (i, j, k) that are connected. Denote $\hat{\mathbb{E}}[(\xi[X_j(n)|H_i^p(n-1)])^2]$ by $\vec{d}(i,j)$. Whenever $\vec{d}(i,j)+\vec{d}(j,k) < \vec{d}(i,k)$, $\vec{d}(i,k)$ is set to zero. In the reduced graph thus obtained, $\vec{d}(\cdot,\cdot)$ is a quasimetric.

This method was used for a simulated system of jointly WSS Gaussian processes with $N = 10$, having the configuration of figure 5.2, and the same parameters as used to demonstrate the multivariate autoregressive approach (figure 5.3). The system was simulated on MATLAB using 50,000 sample realizations. All processes were suitably scaled and shifted to make them zero mean and of equal variance.

It was observed that the method could successfully identify driving processes and accurately reveal the hierarchical order of nodes, although the interconnections were overestimated through the detection of several false edges, in addition to the edges present in the original system. The hierarchical order of nodes determined by the method remained consistent when the original filter parameters were varied within a reasonable range. The results were also affected by the choice of the threshold parameter ϵ_1. A smaller value of ϵ_1 resulted in the detection of an increased number of false edges between driving processes whereas for larger values some of the original edges remained undetected.

For comparison, a second method, involving the notion of directed information was also used for the same system. For a suitable order p, define $\mathbb{R}^p$-valued random vectors $\boldsymbol{X_i^p}$, $\boldsymbol{X_j^p}$ as $\boldsymbol{X}_i^p = [X_i(n) \ \dots \ X_i(n-p+1)]^T$ and $\boldsymbol{X}_j^p = [X_j(n) \ \dots \ X_j(n-p+1)]^T$ respectively. For $k < p$, let $\boldsymbol{X}_i^k = [X_i(n-p+k) \ \dots \ X_i(n-p+1)]^T$. For Gaussian processes, the directed information from $\boldsymbol{X}_j^{p-1}$ to $\boldsymbol{X}_i^p$ simplifies to [53]:

$$I(\boldsymbol{X_j^{p-1}} \to \boldsymbol{X_i^p}) = \frac{1}{2}\sum_{k=1}^{p} \log \left| \frac{\det R(\boldsymbol{X_i^k})/\det R(\boldsymbol{X_i^{k-1}})}{\det R(\boldsymbol{X_i^k}, \boldsymbol{X_j^{k-1}})/\det R(\boldsymbol{X_i^{k-1}}, \boldsymbol{X_j^{k-1}})} \right|$$

where $R(\boldsymbol{X_i^k})$ is the covariance matrix of $[X_i(n) \ \ldots \ X_i(n-k+1)]^T$; $R(\boldsymbol{X_i^k, X_j^l})$ is the covariance matrix of $[X_i(n) \ \ldots \ X_i(n-k+1) \ X_j(n) \ \ldots \ X_j(n-l+1)]^T$ and by definition, $R(\boldsymbol{X_i^0}) = 1$.

Directed information is approximated using estimates of covariance terms in the above expression. Denote the estimated directed information as $\hat{I}(\boldsymbol{X_j^{p-1}} \rightarrow \boldsymbol{X_i^p})$. Directed information rate from process $\{X_j(n)\}$ to $\{X_i(n)\}$ is approximated as

$$\hat{I}(X_j \rightarrow X_i) = \frac{1}{p}\hat{I}(\boldsymbol{X_j^{p-1}} \rightarrow \boldsymbol{X_i^p})$$

To determine the existence of edges in the graph, a procedure analogous to the Wiener filter method is followed. If the directed information computed is significantly close to 0, i.e., if for some threshold of significance $\epsilon_2 > 0$, $|\hat{I}(X_j \rightarrow X_i)| < \epsilon_2$, it is concluded that there is no edge from j to i. Otherwise, if $|\hat{I}(X_j \rightarrow X_i)| > |\hat{I}(X_i \rightarrow X_j)|$, it is concluded that there is an edge from j to i. Finally, to obtain a quasimetric akin to the one in the Wiener filter approach, we replace each $|\hat{I}(X_i \rightarrow X_j)|$ with $\vec{d}(i,j) = K - |\hat{I}(X_i \rightarrow X_j)|$, where $K = \max_{i,j} |\hat{I}(X_i \rightarrow X_j)| + \epsilon_2$, and remove all edges that do not satisfy the triangle inequality with $\vec{d}(\cdot,\cdot)$.

Results of the directed information based method were found to be comparable to those obtained by the Wiener filter approach. Driving processes were accurately identified. For this method too, results were seen to be sensitive to the threshold parameter ϵ_2. A smaller ϵ_2 lead to the false detection of additional edges; while when ϵ_2 was too large, some of the original edges remained undetected.

The structures revealed by the proposed methods are graphically represented in figure 5.5. An edge between two nodes represents a causal filter between the corresponding processes.

The section is concluded with an example of using FIR Wiener filters in determining Granger-causality among time series from real data. We consider a set of currency exchange rates of some of the world's leading economies. Fluctuations in daily exchange rates of these currencies against the Swiss Franc for the period January 1, 2009 to December 31, 2012 were used. The data was obtained from the Bank of Canada website [113]. Causal links were estimated using the Wiener filter based method and is presented in figure 5.6. It is interesting to note that in this example, currencies of economies involved in significant two-way trade indicated stronger dependence. It would have been nicer if there was any way to compare our results with the ground truth in this case.

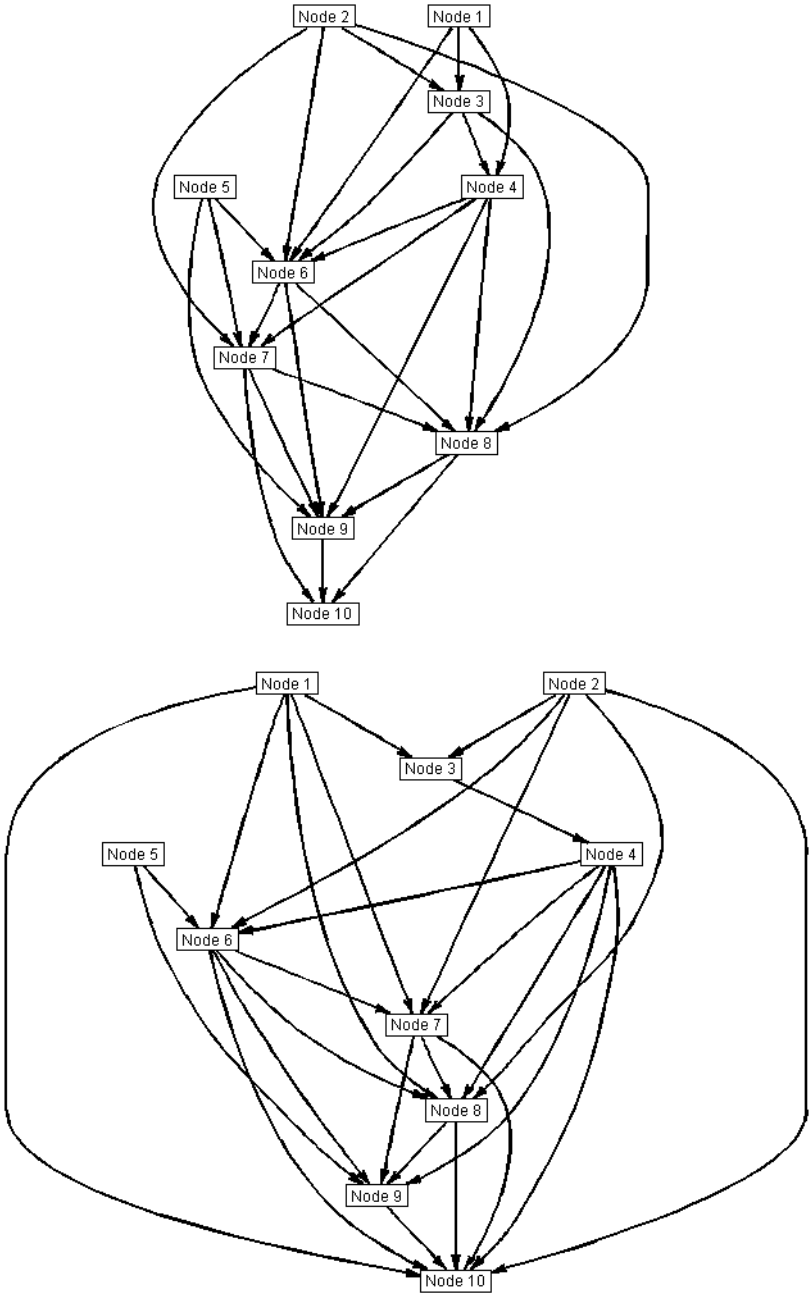

Figure 5.5: A system of 10 processes- top: system recovered through FIR Wiener filter(p=7), bottom: system recovered through directed information(p=7).

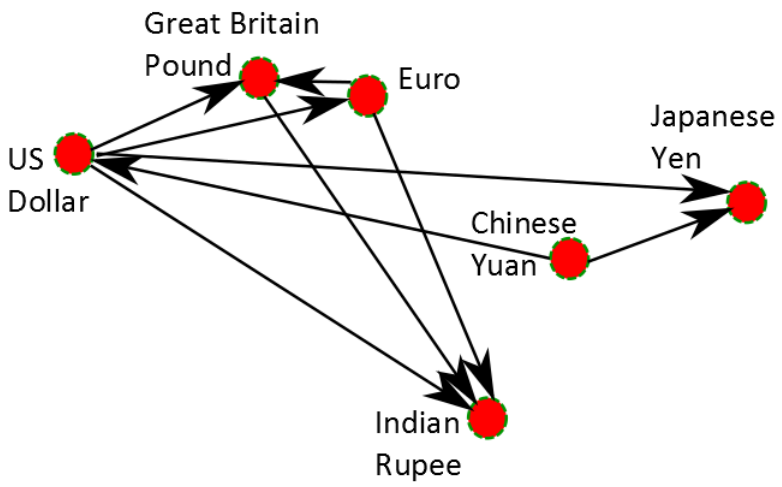

Figure 5.6: Interrelation of currencies inferred using causal Wiener filters (p=10)

5.5 Conclusion

In this chapter, we have presented several results on the utility of pairwise Wiener filtering in determining Granger-causality within a family of stochastic processes. It follows from our results that the method may be employed to obtain useful information on the interdependence relations in a system of jointly WSS random processes. Our theoretical results as well as simulations demonstrate that this method reliably reconstructs the hierarchical structure of nodes and detects most of the edges in the original system.

It is, however clear that there is a limit to the information that may be inferred through such means. The method is not sufficient to unambiguously determine all interdependence relations within a system, and while the general hierarchy of nodes is revealed, there is no direct way of distinguishing between parents and distant ancestors.

Nonetheless, as demonstrated by the simulation results, Wiener filtering can be a quick, efficient tool in obtaining reasonably accurate information on the causal connections of a family of stochastic processes. The method is similar to that proposed in [59]; and yet much easier to implement. For FIR Wiener filters one only needs to estimate a small number of covariance and cross-covariance terms as opposed to spectral densities required for the IIR (infinite impulse response) case.

The performance of the FIR Wiener filter is comparable to a directed information based approach. However, while the latter has gained popularity in recent years; it involves estimating the *distributions* of the processes involved, and is therefore computationally bur-

densome for general processes. In that regard, Wiener filtering is an easier and more robust alternative in detecting causal structures for jointly WSS processes where no information is available on either the distribution or the support set of the processes.

