NEW MODELS AND METHODS FOR TIME SERIES ANALYSIS IN BIG DATA ERA

BY XIALU LIU

A dissertation submitted to the

Graduate School—New Brunswick

Rutgers, The State University of New Jersey

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

Graduate Program in Statistics and Biostatistics

Written under the direction of

Rong Chen and Han Xiao

and approved by

_____

_____

_____

_____

New Brunswick, New Jersey

May, 2015

ABSTRACT OF THE DISSERTATION

New Models and Methods for Time Series Analysis in Big Data Era

by Xialu Liu

Dissertation Director: Rong Chen and Han Xiao

In big data era, available information becomes massive and complex and is often observed over time. Conventional time series models are limited in capability of dealing with these type of data. This dissertation focuses on developing new statistical models, along with their associated estimation procedures, to analyze time series data in functional form, and in high dimension, with linear or nonlinear dynamics, which can be broadly applicable to finance, environment, engineering, biological and medical sciences.

Functional data analysis has became an increasingly popular class of problems in statistical research. However, functional data observed over time with serial dependence remains a less studied area. Motivated by Bosq (2000), who first introduced the functional autoregressive (FAR) models, we propose a convolutional functional autoregressive (CFAR) model, where the function at time t is a result of the sum of convolutions of the past functions with a set of convolution functions, plus a noise process, mimicking the autoregressive process. It provides an intuitive and direct interpretation of the dynamics of a stochastic process. We adopt a sieve estimation procedure based on the B-spline approximation of the convolution functions. We establish convergence rate of the proposed estimator, and investigate its theoretical properties. The model building, model validation, and prediction procedures are also developed.

As for high-dimensional time series data, dimension reduction is an important issue and can be effectively performed by factor analysis. Considering the factor impacts may vary under different conditions, we propose a factor model with regime-switching mechanism, allowing loadings to change across regimes, and combined eigendecomposition and Viterbi algorithm for estimation. We discover that, with multiple states of different 'strength', the convergence rate of loading matrix estimator for strong states is the same as the one-regime case, while the rate improves for weak states, gaining extra information from strong states. The theoretical properties of the procedure are investigated as well.

In addition, we propose a new class of nonparametric seasonal time series models under the framework of the functional coefficient model. The coefficients in the proposed model change over time and consist of the trend and seasonal components to characterize seasonality. A local linear approach is developed to estimate the nonparametric trend and seasonal effect functions. The proposed methodologies are illustrated by two simulated examples and the model is applied to characterizing the seasonality of the monthly number of tourists visiting Hawaii.

Acknowledgements

I went through my Ph.D. studies with help and support from many people, which is the most precious treasure discovered during my days at Rutgers.

First I would like to express my gratitude to my advisor Prof. Rong Chen. Without his support and encouragement, this dissertation could not be possible to be finished. He gives me helpful instructions, when I have some difficulties in my thesis; he inspires my confidence, when I feel frustrated by an unsuccessful attempt of my research. Much more than an advisor, he is a trusted friend, a respectable elder, and a great mentor. He not only shows me what to do, but also shows me how to be. His broad knowledge, his continuous pursuit on advances in statistical research, his willingness to help, his devotion to students influence me, and stimulate me to go further. There is an old saying in China that a day as a teacher, a life as a teacher. For all things Prof. Chen taught me, I will not only benefit, but also appreciate throughout my life.

Secondly I would like to extend my deepest gratitude to Prof. Han Xiao and Prof. Qiwei Yao. Prof. Xiao is the junior professor I admire most. He is always able to convey complex ideas in plain language. Every discussion with him makes my understanding on the problems deeper and better. I am grateful to him for offering me 'hand-over-hand' guidance and brilliant ideas. About Prof. Yao, I already heard of his extraordinary dedication in research and teaching from his students, even before I met him. In the last three years, Prof. Yao gave me continuous instructions on my research by countless emails, although we only met several times. I really appreciate his invaluable direction and patience.

I am so lucky to have such three great teachers, and feel honored to have the opportunity to work with them. They help and witness my growth. They make me interested in exploring a career in academia, to be a researcher and teacher who helps

others as they are.

Then I would like to thank my parents for their unconditional love. I probably would have quit hundreds of times without their support. They give me the courage to pursue the life I want.

I also want to say thanks to Prof. Zongwu Cai for his help with my research, and to Prof. John Kolassa for his support in the past five years. Thanks to my fellow Ph.D. students, Xinyan Chen, Heng Shu and Liwei Wang for their company and morale support. I will cherish these pleasant memories and unique friendship we shared. Thanks to Dungang Liu, Wentao Li, Tingni Sun and my friend Lei Lei for their help with my job hunting.

In addition I would like to thank these nights I spent alone in my apartment, struggling, and thinking. They were beautiful.

After this unforgettable journey full of surprise and joy, I am ready for my new adventure.

Dedication

This dissertation is dedicated to people appearing or passing in my life, that make me who I am today.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Time Series Data

Time series refers to a sequence of data points observed over time, and is widely observed in science, engineering, economics and other fields. Examples include counts of sunspots, airline traffic volume and daily closing value of S &P 500 index. Different from cross-sectional data, successive data points are expected to be dependent. Autoregressive (AR) models are the simplest and most natural class of models for scalar time series data, where the process of interest is decomposed into linear combination of its past values and noise. Whitle (1951) introduced a moving average (MA) part and came up with ARMA(autoregressive-moving-average) models in his thesis. Box and Jenkins (1971) expounded an iterative method for model selection and estimation. Quenouille (1957) extended the models to multivariate cases, and the parameter estimation and model specification have been investigated by Tiao and Box (1981), Tsay and Tiao (1983), Lütkepohl (1985), Tiao and Tsay (1989) and others.

When the linear models have been well-developed and understood, researchers turned their attention to nonlinear (Tong and Lim, 1980; Härdle and Vieu, 1992; Chan, 1993; Härdle, Chen and Lütkepohl, 1997), nonparametric (Chen and Tsay, 1993a; Chen and Tsay, 1993b; Härdle et al., 1997; Xia and Li, 1999b; Cai et al., 2000; Fan and Yao, 2003), and spatial-temporal modelling (Handcock and Wallis, 1994; Cressie and Huang 1999; Gneiting, 2002).

However, modern data becomes complex and massive. This dissertation proposes new models and methods to analyze the time series data in functional form, high dimension, or nonlinear dynamics with seasonality.

1.2 Functional Time Series

Functional data analysis has received much attention in the last few decades, and has been widely applied in many fields, including medical science (Houghton et al., 1980; Gasser et al., 1984; Ratcliffe et al., 2002a; Ratcliffe et al., 2002b), behavioral science (Keselman and Keselman, 1993), and economics (Roberts, 1995; Diebold and Li, 2006). Nonparametric methods, such as spline methods (Silverman, 1994; Brumback and Rice, 1998; Zhou et al., 1998; Cai et al., 2000) and kernel smoothing (Nadaraya, 1964; Watson, 1964; Gasser et al., 1984; Fan and Gijbels, 1996), were often implemented to analyze functional data. Unsupervised learning methods, such as principal component analysis (James et al., 2000) and clustering analysis (James and Sugar, 2003) were extended for functional data. Books by Ramsay and Silverman (2005), Ferraty and Vieu (2006), Horváth and Kokoszka (2012), provide a comprehensive introduction on functional data analysis.

Often, a variety of functional data is observed over time and has serial dependence. For example, in financial industry, the implied volatility of an option as a function of moneyness changes over time. In insurance industry, age-specific mortality rate as a function of age changes over time. In banking industry, term structure of interest rates (yield as a function of time to maturity of a bond) changes over time. In meteorology, daily records of temperature, precipitation and cloud cover for a region, viewed as three related functional surfaces, change over time.

However, scalar or vector time series models cannot be applied directly to functional data. Bosq (2000) first introduced functional autoregressive (FAR) models with order p,

$$X_t = \Delta_1 X_{t-1} + \ldots + \Delta_p X_{t-p} + \varepsilon_t,$$

where $X = (X_t, t \in \mathbb{Z})$ and $\varepsilon = (\varepsilon_t, t \in \mathbb{Z})$ are a sequence of random functions and a functional white noise process respectively, and Δ_i, is a linear operator in Hilbert functional space $\mathbf{H}$. The linear operators can be estimated by performing functional principal component analysis on the sample autocovariance operators. The consistency

of such estimators has been proved (Bosq, 2000; Hörmann et al., 2013). All the theoretical and empirical results in the literature have been developed based on the models and methods in Bosq (2000), including Hörmann and Kokoszka (2010), Horváth et al. (2010), Aue et al. (2012), Horváth et al. (2012), Berkes et al. (2013), and Hörmann et al. (2013). So far functional time series data is still a less studied area.

Sieve methods are a popular set of tools to estimate parameters in an infinite-dimensional space, including the functional space. It optimizes an objective function over a sequence of finite-dimensional parameter subspaces, which are called sieve spaces. Hence, the problem is reduced to a sequence of parametric ones. The consistency of the sequence of the sieve estimators can be derived, under certain conditions of sieve spaces. Sieve methods have been discussed in Chen and Shen (1998), Chen (2008), Halberstam and Richert (2013).

In this dissertation, we develop a new class of functional time series models called the convolutional functional autoregressive (CFAR) models, along with its associated estimation procedure using splines and sieve methods. As a special case of Bosq (2000), our model provides an intuitive and direct interpretation of the dynamics of a stochastic process. It assumes that the function at time t is a result of the sum of convolutions of the past functions with convolution functions plus a noise process, mimicking the autoregressive process commonly used in scalar time series. It is also an extension of vector autoregressive process. We makes contributions to the literature in three aspects. First, we establish the convergence rate of the convolution function estimator in a general case, while Bosq (2000) only considered consistency. Second, in reality functional data are often observed at discrete points, hence nonparametric methods such as splines method are used to obtain a continuous curve. The possible estimation error introduced by these methods was overlooked by Hörmann and Kokoszka (2010), Horváth et al. (2010), Horváth et al. (2012), Hörmann et al. (2013). Under the CFAR model settings, we consider the recovery procedure of functional data, when studying asymptotic properties of the estimators. It can be shown that the estimation error due to discrete observation points is of smaller order, hence as long as the number of observations at each time goes to infinity, the estimator is consistent and the convergence

rate does not depend on the number of discrete samples available directly. Thirdly, we develop model building, model validation and prediction procedures for CFAR models, while the picture of FAR models is less complete due to lack of specific model assumptions.

1.3 Factor Models for High-Dimensional Time Series

Multivariate time series models are often confronted with computational challenges, overparametrization and overfitting issues, especially when dealing with high-dimensional data. Factor analysis is considered as an effective way to alleviate these problems by dimension reduction, starting with Anderson (1963) and Priestley et al. (1974) who applied it to multivariate time series. In the last decades, much attention has been paid to the high-dimensional cases. Chamberlain and Rothschild (1983) and Forni et al. (2000) studied the factor model consisting of common factors and idiosyncratic component with weak cross-sectional and serial dependence. Bai and Ng (2002) and Hallin and Liška (2007) proved that the number of factors can be estimated consistently and established the convergence rate of factor estimators. Peña and Box (1987), Pan and Yao (2008) decomposed the time series into two parts, a latent factor process and a vector white noise process, in which strong cross-sectional dependence is allowed. Lam et al. (2011) and Lam and Yao (2012) developed an approach that takes advantage of information from autocovariance matrices at nonzero lags via eigendecomposition to estimate the factor loading space, and established the asymptotic properties as the dimension goes to infinity with sample size. This innovative method is applicable to nonstationary processes and processes with uncorrelated or endogenous regressors (Chang et al., 2013).

Regime switching (Hamilton, 1989) has been introduced in different models, including threshold models (Tong and Lim, 1980; Tong, 1983) and ARCH models (Hamilton and Susmel, 1994; Hamilton, 1996), and has various applications in economics, including analyzing business circle (Kim and Nelson, 1998), GNP (Hansen, 1992), interest rate (Gray, 1996) and monetary policy (Bernanke and Gertler, 2000; Sims and Zha,

2006). Factor models with regime switching can be tracked back to Diebold and Rudebusch (1994). In this chapter we generalize the factor models of Pan and Yao (2008) and introduce a factor model with an unobserved state variable switching between several regimes in which the mean, factor loadings and the covariance matrices of noise process are all different. By allowing these parameters to switch across regimes, it enhances flexibility in modeling multivariate time series, and provides an effective tool to distinguish and identify the dynamics over time.

For factor models, switching mechanism can be found in many cases. For example, CAPM theory indicates that the expected market return is an important factor for the expected return of an asset, and it is expected that its impact (loadings) on any individual asset may be different depending on whether a stock market is volatile or stable. In economics, risk-free rate, unemployment and economic growth are crucial factors for all economic activities and their performance indicators. Again, the loadings of these factors may vary under different fiscal policies (neutral, expansionary or contractionary) or in different stages of the economic circle (expansion, peak, contraction, or trough); see Kim and Nelson (1998).

In this dissertation, we develop an iterative algorithm for the estimation of model parameters and unobserved time-varying states based on eigendecomposition and Viterbi algorithm. The theoretical properties of the estimators are investigated. As in Lam et al. (2011) whose model is essentially a one-regime model in our case, the convergence rate of estimated loading space depends on the 'strength' of the state. We discover that, with multiple states of different 'strength', the convergence rate of loading space estimator for strong states is the same as the one-regime case, while the rate improves for weak states, gaining extra information from the strong states. Empirical results confirm such observations.

1.4 Functional Coefficient Seasonal Models

Seasonal time series are commonly observed in various applications, including economic and business data, meteorological data and environmental data as well as other fields.

There is a vast literature on seasonal time series analysis, ranging from stochastic seasonality models such as the seasonal ARIMA models (Box et al., 1994; Shumway and Stoffer 2000; Peña et al., 2001), the deterministic seasonal models such as the linear or polynomial additive or multiplicative seasonal component models (Shumway, 1988; Brockwell and Davis, 1991; Franses, 1996, 1998). The books by Hylleberg (1992), Franses (1996, 1998), and Ghysels and Osborn (2001) provide a comprehensive review on the traditional seasonal time series analysis methods. Most of these methods are linear (or polynomial) and parametric in nature. However, it has been documented that time series are often nonlinear (Tong, 1990; Tjøstheim, 1994; Hylleberg, 1992; Franses, 1996, 1998) and often there is not enough information to determine a suitable parametric form for the nonlinear structure. Härdle et al. (2004) discussed and reviewed many the popular statistical nonparametric and semiparametric methods. There was no systematic research done on nonparametric approaches to seasonal time series models, until Burman and Shumway (1998) proposed a nonparametric/semiparametric approach to seasonal time series, which opened the door in this area.

To characterize the seasonality of the monthly number of tourists visiting Hawaii, we propose a nonparametric seasonal time series model with a functional coefficient structure. Different from a linear autoregressive seasonal model with possible regression terms, the coefficients in the proposed model change over time and consist of the trend and seasonal components to characterize the seasonality. This class of models includes, as its special cases, the standard additive trend and seasonal component models as well as other seasonal time series models. We use the local linear approach to estimate the trend and seasonal effect functions nonparametrically and use it to model the Hawaiian tourism data.

Chapter 2

Convolutional Autoregressive Models for Functional Time Series

Motivated by Bosq (2000), who first introduced the functional autoregressive (FAR) models, we propose a convolutional functional autoregressive (CFAR) model, where the function at time t is a result of the sum of convolutions of the past functions with a set of convolution functions, plus a noise process, mimicking the autoregressive process. It provides an intuitive and direct interpretation of the dynamics of a stochastic process. Instead of spectral decomposition approach commonly used in functional data analysis, we adopt a sieve estimation procedure based on B-spline approximation of the convolution functions. We establish convergence rate of the proposed estimator, and investigate its theoretical properties. The model building, model validation, and prediction procedures are also developed. Both simulated and real data examples are presented in this chapter.

2.1 Convolutional Functional Autoregressive Models

We introduce some notations first. For any vector $\boldsymbol{\mu}$, $(\boldsymbol{\mu})_i$ denotes its i-th entry. For any matrix $\mathbf{H}$, $(\mathbf{H})_{ij}$ denotes its (i,j)-th entry. Let

$$\begin{aligned} \text{Lip}^h[-1,1] &= \{f \in [-1,1] : |f(x+\delta) - f(x)| \leq M\delta^h, M < \infty\}, \\ \text{Lip}_2^{r+h}[-1,1] &= \{f \in C^r[-1,1] : f^{(r)} \in \text{Lip}^h[-1,1], r \in \mathbb{N}, h \in (0,1]\}, \end{aligned}$$

If $f \in \text{Lip}_2^{\zeta}[-1,1]$, then ζ is called moduli of smoothness of $f(\cdot)$, which measures the smoothness of $f(\cdot)$.

Without loss of generality, in this chapter, we restrict ourselves only to consider

functional time series in the space $L_2[0,1]$.

If a function $f : \Omega \to \mathbb{R}^{[0,1]}$ is defined on a probability space $(\Omega, \mathcal{F}, \mathcal{P})$, and $f \in L_2[0,1]$, then f is called a random function on $[0,1]$. The space containing all the random functions on $[0,1]$ is denoted by $\mathbf{H}[0,1]$, when L-2 norm is applied.

Definition 2.1. *A sequence of random functions $\varepsilon = (\varepsilon_t, t \in \mathbb{Z})$ in $\mathbf{H}[0,1]$ is said to be a **white noise**, if*

1) $0 < \mathrm{E}\|\varepsilon_t\|_2^2 = \sigma^2 < \infty$, $\mathrm{E}(\varepsilon_t) = 0$, and $\mathrm{Cov}(\varepsilon_t(s_1), \varepsilon_t(s_2))$ does not depend on t for any $s_1, s_2 \in [0,1]$.

2) $\mathrm{Cov}(\varepsilon_t(s_1), \varepsilon_{t+h}(s_2)) = 0$, for $h \neq 0$ and $s_1, s_2 \in [0,1]$.

Definition 2.2. *A sequence of random functions $X = (X_t, t \in \mathbb{Z})$ in $\mathbf{H}[0,1]$ is **(weakly) stationary**, if the mean and covariance functions do not vary with time, i.e., for $\forall h \in \mathbb{Z}, s_1, s_2 \in [0,1]$,*

$$\mathrm{E}(X_t(s_1)) = \mu(s_1), \quad \text{and} \quad \mathrm{Cov}(X_{n+h}(s_1),\, X_{m+h}(s_2)) = \mathrm{Cov}(X_n(s_1),\, X_m(s_2)).$$

A sequence of random functions $X = (X_t, t \in \mathbb{Z})$ in $\mathbf{H}[0,1]$ is called a convolutional functional autoregressive model with order p, CFAR(p), if

$$X_t(s) \;=\; \sum_{i=1}^{p} \int_0^1 \phi_i(s-u) X_{t-i}(u) du + \varepsilon_t(s), \quad s \in [0,1], \tag{2.1}$$

where $\phi_i \in L_2[-1,1]$ for $i = 1, \ldots, p$, are called convolution functions, and $\varepsilon = (\varepsilon_t, t \in \mathbb{Z})$ are i.i.d. Ornstein-Uhlenbeck processes defined on $[0,1]$, following the stochastic differential equation, $d\varepsilon_t(s) = -\rho\varepsilon_t(s)ds + \sigma dW_s$, $\rho > 0$ with W_s being a Wiener process.

Remark 2.1. The skeleton of $X_t(\cdot)$, excluding the noise process, defined as $f_t(\cdot)$, is the sum of convolutions of ϕ_i and X_{t-i}. It can also be viewed as the sum of non-normalized smoothed versions of the p past functions $\{X_{t-1}, \ldots, X_{t-p}\}$. From a pointwise view, $f_t(s)$ is a weighted sum of $\{X_{t-1}, \ldots, X_{t-p}\}$, and the weights $\phi_i(s-u)$ depend on the distance between s and u.

Remark 2.2. We assume the noise process in model (2.1) is an Ornstein-Uhlenbeck

process, which is spatially stationary, Markovian and continuous but not differentiable with the following properties:

$$\varepsilon_t(s_1) \sim N(0, \frac{\sigma^2}{2\rho}), \quad \mathrm{Corr}(\varepsilon_t(\mathrm{s}_1), \varepsilon_t(\mathrm{s}_2)) = \mathrm{e}^{-\rho|\mathrm{s}_1-\mathrm{s}_2|}, \quad \forall \mathrm{s}_1, \mathrm{s}_2 \in [0,1].$$

The variance is constant, and the correlation of the process at s_1 and s_2 is determined by the distance of s_1 and s_2. If all the convolution functions $\{\phi_i(\cdot), i = 1, \ldots, p\}$ are continuous, X_t is also continuous, but not differentiable.

The convolution functions $\{\phi_i(\cdot), i = 1, \ldots, p\}$ determine the pattern of $X_t(\cdot)$ process. Here we show three examples to illustrate the impact of the convolution functions.

Example 2.1. Convolutional functional autoregressive model of order 1, CFAR(1),

$$X_t(s) = \int_0^1 \phi(s-u) X_{t-1}(u) du + \varepsilon_t(s), \quad s \in [0,1], \tag{2.2}$$

$\rho = 5$, $\sigma^2 = 10$. We consider three convolution functions:

(i) $\phi(s) = 1$, $s \in [-1, 1]$, and $X_0(\cdot) = 0$;

(ii) $\phi(s) = 1 - |s|$, $s \in [-1, 1]$, and $X_0(\cdot) = 10$;

(iii) $\phi(s) = I(s > 0)$, $s \in [-1, 1]$, and $X_0(\cdot) = 10$.

Simulated processes of functions for each $\phi(\cdot)$ are shown in the top, middle and bottom panel of Figure 2.1, respectively, for $t = 1, 2, 3, 100$. The solid lines and dashed lines are $X_t(\cdot)$ and $f_t(\cdot)$ respectively. In case(i), since $\phi(\cdot)$ is a constant function, $f_t(\cdot)$ is simply the average of $X_{t-1}(\cdot)$ hence a constant function. In case(ii), $\phi(\cdot)$ is a unimodal function around 0. We start the process with a constant function, $X_0(\cdot) = 10$, and $f_1(s)$ is larger when s is around 0.5. In case (iii), $\phi(\cdot)$ is an indicator function on $(0, 1]$, so the skeleton of $f_t(s)$ would be a partial integration of $X_{t-1}(\cdot)$ on the left of s in $[0, s]$. At $s = 0$, $X_t(0)$ contains no information of X_{t-1}, but only noise; as s increases, the weight functions $\phi(s - \cdot)$ increases and information carried by $X_t(s)$ on $X_{t-1}(\cdot)$ increases as well; at $s = 1$, $X_t(1)$ is the integration of the function $X_{t-1}(\cdot)$ in the entire range of $[0, 1]$ plus noise. It is worth noting that in case(i), we start at $X_0 = 0$, but

$X_t(\cdot)$ gets explosive as time goes by. On the other hand, although we start at a large value, $X_0(s) = 10$ for both cases of (ii) and (iii), the processes become close to 0 when $t = 100$. This is because, the process in case(i) is nonstationary, while the processes in case(ii) and case(iii) are stationary, and their stationary means are a constant function at 0.

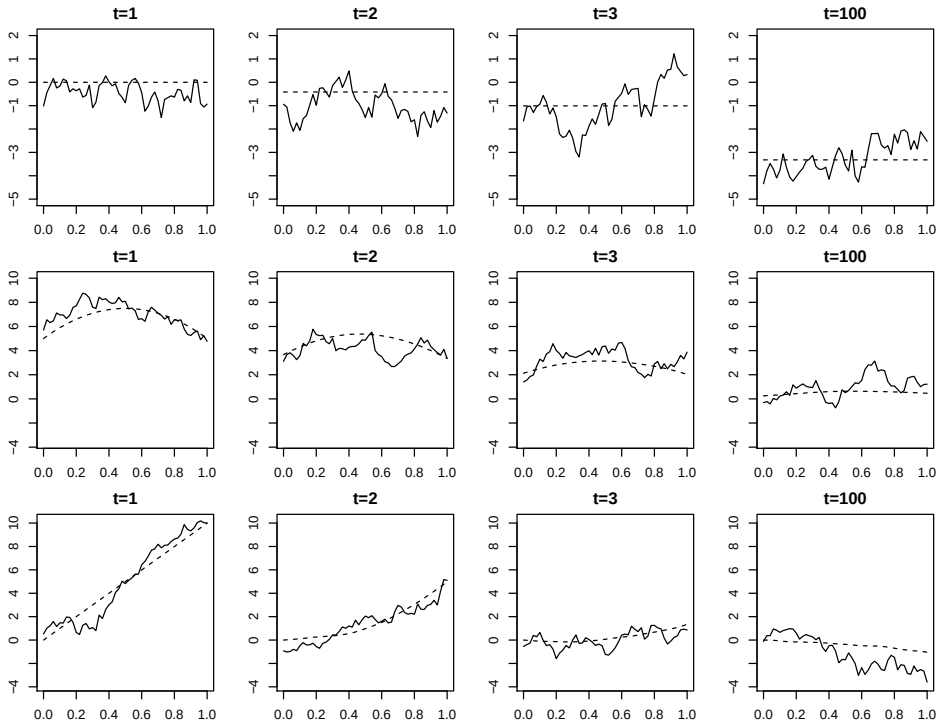


Figure 2.1: Plots of $X_t(\cdot)$ and $f_t(\cdot)$, $t = 1, 2, 3, 100$ for Example 2.1(i) (top panel), Example 2.1(ii) (middle panel) and Example 2.1(iii) (bottom panel).

Theorem 2.1 presents a sufficient condition for the stationarity of CFAR(1) models.

Theorem 2.1. *The CFAR(1) process $X = (X_t, t \in \mathbb{Z})$, defined in (2.1) is (weakly) stationary, if*

$$\kappa = \sup_{0 \le s \le 1} \left(\int_0^1 \phi^2(s-u)du \right)^{1/2} < 1. \tag{2.3}$$

Here weak stationarity is acturally equivalent to strong stationarity, since we assume that the noise process is Gaussian. It is easy to see that $\|\phi\|_2^2 < 1$ is also a sufficient condition for stationarity. However, the condition in Theorem 1 is a weaker one, since

for any s, $\int_0^1 \phi^2(s-u)du \leq \int_{-1}^1 \phi^2(u)du$. Note that $\int_0^1 \phi^2(s-u)du$ is the sum of squares of the convolution weights of $X_{t-1}(\cdot)$ to obtain $f_t(s)$.

Theorem 2.2 presents a sufficient condition for stationarity of CFAR(p) models.

Theorem 2.2. *The CFAR(p) process $X = (X_t, t \in \mathbb{Z})$ defined in (2.1) is (weakly) stationary, if all the roots of the characteristic function*

$$1 - \kappa_1 z - \kappa_2 z^2 - \ldots - \kappa_p z^p = 0, \tag{2.4}$$

are outside the unit circle, where $\kappa_i = \sup_{0\leq s\leq 1} \sqrt{\int_0^1 \phi_i^2(s-u)du}$ for $i = 1 \ldots, p$.

Condition in Theorem 2.2 is similar to the sufficient condition for scalar AR(p) models to be stationary, replacing the AR coefficients by the maximum of the norm of the weights function $\phi_i(s-\cdot)$, $s \in [0,1]$.

Corollary 2.1. *The CFAR(1) process $X = (X_t, t \in \mathbb{Z})$ satisfies (2.3). Define a sequence of functions on $[0,1]^2$:*

$$\Psi_1(s,u) = \phi(s-u), \quad \Psi_\ell(s,u) = \int_0^1 \Psi_{\ell-1}(s,v)\phi(v-u)\,dv, \quad \text{for } \ell \geq 2. \tag{2.5}$$

Then X_t has the following representation

$$X_t(s) = \varepsilon_t(s) + \sum_{\ell=1}^{\infty} \int_0^1 \Psi_\ell(s,u)\varepsilon_{t-\ell}(u)\,du. \tag{2.6}$$

Corollary 2.2. *The CFAR(p) process $X = (X_t, t \in \mathbb{Z})$ satisfies (2.4). Define a sequence of functions on $[0,1]^2$:*

$$\begin{aligned}
\Psi_1(s,u) &= \phi_1(s-u), \\
\Psi_\ell(s,u) &= \sum_{i=1}^{\ell-1} \int_0^1 \phi_i(s-v)\Psi_{\ell-i}(v,u)\,dv + \phi_\ell(s,u), \quad 2 \leq \ell \leq p, \\
\Psi_\ell(s,u) &= \sum_{i=1}^{p} \int_0^1 \Phi_i(s,v)\Psi_{\ell-i}(v,u)\,dv, \quad \ell > p.
\end{aligned}$$

Then X_t has the following representation

$$X_t(s) = \varepsilon_t(s) + \sum_{\ell=1}^{\infty} \int_0^1 \Psi_\ell(s,u)\varepsilon_{t-\ell}(u)\, du. \tag{2.7}$$

Corollary 2.1 and Corollary 2.2 are derived from Theorem 3.1 and Theorem 5.1 of Bosq(2000) directly. This is similar to that of stationary scalar AR(p) case when no constant is in the model. To include nonzero mean function $\mu(\cdot)$, we use $X_t(s) - \mu(s) = \sum_{i=1}^{p} \int_0^1 \phi_i(s-u)(X_{t-i}(u) - \mu(u))du$.

2.2 Estimation, Prediction and Order Determination

Assume that we observe $X_t(\cdot)$ at discrete points, $s = n/N$, $n = 0, \ldots, N$, for time $t = 1, .., T$.

2.2.1 Estimation

Since the convolution operator guarantees continuous path of $X_t(\cdot)$, we approximate it with linear interpolation of the observations. For any $s \in [-1, 1]$, if $s_{n-1} \le s < s_n$, let

$$\widetilde{X}_t(s) = \frac{(s_n - s)X_t(s_{n-1}) + (s - s_{n-1})X_t(s_n)}{1/N}. \tag{2.8}$$

Because $X_t(\cdot)$ is continuous but not differentiable and observations are not subject to noises, linear interpolation suffices to evaluate $\int \phi_i(s-u)X_{t-i}(u)du$.

In a typical nonparametric fashion, we approximate the unknown convolution functions $\phi_i(\cdot)$ with a B-spline approximation. Specifically,

$$\phi_i(\cdot) \approx \widetilde{\phi}_{k,i}(\cdot) = \sum_{j=1}^{k} \widetilde{\beta}_{k,i,j} B_{k,j}(\cdot), \text{ for } i = 1, \ldots, p, \tag{2.9}$$

where $\{B_{k,j}(\cdot), j = 1, \ldots, k\}$ are uniform cubic B-spline functions with degree of freedom k.

Plugging in the linear interpolation of $X_{t-i}(\cdot)$ and B-spline approximation of $\phi_i(\cdot)$

into (2.1), we get

$$\varepsilon_t(s_n) \approx X_t(s_n) - \sum_{i=1}^{p}\sum_{j=1}^{k} \widetilde{\beta}_{k,i,j} \int_0^1 B_{k,j}(s_n - u)\widetilde{X}_{t-i}(u)du. \tag{2.10}$$

Let $\widetilde{\boldsymbol{\varepsilon}}_t' = (\widetilde{\varepsilon}_{t,0}, \ldots, \widetilde{\varepsilon}_{t,N})'$ be an $(N+1) \times 1$ vector. $\widetilde{\varepsilon}_{t,n}$ is the approximated noise at s_n and time t defined as on the right hand side of (2.10). It can be expressed as $\tilde{\boldsymbol{\varepsilon}}_t = X_t(\mathbf{s}) - \mathbf{M}_t\widetilde{\boldsymbol{\beta}}_k$, where $\mathbf{M}_t = (\mathbf{M}_{t,1}, \ldots, \mathbf{M}_{t,p})$, $\mathbf{M}_{t,i}$ is an $(N+1) \times k$ matrix, $(\mathbf{M}_{t,i})_{nj} = \int_0^1 B_{k,j}(s_n - u)\widetilde{X}_{t-i}(u)du$, and $\widetilde{\boldsymbol{\beta}}_k$ is a $pk \times 1$ vector, $\widetilde{\boldsymbol{\beta}}_k = (\widetilde{\boldsymbol{\beta}}_{k,i}', \ldots \widetilde{\boldsymbol{\beta}}_{k,p}')'$, and $(\boldsymbol{\beta}_{k,i})_j = \widetilde{\beta}_{k,i,j}$. Since $B_{k,j}(\cdot)$ are fixed and known functions, $\mathbf{M}_t$ is known, given the observations. Under the Ornstein-Uhenleck process, for equally spaced s_n, $\varepsilon_t(\cdot)$ follows an AR(1) process with AR coefficient $e^{-\rho/N}$, and covariance matrix $\boldsymbol{\Sigma}$, where $(\boldsymbol{\Sigma})_{ij} = e^{-\rho|i-j|/N}\sigma^2/2\rho$. Therefore, $\boldsymbol{\beta} = \{\beta_{ij}, i = 1, \ldots, p, j = 1, \ldots, k\}$, σ^2, and ρ can be estimated by maximizing the approximated log-likelihood function,

$$Q_{k,T,N}(\boldsymbol{\beta}, \sigma^2, \rho) = -\frac{(N+1)(T-p)}{2} \ln\left(\frac{\pi\sigma^2}{\rho}\right) \tag{2.11}$$

$$- \frac{N(T-p)}{2} \ln(1 - e^{-2\rho/N}) - \frac{1}{2}\sum_{t=p+1}^{T} \tilde{\boldsymbol{\varepsilon}}_t' \, \boldsymbol{\Sigma}^{-1} \, \tilde{\boldsymbol{\varepsilon}}_t, \tag{2.12}$$

Let $\widehat{\boldsymbol{\beta}}_k = (\widehat{\boldsymbol{\beta}}_{k,1}', \ldots, \widehat{\boldsymbol{\beta}}_{k,p}')'$ be the estimator of $\widetilde{\boldsymbol{\beta}}$ by maximizing $Q_{k,T,N}(\boldsymbol{\beta}, \sigma^2, \rho)$, where $(\widehat{\boldsymbol{\beta}}_{k,i})_j = \widehat{\beta}_{k,i,j}$.

After reparameterization, the objective function can be written as

$$Q_{k,T,N}(\boldsymbol{\beta}, \varphi, \omega) = -\frac{(N+1)(T-p)}{2} \ln(2\pi\omega) - \frac{N(T-p)}{2} \ln(1 - \varphi^2) - \frac{e(\boldsymbol{\beta}, \varphi)}{2\omega}, \tag{2.13}$$

where $\varphi = e^{-\rho/N}$, $\omega = \sigma^2/2\rho$, $\boldsymbol{\Sigma}_0 = 2\rho\boldsymbol{\Sigma}/\sigma^2$, and $e(\boldsymbol{\beta}, \varphi) = \sum_{t=p+1}^{T} \widehat{\boldsymbol{\varepsilon}}_t'\boldsymbol{\Sigma}_0^{-1}\widehat{\boldsymbol{\varepsilon}}_t$. $\boldsymbol{\Sigma}_0$ is actually the correlation matrix of $\{\varepsilon_t(s_n), n = 1, \ldots, N\}$.

Given $\boldsymbol{\beta}$ and φ, the maximizer of w is

$$\widehat{\omega} = \frac{e(\boldsymbol{\beta}, \varphi)}{(N+1)(T-p)}. \tag{2.14}$$

Hence

$$\max_{\boldsymbol{\beta},\varphi,\omega} Q_{k,T,N}(\boldsymbol{\beta},\varphi,\omega) = \max_{\boldsymbol{\beta},\varphi}\Big(-\frac{(N+1)(T-p)}{2}\ln e(\boldsymbol{\beta},\varphi) - \frac{N(T-p)}{2}\ln(1-\varphi^2)+c\Big)$$
$$= \max_{\varphi}\Big(-\frac{(N+1)(T-p)}{2}\ln e(\boldsymbol{\beta}(\varphi),\varphi) - \frac{N(T-p)}{2}\ln(1-\varphi^2)+c\Big), \tag{2.15}$$

where c is a constant and

$$\widehat{\boldsymbol{\beta}}(\varphi) = \arg\min_{\boldsymbol{\beta}} e(\boldsymbol{\beta},\varphi) = \left(\sum_{t=p+1}^{T}\mathbf{M}_t'\boldsymbol{\Sigma}^{-1}\mathbf{M}_t\right)^{-1}\left(\sum_{t=p+1}^{T}\mathbf{M}_t'\boldsymbol{\Sigma}^{-1}X_t(\mathbf{s})\right). \tag{2.16}$$

Then this problem becomes one parameter optimization, and the estimator of φ can be easily obtained by maximizing (2.15). Together with (2.14) and (2.16), we have the estimators for σ^2, ρ, $\boldsymbol{\beta}$. The convolution function $\phi_i(\cdot)$ can be estimated by $\widehat{\phi}_{k,i}(\cdot) = \sum_{j=1}^{k}\widehat{\beta}_{k,i,j}B_{k,j}(\cdot)$.

Remark 2.3. The above method can also deal with non-equally spaced $\{s_n, n = 0,\ldots,N\}$, different observation points at each time (i.e. $\{s_n, n = 0,\ldots,N\}$ vary with time), or different number of observations at each time. For simplicity, we assume equally spaced $\{s = n/N\}$.

2.2.2 Fitted Values

Given estimated $\phi_i(\cdot)$, the continuous process $X_t(\cdot)$ can be interpolated more precisely than simple linear interpolation. By taking advantage of information from the observed $\{X_t(s_n), n = 0,\ldots,N\}$ and the Markovian property of Ornstein-Uhlenbeck process, we obtain a better approximation of $X_t(\cdot)$. Specifically, for a fixed $s \in (s_n, s_{n+1})$, let

$$\widetilde{f}_t(s) = \sum_{i=1}^{p}\int_0^1 \widehat{\phi}_{k,i}(s-u)\widetilde{X}_{t-i}(u)ds, \tag{2.17}$$

and let

$$\widetilde{X}_t^*(s) = \widetilde{f}_t(s) + \begin{pmatrix} e^{-\widehat{\rho}(s-s_n)} & e^{-\widehat{\rho}(s_{n+1}-s)}\end{pmatrix}\widehat{\boldsymbol{\Sigma}}^{-1}\begin{pmatrix} X_t(s_n)-\widetilde{f}_t(s_n)\\ X_t(s_{n+1})-\widetilde{f}_t(s_{n+1})\end{pmatrix}, \tag{2.18}$$

where $\widehat{\mathbf{\Sigma}}_0$ is the estimated correlation matrix of $\varepsilon_t(s_n)$ and $\varepsilon_t(s_{n+1})$, with diagonal entry 1 and off-diagonal entry $e^{-\widehat{\rho}/N}$. When $s = s_n$, $\widetilde{X}_t^*(s) = X_t(s)$. Here $\widetilde{f}_t(\cdot)$ is the estimated skeleton of $X_t(\cdot)$ process, and $X_t(s_n) - \widetilde{f}_t(s_n)$ is the approximated residual process.

Remark 2.4. Plugging (2.18) into (2.17), we can fit $\widetilde{f}_{t+1}(\cdot)$, and $\widetilde{X}_{t+1}^*(s)$ iteratively. However, empirical results show that its impact is minor with large N.