Chapter 6

Cyclostationary Processes: AR Estimation and Granger-causality

6.1 Introduction

In the previous chapter, we addressed the problem of detecting Granger-causality within a group of WSS processes. While WSS processes are relatively easy to analyze, a number of processes observed in real applications are non-stationary and therefore require more involved treatment. Many processes encountered in various fields of study, including communications and control systems involve parameters that vary periodically with time. A large class of such processes can be appropriately modeled as cyclostationary (CS) processes [114]. In this chapter, the problems of autoregressive estimation and detection of Granger-causality are studied in the context of the latter. After an introductory discussion on some preliminary results, we present a method of estimating CS processes using *time-invariant* autoregressions. Finally, the method is extended in the context of detecting Granger-causality.

We begin with a discussion on some preliminary results on cyclostationary processes. Some of these results can be found in [115, 116, 117] and [118].

6.2 Preliminaries

Let $\{X(n)\}_{n\in\mathbb{Z}}$ be a real-valued, zero-mean, discrete time stochastic process defined over a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. $\{X(n)\}$ is said to be cyclostationary (CS) [114] if the covariance sequence $\mathbb{E}[X(n)X(n-k)] = R(n,k)$ is periodic in the following sense.

$$R(n,k) = R(n+iT_0,k) \tag{6.2.1}$$

where i is an integer. We call T_0 the period of the cyclostationary (CS) process $\{X(n)\}$.

The p-th order autoregressive (AR-p) estimate of $X(n)$ is its best linear predictor based on its p most recent values. It is given by

$$\hat{X}_p(n) = \sum_{k=1}^{p} \beta_p(n,k)X(n-k)$$

where the parameters $\beta_p(n,k)$ are derived by the method of least squares; i.e., the mean squared error

$$\xi_p(n) = \mathbb{E}\left[\left(X(n) - \sum_{k=1}^{p} b_p(n,k)X(n-k)\right)^2\right] \tag{6.2.2}$$

is minimized when $b_p(n,k) = \beta_p(n,k)$. If the instantaneous error corresponding to the AR-p estimate is given by $\nu_p(n)$, then

$$X(n) = \sum_{k=1}^{p} \beta_p(n,k)X(n-k) + \nu_p(n)$$

Lemma 6.2.1. *If $X(n)$ is estimated as an autoregression of order p with time-varying parameters of the form*

$$\hat{X}_p(n) = \sum_{k=1}^{p} \beta_p(n,k)X(n-k)$$

then the parameters $\beta_p(n,k)$ are periodic in n with period T_0, i.e.,

$$\beta_p(n,k) = \beta_p(n+iT_0,k)$$

where i is an integer.

Proof. The proof is trivial. The Yule-Walker equations derived through the method of least squares lead to the following system of simultaneous equations.

$$\boldsymbol{B_p}(n) = \boldsymbol{R_p}^{-1}(n)\boldsymbol{r_p}(n)$$

where

$$\boldsymbol{R_p(n)} = \begin{pmatrix} R(n-1,0) & \cdots & R(n-1,p-1) \\ \vdots & \ddots & \vdots \\ R(n-p,1-p) & \cdots & R(n-p,0) \end{pmatrix}$$

$$\boldsymbol{r_p}(n) = [R(n,1)\ R(n,2)\ \ldots\ R(n,p)]^T$$

and

$$\boldsymbol{B_p}(n) = [\beta_p(n,1)\ \ldots\ \beta_p(n,p)]^T$$

The result follows readily, noting that $\boldsymbol{R_p}(n) = \boldsymbol{R_p}(n+iT_0)$ and $\boldsymbol{r_p}(n) = \boldsymbol{r_p}(n+iT_0)$, due to the periodic property of $R(n,k)$.

□

Theorem 6.2.1. *The error* $\{\nu_p(n)\}$ *is a CS process with*

$$\mathbb{E}[\nu_p(n)\nu_p(n-k)] = \mathbb{E}[\nu_p(n+T_0)\nu_p(n+T_0-k)]$$

Proof. By definition,

$$\begin{aligned} \nu_p(n) &= X(n) - \sum_{k=1}^{p} \beta_p(n,k)X(n-k) \\ &= -\sum_{k=0}^{p} \beta_p(n,k)X(n-k) \end{aligned}$$

with $\beta_p(n,0) = -1$ for all n, p. Then, the covariance of $\{\nu_p(n)\}$ is given by

$$\begin{aligned}
\mathbb{E}[\nu_p(n)\nu_p(n-\tau)] &= \mathbb{E}\left[\left(\sum_{j=0}^{p}\beta_p(n,j)X(n-j)\right)\left(\sum_{k=0}^{p}\beta_p(n-\tau,k)X(n-\tau-k)\right)\right] \\
&= \sum_{j=0}^{p}\sum_{k=0}^{p}\beta_p(n,j)\beta_p(n-\tau,k)R(n-j,k-j+\tau) \\
&= \sum_{j=0}^{p}\sum_{k=0}^{p}\beta_p(n+T_0,j)\beta_p(n-\tau+T_0,k)R(n-j+T_0,k-j+\tau) \\
&= \mathbb{E}[\nu_p(n+T_0)\nu_p(n+T_0-\tau)]
\end{aligned}$$

where the second last step follows from the fact that $\{X(n)\}$ is CS and from lemma 6.2.1.

□

Corollary 6.2.1. *The innovation process associated with a cyclostationary process is an orthogonal sequence with a variance that varies with period T_0.*

The innovation process $\{\nu(n)\}$ of the process $\{X(n)\}$ is the component of $\{X(n)\}$ that is orthogonal to the linear span $H_X(n-1)$. The proof follows readily from lemma 6.2.1, noting that $\nu_p(n) \to \nu(n)$ in quadratic mean [12, Lemma 3.1(b)].

6.3 Representation of a CS process as a vector-valued WSS process

Let the CS process $\{X(m)\}$ be such that, for any m, k,

$$\mathbb{E}[X^2(m)] \geq \mathbb{E}[X(m)X(m-k)] \tag{6.3.3}$$

For each $m \in \mathbb{Z}$, let $n = \left\lceil \frac{m}{T_0} \right\rceil$, and let $i = m - T_0\left(\left\lceil \frac{m}{T_0} \right\rceil - 1\right)$ Define a family of T_0 stochastic processes $\{Y_i(n)\}_{i=1,\ldots,T_0}$ as

$$Y_i(n) = X((n-1)T_0 + i)$$

The cross-covariance of any two processes $\{Y_i(n)\}$ and $\{Y_j(n)\}$ can be found as follows, using (6.2.1).

$$\mathbb{E}[Y_i(n)Y_j(n-k)] = R(i, kT_0 + i - j)$$

Since the expression is independent of n, it follows that $\{Y_i(n)\}$ constitute a family of jointly WSS processes [116]. Furthermore, if $\{\boldsymbol{Y}(n)\}$ is defined as the $\mathbb{R}^{T_0}$-valued process $\boldsymbol{Y}(n) = [Y_1(n) \ \ldots \ Y_{T_0}(n)]^T$ then $\{\boldsymbol{Y}(n)\}$ is a multivariate WSS process. Define $\{\nu_i(n)\}$ as the corresponding family of innovation processes.

$$\nu_i(n) = \nu\left((n-1)T_0 + i\right)$$

Finally, let $\{\boldsymbol{\nu(n)}\}$ be the corresponding vector-valued innovation process. $\boldsymbol{\nu}(n) = [\nu_1(n) \ \ldots \ \nu_{T_0}(n)]^T$.

The original CS process can be expressed in terms of $\{Y_i(n)\}$ in the form of a vector autoregressive (VAR) model (This is similar to the construction in [118]). For some n, i, let $m = (n-1)T_0 + i$. Then,

$$\begin{aligned} Y_i(n) &= \sum_{k=1}^{\infty} \beta(m,k)X(m-k) + \nu(m) \\ &= \sum_{j=1}^{i-1} \gamma_{i,j}(0)Y_j(n) + \sum_{k=1}^{\infty}\sum_{j=1}^{T_0} \gamma_{i,j}(k)Y_j(n-k) + \nu_i(n) \end{aligned}$$

where the new parameters $\gamma_{i,j}(k)$ are given by $\gamma_{i,j}(k) = \beta(i, kT_0 + i - j)$.

The above system of equations represents the individual processes $\{Y_i(n)\}$ in terms of the past of all the processes in the system. However, in this representation, the value of $Y_i(n)$ at time instant n is dependent on those of $\{Y_j(n)\}_{j<i}$ and therefore this cannot be used directly to develop an expression for an AR expansion of the vector-valued process $\{\boldsymbol{Y}(n)\}$. To obtain such an expression, the following adjustments are made.

Define the $\mathbb{R}^{T_0 \times T_0}$-valued lower triangular matrix $\boldsymbol{\Gamma}(0)$ as

$$\boldsymbol{\Gamma}(0) = \begin{pmatrix} 0 & 0 & \cdots & 0 \\ \gamma_{2,1}(0) & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ \gamma_{i,1}(0) & \gamma_{i,2}(0) & \ddots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ \gamma_{T_0,1}(0) & \gamma_{T_0,2}(0) & \cdots & 0 \end{pmatrix}$$

For $k > 0$ define the matrices $\mathbf{\Gamma}(k)$ as $\mathbf{\Gamma}(k) = [\gamma_{i,j}(k)]$.

$$\boldsymbol{Y}(n) = \mathbf{\Gamma}(0)\boldsymbol{Y}(n) + \sum_{k=1}^{\infty} \mathbf{\Gamma}(k)\boldsymbol{Y}(n-k) + \boldsymbol{\nu}(n)$$

and therefore,

$$\left(\boldsymbol{I} - \mathbf{\Gamma}(0)\right)\boldsymbol{Y}(n) = \sum_{k=1}^{\infty} \mathbf{\Gamma}(k)\boldsymbol{Y}(n-k) + \left(\boldsymbol{I} - \mathbf{\Gamma}(0)\right)^{-1}\boldsymbol{\nu}(n)$$

Where $\boldsymbol{I}$ denotes the $T_0 \times T_0$ identity matrix. Note that the determinant of $(\boldsymbol{I} - \mathbf{\Gamma}(0))$ is unity and therefore the matrix is invertible. Rearranging, then, we obtain the following VAR representation of $\{\boldsymbol{Y}(n)\}$.

$$\begin{aligned}
\boldsymbol{Y}(n) &= \sum_{k=1}^{\infty} \left(\boldsymbol{I} - \mathbf{\Gamma}(0)\right)^{-1} \mathbf{\Gamma}(k)\boldsymbol{Y}(n-k) + \left(\boldsymbol{I} - \mathbf{\Gamma}(0)\right)^{-1}\boldsymbol{\nu}(n) \\
&= \sum_{k=1}^{\infty} \mathbf{\Gamma}'(k)\boldsymbol{Y}(n-k) + \boldsymbol{\nu}'(n)
\end{aligned}$$

where $\mathbf{\Gamma}'(k) = \left(\boldsymbol{I} - \mathbf{\Gamma}(0)\right)^{-1}\mathbf{\Gamma}(k)$ for $k > 0$ and $\boldsymbol{\nu}'(n) = \left(\boldsymbol{I} - \mathbf{\Gamma}(0)\right)^{-1}\boldsymbol{\nu}(n)$.

Given N cyclostationary processes with the same period T_0, the above decomposition can be used to express the system as a combination of NT_0 real-valued WSS processes, and the standard tools to detect Granger-causality can be thereby applied. However, this requires a high level of computation that has to be carried out using a large number of samples.

6.4 Time-invariant AR model to estimate CS processes

A CS process can be expressed as both an autoregression with periodically varying parameters and a VAR model. However, using such estimates are computationally burdensome, as compared to a WSS process of the same model order, the number of parameters involved in a CS process is higher by a factor of T_0. Furthermore, larger number of data points are

needed to find reliable estimators of $R(n,k)$ for $n = 1, \ldots, T_0$ and $k = 1, \ldots, p$ where p is the order of the AR model.