2.2.3 Prediction

Given $X_1, \ldots, X_t$, the least squares prediction of $X_{t+1}(s)$ is

$$\widehat{X}_{t+1}(s) = \sum_{i=1}^{p} \int_0^1 \widehat{\phi}_{k,i}(s-u)\widetilde{X}_{t+1-i}^*(u)du, \tag{2.19}$$

where $\widetilde{X}_{t+1-i}^*(\cdot)$ is the fitted processes of $X_{t+1-i}(\cdot)$ in (2.18). The residual is defined as

$$\widehat{\varepsilon}_{t+1}(s) = X_{t+1}(s) - \widehat{X}_{t+1}(s). \tag{2.20}$$

In addition, if $X_{t+1}(\cdot)$ is partially observed at $s = 0, \frac{1}{N}, \ldots, s^*$, the prediction of $X_{t+1}(s)$, for $s \in (s^*, 1]$ can be obtained with,

$$\widehat{X}_{t+1}^*(s) = \widehat{X}_{t+1}(s) + e^{-\widehat{\rho}(s-s^*)}(X_{t+1}(s^*) - \widehat{X}_{t+1}(s^*)), \tag{2.21}$$

where $\widehat{X}_{t+1}(s)$ is from (2.19).

2.2.4 Determination of the B-spline Approximation Order

As in all nonparametric estimation, bandwidth is the key control parameter that balances the estimation bias and variance (Ruppert et al., 1995; Fan and Gijbels, 1996). For spline methods, the control parameter is the degree of freedom k (Zhou et al., 1998; Huang, 2003). We propose to choose k to minimize the out-sample rolling forecasting error instead of the typically used cross validation criterion due to the time series nature of our problem. Specifically, for a given k, and each $t = T_0, \ldots, T$, we use data

$\{X_h(s_n), h = 0, ..., t-1, n = 0, \ldots, N\}$, observed before time t to estimate the convolution functions using B-spline with degree of freedom k, and predict $X_{k,t}(s_n)$ as in (2.19). Define overall squared rolling forecasting error

$$S(k) = \sum_{t=T_0}^{T} \sum_{n=0}^{N} (X_t(s_n) - \widehat{X}_{k,t}(s_n))^2. \tag{2.22}$$

The optimal k is chosen to be the one that minimizes $S(k)$.

2.2.5 AR Order Determination

F-test can be constructed for hypothesis testing of the significance of the convolution functions as well as for AR order determination. For testing $H_0 : \phi_{r+1}(\cdot) = \ldots = \phi_p(\cdot) = 0$ vs $H_1 :$ not H_0, we reject H_0 if

$$F = \frac{(SSE^{(r)} - SSE^{(f)})/[k(p-r)]}{SSE^{(f)}/[(N+1)(T-p) - pk]} > F_{\alpha, k(p-r), (N+1)(T-p)-pk}, \tag{2.23}$$

where $SSE^{(f)}$ and $SSE^{(r)}$ are sum of squared estimated residuals of the approximated model in (2.10) for full CFAR(p) and reduced CFAR(r) model respectively, for an optimally chosen k. Specifically,

$$SSE^{(f)} = \sum_{t=p+1}^{T} \widehat{\boldsymbol{\varepsilon}}'_{t,1} \widehat{\boldsymbol{\Sigma}}_{0,1}^{-1} \widehat{\boldsymbol{\varepsilon}}_{t,1}, \quad SSE^{(r)} = \sum_{t=p+1}^{T} \widehat{\boldsymbol{\varepsilon}}'_{t,2} \widehat{\boldsymbol{\Sigma}}_{0,2}^{-1} \widehat{\boldsymbol{\varepsilon}}_{t,2},$$

where $\widehat{\boldsymbol{\Sigma}}_{0,1}$ and $\widehat{\boldsymbol{\Sigma}}_{0,2}$ are the estimates of $\boldsymbol{\Sigma}_0$, the correlation matrix of $\varepsilon_t(\mathbf{s})$, whose (i,j)-th entry is $e^{-\rho|i-j|/N}$, from full and reduced models, respectively; $\{\widehat{\boldsymbol{\varepsilon}}_{t,i}(s_n), t = p+1, \ldots, T, n = 0, \ldots N\}$, $i = 1, 2$ are the residuals of the full and reduced models, respectively. This test can be used for model specification. We begin with CFAR(1) model, and sequentially add more lags of X_t, until the newly introduced lag is not significant.

In addition, define the cross-sectional residuals of ε_t,

$$e_t(s_n) = \varepsilon_t(s_n) - e^{-\widehat{\rho}/N} \varepsilon_t(s_{n-1}). \tag{2.24}$$

Under our model, $\{e_t(s_n), n = 1, \ldots, N\}$ is a white noise process, for $t = p+1, \ldots, T$. Let $\widehat{e}_t(s_n) = \widehat{\varepsilon}_t(s_n) - e^{-\widehat{\rho}/N}\widehat{\varepsilon}_t(s_{n-1})$. Hence, the features of $\{\widehat{e}_t(s_n), n = 1, \ldots, N\}$, can be studied with standard residual time series analysis for model validation.

2.3 Theoretical Properties

We study the asymptotic properties of our estimator as both N and T go to infinity. The degree of freedom of B-spline approximation k also goes to infinity with N and T.

2.3.1 Sieve Framework

In this section we reformulate our method under the sieve estimation framework.

The sieve method is designed for estimation of a parameter in an infinite-dimensional space $\boldsymbol{\Theta}$. Often optimization the objective function over the parameter space cannot be directly solved. Instead, we optimize the function over a sequence of subspaces $\{\boldsymbol{\Theta}_k, k \in \mathbb{Z}^+\}$, which are called sieve spaces. If the sieve spaces satisfy certain conditions, the sequence of estimators, called sieve estimators, is expected to be consistent, as the complexity of sieve spaces goes to infinity with sample size, see Chen (2008), Halberstam and Richert (2013).

The parameter space $\boldsymbol{\Theta}$ for CFAR(p) models contains all the functions from $L_2[-1, 1]$ that satisfy stationary condition specified in Theorem 2.2. The sieve space, $\boldsymbol{\Theta}_k$, used to approximate the parameter space $\boldsymbol{\Theta}$, is defined as $\boldsymbol{\Theta}_k = \mathbf{S}_k^p \cap \boldsymbol{\Theta}$, where $\mathbf{S}_k^p$ is product of p copies of $\mathbf{S}_k$,

$$\mathbf{S}_k = \{\sum_{j=1}^{k} \beta_j B_{k,j}(\cdot), \boldsymbol{\beta} = (\beta_1, \ldots, \beta_k)' \in \mathbb{R}^k\},$$

and $\{B_{k,j}(\cdot), j = 1, \ldots, k\}$ are the uniform cubic B-spline functions defined on $[-1, 1]$ with degree of freedom k. In other words, $\mathbf{S}_k$ is the cubic spline space with $k-4$ interior knots at $\{-1, -1 + 1/m, \ldots, 1 - 1/m, 1\}$, where $m = (k-3)/2$.

The population objective function to maximize we use here is,

$$Q(\boldsymbol{\phi}, \sigma^2, \rho) = \mathrm{E}\left[-\frac{1}{2\sigma^2}\left(\rho^2 \int_0^1 \varepsilon_t^2(s)ds + \rho\varepsilon_t^2(0) + \rho\varepsilon_t^2(1) - \rho\right)\right], \tag{2.25}$$

where $\varepsilon_t(s) = X_t(s) - \sum_{i=1}^p \int_0^1 \phi_i(s-u)X_{t-i}(u)du$. In fact, (2.25) is the expectation of log-likelihood of the Ornstein-Uhlenbeck process $\varepsilon_t(\cdot)$; see Rao (1999). Define the population objective function $Q_k(\boldsymbol{\beta}, \sigma^2, \rho)$ in sieve space $\boldsymbol{\Theta}_k$,

$$Q_k(\boldsymbol{\beta}, \sigma^2, \rho) = \mathrm{E}\left[-\frac{1}{2\sigma^2}\left(\rho^2 \int_0^1 \varepsilon_t^2(s)\, ds + \rho\varepsilon_t^2(0) + \rho\varepsilon_t^2(1) - \rho\right)\right],$$

where $\varepsilon_t(s) = X_t(s) - \sum_{i=1}^p \sum_{j=1}^k \beta_{i,j} \int_0^1 B_{k,j}(s-u)X_{t-i}(u)\, du$, and $\phi_i(\cdot) = \sum_{j=1}^k \beta_{i,j}B_{k,j}(\cdot)$, $i = 1, \ldots, p$.

When $X_t(\cdot)$ is only observed at discrete points, with (2.8) we have the sample objective function $Q_{k,T,N}(\boldsymbol{\beta}, \sigma^2, \rho)$

$$Q_{k,T,N}(\boldsymbol{\beta}, \sigma^2, \rho) = -\frac{(N+1)(T-p)}{2}\ln\left(\frac{\pi\sigma^2}{\rho}\right) - \frac{N(T-p)}{2}\ln(1-\varphi^2) - \frac{1}{2}\sum_{t=p+1}^{T} \widetilde{\varepsilon}_t' \boldsymbol{\Sigma}^{-1} \widetilde{\varepsilon}_t,$$

where $\widetilde{\varepsilon}_t(s) = X_t(s) - \sum_{i=1}^p \sum_{j=1}^k \beta_{i,j} \int_0^1 B_{k,j}(s-u)\widetilde{X}_{t-i}(u)du$, and $\phi_i(\cdot) = \sum_{j=1}^k \beta_{i,j}B_{k,j}(\cdot)$, $i = 1, \ldots, p$. Then the sieve estimator in space $\boldsymbol{\Theta}_k$ is

$$(\widehat{\boldsymbol{\beta}}_k, \widehat{\sigma}_k^2, \widehat{\rho}_k) = \arg\max Q_{k,T,N}(\boldsymbol{\beta}, \sigma^2, \rho).$$

2.3.2 Asymptotic Properties for CFAR(1) Models

Let $\mathbf{Q}_k$ be the linear operator defined in Section 6.4 of Schumaker (1981), which maps $C[-1, 1]$ into the space $\boldsymbol{\Theta}_k$. Let $\widetilde{\phi}_k = \mathbf{Q}_k\phi = \widetilde{\beta}_{k,1}B_{k,1} + \ldots + \widetilde{\beta}_{k,k}B_{k,k}$, and $r_k(s) = \phi(s) - \widetilde{\phi}_k(s)$.

Theorem 2.3. *X_t is a stationary CFAR(1) process defined in (2.1). $\widehat{\boldsymbol{\beta}}_k$ and $\widetilde{\boldsymbol{\beta}}_k$ are the B-spline coefficients of $\widehat{\phi}_k$ and $\widetilde{\phi}_k$. Assume $\phi \in \mathrm{Lip}_2^{\zeta}[-1, 1]$, and $\zeta > 1$. Then with given σ^2 and ρ, as $N, T \to \infty$,*

$$\sqrt{T}(\widehat{\boldsymbol{\beta}}_k - \widetilde{\boldsymbol{\beta}}_k - \mathbf{b}_k) \xrightarrow{d} N(\mathbf{0}, \boldsymbol{\Sigma}_k),$$

where $\mathbf{\Sigma}_k = \mathbf{\Gamma}_k^{-1}\mathbf{\Upsilon}_k\mathbf{\Gamma}_k^{-1}$, $\mathbf{\Upsilon}_k$ *is the long-run variance of the process* $\mathbf{u}_t + \mathbf{z}_t + \mathbf{A}_t\mathbf{b}_k$, *where* $\mathbf{b}_k = \mathbf{\Gamma}_k^{-1}\boldsymbol{\mu}_k$, *and the entries in* $\mathbf{u}_t$, $\mathbf{z}_t$, $\mathbf{A}_t$, $\boldsymbol{\mu}_k$ *and* $\mathbf{\Gamma}_k$ *are listed as follows*

$$\begin{aligned}
(\mathbf{u}_t)_i &= \frac{1}{\sigma^2}\int_0^1\int_0^1 \mathcal{A}(B_{k,i}, r_k)(u,v)X_{t-1}(u)X_{t-1}(v)(u,v)\,dudv,\\
(\mathbf{z}_t)_i &= \int_0^1 \left[\frac{2\rho}{\sigma^2}B_{k,i}(-u)\varepsilon_t(0) + \frac{1}{\sigma}\int_0^1 (\rho B_{k,i}(v-u) + B'_{k,i}(v-u))\,dW_t(v)\right] X_{t-1}(u)\,du,\\
(\mathbf{A}_t)_{ij} &= \frac{1}{\sigma^2}\int_0^1\int_0^1 \mathcal{A}(B_{k,i}, B_{k,j})(u,v)X_{t-1}(u)X_{t-1}(v)\,dudv,\\
(\boldsymbol{\mu}_k)_i &= \frac{1}{\sigma^2}\int_0^1\int_0^1 \mathcal{A}(B_{k,i}, r_k)(u,v)\gamma(u,v)\,dudv,\\
(\mathbf{\Gamma}_k)_{ij} &= \frac{1}{\sigma^2}\int_0^1\int_0^1 \mathcal{A}(B_{k,i}, B_{k,j})(u,v)\gamma(u,v)\,dudv,
\end{aligned}$$

where $\gamma(u,v) = \mathrm{Cov}(X_t(u), X_t(v))$, *and* $\mathcal{A}$ *is a binary functional operator,* $\mathcal{A} : H_1 \otimes H_1 \to H_2$, $H_1 = \{f : [-1,1] \to \mathbb{R}\}$,$H_2 = \{f : [-1,1]^2 \to \mathbb{R}\}$.

$$\begin{aligned}
\mathcal{A}(f,g)(u,v) = \rho f(-u)g(-v) + \rho f(1-u)g(1-v) + \int_0^1 f'(s-u)g'(s-v)\,ds\\
+ \rho^2\int_0^1 f(s-u)g(s-v)\,ds. \qquad (2.26)
\end{aligned}$$

Theorem 2.3 provides the asymptotic distribution of the estimated B-spline coefficients when the dimension of the sieve space is fixed. It is interesting to note that the distribution does not depend on N. As long as N goes to infinity and $T = o(N^2)$, the estimated B-spline coefficients converges, and the rate depends on T, but is independent of N. It is due to that errors introduced by the linear interpolation has smaller order than others.

With Theorem 2.3, we can easily obtain the asymptotic pointwise estimation error of the convolution function when k is fixed.

Corollary 2.3. *Assume* $\phi \in \mathrm{Lip}^{\zeta}[-1,1]$ *with* $\zeta > 1$, *and* $T = o(N^2)$. *For each fixed* k, *define* $\mathbf{B}_k(s) = (B_{k,1}(s), \ldots, B_{k,k}(s))'$, *and*

$$b_k(s) = \mathbf{B}_k(s)'\mathbf{b}_k - r_k(s), \qquad \sigma_k^2(s) = \mathbf{B}_k(s)'\mathbf{\Sigma}_k\mathbf{B}_k(s).$$

Then

$$\sqrt{T}\left(\widehat{\phi}_k(s) - \phi(s) - b_k(s)\right) \xrightarrow{d} N(0, \sigma_k^2(s)).$$

Theorem 2.4. *Assume $\phi \in C^{\zeta}[-1,1]$, where $\zeta \geq 2$ is an integer. Then as $k \to \infty$,*

$$\|b_k(\cdot)\|_\infty = O(k^{-\zeta_0+3/2}), \quad \|\sigma_k^2(\cdot)\|_\infty = O(k),$$

where $\zeta_0 = \min\{\zeta, 4\}$.

Theorem 2.4 shows the consistency of our proposed estimators as well as the convergence rates of the asymptotic bias and variance. The asymptotic bias depends on the smoothness of the convolution function and complexity of the sieve space used for approximation, while the asymptotic variance only depends on the complexity of the sieve space. It reflects the fact that the complexity of sieve space controls the trade-off between bias and variance. When k is a small number, the bias dominates the error, since there are not enough knots to approximate the convolution function. When k is large enough, the variance dominates, and estimation of extra B-spline coefficients introduces too much error. Theorem 2.4 also reveals the selection principle for the number of knots. When the convolution function is smooth, a small k is favorable; when the sample size is large, we prefer to have more knots.

Remark 2.5. By balancing the bias and variance, the optimal convergence rate of the estimation error is attained at $k \asymp O(T^{1/(2\zeta_0-2)})$, then both the squared bias and variance will be $O(T^{-\frac{2\zeta_0-3}{4\zeta_0-4}})$. If the moduli of smoothness of ϕ is greater than 4 and $k = O(T^{-1/4})$, the optimal convergence rate is $O(T^{-5/12})$. If B-splines with higher order are adopted, the estimator is expected to converge faster.

2.3.3 Asymptotic Properties for CFAR(p)

The asymptotic properties of the estimators of CFAR(p) models are very similar to these of CFAR(1) models. Let $\widetilde{\phi}_{k,i} = \mathbf{Q}_k\phi_i = \widetilde{\beta}_{k,i,1}B_{k,1} + \ldots + \widetilde{\beta}_{k,i,k}B_{k,k}$, and $r_{k,i}(s) = \phi(s) - \widetilde{\phi}_{k,i}(s)$.

Theorem 2.5 provides the bound of the B-spline coefficients estimation error, and is a higher dimension version of Theorem 2.3.

Theorem 2.5. *X_t is a stationary CFAR(p) process defined in (2.1). $\widehat{\boldsymbol{\beta}}_k$ and $\widetilde{\boldsymbol{\beta}}_k$ be the B-spline coefficients of $\{\widehat{\phi}_{k,1},\ldots,\widehat{\phi}_{k,p}\}$ and $\{\widetilde{\phi}_{k,1},\ldots,\widetilde{\phi}_{k,p}\}$ respectively. Assume $\phi_i \in \mathrm{Lip}_2^{\zeta_i}[-1,1]$, and $\zeta_i > 1$, for $i = 1,\ldots,p$. Then with given σ^2 and ρ, as $N,T \to \infty$,*

$$\sqrt{T}(\widehat{\boldsymbol{\beta}}_k - \widetilde{\boldsymbol{\beta}}_k - \mathbf{b}_k) \xrightarrow{d} N\left(\mathbf{0}, \boldsymbol{\Sigma}_k\right).$$

where $\boldsymbol{\Sigma}_k = \boldsymbol{\Gamma}_k^{-1}\boldsymbol{\Upsilon}_k\boldsymbol{\Gamma}_k^{-1}$, $\boldsymbol{\Upsilon}_k$ is the long-run variance of the process $\mathbf{u}_t+\mathbf{z}_t+\mathbf{A}_t\mathbf{b}_k$, where $\mathbf{b}_k = \boldsymbol{\Gamma}_k^{-1}\boldsymbol{\mu}_k$. Here $\mathbf{u}_t = (\mathbf{u}_{t,1}',\ldots,\mathbf{u}_{t,p}')'$, $\mathbf{z}_t = (\mathbf{z}_{t,1}',\ldots,\mathbf{z}_{t,p}')'$, $\boldsymbol{\mu}_k = (\boldsymbol{\mu}_{k,1}',\ldots,\boldsymbol{\mu}_{k,p}')'$, $\boldsymbol{\Gamma}_k$ and $\mathbf{A}_t$ can be partitioned into $p \times p$ blocks.

$$\boldsymbol{\Gamma}_k = \begin{pmatrix} \boldsymbol{\Gamma}_{k,0} & \boldsymbol{\Gamma}_{k,-1} & \ldots & \boldsymbol{\Gamma}_{k,-p+1} \\ \boldsymbol{\Gamma}_{k,1} & \boldsymbol{\Gamma}_{k,0} & \ldots & \boldsymbol{\Gamma}_{k,-p+2} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{\Gamma}_{k,p-1} & \boldsymbol{\Gamma}_{k,p-2} & \ldots & \boldsymbol{\Gamma}_{k,0} \end{pmatrix}, \quad \mathbf{A}_t = \begin{pmatrix} \mathbf{A}_{t,1,1} & \mathbf{A}_{t,1,2} & \ldots & \mathbf{A}_{t,1,p} \\ \mathbf{A}_{t,2,1} & \mathbf{A}_{t,2,2} & \ldots & \mathbf{A}_{t,2,p} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{A}_{t,p,1} & \mathbf{A}_{t,p,2} & \ldots & \mathbf{A}_{t,p,p} \end{pmatrix}.$$

The entries in $\mathbf{u}_t$, $\mathbf{z}_t$, $\mathbf{A}_t$, $\mathbf{b}_k$ and $\boldsymbol{\Gamma}_k$ are listed as follows

$$\begin{aligned}
(\mathbf{u}_{t,h})_i &= \frac{1}{\sigma^2}\sum_{q=1}^{p}\int_0^1\int_0^1 \mathcal{A}(B_{k,i},r_{k,q})(u,v)X_{t-h}(u)X_{t-q}(v)(u,v)\,dudv, \\
(\mathbf{z}_{t,h})_i &= \int_0^1 \left[\frac{2\rho}{\sigma^2}B_{k,i}(-u)\varepsilon_t(0) + \frac{1}{\sigma}\int_0^1 (\rho B_{k,i}(v-u) + B_{k,i}'(v-u))\,dW_t(v)\right] X_{t-h}(u)\,du, \\
(\mathbf{A}_{t,l,h})_{ij} &= \frac{1}{\sigma^2}\int_0^1\int_0^1 \mathcal{A}(B_{k,i},B_{k,j})(u,v)X_{t-l}(u)X_{t-h}(v)\,dudv, \\
(\boldsymbol{\mu}_{k,h})_i &= \frac{1}{\sigma^2}\sum_{q=1}^{p}\int_0^1\int_0^1 \mathcal{A}(B_{k,i},r_{k,q})(u,v)\gamma_{q-h}(u,v)\,dudv, \\
(\boldsymbol{\Gamma}_{k,h})_{ij} &= \frac{1}{\sigma^2}\int_0^1\int_0^1 \mathcal{A}(B_{k,i},B_{k,j})(u,v)\gamma_h(u,v)\,dudv,
\end{aligned}$$

for $i,j = 1,\ldots,k$ and $l,h = 1,\ldots,p$, where $\gamma_q(u,v) = \mathrm{Cov}(X_t(u),\, X_{t+q}(v))$, $q \in \mathbb{Z}$.

With Theorem 2.5, we have the asymptotic pointwise estimation error for each convolution function when k is fixed.

Corollary 2.4. *Assume $\phi_i \in \mathrm{Lip}^{\zeta_i}[-1,1]$, with $\zeta_i > 1$, $i = 1,\ldots,p$, and $T = o(N^2)$.*

For each fixed k, define $\mathbf{B}_{k,i}(s) = (\mathbf{B}'_{k,i,1}, \ldots, \mathbf{B}'_{k,i,k})'$, $\mathbf{B}_{k,i,i} = (B_{k,1}(s), \ldots, B_{k,k}(s))'$, *and* $\mathbf{B}_{k,i,j}$ *is a p-dimensional vector whose entries are all 0, for $j \neq i$. Then*

$$b_{k,i}(s) = \mathbf{B}_{k,i}(s)'\mathbf{b}_k - r_{k,i}(s), \qquad \sigma^2_{k,i}(s) = \mathbf{B}_{k,i}(s)'\mathbf{\Sigma}_k\mathbf{B}_{k,i}(s)), \quad \textit{for } i = 1, \ldots, p,$$

and

$$\sqrt{T}\left(\widehat{\phi}_{k,i}(s) - \phi_i(s) - b_{k,i}(s)\right) \xrightarrow{d} N(0, \sigma^2_{k,i}(s)), \quad \textit{for } i = 1, \ldots, p.$$

Theorem 2.6. *Assume $\phi_i \in C^{\zeta_i}[-1,1]$ with $\zeta_i \geq 2$, and $T = o(N^2)$. It holds that as $k \to \infty$,*

$$\|b_{k,i}(\cdot)\|_\infty = O(k^{-\zeta_0+3/2}), \quad \|\sigma^2_{k,i}(\cdot)\|_\infty = O(k), \quad \textit{for } i = 1, \ldots, p,$$

where $\zeta_0 = \min\{\zeta_1, \ldots, \zeta_p, 4\}$.

2.4 Simulations

In this section, we illustrate the performance of the proposed estimators, and demonstrate the impacts of k, T and N through two simulated examples.

For each model and combination of (k, T, N), we generate the corresponding functional time series 100 times and estimate the convolution functions. The performance of estimation is evaluated by the integrated squared error, $E_i = \|\widehat{\phi}_{k,i} - \phi_i\|_2^2$, for $i = 1, \ldots, p$.

Example 2.2. Consider a simple CFAR(1) model with convolution function $\phi(\cdot)$ being the normal density function with mean 0 and standard deviation 0.1, truncated at $[-1, 1]$. For noise process, we use $\rho = 5$ and $\sigma^2 = 10$. We demonstrate various features of the model with a typical data set whose estimation error is the median value of the 100 simulated sets for $k = 11$, $N = 100$, and $T = 100$.

Table 2.1 shows the average of E for different k, with fixed $T = 100$ and $N = 100$. The second column shows the distance between two adjacent knots. As k increases,

Table 2.1: The average (sd) of errors when $k = 7, 11, 19$ for Example 2.2

k	distance	N	T	E
7	1/2	100	100	0.4257 (0.0955)
11	1/4	100	100	0.1685 (0.0954)
19	1/8	100	100	0.3318 (0.1370)

average of E decreases first, due to improvement of the sieve approximation, and then increases, because of the introduction of extra B-spline coefficients. It reaches the minimum when $k = 11$.

Table 2.2: The average (sd) of errors when $T = 10, 100, 1000$ for Example 2.2

k	N	T	E
11	100	10	4.3748 (3.3497)
11	100	100	0.1685 (0.0954)
11	100	1000	0.0312 (0.0187)

Table 2.2 shows the estimation performance as the sample size T changes from 10, 100 to 1000, with fixed k and N. As T increases, the means and standard deviations of E all decrease, and the estimates become more accurate and stable. It can be seen that E decreases approximately at the rate of $1/T$.

Table 2.3: The average (sd) of errors when $N = 10, 100, 1000$ for Example 2.2

k	N	T	E
11	10	100	0.2015 (0.1001)
11	100	100	0.1685 (0.0954)
11	1000	100	0.1687 (0.0968)

Table 2.3 summarizes the estimation performance for $k = 11$ and $T = 100$ with different N. It is seen that when N is small, increasing N improves the performance of the estimators as they benefit from the increase of information. However, when N is large enough, E stays at the same level, because there are sufficient observations to get an accurate approximation of $X_t(\cdot)$ by linear interpolation. Due to the strong spatial correlation in the error process, dense observations do not provide much extra information.

Figure 2.2 displays the predicted $X_t(\cdot)$ using (2.19) from time 2 to 9, using $k = 15$, $T = 100$ and $N = 100$ for the typical data set. Because the estimated $\phi(\cdot)$ is very close

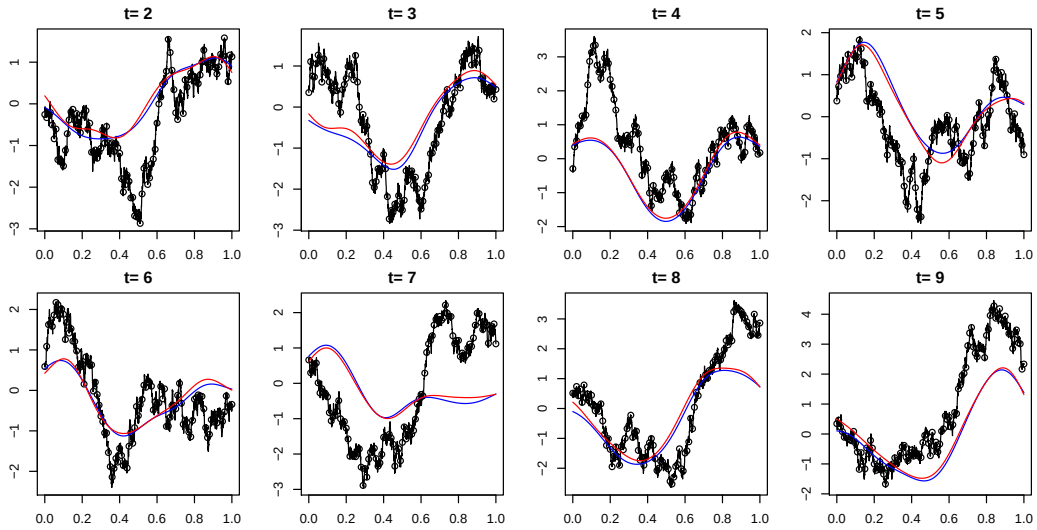


Figure 2.2: Plots of predicted $X_t(\cdot)$ when $k = 15$, $N = 100$, and $T = 100$, $t = 2, 3, 4, 5, 6, 7, 8, 9$, for the typical data set from Example 2.2. The black lines are $X_t(\cdot)$; the blue lines are $f_t(\cdot)$; the red lines are the predictions; the black circles are the observations.

to the true convolution function, the predictions (red lines) are also close to the skeleton (blue lines). Also note that the noisy process is large in magnitude with strong spatial correlation, hence simple smoothing of $X_t(\cdot)$ will be far away from the true skeleton $f_t(\cdot)$.

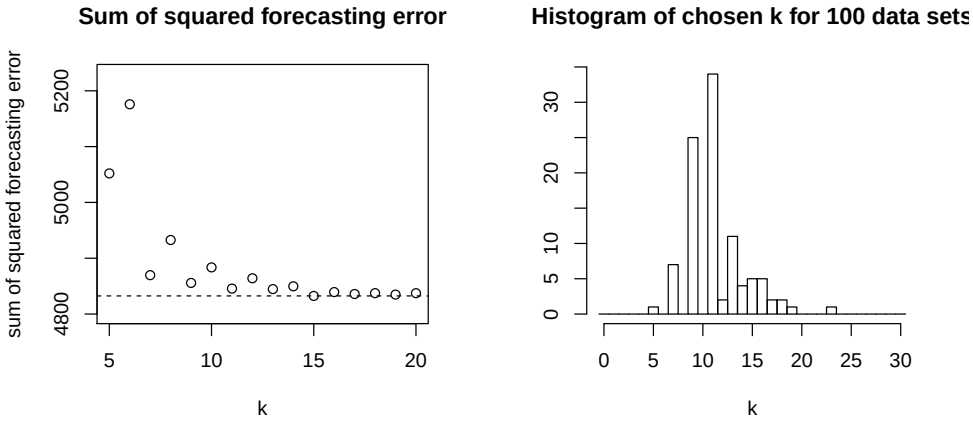


Figure 2.3: Left panel: plot of the mean squared out-sample forecasting error against k for the typical data set when $T = 100$ and $N = 100$ from Example 2.2; right panel: histogram of chosen k for 100 data sets when $T = 100$ and $N = 100$ from Example 2.2.

Figure 2.3 shows the performance of using out-sample forecasting to choose k, for $T = 100$, $T_0 = T/2$ and $N = 100$. The left panel in Figure 2.3 displays the sum of

squared forecasting error across k for the typical data set. The minimum is achieved at $k = 15$. It is seen that the sum of squared forecasting error experiences two different phases as k increases. At the beginning, it is very large, due to the lack of knots to approximate $\phi(\cdot)$ well. As the complexity of sieve space k increases, the squared forecasting error reduces very quickly, and the performance of estimator improves, until its minimum is reached. In this phase, the bias dominates the estimation error. After that, the error increases gradually with k because the fixed sample size T is not sufficient for the extra knots estimation. In this phase the variance dominates the error.

The right panel in Figure 2.3 shows the histogram of chosen k for the 100 data sets. It is truncated at $k = 5$, and most data sets require the dimension of the sieve space to be greater than 8. Note that when k is odd, the prediction error is smaller than that when k is even. This is because $\phi(\cdot)$ reaches its maximum value and maximum second derivatives at 0. The estimate benefits from a knot at 0.

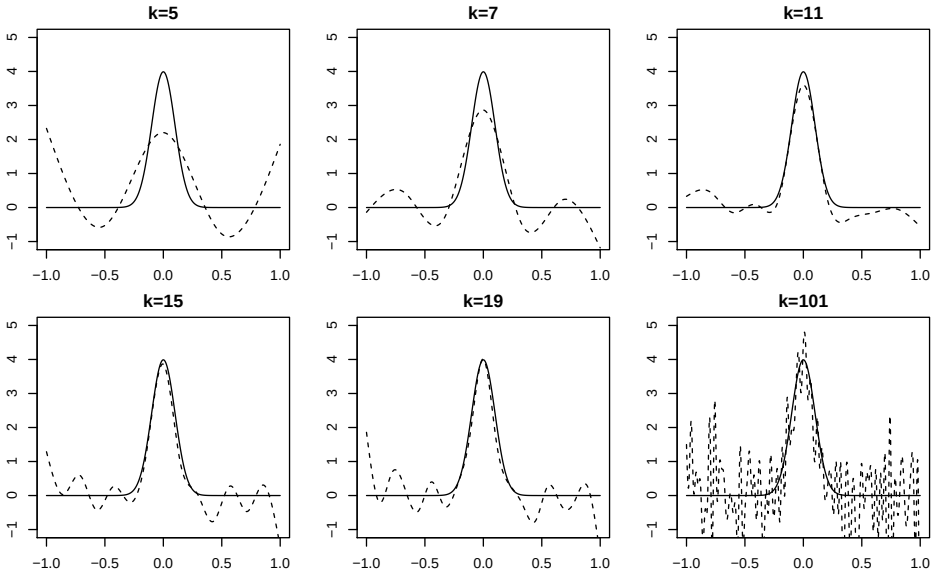


Figure 2.4: Plots of $\phi(\cdot)$ (solid line) and $\widehat{\phi}(\cdot)$ (dashed line) with $T = 100$, $N = 100$, and $k = 5, 7, 11, 15, 19, 101$ for the typical data set from Example 2.2.

Figure 2.4 shows the estimated $\phi(\cdot)$ for the typical data set when $k = 5, 7, 11, 15, 19, 101$. As expected, the accuracy of the estimate improves first and then becomes worse as k increases. The top panel shows the first phase where the bias dominates, and the bottom panel shows the second phase where the variance dominates. When $k = 11$,

$\widehat{\phi}(\cdot)$ is very close to the true convolution function, shown in the top right plot of Figure 2.4.

Table 2.4 reports the performance of the F-test, where the first row is for testing $H_0 : \phi = 0$ under CFAR(1) model, and the second row is for testing $H_0 : \phi_2 = 0$ under CFAR(2) model, comparing with CFAR(1) model. The second column shows the p-values for the typical data set. The last 5 columns show the minimum, mean, maximum and standard deviation of the p-values, and frequency of rejection at 5% level, respectively. $\phi_1(\cdot)$ is significant for all the data set, and $\phi_2(\cdot)$ is not significant for most of them.

Table 2.4: The p-values for $\phi_i = 0$, when $k = 11$, $T = 100$, $N = 100$ for Example 2.2

H_0	Example	Overall				Significance
		Min	Mean	Max	Sd	frequency
$\phi_1 = 0$	< 0.0001	< 0.0001	< 0.0001	< 0.0001	< 0.0001	100%
$\phi_2 = 0$	0.8442	0.0286	0.5698	0.9832	0.2679	2%

Example 2.3. A CFAR(2) model is considered in this example. Here we use $\phi_1(s) = \frac{\sqrt{50}}{2\sqrt{\pi}} e^{-50s^2}$, and $\phi_2(s) = \frac{1}{2}\cos(2\pi s)$, for $s \in [-1, 1]$. We fix $\rho = 5$, and $\sigma^2 = 10$. Again, a typical data is selected whose estimation error is the median of these of 100 data sets when $k = 11$, $T = 500$ and $N = 100$.

Table 2.5: The average (sd) of errors when $k = 5, 7, 11, 19$ for Example 2.3

k	distance	N	T	E_1	E_2
5	1	100	500	0.2711 (0.0406)	0.0464 (0.0216)
7	1/2	100	500	0.1227 (0.0317)	0.0446 (0.0226)
11	1/4	100	500	0.0623 (0.0335)	0.0596 (0.0286)
19	1/8	100	500	0.1029 (0.0434)	0.0972 (0.0410)

Table 2.6: The average (sd) of errors when $T = 100, 200, 500$ for Example 2.3

k	N	T	E_1	E_2
11	100	100	0.3000 (0.1626)	0.3128 (0.1822)
11	100	200	0.1493 (0.0693)	0.1406 (0.0838)
11	100	500	0.0623 (0.0335)	0.0596 (0.0286)

Tables 2.5-2.7 show the average and standard deviation of E_1 and E_2 with different combination of (k, T, N). The pattern is similar to that in Example 2.2. As k increases,

Table 2.7: The average (sd) of errors when $N = 100, 200, 500$ for Example 2.3

k	N	T	E_1	E_2
11	100	500	0.0623 (0.0335)	0.0596 (0.0286)
11	200	500	0.0623 (0.0334)	0.0595 (0.0285)
11	500	500	0.0622 (0.0330)	0.0593 (0.0284)

the errors decrease first, reach the minimum, then increase. E_1 and E_2 are roughly linear in T, but does not change too much with N.

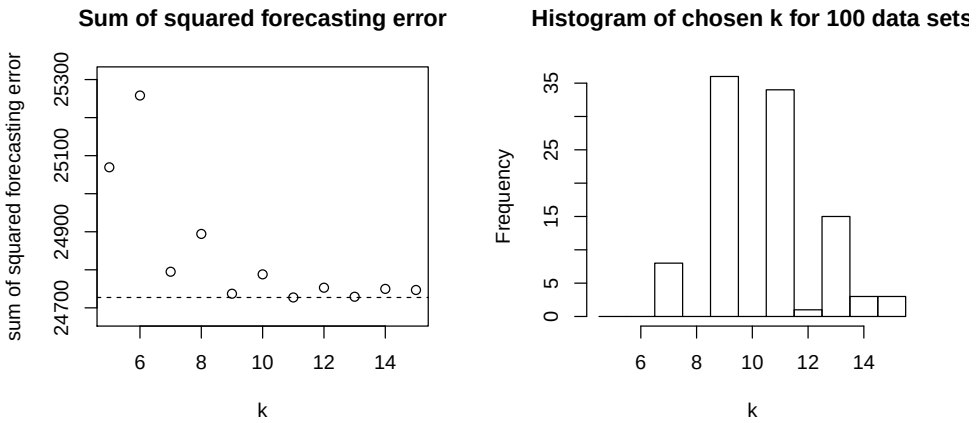


Figure 2.5: Left panel: plot of the mean squared out-sample forecasting error against k for the typical data set when $T = 500$ and $N = 100$ from Example 2.3; right panel: histogram of chosen k for 100 data sets when $T = 500$ and $N = 100$ from Example 2.3.