An alternative is to *treat* the CS process as a WSS process and obtain a *time-invariant* AR model. Define

$$\hat{R}_N(k) = \frac{1}{N} \sum_{n=|k|+1}^{N} X(n)X(n-k)$$

For $i = 1, \ldots, T_0$, define

$$\hat{R}_{i,N}(k) = \frac{1}{N} \sum_{n=|k|+1}^{N} Y_i(n)Y_i(n-k)$$

$\hat{R}_{i,N}(k)$ are empirical covariances of the WSS processes $\{Y_i(n)\}$. Assume the processes to be covariance-ergodic.

$$\lim_{N\to\infty} \hat{R}_{i,N}(k) = \mathbb{E}[Y_i(n)Y_i(n-k)]$$

Finally, define

$$\widetilde{R}(k) = \frac{1}{T_0} \sum_{i=1}^{T_0} \mathbb{E}[X(i)X(i-k)]$$

$\widetilde{R}(k)$ so defined, does not depend on i. Also, by (6.3.3), $\widetilde{R}(0) \geq \widetilde{R}(k)$.

Lemma 6.4.1. *For all k, $\lim_{N\to\infty} \hat{R}_N(k) = \widetilde{R}(k)$.*
Proof.

$$\lim_{N\to\infty} \hat{R}_N(k) = \lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{T_0} \sum_{n=|k|+1}^{n_i(N)} Y_i(n)Y_i(n-k)$$

where $n_i(N) = \left\lceil \frac{N-i}{T_0} \right\rceil$. Note that as $N \to \infty$, $n_i(N) \to \infty$ for all i and

$$\lim_{N\to\infty} \frac{n_i(N)}{N} = \frac{1}{T_0}$$

for all i. Therefore,

$$\begin{aligned}\lim_{N\to\infty}\hat{R}_N(k) &= \sum_{i=1}^{T_0}\left(\lim_{N\to\infty}\frac{n_i(N)}{N}\right)\left(\lim_{n_i(N)\to\infty}\frac{1}{n_i(N)}\sum_{n=|k|+1}^{n_i(N)}Y_i(n)Y_i(n-k)\right)\\ &= \frac{1}{T_0}\sum_{i=1}^{T_0}\mathbb{E}[Y_i(n)Y_i(n-k)]\\ &= \widetilde{R}(k)\end{aligned}$$

□

The limit of $\hat{R}_N(k)$ gives the arithmetic mean of the different covariance values of the CS process at the same lag k. The terms $\widetilde{R}(k)$ can be estimated through $\hat{R}_N(k)$, and can be used to determine time-invariant AR parameters $\widetilde{\boldsymbol{B}}_{\boldsymbol{p}} = [\widetilde{b}_p(1)\ \ldots\ \widetilde{b}_p(T_0)]^T$ by solving the Yule-Walker equations

$$\widetilde{\boldsymbol{B}}_{\boldsymbol{p}} = \widetilde{\boldsymbol{R}}_{\boldsymbol{p}}^{-1}\widetilde{\boldsymbol{r}}_{\boldsymbol{p}}$$

The parameters so obtained essentially minimize the limiting mean squared error

$$\lim_{N\to\infty}\frac{1}{N}\sum_{n=1}^{N-p}\left(X(n)-\sum_{k=1}^{p}\widetilde{b}_p(k)X(n-k)\right)^2$$

Note that as the sequence $\{\widetilde{R}(k)\}$ is positive semidefinite by construction, it is a valid covariance sequence, i.e., there exists some WSS stochastic process $\{Z(n)\}$ with covariance sequence $\{\widetilde{R}(k)\}$. The time-invariant AR model obtained using $\{\widetilde{R}(k)\}$ is essentially the AR model corresponding to $\{Z(n)\}$.

6.5 Granger-causality between CS processes

In this section, we explore the problem of inferring causal relations among CS processes with the same period T_0. Two CS processes $\{X(n)\}$, $\{Y(n)\}$ with period T_0 are said to be jointly CS if

$$\mathbb{E}[X(n)Y(n-k)] = \mathbb{E}[X(n+iT_0)Y(n+iT_0-k)]$$

for any integer values of k, i. In other words, the cross-covariance terms are periodic with period T_0.

Consider two jointly CS processes $\{X_i(n)\}$ and $\{X_j(n)\}$ having the same period T_0. The definition of Granger-causality given in chapters 1 and 5 can be slightly modified to accommodate CS processes. Unlike the WSS case, here, the estimation parameters and error variances will no longer be stationary but will vary with period T_0. Again, for a model order of p, $\{X_i(n)\}$ is first modeled as an univariate AR process with error $\xi_{i|i}$; i.e.,

$$X_i(n) = \sum_{i=1}^{p} \alpha_{i,i}(n,k) X_i(n-k) + \xi_{i|i}(n)$$

and then modeled as an autoregression that also includes past observations of $\{X_j(n)\}$ with error $\xi_{i|i,j}$:-

$$X_i(n) = \sum_{i=1}^{p} \beta_{i,i}(n,k) X_i(n-k) + \sum_{i=1}^{p} \beta_{i,j}(n,k) X_j(n-k) + \xi_{i|i,j}(n)$$

We say that $\{X_j(n)\}$ Granger-causes $\{X_i(n)\}$ if for some $n \in \{1,\ldots,T_0\}$,

$$\mathbb{E}[\xi^2_{i|i}(n)] > \mathbb{E}[\xi^2_{i|i,j}(n)]$$

In terms of the projection notation, $\{X_j(n)\}$ Granger-causes $\{X_i(n)\}$ if for some $n \in \{1,\ldots,T_0\}$,

$$\mathbb{E}\left[(\xi[X_i(n)|H_i(n-1)])^2\right] > \mathbb{E}\left[(\xi[X_i(n)|H_{i,j}(n-1)])^2\right]$$

For a family of N jointly CS processes $\{X_k(n)\}_{k=1,\ldots,N}$, $\{X_i(n)\}$ is said to be Granger-caused by $\{X_j(n)\}$ if there exists some $n \in \{1,\ldots,T_0\}$ such that

$$\mathbb{E}\left[(\xi[X_i(n)|H_{-j}(n-1)])^2\right] > \mathbb{E}\left[(\xi[X_i(n)|H(n-1)])^2\right]$$

To check for the above condition, one requires to solve for the least square parameters for each n, from 1 to T_0. When the period T_0 is large, this becomes computationally intensive.

For a system consisting of a large number of jointly CS processes with the same period, the ideal approach to determine causal relations will be to fit a multivariate autoregressive (MVAR) model. However, because of the cyclostationarity of the processes involved, the MVAR parameters will vary periodically, and so the number of equations required to solve will be multiplied by a factor of T_0, compared to the WSS case.

Following the ideas of chapter 5, a pairwise Wiener filter approach can be used to detect Granger-causality for CS processes as well. However, because the processes are CS with

the same period T_0, the Wiener filter parameters, too, will vary periodically. As a result, one has to separately solve for Wiener filter parameters for each $n = 1, \ldots, T_0$. While such a method will lead to accurate estimation of the processes, this, too, would involve a high level of computation.

Alternatively, a time-invariant Wiener filter estimate may be used, analogous to the time-invariant AR estimates discussed in section 6.4, which finds the best time-invariant estimate of $X_i(n)$ in terms of a linear combination of the past values of $X_j(n)$. In that case, the time-invariant average cross-covariance $\widetilde{R}_{i,j}(\tau)$ is computed as

$$\widetilde{R}_{i,j}(\tau) = \lim_{T\to\infty} \frac{1}{T} \sum_{n=\tau+1}^{T} X_i(n) X_j(n-\tau)$$

The above quantity is equal to the arithmetic mean of the quantities $\{R_{i,j}(n,\tau)\}_{n=1,\ldots,T_0}$. The Wiener-Hopf equations are solved by using this average cross-covariance $\widetilde{R}_{i,j}(\tau)$. Let $\overline{\xi}_{i|j}$ be the corresponding error.

The following result shows that for CS processes, this time-invariant Wiener filter can be used to determine if $\{X_i(n)\}$ is Granger-caused by $\{X_j(n)\}$.

Theorem 6.5.1. *Consider a system of two jointly cyclostationary processes $\{X_i(n)\}$ and $\{X_j(n)\}$, with period T_0. If $\{X_j(n)\}$ Granger-causes $\{X_i(n)\}$, then*

$$\mathbb{E}[\overline{\xi}^2_{i|j}] < \widetilde{R}_i(0)$$

Proof. Note that for the required condition to be satisfied, we need to show the existence of some $\tau > 0$ for which $\widetilde{R}_{i,j}(\tau) \neq 0$.

Since $\{X_j(n)\}$ Granger-causes $\{X_i(n)\}$, there exists some $m \in \{1, \ldots, T_0\}$ for which

$$X_i(m) = \sum_{k=1}^{p} \beta_{i,j}(m,k) X_j(m-k) + \sum_{k=1}^{p} \beta_{i,i}(m,k) X_i(m-k) + \xi_{i|i,j}(m)$$

Taking z-transforms on both sides, and re-arranging

$$X_i(z) = \left(1 - \sum_{k=1}^{p} \beta_{i,i}(m,k) z^{-k}\right)^{-1} \left(\sum_{k=1}^{p} \beta_{i,j}(m,k) z^{-k} X_j(z) + \xi_{i|i,j}(z)\right)$$

Shifting back to the time-domain, the above becomes

$$X_i(m) = \sum_{k=1}^{\infty} c_{i,j}(m,k) X_j(m-k) + \sum_{k=0}^{\infty} d_{i,j}(m,k) \xi_{i|i,j}(m-k)$$

where the parameters $c_{i,j}(m,k)$s and $d_{i,j}(m,k)$s can be derived from the $\beta_{i,i}(m,k)$s and $\beta_{i,j}(m,k)$s, and in general, $\beta_{i,j}(m,k) \neq 0$. Thus, $\sum_{k=1}^{\infty} c_{i,j}(m,k)X_j(m-k)$ is a non-trivial estimator of $X_i(m)$. It follows then, that the time-varying causal Wiener filter with the past values of $X_j(n)$ as input and $X_i(n)$ as output, will be non-trivial as well for $n = m$, and the estimation error will be less or equal to that corresponding to $\sum_{k=1}^{\infty} c_{i,j}(m,k)X_j(m-k)$. Then, $X_i(m)$ is not orthogonal to the past values of $X_j(m)$, and there exists some τ such that

$$\mathbb{E}[X_i(m)X_j(m-\tau)] \neq 0$$

Therefore, $R_{i,j}(m,\tau) \neq 0$. It follows, then, that $\widetilde{R}_{i,j}(\tau)$ is also non-zero and the result follows. □

The above result has a similarity with theorem 5.3.1. Like theorem 5.3.1, here too, the reverse, in general, is not true.

6.6 Results with real data

In this section, the efficacy of Wiener filters in determining Granger-causality is studied in the context of data obtained from a practical application. Fluctuations in many of the variables involved in climate and weather may be characterized as cyclostationary [117]. In this example, daily mean temperatures of several cities in Ontario for the year 2010, obtained from the Canadian climate data website [119] were used. Dependence relations were inferred through the pairwise Wiener filter-based method proposed in chapter 5. Causal connections inferred using our approach is presented in figure 6.6.

The pattern of inter-relations detected by our method relates closely to the geographical locations of the cities considered. Temperatures of cities that are close to each other are seen to reflect stronger dependence relations. Also, in general, the direction of causality is observed to be from West to East.

6.7 Conclusion

In this chapter, the problem of estimating CS time series through AR approximations has been studied. Starting with a brief discussion on the background of the theory of estimating CS processes, we have developed a *time-invariant* AR estimator of such processes.

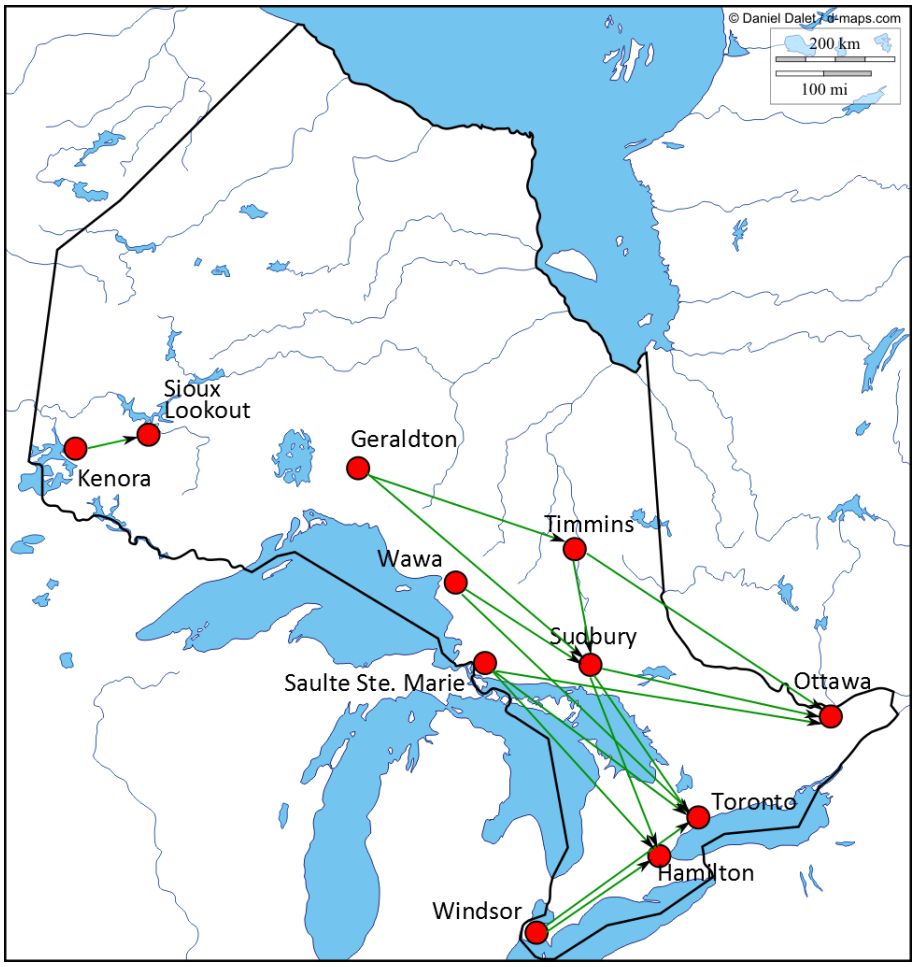

Figure 6.1: Interrelation of daily mean temperature in cities of Ontario

Furthermore, we have shown that a time-invariant causal Wiener filter, based on the same principles, can be used to detect causality among several CS processes having the same period. While for a pair of CS processes, time-varying Wiener filters are more accurate for estimation, it is interesting to note that the easy-to-compute time-invariant version can also be used for the detection of Granger-causality.

Our result further shows that the results on the pairwise causal Wiener filter presented in chapter 5 are also applicable to CS processes. Moreover, the technique proposed in section 5.4 can be used to determine interdependence relations within a family of CS time series as well, as demonstrated through the example with temperature related data.

Chapter 7

Detecting Causality Under Sparsity Constraints

7.1 Introduction

Recall the problem of determining the underlying causal connections for a family of a number of WSS processes from their dynamic behaviour. From a theoretical perspective, the best solution is obtained by solving equations for the multivariate autoregressive (MVAR) model, which determines the ordinary least square predictor of the system. In terms of practical usability, however, this approach has several limitations.

In almost all practical scenarios, true covariance and cross-covariance terms are not directly available and have to be estimated from data. These estimates are likely to deviate at least slightly from their true values which would result in the detection of additional spurious links of causality. Often the observed values of the time series themselves are contaminated with additional noise, which exacerbates the problem. Lack of knowledge about the exact model order leads to further complications. Finally, characteristics of time series observed in the real world may not adhere strictly to those of a WSS process, which would again adversely affect the outcome of the method. As a result, the configuration identified by the MVAR model is often a complicated mesh with causal links between most of the pairs of nodes.

A model that indicates causal linkage between almost every pair of processes is difficult to interpret. It is more useful to identify and keep the more significant links, i.e., the links that reflect the strongest dependences, and remove the weaker ones. Finally, the

least square estimates often tend to have high variances, and in those cases, more accurate prediction can be made by setting some of the smaller parameters to zero [88]. For the above reasons, a sparse model with fewer non-zero parameters is preferable in the context of multivariate prediction.

A causal network with fewer edges is easier to interpret and more conducive to prediction. In this chapter, we suggest a technique that modifies the MVAR approach with a constraint that reduces the number of edges entering a node. In other words, it restricts the number of processes affecting (or Granger-causing) any given process.

7.2 Problem formulation

Consider a system of N discrete time, real-valued, zero-mean, regular stochastic processes defined over a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Let each process be denoted by $\{X_i(n)\}$ for $i = \{1, \ldots, N\}$. Let $\boldsymbol{X}(n) = [X_1(n) \ \ldots \ X_N(n)]^T$. The objective is to find a causal linear MMSE estimate of the $\mathbb{R}^N$-valued process $\boldsymbol{X}(n)$ using the most recent p past observations, under the restriction that the number of processes directly influencing any process is small, i.e., we are interested in finding a multivariate AR model that preserves only the *strongest* causal links in the system. The required estimate $\breve{\boldsymbol{X}}^p$ is a linear combination of the past values of $\boldsymbol{X}(n)$.

$$\breve{\boldsymbol{X}}^p = \sum_{k=1}^{p} \breve{B}^p(k)\boldsymbol{X}(n-k)$$

where $\breve{B}^p(k) = [\breve{b}^p_{i,j}(k)]$ is a matrix in $\mathbb{R}^{p\times p}$. For each individual process $X_i(n)$, the estimate is given by

$$\breve{X}^p_i(n) = \sum_{k=1}^{p}\sum_{j=1}^{N} \breve{b}_{i,j}(k)X_j(n-k) \tag{7.2.1}$$

Define $\breve{B}^p_i$ as the matrix whose elements are the optimal parameters corresponding to the process $\{X_i(n)\}$.

$$\breve{B}^p_i = \begin{pmatrix} \breve{b}^p_{i,1}(1) & \breve{b}^p_{i,2}(1) & \cdots & \breve{b}^p_{i,N}(1) \\ \breve{b}^p_{i,1}(2) & \breve{b}^p_{i,2}(2) & \cdots & \breve{b}^p_{i,N}(2) \\ \vdots & \vdots & \ddots & \vdots \\ \breve{b}^p_{i,1}(p) & \breve{b}^p_{i,2}(p) & \cdots & \breve{b}^p_{i,N}(p) \end{pmatrix}$$

Each column of the above matrix represents an edge. The suffix j correspond to the influencing process, while k is the time lag. The parameters $\{\breve{b}^p_{i,j}(k)\}$ minimize the mean squared error $\mathbb{E}[(X_i(n) - \breve{X}^p_i(n))^2]$ for each i, under the following condition of sparsity: for any i, the parameters $\breve{b}^p_{i,j}(k) = 0$ for all k, for most js.

Ideally, conditions of sparsity can be achieved by restricting the number of non-zero parameters of the predictor, i.e., the size of the support set of the parameters, often loosely termed as the "ℓ_0 norm" of the parameters. But a constraint on the support set of the parameters is non-convex and difficult to implement. A more tractable alternative is to use a constraint on the ℓ_1 *norm* of the parameters, i.e., their absolute sum [120, 121, 122].

Suppose, the goal is to find a set of optimal parameters $\{a_1 \dots a_N\}$ that minimizes the objective function $f(a_1 \dots a_N)$, subject to the condition that the support of the a_is is small, i.e.,

$$\sum_{i=1}^{N} \mathbb{1}_{\{a_i \neq 0\}} \leq C < N$$

where $\mathbb{1}$ denotes the indicator function. This problem can be reformulated to set the constraint as

$$\sum_{i=1}^{N} |a_i| \leq C_1$$

If the function $f(\cdot)$ is convex, this is now a convex optimization problem which can be addressed readily.

A popular technique that uses ℓ_1 constraints to restrict the number of non-zero parameters in the MMSE predictor is known as the "least absolute shrinkage and selection operator", abbreviated as lasso ([88]). Essentially, the method determines a linear estimator by minimizing the residual sum of squares, subject to the absolute sum of the parameters being bounded by a constant.

Suppose the data consists of $\{Z(n), \mathbf{Y}(n)\}_{n=1\dots T}$, where $\mathbf{Y}(n) = [Y_1(n) \dots Y_K(n)]^T$. Without loss of generality, let the sample means be 0, i.e.,

$$\frac{1}{T} \sum_{n=1}^{T} Z(n) = 0$$

and

$$\frac{1}{T}\sum_{n=1}^{T} Y_i(n) = 0 \text{ for } i = 1, \ldots, K$$

The lasso estimates $\mathbf{b} = [b_1 \ \ldots \ b_K]^T$ are given by

$$\mathbf{b} = \arg\min \left\{ \sum_{n=1}^{T} \left(Z(n) - \sum_{i=1}^{K} b_i Y_i(n) \right)^2 \right\} \text{ subject to } \sum_{i=1}^{K} |b_i| \leq \theta \tag{7.2.2}$$

Note that the residual sum of squares $\sum_{n=1}^{T} \left(Z(n) - \sum_{i=1}^{K} b_i Y_i(n) \right)$ is an estimator of the mean squared error $\mathbb{E}\left[\left(Z(n) - \sum_{i=1}^{K} b_i Y_i(n) \right)^2 \right]$. The lasso estimates are thus the MMSE parameters under an ℓ_1 constraint. The optimal parameters are often computed by solving an unconstrained minimization problem where $\sum_{i=1}^{K} |b_i|$ is the penalty function:

$$\mathbf{b} = \arg\min \left\{ \sum_{n=1}^{T} \left(Z(n) - \sum_{i=1}^{K} b_i Y_i(n) \right)^2 + \lambda \sum_{i=1}^{K} |b_i| \right\}$$

7.3 The lasso and group lasso methods in the context of multivariate AR models

In the simplest case, when $p = 1$, the objective is to estimate $X_i(n)$ (for any i) as a multivariate AR-1 model, with a restriction on the number of non-zero parameters. In this model, each parameter represents an individual edge, and therefore, restricting the number of edges is equivalent to restricting the number of non-zero (or significant) parameters. The constraint on the support of the set of parameters can be replaced by a constraint on their absolute sum. The problem, then, is identical to that formulated in (7.2.2) and the optimal parameters $\breve{\mathbf{b}}_i^1 = [\breve{b}_{i,1}^1(1) \ \ldots \ \breve{b}_{i,N}^1(1)]^T$ are given by

$$\breve{\mathbf{b}}_i^1 = \arg\min \left\{ \sum_{n=1}^{T-1} \left(X_i(n) - \sum_{j=1}^{N} \breve{b}_{i,j}^1(1) X_j(n-1) \right)^2 \right\} \text{ subject to } \sum_{j=1}^{N} |\breve{b}_{i,j}^1(1)| \leq \theta$$

Even though the lasso method serves the purpose in the AR-1 case, the scenario is more complicated for $p > 1$. Recall that the objective function in this case is

$$\sum_{n=1}^{T-p} \left(X_i(n) - \sum_{j=1}^{N} \sum_{k=1}^{p} \breve{b}_{i,j}^{p}(k) X_j(n-k) \right)^2$$

Extending the lasso approach, one can choose to replace the terms $|\breve{b}_{i,j}^{1}(1)|$ in the constraint conditions of the AR-1 model with the absolute sums along the corresponding edges, i.e.,

$$\sum_{k=1}^{p} |\breve{b}_{i,j}^{p}(k)|$$

However, this approach does not distinguish among parameters in different edges. The constraint

$$\sum_{j=1}^{N} \sum_{k=1}^{p} |\breve{b}_{i,j}^{p}(k)| \leq \theta$$

puts a bound on the sum of all the parameters in all edges as a whole. While this translates to a reduced number of significant parameters, it does not implement a restriction on the number of edges. Indeed, it is quite possible that in spite of the lasso approach reducing many of the Np parameters close to zero, the number of edges remain significantly large. For instance, consider the case where for some i,

$$\breve{b}_{i,j}^{p}(k) = \begin{cases} \rho \text{ for } k = 1 \text{ for } j = 1, \ldots, N \\ = 0 \text{ otherwise} \end{cases}$$

In the above example, the matrix $\breve{B}_i^p$ has few non-zero elements and the absolute sum of all the elements is low; and yet the process $\{X_i(n)\}$ depends on the past values of *all* the N processes in the system. The ℓ_1 norm, then, is not a suitable penalty function to use in this context.