Figure 2.5 reads similarly to Figure 2.3. For most of the data sets, $k = 9$, or $k = 11$ are selected based on the out-sample prediction criterion, when $T = 500$, $T_0 = T/2$, and $N = 100$. The estimated $\phi_1(\cdot)$ and estimated $\phi_2(\cdot)$ for the typical data set are shown in Figure 2.6, when $k = 11$, $T = 500$ and $N = 100$. The estimated function is more accurate around $s = 0$ than that around the boundary. This is because $\phi_i(u)$'s are used as weights for obtaining $X_t(s)$, when $u \in [-1 + s, s]$. Specifically $\phi_i(0)$'s are the convolution weights for $X_t(s)$ with $u \in [-1+s, s]$. Every $X_t(s)$, $s \in [-1, 1]$ contains information of $\phi_i(0)$ but only $X_t(-1)$ contains information about $\phi_i(-1)$. Hence, the estimate of $\phi_i(\cdot)$ in the center exploits data more efficiently than that close to the boundary.

Table 2.8 summarizes the p-values of a sequence of tests. The first row is for testing $H_0 : \phi = 0$ under CFAR(1) model. The second row is for testing $H_0 : \phi_2 = 0$ under

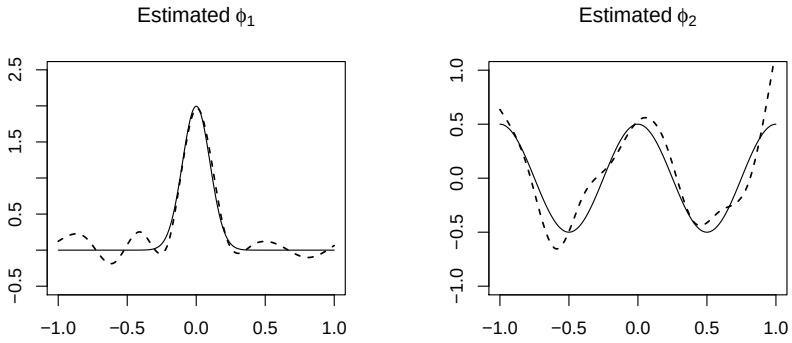


Figure 2.6: Plots of estimated $\phi_1(\cdot)$ and $\phi_2(\cdot)$ when $k = 11$, $T = 500$ and $N = 100$ for the typical data set from Example 2.3.

CFAR(2) model, comparing with CFAR(1) model in the F-test. The third row is defined similarly. Based on the p-values shown in the second column, CFAR(2) would be correctly specified for the typical data set. The last 5 columns show that F-tests perform well in determining the AR order.

Table 2.8: The p-values for $\phi_i = 0$, when $k = 11$, $T = 500$ $N = 100$ from Example 2.3

p-value	Example	Overall				Significance
		Min	Mean	Max	Sd	frequency
$\phi_1 = 0$	< 0.0001	< 0.0001	< 0.0001	< 0.0001	< 0.0001	100%
$\phi_2 = 0$	< 0.0001	< 0.0001	< 0.0001	< 0.0001	< 0.0001	100%
$\phi_3 = 0$	0.0797	0.0119	0.4906	0.9966	0.2678	2%

2.5 Real Data Analysis

We apply our method to the S&P 500 index European call option data in this section. It is well-known that the implied volatility of an option is a function of its strike prices, and this phenomenon is called volatility smiles. In this dissertation, we treat the implied volatility as a function of moneyness, which is the relative strike price with respective to the price of underlying asset. The option data is collected from July 9, 2004 to September 20, 2004, and the expiration date of the options is December 18, 2004. Hence, the time series is $T = 51$ long. We select the options with strikes between 950 and 1550, which were actively traded within this period, and have a solution for implied volatility derived from Black-Scholes model most of the time. Number of observations

at each time varies from 43 to 48. Our aim is to construct the volatility curve against moneyness, study its evolution mechanism, and predict implied volatility in the future.

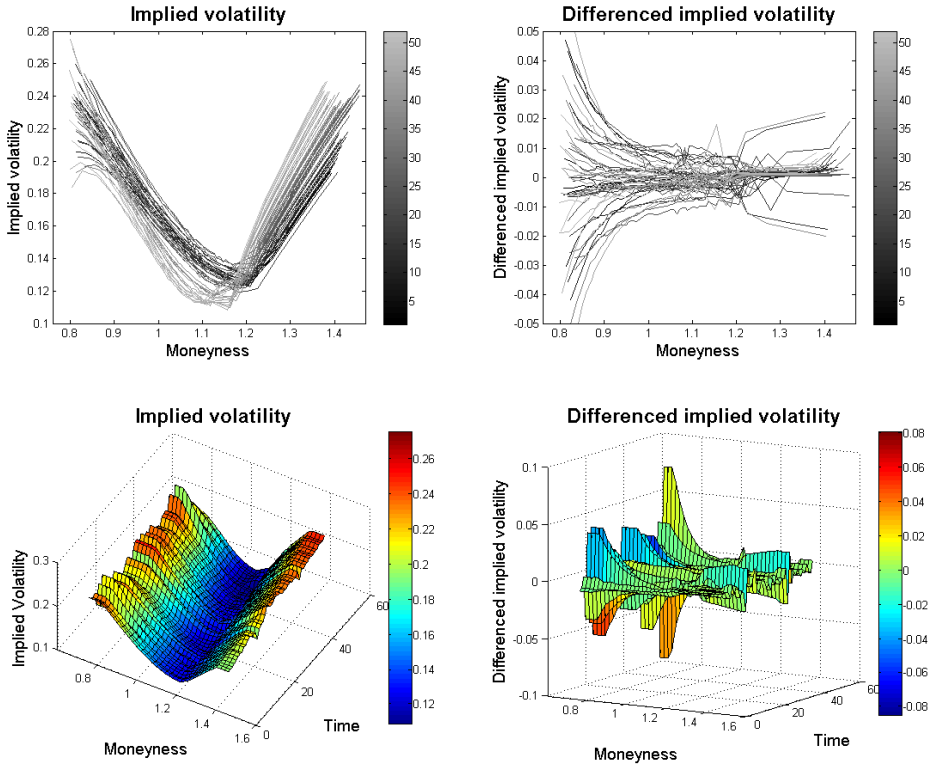


Figure 2.7: Plots of implied volatilities against moneyness (top left panel) across time (bottom left panel) and plots of differenced implied volatility against moneyness (top right panel) across time (bottom right panel). The colorbars on the right show the time (top panel) and implied volatility (bottom panel) corresponding to different color scales.

The left panel of Figure 2.7 shows the implied volatility curves as functions of moneyness over time, where the volatility smiles are clearly demonstrated, since deep in-the-money or out-the-money options have larger volatility than others. As expiration approaches, the implied volatilities of at-the-money options decreases, the implied volatilities of out-the-money options increase, and the location where minimum reaches approaches to 1 seen from the top left panel of Figure 2.7. Hence, we take a difference of the implied volatilities, and model them under CFAR settings. Specifically, let $Y_t(s_n) = X_t(s_n) - \widetilde{X}_{t-1}(s_n)$, for $n = 1, \ldots, N_t$, $t = 2, \ldots, T$, where $X_t(s_n)$ is the implied volatility at time t with moneyness s_n, $\widetilde{X}_t(\cdot)$ is the linear interpolation of $X_t(\cdot)$ defined in (2.8), and N_t is number of observations at time t. Data after taking a difference is

plotted in the right panel of Figure 2.7. The implied volatilities of deep in-the-money and out-the-money options are much more volatile than these of at-the-money options.

Table 2.9: The p-values for $H_0 : \phi_i = 0$, when $k = 11$

H_0	$\phi_1 = 0$	$\phi_2 = 0$
p-value	< 0.0001	0.1027

The sequential F-tests of testing $H_0 : \phi_i = 0$ vs $H_1 : \phi_i \neq 0$ for $i = 1, 2$ in Table 2.9 indicate that CFAR(1) model should be chosen to fit the data. The estimates of $\phi(\cdot)$ for $k = 12$ (top left panel), $k = 16$ (top right panel) and $k = 18$ (bottom left panel) are shown in Figure 2.8. It can be seen that $\widehat{\phi}(\cdot)$ is not very sensitive to the number of knots. The bottom right panel of Figure 2.8 displays the sum of squared out-sample forecasting errors against k when $T_0 = 4/5T$. When $k = 18$, the errors reaches the minimum value, hence we select $k = 18$. $\widehat{\phi}(s)$ is negative when $s \leq 0.1$ or $0.65 \leq s \leq 0.9$; otherwise, it is positive, for $k = 18$ from bottom right panel of Figure 2.8. Hence, the implied volatility with moneyness s at time t is likely to increase, when these with moneyness u, $u \geq s-0.1$, $s-0.65 \leq u \leq s-0.9$ decrease, and others increase at time $t - 1$.

Define

$$\widehat{e}_t(n) = \widehat{\varepsilon}_t(s_n) - e^{-\widehat{\rho}|s_n - s_{n-1}|}\widehat{\varepsilon}_t(s_{n-1}), \quad n = 2, \ldots, N_t.$$

The sample autocorrelation functions of $\widehat{e}_t(n)$ are shown in Figure 2.9 for $t = 2, \ldots, 37$. Most of the time, the sample autocorrelations of cross-sectional errors lie within confidence intervals, except $t = 8, 14, 15, 30, 37$. Overall, most of the sample autocorrelations of $\widehat{e}_t(s_n)$ are not significant, which implies that the CFAR(1) model fits data well.

One-step-ahead out-sample forecasting error of $X_t(\cdot)$ is used to compare the prediction performance of the estimated CFAR(1) model with a functional random walk model , where $\widehat{X}_{t+1}(s_n) = 0$, $\forall s_n[0, 1]$. Table 2.10 shows the estimated ρ and σ^2 using CFAR(1) model, and the mean squared out-sample forecasting errors for both models. CFAR(1) outperforms the functional random walk model, reducing the MSE by about 9%.

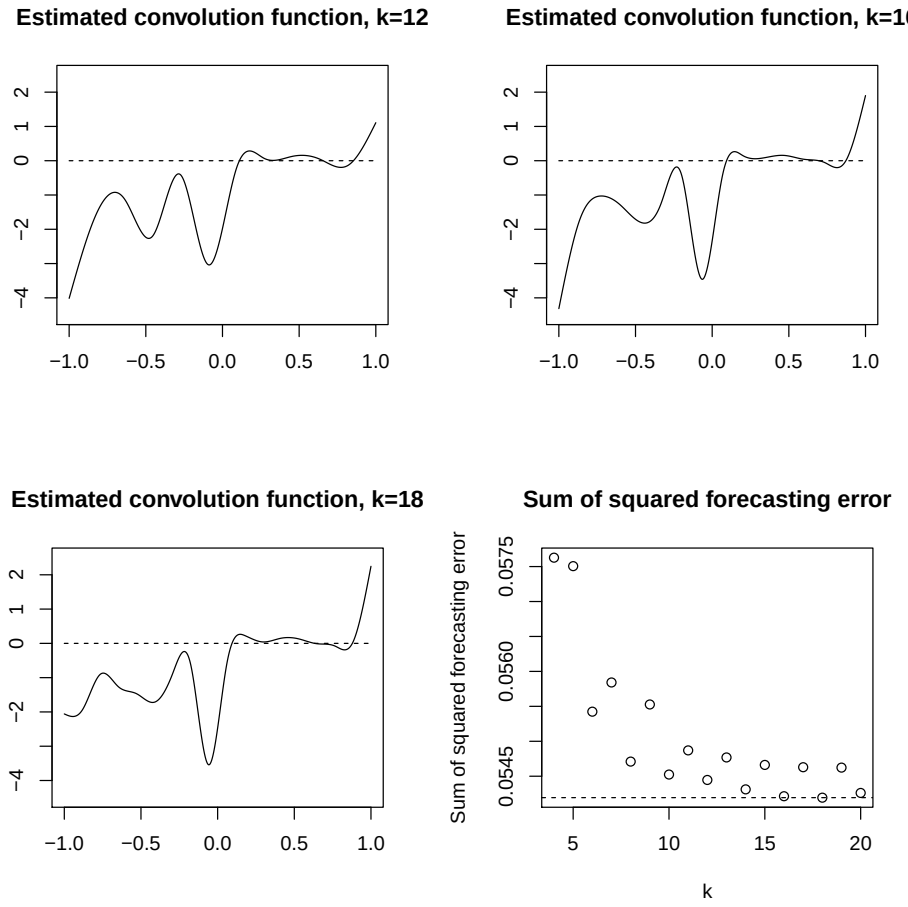


Figure 2.8: Plots of $\widehat{\phi}(\cdot)$ for different k, $k = 12$ (top left panel), $k = 16$ (top right panel) and $k = 18$ (bottom left panel), and plot of the sum of squared out-sample forecasting error against k (bottom right panel).

Table 2.10: Parameter estimates for CFAR model and out-sample forecasting MSEs for both models

$\widehat{\rho}$	$\widehat{\sigma}^2$	MSE1(CFAR model)	MSE2(random walk)
0.0001	$1.5176e-04$	$1.167e-04$	$1.2755e-04$

2.6 Proofs

Proof of Theorem 2.1: By Cauchy-Schwarz inequality, for any function $x \in \mathrm{L}_2[0,1]$,

$$\begin{aligned}\Big\| \int_0^1 \phi(s-u)x(u)du \Big\|_2^2 &= \int_0^1 \left(\int_0^1 \phi(s-u)x(u)du \right)^2 ds \\ &\leq \int_0^1 \left(\int_0^1 \phi^2(s-u)du \right) \left(\int_0^1 x^2(u)du \right) ds \leq \kappa^2 \|x\|.\end{aligned}$$

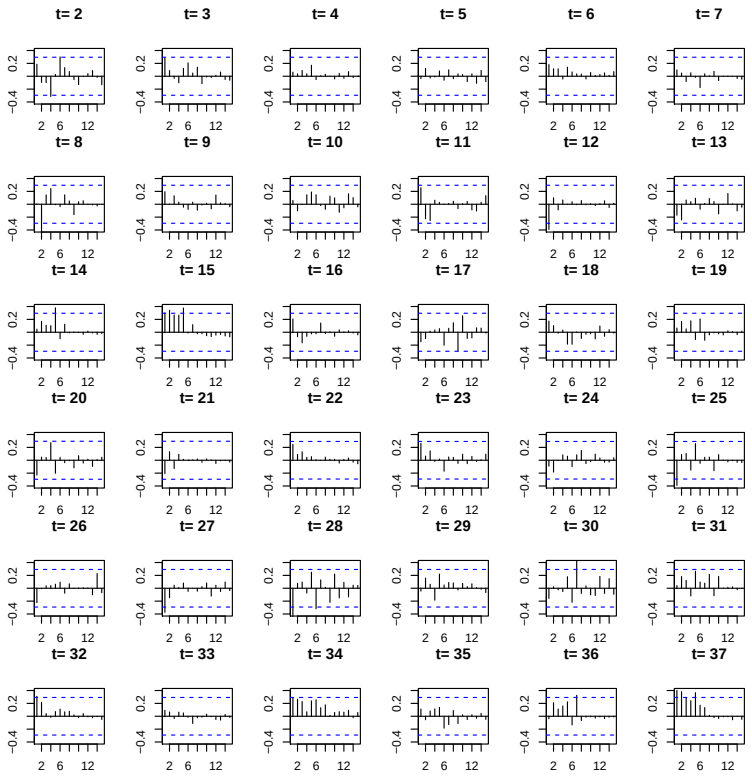


Figure 2.9: Sample autocorrelations of $\widehat{e}_t$ with lag 0 autocorrelation removed when $t = 2, \ldots, 37$.

It follows that

$$\sup_{\|x\| \leq 1} \| \int_0^1 \phi(s-u)x(u)du\|_2^2 \leq \kappa^2 < 1.$$

By Lemma 3.1 in Bosq (2000), X_t is stationary. ∎

Proof of Theorem 2.2: Let $\mathbf{H}^p[0,1]$ be the Cartesian product of p copies of the random function space $\mathbf{H}[0,1]$. The norm in $\mathbf{H}^p[0,1]$ is defined as

$$\|(X_1, \ldots, X_p)\|_p = \sqrt{\sum_{i=1}^{p} \|X_1\|_2^2}, \quad \text{where} \quad X_1, \ldots, X_p \in \mathbf{H}[0,1].$$

Consider

$$\mathbf{\Delta} = \begin{bmatrix} \Delta_1 & \Delta_2 & \ldots & \Delta_p \\ I & 0 & \ldots & 0 \\ 0 & I & \ldots & 0 \\ 0 & \ldots & I & 0 \end{bmatrix}, \quad \mathbf{K} = \begin{bmatrix} \kappa_1 & \kappa_2 & \ldots & \kappa_p \\ 1 & 0 & \ldots & 0 \\ 0 & 1 & \ldots & 0 \\ 0 & \ldots & 1 & 0 \end{bmatrix}, \tag{2.27}$$

where I denotes the identity operator, Δ_i is a convolution operator associating with ϕ_i, i.e. $\Delta_i X = \int_0^1 \phi_i(\cdot - u)X(u)du$. Note that $\|\mathbf{\Delta X}\|_p \leq \|\mathbf{KX}\|_p$. And all the roots of the characteristic function are eigenvalues of the matrix $\mathbf{K}$. Let $\lambda_{\max}$ be the maximum eigenvalue in modulus of $\mathbf{K}$, and $|\lambda_{\max}| < 1$. Hence, there exists an integer j, such that $\|\mathbf{\Delta}^j\| \leq \|\mathbf{K}^j\| < 1$. By Theorem 5.1 in Bosq (2000), CFAR(p) model is stationary if (2.4) is satisfied. ∎

Lemma 2.1. *Let $f, g \in \text{Lip}_2^{\zeta}[-1,1]$, and $\zeta > 1$, $\{s_0 = 0, s_1, \ldots, s_N = 1\}$ is an equally spaced partition of $[0,1]$. For any $u, v \in [0,1]$, let $\mathbf{f}_u$ and $\mathbf{g}_v$ be $(N+1) \times 1$ vectors, $\mathbf{f}_u = (f(s_0 - u), f(s_1 - u), \ldots, f(s_N - u))'$, $\mathbf{g}_v = (g(s_0 - v), g(s_1 - v), \ldots, g(s_N - v))'$. $\mathbf{\Sigma}$ is defined in Section 3.1, where $(\mathbf{\Sigma})_{ij} = e^{-\rho|i-j|/N}\sigma^2/2\rho$. Define $A_N(f,g)(u,v) = \mathbf{f}_u'\mathbf{\Sigma}^{-1}\mathbf{g}_v$, then*

$$A_N(f,g)(u,v) = \frac{1}{\sigma^2}\mathcal{A}(f,g)(u,v) + O(N^{-\zeta_0+1}), \tag{2.28}$$

where $\zeta_0 = \min\{\zeta, 2\}$.

Proof: Let $\varphi = e^{-\rho/N}$, and it is easy to show that $\mathbf{\Sigma}^{-1} = \frac{2\rho}{\sigma^2}\mathbf{U}'\mathbf{U}$, where

$$\mathbf{U} = \frac{1}{\sqrt{1-\varphi^2}} \begin{pmatrix} \sqrt{1-\varphi^2} & 0 & \cdots & 0 \\ -\varphi & 1 & \cdots & 0 \\ 0 & \cdots & -\varphi & 1 \end{pmatrix}. \tag{2.29}$$

And we have

$$\mathbf{U}\mathbf{f}_u = \frac{1}{\sqrt{1-\varphi^2}}$$
$$\begin{pmatrix} \sqrt{1-\varphi^2}f(s_0-u) & f(s_1-u) - \varphi f(s_0-u) & \ldots & f(s_N-u) - \varphi f(s_{N-1}-u) \end{pmatrix}',$$

$\mathbf{U}\mathbf{g}_v$ has a similar expression, hence

$$\begin{aligned}
\mathbf{f}_u'\boldsymbol{\Sigma}^{-1}\mathbf{g}_v &= \frac{2\rho}{\sigma^2}f(-u)g(-v) + \frac{2\rho}{(1-\varphi^2)\sigma^2} \\
&\quad \sum_{n=1}^{N}\left(f(s_n-u)-\varphi f(s_{n-1}-u)\right)\left(g(s_n-v)-\varphi g(s_{n-1}-v)\right) \\
&= \frac{2\rho}{\sigma^2}f(-u)g(-v) + \frac{2\rho}{(1-\varphi^2)\sigma^2} \\
&\quad + \sum_{n=1}^{N}\left[f(s_n-u)-f(s_{n-1}-u)\right]\left[g(s_n-v)-g(s_{n-1}-v)\right] \\
&\quad + \frac{2\rho}{(1-\varphi^2)\sigma^2}\sum_{n=1}^{N}(1-\varphi)\left[f(s_n-u)-f(s_{n-1}-u)\right]g(s_{n-1}-v) \\
&\quad + \frac{2\rho}{(1-\varphi^2)\sigma^2}\sum_{n=1}^{N}(1-\varphi)f(s_{n-1}-v)\left[g(s_n-v)-g(s_{n-1}-v)\right] \\
&\quad + \frac{2\rho}{(1-\varphi^2)\sigma^2}\sum_{n=1}^{N}(1-\varphi)^2 f(s_{n-1}-v)g(s_{n-1}-v) \\
&= L_1+L_2+L_3+L_4+L_5.
\end{aligned}$$

Since $f, g \in \mathrm{Lip}_2^{\zeta}[-1,1]$, $|f'(u)-f'(v)| \le M|u-v|$, if $\zeta \ge 2$; $|f'(u)-f'(v)| \le M|u-v|^{-\zeta+1}$, if $\zeta < 2$, for some positive constant M. It follows that

$$\begin{aligned}
L_2 &= \frac{2\rho}{N(1-\varphi^2)\sigma^2}\sum_{n=1}^{N}\frac{1}{N}\left(\frac{f(s_n-u)-f(s_{n-1}-u)}{1/N}\right)\left(\frac{g(s_n-v)-g(s_{n-1}-v)}{1/N}\right) \\
&= \frac{2\rho}{N(1-\varphi^2)\sigma^2}\sum_{n=1}^{N}\frac{1}{N}(f'(s_{n-1}-u)+O(N^{-\zeta_0+1}))(g'(s_{n-1}-u)+O(N^{-\zeta_0+1})) \\
&= \frac{1}{\sigma^2}\int_0^1 f'(s-u)g'(s-v)\,ds + O(N^{-\zeta_0+1}).
\end{aligned}$$

Similarly, we obtain $L_3 = \frac{\rho}{\sigma^2}\int_0^1 f'(s-u)g(s-v)ds + O(N^{-\zeta_0+1})$, $L_4 = \frac{\rho}{\sigma^2}\int_0^1 f(s-u)g'(s-v)ds + O(N^{-\zeta_0+1})$, and $L_5 = \frac{\rho^2}{\sigma^2}\int_0^1 f(s-u)g(s-v)ds + O(N^{-\zeta_0+1})$.

Using integration by parts, we have

$$\begin{aligned}
\mathbf{f}_u'\boldsymbol{\Sigma}^{-1}\mathbf{g}_v &= \frac{1}{\sigma^2}\Big(\rho f(-u)g(-v) + \rho f(1-u)g(1-v) \\
&\quad + \int_0^1 f'(s-u)g'(s-v)\,ds + \rho^2\int_0^1 f(s-u)g(s-v)\,ds\Big) + O(N^{-\zeta_0+1}).
\end{aligned}$$

■

Lemma 2.2. *If the CFAR(1) process $X = (X_t, t \in \mathbb{Z})$ satisfies (2.3) and $\phi \in C^2[-1,1]$, then there exists a positive constant ι which only depends on ϕ and ρ such that:*

(i) $\|\Psi_h\|_\infty \le \kappa^h$ *for* $h \ge 2$.

(ii) $\|\gamma_h\|_\infty \le \sigma^2 \iota \kappa^{|h|}$ *for* $h \in \mathbb{Z}$.

(iii) For $h \ge 1$,

$$\max_{u,v} \left\{ \left| \frac{\partial \gamma_h(u,v)}{\partial u} \right|, \left| \frac{\partial \gamma_h(u,v)}{\partial v} \right| \right\} \le \sigma^2 \iota \kappa^h,$$

and when $h = 0$, *for any* $u_1, u_2, v \in [0,1]$,

$$|\gamma(u_1, v) - \gamma(u_2, v)| \le \sigma^2 \iota |u_1 - u_2|.$$

(iv) For $h \ge 1$,

$$\max_{u,v} \left\{ \left| \frac{\partial^2 \gamma_h(u,v)}{\partial u^2} \right|, \left| \frac{\partial^2 \gamma_h(u,v)}{\partial u \partial v} \right|, \left| \frac{\partial^2 \gamma_h(u,v)}{\partial v^2} \right| \right\} \le \sigma^2 \iota \kappa^h,$$

and when $h = 0$,

$$\max_{u \ne v} \left\{ \left| \frac{\partial^2 \gamma(u,v)}{\partial u^2} \right|, \left| \frac{\partial^2 \gamma(u,v)}{\partial u \partial v} \right|, \left| \frac{\partial^2 \gamma(u,v)}{\partial v^2} \right| \right\} \le \sigma^2 \iota.$$

Proof: By Cauchy-Schwarz inequality, for any $u, v \in [0,1]$,

$$|\Psi_2(u,v)| = \left| \int_0^1 \phi(u-s)\phi(s-v)\, ds \right| \le \left(\int_0^1 \phi^2(u-s) ds \right)^{1/2} \cdot \left(\int_0^1 \phi^2(s-v) ds \right)^{1/2} \le \kappa,$$

$$|\Psi_h(u,v)| \le \left(\int_0^1 \Psi_{h-1}^2(u,s) ds \right)^{1/2} \left(\cdot \int_0^1 \phi^2(s-v) ds \right)^{1/2} \le \kappa \|\Psi_{h-1}\|_\infty, \text{ for } h \ge 3.$$

Hence, Lemma 2.2(i) can be proved inductively.

By the definition of Ψ, for $h \ge 2$,

$$\Psi_h(u,v) = \int_{[0,1]^h} \phi(u-s_1)\phi(s_1-s_2)\ldots\phi(s_{h-1}-v)\, ds_1 \ldots ds_{h-1},$$

then we have

$$\frac{\partial \Psi_h(u,v)}{\partial u} = \int_{[0,1]^h} \phi'(u-s)\Psi_{h-1}(s,v)\,ds, \quad \frac{\partial \Psi_h(u,v)}{\partial v} = \int_{[0,1]^h} \Psi_{h-1}(u,s)\phi'(s-v)\,ds,$$

$$\frac{\partial^2 \Psi_h(u,v)}{\partial u^2} = \int_{[0,1]^h} \phi''(u-s)\Psi_{h-1}(s,v)\,ds, \quad \frac{\partial^2 \Psi_h(u,v)}{\partial v^2} = \int_{[0,1]^h} \Psi_{h-1}(u,s)\phi''(s-v)\,ds,$$

Let $d_1 = \max_{-1\le s\le 1} |\phi'(s)|$ and $d_2 = \max_{-1\le s\le 1} |\phi''(s)|$, and it holds that

$$\max_{u,v} \Big\{ \Big| \frac{\partial \Psi_h(u,v)}{\partial u} \Big|, \quad \Big| \frac{\partial \Psi_h(u,v)}{\partial v} \Big| \Big\} \le d_1 \kappa^{h-1}, \text{ for } h \ge 1,$$

$$\max_{u\neq v} \Big\{ \Big| \frac{\partial^2 \Psi_h(u,v)}{\partial u^2} \Big|, \quad \Big| \frac{\partial^2 \Psi_h(u,v)}{\partial v^2} \Big| \Big\} \le \quad d_2 \kappa^{h-1}, \text{ for } h \ge 1.$$

By Corollary 2.1, the covariances $\gamma(u,v)$ have the expression:

$$\gamma(u,v) = \frac{\sigma^2}{2\rho} e^{-\rho|u-v|} + \frac{\sigma^2}{2\rho} \sum_{\ell=1}^{\infty} \int_0^1 \int_0^1 \Psi_\ell(u,w)\Psi_\ell(v,z) e^{-\rho|w-z|}\,dwdz.$$

It follows that, for any $u, v \in [0,1]$,

$$\|\gamma\|_\infty \le \sigma^2/[2\rho(1-\kappa^2)],$$

$$|\gamma(u_1,v) - \gamma(u_2,v)| \le \frac{\sigma^2}{2\rho}(\rho + \frac{d_1\kappa}{1-\kappa^2})|u_1 - u_2|,$$

$$\max_{u\neq v} \left\{ \left| \frac{\partial^2 \gamma(u,v)}{\partial u^2} \right|, \left| \frac{\partial^2 \gamma(u,v)}{\partial u \partial v} \right|, \left| \frac{\partial^2 \gamma(u,v)}{\partial v^2} \right| \right\} \le \frac{\sigma^2}{2\rho}(\rho^2 + \frac{d}{1-\kappa^2}),$$

where $d = \max\{d_1^2, d_2\kappa\}$.

For $h \ge 1$, the autocovariances $\gamma_h(u,v) = \mathrm{Cov}(X_t(u), X_{t+h}(v))$ is given by

$$\gamma_h(u,v) = \frac{\sigma^2}{2\rho} \int_0^1 e^{-\rho|u-s|} \Psi_h(v,s)\,ds + \frac{\sigma^2}{2\rho} \sum_{\ell=1}^{\infty} \int_0^1 \int_0^1 \Psi_\ell(u,w)\Psi_{\ell+h}(v,z) e^{-\rho|w-z|}\,dwdz.$$

Hence, for $h \ge 1$, we have

$$\|\gamma_h\|_\infty \le \sigma^2 \kappa^{|h|}/[2\rho(1-\kappa^2)],$$

$$\|\frac{\partial\gamma_h(u,v)}{\partial u}\|_\infty \le \frac{\sigma^2}{2\rho}\left(\rho+\frac{d_1\kappa}{1-\kappa^2}\right)\kappa^h,$$

$$\|\frac{\partial\gamma_h(u,v)}{\partial v}\|_\infty \le \frac{\sigma^2}{2\rho}\cdot\frac{d_1\kappa^{h-1}}{1-\kappa^2},$$

$$\max_{u\neq v}\left\{\left|\frac{\partial^2\gamma_h(u,v)}{\partial u^2}\right|,\ \left|\frac{\partial^2\gamma_h(u,v)}{\partial u\partial v}\right|,\ \left|\frac{\partial^2\gamma_h(u,v)}{\partial v^2}\right|\right\} \le \frac{\sigma^2}{2\rho}(\rho^2+\frac{d}{1-\kappa^2})\kappa^h,$$

Since $\gamma_h(u,v)=\gamma_{-h}(v,u)$, the proof of Lemma 2.2(ii), 2.2(iii) and 2.2(iv) is complete. ∎

Lemma 2.3. *The CFAR(1) process $X=(X_t,t\in\mathbb{Z})$ satisfies (2.3). Let ε_0^* be an i.i.d. copy of ε_0, and X_t^* be obtained by replacing ε_0 with ε_0^* in the definition of X_t. If $f:[0,1]^2\to\mathbb{R}$ be a continuous function, and set $\|f\|_\infty:=\max_{u,v}|f(u,v)|$. Define*

$$Y_t=\int_0^1\int_0^1 f(u,v)X_t(u)X_t(v)\,dudv, \quad \text{and} \quad Y_t^*=\int_0^1\int_0^1 f(u,v)X_t^*(u)X_t^*(v)\,dudv.$$

Then, for each $q\in\mathbb{N}$ and $q\ge 1$, and $t\ge 0$,

$$\|Y_t-Y_t^*\|_q \le \frac{\sqrt{2}\sigma^2}{\rho\sqrt{1-\kappa^2}}\left[(2q-1)!!\right]^{1/q}\|f\|_\infty\cdot\kappa^t.$$

Proof: By Cauchy-Schwarz inequality,

$$\begin{aligned}
\|Y_t-Y_t^*\|_q &\le 2\Big\|\int_0^1\int_0^1 f(u,v)[X_t(u)-X_t^*(u)]X_t(v)\,dudv\Big\|_q\\
&\le 2\|f\|_\infty\int_0^1\int_0^1\big\|[X_t(u)-X_t^*(u)]X_t(v)\big\|_q\,dudv\\
&\le 2\|f\|_\infty\int_0^1\int_0^1\big\|[\int_0^1\Psi_t(u,w)(\varepsilon_t(w)-\varepsilon_t^*(w)dw\big\|_{2q}\cdot\big\|X_t(v)\|_{2q}\,dudv\\
&\le \frac{\sqrt{2}\sigma^2}{\rho\sqrt{(1-\kappa^2)}}[(2q-1)!!]^{1/q}\|f\|_\infty\cdot\kappa^t.
\end{aligned}$$

∎

Proof of Theorem 2.3: Define $\delta_t(\cdot)=X_t(\cdot)-\widetilde{X}_t(\cdot)$. For any function $f\in C[-1,1]$, denote $(f*g)(u)=\int_0^1 f(u-v)g(v)dv$. We first observe that X_t can be decomposed as

$$X_t(s)=\sum_{j=1}^k\beta_{k,j}(B_{k,j}*\widetilde{X}_{t-1})(s)+(r_k*\widetilde{X}_{t-1})(s)+\varepsilon_t(s)+(\phi*\delta_t)(s).$$

Let $\widetilde{\mathbf{v}}_t$ and $\widetilde{\mathbf{w}}_t$ be two $(N+1)$-dimensional vectors whose i-th entries are $(r_k * \widetilde{X}_{t-1})((i-1)/N)$ and $(\phi * \delta_t)((i-1)/N)$ respectively. Let $\boldsymbol{\varepsilon}_t$ be the $(N+1)$-dimensional vector whose i-th entry is $\varepsilon_t((i-1)/N)$. The estimator $\widehat{\boldsymbol{\beta}}_k$ can be decomposed as

$$\widehat{\boldsymbol{\beta}}_k = \widetilde{\boldsymbol{\beta}}_k + \left(\sum_{t=2}^{T} \mathbf{M}_t' \boldsymbol{\Sigma}^{-1} \mathbf{M}_t\right)^{-1} \left(\sum_{t=2}^{T} \mathbf{M}_t' \boldsymbol{\Sigma}^{-1} (\widetilde{\mathbf{v}}_t + \boldsymbol{\varepsilon}_t + \widetilde{\mathbf{w}}_t)\right).$$

We claim that under the condition $T = o(N^2)$

$$\frac{1}{\sqrt{T}} \sum_{t=2}^{T} \mathbf{M}_t' \boldsymbol{\Sigma}^{-1} \widetilde{\mathbf{w}}_t = o_p(1). \tag{2.30}$$

Let $\boldsymbol{\mu}_{k,N} = E(\mathbf{M}_t' \boldsymbol{\Sigma}^{-1} \widetilde{\mathbf{v}}_t)$, $\boldsymbol{\Gamma}_{k,N} = E(\mathbf{M}_t' \boldsymbol{\Sigma}^{-1} \mathbf{M}_t)$, and $\mathbf{b}_{k,N} = \boldsymbol{\Gamma}_{k,N}^{-1} \boldsymbol{\mu}_{k,N}$, then

$$\begin{aligned} \widehat{\boldsymbol{\beta}}_k =& \widetilde{\boldsymbol{\beta}}_k + \mathbf{b}_{k,N} + \boldsymbol{\Gamma}_{k,N}^{-1} \frac{1}{T} \sum_{t=2}^{T} (\mathbf{M}_t' \boldsymbol{\Sigma}^{-1} \widetilde{\mathbf{v}}_t - \boldsymbol{\mu}_{k,N}) + \boldsymbol{\Gamma}_{k,N}^{-1} \frac{1}{T} \sum_{t=2}^{T} \mathbf{M}_t' \boldsymbol{\Sigma}^{-1} \boldsymbol{\varepsilon}_t \\ & - \boldsymbol{\Gamma}_{k,N}^{-1} \frac{1}{T} \sum_{t=2}^{T} (\mathbf{M}_t' \boldsymbol{\Sigma}^{-1} \mathbf{M}_t - \boldsymbol{\Gamma}_{k,N}) \mathbf{b}_{k,N} + o_p(T^{-1/2}). \end{aligned} \tag{2.31}$$

We claim that

$$\begin{aligned} \widehat{\boldsymbol{\beta}}_k =& \widetilde{\boldsymbol{\beta}}_k + \mathbf{b}_k + \boldsymbol{\Gamma}_k^{-1} \frac{1}{T} \sum_{t=2}^{T} (\mathbf{u}_t - \boldsymbol{\mu}_k) + \boldsymbol{\Gamma}_k^{-1} \frac{1}{T} \sum_{t=2}^{T} \mathbf{z}_t \\ & - \boldsymbol{\Gamma}_k^{-1} \frac{1}{T} \sum_{t=2}^{T} (\mathbf{A}_t - \boldsymbol{\Gamma}_k) \mathbf{b}_k + o_p(T^{-1/2}). \end{aligned} \tag{2.32}$$

If we represent $\varepsilon_t(u)$ as

$$\varepsilon_t(u) = \varepsilon_t(0) e^{-\rho u} + \sigma \int_0^u e^{-\rho(u-v)} \, dW_t(v),$$

then similar to proof of Lemma 2.1, we get

$$(\mathbf{z}_t)_j = \int_0^1 \left[\frac{2\rho}{\sigma^2} B_{k,j}(-u) \varepsilon_t(0) + \frac{1}{\sigma} \int_0^1 (\rho B_{k,j}(v-u) + B_{k,j}'(v-u)) \, dW_t(v)\right] X_{t-1}(u) \, du. \tag{2.33}$$

The vector-valued process $\mathbf{u}_t + \mathbf{z}_t + \mathbf{A}_t \mathbf{b}_k$ is a stationary process. For any fixed unit

vector $\boldsymbol{\theta} \in \mathbb{R}^k$, by Lemma 2.3, the physical dependence measures defined in Wu (2005) of the process $\boldsymbol{\theta}'(\mathbf{u}_t + \mathbf{z}_t + \mathbf{A}_t\mathbf{b}_k)$ decay geometrically fast, and therefore the martingale approximation used in Wu (2005) can be applied to obtain the central limit theorem for $\sum_{t=2}^{T} \boldsymbol{\theta}'(\mathbf{u}_t + \mathbf{z}_t + \mathbf{A}_t\mathbf{b}_k)$. By the Cramer-Wold device, we have

$$\sqrt{T}(\widehat{\boldsymbol{\beta}}_k - \widetilde{\boldsymbol{\beta}}_k - \mathbf{b}_k) \xrightarrow{d} N(\mathbf{0}, \boldsymbol{\Gamma}_k^{-1}\boldsymbol{\Upsilon}_k\boldsymbol{\Gamma}_k^{-1}),$$

where $\boldsymbol{\Upsilon}_k$ is the long-run variance of the process $\mathbf{u}_t + \mathbf{z}_t + \mathbf{A}_t\mathbf{b}_k$.