To successfully implement a method that would reduce the number of edges, it is necessary to *decouple* the parameters associated with different edges, while grouping together parameters belonging to the same edge. Instead of using a constraint that restricts *all* the parameters together, then, it may be possible to restrict parameters within individual edges.

To address the limitation in the lasso method, the group lasso (glasso) method was proposed in [90], which incorporates the above idea. In this technique, the objective

function is minimized using the parameters in groups. The penalty function is defined as the absolute sum of the ℓ_2 norm of each individual group. In the context of our problem, the optimal parameters in the group lasso method are given by

$$\breve{\mathbf{b}}_i^1 = \arg\min \left\{ \sum_{n=1}^{T-p} \left(X_i(n) - \sum_{j=1}^{N} \sum_{k=1}^{p} \breve{b}_{i,j}^p(k) X_j(n-k) \right)^2 + \lambda \sum_{j=1}^{N} \sqrt{\sum_{k=1}^{p} |\breve{b}_{i,j}^p(k)|^2} \right\}$$

This method successfully decouples parameters based on their corresponding groups. The technique is particularly useful when each group of parameters is expected to be either all zero or all non-zero.

7.4 A new method for detecting causality under sparsity constraints

In this section, we propose a novel method that can achieve the goal of restricting the number of edges in the graphical representation of a system of several time series.

Recall that the p-th order multivariate AR estimator of $X_i(n)$ is given by

$$\breve{X}_{i,p}(n) = \sum_{k=1}^{p} \sum_{j=1}^{N} \breve{b}_{i,j}^p(k) X_j(n-k)$$

Taking Z-transforms on both sides,

$$\begin{aligned} \breve{X}_{i,p}(z) &= \sum_{j=1}^{N} \sum_{k=1}^{p} \breve{b}_{i,j}^p(k) z^{-k} X_j(z) \\ &= \sum_{j=1}^{N} \breve{b}_{i,j}^p(z) X_j(z) \end{aligned}$$

where

$$\breve{b}_{i,j}^p(z) = \left(\sum_{k=1}^{p} \breve{b}_{i,j}^p(k) z^{-k} \right)$$

$\breve{b}^p_{i,j}(z)$ gives the z-domain representation of the parameters corresponding to the edge from node j to i. Motivated by this inherent structure, we define the following function $g_p(\cdot) : \mathbb{R}^p \to \mathbb{R}$ as follows. For $\mathbf{c} = [c(1) \ \dots \ c(p)]^T$, let

$$g_p(\mathbf{c}) = \int_{-\frac{1}{2}}^{\frac{1}{2}} \left| \sum_{k=1}^{p} c(k) e^{-2\pi\imath\lambda k} \right| d\lambda \tag{7.4.3}$$

We drop the suffix $_p$ when there is no scope of confusion.

Lemma 7.4.1.

$$g(\mathbf{c}) = \int_{-\frac{1}{2}}^{\frac{1}{2}} \sqrt{\sum_{k=1}^{p} c^2(k) + 2 \sum_{\substack{j=1 \\ j>k}}^{p} \sum_{k=1}^{p} c(k)c(j) \left(cos(2\pi\lambda(k-j))\right)} d\lambda$$

Proof. The result is derived through a simplification of the expression of (7.4.3).

$$\begin{aligned}
g(\mathbf{c}) &= \int_{-\frac{1}{2}}^{\frac{1}{2}} \left| \sum_{k=1}^{p} c(k) e^{-2\pi\imath\lambda k} \right| d\lambda \\
&= \int_{-\frac{1}{2}}^{\frac{1}{2}} \sqrt{\left(\sum_{k=1}^{p} c(k) cos(2\pi\lambda k) \right)^2 + \left(\sum_{k=1}^{p} c(k) sin(2\pi\lambda k) \right)^2} d\lambda \\
&= \int_{-\frac{1}{2}}^{\frac{1}{2}} \Bigg(\sum_{k=1}^{p} c^2(k)(cos^2(2\pi\lambda k) + sin^2(2\pi\lambda k)) \\
&\quad + 2 \sum_{\substack{j=1 \\ j>k}}^{p} \sum_{k=1}^{p} c(k)c(j) \left(cos(2\pi\lambda k) cos(2\pi\lambda j) + sin(2\pi\lambda k) sin(2\pi\lambda j)\right) \Bigg)^{1/2} d\lambda \\
&= \int_{-\frac{1}{2}}^{\frac{1}{2}} \sqrt{\sum_{k=1}^{p} c^2(k) + 2 \sum_{\substack{j=1 \\ j>k}}^{p} \sum_{k=1}^{p} c(k)c(j) \left(cos(2\pi\lambda(k-j))\right)} d\lambda
\end{aligned}$$

□

It is interesting to note that the above expression is very similar to the ℓ_2 norm of the vector $\mathbf{c}$ used in the glasso technique. The difference is due to the additional terms of the form $c(k)c(j)cos(2\pi\lambda(k-j))$. Unlike the glasso method, however, this technique inherits the temporal structure of the lag parameters through its formulation in the frequency domain.

Let $\breve{\mathbf{b}}^p_{i,j} = [\breve{b}^p_{i,j}(1) \dots \breve{b}^p_{i,j}(p)]^T$. In the context of our problem, $g(\breve{\mathbf{b}}^p_{i,j})$ can be interpreted as a measure of how "strong" the edge is from node j to node i. A bound on the absolute sum $\sum_{j=1}^{N} g(\breve{\mathbf{b}}^p_{i,j})$ forces some of the less significant edges entering node i to vanish, while parameters on different edges remain decoupled.

Lemma 7.4.2. *$g(\cdot) : \mathbb{R}^p \to \mathbb{R}$ is a convex function.*

Proof. Let $t \in [0,1]$. Let $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^p$, $\boldsymbol{a} = [a(1) \dots a(p)]^T$, $\boldsymbol{b} = [b(1) \dots b(p)]^T$. For all $\lambda \in \left(-\frac{1}{2}, \frac{1}{2}\right]$

$$\begin{aligned} & \left| \sum_{k=1}^{p} t a(k) e^{-2\pi\imath\lambda k} + \sum_{k=1}^{p} (1-t) b(k) e^{-2\pi\imath\lambda k} \right| \\ \leq & \left| \sum_{k=1}^{p} t a(k) e^{-2\pi\imath\lambda k} \right| + \left| \sum_{k=1}^{p} (1-t) b(k) e^{-2\pi\imath\lambda k} \right| \\ = & \; t \left| \sum_{k=1}^{p} a(k) e^{-2\pi\imath\lambda k} \right| + (1-t) \left| \sum_{k=1}^{p} b(k) e^{-2\pi\imath\lambda k} \right| \end{aligned}$$

Integrating both sides over $\lambda \in \left(-\frac{1}{2}, \frac{1}{2}\right]$, we obtain

$$g(t\boldsymbol{a} + (1-t)\boldsymbol{b}) \leq t g(\boldsymbol{a}) + (1-t) g(\boldsymbol{b})$$

Thus, $g(\cdot)$ is convex. □

It follows that minimization of the estimated error using $g(\cdot)$ as a penalty function is a convex optimization problem and therefore tractable.

For the multivariate AR-p estimator, we propose the use of

$$G_i = \sum_{j=1}^{N} g(\breve{\mathbf{b}}^p_{i,j})$$

as the penalty function, where $\breve{\mathbf{b}}_{i,j}^p$ denotes the jth column of $\breve{B}_i^p$. The optimal parameters are given by

$$\breve{B}_i^p = \arg\min \left\{ \sum_{n=1}^{T-p} \left(X_i(n) - \sum_{k=1}^{p} \sum_{j=1}^{N} \breve{b}_{i,j}^p(k) X_j(n-k) \right)^2 \right\} \text{ subject to } \sum_{j=1}^{N} g(\breve{\mathbf{b}}_{i,j}^p) \leq \theta$$

The parameters can be computed by solving the following unconstrained optimization problem:

$$\breve{B}_i^p = \arg\min \left\{ \sum_{n=1}^{T-p} \left(X_i(n) - \sum_{k=1}^{p} \sum_{j=1}^{N} \breve{b}_{i,j}^p(k) X_j(n-k) \right)^2 + \lambda \sum_{j=1}^{N} g(\breve{\mathbf{b}}_{i,j}^p) \right\}$$

Finally, the existence of edges is determined as follows. For each pair of nodes i, j if $G_i < K$ for some threshold K, we conclude that the causal effect of node j on node i is not significant, or there is no edge from j to i.

7.5 Comparison with the glasso method: A simple example

In this section we compare the efficacy of the proposed method to achieve the objective of having fewer edges, with that of the glasso method for a system with $N = p$ nodes. We consider the following cases.

1. All the edges entering node i have exactly one non-zero parameter.

$$\breve{\mathbf{b}}_{i,j}^p = [c\ 0 \ldots\ 0]^T \text{ for } j = 1, 2, \ldots, p$$

2. Two of the edges entering node i have non-zero parameters, which are evenly distributed between the two edges, as indicated below.

$$\breve{\mathbf{b}}_{i,j}^p = \begin{cases} \left[c\ c\ \ldots \left(\left\lfloor \frac{p}{2} \right\rfloor \text{ times}\right)\ 0\ 0\ \ldots \left(p - \left\lfloor \frac{p}{2} \right\rfloor \text{ times}\right)\right]^T \\ \text{for } j = 1 \\ \left[c\ c\ \ldots \left(p - \left\lfloor \frac{p}{2} \right\rfloor \text{ times}\right)\ 0\ 0\ \ldots \left(\left\lfloor \frac{p}{2} \right\rfloor \text{ times}\right)\right]^T \\ \text{for } j = 2 \\ [0\ 0\ \ldots\ 0]^T \text{ for } 2 < j \leq p \end{cases}$$

3. Only one edge entering node i has non-zero parameters.

$$\breve{\mathbf{b}}_{i,j}^{p} = \begin{cases} [c\ c\ \ldots\ c]^T \text{ for } j = 1 \\ [0\ 0\ \ldots\ 0]^T \text{ for } 1 < j \leq p \end{cases}$$

For all the above cases, the lasso penalty function, i.e., the absolute sum of the parameters is the same, viz. $p|c|$. From the perspective of having fewer edges, however, the penalty function should be the greatest for the first case and the least for the third.

The values of $\sum_{i=1}^{N} g(\breve{\mathbf{b}}_{i,j}^{p})$ are plotted for the above three cases in figure 7.1 by varying the order of autoregression (and the number of nodes) p from 1 to 15, with $c = 1$. For comparison, the corresponding plots for the glasso penalty function $\sum_{j=1}^{N} ||\breve{\mathbf{b}}_{i,j}^{p}||_2$ are also presented.