Now we prove the claim (2.30). We proceed by showing that each entry of the vector converges to zero in probability. The j-th entry of $\mathbf{M}_t\boldsymbol{\Sigma}^{-1}\widetilde{\mathbf{w}}_t$ can be written as

$$\int_0^1 \int_0^1 A_N(B_{k,j}, \phi)(u, v)\widetilde{X}_{t-1}(u)\delta_t(v)\, dudv.$$

By Lemma 2.1 , there exists a constant $c_1 > 0$ such that

$$|A_N(B_j, \phi)(u, v)| \leq c_1, \quad \text{for all } u, v, N. \tag{2.34}$$

Let

$$\widetilde{\gamma}_h(u, v) = \mathrm{Cov}(X_t(u), \delta_{t+h}(v)), \quad \bar{\gamma}_h(u, v) = \mathrm{Cov}(\delta_t(u), \delta_{t+h}(v)),$$

By Lemma 2.2, there exists a constant $c_2 > 0$ such that

$$\|\tilde{\gamma}_h\|_\infty \leq c_2\kappa^h/N, \qquad \|\bar{\gamma}_h\|_\infty \leq c_2\kappa^h/N.$$

It follows that

$$\mathrm{E}\int_0^1 \int_0^1 |A_N(B_{k,j}, \phi)(u, v)\delta_t(u)\delta_t(v)|\, dudv = O(N^{-1}),$$

so it suffices to prove $\sum_{t=2}^{T} y_{jt} = o_p(T^{1/2})$, where

$$y_{jt} = \int_0^1 \int_0^1 A_N(B_{k,j}, \phi)(u, v)X_{t-1}(u)\delta_t(v)\, dudv.$$

By (2.34),

$$\mathrm{E}y_{jt} = \int_0^1 \int_0^1 A_N(B_{k,j}, \phi)(u,v)[\gamma_1(u,v) - \widetilde{\gamma}_1(u,v)]\, dudv = O(1/N). \tag{2.35}$$

The autocovariance is

$$\mathrm{Cov}(y_{jt}, y_{j,t+h}) = \int_{[0,1]^4} A_N(B_{k,j}, \phi)(u_1,v_1) A_N(B_{k,j}, \phi)(u_2,v_2) \\ \times [\gamma_h(u_1,u_2)\bar{\gamma}_h(v_1,v_2) + \widetilde{\gamma}_{h+1}(u_1,v_2)\widetilde{\gamma}_{-h+1}(u_2,v_1)]\, du_1 du_2 du_3 du_4.$$

By (2.34) and Lemma 2.2, we know there exist a constant $c_3 > 0$ such that

$$|\mathrm{Cov}(w_{jt}, w_{j,t+h})| \le c_3 \kappa^{2h}/N.$$

It follows that

$$\mathrm{Var}\left(\sum_{t=2}^{T} y_{jt}\right) = O(T/N). \tag{2.36}$$

Combining (2.35) and (2.36), and using the condition $T = o(N^2)$, we have $\sum_{t=2}^{T} y_{jt} = o_p(T^{1/2})$, and the proof of claim (2.30) is complete.

The difference between the two expressions in (2.31) and (2.32) is due to the approximation of X_t by $\widetilde{X}_t$, so the proof of claim (2.32) is in the same fashion of that of claim (2.30). We omit the details. ∎

Lemma 2.4. *The distance of two adjacent knots for uniform cubic B-spline functions defined on* $[-1,1]$ *with degree of freedom* k *is* $1/m$*, and* $m = (k-3)/2$*. Define two* $(m+4)\times(m+4)$ *matrices* $\mathbf{P}_s$ *and* $\mathbf{V}_s$*:*

$$(\mathbf{P}_s)_{qj} = \int_{s/m}^{1+s/m} \int_{s/m}^{1+s/m} e^{-\rho|u-v|} B'_{k,q+m-1}(u) B'_{k,j+m-1}(v)\, dudv,$$
$$(\mathbf{V}_s)_{qj} = \rho^2 \int_{s/m}^{1+s/m} \int_{s/m}^{1+s/m} e^{-\rho|u-v|} B_{k,q+m-1}(u) B_{k,j+m-1}(v)\, dudv,$$

for $q, j = 1, \ldots, m+4$*. Let* $\mathbf{b} = (b_0, \ldots, b_{m+3})'$ *be a unit vector. There exists constants*

c_1 and c_2 such that if $s \in I_\alpha$, $\sum_{j=0}^{3} b_j^2 \geq \frac{1}{3m}$, and $m \geq c_1$, then

$$\mathbf{b}'(\mathbf{P}_s + \mathbf{V}_s)\mathbf{b} \geq c_2 \sum_{j=0}^{3} b_j^2.$$

Proof: Rescale the uniform B-spline functions,

$$B_q(s) = B_{k,m+q}(s/m) = [q, q+1, \ldots, q+4](\cdot - s)_+^3, \text{ for } q = 0, \ldots, m+3.$$

The support of $B_q(\cdot)$ is $[q, q+4]$ and we denote $B(u) \equiv B_{-2}(u)$. $\mathbf{P}_s$ and $\mathbf{V}_s$ can be written as

$$\begin{aligned}(\mathbf{P}_s)_{qj} &= \int_s^{m+s} \int_s^{m+s} e^{-\rho|u-v|/m} B'_{q-1}(u) B'_{j-1}(v)\, dudv, \\ (\mathbf{V}_s)_{qj} &= \frac{\rho^2}{m^2} \int_s^{m+s} \int_s^{m+s} e^{-\rho|u-v|/m} B_{q-1}(u) B_{j-1}(v)\, dudv,\end{aligned}$$

for $q, j = 1, \ldots, m+4$. For a fixed $3 < s < 4$, there are $m+4$ spline functions which are not identically zero on the interval $[s, m+s]$: $B_0(u), B_1(u), \ldots, B_{m+3}(u)$. For any numbers b_0, b_1, b_2, b_3, there exists a constant $c_3 > 0$ such that $\int_3^4 \left[\sum_{j=0}^{3} b_j B_j(s)\right]^2 ds \geq c_3 \sum_{j=0}^{3} b_j^2$. Thus there exists an $s \in [3, 4]$ such that

$$\left|\sum_{j=0}^{3} b_j B_j(s)\right| \geq \sqrt{c_3} \cdot \left(\sum_{j=0}^{3} b_j^2\right)^{1/2}.$$

On the other hand, there exists a constant c_4 such that for all $s \in [3, 4]$, $\left|\sum_{j=0}^{3} b_j B'_j(s)\right| \leq c_4 \cdot \left(\sum_{j=0}^{3} b_j^2\right)^{1/2}$. As a result, there exists an interval I_α of length $c_5 > 0$, which is contained in $[3, 4]$, such that for each s in this interval

$$\left|\sum_{j=0}^{3} b_j B_j(s)\right| \geq c_6 \cdot \left(\sum_{j=0}^{3} b_j^2\right)^{1/2},$$

where $0 < c_6 < 1$ is an absolute constant.

Define the functions:

$$f_1(u) = \sum_{j=0}^{3} b_j B_j(u), \quad f_2(u) = \sum_{j=4}^{m-1} b_j B_j(u), \quad f_3(u) = \sum_{j=m}^{m+3} b_j B_j(u),$$

and let $f(u) = f_1(u) + f_2(u) + f_3(u)$. Using the identity

$$e^{-\rho|u|/m} = \frac{1}{\pi} \int_{-\infty}^{\infty} e^{iu\lambda} \cdot \frac{m/\rho}{1 + (m\lambda/\rho)^2} \, d\lambda,$$

we have

$$\begin{aligned}
\mathbf{b}'\mathbf{P}_s\mathbf{b} &= \frac{1}{\pi} \int_s^{m+s} \int_s^{m+s} f'(u) f'(v) \int_{-\infty}^{\infty} e^{i(u-v)\lambda} \cdot \frac{m/\rho}{1 + (m\lambda/\rho)^2} \, d\lambda du dv \\
&= \frac{1}{\pi} \int_{-\infty}^{\infty} \left| \int_s^{m+s} f'(u) e^{iu\lambda} \, du \right|^2 \frac{m/\rho}{1 + (m\lambda/\rho)^2} \, d\lambda \\
&= \frac{1}{\pi} \int_{-\infty}^{\infty} \left| -f_1(s) e^{is\lambda} + f_3(m+s) e^{i(m+s)\lambda} - (i\lambda) \int_s^{m+s} f(u) e^{iu\lambda} \, du \right|^2 \frac{m/\rho}{1 + (m\lambda/\rho)^2} \, d\lambda.
\end{aligned}$$

Similarly,

$$\mathbf{b}'\mathbf{V}_s\mathbf{b} = \frac{\rho^2}{\pi m^2} \int_{-\infty}^{\infty} \left| \int_s^{m+s} f(u) e^{iu\lambda} \, du \right|^2 \frac{m/\rho}{1 + (m\lambda/\rho)^2} \, d\lambda.$$

Neuman (1981) showed the Fourier transform of the central spline $B_{-2}(u)$ is

$$\int_{-2}^{2} B_{-2}(u) e^{iu\lambda} \, du = \left[\frac{2\sin(\lambda/2)}{\lambda} \right]^4.$$

Let $g(\lambda) = \sum_{j=4}^{m-1} b_j e^{i(j+2)\lambda}$, we have

$$\begin{aligned}
\int_s^{m+s} f(u) e^{iu\lambda} \, du &= \int_s^{m+s} f_1(u) e^{iu\lambda} \, du + \int_s^{m+s} f_3(u) e^{iu\lambda} \, du + g(\lambda) \int_0^4 B_{-2}(u) e^{iu\lambda} \, du \\
&= \int_s^{m+s} f_1(u) e^{iu\lambda} \, du + \int_s^{m+s} f_3(u) e^{iu\lambda} \, du + g(\lambda) \left[\frac{2\sin(\lambda/2)}{\lambda} \right]^4.
\end{aligned} \tag{2.37}$$

Note that

$$\left|\int_s^{m+s} f_1(u)e^{iu\lambda}\,du\right| \le \int_3^7 \sum_{j=0}^{3} |b_j| B_j(u)\,du \le 2, \tag{2.38}$$

and a similar inequality holds for the Fourier transform of $f_3(u)$.

We will consider two cases.

Case 1: $\max_{\lambda\in[0,2/m]} |g(\lambda)| \le m\,|f_1(s)|/3$. on the interval $\lambda \in [\pi/(6m), 11\pi/(6m)]$, by law of cosines, it holds that

$$\left|-f_1(s)e^{is\lambda} + f_3(m+s)e^{i(m+s)\lambda}\right| \ge |f_1(s)\sin(m\lambda)| \ge |f_1(s)|/2,$$

and with (2.37) and (2.38), we have

$$\begin{aligned}\left|-f_1(s)e^{is\lambda} + f_3(m+s)e^{i(m+s)\lambda} - (i\lambda)\int_s^{m+s} f(u)e^{iu\lambda}\,du\right| &\ge \frac{|f_1(s)|}{2} - \frac{22\pi}{3m} - \frac{|f_1(s)|}{3} \\ &= \frac{|f_1(s)|}{6} - \frac{22\pi}{3m}.\end{aligned}$$

It follows that

$$\begin{aligned}\mathbf{b}'\mathbf{P}_s\mathbf{b} &\ge \frac{1}{\pi}\int_{\pi/(6m)}^{11\pi/(6m)} \left(\frac{|f_1(s)|}{6} - \frac{22\pi}{3m}\right)^2 \frac{m/\rho}{1+(m\lambda/\rho)^2}\,d\lambda \\ &= \frac{1}{\pi}\int_{\pi/6}^{11\pi/6} \left(\frac{|f_1(s)|}{6} - \frac{22\pi}{3m}\right)^2 \frac{\rho}{\rho^2+\lambda^2}\,d\lambda \ge \frac{60\rho}{36\rho^2+121\pi^2}\left(\frac{|f_1(s)|}{6} - \frac{22\pi}{3m}\right)^2.\end{aligned}$$

Recall that $\sum_{j=0}^{3} b_j^2 \ge \frac{1}{3m}$, so

$$|f_1(s)| \ge c_4\left(\sum_{j=0}^{3} b_j^2\right)^{1/2} \ge c_6/\sqrt{3m}.$$

Therefore, for Case 1, there exists a constant $c_7 > 0$, such that when $m \ge c_7$, it holds that

$$\mathbf{b}'\mathbf{P}_s\mathbf{b} \ge \frac{5\rho}{109\rho^2+364\pi^2}|f_1(s)|^2 \ge \frac{5\rho c_6^2}{109\rho^2+364\pi^2}\sum_{j=0}^{3} b_j^2.$$

Case 2: $\max_{\lambda\in[0,2/m]}|g(\lambda)| > m\,|f_1(s)|/3$. This implies that there exists a $\lambda_0 \in [0, 2/m]$ such that $|g(\lambda_0)| \geq m\,|f_1(s)|/3 \geq c_6\sqrt{m}/6$. Since $\mathbf{b}$ is a unit vector, $|g(\lambda)| \leq \sqrt{m}$, and it follows that $f_1(s) \leq 3/\sqrt{m}$. By Zygmund (2002), the derivative of $g(\lambda)$ satisfies $|g'(\lambda)| \leq m^{3/2}$. Therefore, on an interval $I_1 \subset [0, 3/m]$ of length $c_6/(12m)$, $|g(\lambda)| \geq c_6\sqrt{m}/12$. On this interval, it holds that

$$\left|\int_s^{m+s} f(u)e^{iu\lambda}\,du\right| \geq \frac{c_6\sqrt{m}}{12}\left[\frac{2\sin(\lambda/2)}{\lambda}\right]^4 - 4.$$

There exists a $c_8 > 0$ such that when $m \geq c_8$,

$$\left|\int_s^{m+s} f(u)e^{iu\lambda}\,du\right| \geq \frac{c_6\sqrt{m}}{13}.$$

It follows that

$$\begin{aligned}\mathbf{b}'\mathbf{V}_s\mathbf{b} &\geq \frac{\rho^2}{\pi m^2}\int_{I_1}\left|\int_s^{m+s} f(u)e^{iu\lambda}\,du\right|^2 \frac{m/\rho}{1+(m\lambda/\rho)^2}\,d\lambda \geq \frac{\rho^2}{\pi m^2}\int_{mI_1}\frac{c_4^2 m}{169}\frac{\rho}{\rho^2+\lambda^2}\,d\lambda \\ &\geq \frac{c_6^3\rho^3}{2028\pi(\rho^2+9)}\cdot\frac{1}{m} \geq \frac{c_6^3\rho^3}{18252\pi(\rho^2+9)}\cdot|f_1(s)|^2 \geq \frac{c_6^5\rho^3}{18252\pi(\rho^2+9)}\cdot\sum_{j=0}^{3} b_j^2.\end{aligned}$$

Note that

$$\begin{aligned}\mathbf{b}'\mathbf{P}_s\mathbf{b} &= \sum_{q=0}^{m+3}\sum_{j=0}^{m+3}\int_{s/m}^{1+s/m}\int_{s/m}^{1+s/m} b_i b_j e^{-\rho|u-v|/m}B'_{q+m}(u)B'_{j+m}(v)\,dudv \\ &= \mathrm{E}\left(\sum_{q=0}^{m+3}\int_s^{1+s/m} b_q\varepsilon_t(u)B'_{q+m}(u)\,du\right)^2 \geq 0.\end{aligned}$$

Similarly, we have $\mathbf{b}'\mathbf{V}_s\mathbf{b} \geq 0$, so both $\mathbf{P}_s$ and $\mathbf{V}_s$ are non-negative definite. Hence, the proof is completed by setting $c_1 = \max\{c_7, c_8\}$, and

$$c_2 = \frac{c_6^5\rho\min\{\rho^2, 1\}}{18252\pi(\rho^2+9)}.$$

■

Lemma 2.5. *The minimum eigenvalue of $\mathbf{\Gamma}_k$, defined in Theorem 3.3 is $O(k^{-1})$.*

Proof: $\mathbf{\Gamma}_k$ could be written as sum of the following four matrices. $\mathbf{\Gamma}_k = \sum_{i=1}^4 \mathbf{G}_{k,i}$.

$$(\mathbf{G}_{k,1})_{ij} = \int_0^1 \int_0^1 \int_0^1 \gamma_0(u,v) B'_{k,i}(s-u) B'_{k,j}(s-v)\, dudvds,$$
$$(\mathbf{G}_{k,2})_{ij} = \rho^2 \int_0^1 \int_0^1 \int_0^1 \gamma_0(u,v) B_{k,i}(s-u) B_{k,j}(s-v)\, dudvds,$$

$$(\mathbf{G}_{k,3})_{ij} = \rho \int_0^1 \int_0^1 \gamma_0(u,v) B_{k,i}(-u) B_{k,j}(-v)\, dudv,$$
$$(\mathbf{G}_{k,4})_{ij} = \rho \int_0^1 \int_0^1 \gamma_0(u,v) B_{k,i}(1-u) B_{k,j}(1-v)\, dudv,$$

for $i, j = 1, \ldots, k$. Define $\mathbf{W}_{k,1}$ and $\mathbf{W}_{k,2}$.

$$(\mathbf{W}_{k,1})_{ij} = \frac{\sigma^2}{2\rho} \int_0^1 \int_0^1 \int_0^1 e^{-\rho|u-v|} B'_{k,i}(s-u) B'_{k,j}(s-v)\, dudvds,$$
$$(\mathbf{W}_{k,2})_{ij} = \frac{\rho\sigma^2}{2} \int_0^1 \int_0^1 \int_0^1 e^{-\rho|u-v|} B_{k,i}(s-u) B_{k,j}(s-v)\, dudvds,$$

for $i, j = 1, \ldots, k$. For any $\mathbf{b} = \{b_1, \ldots, b_k\} \in \mathbb{R}^k$,

$$\begin{aligned}
\mathbf{b}'\mathbf{G}_{k,1}\mathbf{b} &= \int_0^1 \mathrm{E}\left(\sum_{i=1}^k \int_0^1 b_i X_t(u) B'_{k,i}(s-u) du \right)^2 ds \\
&= \int_0^1 \mathrm{E}\left[\sum_{i=1}^k \int_0^1 b_i \left(\int_0^1 \phi(u-v) X_{t-1}(v)\, dv + \varepsilon_{t-1}(u) \right) B'_{k,i}(-u)\, du \right]^2 ds \\
&\geq \int_0^1 \mathrm{E}\left[\sum_{i=1}^k \int_0^1 b_i \varepsilon_{t-1}(u) B'_{k,i}(-u)\, du \right]^2 ds \geq \mathbf{b}'\mathbf{W}_{k,1}\mathbf{b} \geq 0.
\end{aligned}$$

So $\mathbf{G}_{k,1} \succeq \mathbf{W}_{k,1} \succeq \mathbf{0}$. In a similar way, we can show that $\mathbf{G}_{k,2} \succeq \mathbf{W}_{k,2} \succeq \mathbf{0}$. Thus $\mathbf{\Gamma}_k \succeq \mathbf{W}_{k,1} + \mathbf{W}_{k,2}$. Let $\mathbf{b} = (b_0, \ldots, b_{2m+2})$ be a unit vector, and

$$\mathcal{D}_1 = \{0 \leq j \leq 2m-1 : \sum_{l=j}^{j+3} b_l^2 \geq 1/(3m)\},$$

and $\mathcal{D}_2 = \{0, 1, \ldots, 2m-1\} \setminus \mathcal{D}_1$. Since $\sum_{j \in \mathcal{D}_2} \sum_{l=j}^{j+3} b_l^2 \leq 2m \cdot 1/(3m) = 2/3$, it holds that $\sum_{j \in \mathcal{D}_1} \sum_{l=j}^{j+3} b_l^2 \geq 1/3$. For each $j \in \mathcal{D}_1$ and $j \leq m-1$, by Lemma 2.4 , There exists an interval $I_j \subset [-1 + j/m, -1 + (j+1)/m]$ of length c_5/m, such that for each

$s \in I_j$

$$\int_0^1 \int_0^1 \sum_{h,l=0}^{2m+2} b_h b_l \left[B'_{k,h}(s-u)B'_{k,l}(s-v) + \rho^2 B_{k,h}(s-u)B_{k,l}(s-v) \right] e^{-\rho|u-v|}\, dudv \geq c_6 \sum_{l=j}^{j+3} b_l^2.$$

It follows that

$$\int_0^1 \int_0^1 \int_0^1 \sum_{h,l=0}^{2m+2} b_h b_l \left[B'_{k,h}(s-u)B'_{k,l}(s-v) + \rho^2 B_{k,h}(s-u)B_{k,l}(s-v) \right] e^{-\rho|u-v|}\, dudvds$$
$$\geq \frac{c_3 c_6}{m} \sum_{j \in \mathcal{D}_1, j \leq m-1} \sum_{l=j}^{j+3} b_l^2.$$

By applying reverting the order of the rows and the columns, we can show that for each $j \in \mathcal{D}_1$ and $j \geq m$, there exists an interval $I_j \subset [-1 + j/m, -1 + (j+1)/m]$ of length c_5/m, such that for each $s \in I_j$

$$\int_0^1 \int_0^1 \sum_{h,l=0}^{2m+2} b_h b_l \left[B'_{k,h}(s-u)B'_{k,l}(s-v) + \rho^2 B_{k,h}(s-u)B_{k,l}(s-v) \right] e^{-\rho|u-v|}\, dudv \geq c_6 \sum_{l=j}^{j+3} b_l^2.$$

So similarly,

$$\int_0^1 \int_0^1 \int_0^1 \sum_{h,l=0}^{2m+2} b_h b_l \left[B'_{k,h}(s-u)B'_{k,l}(s-v) + \rho^2 B_{k,h}(s-u)B_{k,l}(s-v) \right] e^{-\rho|u-v|}\, dudvds$$
$$\geq \frac{c_3 c_6}{m} \sum_{j \in \mathcal{D}_1, j \geq m} \sum_{l=j}^{j+3} b_l^2.$$

Therefore,

$$\mathbf{b}'(\mathbf{W}_{k,1} + \mathbf{W}_{k,2})\mathbf{b} \geq \frac{c_3 c_6}{6m},$$

and the proof is complete. ∎

Lemma 2.6. *Assume* $\Xi : [0,1]^2 \to [0,1]$ *is a function, and satisfies*

$$\begin{aligned}
\|\Xi^{(1)}\|_\infty &= \sup_{0\le u_1,u_2,v\le 1, u_1\neq u_2} \left\{ \left| \frac{\Xi(u_1,v)-\Xi(u_2,v)}{u_1-u_2} \right|, \left| \frac{\Xi(v,u_1)-\Xi(v,u_2)}{u_1-u_2} \right| \right\} < \infty; \\
\|\Xi^{(2)}\|_\infty &= \sup_{0\le u,v\le 1, u\neq v} \left\{ \left| \frac{\partial^2 \Xi(u,v)}{\partial u^2} \right|, \left| \frac{\partial^2 \Xi(u,v)}{\partial u \partial v} \right|, \left| \frac{\partial^2 \Xi(u,v)}{\partial v^2} \right| \right\} < \infty.
\end{aligned} \tag{2.39}$$

Define the $k \times k$ *matrix* $\mathbf{\Lambda}$ *with the* (j,l)*-th entry*

$$\int_0^1 \int_0^1 \int_0^1 B'_{k,j}(s-u) B'_{k,l}(s-v) \Xi(u,v)\, dudvds. \tag{2.40}$$

Then $\|\mathbf{\Lambda}\| = O(k^{-1})$.

Proof: Let $\mathcal{S}_j$ be the support of $B_{k,j}$, and let $\mathcal{S}_{jl}$ be the set of s which makes the integral

$$\int_0^1 \int_0^1 B'_{k,j}(s-u) B'_{k,l}(s-v) \Xi(u,v)\, dudv \tag{2.41}$$

nonzero. Define two subsets of $\mathcal{S}_{jl}$,

$$\mathcal{S}_{jl1} := \{ s : \ s \in \mathcal{S}_{jl}, (\mathcal{S}_j \cup \mathcal{S}_l) \not\subset [s-1, s] \},$$

and $\mathcal{S}_{jl2} := \mathcal{S}_{jl} \setminus \mathcal{S}_{jl1}$. Use λ to denote the Lebesgue measure. Let $k = 2m+3$, and note that $1/m$ is the mesh size. For each $1 \le j \le k$, we consider the entries in the j-th row of the matrix $\mathbf{\Lambda}$. There are three cases.

Case 1. If $||l-j|-m| \le 3$, then $\lambda(\mathcal{S}_{jl}) \le 7/m$, and it follows that $|(\mathbf{\Lambda})_{jl}| \le C_1/k$, where C_1 is a constant depending only on Ξ. Observe that for each j, the inequality $||l-j|-m| \le 3$ only holds for at most 14 different values of l.

Case 2. If $|j-l| \le 3$, then $\lambda(\mathcal{S}_{jl1}) \le 7/m$, and integrating (2.41) over $s \in \mathcal{S}_{jl1}$ leads to a value bounded by $C_1/(2k)$. For $s \in \mathcal{S}_{jl2}$, the integral (2.41) can be written as

$$\int_0^1 \int_0^1 B'_{k,j}(s-u) B'_{k,l}(s-v) \Xi(u,v)\, dudv = \int_{\mathcal{S}_j} \int_{\mathcal{S}_l} B'_{k,j}(u) B'_{k,l}(v) \Xi(s-u, s-v)\, dudv.$$

Let u_j^* be the left boundary of $\mathcal{S}_j$, then

$$\begin{aligned}
&\left|\int_{\mathcal{S}_j}\int_{\mathcal{S}_l} B'_{k,j}(u)B'_{k,l}(v)\Xi(s-u,s-v)\,dudv\right| \\
&\qquad = \left|\int_{\mathcal{S}_j}\int_{\mathcal{S}_l} B'_{k,j}(u)B'_{k,l}(v)[\Xi(s-u,s-v)-\Xi(s-u_j^*,s-v)\,dudv]\right| \\
&\qquad \le \int_{\mathcal{S}_j}\int_{\mathcal{S}_l} |B'_{k,j}(u)B'_{k,l}(v)|\cdot\|\Xi^{(1)}\|_\infty(u-u_j^*)\,dudv \le C_1/(2k).
\end{aligned} \tag{2.42}$$

Since $\lambda(\mathcal{S}_{jl2})\le 1$, we have $|(\mathbf{\Lambda})_{jl}|\le C_1/k$.

Case 3. If $3<|j-l|<m-3$, then for each $s\in\mathcal{S}_{jl1}$, it must holds that either $\mathcal{S}_j\subset[s-1,s]$ or $\mathcal{S}_l\subset[s-1,s]$. Similar as (2.42), it holds that for each $s\in\mathcal{S}_{jl1}$, the integral (2.41) has an absolute value less than C_1/k. For $s\in\mathcal{S}_{jl2}$, since $\mathcal{S}_j$ and $\mathcal{S}_l$ has no intersection, we have

$$\begin{aligned}
\Xi(s-u,s-v) =& \Xi(s-u_j^*,s-v) - \frac{\partial\Xi(s-u_j^*,s-v_l^*)}{\partial u}(u-u_j^*) \\
&- \frac{\partial\Xi(s-u_j^*,s-v_l^*)}{\partial v}(v-v_l^*) + R(u,v)
\end{aligned}$$

where

$$|R(u,v)|\le\|\Xi^{(2)}\|_\infty(u-u_j^*+v-v_l^*)^2,\quad u\in\mathcal{S}_j,\ v\in\mathcal{S}_l.$$

Therefore

$$\begin{aligned}
&\left|\int_{\mathcal{S}_j}\int_{\mathcal{S}_l} B'_{k,j}(u)B'_{k,l}(v)\Xi(s-u,s-v)\,dudv\right| = \left|\int_{\mathcal{S}_j}\int_{\mathcal{S}_l} B'_{k,j}(u)B'_{k,l}(v)R(u,v)\,dudv\right| \\
&\qquad \le \int_{\mathcal{S}_j}\int_{\mathcal{S}_l} |B'_{k,j}(u)B'_{k,l}(v)|\cdot\|\Xi^{(2)}\|_\infty(u-u_j^*+v-v_l^*)^2\,dudv \le C_1/k^2.
\end{aligned} \tag{2.43}$$

Since $\lambda(\mathcal{S}_{jl1})\le 8/m$ and $\lambda(S_{jl2})\le 1$, it holds that

$$|(\mathbf{\Lambda})_{jl}|\le\frac{C_1}{k}\cdot\frac{8}{m}+\frac{C_1}{k^2}\le\frac{C_2}{k^2},$$

where C_2 is a constant which only depends on Ξ.

Combining these three cases, we see that there exists a constant C_3 which only depends on Ξ, such that $\|\mathbf{\Lambda}\|_\infty \le C_3/k$, and $\|\mathbf{\Lambda}\|_1 \le C_3/k$. Therefore,

$$\|\mathbf{\Lambda}\| \le \|\mathbf{\Lambda}\|_\infty^{1/2}\|\mathbf{\Lambda}\|_1^{1/2} \le C_3/k.$$

■

Lemma 2.7. *The CFAR(p) process $X = (X_t, t \in \mathbb{Z})$ satisfies (2.4). Let ε_0^* be an i.i.d. copy of ε_0, and X_t^* be obtained by replacing ε_0 with ε_0^* in the definition of X_t. If $f : [0,1]^2 \to \mathbb{R}$ be a continuous function, and set $\|f\|_\infty := \max_{u,v} |f(u,v)|$, then for $-p \le h \le p$,*

$$Y_{t,h} = \int_0^1 \int_0^1 f(u,v) X_t(u) X_{t-h}(v)\, dudv, \quad \text{and} \quad Y_{t,h}^* = \int_0^1 \int_0^1 f(u,v) X_t^*(u) X_{t-h}^*(v)\, dudv.$$

Then, for each $q \in \mathbb{N}$, there exists a constant ι which depends on $\{\phi_i, i = 1, \ldots, p\}$, ρ, σ^2, p and $f(u,v)$, such that

$$\|Y_{t,h} - Y_{t,h}^*\|_q \le \iota \lambda^t.$$

Proof: By Cauchy-Schwartz inequality and Lemma 2.7, for $t \ge t_0 + p$,

$$\begin{aligned}
\|Y_{t,h} - Y_{t,h}^*\|_q &\le \Big\| \int_0^1 \int_0^1 f(u,v)[X_t(u) - X_t^*(u)] X_{t-h}^*(v)\, dudv \Big\|_q \\
&\quad + \Big\| \int_0^1 \int_0^1 f(u,v)[X_{t-h}(v) - X_{t-h}^*(v)] X_t(u)\, dudv \Big\|_q \\
&\le \|f\|_\infty \int_0^1 \int_0^1 \big\| \mathbf{\Delta}^t \boldsymbol{\varepsilon}_t(u) - \mathbf{\Delta}^t \boldsymbol{\varepsilon}_t^*(u) \big\|_{2q} \cdot \|X_{t-h}^*(v)\|_{2q}\, dudv \\
&\quad + \|f\|_\infty \int_0^1 \int_0^1 \big\| \mathbf{\Delta}^{t-h} \boldsymbol{\varepsilon}_{t-h}(u) - \mathbf{\Delta}^{t-h} \boldsymbol{\varepsilon}_{t-h}^*(u) \big\|_{2q} \cdot \|X_t(v)\|_{2q}\, dudv \\
&\le 2\|f\|_\infty \cdot \lambda^{t-p} \left(\left(\frac{\sigma^2}{\rho} \right)^q [(2q-1)!!] \right)^{1/(2q)} (\gamma_{\max}^q [(2q-1)!!])^{1/(2q)} \\
&\le \frac{2\sigma \gamma_{\max}^{1/2}}{\rho^{1/2}} [(2q-1)!!]^{1/q} \|f\|_\infty \cdot \lambda^{t-p}.
\end{aligned}$$

where $\gamma_{\max} = \max_{s \in [0,1]} \mathrm{Var}(X_t(s))$. ■

Proof of Theorem 3.5: The proof is similar to that of Theorem 3.3. X_t can be

decomposed as

$$X_t(s) = \sum_{i=1}^{p}\sum_{j=1}^{k}\widetilde{\beta}_{k,i,j}(B_{k,j} * \widetilde{X}_{t-i})(s) + \sum_{i=1}^{p}(r_{k,i} * \widetilde{X}_{t-i})(s) + \varepsilon_t(s) + \sum_{i=1}^{p}(\phi_i * \delta_t)(s).$$

Let $\widetilde{\mathbf{v}}_t$ and $\widetilde{\mathbf{w}}_t$ be two $(N+1)$-dimensional vectors whose j-th entries are $\sum_{i=1}^{p}(r_{k,i} * \widetilde{X}_{t-i})((j-1)/N)$ and $\sum_{i=1}^{p}(\phi_i * \delta_t)((j-1)/N)$ respectively. Let $\boldsymbol{\varepsilon}_t$ be the $(N+1)$-dimensional vector whose j-th entry is $\varepsilon_t((j-1)/N)$. The estimate $\widehat{\boldsymbol{\beta}}_k$ can be decomposed as

$$\widehat{\boldsymbol{\beta}}_k = \widetilde{\boldsymbol{\beta}}_k + \left(\sum_{t=p+1}^{T}\mathbf{M}_t'\boldsymbol{\Sigma}^{-1}\mathbf{M}_t\right)^{-1}\left(\sum_{t=p+1}^{T}\mathbf{M}_t'\boldsymbol{\Sigma}^{-1}(\widetilde{\mathbf{v}}_t + \boldsymbol{\varepsilon}_t + \widetilde{\mathbf{w}}_t)\right).$$

We claim that under the condition $T = o(N^2)$

$$\frac{1}{\sqrt{T}}\sum_{t=p+1}^{T}\mathbf{M}_t'\boldsymbol{\Sigma}^{-1}\widetilde{\mathbf{w}}_t = o_p(1). \tag{2.44}$$

Let $\boldsymbol{\mu}_{k,N} = E(\mathbf{M}_t'\boldsymbol{\Sigma}^{-1}\widetilde{\mathbf{v}}_t)$, $\boldsymbol{\Gamma}_{k,N} = E(\mathbf{M}_t'\boldsymbol{\Sigma}^{-1}\mathbf{M}_t)$, and $\mathbf{b}_{k,N} = \boldsymbol{\Gamma}_{k,N}^{-1}\boldsymbol{\mu}_{k,N}$, then

$$\begin{aligned}\widehat{\boldsymbol{\beta}}_k =& \widetilde{\boldsymbol{\beta}}_k + \mathbf{b}_{k,N} + \boldsymbol{\Gamma}_{k,N}^{-1}\frac{1}{T}\sum_{t=p+1}^{T}(\mathbf{M}_t'\boldsymbol{\Sigma}^{-1}\widetilde{\mathbf{v}}_t - \boldsymbol{\mu}_{k,N}) + \boldsymbol{\Gamma}_{k,N}^{-1}\frac{1}{T}\sum_{t=p+1}^{T}\mathbf{M}_t'\boldsymbol{\Sigma}^{-1}\boldsymbol{\varepsilon}_t \\ &- \boldsymbol{\Gamma}_{k,N}^{-1}\frac{1}{T}\sum_{t=p+1}^{T}(\mathbf{M}_t'\boldsymbol{\Sigma}^{-1}\mathbf{M}_t - \boldsymbol{\Gamma}_{k,N})\mathbf{b}_{k,N} + o_p(T^{-1/2}).\end{aligned} \tag{2.45}$$

We claim that

$$\begin{aligned}\widehat{\boldsymbol{\beta}}_k =& \widetilde{\boldsymbol{\beta}}_k + \mathbf{b}_k + \boldsymbol{\Gamma}_k^{-1}\frac{1}{T}\sum_{t=p+1}^{T}(\mathbf{u}_t - \boldsymbol{\mu}_k) + \boldsymbol{\Gamma}_k^{-1}\frac{1}{T}\sum_{t=p+1}^{T}\mathbf{z}_t \\ &- \boldsymbol{\Gamma}_k^{-1}\frac{1}{T}\sum_{t=p+1}^{T}(\mathbf{A}_t - \boldsymbol{\Gamma}_k)\mathbf{b}_k + o_p(T^{-1/2}).\end{aligned} \tag{2.46}$$

By Lemma 2.7, we can finish this proof similar to that of Theorem 3.3. ∎

Lemma 2.8. *Let $q \geq 2$ be an integer. Consider two $(qk) \times (qk)$ symmetric positive*

semi-definite matrices:

$$\boldsymbol{\Omega}_1 = \begin{pmatrix} \boldsymbol{\Omega} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}, \qquad \boldsymbol{\Omega}_2 = \begin{pmatrix} \boldsymbol{\Omega}_{11} & \boldsymbol{\Omega}_{12} \\ \boldsymbol{\Omega}_{21} & \boldsymbol{\Omega}_{22} \end{pmatrix},$$

where $\boldsymbol{\Omega}$ *and* $\boldsymbol{\Omega}_{11}$ *are* $(q-1)k$*-dimensional square matrices. Assume there exist positive constants* $c_1 \leq c_2$ *such that*

$$\|\boldsymbol{\Omega}_{11}\| \leq c_2, \quad \boldsymbol{\Omega} \succeq c_1 \mathbf{I}_{(q-1)k}, \quad \boldsymbol{\Omega}_{22} \succeq c_1 \mathbf{I}_k.$$

Then

$$\boldsymbol{\Omega}_1 + \boldsymbol{\Omega}_2 \succeq \frac{c_1^2}{c_1 + 2c_2} \mathbf{I}_{qk}.$$

Proof: Let $0 < c < 1$, and define

$$\boldsymbol{\Omega}_3 = \begin{pmatrix} c^{-1/2}\mathbf{I}_{(q-1)k} & \mathbf{0} \\ \mathbf{0} & c^{1/2}\mathbf{I}_k \end{pmatrix} \begin{pmatrix} \boldsymbol{\Omega}_{11} & \boldsymbol{\Omega}_{12} \\ \boldsymbol{\Omega}_{21} & \boldsymbol{\Omega}_{22} \end{pmatrix} \begin{pmatrix} c^{-1/2}\mathbf{I}_{(q-1)k} & \mathbf{0} \\ \mathbf{0} & c^{1/2}\mathbf{I}_k \end{pmatrix} = \begin{pmatrix} c^{-1}\boldsymbol{\Omega}_{11} & \boldsymbol{\Omega}_{12} \\ \boldsymbol{\Omega}_{21} & c\,\boldsymbol{\Omega}_{22} \end{pmatrix}.$$

We see that $\boldsymbol{\Omega}_3$ is positive semi-definite, and

$$\boldsymbol{\Omega}_1 + \boldsymbol{\Omega}_2 = \boldsymbol{\Omega}_3 + \begin{pmatrix} \boldsymbol{\Omega} - (c^{-1} - 1)\boldsymbol{\Omega}_{11} & \mathbf{0} \\ \mathbf{0} & (1-c)\boldsymbol{\Omega}_{22} \end{pmatrix}$$

Since $\boldsymbol{\Omega} \succeq c_1 \mathbf{I}_{(q-1)k}$ and $\|\boldsymbol{\Omega}_{11}\| \leq c_2$, it holds that

$$\boldsymbol{\Omega} - (c^{-1} - 1)\boldsymbol{\Omega}_{11} \succeq (c_1 - (c^{-1} - 1)c_2)\mathbf{I}_{(q-1)k}.$$

By taking $c = 2c_2/(c_1 + 2c_2)$, we have

$$\boldsymbol{\Omega}_1 + \boldsymbol{\Omega}_2 \succeq \boldsymbol{\Omega}_3 + \frac{c_1^2}{c_1 + 2c_2} \mathbf{I}_{qk},$$

and the proof is complete. ∎

Lemma 2.9. *Assume that* $\phi_1, \ldots, \phi_p \in C^2[-1, 1]$*, The minimum eigenvalue of* $\boldsymbol{\Gamma}_k$*,*

defined in Theorem 3.5 is $O(k^{-1})$.

Proof: For any function Ψ_i, denote by $(\Psi_i\varepsilon_t)(u)$ the random variable $\int_0^1 \Psi_i(u,v)\varepsilon_t(v)\,dv$. Define $\gamma_{11}(u,v) := \sigma^2 e^{-\rho|u-v|}$. For each $2 \le i \le p$, let $\gamma_{1i}(u,v) := \mathrm{Cov}[\varepsilon_t(u), (\Psi_{i-1}\varepsilon_t)(v)]$ and $\gamma_{i1}(u,v) := \gamma_{1i}(v,u)$. For every pair $2 \le i,j \le p$, define

$$\gamma_{ij}(u,v) = \mathrm{Cov}[(\Psi_{i-1}\varepsilon_t)(u), (\Psi_{j-1}\varepsilon_t)(v)].$$

Now for each pair $1 \le i,j \le p$, define a $k \times k$ matrix $\mathbf{\Xi}_{ij}$, with (h,l)-th entry

$$(\mathbf{\Xi}_{ij})_{hl} = \int_0^1 \int_0^1 \int_0^1 B'_{k,h}(s-u)B'_{k,l}(s-v) + \rho^2 B_{k,h}(s-u)B_{k,l}(s-v)\,ds\,\gamma_{ij}(u,v)dudv.$$

For each $1 \le i \le p$, define the $p \times p$ block matrix

$$\mathbf{\Gamma}_{p,i} = \begin{pmatrix} \mathbf{\Xi}_{ii} & \mathbf{\Xi}_{i,i-1} & \dots & \mathbf{\Xi}_{i,1} & \mathbf{0} \\ \mathbf{\Xi}_{i-1,i} & \mathbf{\Xi}_{i-1,i-1} & \dots & \mathbf{\Xi}_{i-1,1} & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \mathbf{\Xi}_{1,i} & \mathbf{\Xi}_{1,i-1} & \dots & \mathbf{\Xi}_{11} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \dots & \mathbf{0} & \mathbf{0} \end{pmatrix}.$$

Using the representation (2.7), we know

$$\mathbf{\Gamma}_k \succeq \sum_{i=1}^{p} \mathbf{\Gamma}_{p,i}.$$

By Lemma 2.5, there exists a constant $c_1 > 0$ such that

$$\mathbf{\Xi}_{11} \succeq c_1/k \cdot \mathbf{I}_k.$$

The condition that $\phi_1, \dots, \phi_p \in C^2[-1,1]$ implies that all functions $\gamma_{ij}(u,v)$ satisfies the assumption (2.39), and then by Lemma 2.6, the operator norms of all matrices Ξ_{ij} have the order $O(k^{-1})$ uniformly. Then we can apply Lemma 2.8 inductively to complete the proof. ∎

Proof of Theorem 2.4: By Corollary 6.26 of Schumaker (1981),

$$\|\phi - \widetilde{\phi}_k\|_\infty \le c_1 k^{-\zeta}, \qquad \|\phi'(\cdot) - \widetilde{\phi}'_k(\cdot)\|_\infty \le c_1 k^{-\zeta+1}, \tag{2.47}$$

where c_1 is an absolute constant. There are four terms in the definition of $\mathcal{A}(B_{k,j}, r_k)$. We first consider

$$\int_0^1 \int_0^1 \int_0^1 B'_{k,j}(s-u) r'_k(s-v)\gamma(u,v)\, dudvds.$$

Let

$$\mathbb{S}_j = \{s:\ \lambda(\mathcal{S}_j \cap [s-1, s]) > 0\},$$

and $\mathbb{S}_{j1} = \{s \in \mathbb{S}_j :\ \mathcal{S}_j \not\subset [s-1, s]\}$, $\mathbb{S}_{j2} = \mathbb{S}_j \setminus \mathbb{S}_{j1}$. When $s \in \mathbb{S}_{j1}$, it holds that

$$\left| \int_0^1 \int_0^1 B'_{k,j}(s-u) r'_k(s-v)\gamma(u,v)\, dudv \right| \le C_\rho k^{-\zeta+1}.$$

where C is constant which only depends on ρ, σ^2 and ϕ. When $s \in \mathbb{S}_{j2}$ and $3 < j < k-2$, using the fact that $\int_{\mathcal{S}_j} B'_{k,j}(u)\, du = 0$ and Lemma 2.2, we have

$$\begin{aligned}
&\left| \int_0^1 \int_0^1 B'_{k,j}(s-u) r'_k(s-v)\gamma(u,v)\, dudv \right| \\
&\quad = \left| \int_0^1 \int_{\mathcal{S}_j} B'_{k,j}(u) r'_k(s-v) \left[\gamma(s-u, v) - \gamma(s-u_j^*, v)\right]\, dudv \right| \\
&\quad \le Ck^{-\zeta}.
\end{aligned}$$

Noticing that $\lambda(\mathbb{S}_{j1}) \le 4/m$, and combining the previous two cases, we see that

$$\left| \int_0^1 \int_0^1 \int_0^1 B'_{k,j}(s-u) r'_k(s-v)\gamma(u,v)\, dudvds \right| \le Ck^{-\zeta}.$$

It can be shown that the other three terms in the definition of $\mathcal{A}(B_{k,j}, r_k)$ have the same order $O(k^{-\zeta})$, and hence

$$|(\boldsymbol{\mu}_k)_j| \le Ck^{-\zeta}.$$

By Lemma 2.5,

$$\|\mathbf{b}_k\| = \|\mathbf{\Gamma}_k^{-1}\boldsymbol{\mu}_k\| = O(k^{-\zeta+3/2}).$$

It follows that

$$\|b_k(\cdot)\|_\infty \leq \|\mathbf{B}_k(\cdot)'\mathbf{b}_k\|_\infty + \|r_k(\cdot)\|_\infty = O(k^{-\zeta+3/2}).$$

For the statement regarding $\|\sigma_k^2(\cdot)\|_\infty$, it suffices to show that

$$\|\mathbf{\Sigma}_k\| = O(k).$$

We proceed by calculating the long-run variances of the three processes $\mathbf{u}_t$, $\mathbf{z}_t$ and $\mathbf{A}_t\mathbf{b}_k$. First we observe that $\mathbf{z}_t$ is a martingale, and

$$\mathrm{Var}(\mathbf{z}_t) = \mathbf{\Gamma}_k.$$

The covariance between u_{tj} and $u_{t+h,l}$ is

$$\begin{aligned}\mathrm{Cov}(u_{tj}, u_{t+h,l}) = \int_{[0,1]^4} &\mathcal{A}(B_{k,j}, r_k)(u_1, v_1)\mathcal{A}(B_{k,l}, r_k)(u_2, v_2) \\ &\times [\gamma_h(u_1, u_2)\gamma_h(v_1, v_2) + \gamma_h(u_1, v_2)\gamma_h(v_1, u_2)]\, du_1 du_2 du_3 du_4.\end{aligned}$$

By (2.47), we have

$$\mathcal{A}(B_{k,j}, r_k)(u, v) \leq C_\rho k^{-\zeta+1},$$

where C_ρ is constant which only depends on ρ. The preceding bound, together with Lemma 2.2 implies that,

$$|\mathrm{Cov}(u_{tj}, u_{t+h,l})| \leq 2C_\rho^2/(1-\kappa_2^2)^2\sigma^4 \cdot k^{-2\zeta+2}\kappa^{2|h|}.$$

Therefore, the operator norm of the long-run variance of $\mathbf{u}_t$ is of the order $O(k^{-2\zeta+3})$.

The (j_1, j_2)-th entry of the matrix $\mathrm{E}(\mathbf{A}_t \mathbf{b}_k \mathbf{b}_k' \mathbf{A}_{t+h}')$ is

$$\sum_{l_1,l_2=1}^{k} \int_{[0,1]^4} \mathcal{A}(B_{k,l_1}, B_{k,j_1})(u_1, v_1)(\mathbf{b}_k)_{l_1}(\mathbf{b}_k)_{l_2} \mathcal{A}(B_{k,l_2}, B_{k,j_2})(u_2, v_2) \times [\gamma_h(u_1, u_2)\gamma_h(v_1, v_2) + \gamma_h(u_1, v_2)\gamma_h(v_1, u_2)]\, du_1 du_2 du_3 du_4. \tag{2.48}$$

By the definition of $\mathcal{A}(\cdot, \cdot)$, the product $\mathcal{A}(B_{k,j_1}, B_{k,h_1})\mathcal{A}(B_{k,j_2}, B_{k,h_2})$ can be expanded to 16 terms. We first consider the term

$$\int_{[0,1]^6} B'_{k,l_1}(s_1 - u_1) B'_{k,j_1}(s_1 - v_1) B'_{k,l_2}(s_2 - u_2) B'_{k,j_2}(s_2 - v_2) \times [\gamma_h(u_1, u_2)\gamma_h(v_1, v_2) + \gamma_h(u_1, v_2)\gamma_h(v_1, u_2)]\, ds_1 ds_2 du_1 dv_1 du_2 dv_2. \tag{2.49}$$

Let $\mathcal{B}_k$ be the set of the pairs (j, l) such that either $|j - l| \leq 3$, or $||j - l| - m| \leq 3$. Lemma 2.2 gives bounds on derivatives of $\gamma_h(u, v)$. Using these bounds, and a similar argument as the proof of Lemma 2.6, we can show that there exists a constant $c_2 > 0$ such that the absolute value of (2.49) can be controlled as

$$\begin{cases} c_2 \kappa^{2h}/k^2 & \text{if } (j_1, l_1) \in \mathcal{B}_k \text{ and } (j_2, l_2) \in \mathcal{B}_k; \\ c_2 \kappa^{2h}/k^3 & \text{if one and only one of } (j_1, l_1) \text{ and } (j_2, l_2) \text{ belongs to } \mathcal{B}_k; \\ c_2 \kappa^{2h}/k^4 & \text{if } (j_1, l_1) \notin \mathcal{B}_k \text{ and } (j_2, l_2) \notin \mathcal{B}_k. \end{cases}$$

It can be shown that the other 15 terms have the same bounds, and we omit the details. Let $\mathbf{C}$ be a $k \times k$ matrix whose (j, l)-th entry is $1/k$ when $(j, l) \in \mathcal{B}_k$, and $1/k^2$ when $(j, l) \notin \mathcal{B}_k$, then

$$\|E(\mathbf{A}_t \mathbf{b}_k \mathbf{b}_k' \mathbf{A}_{t+h}')\| \leq c_3 \kappa_2^{2h} \cdot \|\mathbf{C} \cdot |\mathbf{b}_k| \cdot |\mathbf{b}_k'| \cdot \mathbf{C}\| \leq c_4 \kappa^{2h} k^{-2\zeta+3},$$

where c_3 and c_4 are absolute constants, and $|\mathbf{b}_k|$ consists of entry-wise absolute values of $\mathbf{b}_k$. Therefore, the long-run variance of $\mathbf{A}_t \mathbf{b}_k$ is of the order $O(k^{-2\zeta+3})$, and the proof is complete. ∎

Lemma 2.10. *If the CFAR(p) process $X = (X_t, t \in \mathbb{Z})$ satisfies (2.4) and $\phi_i \in C^2[-1, 1]$ for $i = 1, \ldots, p$, then $\mathbf{X}_t = \sum_{j=0}^{\infty} \mathbf{\Delta}^j \boldsymbol{\varepsilon}_{t-j}$ is a stationary process, where*

$\mathbf{X}_t = (X_t, X_{t-1}, \ldots, X_{t-p+1})'$, $\boldsymbol{\varepsilon}_t = (\varepsilon_t, 0, \ldots, 0)'$, *and* $\boldsymbol{\Delta}$ *is defined in (2.27). There exist three constants* ι, t_0 *and* λ *which only depend on* $\{\phi_i, i = 1, \ldots, p\}$ *and* ρ *such that:*

(i) $\|\boldsymbol{\Delta}^h\| \leq \iota\lambda^h$ *for* $h \geq 2$.

(ii) $\|\gamma_h\|_\infty \leq \sigma^2\iota\lambda^{|h|}$ *for* $h \in \mathbb{Z}$.

(iii) For $h \geq 1$,

$$\max_{u,v}\left\{\left|\frac{\partial\gamma_h(u,v)}{\partial u}\right|, \left|\frac{\partial\gamma_h(u,v)}{\partial v}\right|\right\} \leq \sigma^2\iota\kappa^h.$$

(iv) For $h \geq 1$,

$$\max_{u,v}\left\{\left|\frac{\partial^2\gamma_h(u,v)}{\partial u^2}\right|, \left|\frac{\partial^2\gamma_h(u,v)}{\partial u\partial v}\right|, \left|\frac{\partial^2\gamma_h(u,v)}{\partial v^2}\right|\right\} \leq \sigma^2\iota\kappa^h.$$

Proof: By Lemma 5.1 and Theorem 5.1 in Bosq (2000), it is straightforward to see that $\mathbf{X}_t$ has the following expression: $\mathbf{X}_t = \sum_{j=0}^{\infty}\boldsymbol{\Delta}^j\boldsymbol{\varepsilon}_{t-j}$. Since we show that $\|\boldsymbol{\Delta}^j\|^{1/j} \to \lambda_{\max} < 1$ in the proof of Theorem 2, there exists an integer t_0 such that $\|\boldsymbol{\Delta}^h\| \leq \lambda^h$, for $h \geq t_0$, where $\lambda = (1 + \lambda_{\max})/2$.

For $h \geq t_0$, any $u, v \in [0, 1]$,

$$\begin{aligned}\gamma_h(u,v) &= \frac{1}{p}\mathrm{E}(\mathbf{X}_t(u)'\mathbf{X}_{t+h}(v)) = \frac{1}{p}\sum_{j=0}^{\infty}\mathrm{E}\left[(\boldsymbol{\Delta}^j\boldsymbol{\varepsilon}_{t-j})(u)'(\boldsymbol{\Delta}^{j+h}\boldsymbol{\varepsilon}_{t-j})(v)\right] \\ &\leq \frac{1}{p}\sum_{j=0}^{\infty}\|(\boldsymbol{\Delta}^j\boldsymbol{\varepsilon}_{t-j})(u)\|_2 \cdot \|(\boldsymbol{\Delta}^{j+h}\boldsymbol{\varepsilon}_{t-j})(v)\|_2 \leq \frac{\delta\sigma^2\lambda^h}{2\rho p(1-\lambda)},\end{aligned}$$

where $\delta = \max_{j\geq 0}\|\boldsymbol{\Delta}^j\|$. By the definition of Δ_i, we have

$$\Delta_{q_1}\Delta_{q_2}\ldots\Delta_{q_n}f(s) = \int_{[0,1]^n}\phi_{q_1}(s-u_1)\phi_{q_2}(u_1-u_2)\ldots\phi_{q_n}(u_{q_n-1}-u_{q_n})f(u_n)\,du_1\ldots du_{q_n}.$$

The entries in i-th row of $\boldsymbol{\Delta}^p$ can be regarded a polynomial in $\{\Delta_1, \ldots, \Delta_p\}$ of degree $p - i + 1$, without intercept (the identity operator). Hence, we can define two linear operators $\boldsymbol{\Delta}^{(1)}$ and $\boldsymbol{\Delta}^{(2)}$ in $\mathbf{H}^p[0, 1]$ such that $\boldsymbol{\Delta}^{(1)}f(s) = [\boldsymbol{\Delta}^p f(s)]'$, and $\boldsymbol{\Delta}^{(2)}f(s) =$

$[\mathbf{\Delta}^p f(s)]''$. Specifically, assume that the (i,j)-th entry in $\mathbf{\Delta}$ is $\Delta_{q_1}\Delta_{q_2}\dots\Delta_{q_n}$, and we define (i,j)-th entries in $\mathbf{\Delta}^{(1)}$ and $\mathbf{\Delta}^{(2)}$ as follows

$$(\mathbf{\Delta}^{(1)} f(s))_{ij} = \int_{[0,1]^n} \phi'_{q_1}(s-u_1)\phi_{q_2}(u_1-u_2)\dots\phi_{q_n})(u_{q_n-1}-u_{q_n})f(u_n)\,du_1\dots du_{q_n},$$

$$(\mathbf{\Delta}^{(2)} f(s))_{ij} = \int_{[0,1]^n} \phi''_{q_1}(s-u_1)\phi_{q_2}(u_1-u_2)\dots\phi_{q_n})(u_{q_n-1}-u_{q_n})f(u_n)\,du_1\dots du_{q_n}.$$

Then, the operator norms of $\mathbf{\Delta}^{(1)}$ and $\mathbf{\Delta}^{(2)}$ are bounded by $pd_{p,1}\kappa_{\max}^{p-1}$ and $pd_{p,2}\kappa_{\max}^{p-1}$ respectively, where $\kappa_{\max} = \max\{\kappa_1,\dots,\kappa_p,1\}$, $d_{p,1} = \max_{1\le j\le p,0\le s\le 1}|\phi'_j(s)|$, and $d_{p,2} = \max_{1\le j\le p,0\le s\le 1}|\phi''_j(s)|$.

On the other hand, by the definition of $\mathbf{\Delta}$, it suffices to show that the operator norms of $\partial(\mathbf{\Delta}^j f(s))/\partial s$ and $\partial(\mathbf{\Delta}^j f(s))/\partial s^2$ are bounded by $pd_{p,1}\kappa_{\max}^{p-1}\|f\|_\infty$ and $pd_{p,2}\kappa_{\max}^{p-1}\|f\|_\infty$ respectively, for $1\le j<p$.

It follows that,

$$\frac{\partial\gamma_h(u,v)}{\partial u} \le \frac{\sigma^2\delta_1\lambda^h}{2\rho p(1-\lambda)}, \quad \frac{\partial\gamma_h(u,v)}{\partial v} \le \frac{\sigma^2\delta d_{p,1}\kappa_{\max}^{p-1}\lambda^{h-p}}{2\rho(1-\lambda)},$$

$$\frac{\partial^2\gamma_h(u,v)}{\partial u^2} \le \frac{\sigma^2 d\delta_2\lambda^h}{2\rho p(1-\lambda)}, \quad \frac{\partial^2\gamma_h(u,v)}{\partial v^2} \le \frac{\sigma^2\delta d_{p,2}\kappa_{\max}^{p-1}\lambda^{h-p}}{2\rho(1-\lambda)},$$

$$\frac{\partial^2\gamma_h(u,v)}{\partial u\partial v} \le \frac{\sigma^2\delta_2\delta d_{p,2}\kappa_{\max}^{p-1}\lambda^{h-p}}{2\rho(1-\lambda)},$$

for $h\ge t_0+p$, where $\delta_1 = \max\{\rho, pd_{p,1}\kappa_{\max}^{p-1}\delta\}$, and $\delta_2 = \max\{\rho^2, pd_{p,2}\kappa_{\max}^{p-1}\delta\}$.

The proof is complete. ∎

Proof of Theorem 2.6: By Lemma 2.9 and Lemma 2.10, we can complete the proof of Theorem 2.6, following the proof of Theorem 2.4. ∎

Chapter 3

Regime-Switching Factor Models for High-Dimensional Time Series

In this chapter, we consider a factor model for high-dimensional time series with regime-switching dynamics. The switching is assumed to be driven by an unobserved Markov chain; the mean, factor loading matrix and covariance matrix of the noise process are different among the regimes. The model is an extension of the traditional factor models for time series and provides flexibility in dealing with real applications in which underlying states may be changing over time. We propose an iterative approach to estimate the loading space of each regime and cluster the data points, by combining eigenanalysis and Viterbi algorithm. The theoretical properties of the procedure are investigated. Simulation results and the analysis of a real example are presented.

3.1 Switching Factor Models

We introduce some notation first. For any matrix $\mathbf{H}$, $\|\mathbf{H}\|_F$, and $\|\mathbf{H}\|_2$ denote the Frobenius norm and L-2 norm of $\mathbf{H}$; $\mathrm{tr}(\mathbf{H})$ and $\lambda_{\max}(\mathbf{H})$ are trace, and the largest nonzero eigenvalue values of a square matrix $\mathbf{H}$ respectively, and $\|\mathbf{H}\|_{\min}$ is the square root of minimum nonzero eigenvalue of $\mathbf{H}'\mathbf{H}$. We use $a \asymp b$, if $a = O(b)$ and $b = O(a)$.

Let $\mathbf{y}_t$ be a $p \times 1$ observed time series and z_t be a homogenous and stationary hidden Markov chain taking values in $\{1, 2, \ldots, m\}$ with transition probabilities

$$\pi_{k,j} = P(z_{t+1} = j \mid z_t = k) \qquad k, j = 1, \cdots, m, \tag{3.1}$$

and number of states m is known. We assume that, for $t = 1, \ldots, n$, when $z_t = k$,

$$\mathbf{y}_t = \boldsymbol{\mu}_k + \mathbf{A}_k \mathbf{x}_t + \boldsymbol{\varepsilon}_t^{(k)} \quad \text{and} \quad \boldsymbol{\varepsilon}_t^{(k)} \sim N(\mathbf{0}, \boldsymbol{\Sigma}_k), \tag{3.2}$$

where $\mathbf{x}_t$ is a $d \times 1$ latent factor process with d fixed and (much) smaller than p, independent of $\mathbf{z} = \{z_1, \ldots, z_n\}$, $\mathrm{E}(\mathbf{x}_t) = 0$. $\boldsymbol{\mu}_k$ is the mean of the process, $\mathbf{A}_k$ is the unknown loading matrix, and $\boldsymbol{\Sigma}_k$ is the covariance matrix of the noise process for state k. We also assume that $\{\boldsymbol{\varepsilon}_t^{(1)}\}, \cdots, \{\boldsymbol{\varepsilon}_t^{(m)}\}$ are m uncorrelated white noise processes, and are independent of $\{(\mathbf{x}_t, z_t), t \in \mathbb{Z}\}$. Our model is a generalization of factor models of Lam et al. (2011). The dynamics of $\mathbf{y}_t$ are driven by the factor process $\mathbf{x}_t$ according to m states controlled by the switch variable z_t.

As noted in Lam and Yao (2012), $\mathbf{A}_k$ is not uniquely defined, since $(\mathbf{A}_k, \mathbf{x}_t)$ in (3.2) can be replaced by $(\mathbf{A}_k \mathbf{U}_k, \mathbf{U}_k^{-1} \mathbf{x}_t)$ for any $d \times d$ non-singular matrix $\mathbf{U}_k$. Denote the linear space spanned by the columns of a matrix $\mathbf{A}$ as $\mathcal{M}(\mathbf{A})$. It is easily seen that $\mathcal{M}(\mathbf{A}_k)$, the factor loading space for state k, is uniquely defined by (3.2). Hence, we can find a $p \times d$ matrix $\mathbf{Q}_k$ and a $d \times d$ non-singular matrix $\boldsymbol{\Gamma}_k$ satisfying

$$\mathbf{Q}_k' \mathbf{Q}_k = \mathbf{I}_d, \text{ and } \mathbf{A}_k = \mathbf{Q}_k \boldsymbol{\Gamma}_k, \quad k = 1, \cdots, m. \tag{3.3}$$

It follows that $\mathcal{M}(\mathbf{Q}_k) = \mathcal{M}(\mathbf{A}_k)$. The columns of $\mathbf{Q}_k$ are d orthonormal vectors, and the column space spanned by $\mathbf{Q}_k$ is the same as the column space spanned by $\mathbf{A}_k$. In addition, let $\mathbf{B}_k = (\mathbf{b}_{k,1}, \ldots, \mathbf{b}_{k,p-d})$ be an orthnormal basis such that $\mathcal{M}(\mathbf{B}_k)$ is the orthogonal complement space of $\mathcal{M}(\mathbf{Q}_k)$. Hence $(\mathbf{Q}_k, \mathbf{B}_k)$ forms a $p \times p$ matrix with orthogonal columns, $\mathbf{Q}_k' \mathbf{B}_k = \mathbf{0}$, and $\mathbf{B}_k' \mathbf{B}_k = \mathbf{I}_{p-d}$. In practice,

$$\mathbf{A}_k' \mathbf{B}_k = \mathbf{0}. \tag{3.4}$$

Assume that the loading spaces are different across regimes, our goal is to cluster the data by regimes, and estimate d and $\mathcal{M}(\mathbf{Q}_k)$, for $k = 1, \ldots, m$.

Remark 3.1. As mentioned above, $(\mathbf{A}_k, \mathbf{x}_t)$ in (3.2) can be replaced by $(\mathbf{A}_k \mathbf{U}_k, \mathbf{U}_k^{-1} \mathbf{x}_t)$

for any $d \times d$ non-singular matrix $\mathbf{U}_k$. Hence, the factor process may not be stationary after such nonsingular transformations across regimes, if $\{\mathbf{U}_k, k = 1, \ldots, m\}$ are different. However, it does not directly affect the underlying process or the estimation procedure, since we do not impose the stationarity on the latent process $\mathbf{x}_t$.

For factor models in high-dimensional cases, it is common to assume that the squared L-2 norm of the $p \times d$ loading matrix grows with the dimension p (Bai and Ng, 2002; Doz et al., 2011), and the growth rate is defined as the strength of the factors in Lam et al. (2011). In our multi-regime factor model in (3.2), the strength of the factors may be different across regimes. Assume that

$$\|\mathbf{A}_k\|_2^2 \asymp \|\mathbf{A}_k\|_{\min}^2 \asymp p^{1-\delta_k},$$

where $\|\mathbf{A}_k\|_{\min}^2$ is the minimum nonzero eigenvalue of $\mathbf{A}_k'\mathbf{A}_k$. If $\delta_k = 0$, the factors are 'strong' for state k and we call state k a *strong state* and $\mathbf{A}_k$ is a dense loading matrix. If $\delta_k > 0$, the factors are 'weak' for state k and we call state k a *weak state* and $\mathbf{A}_k$ is a sparse loading matrix. The strength of the state is an indicator of signal-to-noise ratio. It measures the relative growth rate of the amount of information which the observed process $\mathbf{y}_t$ carries about the common factors $\mathbf{x}_t$ as p increases, with respective to the growth rate of the amount of noise process. When the state is weak, the information contained in $\mathbf{y}_t$ about the factors grows slower than the noises introduced as p increases, hence the proportion of information is diluted by the noise. When the state is strong, the signal-to-noise ratio remains constant.

Define

$$\mathbf{R}_t = \sum_{k=1}^{m} \mathbf{\Gamma}_k \mathbf{x}_t I(z_t = k), \tag{3.5}$$

and the switching factor model can be written as

$$\mathbf{y}_t = \sum_{k=1}^{m} I(z_t = k) \left(\boldsymbol{\mu}_k + \mathbf{A}_k \mathbf{x}_t + \boldsymbol{\varepsilon}_t^{(k)}\right) = \sum_{k=1}^{m} I(z_t = k) \left(\boldsymbol{\mu}_k + \mathbf{Q}_k \mathbf{\Gamma}_k \mathbf{x}_t + \boldsymbol{\varepsilon}_t^{(k)}\right) \tag{3.6}$$

$$= \sum_{k=1}^{m} I(z_t = k) \left(\boldsymbol{\mu}_k + \mathbf{Q}_k \mathbf{R}_t + \boldsymbol{\varepsilon}_t^{(k)}\right). \tag{3.7}$$

The above equation reveals different ways to decompose the dynamic part of the process. In (3.6), $\mathbf{Q}_k$ is the standardized loadings, $\mathbf{x}_t$ is the factor latent process, and $\mathbf{\Gamma}_k$ reflects the strength of the state, controlling the amount of information $\mathbf{y}_t$ carries on the latent factors. When the dynamic part is divided into two parts in (3.7), $\mathbf{R}_t$ can be regarded as another latent factor process but with standardized loadings, and the L-2 norm of its variance matrix increases with p at different rates across regimes.

3.2 Estimation Procedure

In Section 3.2.1 we introduce a method taking advantage of the autocovariance matrices to estimate of the loading spaces when the state variable $\mathbf{z}$ is known; in Section 3.2.2 we propose a method using Viterbi algorithm to estimate the hidden state variable when the loading spaces are known. Combining the two methods, we propose an iterative algorithm to estimate all the model parameters in Section 3.2.3.

3.2.1 Estimation of $\mathbf{B}_k$, $\boldsymbol{\mu}_k$, d and the Transition Probabilities Given State Indicator z

If the states $z_1, \ldots, z_n$ are given, the transition probability can be estimated by

$$\widehat{\pi}_{k,j} = \frac{\sum_{t=1}^{n-1} I(z_t = k, z_{t+1} = j)}{\sum_{t=1}^{n-1} I(z_t = k)}, \quad \text{for} \quad k, j = 1, ..., m, \quad \text{and}$$

$$\widehat{\pi}_k = \frac{\sum_{t=1}^{n} I(z_t = k)}{n}, \quad \text{for} \quad k = 1, ..., m.$$

For the estimation of factor loading spaces, we adopt the procedure proposed by Lam et al. (2011), Lam and Yao (2012), Chang et al. (2013). It is based on the observation that, since the idiosyncratic noise $\boldsymbol{\varepsilon}_t^{(k)}$ is white, the dynamic of $\mathbf{y}_t$ (autocovariance) only comes from the dynamics of the factor $\mathbf{x}_t$. Hence we can retrieve the factor loading space through an analysis of the autocovariance structure of $\mathbf{y}_t$. Let

$$\mathbf{\Sigma}_x(l) = \frac{1}{n-l} \sum_{t=1}^{n-l} \mathrm{Cov}(\mathbf{x}_t, \mathbf{x}_{t+l}),$$

$$\mathbf{\Sigma}_{y,k}(l) = \frac{1}{n-l}\sum_{t=1}^{n-l} \mathrm{Cov}(\mathbf{y}_t,\, \mathbf{y}_{t+l} \mid z_t = k).$$

Here $\mathbf{\Sigma}_x(l)$ and $\mathbf{\Sigma}_{y,k}(l)$ are the averages of autocovariance matrices of $\mathbf{x}_t$, and $\mathbf{y}_t$ at lead l from time 1 to $n-l$ respectively, given that the current state is k. It follows that

$$\begin{aligned}\mathbf{\Sigma}_{y,k}(l) &= \frac{1}{n-l}\sum_{t=1}^{n-l}\sum_{j=1}^{m} \pi_{k,j}^{(l)} \mathrm{Cov}(\mathbf{y}_t,\, \mathbf{y}_{t+l} I(z_{t+l}=j) \mid z_t = k) \\ &= \mathbf{A}_k \mathbf{\Sigma}_x(l) \sum_{j=1}^{m} \pi_{k,j}^{(l)} \mathbf{A}_j', \end{aligned} \tag{3.8}$$

where $\pi_{k,j}^{(l)} = P(z_{t+l} = j \mid z_t = k)$, the transition probability from state k to state j in l steps. Note that if $\mathbf{x}_t$ is stationary, then $\mathbf{\Sigma}_x(l)$ is the autocovariance matrix of $\mathbf{x}_t$ at lead l.

For a fixed prescribed integer l_0, define

$$\mathbf{M}_k = \sum_{l=1}^{l_0} \mathbf{M}_{k,l}, \tag{3.9}$$

where $\mathbf{M}_{k,l} = \mathbf{\Sigma}_{y,k}(l)\mathbf{\Sigma}_{y,k}(l)'$ is a quadratic version of the autocovariance matrix $\mathbf{\Sigma}_{y,k}(l)$. Because of (3.4) and (3.8), we have $\mathbf{M}_{k,l}\mathbf{B}_k = \mathbf{0}$ for all k. If $\sum_{j=1}^{m} \pi_{k,j}^{(l)} \mathbf{A}_j'$ is of full rank, then $\mathbf{M}_k$ is a non-negative definite matrix sandwiched by $\mathbf{A}_k$ and $\mathbf{A}_k'$ with rank d. Then the d unit eigenvectors of $\mathbf{M}_k$ corresponding to its d non-zero eigenvalues form the space $\mathcal{M}(\mathbf{A}_k)$, the space spanned by the columns of $\mathbf{A}_k$.

We define the sample version of the above statistics, given $\mathbf{z} = \{z_1, \ldots, z_n\}$, for $k = 1, \ldots, m$,

$$\widehat{\mathbf{\Sigma}}_{y,k}(l) = \frac{\sum_{t=1}^{n-l}\sum_{j=1}^{m}(\mathbf{y}_t - \widehat{\boldsymbol{\mu}}_k)(\mathbf{y}_{t+l} - \widehat{\boldsymbol{\mu}}_j)' I(z_t = k, z_{t+l} = j)}{\sum_{t=1}^{n-l} I(z_t = k)}, \quad \widehat{\boldsymbol{\mu}}_k = \frac{\sum_{t=1}^{n} \mathbf{y}_t I(z_t = k)}{\sum_{t=1}^{n} I(z_t = k)} \tag{3.10}$$

$$\widehat{\mathbf{M}}_{k,l} = \widehat{\mathbf{\Sigma}}_{y,k}(l)\widehat{\mathbf{\Sigma}}_{y,k}(l)', \quad \widehat{\mathbf{M}}_k = \sum_{l=1}^{l_0} \widehat{\mathbf{M}}_{k,l}.$$

Let $\widehat{\lambda}_{k,1} \geq \widehat{\lambda}_{k,2} \geq \ldots \geq \widehat{\lambda}_{k,p}$ be the p eigenvalues of $\widehat{\mathbf{M}}_k$ and $\widehat{\mathbf{q}}_{k,1}, \ldots, \widehat{\mathbf{q}}_{k,p}$ be the set of

corresponding orthonormal eigenvectors. Define $\widehat{\mathbf{Q}}_k$ and $\widehat{\mathbf{B}}_k$ as

$$\widehat{\mathbf{Q}}_k = (\widehat{\mathbf{q}}_{k,1}, \ldots, \widehat{\mathbf{q}}_{k,d}), \qquad \text{and} \quad \widehat{\mathbf{B}}_k = (\widehat{\mathbf{q}}_{k,d+1}, \ldots, \widehat{\mathbf{q}}_{k,p}), \tag{3.11}$$

then $\mathcal{M}(\mathbf{Q}_k)$ and $\mathcal{M}(\mathbf{B}_k)$ can be estimated by $\mathcal{M}(\widehat{\mathbf{Q}}_k)$ and $\mathcal{M}(\widehat{\mathbf{B}}_k)$, respectively. To estimate the number of factors with data in each regime, we use the eigenvalue-ratio method of Lam and Yao (2012). Specifically, let

$$\widehat{d}_k = \arg\min_{1 \le j \le c} \widehat{\lambda}_{k,j+1}/\widehat{\lambda}_{k,j}. \tag{3.12}$$

We set c to $p/2$, since the minimum eigenvalues of $\mathbf{M}_k$ may be practically 0, especially when n is small and p is large; see Lam and Yao (2012).

Corollary 3.1 in Section 3.3 shows that under some mild conditions, $\widehat{d}_1, \ldots, \widehat{d}_m$ are all reasonable estimates of the number of factors d. Since d is common to all regimes, we choose the one from the strongest state, as the theoretical results show that the estimated nonzero eigenvalues from a stronger state have a faster convergence rate. Hence, we use $\widehat{d} = \widehat{d}_{\tilde{k}}$ to estimate d, where $\tilde{k} = \arg\max \|\widehat{\mathbf{M}}_k\|_2$.

Let $\mathbf{f}_t$ be the dynamic part of $\mathbf{y}_t$, i.e. $\mathbf{f}_t = \sum_{k=1}^m \mathbf{A}_k \mathbf{x}_t I(z_t = k)$. Since the column space of $\mathbf{A}_k$ is identifiable only up to a nonsingular transformation across regimes, we cannot recover $\mathbf{x}_t$ directly, but we have natural estimators of $\mathbf{R}_t$ and $\mathbf{f}_t$,

$$\widehat{\mathbf{R}}_t = \sum_{k=1}^m \widehat{\mathbf{Q}}_k'(\mathbf{y}_t - \widehat{\boldsymbol{\mu}}_k) I(z_t = k), \quad \widehat{\mathbf{f}}_t = \sum_{k=1}^m \widehat{\mathbf{Q}}_k \widehat{\mathbf{Q}}_k'(\mathbf{y}_t - \widehat{\boldsymbol{\mu}}_k) I(z_t = k), \tag{3.13}$$

and the residuals are

$$\widehat{\boldsymbol{\varepsilon}}_t = \sum_{k=1}^m (\mathbf{I}_p - \widehat{\mathbf{Q}}_k \widehat{\mathbf{Q}}_k')(\mathbf{y}_t - \widehat{\boldsymbol{\mu}}_k) I(z_t = k). \tag{3.14}$$

Remark 3.2. Our method works under weaker assumptions in which the dependence between $\mathbf{x}_t$ and $\boldsymbol{\varepsilon}_s$ when $t > s$ is allowed. If $\mathrm{E}(\boldsymbol{\varepsilon}_s^{(k)} \mathbf{x}_t') = 0$ only for $t \le s$, we can still follow the same procedure to estimate $\mathcal{M}(\mathbf{Q}_k)$, but construct $\mathbf{M}_k$ slightly differently.

Specifically, we can use

$$\mathbf{\Sigma}_{y,k}(l) = \frac{1}{n-l}\sum_{t=1}^{n-l} \mathrm{Cov}(\mathbf{y}_{t+l},\, \mathbf{y}_t \mid z_{t+l} = k),$$

and $\mathbf{M}_k = \sum_{l=1}^{l_0} \mathbf{\Sigma}_{y,k}(l)\mathbf{\Sigma}_{y,k}(l)'$. In this dissertation we assume that $\{\boldsymbol{\varepsilon}_t^{(k)}, t \in \mathbb{Z}\}$ and $\{\mathbf{x}_t, t \in \mathbb{Z}\}$ for all $k = 1, \ldots, m$, are independent for simplicity.

Remark 3.3. This estimation procedure has been used for one-regime factor model with stationary processes in Tao et al. (2011) and Lam et al. (2011), and with non-stationary processes in Chang et al. (2013). Many numerical results show that the estimation of the loading space is not sensitive to the choice of l_0; see Neil et al. (2010), Lam et al. (2011), Lam and Yao (2012) and Chang et al. (2013). Although the estimator works with any $l_0 \geq 1$ both theoretically and numerically, the extra terms in $\mathbf{M}_k$ of (3.9) are very useful when the sample size is small and the variability in the estimation of the autocovariance matrices is large. Nevertheless, as the autocorrelation is often at its strongest at small time lags, a relatively small l_0 is usually adopted.