It is seen that for both methods, the penalty is the highest for the first case, where the number of input edges is the maximum possible (N), and is the least for the third case where there is only one input edge.

For case 1, where each of the p edges has a single parameter with the same value c, the penalty function for both the methods is simply $p|c|$, the sum of the individual parameters. For the other two cases, the penalty functions corresponding to the two approaches reflect similar characteristics. For case 2, where the same number of non-zero parameters (with the same absolute sum) is distributed evenly between two edges, the penalty function is lower than that for case 1. Finally, for the third case, where all non-zero parameters appear on a single edge, the penalty function is significantly lower than the first two cases. The difference in the penalty functions for cases 2 and 3 are seen to be comparable for the two methods.

This example attests that like the glasso penalty function, the proposed penalty function is an appropriate choice when the objective is to restrict the number of causal connections. In both methods, a higher number of edges is aptly penalized.

It is observed that for both cases 2 and 3, the glasso penalty increments more significantly with the number of non-zero parameters within an edge, compared to the proposed alternative. The latter flattens out eventually, indicating insensitivity to higher model orders. As seen from the plot, for $p = 15$, the penalty functions corresponding to the all 1s vector of case 3 for the glasso and the proposed method are 3.8730 and 2.087 respectively. For $p = 10,000$, the glasso penalty is 100, while that for the latter is only 4.1956. The penalty function of our proposed method, thus, is more affected by the number of existing *edges* and less by the number of non-zero parameters within the edges.

When a system has a long memory, the present value depends on a large number of past values, i.e., the autoregressive parameters decay slowly with increasing lag, and the model order p is large. In other words, for such systems, the "influential" edges contain a large number of non-zero parameters. As such, when determining the optimal parameters for such processes, the method should be such, that while a higher number of edges is severely penalized, a higher number of parameters along the *same edge* is not penalized significantly. The proposed method satisfies this requirement more tightly than the glasso, and will therefore be more appropriate for detecting causal interdependence relations.

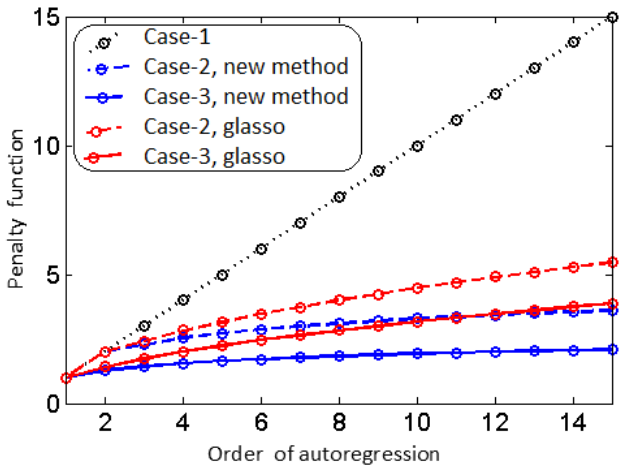

Figure 7.1: Proposed penalty functions for different examples, compared with those for glasso

7.6 Simulation Results

In this section we demonstrate the utility of the proposed method in inferring interdependence relations among several stochastic processes. A system of six Gaussian, zero-mean WSS processes with equal variances and model order $p = 4$ was simulated using $T = 800$ samples for each process. The number of samples T was deliberately chosen to be small so that the interdependence relations among the processes are not clearly detectable through the MVAR approach. Interdependence relations of the original system are graphically represented in Fig. 7.6.

Gradient descent method with line search [123] was used for optimization and a model order $p = 4$ was used. The estimation parameters were first computed by fitting the regular

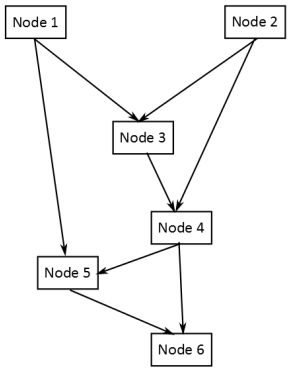

Figure 7.2: A system of six interdependent WSS processes

MVAR model and then they were further optimized using the proposed method. Values of $g(\breve{\mathbf{b}}_{i,j}^{p})$ for the regular MVAR model and the proposed technique are tabulated below in tables 7.1 and 7.2 respectively. The corresponding values of $||\breve{\mathbf{b}}_{i,j}^{p}||_2$ for the glasso method are presented in tables 7.3 and 7.4. The edges that exist in the original system are depicted in bold font.

Table 7.1: $g(\breve{\mathbf{b}}_{i,j}^{p})$ computed from the MVAR parameters estimated directly

	j=1	j=2	j=3	j=4	j=5	j=6
i=1	**0.7611**	0.1444	0.1361	0.1737	0.1010	0.2278
i=2	0.0559	**0.5832**	0.0771	0.1370	0.1742	0.0965
i=3	**0.3674**	**0.8408**	0.0301	0.0428	0.0808	0.1609
i=4	0.0188	**0.5071**	**0.2185**	0.0399	0.0556	0.1297
i=5	**0.3323**	0.0587	0.0612	**0.7161**	0.0916	0.0658
i=6	0.0129	0.0384	0.0315	**0.1999**	**0.4758**	0.0601

It is seen that both methods successfully detected the original edges in the system. In table 7.1, where the estimates were computed directly through the method of least squares, due to the limited number of samples, parameters were inaccurate and the original edges were not easily determined. After the implementation of the proposed method, however, the values of $g(\breve{\mathbf{b}}_{i,j}^{p})$ were significantly reduced for the cases where there is no edge along $\overrightarrow{(j,i)}$, while those corresponding to the actual edges were affected only slightly. As a result, in table 7.2, *actual* edges distinctly stood out from the spurious ones, and the original

Table 7.2: $g(\breve{\mathbf{b}}_{i,j}^{p})$ after optimizing through the proposed method

	j=1	j=2	j=3	j=4	j=5	j=6
i=1	**0.6177**	0.0460	0.0109	0.0369	0.0112	0.1153
i=2	0.0281	**0.4630**	0.0234	0.0223	0.0453	0.0098
i=3	**0.2707**	**0.7305**	0.0068	0.0049	0.0042	0.0346
i=4	0.0188	**0.5071**	**0.2185**	0.0399	0.0556	0.1297
i=5	**0.2570**	0.0084	0.0109	**0.6304**	0.0275	0.0101
i=6	0.0129	0.0384	0.0315	**0.1999**	**0.4758**	0.0601

Table 7.3: $||\breve{\mathbf{b}}_{i,j}^{p}||_2$ computed from the MVAR parameters estimated directly

	j=1	j=2	j=3	j=4	j=5	j=6
i=1	**0.5107**	0.0626	0.0261	0.0330	0.0113	0.0195
i=2	0.0125	**0.5326**	0.0162	0.0506	0.1114	0.0754
i=3	**0.3298**	**0.7869**	0.0021	0.0022	0.0111	0.0049
i=4	0.0893	**0.6216**	**0.1485**	0.0595	0.0724	0.0633
i=5	**0.2915**	0.0087	0.0105	**0.7035**	0.0311	0.0212
i=6	0.0334	0.0421	0.0520	**0.2259**	**0.5861**	0.0453

configuration could be easily recovered. The performance of the proposed method is seen to be similar to that of the glasso method in this example.

As a second example, we consider the currency exchange rates used in chapter 5. Fluctuations in daily exchange rates of the currencies of some of the world's leading economies against the Swiss Franc, for the period January 1, 2009 to December 31, 2012, obtained from the Bank of Canada website [113] were used. The data was used to find the interdependence relations among the different conversion rates using our proposed method. The optimal values of $g(\breve{\mathbf{b}}_{i,j}^{p})$ are presented below in table 7.5.

The most significant interdependence relations detected through the proposed method are illustrated in figure 7.3. For comparison, we also present here the interdependence relations indicated by the pairwise Wiener filter based approach from chapter 5 in figure 7.4. It is interesting to note that while the interconnections are slightly different from those in figure 7.4, the basic pattern bears a strong resemblance.

Table 7.4: $||\breve{\mathbf{b}}^p_{i,j}||_2$ computed after optimizing through glasso

	j=1	j=2	j=3	j=4	j=5	j=6
i=1	**0.6820**	0.2790	0.1689	0.2176	0.1939	0.2151
i=2	0.1057	**0.6413**	0.1127	0.1761	0.2677	0.2015
i=3	**0.3987**	**0.8770**	0.0730	0.0619	0.1293	0.1358
i=4	0.0893	**0.6216**	**0.1485**	0.0595	0.0724	0.0633
i=5	**0.3429**	0.0620	0.0367	**0.7482**	0.0715	0.0745
i=6	0.0334	0.0421	0.0520	**0.2259**	**0.5861**	0.0453

Table 7.5: $g(\breve{\mathbf{b}}^p_{i,j})$ for currency conversion rates

	USD	GBP	Euro	Yuan	Yen	INR
US Dollar	1.1608	0.0423	0.1557	0.3874	0.0555	0.0715
Great Britain Pound	0.6959	0.7090	0.1794	0.3049	0.1295	0.0875
Euro	0.5021	0.0389	1.0315	0.2432	0.0476	0.0381
Chinese Yuan	0.8819	0.0715	0.2598	1.0154	0.0320	0.1485
Japanese Yen	1.0380	0.1179	0.2365	0.5786	0.7054	0.0858
Indian Rupee	0.5294	0.0467	0.1225	0.2860	0.0426	0.8537

7.7 Conclusion

In this chapter, we have presented a new method that detects causal interconnections within a group of time series under a constraint that restricts the number of edges. The proposed penalty function is derived using the frequency domain representation of each edge, and can be intuitively interpreted as the "strength" of the edge. It is seen that the expression has an interesting similarity with that of the penalty function corresponding to the glasso method, and differs only in the additional product terms.

Implementation of the method on the simulated example and real data indicates that the technique is comparable to the glasso method in terms of performance. With an appropriate choice of the parameter λ, desired results can be obtained. Our method, however, is more computationally intensive, as it requires the computation of an integration numerically.

However, as indicated in the plots of figure 7.1, the proposed penalty function exhibits an advantage over the glasso method in the context of long memory processes. The former is less sensitive to a higher number of non-zero parameters in the same edge, compared to that of the glasso technique. Due to this property, optimization carried out under a constraint on G_i will preserve the significant edges, without severely affecting the individual parameters within the significant edges.

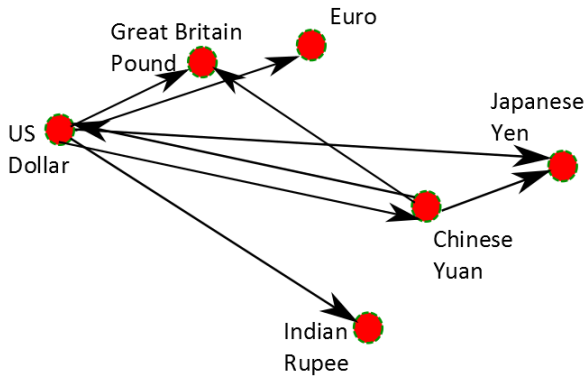

Figure 7.3: Interrelation of currencies inferred using the proposed method ($p = 10$)

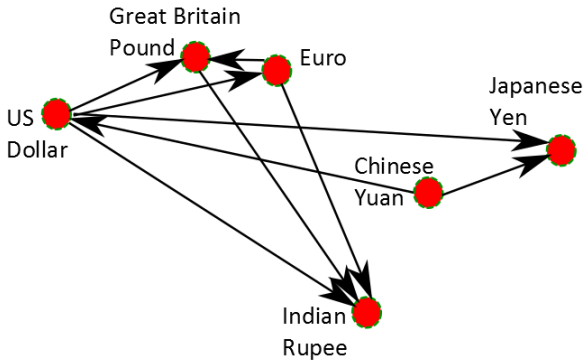

Figure 7.4: Interrelation of currencies inferred using pairwise Wiener filters ($p = 10$), for comparison

Chapter 8

Conclusion

8.1 Summary

In this dissertation, we focused on two problems related to the estimation of time series through linear MMSE approximations. First, we studied the asymptotic behaviour of AR and MA estimates of WSS time series and presented results on the convergence of the spectral density of the approximating sequences. Next, we analyzed the problem of detecting causal interdependence relations within a family of WSS time series.