3.2.2 Estimation of the Hidden State z Given Loading Spaces and Other Model Parameters

Although $\{\mathbf{B}_k, k = 1, \ldots, m\}$ are only uniquely identifiable up to orthogonal transformations, the density function of $\mathbf{B}'_{z_t}\mathbf{y}_t$ is invariant to such transformations. Based on this observation, given $\pi_{k,j}$, π_k, $\boldsymbol{\mu}_k$, and $\mathbf{B}_k$, $k, j = 1, \ldots, m$, the state variables $z_1, \ldots, z_n$ can be estimated by maximizing $G_n(\mathbf{z})$, the logarithm of the probability density function of $\{\mathbf{B}'_{z_t}\mathbf{y}_t, t = 1, \ldots, n\}$. Specifically, under the assumption that the noise process is normally distributed,

$$G_n(\mathbf{z}) = \log\left(\pi_{z_1} f(\mathbf{B}'_{z_1}\mathbf{y}_1)\right) + \sum_{t=2}^{n} \log\left(\pi_{z_{t-1},z_t} f(\mathbf{B}'_{z_t}\mathbf{y}_t)\right), \tag{3.15}$$

where

$$f(\mathbf{B}'_{z_t}\mathbf{y}_t) = -\frac{1}{\sqrt{(2\pi)^{p-d}|\mathbf{\Sigma}_{B,z_t}|}} - \exp\left[-\frac{(\mathbf{B}'_{z_t}(\mathbf{y}_t - \boldsymbol{\mu}_{z_t}))'\mathbf{\Sigma}_{B,z_t}^{-1}\mathbf{B}'_{z_t}(\mathbf{y}_t - \boldsymbol{\mu}_{z_t})}{2}\right]. \tag{3.16}$$

Note that $G_n(\mathbf{z})$ is a sum of n functions in the form of $G_n(\mathbf{z}) = g_1(z_1)+\sum_{t=2}^{n} g_t(z_{t-1}, z_t)$, due to the Markovian structure of $\mathbf{z}$, where

$$g_1(z_1) = \log\left(\pi_{z_1} f(\mathbf{B}'_{z_1}\mathbf{y}_1)\right), \text{ and } g_t(z_{t-1}, z_t) = \log\left(\pi_{z_{t-1},z_t} f(\mathbf{B}'_{z_t}\mathbf{y}_t)\right), \quad t = 2, \ldots, n.$$

Hence $G_n(\mathbf{z})$ can be maximized by the Viterbi algorithm (Viterbi, 1967 and Forney, 1973). The maximizer of the state sequence $z_1, \ldots, z_n$ is given by the following recurrence relations:

$$\begin{aligned} S_{1,k} &= k, \\ x_{t,k} &= \arg\max_{1\le j\le m} \left[g_t(z_{t-1} = j, z_t = k) + G_{t-1}(S_{t-1,j})\right], \\ \mathbf{S}_{t,k} &= (\mathbf{S}_{t-1,x_{t,k}},\, k), \end{aligned}$$

where $\mathbf{S}_{t,k}$ is a $t \times 1$ vector and the maximizer of $G_t(z_1, \ldots, z_{t-1}, z_t = k)$ for the first t observations that has k as its final state. In each iteration there are m evaluations $g_t(\cdot, k)$ to update $G_t(\mathbf{S}_{t,k})$ for $k = 1, \ldots, m$, and m possible paths $\{\mathbf{S}_{t,1}, \ldots, \mathbf{S}_{t,m}\}$ to be compared. So the complexity of Viterbi algorithm is $O(m^2 n)$. The state variable can be estimated by

$$\widehat{\mathbf{z}} = \arg\max_{1\le k\le m} G_n(\mathbf{S}_{n,k}).$$

The covariance matrix of $\mathbf{B}'_k\mathbf{y}_t$ given $z_t = k$, $\mathbf{\Sigma}_{B,k} = \mathbf{B}'_k\mathbf{\Sigma}_k\mathbf{B}_k$ can be estimated by

$$\widehat{\mathbf{\Sigma}}_{B,k} = \sum_{t=1}^{n} \mathbf{B}'_k(\mathbf{y}_t - \boldsymbol{\mu}_k)(\mathbf{y}_t - \boldsymbol{\mu}_k)'\mathbf{B}_k I(\widehat{z}_t = k)/[\sum_{t=1}^{n} I(\widehat{z}_t = k) - 1]. \tag{3.17}$$

Remark 3.4. One would prefer to construct the density of $\mathbf{y}_t$ given z_t with (3.14). However, since $\mathbf{I}_p - \mathbf{Q}_k\mathbf{Q}'_k = \mathbf{B}_k\mathbf{B}'_k$, it follows that

$$(\mathbf{I}_p - \mathbf{Q}_k\mathbf{Q}'_k)(\mathbf{y}_t - \boldsymbol{\mu}_k) \sim N(\mathbf{0}, \mathbf{B}_k\mathbf{B}'_k\mathbf{\Sigma}_k\mathbf{B}_k\mathbf{B}'_k).$$

This is a p-variate normal distribution degenerated into a $(p-d)$-dimensional space,

while

$$\mathbf{B}_k'(\mathbf{y}_t - \boldsymbol{\mu}_k) \sim N(\mathbf{0}, \boldsymbol{\Sigma}_{B,k}), \text{ where } \boldsymbol{\Sigma}_{B,k} = \mathbf{B}_k'\boldsymbol{\Sigma}_k\mathbf{B}_k,$$

is a non-degenerated representation of the above distribution restricted on the $(p-d)$-dimensional space.

Remark 3.5. There are several advantages to use the density function of $\{\mathbf{B}_{z_t}'\mathbf{y}_t, t = 1, \ldots, n\}$, instead of $\{\mathbf{y}_t, t = 1, \ldots, n\}$. First, we do not need to estimate $\mathbf{A}_k$, since we can only estimate the space it spans. Second, in order to compute the density of $\mathbf{y}_t$, we would need to to assume a specific model for the latent process $\mathbf{x}_t$. Although there are a vast literature in dynamic factor models (Forni et al. 2000; Bai and Ng, 2002; Hallin and Liška, 2007), here we choose to avoid the difficulty.

3.2.3 An Iterative Algorithm

We adopt the distance measure used in Chang et al. (2013). For any $p \times d_1$ orthonromal matrix $\mathbf{H}_1$ and $p \times d_2$ orthonormal $\mathbf{H}_2$,

$$\mathcal{D}(\mathbf{H}_1, \mathbf{H}_2) = \{1 - \frac{1}{\max\{d_1, d_2\}}\text{tr}(\mathbf{H}_1\mathbf{H}_1'\mathbf{H}_2\mathbf{H}_2')\}^{1/2}. \tag{3.18}$$

Note that $\mathcal{D}(\mathbf{H}_1, \mathbf{H}_2) \in [0, 1]$, $\mathcal{D}(\mathbf{H}_1, \mathbf{H}_2) = 0$ if and only if $d_1 = d_2$, $\mathcal{M}(\mathbf{H}_1) = \mathcal{M}(\mathbf{H}_2)$, and $\mathcal{D}(\mathbf{H}_1, \mathbf{H}_2) = 1$ if and only if $\mathcal{M}(\mathbf{H}_1) \perp \mathcal{M}(\mathbf{H}_2)$. Now we are ready to state the algorithm.

Step 1. Begin with some initial values $\widehat{\mathbf{z}}$.

Step 2. Given $\widehat{\mathbf{z}}$, obtain $\widehat{d}$, $\widehat{\pi}_k$, $\widehat{\pi}_{j,k}$, $\widehat{\boldsymbol{\mu}}_k$, $\widehat{\mathbf{Q}}_k$ and $\widehat{\mathbf{B}}_k$ based on the methods in Section 3.1, for $j, k = 1, \ldots, m$.

Step 3. Given the estimates obtained in Step 2, estimate $\mathbf{z}$ by maximizing $G_n(\mathbf{z})$, using Viterbi algorithm in Section 3.2.

Step 4. Repeat Step 2 and Step 3 until either a maximum number of iterations is reached, or both of the following two conditions are satisfied.

$$\frac{1}{m}\sum_{k=1}^{m}\mathcal{D}(\widehat{\mathbf{Q}}_k^{(1)}, \widehat{\mathbf{Q}}_k^{(2)}) < c_1, \quad \text{and} \quad \Big|\frac{G_n(\widehat{\mathbf{z}}^{(1)}) - G_n(\widehat{\mathbf{z}}^{(2)})}{G_n(\widehat{\mathbf{z}}^{(1)})}\Big| < c_2,$$

where $c_1, c_2 \in (0,1)$ are prescribed small constants, and $\widehat{\mathbf{Q}}_k^{(1)}$, $\widehat{\mathbf{Q}}_k^{(2)}$, $\widehat{\mathbf{z}}^{(1)}$, and $\widehat{\mathbf{z}}^{(2)}$ are successive estimates for $\mathbf{Q}_k$ and $\mathbf{z}$, respectively.

Remark 3.6. Since the two iterative steps do not minimize the same objective function, the algorithm does not guarantee to reach a fixed point solution given a finite sample. However, we note that the objectives of the two steps are consistent. In step 2, we try to extract common factors when estimating the loading spaces, hence try to reduce the remainder error terms in the factor models. In step 3, maximizing of the density function is equivalent to minimizing the errors in the factor models for normally distributed errors. We also note that such issues are common in the estimation procedures of dynamic factor models, and an iterative algorithm is widely used (Watson and Engle, 1983; Stock and Watson, 2005; Doz et al., 2011).

Remark 3.7. We estimate the state variables instead of the transition probabilities in Step 3, because finding the maximizer of $\{\pi_{k,j}\}$ of density function of $\{\mathbf{B}'_{z_t}\mathbf{y}_t\}$ is much more computationally expensive than the estimation of $\mathbf{z}$ by Viterbi algorithm, due to the dependence of the state variables. Although the misclassification occurs because of the nature of hard clustering (Kearns et al. 1998; Hastie et al. 2009), it does not have too much bad influence on the estimation of the loading spaces, since it often occurs when the data points lie near the intersection of the loading spaces. The numerical results show that our algorithm is able to cluster the data by regimes efficiently and estimate the loading spaces effectively. To obtain consistent estimators of transition probabilities and avoid misclassification, discarding algorithm can be applied (Chen 1995), in which only the data points that can be 'clearly separated' by the objective function, e.g. $f(\mathbf{B}'_k\mathbf{y})/f(\mathbf{B}'_j\mathbf{y}) > c_0$, are used for estimation. Since a small number of observations are used, the consistency of the estimators can be achieved with lower efficiency, as c_0 goes to infinity together with the sample size.

3.2.4 Initial Values of z

Our experience shows that the initial values of the state variable $\mathbf{z}$ are crucial for the estimation procedure. Here we provide a method for finding reasonable initial values

of $\mathbf{z}$. Define

$$\mathbf{\Sigma}_y(l) = \frac{1}{n-l}\sum_{t=1}^{n-l}\mathrm{Cov}(\mathbf{y}_t, \mathbf{y}_{t+l}) = \frac{1}{n-l}\sum_{t=1}^{n-l}\sum_{k=1}^{m}\pi_k \mathrm{Cov}(\mathbf{y}_t, \mathbf{y}_{t+l} \mid z_t = k) = \sum_{k=1}^{m}\pi_k \mathbf{\Sigma}_{y,k}(l).$$

As a sum of k rank-d matrices, the matrix $\mathbf{\Sigma}_y(l)$ should have a rank between 1 and dm. If the transition probabilities between any two states are all equal, $\pi_{j,k} = 1/m$ for $j, k = 1, \ldots, m$, we have

$$\mathbf{\Sigma}_y(l) = \frac{1}{m^2}\sum_{k=1}^{m}\mathbf{A}_k \mathbf{\Sigma}_x(l) \sum_{j=1}^{m}\mathbf{A}_j'.$$

Hence $\mathbf{M} = \sum_{l=1}^{l_0}\mathbf{\Sigma}_y(l)\mathbf{\Sigma}_y(l)'$ is a matrix sandwiched by $\sum_{k=1}^{m}\mathbf{A}_k$ and $\sum_{k=1}^{m}\mathbf{A}_k'$ with rank smaller or equal to d. We find the eigenvalues of $\mathbf{M}$ and use the ratio estimator in (3.12) for the initial estimate of d. Specifically, let

$$\widehat{\mathbf{\Sigma}}_y(l) = \frac{1}{n-l}\sum_{t=1}^{n-l}(\mathbf{y}_t - \widehat{\boldsymbol{\mu}})(\mathbf{y}_{t+l} - \widehat{\boldsymbol{\mu}}'), \quad \widehat{\boldsymbol{\mu}} = \frac{1}{n}\sum_{t=1}^{n}\mathbf{y}_t, \quad \widehat{\mathbf{M}} = \sum_{l=1}^{l_0}\widehat{\mathbf{\Sigma}}_y(l)\widehat{\mathbf{\Sigma}}_y(l)'.$$

Let $\widehat{\lambda}_1, \ldots, \widehat{\lambda}_p$ be the eigenvalues of $\widehat{\mathbf{M}}$ in descending order. We use $d_0 = \arg\min_{1 \le j \le p/2}\widehat{\lambda}_{j+1}/\widehat{\lambda}_j$ as the initial value of d.

The dynamic part of the observed process at time t lies in the column space of $\mathbf{A}_k$ if $z_t = k$. Therefore, $\{\mathbf{y}_1, \ldots, \mathbf{y}_n\}$ should be located nearby the m d-dimensional subspaces $\mathcal{M}(\mathbf{A}_1), \ldots, \mathcal{M}(\mathbf{A}_m)$. With d_0, we perform a principal component analysis on $\mathbf{y}_t$ to find the $d_0 m$ directions, $\widehat{\mathbf{q}}_1, \ldots, \widehat{\mathbf{q}}_{d_0 m}$ in descending order that account for the most variation of $\{\mathbf{y}_1, \ldots, \mathbf{y}_n\}$. We can then construct the m subspaces, $\mathcal{S}_1, \ldots, \mathcal{S}_m$ by dividing the set $\{\widehat{\mathbf{q}}_1, \ldots, \widehat{\mathbf{q}}_{d_0 m}\}$ into m groups that minimizes the squared distance between $\mathbf{y}_t$ and its closest subspace, i.e. maximizes the squared projection of $\mathbf{y}_t$ onto its closest subspace. Specifically, let $\mathbf{s}_t = \{s_{1,t}, \ldots, s_{d_0 m,t}, \ldots, s_{p,t}\}$ be the principal component scores of $\mathbf{y}_t$, and $\{\mathcal{K}_1, \ldots, \mathcal{K}_m\}$ be a partition of the index set of $\{1, \ldots, d_0 m\}$, each $\mathcal{K}_i$ contains d_0 elements. Define

$$W(\mathcal{K}_1, \ldots, \mathcal{K}_m) = \sum_{t=1}^{n}\max_{1 \le i \le m}\sum_{j \in \mathcal{K}_i} s_{j,t}^2, \tag{3.19}$$

and select the partition $\{\mathcal{K}_1^*, \ldots, \mathcal{K}_m^*\}$ that maximizes W. Note that $\sum_{j\in\mathcal{K}_i}(s_{j,t})^2$ is the squared norm of the projection of $\mathbf{y}_t$ onto the space $\mathcal{S}_i$, where $\mathcal{S}_i = \mathcal{M}(\mathbf{q}_j,\, j \in \mathcal{K}_i)$, and it is maximized by the index corresponding to the subspace to which $\mathbf{y}_t$ is the closest, from $\{\mathcal{S}_1, \ldots, \mathcal{S}_m\}$. Hence, the initial values of state variables can be set as

$$\widehat{z}_t = \arg \max_{1\le i\le m} \sum_{j\in\mathcal{K}_i^*} s_{j,t}^2.$$

Finding the optimal partition is computationally extensive, unless $d_0 m$ is small. With large $d_0 m$, one can use a procedure similar to K-mean clustering to find a tentative solution, as the procedure is only for searching a good set of initial values.

Since the directions obtained by principal component analysis are orthogonal to each other, the constructed subspaces $\mathcal{S}_1, \ldots, \mathcal{S}_m$ are also orthogonal. However, we do not assume that for $\mathcal{M}(\mathbf{A}_1), \ldots, \mathcal{M}(\mathbf{A}_m)$, so the constructed subspaces $\mathcal{S}_1, \ldots, \mathcal{S}_m$ are not necessarily good estimates for loading spaces. It follows that the $d_0 m$ orthogonal directions are often more than needed and there may be states which only a few observations are assigned to. If it happens, a smaller d_0 can be used.

3.3 Theoretical Properties

In this section, we first investigate the convergence rates of the proposed estimator $\mathcal{M}(\widehat{\mathbf{Q}}_k)$ and $\widehat{d}$ as n and p go to infinity, given true state classification $\mathbf{z}$, under the setting in Section 3.2.1. Second, we introduce a theorem regarding misclassification under the setting in Section 3.2.2.

Some regularity conditions are needed.

Condition 1. The process $\mathbf{x}_t$ is α-mixing with mixing coefficients satisfying $\sum_{t=1}^{\infty} \alpha(t)^{1-2/\gamma} < \infty$, for some $\gamma > 2$, where

$$\alpha(t) = \sup_i \sup_{A\in\mathcal{F}_{-\infty}^i, B\in\mathcal{F}_{i+t}^{\infty}} |P(A\cap B) - P(A)P(B)|,$$

and $\mathcal{F}_i^j$ is the σ-field generated by $\{\mathbf{x}_t : i \le t \le j\}$.

Condition 2. For any $j = 1, ..., d$, and $t = 1, ..., n$, $\mathrm{E}(|x_{j,t}|^{2\gamma}) \le C$, where $x_{j,t}$ is the

j-th element of $\mathbf{x}_t$, $C > 0$ is a constant and γ is given in Condition 1. For $l = 1, \ldots, l_0$, $\mathbf{\Sigma}_x(l)$ is of full rank, and $\|\mathbf{\Sigma}_x(l)\|_2 \asymp O(1) \asymp \|\mathbf{\Sigma}_x(l)\|_{\min}$.

Condition 3. Each element of $\mathbf{\Sigma}_k$, for $k = 1, ..., m$, remains bounded as p increases to infinity.

Condition 4. For each k, $k = 1, \ldots, m$, there exists a constant $\delta_k \in [0, 1]$ such that $\|\mathbf{A}_k\|_2^2 \asymp p^{1-\delta_k} \asymp \|\mathbf{A}_k\|_{\min}^2$, as p goes to infinity.

Condition 5. The Markov chain $\mathbf{z}$ is irreducible, positive recurrent and aperiodic.

Condition 6. For each $k = 1, \ldots, m$, define $\mathcal{C} = \{j \mid \delta_j = \min_{1 \le k \le m} \delta_k\}$ containing all the indices of the strongest states, and for each state k there exists an integer l_k, satisfying that $l_k \le l_0$, $\sum_{j \in \mathcal{C}} \pi_{k,j}^{(l_k)} \mathbf{A}_j$ is of rank d, and

$$\Big\| \sum_{j \in \mathcal{D}} \pi_{k,j}^{(l_k)} \mathbf{A}_j \Big\|_{\min}^2 \asymp p^{1-\delta_{\min}}. \tag{3.20}$$

Condition 7. For $k = 1, \ldots, m$, $\mathbf{M}_k$ in (3.9) has d distinct positive eigenvalues. For $j \neq k$, $\mathcal{D}(\mathbf{Q}_j, \mathbf{Q}_k) \neq 0$, where $\mathcal{D}(\cdot, \cdot)$ is defined in (3.18).

Remark 3.8. The stationarity of the latent process is not required, though we do require the mixing conditions stated in Condition 1.

Remark 3.9. $\mathbf{M}_k$, the quadratic form of autocovariance matrices of $\mathbf{y}_t$ depends on observations from other regimes, including the strongest regime. Hence the most dense loading matrices influence the estimation of the loading space for each state. Condition (3.20) requires that at one of the nonzero lags, the impact of the dense loading matrices do not be cancelled out each other. It is also used to bound $\|\mathbf{M}_k\|_{\min}$.

Remark 3.10. Condition 7 makes $\mathbf{Q}_k$ uniquely defined and identifiable, where $\mathbf{Q}_k = (\mathbf{q}_{k,1}, \ldots, \mathbf{q}_{k,d})$, where $\mathbf{q}_{k,1}, \ldots, \mathbf{q}_{k,d}$ are the d orthonormal eigenvectors of $\mathbf{M}_k$ corresponding to the d nonzero eigenvalues $\lambda_{k,1} > \ldots, > \lambda_{k,d}$.

Theorem 3.1. *If Conditions 1-7 hold, with given observed state* $\mathbf{z}$ *and true* d*, for* $k = 1, \ldots, m$, $p^{\delta_k/2 + \delta_{\min}/2} n^{-1/2} \to 0$ *as* $n, p \to \infty$*, we have*

$$\|\widehat{\mathbf{Q}}_k - \mathbf{Q}_k\|_2 = O_p(p^{\delta_k/2 + \delta_{\min}/2} n^{-1/2}),$$

where $\delta_{\min} = \min_{1\le k\le m} \delta_k$.

When there is only one state $m = 1$, a special case of our setting, Lam et al. (2011) proved that the convergence rate of the estimator of the loading space is $O_p(p^{\delta} n^{-1/2})$. For the regime switching model with $m > 1$, our results above show that, except for the strongest states (with $\delta_{\min}$), the estimators of loading spaces for all the weaker states converge faster than $p^{\delta_k} n^{-1/2}$. In other words, the estimators of the loading spaces for the strongest states retain the same convergence rate, while these for other states gain some efficiency from regime switching mechanism. The main reason is that our approach depends on the autocovariance matrices of $\mathbf{y}_t$ given $z_t = k$ at leads $1, \ldots, l_0$. It is a linear combination of autocovariance matrices given current state k switching to all the states. The autocovariance matrices switching to the strongest states have the leading order and all other terms are of smaller order.

Theorem 3.2. *If Conditions 1-7 hold, with observed state* $\mathbf{z}$ *and true* d*, for* $k = 1, \ldots, m$*,* $p^{\delta_k/2+\delta_{\min}/2} n^{-1/2} \to 0$*, and* $\|\mathbf{\Sigma}_k\|_2$ *is bounded, as* $n, p \to \infty$*, we have*

$$p^{-1/2} \|\widehat{\mathbf{f}}_t - \mathbf{f}_t\|_2 = O_p(p^{\delta_{\min}/2} n^{-1/2} + p^{-1/2}).$$

Theorem 3.2 provides the convergence of the extracted factor term, and the rate does not vary across regimes, free of δ_k. It shows that by introducing stronger states, the estimated dynamic part of the observed process shows an overall improvement.

If the distance measure in (3.18) is adopted for the loading space $\mathcal{M}(\mathbf{Q}_k)$, then we have the following theorem for its estimation error.

Theorem 3.3. *If Conditions 1-7 hold, with observed state* $\mathbf{z}$ *and true* d*, for* $k = 1, \ldots, m$*,* $p^{\delta_k/2+\delta_{\min}/2} n^{-1/2} \to 0$ *as* $n, p \to \infty$*, we have*

$$\mathcal{D}(\widehat{\mathbf{Q}}_k, \mathbf{Q}_k) = O_p(p^{\delta_k/2+\delta_{\min}/2} n^{-1/2}).$$

Theorem 3.3 shows that the error for estimated loading space is on the same order as that for the estimated $\mathbf{Q}_k$ when $\mathbf{Q}_k$ is uniquely defined as in Remark 3.10.

Theorem 3.4. *If Conditions 1-7 hold, with observed states* $\mathbf{z}$*, for* $k = 1, \ldots, m$*,* $h_{n,k} = p^{\delta_k/2+\delta_{\min}/2} n^{-1/2} \to 0$ *as* $n,\ p \to \infty$ *, the eigenvalues* $\{\widehat{\lambda}_{k,1}, \ldots, \widehat{\lambda}_{k,p}\}$ *of* $\widehat{\mathbf{M}}_k$ *satisfy*
(1) $|\widehat{\lambda}_{k,i} - \lambda_{k,i}| = O_p(p^{2-\delta_k/2-\delta_{\min}/2} n^{-1/2})$ *for* $i = 1, ..., d$*, and*
(2) $\widehat{\lambda}_{k,j} = O_p(p^2 n^{-1})$ *for* $j = d+1, \ldots, p$*.*

Corollary 3.1. *Under the conditions of Theorem 3.4, we have*
$\widehat{\lambda}_{k,j+1}/\widehat{\lambda}_{k,j} \asymp 1$ *for* $j = 1, \ldots, d$*, and* $\widehat{\lambda}_{k,d+1}/\widehat{\lambda}_{k,d} = O_p(p^{\delta_k+\delta_{\min}} n^{-1})$*, for* $k = 1, \ldots, m$*.*

Theorem 3.4 shows that the estimators for the d nonzero eigenvalues of $\mathbf{M}_k$ converge slower than those for the $p - d$ zero eigenvalues. Corollary 3.1 gives the order of ratio of the estimated eigenvalues. Hence, it provides partial theoretical support for the ratio estimator proposed in Section 3.2.1. Because of differences in δ_k, the stronger the state k is, the faster convergence rate $\widehat{\lambda}_{k,d+1}/\widehat{\lambda}_{k,d}$ has. Therefore, we choose $\widehat{d}_{\tilde{k}}$ as the estimator of the number of factors using the state $\tilde{k}$ that the maximizing $\|\widehat{\mathbf{M}}_k\|_2$, since it is related through $\|\widehat{\mathbf{M}}_k\|_2 = O_p(p^{2-\delta_k-\delta_{\min}})$, which is proved by Lemma 3.3 and Lemma 3.4 in Appendix.

Remark 3.11. The asymptotics of $\widehat{\lambda}_{k,i+1}/\widehat{\lambda}_{k,i}$ with $i > d$ are difficult to obtain, even when $m = 1$; see Remark 2 in Lam and Yao (2012). Chang et al. (2013) adjusted the ratio estimator as follows

$$\widehat{d} = \arg \min_{1 \le j \le p/2} \{ \frac{\widehat{\lambda}_{j+1} + C_T}{\widehat{\lambda}_j + C_T} \}, \tag{3.21}$$

where $C_t = p^{2-\delta} n^{-1/2} \log n$ for one-regime model, and proved it is a consistent estimator for d. However, the adjusted ratio estimator in (3.21) can not be used for data analysis as δ is unknown. In practice, the ratio estimator in (3.12) is used; see Lam et al. (2011), Lam and Yao (2012) and Chang et al. (2013).

Next we investigate the performance of the estimator of the state $\mathbf{z}$. To simplify the investigation, in the following we assume $\pi_{k,j}$ for $k, j = 1, \ldots, m$ are all equal, hence estimating z_t can be done separately for each t, instead of relying on the Veterbi algorithm. It is also equivalent to pure classification without the Markov chain mechanism. The setting is not exactly what we assumed in Section 3.2.2, but the results reveal how

misclassification occurs and its impact on the estimation of the rest of the parameters.

Let

$$w_{t,k,j} = \log[f(\mathbf{B}_k'\mathbf{y}_t)] - \log[f(\mathbf{B}_j'\mathbf{y}_t)] = l(k|\mathbf{y}_t) - l(j|\mathbf{y}_t), \tag{3.22}$$

where

$$\begin{aligned} l(k|\mathbf{y}_t) &= \log(f(\mathbf{B}_k'\mathbf{y}_t)) \\ &= -\frac{p-d}{2}\log(2\pi) - \frac{1}{2}\log|\mathbf{\Sigma}_{B,k}| - \frac{(\mathbf{B}_k'(\mathbf{y}_t - \boldsymbol{\mu}_k))'\mathbf{\Sigma}_{B,k}^{-1}\mathbf{B}_k'(\mathbf{y}_t - \boldsymbol{\mu}_k)}{2}. \end{aligned} \tag{3.23}$$

The estimator of z_t under equal transition probability assumption can be rewritten as $\widehat{z}_t = k$ if $w_{t,k,j} > 0$ for all $j \neq k$. Hence misclassification occurs when there exists a j such that $w_{t,z_t,j} < 0$. Specifically, the probability of misclassification, when $z_t = k$, is

$$P(\widehat{z}_t \neq z_t = k) = P(\min_{j \neq k} w_{t,k,j} < 0).$$

Note that, when $z_t = k$,

$$\begin{aligned} l(k|\mathbf{y}_t) &= -\frac{p-d}{2}\log(2\pi) - \frac{1}{2}\log|\mathbf{\Sigma}_{B,k}| - \frac{1}{2}(\mathbf{B}_k'(\mathbf{y}_t - \boldsymbol{\mu}_k))'{\mathbf{\Sigma}_{B,k}}^{-1}\mathbf{B}_k'(\mathbf{y}_t - \boldsymbol{\mu}_k) \\ &= -\frac{p-d}{2}\log(2\pi) - \frac{1}{2}\log|\mathbf{\Sigma}_{B,k}| - \frac{1}{2}(\boldsymbol{\varepsilon}_t^{(k)\prime}\mathbf{B}_k{\mathbf{\Sigma}_{B,k}}^{-1}\mathbf{B}_k'\boldsymbol{\varepsilon}_t^{(k)}), \end{aligned}$$

and

$$\begin{aligned} l(j|\mathbf{y}_t) &= -\frac{p-d}{2}\log(2\pi) - \frac{1}{2}\log|\mathbf{\Sigma}_{B,j}| \\ &\quad -\frac{1}{2}\left(\mathbf{B}_j'(\mathbf{A}_k\mathbf{x}_t + \boldsymbol{\mu}_k - \boldsymbol{\mu}_j + \boldsymbol{\varepsilon}_t^{(k)})\right)'{\mathbf{\Sigma}_{B,j}}^{-1}\left(\mathbf{B}_j'(\mathbf{A}_k\mathbf{x}_t + \boldsymbol{\mu}_k - \boldsymbol{\mu}_j + \boldsymbol{\varepsilon}_t^{(k)})\right). \end{aligned}$$

Hence,

$$\begin{aligned} w_{t,k,j} &= \frac{1}{2}(\log|\mathbf{\Sigma}_{B,j}| - \log|\mathbf{\Sigma}_{B,k}|) + \frac{1}{2}\boldsymbol{\varepsilon}_t^{(k)\prime}(\mathbf{B}_j{\mathbf{\Sigma}_{B,j}}^{-1}\mathbf{B}_j' - \mathbf{B}_k{\mathbf{\Sigma}_{B,k}}^{-1}\mathbf{B}_k')\boldsymbol{\varepsilon}_t^{(k)} \\ &\quad + \frac{1}{2}\left(\mathbf{B}_j'\mathbf{A}_k\mathbf{x}_t + \mathbf{B}_j'(\boldsymbol{\mu}_k - \boldsymbol{\mu}_j)\right)'{\mathbf{\Sigma}_{B,j}}^{-1}\left(\mathbf{B}_j'\mathbf{A}_k\mathbf{x}_t + \mathbf{B}_j'(\boldsymbol{\mu}_k - \boldsymbol{\mu}_j)\right) \\ &\quad + (\mathbf{B}_j'(\mathbf{A}_k\mathbf{x}_t + \boldsymbol{\mu}_k - \boldsymbol{\mu}_j))'\mathbf{\Sigma}_{B,j}^{-1}\mathbf{B}_j'\boldsymbol{\varepsilon}_t^{(k)} \\ &= I_1 + I_2 + I_3 + I_4. \end{aligned} \tag{3.24}$$

Here I_1 is a given constant, measuring the differences in variation of the two states. I_2 reflects the impact of the noise, after being projected into the space $\mathbf{B}_k$ and $\mathbf{B}_j$. The third term I_3 shows the size of the noises needed for misclassification. In addition, $\mathbf{B}_j'\mathbf{A}_k\mathbf{x}_t$ is the projection of $\mathbf{x}_t$ to the intersection of $\mathcal{M}(\mathbf{B}_j)$ and $\mathcal{M}(\mathbf{A}_k)$. If $\mathcal{M}(\mathbf{A}_k)$ and $\mathcal{M}(\mathbf{A}_j)$ are less common, then $\mathbf{B}_j'\mathbf{A}_k\mathbf{x}_t$ is larger (in magnitude), and the chance for misclassification is less. Of course, if the difference in the mean $\boldsymbol{\mu}_k - \boldsymbol{\mu}_j$ is larger, then misclassification probability will be smaller. I_4 is the cross term of I_2 and I_3.

Misclassification of z_t may not have large impact on the estimation of the loading spaces. Small $\mathbf{B}_j'\mathbf{A}_k\mathbf{x}_t$ will lead to misclassifying the observation from state k to state j. It happens in two situations. For some observations, $\mathbf{A}_k\mathbf{x}_t$ are close to the column space of $\mathbf{A}_j$, which makes $\mathbf{B}_j'\mathbf{A}_k\mathbf{x}_t$ small values. Such observations hence lie close to the space $\mathcal{M}(\mathbf{A}_j)$ and do not have large impact in the estimation of $\mathcal{M}(\mathbf{A}_j)$. Other possibility is that these misclassified observations have a small signal-to-noise ratio which makes $\mathbf{B}_j'\mathbf{A}_k\mathbf{x}_t$ small, hence they are less influential for estimation of $\mathcal{M}(\mathbf{A}_j)$.

From a different angle, we can calculate the expectation and variance of $w_{t,k,j}$.

Theorem 3.5. *If $\mathbf{x}_t$ is a normally distributed random process, given true $\mathbf{B}_k$, $\boldsymbol{\mu}_k$, and $\mathbf{\Sigma}_{B,k}$, $k = 1, \ldots, m$, we have*

$$\mathrm{E}(w_{t,k,j}) = \frac{1}{2}(\log|\mathbf{\Sigma}_{B,j}| - \log|\mathbf{\Sigma}_{B,k}|) + \frac{1}{2}\mathrm{tr}\left((\mathbf{\Sigma}_k + \mathbf{\Sigma}_{f,k,t} + \mathbf{U}_{k,j})\mathbf{W}_j\right) - \frac{(p-d)}{2}, \tag{3.25}$$

$$\begin{aligned}
&\mathrm{Var}(w_{t,k,j}) \\
&= \frac{1}{2}\|\mathbf{\Sigma}_k^{1/2}(\mathbf{W}_j - \mathbf{W}_k)\mathbf{\Sigma}_k^{1/2}\|_F^2 + \frac{1}{2}\|\mathbf{\Sigma}_k^{1/2}\mathbf{A}_k'\mathbf{W}_j\mathbf{A}_k\mathbf{\Sigma}_k^{1/2}\|_F^2 + \frac{1}{2}\mathrm{tr}\left(\mathbf{\Sigma}_{f,k,t}\mathbf{W}_j\mathbf{U}_{k,j}\mathbf{W}_j\right) \\
&+ \mathrm{tr}\left((\mathbf{\Sigma}_{f,k,t} + \mathbf{U}_{k,j})\mathbf{W}_j\mathbf{\Sigma}_k\mathbf{W}_j\right). \qquad (3.26)
\end{aligned}$$

where $\mathbf{\Sigma}_{f,k,t} = \mathrm{Var}(\mathbf{A}_k\mathbf{x}_t)$, *and* $\mathbf{U}_{k,j} = (\boldsymbol{\mu}_k - \boldsymbol{\mu}_j)'(\boldsymbol{\mu}_k - \boldsymbol{\mu}_j)$, $\mathbf{W}_j = \mathbf{B}_j\mathbf{\Sigma}_{B,j}^{-1}\mathbf{B}_j'$.

The mean and variance of $\omega_{t,k,j}$ increase with p. As expected, misclassification is unavoidable. For weak states, as p increases, the accumulated noises tend to overwhelm the difference in $\mathbf{A}_k\mathbf{x}_t + \boldsymbol{\mu}_k - \boldsymbol{\mu}_j$. Hence classification error may increase with p. On the other hand, for strong states, signal remains strong and misclassification rate will be much better than these for weak states.

The theoretical investigation above only helps to reveal the important features of each step and provides general guidance in implementations.

3.4 Simulations

In this section, we illustrate the performance of the proposed estimators with some numerical experiments, compare their convergence rates for states with different strength, and explore the interactions among states. The performance of the estimators of $\mathcal{M}(\mathbf{Q}_k)$ and d, and the performance of clustering are presented separately.

With two switching regimes $m = 2$, we consider three models. In Model 1, both states are strong, with $\delta_1 = \delta_2 = 0$. In Model 2, one of the states is strong and one is weak, with $\delta_1 = 0$, and $\delta_2 = 1$. In Model 3, both states are weak, with $\delta_1 = \delta_2 = 1$. The transition probabilities between the two states are set to 0.5. In the simulation, all $p \times d$ entries in $\mathbf{A}_k$ are generated independently from the uniform distribution on $[-p^{-\delta_k/2}, p^{-\delta_k/2}]$ with strength δ_k. The mean of observed process $\boldsymbol{\mu}_k$ is a $p \times 1$ vector with all entries zero, for $k = 1, 2$. Different values of d, and different structures of the latent process and noises are used. In all the examples, we use $l_0 = 1$. Estimation error of $\mathcal{M}(\widehat{\mathbf{Q}}_k)$ is defined as $\mathcal{D}(\widehat{\mathbf{Q}}_k, \mathbf{Q}_k)$.

3.4.1 The Performance of $\mathcal{M}(\widehat{\mathbf{Q}}_k)$

In this experiment d is set to 1 and we estimate the loading spaces using true d. The factor process x_t is from an AR(1) process with AR coefficient 0.9 and $N(0, 4)$ noises. The noise process $\{\boldsymbol{\varepsilon}_t^{(1)}, \ldots, \boldsymbol{\varepsilon}_t^{(m)}\}$ are m independent vector white noise processes whose covariance matrix has 1 on the diagonal and 0.95 as the off-diagonal entries. Set the pre-specified controls t_0, c_1, and c_2 in the iterative algorithm in Section 3.3 to 50, 0.001 and 0.001 respectively.

We repeat the simulation 100 times with sample size $n = 1000$. Let $p = 20, 40, 80$. Table 3.1 and Figure 3.1 show the results for when $\mathbf{z}$ is observed and when $\mathbf{z}$ is unobserved.

Table 3.1: Means of the estimation errors $\mathcal{D}(\widehat{\mathbf{Q}}_k, \mathbf{Q}_k)$

		$\mathbf{z}$ observed			$\mathbf{z}$ unobserved		
		$p = 20$	$p = 40$	$p = 80$	$p = 20$	$p = 40$	$p = 80$
Model 1	State 1 ($\delta_1 = 0$)	0.0159	0.0164	0.0161	0.0438	0.0606	0.1055
	State 2 ($\delta_2 = 0$)	0.0143	0.0155	0.0169	0.0445	0.0711	0.0958
Model 2	State 1 ($\delta_1 = 0$)	0.0203	0.0216	0.0207	0.0225	0.0274	0.0304
	State 2 ($\delta_2 = 1$)	0.0856	0.1274	0.2131	0.0977	0.1495	0.2689
Model 3	State 1 ($\delta_1 = 1$)	0.0796	0.1489	0.4149	0.2424	0.5563	0.6067
	State 2 ($\delta_2 = 1$)	0.0803	0.1453	0.4226	0.2626	0.5091	0.6614

When the state variable $\mathbf{z}$ is observed, by comparing the results of Model 2 to these of Model 1, we can see that the estimation for the strong state is slightly worse after a weak state is introduced to the model. However, by comparing the results of Model 2 to these of Model 3, the estimation for the weaker state is much better due to the existence of a strong state, especially when p is large. It shows that the estimation for weak states benefits from the stronger states as our theory indicates.