In chapter 3, it was shown that the spectral density of both the MA and the AR type approximations converge in L_2 when the covariance sequence is summable and when the spectral density is strictly positive. It was also established that under the same conditions, the time average variance constant (TAVC) of a WSS time series converges to that of the infinite order AR approximation of the series. These conditions for convergence are fairly general and are satisfied by a large class of stochastic processes.

Furthermore, in chapter 4, we considered asymptotic behavior of AR approximations when empirical covariances, computed from a sample of size N, are used to estimate the AR parameters in lieu of the true covariance sequence. Under some additional regularity conditions and a mild assumption, a result on the convergence in quadratic mean of the empirical AR parameters was derived when the model order $p = o\{N^{\frac{1}{3}}\}$. The spectral density and the TAVC of the approximating AR sequence were shown to converge under the same conditions.

In chapter 5, we studied the utility of pairwise causal Wiener filters in detecting interdependence relations among several jointly WSS time series. We presented some results

that linked the causal Wiener filter with Granger-causality, a tool that is used to identify causal connections. We also proposed a simple technique that uses the FIR Wiener filter to detect Granger-causality, the performance of which was compared to that of directed information. Our results indicated that while ideally, such interdependence relations should be derived through the simultaneous consideration of all processes involved, pairwise estimation techniques like the Wiener filter can be useful to obtain suboptimal results at low computational costs.

Noting that many processes encountered in practice are non-stationary, in chapter 6, we reviewed the problem of AR approximation of cyclostationary processes and presented a time-invariant AR estimation technique for the same. It was also shown that the former may be extended to develop a time-invariant Wiener filter to detect Granger-causality, the performance of which was demonstrated using climate related data.

Finally, in chapter 7, we considered the case where a multivariate AR model is to be derived for a group of random processes under the condition that each process is influenced by a small number of other processes. A new method was proposed in this regard which is based on the frequency-domain representation of the parameters along each edge. This method was compared to the group lasso method.

8.2 Extensions

Following the results presented in this dissertation, there are several directions in which future research may be pursued. Some of the possible extensions are discussed below.

The estimation of the TAVC, addressed in chapters 3 and 4 arise mainly in the context of steady-state simulation. Often the process being simulated is Markov. We would like to investigate whether there is an easier characterization of the condition of strict-positivity of the spectral density for Markov processes.

In chapter 4, we identified conditions for the spectral density of the approximating AR sequence to converge in mean. More interesting would be to find conditions on p and N for the spectral density to converge in quadratic mean, i.e., in $L_2(\mathbb{P})$. This, however, requires the fourth moment of the sum $(\sum_{k=1}^{p} |b_{k,p,N} - b_{k,p}|)$ to converge, which would necessitate the existence of a higher moment of the innovation sequence and a stronger restriction on p with respect to N. It would also be worthwhile to identify a class of random processes that satisfy the conditions imposed in chapter 4. Furthermore, it would be interesting to see if there can be more relaxed conditions for these convergence results than the ones imposed by us.

An interesting extension of this work would be to find conditions on p and N for the spectral density to converge almost surely when the original sequence is only assumed to be in $L_2(\mathbb{P})$. As discussed in our literature survey, results on almost sure convergence of the AR estimates are available only under the assumption that the associated innovation is Martingale difference. It would be interesting to find out if the same can be derived under more relaxed conditions.

In chapters 5 and 6 it was shown that the pairwise Wiener filter can be useful in gathering reasonably accurate information on the underlying causal structure of a group of WSS or CS processes at low computational cost. An interesting extension would be to somehow quantify the accuracy of this method. For instance, given that the technique detects an edge between two nodes, we would like to know the probability that there is an edge between the corresponding nodes in the original system. The question of convergence is relevant in this context as well. It would be interesting to see under what conditions the pairwise FIR Wiener filters converge to their causal IIR counterparts.

The performance of directed information and the Wiener filter were seen to be comparable in detecting Granger-causality in our simulations for the Gaussian example. Although the former has gained considerable popularity in this area, its computation involves the estimation of conditional probability density functions, which is, in general, not easy. It would be of interest to investigate whether directed information can be estimated under more general settings, and then compare the two techniques under such relaxed conditions.

In chapter 7, we proposed a new method to determine causal connections in a family of time series where a restriction on the number of edges is imposed through a penalty function that represents the "strength" of edges. While it is seen that the new method has some similarity with the group lasso technique, it would be interesting to find an analytical relation between the two penalty functions. This would provide more insight on which technique to choose depending on the estimation problem at hand. We would also like to identify fast, efficient algorithms to solve the optimization problem we formulated.

The results presented in chapters 5, 6 and 7 chiefly deal with applications, and as such, we would like to see how these methods perform in various practical scenarios. Through extensive experimentation with real data, these techniques may be modified and perfected in the future in accordance with the specific field of application.

References

[1] H. Damerdji, S. G. Henderson, and P. W. Glynn. Computational efficiency evaluation in output analysis. In *Proceedings of the 29th conference on Winter simulation*, pages 208–215, 1997.

[2] W. B. Wu. Recursive estimation of time-average variance constants. *Annals of Applied Probability*, 19(4):1529–1552, 2009.

[3] S. Asmussen and P.W. Glynn. *Stochastic Simulation*. Springer, 2007.

[4] B. Picinbono. *Random Signals and Systems*. Prentice Hall, 1993.

[5] P. Bremaud. *Mathematical Principles of Signal Processing*. Springer, 2002.

[6] T. Kailath, A. H. Sayed, and B. Hassibi. *Linear Estimation*. Prentice Hall, 2000.

[7] M.H. Hayes. *Statistical Digital Signal Processing and Modeling*. Wiley, 1996.

[8] C. W. J. Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*, 37(3):424–438, August 1969.

[9] C. W. J. Granger. Testing for causality : a personal viewpoint. *Journal of Economic Dynamics and Control*, 2:329–352, 1980.

[10] J. Geweke. Measurement of linear dependence and feedback between multiple time series. *Journal of the American Statistical Association*, 77:304–313, 1982.

[11] J. Geweke. Measures of conditional linear dependence and feedback between time series. *Journal of the American Statistical Association*, 79:907–915, 1984.

[12] P. Caines. *Linear Stochastic Systems*. John Wiley & Sons, 1988.

[13] W. Rudin. *Real and Complex Analysis*. McGraw-Hill Science/Engineering/Math, 1986.

[14] G.U. Yule. On a method of investigating periodicities in disturbed series, with special reference to Wolfer's sunspot numbers. *Philosophical Transactions of the Royal Society of London, Ser. A*, 226:267–298, 1927.

[15] G. Walker. On periodicity in series of related terms. *Proceedings of the Royal Society of London, Ser. A*, 131:518–532, 1931.

[16] E. Slutsky. The summation of random causes as the source of cyclic processes. *Econometrica*, 5, 1937.

[17] N. Wiener. The extrapolation, interpolation, and smoothing of stationary time series. *Report of the Services 19, Research Project DIC-6037*, 1942.

[18] P.M.T. Broersen. *Automatic Autocorrelation and Spectral Analysis*. Springer, 2006.

[19] S.M. Kay and S.L. Marple Jr. Spectrum analysis; a modern perspective. *Proceedings of the IEEE*, 69(11):1380 – 1419, 1981.

[20] H. Akaike. Fitting autoregressive models for prediction. *Ann. Inst. Statist. Math.*, 21:243–247, 1969.

[21] H. Akaike. Statistical predictor identification. *Ann. Inst. Statist. Math.*, 22:203–207, 1970.

[22] H. Akaike. A new look at the statistical model identification. *Automatic Control, IEEE Transactions on*, 19(6):716 – 723, 1974.

[23] E. Parzen. Multiple time series modelling. *Multivariate Analysis II*, pages 398–410, 1969.

[24] E. Parzen. Solutions to the stationary time series modeling and prediction problem. In *Decision and Control including the 13th Symposium on Adaptive Processes, 1974 IEEE Conference on*, volume 13, pages 468 –473, 1974.

[25] E. Parzen. Some recent advances in time series modeling. *Automatic Control, IEEE Transactions on*, 19(6):723 – 730, 1974.

[26] R. J. Bhansali. The criterion autoregressive transfer function of Parzen. *Journal of Time Series Analysis*, 7(2):79–104, 1986.

[27] E. J. Hannan and B. G. Quinn. The determination of the order of an autoregression. *Journal of the Royal Statistical Society*, 41:190195, 1979.

[28] G. E. Schwarz. Estimating the dimension of a model. *Annals of Statistics*, 6(2):461–464, 1978.

[29] J. Rissanen. Modeling by shortest data description. *Automatica*, 14:445–471, 1978.

[30] N. Beamish and M. B. Priestley. A study of autoregressive and window spectral estimation. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 30(1):41–58, 1981.

[31] O.I. Shittu and M.J. Asemota. Comparison of criteria for estimating the order of autoregressive process: A monte carlo approach. *European Journal of Scientific Research*, 30(3):409–416, 2009.

[32] S. Khorshidi and M. Karimi. Finite sample FPE and AIC criteria for autoregressive model order selection using same-realization predictions. *EURASIP Journal on Advances in Signal Processing*, 2009.

[33] S. Khorshidi and M. Karimi. Modified AIC and FPE criteria for autoregressive (AR) model order selection by using LSFB estimation method. In *Advances in Computational Tools for Engineering Applications, 2009. ACTEA '09. International Conference on*, pages 374 –379, 2009.

[34] R. Shibata. Asymptotically efficient selection of the order of the model for estimating parameters of a linear process. *Annals of Statistics*, 8(1):147–164, 1980.

[35] C. Ing and Wei C. On same-realization prediction in an infinite-order autoregressive process. *Journal of Multivariate Analysis*, 85(1):130–155, 2003.

[36] C. Ing and Wei C. Order selection for same-realization predictions in autoregressive processes. *Annals of Statistics*, 33(5):2423–2474, 2005.

[37] S. Degerine. Partial autocorrelation function for a scalar stationary discrete-time process. In *Alternative Approaches to Time Series Analysis, Proceedings of the 3rd Franco-Belgian Metting of Statisticians*, pages 79–94, 1982.

[38] G. Baxter. An asymptotic result for the finite predictor. *Math. Scand.*, 10:137–144, 1962.

[39] K. N. Berk. Consistent autoregressive spectral estimates. *Annals of Statistics*, 2(3):489–502, 1974.

[40] M. Pourahmadi. On the convergence of finite linear predictors of stationary processes. *Journal of Multivariate Analysis*, 30(2):167–180, 1989.

[41] R. Cheng and M. Pourahmadi. Baxter's inequality and convergence of finite predictors of multivariate stochastic processess. *Probability Theory and Related Fields*, 95:115–124, 1993.

[42] M. Pourahmadi. *Foundations of Time Series Analysis and Prediction Theory*. John Wiley, 2001.

[43] D. F. Findley. Convergence of finite multistep predictors from incorrect models and its role in model selection. *Note di Matematica*, XI:145–155, 1991.

[44] D. S. Poskitt. A note on autoregressive modeling. *Econometric Theory*, 10(5):884–899, 1994.

[45] R. J. Bhansali. Effects of not knowing the order of an autoregressive process. *Journal of the American Statistical Association*, 76(375):588–597, 1981.

[46] A. Hong-Zhi, C. Zhao-Guo, and E. J. Hannan. Autocorrelation, autoregression and autoregressive approximation. *Annals of Statistics*, 10(3):926–936, 1982.

[47] Y. Gel and A. Barabanov. Banded regularization of autocovariance matrices in application to parameter estimation and forecasting of time series. *Journal of Statistical Planning and Inference*, 137:1260–1277, 2007.

[48] W. B. Wu and M. Pourahmadi. Banding sample autocovariance matrices of stationary processes. *Statist. Sin.*, 19:1755–1768, 2009.

[49] P. J. Bickel and Y. R. Gel. Banded regularization of autocovariance matrices in application to parameter estimation and forecasting of time series. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(5):711–728, 2011.

[50] A. Barrat, M. Barthelemy, and A Vespigniani. *Dynamics on complex networks*. Cambridge University Press, 2008.

[51] M. E. J. Newman. The structure and function of complex networks. *SIAM reviews*, 43:167–256, 2003.

[52] S. Boccaletti, V. Latora, M. Chavez, and D. U. Hwang. Complex networks: Structure and dynamics. *Physics Reports*, 424(22):175–308, 2006.

[53] P. Amblard and O. Michel. Relating Granger causality to directed information theory for networks of stochastic processes. *Arxiv preprint arXiv:0911.2873*, 2009.

[54] P. Amblard and O. Michel. On directed information theory and Granger causality graphs. *Journal of computational neuroscience*, 30:7–16, 2009.

[55] S.H. Strogatz. Exploring complex networks. *Nature (London)*, 410:268–276, 2001.

[56] I. Stewart. Networking opportunity. *Nature (London)*, 427:601–604, 2004.

[57] M. Timme. Revealing network connectivity from response dynamics. *Phys. Rev. Lett.*, 98(22), 2007.

[58] D. Materassi and G. Innocenti. Topological identification in networks of dynamical systems. *Automatic Control, IEEE Transactions on*, 55(8):1860–1871, 2010.

[59] D. Materassi and M.V. Salapaka. On the problem of reconstructing an unknown topology via locality properties of the Wiener filter. *Automatic Control, IEEE Transactions on*, 57(7):1765–1777, 2012.

[60] D. Napoletani and T.D. Sauer. Reconstructing the topology of sparsely connected dynamical networks. *Phys. Rev. E*, 77(2), 2008.

[61] C. D. Michener and R. R. Sokal. A quantitative approach to a problem in classification. *Evolution*, 11(2), 1957.

[62] R.P. Freckleton, P.H. Harvey, and Pagel M. Phylogenetic analysis and comparative data: a test and review of evidence. *American Naturalist*, 160:712–726, 2002.

[63] J. Felsenstein. Phylogenies and the comparative method. *American Naturalist*, 125:1–15, 1985.

[64] T. Jr. Garland, A.W. Dickerman, C.M. Janis, and J.A. Jones. Phylogenetic analysis of covariance by computer simulation. *Systematic Biology*, 42:265–292, 1985.

[65] R.N. Mantegna and H. Stanley. *An Introduction to Econophysics: Correlations and Complexity in Finance*. Cambridge University Press, 2000.

[66] R.N. Mantegna. Hierarchical structure in financial markets. *Eur. Phys. J. B.*, 11:193–197, 1999.

[67] O. Ledoit and M. Wolf. Honey, I shrunk the sample covariance matrix. *Journal of. Portfolio Management*, pages 110–119, 2004.

[68] B. Toth and J. Kertesz. Accurate estimator of correlations between asynchronous signals. *Physica A*, pages 1696–1705, 2009.

[69] M. Tumminello, F. Lillo, and R.N. Mantenga. Shrinkage and spectral filtering of correlation matrices: a comparison via the Kullback-Leibler distance. *Acta Physica Polonica B*, pages 4079–4088, 2008.

[70] C. W. J. Granger. Some recent developments in a concept of causality. *Journal of Econometrics*, 39:199–211, 1988.

[71] A. C. Sims. Money, income, and causality. *The American Economic Review*, 62(4):540–552, September 1972.

[72] Y. Liu and M. T. Bahadori. A survey on Granger causality: A computational view. *in preparation*, 2012.

[73] S. Boccaletti, M. Ivanchenko, V. Latora, A. Pluchino, and A. Rapisarda. Detecting complex network modularity by dynamical clustering. *Phys. Rev. E*, 75, 2007.

[74] S.L. Bressler and A.K. Seth. Wiener-Granger causality: A well established methodology. *NeuroImage*, 58(2):323–329, 2011.

[75] T.M. Cover and J.A. Thomas. *Elements of Information Theory*. John Wiley & Sons, 2006.

[76] H. Marko. The bidirectional communication theory – a generalization of information theory. *Communications, IEEE Transactions on*, 21(12):1345 – 1351, dec 1973.

[77] James L. Massey. Causality, Feedback, and Directed Information. In *International Symposium on Information Theory and its Applications (ISITA-90)*, pages 303–205, November 1990.

[78] T. Schreiber. Measuring information transfer. *Physics Review Letters*, 85(2):461–465, 2000.

[79] A. Kaiser and T. Schreiber. Information transfer in continuous processes. *Physica D*, 166:43–62, 2002.

[80] T. Weissman, Y. Kim, and H.H. Permuter. Directed information, causal estimation, and communication in continuous time. *submitted to IEEE Transactions on Information theory*, 2011.

[81] H.H. Permuter, Y. Kim, and T. Weissman. Interpretations of directed information in portfolio theory, data compression, and hypothesis testing. *Information Theory, IEEE Transactions on*, 57(6):3248–3259, june 2011.

[82] C.J. Quinn, T.P. Coleman, N. Kiyavash, and N.G. Hatsopoulos. Estimating the directed information to infer causal relationships in ensemble neural spike train recordings. *Journal of Computational Neuroscience*, 2011.

[83] L. Zhao, H. Permuter, Y. Kim, and T. Weissman. Universal estimation of directed information. In *Information Theory Proceedings (ISIT), 2010 IEEE International Symposium on*, pages 1433 –1437, june 2010.

[84] L. Barnett, A. B. Barrett, and A. K. Seth. Granger causality and transfer entropy are equivalent for gaussian variables. *Phys. Rev. Lett.*, 103, Dec 2009.

[85] C.J. Quinn, N. Kiyavash, and T.P. Coleman. Efficient methods to compute optimal tree approximations of directed information graphs. *Signal Processing, IEEE Transactions on*, 61(12):3173–3182, 2013.

[86] D. Materassi and G. Innocenti. Unveiling the connectivity structure of financial networks via high-frequency analysis. *Physica A: Statistical Mechanics and its Applications*, 388(18):38663878, June 2009.

[87] D. Materassi, G. Innocenti, L. Giarr, and M.V. Salapaka. Model identification of a network as compressing sensing. *submitted to System and Control letters*, 2011.

[88] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, 58:267–288, 1994.

[89] H. Zou. The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101:1418–1429, 2006.

[90] M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67, 2006.

[91] J. Friedman, T. Hastie, and R. Tibshirani. A note on the group lasso and a sparse group lasso. Technical Report arXiv:1001.0736, Jan 2010.

[92] A. K. Bolstad, B. D. Van Veen, and R. Nowak. Causal network inference via group sparse regularization. *IEEE Transactions on Signal Processing*, 59(6):2628–2641, 2011.

[93] A. Shojaie and G. Michailidis. Discovering graphical Granger causality using the truncating lasso penalty. *Bioinformatics*, 26:i517–i523, 2010.

[94] A. Arnold, Y. Liu, and N. Abe. Temporal causal modeling with graphical Granger methods. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '07, pages 66–75, 2007.

[95] A. C. Lozano, N. Abe, Yan Liu, and S. Rosset. Grouped graphical Granger modeling methods for temporal causal modeling. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 577–586. ACM, 2009.

[96] S. Haufe, K. Mller, G. Nolte, and N. Krmer. Sparse causal discovery in multivariate time series. *Journal of Machine Learning Research - Proceedings Track*, 6:97–106, 2010.

[97] S. Datta Gupta, R. R. Mazumdar, and P. W. Glynn. On the asymptotic behavior of the spectral density of autoregressive estimates. In *49th Annual Allerton Conference on Communication, Control and Computing*, 2011.

[98] S. Datta Gupta and R. R. Mazumdar. On the convergence of the spectral density of autoregressive approximations via empirical covariance estimates. In *46th Annual Conference on Information Sciences and Systems*, 2012.

[99] S. Datta Gupta, R. Mazumdar, and P. Glynn. On the convergence of the spectrum of finite order approximations of stationary time series. *Journal of Multivariate Analysis*, 121:1–21, October 2013.

[100] S. Datta Gupta and R. R. Mazumdar. Inferring causality in networks of wss time series by pairwise estimation methods. In *Proceedings of the 2013 Information Theory and Applications Workshop (ITA 2013)*, 2013.

[101] A. N. Shiryaev. *Probability*. Springer, 1996.

[102] I.A. Ibragimov and J.V. Linnik. *Independent and stationary sequences of random variables*. Groningen, Wolters-Noordhoff, 1971.

[103] B.D. Craven. The spectral density of a Markov process. *Journal of Applied Probability*, 10(3):52–527, 1973.

[104] D. R. Brillinger. *Time Series Data Analysis and Theory*. Holt, Rinehart and Winston, Inc., 1975.

[105] C. C. Heyde. *Quasi-Likelihood and its Application: A General Approach to Optimal Parameter Estimation*. Springer, 1997.

[106] R. J. Bhansali and P. S. Kokoszka. Prediction of long memory time series: An overview. *Estadistica*, 53:41–96, 2001.

[107] E. J. Hannan. *Time Series Analysis*. Springer, 1967.

[108] G. Szego and U. Grenander. *Toeplitz Forms and Their Applications*. University of California Press, 1958.

[109] D. S. G. Pollock. *A handbook of time-series analysis, signal processing and dynamics, Volume 1*. Academic Press, 1999.

[110] J. Schur. ber potenzreihen, die im innern des einheitskreises beschrnkt sind. *Journal für die reine und angewandte Mathematik (Crelle's Journal)*, 1917:205–232, 1917.

[111] S. Elaydi. *An Introduction to Difference Equations*. Springer, 2005.

[112] G. Kramer. Directed information for channels with feedback. *Ph.D. dissertation, Swiss Fed. Inst. Technol. (ETH), Zurich, Switzerland*, 1998.

[113] Bank of Canada. Exchange rates - Bank of Canada. http://www.bankofcanada.ca/rates/exchange/, 2013.

[114] W.A. Gardner and L. Franks. Characterization of cyclostationary random signal processes. *Information Theory, IEEE Transactions on*, 21(1):4–14, 1975.

[115] E. Gladyshev. Periodically correlated random sequences. *Soviet Math. Dokl.*, pages 385–388, 1961.

[116] M. Pagano. On periodic and multiple autoregressions. *Annals of Statistics*, 6(6):1310–1317, 1978.

[117] H. R. Jones and W. M. Brelsford. Time series with periodic structure. *Biometrika*, 54(3/4):403–408, 1967.

[118] B. M. Troutman. Some results in periodic autoregression. *Biometrika*, 66(2):219–228, 1979.

[119] Government of Canada. Climate. http://climate.weather.gc.ca, 2013.

[120] E.J. Candes and T. Tao. Decoding by linear programming. *Information Theory, IEEE Transactions on*, 51(12):4203–4215, 2005.

[121] E.J. Candes, J. K. Romberg, and T. Tao. Stable signal recovery from incomplete and inaccurate measurements. *Comm. Pure Appl. Math*, 59:1207–1223, 2006.

[122] C. Ramirez, V. Kreinovich, and M. Argaez. Why l_1 is a good approximation to l_0: A geometric explanation. *Journal of Uncertain Systems*, 7, 2013.

[123] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, New York, NY, USA, 2004.