The top panels in Figure 3.1 display the boxplots of estimation errors for different p when $\mathbf{z}$ is observed in each model on the same scale. In addition to what we can see from Table 3.1, it shows that the estimation variation increases with p and is larger for the weak states.

When $\mathbf{z}$ is unobserved, the estimation errors of the loading spaces for each model with different p have similar pattern to these in the case when $\mathbf{z}$ is known shown in

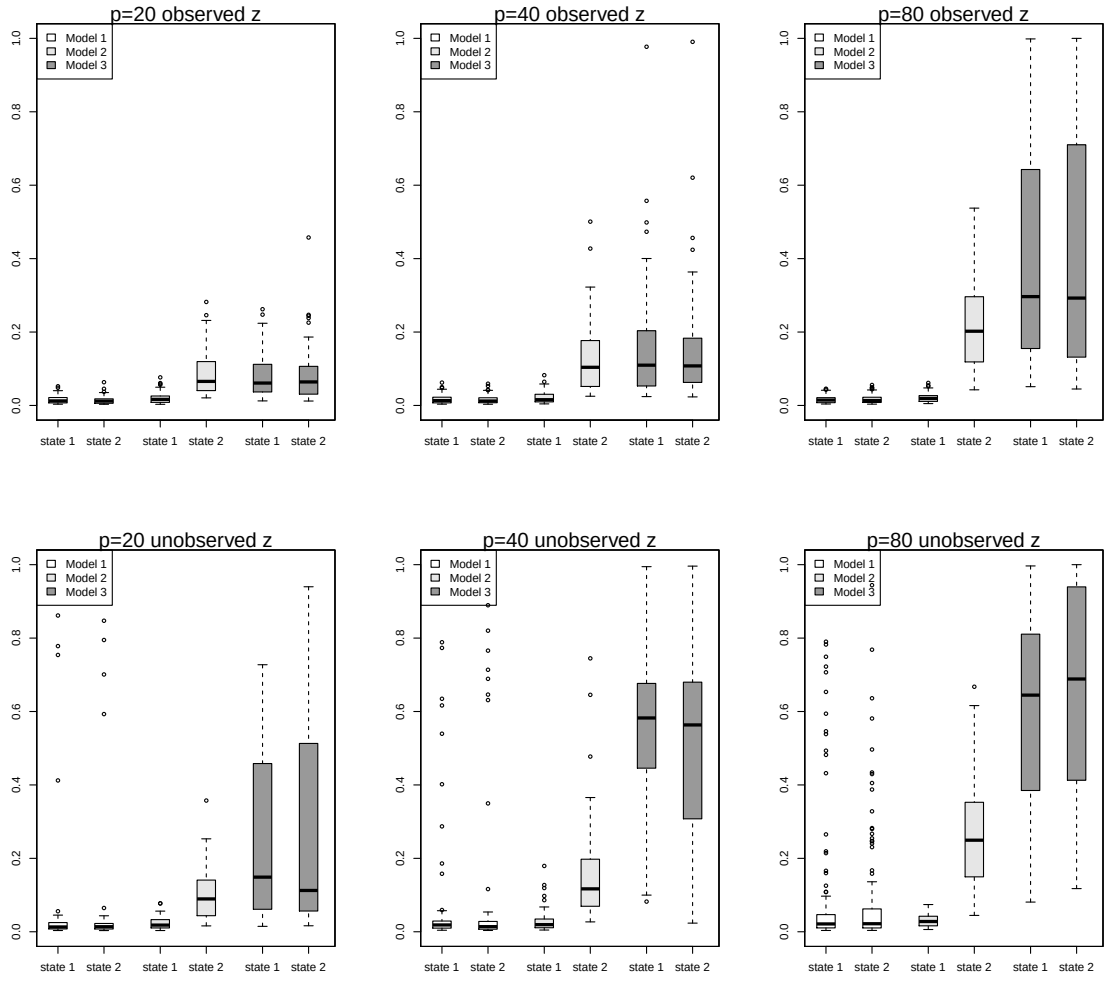


Figure 3.1: Boxplots of estimation errors of $\mathcal{M}(\widehat{\mathbf{Q}}_k)$ for $p = 20, 40, 80$ when $\mathbf{z}$ is observed (top panels), and when $\mathbf{z}$ is unobserved (bottom panels), under true d with model described in Section 3.4.1.

Table 3.1 and the bottom panels in Figure 3.1. The estimation of strong states still has a good performance in absence of weak states; for weak states, it benefits from the existence of strong states as well. The estimators are less accurate if the state variable is unobserved. As p increases, even the strong states suffer from unobserved $\mathbf{z}$. Because of lack of information on $\mathbf{z}$, it happens that our algorithm is trapped in a local maximum.

3.4.2 The Clustering Performance

In this experiment we use the settings in Section 3.4.1 for a comparison of the clustering performance among models with different strength. Results of misclassification rates and transition probabilities are summarized in Tables 3.2 and 3.3, respectively.

Table 3.2: Means(sd) of misclassification rates of the hidden states

	$p = 20$	$p = 40$	$p = 80$
Model 1	0.0334(0.0909)	0.0485(0.1169)	0.1153(0.1418)
Model 2	0.0452(0.0128)	0.0635(0.0609)	0.1525(0.0723)
Model 3	0.2155(0.1905)	0.4439(0.0482)	0.4686(0.0434)

Table 3.2 shows the misclassification rates for each model with different p. It is seen that misclassification occurs very often when all the states are weak, but occur sometime in the presence of at least one strong state.

Table 3.3: Means(sd) of estimated transition matrices. The true values are all 0.5.

	$p = 20$		$p = 40$		$p = 80$	
Model 1	0.5110	0.4890 (0.1008)	0.6435	0.3565 (0.0689)	0.5273	0.4727 (0.1134)
	0.5000	0.5000 (0.0877)	0.6133	0.3867 (0.0582)	0.4918	0.5082 (0.1021)
Model 2	0.5322	0.4678 (0.0296)	0.5423	0.4577 (0.0431)	0.4996	0.5004 (0.0558)
	0.5219	0.4781 (0.0260)	0.5373	0.4627 (0.0422)	0.4739	0.5261 (0.0601)
Model 3	0.5153	0.4847 (0.2145)	0.5861	0.4139 (0.3114)	0.4240	0.5760(0.3339)
	0.4799	0.5201 (0.2201)	0.5200	0.4800 (0.3129)	0.3686	0.6314(0.3388)

Table 3.3 shows the means and standard deviations of the estimated transition probabilities, where the true transition probabilities are all 0.5. For Model 3, because of random noises and lack of information, some observations are misclassified to each state with larger probability comparing to Model 1 and Model 2, since the standard deviations of estimates of transition probabilities for Model 3 are much larger than these for Model 1 and Model 2.

3.4.3 The Performance of $\widehat{d}$

In this experiment we set the number of factors to 3 ($d = 3$) and investigate the performance of the proposed estimator for d, under true $\mathbf{z}$. Here the latent process $\mathbf{x}_t$ is set to be three independent AR(1) processes with $N(0, 4)$ noises and AR coefficients 0.6,

-0.5 and 0.8, respectively. $\{\boldsymbol{\varepsilon}_t^{(1)}, \ldots, \boldsymbol{\varepsilon}_t^{(m)}\}$ are m white noise process whose covariance matrix has 1 on the diagonal and 0.2 as the off-diagonal entries. Let $n = 50$, 100, 200, 500, 1000, and $p = 0.1n$, $0.5n$, $0.8n$. We repeat the simulation 200 times for each (n, p) setting and the relative frequencies of correct estimates of d are reported in Table 3.4.

Table 3.4: The relative frequency estimates of $\widehat{d} = d$

	n	50	100	200	500	1000
Model 1	$p = 0.1n$	0.180	0.385	0.725	0.995	1
	$p = 0.5n$	0.380	0.610	0.850	0.995	1
	$p = 0.8n$	0.390	0.585	0.855	0.995	1
Model 2	$p = 0.1n$	0.200	0.405	0.820	1	1
	$p = 0.5n$	0.365	0.605	0.915	1	1
	$p = 0.8n$	0.380	0.620	0.905	1	1
Model 3	$p = 0.1n$	0.115	0.125	0.075	0.075	0.275
	$p = 0.5n$	0.055	0.100	0.200	0.065	0
	$p = 0.8n$	0.080	0.080	0.060	0	0

From Table 3.4 we can see that the existence of a strong state, no matter whether or not there is a weaker state, produces much more accurate estimates for the number of factors d. As n increases, the estimations all improve in the presence of a strong state. Regarding the impact of p, it is seen that the estimation of d benefits from 'blessing of dimensionality' when one or more strong states exist, and performs better as p increases. However, when all states are extremely weak ($\delta_1 = \delta_2 = 1$), the number of correct estimation goes to 0 as n increases. The features do not change much with p, partially because the increase of information in n offsets the increase of noise introduced as p increases.

3.5 Real Data Analysis

We apply our approach to the daily returns of 123 stocks from January 2, 2002 to July 11, 2008. These stocks were selected among those included in the S&P 500 and traded every day during the period. The returns were calculated in percentages based on daily closing prices. This data was analyzed by Lam and Yao (2012) and Chang et al. (2013). We have the sample size $n = 1642$ and the dimension of observations $p = 123$. We assume that there are two regimes $m = 2$, and set $l_0 = 1$, $t_0 = 200$, $c_0 = c_1 = 0.001$.

Varying the value of l_0 does not change the estimation results significantly.

The proposed iterative procedure yields $\widehat{d} = 1$. Different from the number of factors estimated in Lam and Yao (2012), allowing the loading matrix to change across two regimes reduces the number of factors needed. This only factor accounts for 25.92% of the total variation of stock returns. The residuals $\widehat{\boldsymbol{\varepsilon}}_t$ are computed with (3.14). The sample cross-autocorrelations of $\widehat{\boldsymbol{\varepsilon}}_t$ for the first 7 stocks are plotted in Figure 3.2. There are almost no significant nonzero autocorrelations for $\widehat{\boldsymbol{\varepsilon}}_t$, showing that after extracting the latent factor, little serial dependence is left in the data. Our results indicate that only one factor drives the 123 stocks, but the factor loadings switch between two states. Ignoring the switching structure as in Lam and Yao (2012), it would appear that there are two different factors.

Note that even with $d = 1$, $\mathbf{Q}_k$ is still not unique due to a trivial replacement of $(\mathbf{Q}_k, \mathbf{R}_t I(z_t = k))$ by $(-\mathbf{Q}_k, -\mathbf{R}_t I(z_t = k))$ for either $k = 1$ or $k = 2$ or both in (3.7). According to (3.3) and (3.5), let

$$\mathbf{A}_k = \gamma_k \mathbf{Q}_k \quad \text{and} \quad \mathbf{R}_t I(z_t = k) = \gamma_k \mathbf{x}_t,$$

where we could set γ_k to 1 or -1 when the dimension is fixed. Here we choose γ_k which makes the majority of the entries in $\mathbf{Q}_k$ positive, hence $\mathbf{y}_t$ is mostly positively correlated with the corresponding latent factor $\mathbf{x}_t$ and $\mathbf{R}_t$, since

$$\mathbf{y}_t = \boldsymbol{\mu}_k + \mathbf{A}_k \mathbf{x}_t + \boldsymbol{\varepsilon}_t^{(k)} = \boldsymbol{\mu}_k + \mathbf{Q}_k \mathbf{R}_t + \boldsymbol{\varepsilon}_t^{(k)}, \text{ when } z_t = k.$$

Specifically, let $\widehat{\mathbf{Q}}_k$ and $\widehat{\mathbf{R}}_t$ be the estimate of $\mathbf{Q}_k$ and $\mathbf{R}_t$ without the above sign consideration. We adjust the sign of $\widehat{\mathbf{Q}}_k$ as follows,

$$\widehat{\mathbf{Q}}_{\text{adj},k} = \begin{cases} -\widehat{\mathbf{Q}}_k, & \text{if} \quad \sum_{i=1}^{p} I(\widehat{q}_{k,i} > 0) > \sum_{i=1}^{p} I(-\widehat{q}_{k,i} > 0), \\ \widehat{\mathbf{Q}}_k, & \text{otherwise}, \end{cases} \tag{3.27}$$

where $\widehat{q}_{k,i}$ is the i-th entry in $\widehat{\mathbf{Q}}_k$ for $k = 1, 2$. The adjusted $\widehat{\mathbf{Q}}_k$ makes most of its entries positive. $\widehat{\mathbf{R}}_{\text{adj},t}$ is obtained according to $\widehat{\mathbf{Q}}_{\text{adj},1}$ and $\widehat{\mathbf{Q}}_{\text{adj},2}$.

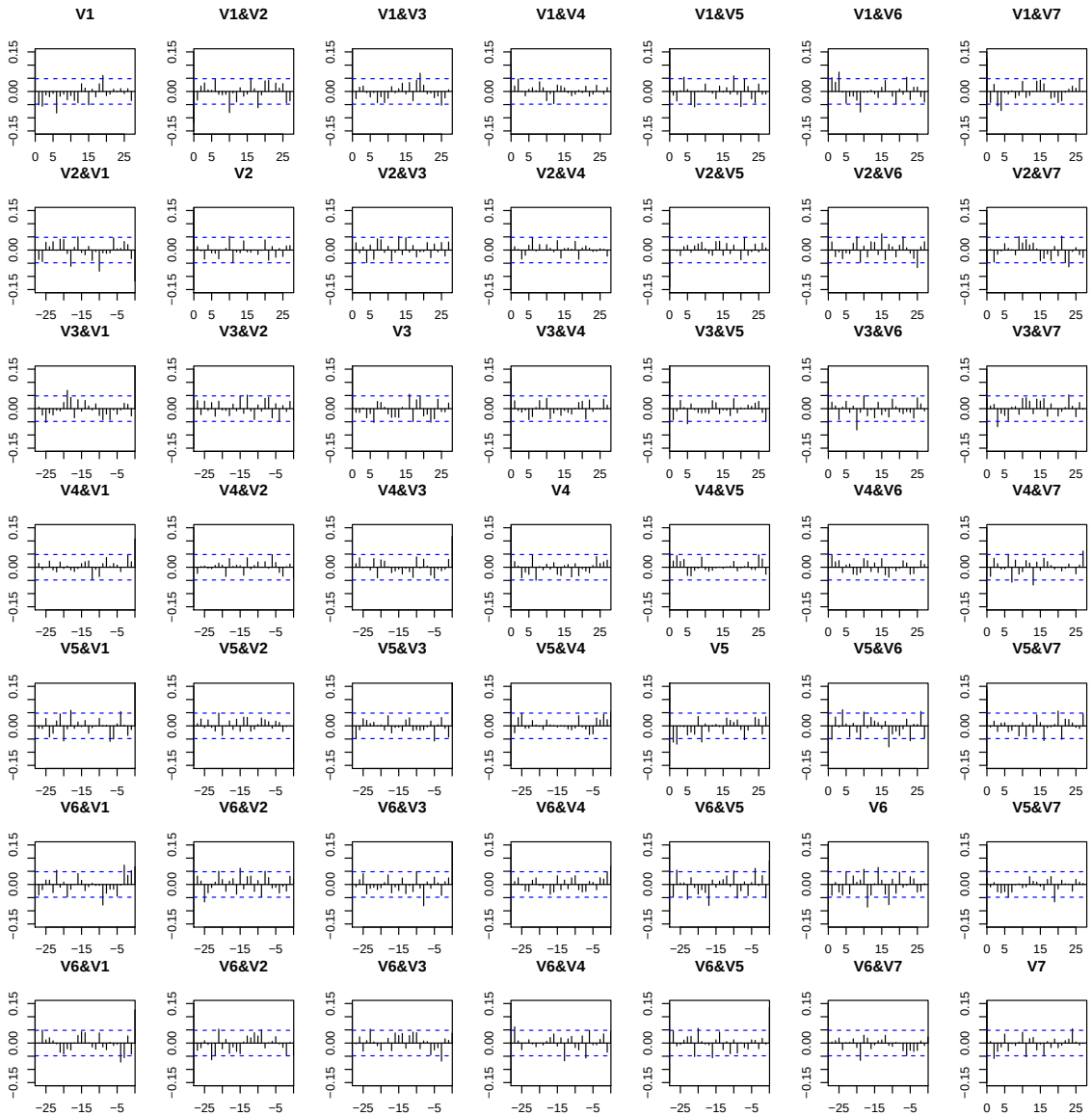


Figure 3.2: Plots of the sample cross-autocorrelations of $\widehat{\varepsilon}_t$ of the first 7 stocks with lag 0 autocorrelation removed.

Figure 3.3 displays the time series plots of $\widehat{\mathbf{R}}_{\text{adj},t}$ in the top panel and returns of the S&P 500 index in the bottom panel. $\widehat{\mathbf{R}}_{\text{adj},t}$ changes along with the S&P 500 index in this period, except for a few days around July 22, 2002, and it explains 76.27% of the total variation in the S&P 500 index. Hence, this factor can be regarded as a representation of market performance. Because index funds, which aim to replicate the movements of an index of a financial market, build their investment portfolio with all the stocks in the index and trade them together, it causes synchronous oscillations

between the market and the stocks. The popularity of index funds provides a reason that the market factor accounts for a large percentage of the total variation of stock returns.

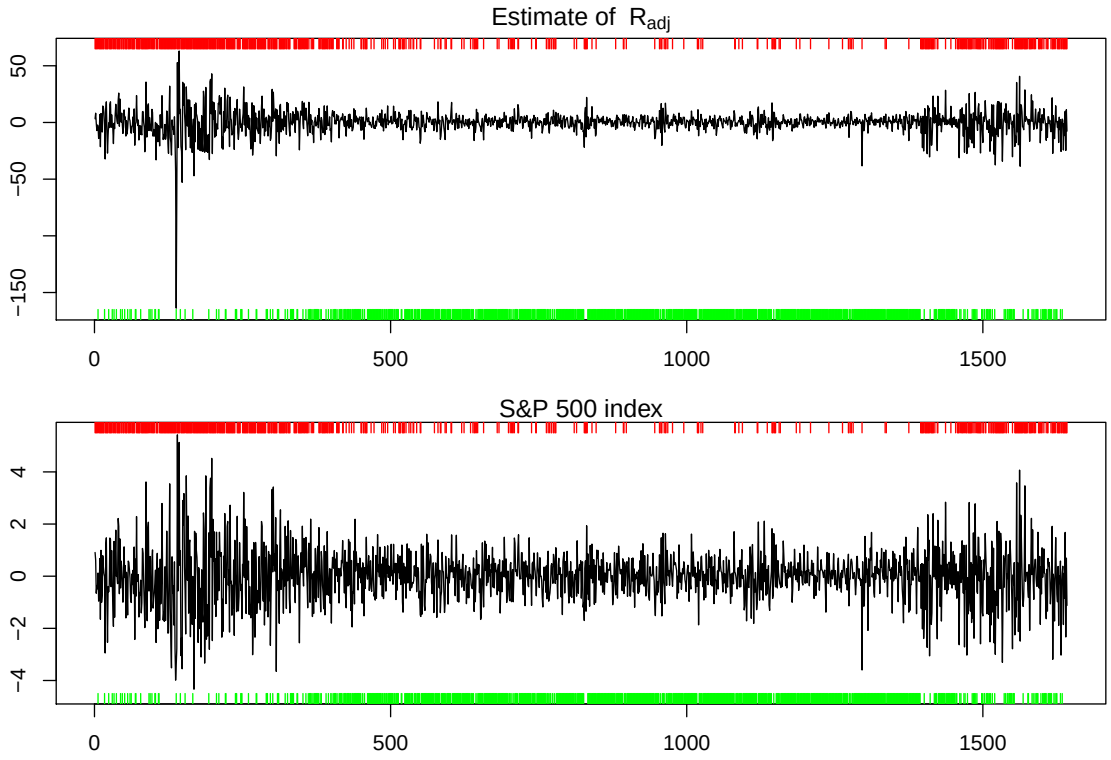


Figure 3.3: Time series plots of $\widehat{\mathbf{R}}_{\text{adj},t}$ (top panel) and the return series of the S&P 500 index (bottom panel) in the same period. Indicators of the estimated states of the observations $I(\widehat{z}_t = k)$ for $k = 1, 2$, are shown in the rug plots, on the top for State 1 and at the bottom for State 2.

The indicators of the estimated state variable $I(\widehat{z}_t = k)$, for $k = 1, 2$, are shown in the rug plots of both panels in Figure 3.4, State 1 on the top and State 2 at the bottom. It is obvious that the state variable is strongly correlated to the volatility of the market. The standard deviation of the S&P 500 index is 1.4642 given $\widehat{z}_t = 1$, while the standard deviation of the S&P 500 index is 0.6649 given $\widehat{z}_t = 2$. When the S&P 500 index was volatile in 2002, 2003 and 2007 due to internet bubble, invasion of Iraq, and subprime crisis respectively, the observations are more likely to belong to State 1; when the S&P 500 index was stable in 2004-2006, the observations tend to be assigned to State 2.

For State 1, the factor accounts for 34.89% of the total variation in $\mathbf{y}_t$, while for

State 2, it only accounts for 15.75%. A possible explanation is that investors may prefer passive management, such as index-tracking funds, to avoid nonsystematic risk when the market is volatile.

Table 3.5: Estimated transition matrix and stationary probabilities

	State 1	State 2	π_k
State 1	0.6758	0.3242	0.3782
State 2	0.1969	0.8031	0.6281

The estimated transition probabilities are shown in Table 3.5. During this period, about two third of the time the system stays in State 2. The transition between the states are quite often, especially from State 1 to State 2.

3.6 Proofs

Here we use Cs to denote the generic uniformly positive constants. Define

$$\mathbf{\Sigma}_{x,k,j}(l) = \frac{1}{n-l}\sum_{t=1}^{n-l} \mathrm{Cov}(\mathbf{x}_t,\, \mathbf{x}_{t+l} I(z_{t+l}=j) \mid z_t = k) = \pi_{k,j}^{(l)} \mathbf{\Sigma}_x(l),$$

$$\mathbf{\Sigma}_{f,k,j}(l) = \frac{1}{n-l}\sum_{t=1}^{n-l} \mathrm{Cov}\{\mathbf{f}_t,\, \mathbf{f}_{t+l} I(z_{t+l}=j) \mid z_t = k)\},$$

$$\widehat{\mathbf{\Sigma}}_{x,k,j}(l) = \frac{\sum_{t=1}^{n-l}(\mathbf{x}_t - \bar{\mathbf{x}}_k)(\mathbf{x}_{t+l} - \bar{\mathbf{x}}_j)' I(z_t = k, z_{t+l} = j)}{\sum_{t=1}^{n-l} I(z_t = k)},$$

$$\widehat{\mathbf{\Sigma}}_{\varepsilon,k,j} = \frac{\sum_{t=1}^{n-l}(\boldsymbol{\varepsilon}_t^{(k)} - \bar{\boldsymbol{\varepsilon}}_k)(\boldsymbol{\varepsilon}_{t+l}^{(j)} - \bar{\boldsymbol{\varepsilon}}_j)' I(z_t = k, z_{t+l} = j)}{\sum_{t=1}^{n-l} I(z_t = k)},$$

$$\widehat{\mathbf{\Sigma}}_{f,k,j}(l) = \frac{\sum_{t=1}^{n-l}(\mathbf{f}_t - \bar{\mathbf{f}}_k)(\mathbf{f}_{t+l} - \bar{\mathbf{f}}_j)' I(z_t = k, z_{t+l} = j)}{\sum_{t=1}^{n-l} I(z_t = k)},$$

$$\widehat{\mathbf{\Sigma}}_{f,\varepsilon,k,j}(l) = \frac{\sum_{t=1}^{n-l}(\mathbf{f}_t - \bar{\mathbf{f}}_k)(\boldsymbol{\varepsilon}_{t+l}^{(j)} - \bar{\boldsymbol{\varepsilon}}_j)' I(z_t = k, z_{t+l} = j)}{\sum_{t=1}^{n-l} I(z_t = k)},$$

$$\widehat{\mathbf{\Sigma}}_{\varepsilon,f,k,j}(l) = \frac{\sum_{t=1}^{n-l}(\boldsymbol{\varepsilon}_t^{(k)} - \bar{\boldsymbol{\varepsilon}}_k)(\mathbf{f}_{t+l} - \bar{\mathbf{f}}_j)' I(z_t = k, z_{t+l} = j)}{\sum_{t=1}^{n-l} I(z_t = k)},$$

where $\bar{\mathbf{x}}_k = \sum_{t=1}^{n} \mathbf{x}_t I(z_t = k) / \sum_{t=1}^{n} I(z_t = k)$, $\bar{\mathbf{f}}_k = \sum_{t=1}^{n} \mathbf{f}_t I(z_t = k) / \sum_{t=1}^{n} I(z_t = k)$, and $\bar{\boldsymbol{\varepsilon}}_k = \sum_{t=1}^{n} \boldsymbol{\varepsilon}_t^{(k)} I(z_t = k) / \sum_{t=1}^{n} I(z_t = k)$, for $k = 1, \ldots, m$.

We introduce some lemmas first.

Lemma 3.1. *Under Conditions 1-2 and Condition 5, if* $\pi_{k,j}^{(l)} > 0$*, we have*

$$\|\widehat{\mathbf{\Sigma}}_{x,k,j}(l) - \mathbf{\Sigma}_{x,k,j}(l)\|_2 = O_p(n^{-1/2}), \quad \text{for} \quad k, j = 1, \ldots, m. \tag{3.28}$$

Proof: Since $\mathbf{z}$ is irreducible, positive current and aperiodic under Condition 5, by Theorem 3.5 in Bradley (2005) and Theorem 17.0.1 in Meyne and Tweedie (2009), it follows

$$\frac{n}{\sum_{t=1}^{n-l} I(z_t = k)} - \frac{1}{\pi_k} = O_p(n^{-1/2}), \tag{3.29}$$

$$\frac{\sum_{t=1}^{n-l} I(z_t = k, z_{t+l} = j)}{\sum_{t=1}^{n-l} I(z_t = k)} - \pi_{k,j}^{(l)} = O_p(n^{-1/2}). \tag{3.30}$$

$$\begin{aligned}
&\widehat{\mathbf{\Sigma}}_{x,k,j}(l) - \mathbf{\Sigma}_{x,k,j}(l) \\
&= \frac{\sum_{t=1}^{n-l} (\mathbf{x}_t - \bar{\mathbf{x}}_k)(\mathbf{x}_{t+l} - \bar{\mathbf{x}}_j)' I(z_t = k, z_{t+l} = j)}{\sum_{t=1}^{n-l} I(z_t = k)} - \frac{\sum_{t=1}^{n-l} \pi_{k,j}^{(l)} \mathrm{E}(\mathbf{x}_t \mathbf{x}_{t+l}')}{n-l} \\
&= \frac{\sum_{t=1}^{n-l} \left[(\mathbf{x}_t - \bar{\mathbf{x}}_k)(\mathbf{x}_{t+l} - \bar{\mathbf{x}}_j)' - \mathrm{E}(\mathbf{x}_t \mathbf{x}_{t+l}') I(z_t = k, z_{t+l} = j) \right]}{\sum_{t=1}^{n-l} I(z_t = k)} \\
&+ \sum_{t=1}^{n-l} \left[\left(\frac{I(z_t = k, z_{t+l} = j)}{\sum_{t=1}^{n-l} I(z_t = k)} - \frac{\pi_{k,j}^{(l)}}{n-l} \right) \mathrm{E}(\mathbf{x}_t \mathbf{x}_{t+l}') \right] \\
&= I_1 + I_2,
\end{aligned}$$

where

$$\begin{aligned}
I_1 \;=\; & \frac{\sum_{t=1}^{n-l} (\mathbf{x}_t \mathbf{x}_{t+l}' - \mathrm{E}\mathbf{x}_t \mathbf{x}_{t+l}') I(z_t = k, z_{t+l} = j)}{\sum_{t=1}^{n-l} I(z_t = k)} - \frac{\sum_{t=1}^{n-l} \mathbf{x}_t \bar{\mathbf{x}}_j' I(z_t = k, z_{t+l} = j)}{\sum_{t=1}^{n-l} I(z_t = k)} \\
& - \frac{\sum_{t=1}^{n-l} \bar{\mathbf{x}}_k \mathbf{x}_{t+l}' I(z_t = k, z_{t+l} = j)}{\sum_{t=1}^{n-l} I(z_t = k)} + \frac{\sum_{t=1}^{n-l} \bar{\mathbf{x}}_k \bar{\mathbf{x}}_j' I(z_t = k, z_{t+l} = j)}{\sum_{t=1}^{n-l} I(z_t = k)} \\
=\; & L_1 + L_2 + L_3 + L_4.
\end{aligned}$$

For L_1, since $\mathbf{x}$ and $\mathbf{z}$ are independent, for $i, q = 1, \ldots, d$, by (3.29) and Davydov inequality, under Condition 1, it follows that,

$$\begin{aligned}
&\mathrm{E}\left\{\left[\frac{\sum_{t=1}^{n-l}\left(x_{i,t}x_{q,t+l} - \mathrm{E}(x_{i,t}x_{q,t+l})\right) I(z_t = k, z_{t+l} = j)}{\sum_{t=1}^{n-l} I(z_t = k)}\right]^2\right\} \\
&\leq \frac{C}{n^2}\mathrm{E}\left\{\left[\sum_{t=1}^{n-l}\left(x_{i,t}x_{q,t+l} - \mathrm{E}(x_{i,t}x_{q,t+l})\right) I(z_t = k, z_{t+l} = j)\right]^2\right\} \\
&\leq \frac{C}{n^2}\sum_{|t_1 - t_2| > l}\left|\mathrm{E}\left\{\left[x_{i,t_1}x_{q,t_1+l} - \mathrm{E}(x_{i,t_1}x_{q,t_1+l})\right]\cdot\left[x_{i,t_2}x_{q,t_2+l} - \mathrm{E}(x_{i,t_2}x_{q,t_2+l})\right]\right\}\right| \\
&\quad + \frac{C}{n^2}\sum_{|t_1 - t_2| \leq l}\left|\mathrm{E}\left\{\left[x_{i,t_1},x_{q,t_1+l} - \mathrm{E}(x_{i,t_1}x_{q,t_1+l})\right]\cdot\left[x_{i,t_2}x_{q,t_2+l} - \mathrm{E}(x_{i,t_2}x_{q,t_2+l})\right]\right\}\right| \\
&\leq \frac{C}{n^2}\sum_{t_1 \neq t_2}\alpha(|t_1 - t_2|)^{1-2/\gamma} + \frac{C}{n} = O(1/n).
\end{aligned}$$

Then $\mathrm{E}\left(\|L_1\|_F^2\right) = O(1/n)$. Since $\|L_1\|_2 \leq \|L_1\|_F \leq \sqrt{d}\|L_1\|_2$, it follows that $\|L_1\|_2 = O_p(n^{-1/2})$.

For L_2, under Conditions 1 and 2, by (3.29) and Davydov inequality, we have

$$\begin{aligned}
\mathrm{E}\|\bar{\mathbf{x}}_j\|_2^2 &= \sum_{q=1}^{d}\mathrm{E}\left(\frac{\sum_{t=1}^{n}x_{q,t}I(z_t = j)}{\sum_{t=1}^{n}I(z_t = j)}\right)^2 \\
&\leq \frac{C}{n^2}\sum_{q=1}^{d}\left(\sum_{t=1}^{n}\mathrm{E}(x_{q,t}^2) + \sum_{t_1 \neq t_2}^{n}\left|\mathrm{Cov}(x_{q,t_1}, x_{q,t_2})\right|\right) = O(1/n), \qquad (3.31)
\end{aligned}$$

and

$$\begin{aligned}
\mathrm{E}\left\|\frac{\sum_{t=1}^{n-l}\mathbf{x}_t I(z_t = k, z_{t+l} = j)}{\sum_{t=1}^{n-l}I(z_t = k)}\right\|_2^2 &= \sum_{q=1}^{d}\mathrm{E}\left(\frac{\sum_{t=1}^{n-l}x_{q,t}I(z_t = k, z_{t+l} = j)}{\sum_{t=1}^{n-l}I(z_t = k)}\right)^2 \\
&\leq \frac{C}{n^2}\sum_{q=1}^{d}\left(\sum_{t=1}^{n-l}\mathrm{E}(x_{q,t}^2) + \sum_{t_1 \neq t_2}^{n-l}\left|\mathrm{Cov}(x_{q,t_1}, x_{q,t_2})\right|\right) = O(1/n). \qquad (3.32)
\end{aligned}$$

Hence,

$$\|L_2\|_2 \leq \|\bar{\mathbf{x}}_j\|_2 \cdot \left\|\frac{\sum_{t=1}^{n-l}\mathbf{x}_t I(z_t = k, z_{t+l} = j)}{\sum_{t=1}^{n-l}I(z_t = k)}\right\|_2 = O_p(1/n).$$

Similarly $\|L_3\|_2 = O_p(1/n)$ and $\|L_4\|_2 = O_p(1/n^2)$.

For $i, q = 1, \ldots, d$, since the second moment of $\mathbf{x}_t$ is bounded under Condition 2, together with (3.30),

$$\mathrm{E}\left\{\sum_{t=1}^{n-l}\left[\left(\frac{I(z_t=k, z_{t+l}=j)}{\sum_{t=1}^{n-l} I(z_t=k)} - \frac{\pi_{k,j}^{(l)}}{n-l}\right)\mathrm{E}(x_{i,t}x_{q,t+l})\right]\right\}^2 \leq \frac{C}{n}.$$

(3.28) follows by combining the above results. ∎

Lemma 3.2. *Under Conditions 1-5, if $\pi_{k,j}^{(l)} > 0$, we have, for $k, j = 1, \ldots, m$.*

$$\|\widehat{\mathbf{\Sigma}}_{f,k,j}(l) - \mathbf{\Sigma}_{f,k,j}(l)\|_2 = O_p(p^{1-\delta_k/2-\delta_j/2}n^{-1/2}), \tag{3.33}$$

$$\|\widehat{\mathbf{\Sigma}}_{f,\varepsilon,k,j}(l)\|_2 = O_p(p^{1-\delta_k/2}n^{-1/2}), \tag{3.34}$$

$$\|\widehat{\mathbf{\Sigma}}_{\varepsilon,f,k,j}(l)\|_2 = O_p(p^{1-\delta_j/2}n^{-1/2}). \tag{3.35}$$

Proof:

$$\begin{aligned}
&\widehat{\mathbf{\Sigma}}_{f,k,j}(l) - \mathbf{\Sigma}_{f,k,j}(l) \\
&= \frac{\sum_{t=1}^{n-l}\mathbf{A}_k(\mathbf{x}_t - \bar{\mathbf{x}}_k)(\mathbf{x}_{t+l} - \bar{\mathbf{x}}_j)'\mathbf{A}_j' I(z_t=k, z_{t+l}=j)}{\sum_{t=1}^{n-l} I(z_t=k)} - \pi_{k,j}^{(l)}\mathbf{A}_k\mathbf{\Sigma}_{x,k,j}(l)\mathbf{A}_j' \\
&= \frac{\mathbf{A}_k\sum_{t=1}^{n-l}\left(\widehat{\mathbf{\Sigma}}_{x,k,j}(l) - \mathbf{\Sigma}_{x,k,j}(l)\right)\mathbf{A}_j'}{\sum_{t=1}^{n-l} I(z_t=k)} \\
&\quad + \mathbf{A}_k\mathbf{\Sigma}_{x,k,j}(l)\mathbf{A}_j'\left(\frac{\sum_{t=1}^{n} I(z_t=k, z_{t+l}=j)}{\sum_{t=1}^{n-l} I(z_t=k)} - \pi_{k,j}^{(l)}\right).
\end{aligned}$$

Hence, under Conditions 2 and 4, by Lemma 3.1 and (3.30), for $k = 1, \ldots, m$,

$$\begin{aligned}
\|\widehat{\mathbf{\Sigma}}_{f,k,j}(l) - \mathbf{\Sigma}_{f,k,j}(l)\|_2 &\leq \|\mathbf{A}_k\|_2 \cdot \|\widehat{\mathbf{\Sigma}}_{x,k,j}(l) - \mathbf{\Sigma}_{x,j,k}(l)\|_2 \cdot \|\mathbf{A}_j\|_2 \\
&\quad + \|\mathbf{A}_k\|_2 \cdot \|\mathbf{\Sigma}_{x,k,j}(l)\|_2 \cdot \|\mathbf{A}_j\|_2 \cdot O(n^{-1/2}) \\
&= O_p(p^{1-\delta_k/2-\delta_j/2}n^{-1/2}).
\end{aligned}$$

For (3.34), we expand $\widehat{\mathbf{\Sigma}}_{f,\varepsilon,k,j}(l)$,

$$\begin{aligned}
\widehat{\mathbf{\Sigma}}_{f,\varepsilon,k,j}(l) &= \frac{\mathbf{A}_k \sum_{j=1}^{m}\sum_{t=1}^{n-l}(\mathbf{x}_t-\bar{\mathbf{x}}_k)(\boldsymbol{\varepsilon}_{t+l}^{(j)}-\bar{\boldsymbol{\varepsilon}}_j)' I(z_t=k,z_{t+l}=j)}{\sum_{t=1}^{n-l} I(z_t=k)} \\
&= \frac{\mathbf{A}_k \sum_{j=1}^{m}\sum_{t=1}^{n-l}\mathbf{x}_t\boldsymbol{\varepsilon}_{t+l}^{(j)\prime} I(z_t=k,z_{t+l}=j)}{\sum_{t=1}^{n-l} I(z_t=k)} \\
&\quad -\frac{\mathbf{A}_k \sum_{j=1}^{m}\sum_{t=1}^{n-l}\mathbf{x}_t\bar{\boldsymbol{\varepsilon}}_j' I(z_t=k,z_{t+l}=j)}{\sum_{t=1}^{n-l} I(z_t=k)} \\
&\quad -\frac{\mathbf{A}_k \sum_{j=1}^{m}\sum_{t=1}^{n-l}\bar{\mathbf{x}}_k\boldsymbol{\varepsilon}_{t+l}^{(j)\prime} I(z_t=k,z_{t+l}=j)}{\sum_{t=1}^{n-l} I(z_t=k)} \\
&\quad +\frac{\mathbf{A}_k \sum_{j=1}^{m}\sum_{t=1}^{n-l}\bar{\mathbf{x}}_k\bar{\boldsymbol{\varepsilon}}_j' I(z_t=k,z_{t+l}=j)}{\sum_{t=1}^{n-l} I(z_t=k)} \\
&= L_1+L_2+L_3+L_4.
\end{aligned}$$

For L_1, since $\mathbf{x}$, $\boldsymbol{\varepsilon}^{(j)}$ and $\mathbf{z}$ are independent, $j=1,\ldots,m$, under Conditions 1-3, for $i=1,\ldots,d$, $q=1,\ldots,p$,

$$\begin{aligned}
&\mathrm{E}\left[\left(\frac{\sum_{t=1}^{n-l} x_{i,t}\varepsilon_{q,t+l}^{(j)} I(z_t=k,z_{t+l}=j)}{\sum_{t=1}^{n-l} I(z_t=k)}\right)^2\right] \\
&\le \frac{C}{n^2}\mathrm{E}\left[\left(\sum_{t=1}^{n-l} x_{i,t}\varepsilon_{q,t+l}^{(j)} I(z_t=k,z_{t+l}=j)\right)^2\right] \\
&\le \frac{C}{n^2}\sum_{t=1}^{n-l}\mathrm{E}\left[x_{i,t}^2\left(\varepsilon_{q,t+l}^{(j)}\right)^2\right] + \sum_{t_1\neq t_2}^{n-l}\left|\mathrm{Cov}\left(x_{i,t_1}\varepsilon_{q,t_1+l}^{(j)}, x_{i,t_2}\varepsilon_{q,t_2+l}^{(j)}\right)\right| = O(1/n).
\end{aligned}$$

So $\mathrm{E}\left\|\sum_{j=1}^{m}\sum_{t=1}^{n-l}\mathbf{x}_t\boldsymbol{\varepsilon}_{t+l}^{(j)\prime} I(z_t=k,z_{t+l}=j)/\sum_{t=1}^{n-l} I(z_t=k)\right\|_F^2 = O(pn^{-1})$. Under Condition 4, we have

$$\begin{aligned}
\|L_1\|_2 &\le \|\mathbf{A}_k\|_2 \cdot \left\|\frac{\sum_{j=1}^{m}\sum_{t=1}^{n}\mathbf{x}_t\boldsymbol{\varepsilon}_{t+l}^{(j)\prime} I(z_t=k,z_{t+l}=j)}{\sum_{t=1}^{n} I(z_t=k)}\right\|_F \\
&= O(p^{1/2-\delta_k/2})O_p(p^{1/2}n^{-1/2}) = O_p(p^{1-\delta_k/2}n^{-1/2}).
\end{aligned}$$

For L_2, with (3.29) and independence of $\mathbf{z}$ and $\boldsymbol{\varepsilon}_t^{(j)}$, under Condition 3,

$$\begin{aligned}
\mathrm{E}\|\bar{\boldsymbol{\varepsilon}}_j\|_2^2 &= \sum_{q=1}^{p} \mathrm{E}\left(\frac{\sum_{t=1}^{n} \boldsymbol{\varepsilon}_{q,t}^{(j)} I(z_t = j)}{\sum_{t=1}^{n} I(z_t = j)}\right)^2 \leq \sum_{q=1}^{p} \frac{C}{n^2} \mathrm{E}\left[\left(\sum_{t=1}^{n} \varepsilon_{q,t}^{(j)} I(z_t = j)\right)^2\right] \\
&\leq \sum_{q=1}^{p} \frac{C}{n^2}\left[\sum_{t=1}^{n} \mathrm{E}\left(\varepsilon_{q,t}^{(j)}\right)^2 + \sum_{t_1 \neq t_2} \left|\mathrm{E}(\varepsilon_{q,t_1}\varepsilon_{q,t_2})\right|\right] \leq \frac{Cp}{n} = O(pn^{-1}),
\end{aligned}$$

and with (3.31) and (3.32), under Condition 4, we have $\|L_2\|_2 = O_p(p^{1-\delta_k/2}n^{-1})$, $\|L_3\|_2 = O_p(p^{1-\delta_k/2}n^{-1})$, and $\|L_4\|_2 = O_p(p^{1-\delta_k/2}n^{-2})$. Hence (3.34) follows. Similar to the proof of (3.34), we can prove (3.35). ∎

Lemma 3.3. *Under Conditions 4-7,*

$$\lambda_{\min}(\mathbf{M}_k) = O(p^{2-\delta_k-\delta_{\min}}), \text{ for } k = 1, \ldots, m. \tag{3.36}$$

where $\mathbf{M}_k$ is defined in (3.9) and $\lambda_{\min}(\mathbf{M})_k$ is the minimum eigenvalue of $\mathbf{M}_k$.

Proof: Let $\sigma_{\max}(\mathbf{H})$ and $\sigma_{\min}(\mathbf{H})$ denote the maximum and minimum singular value of $\mathbf{H}$. Under Condition 6, using the inequality about the singular values in Merikoski and Kumar (2004) we can prove that $\sigma_{\min}(\mathbf{A}_k\boldsymbol{\Sigma}_x(l_k)\sum_{j\in\mathcal{C}}\pi_{k,j}^{(l_k)}\mathbf{A}_j') = O(p^{1-\delta_k/2-\delta_{\min}/2})$. Using the fact that $\sigma_{\max}(\mathbf{A}_k\boldsymbol{\Sigma}_x(l_k)\sum_{j\notin\mathcal{C}}\pi_{k,j}^{(l_k)}\mathbf{A}_j') = o(p^{1-\delta_k/2-\delta_{\min}/2})$ under Conditions 4 and 6, we have

$$\begin{aligned}
&\sigma_{\min}\left(\mathbf{A}_k\boldsymbol{\Sigma}_x(l_k)\sum_{j=1}^{m}\pi_{k,j}^{(l_k)}\mathbf{A}_j'\right) \\
&\geq \sigma_{\min}\left(\mathbf{A}_k\boldsymbol{\Sigma}_x(l_k)\sum_{j\in\mathcal{C}}\pi_{k,j}^{(l_k)}\mathbf{A}_j'\right) - \sigma_{\max}\left(\mathbf{A}_k\boldsymbol{\Sigma}_x(l_k)\sum_{j\notin\mathcal{C}}\pi_{k,j}^{(l_k)}\mathbf{A}_j'\right) \\
&= O(p^{1-\delta_k/2-\delta_{\min}/2}).
\end{aligned} \tag{3.37}$$

It follows that

$$\lambda_{\min}(\mathbf{M}_k) \geq \max_{1\leq l\leq l_0} \sigma_{\min}^2\left(\mathbf{A}_k\boldsymbol{\Sigma}_x(l)\sum_{j=1}^{m}\pi_{k,j}^{(l)}\mathbf{A}_j'\right) = O(p^{2-\delta_k-\delta_{\min}}).$$

∎

Lemma 3.4. *Under Conditions 1-7,*

$$\|\widehat{\mathbf{M}}_k - \mathbf{M}_k\|_2 = O_p(p^{2-\delta_k/2-\delta_{\min}/2}\, n^{-1/2}),\ \textit{for } k = 1, \ldots, m.$$

Proof:

$$\|\widehat{\mathbf{M}}_k - \mathbf{M}_k\|_2 \leq \sum_{l=1}^{l_0} \left(\|\widehat{\mathbf{\Sigma}}_{y,k}(l) - \mathbf{\Sigma}_{y,k}(l)\|_2^2 + 2\|\mathbf{\Sigma}_{y,k}(l)\|_2 \cdot \|\widehat{\mathbf{\Sigma}}_{y,k}(l) - \mathbf{\Sigma}_{y,k}(l)\|_2 \right). \quad (3.38)$$

Conditions 5, 6 and 7 indicate that

$$\|\mathbf{\Sigma}_{y,k}(l_k)\|_2 = O_p(p^{1-\delta_k/2-\delta_{\min}/2}). \quad (3.39)$$

When $\pi_{k,j}^{(l)} > 0$, for any l,

$$\begin{aligned}
&\widehat{\mathbf{\Sigma}}_{y,k}(l) \\
&= \frac{\sum_{j=1}^m \sum_{t=1}^{n-l} (\mathbf{y}_t - \widehat{\boldsymbol{\mu}}_k)(\mathbf{y}_{t+l} - \widehat{\boldsymbol{\mu}}_j)' I(z_t = k, z_{t+l} = j)}{\sum_{t=1}^{n-l} I(z_t = k)} \\
&= \frac{\sum_{j=1}^m \sum_{t=1}^{n-l} (\mathbf{y}_t - \boldsymbol{\mu}_k - \mathbf{A}_k \bar{\mathbf{x}}_k - \bar{\boldsymbol{\varepsilon}}_k)(\mathbf{y}_{t+l} - \boldsymbol{\mu}_j - \mathbf{A}_j \bar{\mathbf{x}}_j - \bar{\boldsymbol{\varepsilon}}_j)' I(z_t = k, z_{t+l} = j)}{\sum_{t=1}^{n-l} I(z_t = k)} \\
&= \sum_{j=1}^m \left(\widehat{\mathbf{\Sigma}}_{f,k,j}(l) + \widehat{\mathbf{\Sigma}}_{\varepsilon,k,j}(l) + \widehat{\mathbf{\Sigma}}_{f,\varepsilon,k,j}(l) + \widehat{\mathbf{\Sigma}}_{\varepsilon,f,k,j}(l) \right).
\end{aligned}$$

By Lemma 3.2 we have

$$\begin{aligned}
&\|\widehat{\mathbf{\Sigma}}_{y,k}(l_k) - \mathbf{\Sigma}_{y,k}(l_k)\|_2 \\
&= \sum_{j=1}^m \left(\|\widehat{\mathbf{\Sigma}}_{f,k,j}(l_k) - \mathbf{\Sigma}_{f,k,j}(l_k)\|_2 + \|\widehat{\mathbf{\Sigma}}_{f,\varepsilon,k,j}(l_k)\|_2 + \|\widehat{\mathbf{\Sigma}}_{\varepsilon,f,k,j}(l_k)\|_2 + \|\widehat{\mathbf{\Sigma}}_{\varepsilon,k,j}(l_k)\|_2 \right) \\
&= O_p(p^{1-\delta_k/2-\delta_{\min}/2} n^{-1/2} + p^{1-\delta_k/2} n^{-1/2} + p^{1-\delta_{\min}/2} n^{-1/2} + \sum_{k=1}^m \|\mathbf{\Sigma}_{\varepsilon,k,j}(l_k)\|_2) \quad (3.40)
\end{aligned}$$

Since $\boldsymbol{\varepsilon}_t^{(k)}$ are independent noises, we have $\|\widehat{\mathbf{\Sigma}}_{\varepsilon,k,j}(l_k)\|_2 \leq \|\widehat{\mathbf{\Sigma}}_{\varepsilon,k,j}(l_k)\|_F = O_p(pn^{-1/2})$, which implies from (3.40) that

$$\|\widehat{\mathbf{\Sigma}}_{y,k}(l_k) - \mathbf{\Sigma}_{y,k}(l_k)\|_2 = O_p(pn^{-1/2}). \quad (3.41)$$

Together with (3.38), (3.39) and (3.41), the lemma follows. ∎

Proof of Theorem 3.1: By Lemmas 3.1-3.4, and Lemma 3 in Lam et al. (2011), we can easily reach the conclusion of Theorem 1. ∎

Proof of Theorem 3.2: From (3.13), when $z_t = k$,

$$\begin{aligned}
\widehat{\mathbf{f}}_t - \mathbf{f}_t &= \widehat{\mathbf{Q}}_k\widehat{\mathbf{R}}_t - \mathbf{Q}_k\mathbf{R}_t = \widehat{\mathbf{Q}}_k\widehat{\mathbf{Q}}_k'(\mathbf{y}_t - \widehat{\boldsymbol{\mu}}_k) - \mathbf{Q}_k\mathbf{R}_t \\
&= \widehat{\mathbf{Q}}_k\widehat{\mathbf{Q}}_k'(\mathbf{Q}_k\mathbf{R}_t + \boldsymbol{\varepsilon}_t^{(k)} + \boldsymbol{\mu}_k - \widehat{\boldsymbol{\mu}}_k) - \mathbf{Q}_k\mathbf{R}_t \\
&= (\widehat{\mathbf{Q}}_k\widehat{\mathbf{Q}}_k' - \mathbf{Q}_k\mathbf{Q}_k')\mathbf{Q}_k\mathbf{R}_t + \widehat{\mathbf{Q}}_k(\widehat{\mathbf{Q}}_k - \mathbf{Q}_k)'(\boldsymbol{\varepsilon}_t^{(k)} + \boldsymbol{\mu}_k - \widehat{\boldsymbol{\mu}}_k) \\
&\quad + \widehat{\mathbf{Q}}_k\mathbf{Q}_k'(\boldsymbol{\varepsilon}_t^{(k)} + \boldsymbol{\mu}_k - \widehat{\boldsymbol{\mu}}_k) \\
&= I_1 + I_2 + I_3.
\end{aligned}$$

Note that when $z_t = k$, $\|\mathbf{R}_t\|_2 = \|\mathbf{A}_k\|_2 = O(p^{1/2-\delta_k/2})$ defined in (3.3), so $\|I_1\|_2 \leq 2\|\widehat{\mathbf{Q}}_k - \mathbf{Q}_k\|_2\|\mathbf{R}_t\|_2 = O_p(p^{1/2-\delta_k/2}\|\widehat{\mathbf{Q}}_k - \mathbf{Q}_k\|_2) = O_p(p^{1/2+\delta_{\min}/2}n^{-1/2})$. I_2 is dominated by I_3 in probability.

$$\mathrm{E}(\|\widehat{\mathbf{Q}}_k\mathbf{Q}_k'\boldsymbol{\varepsilon}_t^{(k)}\|_2^2) = \sum_{i=1}^{d}\mathrm{E}[(\mathbf{q}_i'\boldsymbol{\varepsilon}_t^{(k)})^2] \leq d\lambda_{max}(\boldsymbol{\Sigma}_k) < \infty. \tag{3.42}$$

$$\widehat{\boldsymbol{\mu}}_k - \boldsymbol{\mu}_k = \frac{\sum_{t=1}^n \mathbf{y}_t I(z_t = k)}{\sum_{t=1}^n I(z_t = k)} - \boldsymbol{\mu}_k = \frac{\sum_{t=1}^n (\mathbf{A}_k\mathbf{x}_t + \boldsymbol{\varepsilon}_t^{(k)})I(z_t = k)}{\sum_{t=1}^n I(z_t = k)}.$$

By (3.42), we can easily have

$$\Big\|\frac{\sum_{t=1}^n (\widehat{\mathbf{Q}}_k\mathbf{Q}_k'\boldsymbol{\varepsilon}_t^{(k)} I(z_t = k)}{\sum_{t=1}^n I(z_t = k)}\Big\|_2 = O_p(n^{-1/2}).$$

Under Condition 4, $\|\widehat{\mathbf{Q}}_k\mathbf{Q}_k'(\widehat{\boldsymbol{\mu}}_k - \boldsymbol{\mu}_k)\|_2 = O_p(p^{1/2-\delta_k/2}n^{-1/2}) + O_p(n^{-1/2})$. Hence, with (3.42), $\|I_3\|_2 = O_p(p^{1/2-\delta_k/2}n^{-1/2}) + O_p(1)$.

We have $p^{-1/2}\|\widehat{\mathbf{f}}_t - \mathbf{f}_t\|_2 = O_p(p^{\delta_{\min}/2}n^{-1/2} + p^{-1/2})$. ∎

Proof of Theorem 3.3: We assume that $\mathbf{Q}_k$ is uniquely defined as in Remark 3.10

under Condition 7. Then

$$\text{tr}\left[\mathbf{Q}_k'(\mathbf{I}_p - \widehat{\mathbf{Q}}_k\widehat{\mathbf{Q}}_k')\mathbf{Q}_k\right] = \text{tr}(\mathbf{I}_d - \mathbf{Q}_k'\widehat{\mathbf{Q}}_k\widehat{\mathbf{Q}}_k'\mathbf{Q}_k) = d\left[\mathcal{D}(\mathcal{M}(\widehat{\mathbf{Q}}_k), \mathcal{M}(\mathbf{Q}_k))\right]^2. \quad (3.43)$$

On the other hand,

$$\begin{aligned}\text{tr}\left[\mathbf{Q}_k'(\mathbf{I}_p - \widehat{\mathbf{Q}}_k\widehat{\mathbf{Q}}_k')\mathbf{Q}_k\right] - \text{tr}\left[\mathbf{Q}_k'(\mathbf{I}_p - \mathbf{Q}_k\mathbf{Q}_k')\mathbf{Q}_k\right] &= \text{tr}\left[\mathbf{Q}_k'(\mathbf{Q}_k\mathbf{Q}_k' - \widehat{\mathbf{Q}}_k\widehat{\mathbf{Q}}_k')\mathbf{Q}_k\right] \\ &\leq d\|\mathbf{Q}_k'(\mathbf{Q}_k\mathbf{Q}_k' - \widehat{\mathbf{Q}}_k\widehat{\mathbf{Q}}_k')\mathbf{Q}_k\|_2.\end{aligned}$$

And since the diagonal entries in $\mathbf{Q}_k'\widehat{\mathbf{Q}}_k\widehat{\mathbf{Q}}_k'\mathbf{Q}_k$ are between 0 and 1,

$$\begin{aligned}\text{tr}\left[\mathbf{Q}_k'(\mathbf{I}_p - \widehat{\mathbf{Q}}_k\widehat{\mathbf{Q}}_k')\mathbf{Q}_k\right] - \text{tr}\left[\mathbf{Q}_k'(\mathbf{I}_p - \mathbf{Q}_k\mathbf{Q}_k')\mathbf{Q}_k\right] &= \text{tr}\left[\mathbf{Q}_k'(\mathbf{Q}_k\mathbf{Q}_k' - \widehat{\mathbf{Q}}_k\widehat{\mathbf{Q}}_k')\mathbf{Q}_k\right] \\ &\geq \|\mathbf{Q}_k'(\mathbf{Q}_k\mathbf{Q}_k' - \widehat{\mathbf{Q}}_k\widehat{\mathbf{Q}}_k')\mathbf{Q}_k\|_2.\end{aligned}$$

Note that $\text{tr}\left[\mathbf{Q}_k'(\mathbf{I}_p - \mathbf{Q}_k\mathbf{Q}_k')\mathbf{Q}_k\right] = 0$. Hence,

$$\left[\mathcal{D}(\mathcal{M}(\widehat{\mathbf{Q}}_k), \mathcal{M}(\mathbf{Q}_k))\right]^2 \asymp \|\mathbf{Q}_k'(\mathbf{Q}_k\mathbf{Q}_k' - \widehat{\mathbf{Q}}_k\widehat{\mathbf{Q}}_k')\mathbf{Q}_k\|_2.$$

Since

$$\mathbf{Q}_k'(\mathbf{Q}_k\mathbf{Q}_k' - \widehat{\mathbf{Q}}_k\widehat{\mathbf{Q}}_k')\mathbf{Q}_k = -\mathbf{Q}_k'(\mathbf{Q}_k - \widehat{\mathbf{Q}}_k)(\mathbf{Q}_k - \widehat{\mathbf{Q}}_k)'\mathbf{Q}_k + (\mathbf{Q}_k - \widehat{\mathbf{Q}}_k)'(\mathbf{Q}_k - \widehat{\mathbf{Q}}_k),$$

which is bounded by $2\|\widehat{\mathbf{Q}}_k - \mathbf{Q}\|_2^2$, with (3.43) we have

$$\mathcal{D}(\widehat{\mathbf{Q}}_k, \mathbf{Q}_k) = O_p(\|\widehat{\mathbf{Q}}_k - \mathbf{Q}_k\|_2).$$

By Theorem 3.1, we have proved Theorem 3.3. ∎

Proof of Theorem 3.4: The proof is quite similar to that of Theorem 1 of Lam and Yao (2012). We denote $\lambda_{k,j}$ and $\widehat{\mathbf{q}}_{k,j}$ for the j-th largest eigenvalues of $\widehat{\mathbf{M}}_k$ and its corresponding orthonormal eigenvectors, respectively, for $k = 1, \ldots, m$. The corresponding population values are denoted by $\lambda_{k,j}$ and $\mathbf{q}_{k,j}$ for the matrix $\mathbf{M}_k$. Let

$\widehat{\mathbf{Q}}_k = (\widehat{\mathbf{q}}_{k,1}, \ldots, \widehat{\mathbf{q}}_{k,d})$ and $\mathbf{Q}_k = (\mathbf{q}_{k,1}, \ldots, \mathbf{q}_{k,d})$. We have

$$\lambda_{k,j} = \mathbf{q}_{k,j}' \mathbf{M}_k \mathbf{q}_{k,j}, \quad \text{and} \qquad \widehat{\lambda}_{k,j} = \widehat{\mathbf{q}}_{k,j}' \widehat{\mathbf{M}}_k \widehat{\mathbf{q}}_{k,j}, \qquad j = 1, \ldots, p.$$

We can decompose $\widehat{\lambda}_{k,j} - \lambda_{k,j}$ by

$$\widehat{\lambda}_{k,j} - \lambda_{k,j} = \widehat{\mathbf{q}}_{k,j}' \widehat{\mathbf{M}}_k \widehat{\mathbf{q}}_{k,j} - \mathbf{q}_{k,j}' \mathbf{M}_k \mathbf{q}_{k,j} = I_1 + I_2 + I_3 + I_4 + I_5,$$

where

$$I_1 = (\widehat{\mathbf{q}}_{k,j} - \mathbf{q}_{k,j})'(\widehat{\mathbf{M}}_k - \mathbf{M}_k)\widehat{\mathbf{q}}_{k,j}, \quad I_2 = (\widehat{\mathbf{q}}_{k,j} - \mathbf{q}_{k,j})'\mathbf{M}(\widehat{\mathbf{q}}_{k,j} - \mathbf{q}_{k,j}),$$

$$I_3 = (\widehat{\mathbf{q}}_{k,j} - \mathbf{q}_{k,j})'\mathbf{M}_k \mathbf{q}_{k,j}, \quad I_4 = \mathbf{q}_{k,j}'(\widehat{\mathbf{M}}_k - \mathbf{M}_k)\widehat{\mathbf{q}}_{k,j}, \quad I_5 = \mathbf{q}_{k,j}'\mathbf{M}_k(\widehat{\mathbf{q}}_{k,j} - \mathbf{q}_{k,j}).$$

For $j = 1, \ldots, d$, $\|\widehat{\mathbf{q}}_{k,j} - \mathbf{q}_{k,j}\|_2 \leq \|\widehat{\mathbf{Q}}_k - \mathbf{Q}_k\|_2 = O_p(h_{n,k})$, where $h_{n,k} = p^{\delta_k/2 + \delta_{\min}/2} n^{-1/2}$ by Theorem 3.1, and $\|\mathbf{M}_k\|_2 \leq \sum_{l=1}^{l_0} \|\mathbf{\Sigma}_y(l)\|^2 = O_p(p^{2-\delta_k-\delta_{\min}})$. By Lemma 3.2 and Lemma 3.4, we have $\|I_1\|_2$ and $\|I_2\|_2$ are of order $O_p(p^{2-\delta_k-\delta_{\min}} h_{n,k}^2)$ and $\|I_3\|_2$, $\|I_4\|_2$ and $\|I_5\|_2$ are of order $O_p(p^{2-\delta_k-\delta_{\min}} h_{n,k})$. So $|\widehat{\lambda}_{k,j} - \lambda_{k,j}| = O_p(p^{2-\delta_k-\delta\min} h_{n,k}) = O_p(p^{2-\delta_k/2-\delta_{\min}/2} n^{-1/2})$.

For $j = d+1, \ldots, p$, define,

$$\widetilde{\mathbf{M}}_k = \sum_{l=1}^{l_0} \widehat{\mathbf{\Sigma}}_{y,k}(l) \mathbf{\Sigma}_{y,k}(l)', \quad \widehat{\mathbf{B}}_k = (\widehat{\mathbf{q}}_{k,d+1}, \ldots, \widehat{\mathbf{q}}_{k,p}), \quad \text{and} \quad \mathbf{B}_k = (\mathbf{q}_{k,d+1}, \ldots, \mathbf{q}_{k,p}).$$

It can be shown that $\|\widehat{\mathbf{B}}_k - \mathbf{B}_k\|_2 = O_p(h_{n,k})$, similar to proof of Theorem 3.1 with Lemma 3 in Lam et al. (2011). Hence, $\|\widehat{\mathbf{q}}_{k,j} - \mathbf{q}_{k,j}\|_2 \leq \|\widehat{\mathbf{B}}_k - \mathbf{B}_k\|_2 = O_p(h_{n,k})$.

Since $\lambda_j = 0$, for $j = d+1, \ldots, p$, consider the decomposition

$$\widehat{\lambda}_j = \widehat{\mathbf{q}}_{k,j}' \widehat{\mathbf{M}}_k \widehat{\mathbf{q}}_{k,j} = K_1 + K_2 + K_3,$$

where

$$K_1 = \widehat{\mathbf{q}}_{k,j}'(\widehat{\mathbf{M}}_k - \widetilde{\mathbf{M}}_k - \widetilde{\mathbf{M}}_k' + \mathbf{M}_k)'\widehat{\mathbf{q}}_{k,j}, \quad K_2 = 2\widehat{\mathbf{q}}_{k,j}'(\widetilde{\mathbf{M}}_k - \mathbf{M}_k)(\widehat{\mathbf{q}}_{k,j} - \mathbf{q}_{k,j}),$$

$$K_3 = (\widehat{\mathbf{q}}_{k,j} - \mathbf{q}_{k,j})'\mathbf{M}_k(\widehat{\mathbf{q}}'_{kj} - \mathbf{q}_{k,j}).$$

By Lemma 3.2 and Lemma 3.4,

$$\begin{aligned}
K_1 &= \sum_{l=1}^{l_0} \|(\widehat{\mathbf{\Sigma}}_{y,k}(l) - \mathbf{\Sigma}_{y,k}(l))\widehat{\mathbf{q}}_{k,j}\|_2^2 \leq \sum_{l=1}^{l_0} \|\widehat{\mathbf{\Sigma}}_{y,k}(l) - \mathbf{\Sigma}_{y,k}(l)\|_2^2 = O_p(p^2 n^{-1}), \\
|K_2| &= O_p(\|\widetilde{\mathbf{M}}_k - \mathbf{M}_k\|_2 \cdot \|\widehat{\mathbf{q}}_{k,j} - \mathbf{q}_{k,j}\|_2) = O_p(\|\widetilde{\mathbf{M}}_k - \mathbf{M}_k\|_2 \cdot \|\widehat{\mathbf{B}}_k - \mathbf{B}_k\|_2) \\
&= O_p(p^2 n^{-1}), \\
|K_3| &= O_p(\|\widehat{\mathbf{B}}_k - \mathbf{B}_k\|_2^2 \cdot \|\mathbf{M}_k\|_2) = O_p(p^{2-\delta_k-\delta_{\min}} h_n^2) = O_p(p^2 n^{-1}).
\end{aligned}$$

Hence $\lambda_{k,j} = O_p(p^2 n^{-1})$. ∎

Proof of Corollary 3.1: The proof is similar to the proof of Corollary 1 of Lam and Yao (2012).

By Lemma 3.3 and Lemma 3.4, we have

$$\lambda_{k,1} = \|\mathbf{M}_k\|_2 = O(p^{2-\delta_k-\delta_{\min}}) \quad \text{and} \quad \lambda_{k,d} = O(p^{2-\delta_k-\delta_{\min}}).$$

So we have $\lambda_{k,i} \asymp p^{2-\delta_k-\delta_{\min}}$, for $i = 1, \ldots, d$. From Theorem 3.4(i), we have $|\widehat{\lambda}_{k,i} - \lambda_{k,i}| = O_p(p^{2-\delta_k-\delta_{\min}} n^{-1/2})$, then $\widehat{\lambda}_{k,i} = O_p(p^{2-\delta_k-\delta_{\min}})$ for $i = 1, \ldots, d$. It implies that $\widehat{\lambda}_{k,i+1}/\widehat{\lambda}_{k,i} \asymp 1$ for $i = 1, \ldots, d-1$. By Theorem 3.4(ii),

$$\widehat{\lambda}_{k,d+1}/\widehat{\lambda}_{k,d} = O_p(p^2 n^{-1}/p^{2-\delta_k-\delta_{\min}}) = O_p(p^{\delta_k+\delta_{\min}} n^{-1}).$$

∎

Proof of Theorem 3.5:

$$\begin{aligned}
w_{t,k,j} &= \frac{1}{2}(\log|\mathbf{\Sigma}_{B,j}| - \log|\mathbf{\Sigma}_{B,k}|) + \frac{1}{2}\boldsymbol{\varepsilon}_t^{(k)\prime}(\mathbf{B}_j{\mathbf{\Sigma}_{B,j}}^{-1}\mathbf{B}'_j - \mathbf{B}_k{\mathbf{\Sigma}_{B,k}}^{-1}\mathbf{B}'_k)\boldsymbol{\varepsilon}_t^{(k)} \\
&\quad + \frac{1}{2}\left(\mathbf{B}'_j\mathbf{A}_k\mathbf{x}_t + \mathbf{B}'_j(\boldsymbol{\mu}_k - \boldsymbol{\mu}_j)\right)'{\mathbf{\Sigma}_{B,j}}^{-1}\left(\mathbf{B}'_j\mathbf{A}_k\mathbf{x}_t + \mathbf{B}'_j(\boldsymbol{\mu}_k - \boldsymbol{\mu}_j)\right) \\
&\quad + (\mathbf{B}'_j(\mathbf{A}_k\mathbf{x}_t + \boldsymbol{\mu}_k - \boldsymbol{\mu}_j))'\mathbf{\Sigma}_{B,j}^{-1}\mathbf{B}'_j\boldsymbol{\varepsilon}_t^{(k)} \\
&= L_1 + L_2 + L_3 + L_4. \qquad (3.44)
\end{aligned}$$

We have

$$\begin{aligned}\mathrm{E}(L_2) &= \frac{1}{2}\mathrm{tr}\left[\boldsymbol{\Sigma}_k^{1/2}\left(\mathbf{B}_j\boldsymbol{\Sigma}_{B,j}^{-1}\mathbf{B}_j' - \mathbf{B}_k\boldsymbol{\Sigma}_{B,k}^{-1}\mathbf{B}_k'\right)\boldsymbol{\Sigma}_k^{1/2}\right] \\ &= \frac{1}{2}\mathrm{tr}\left(\mathbf{B}_j'\boldsymbol{\Sigma}_k\mathbf{B}_j\boldsymbol{\Sigma}_{B,j}^{-1} - \mathbf{I}_{p-d}\right) = \frac{1}{2}\mathrm{tr}(\mathbf{B}_j'\boldsymbol{\Sigma}_k\mathbf{B}_j\boldsymbol{\Sigma}_{B,j}^{-1}) - \frac{(p-d)}{2},\end{aligned}$$

$\mathrm{E}\left(L_3\right) = \frac{1}{2}\mathrm{tr}\left(\mathbf{B}_j'(\boldsymbol{\Sigma}_{f,k,t} + \mathbf{U}_{k,j})\mathbf{B}_j\boldsymbol{\Sigma}_{B,j}^{-1}\right)$, and $\mathrm{E}(L_4) = \mathbf{0}$, so we obtain (3.25).

To prove (3.26) , we refer to a fact about multivariate normal random vector. Let $\mathbf{v} \sim N(0, \mathbf{I}_p)$, then for a symmetric matrix $\boldsymbol{\Sigma}$, $\mathrm{Var}(\mathbf{v}'\boldsymbol{\Sigma}\mathbf{v}) = 2\|\boldsymbol{\Sigma}\|_F^2$. Note that $\mathrm{Cov}(L_2, L_3) = \mathrm{Cov}(L_2, L_4) = \mathrm{Cov}(L_3, L_4) = 0$ in (3.24), define $\mathbf{W}_j = \mathbf{B}_j\boldsymbol{\Sigma}_{B,j}^{-1}\mathbf{B}_j$, and we have

$$\begin{aligned}\mathrm{Var}(w_{t,k,j}) &= \mathrm{Var}(L_2) + \mathrm{Var}(L_3) + \mathrm{Var}(L_4) \\ &= \frac{1}{2}\|\boldsymbol{\Sigma}_k^{1/2}(\mathbf{W}_j - \mathbf{W}_k)\boldsymbol{\Sigma}_k^{1/2}\|_F^2 + \frac{1}{4}\mathrm{Var}\left(\mathbf{x}_t'\mathbf{A}_k'\mathbf{W}_j\mathbf{A}_k\mathbf{x}_t\right) \\ &+ \frac{1}{2}\mathrm{Var}\left(\mathbf{x}_t'\mathbf{A}_k'\mathbf{W}_j(\boldsymbol{\mu}_k - \boldsymbol{\mu}_j)\right) + \mathrm{Var}\left(\mathbf{x}_t'\mathbf{A}_k'\mathbf{W}_j\boldsymbol{\varepsilon}_t^{(k)}\right) \\ &+ \mathrm{Var}\left((\boldsymbol{\mu}_k - \boldsymbol{\mu}_j)'\mathbf{W}_j\boldsymbol{\varepsilon}_t^{(k)}\right) \\ &= \frac{1}{2}\|\boldsymbol{\Sigma}_k^{1/2}(\mathbf{W}_j - \mathbf{W}_k)\boldsymbol{\Sigma}_k^{1/2}\|_F^2 + \frac{1}{2}\|\boldsymbol{\Sigma}_k^{1/2}\mathbf{A}_k'\mathbf{W}_j\mathbf{A}_k\boldsymbol{\Sigma}_k^{1/2}\|_F^2 \\ &+ \frac{1}{2}\mathrm{tr}\left(\boldsymbol{\Sigma}_{f,k,t}\mathbf{W}_j\mathbf{U}_{k,j}\mathbf{W}_j\right) + \mathrm{tr}\left((\boldsymbol{\Sigma}_{f,k,t} + \mathbf{U}_{k,j})\mathbf{W}_j\boldsymbol{\Sigma}_k\mathbf{W}_j\right).\end{aligned}$$

■

Chapter 4

Functional Coefficient Seasonal Time Series Models

In this chapter, motivated by an analysis of the monthly number of tourists visiting Hawaii, we propose a new class of nonparametric seasonal time series models under the framework of the functional coefficient model. The coefficients change over time and consist of the trend and seasonal components to characterize seasonality. A local linear approach is developed to estimate the nonparametric trend and seasonal effect functions. The proposed methodologies are illustrated by two simulated examples and the model is applied to characterizing the seasonality of the monthly number of tourists visiting Hawaii.

4.1 The Model

Denote a seasonal time series as

$$y_{t1}, \ldots, y_{td}, \qquad t = 1, 2, \ldots n, \tag{4.1}$$

where d is the number of seasons within a period and n is the number of periods. We assume that there exist p other time series $\{x_{ktj}\}$, $k = 1, \ldots, p$, and $j = 1, \ldots, d$ that are related to the time series y_{tj}, and indexed according to y_{tj}. Those time series can be the lagged series of y_{tj} (in an AR fashion), or some exogenous variables.

The proposed *functional-coefficient seasonal time series model* assumes the form as

$$y_{tj} = \sum_{k=1}^{p} [\alpha_k(t) + \beta_{kj}(t)]\, x_{ktj} + e_{tj}, \tag{4.2}$$

where $\{\alpha_k(\cdot)\}$ are the trend functions for the coefficients, and $\{\beta_{jk}(\cdot)\}$ are the seasonal

effect functions in the coefficient functions, satisfying constraints for the identification,

$$\sum_{j=1}^{d} \beta_{kj}(t) = 0, \quad \text{for each } 1 \le k \le p \text{ and all } t,$$

and the error term $\{e_{tj}\}$ is stationary and satisfies $E(e_{tj} \,|\, \mathbf{X}_{tj}) = 0$, and with $\mathbf{X}_{tj} = (X_{1tj}, \ldots, X_{ptj})'$.

Remark 4.1. There is another way to denote seasonal time series with only one subscript as

$$y_1, \ldots y_m, \ldots, y_T, \qquad m = 1, 2, \ldots, T = dn. \tag{4.3}$$

Both (4.1) and (4.3) are used in this paper exchangeably, identified by the number of subscripts. Time series denoted by the two different indexed methods satisfies the formula as $y_m = y_{tj}$, where $m = d(t-1) + j$ for $1 \le t \le n$ and $1 \le j \le d$.

Model (4.2), where coefficients combine of nonlinear trend and seasonal effect changing over time, is a generalization of the functional-coefficient time series model, a popular nonlinear time series model in the time series literature (Chen and Tsay 1993a; Xia and Li 1999a; Cai, Fan, and Yao 2000; Cai and Tiwari 2000), and the varying-coefficient model (Hastie and Tibshirani 1993, Yang, Park, Xue, and Härdle 2006) for i.i.d. samples.

This model is also motivated by the standard additive time trend and seasonal component model as

$$y_{tj} = T_t + S_{tj} + e_{tj}; \tag{4.4}$$

see Cleveland, Cleveland, McRae, and Terpenning (1990) and Cai and Chen (2006) where T_t is the common trend same to different seasons within a period, and S_{tj} is the seasonal effect, satisfying $\sum_{j=1}^{d} S_{tj} = 0$. A standard parametric model assumes a parametric function for the common trend T_t, such as linear or polynomial functions. The seasonal effects are usually assumed to be the same for different periods; that is, $S_{tj} = S_j$ for $j = 1, \ldots, d$ and all t. Note that if $p = 1$ and $x_{1tj} = 1$ for all t and j, then model (4.2) becomes

$$y_{tj} = \alpha(t) + \beta_j(t) + e_{tj}, \tag{4.5}$$

where $\{\beta_j(t)\}$ satisfy the condition $\sum_{j=1}^{d} \beta_j(t) = 0$; see Cai and Chen (2006) for details. This is the exact same as (4.4). Here, we assume nonparametric forms for both trend and seasonal component. If we further assume that $\beta_j(t) = \gamma_j \beta(t)$, then we obtained the model proposed by Burman and Shumway (1998), where $\{\gamma_j\}$ are seasonal factors. Hence, the overall seasonal effect changes over periods in accordance with the modulating function $\beta(t)$. Implicitly, this model assumes that the seasonal effect curves have the same shape (up to a multiplicative constant) for all seasons.

The AR model with trend and seasonal component is also commonly used in modeling seasonal time series (e.g., Hylleberg 1992; Franses 1996, 1998; Ghysels and Osborn 2001),

$$y_{tj} = T_t + S_{tj} + \phi y_{tj-1} + e_{tj}. \tag{4.6}$$

Our model allows both AR terms and exogenous variables entering the model in a linear fashion. The AR coefficients and the coefficients of the exogenous variables are commonly assumed to be constant over different periods. However, for seasonal time series models, it is difficult to justify that the relationships between y_t and its lag variables and exogenous variables are the same for different periods. Allowing different functions for different periods (hence seasonality) has an ability to enhance the model to adopt the nature of the underlying time series and to capture the seasonality better.

In addition, if $p = 1$ and x_t is the lag d variable of y_t, say $x_t = y_{t-d}$, or $x_{tj} = y_{(t-1)j}$, then this model assumes a pure seasonal AR model with d different series, each with seasonality of 1 as

$$y_{tj} = (\alpha(t) + \beta_j(t)) y_{(t-1)j} + e_{tj}, \;\; j = 1, \ldots, d, \tag{4.7}$$

where the coefficients change over time, with $\alpha(t)$ being the common trend and $\beta_j(t)$ being the seasonal effect, special to each season j in the period. Both trend and seasonal effect functions are nonparametric. An extreme case is that $\beta_j(t) = 0$ and $\alpha(t) = \alpha$, a constant. In this case, equation (4.7) becomes

$$y_m - \alpha y_{m-d} = e_t,$$

which is a pure seasonal AR model. Therefore, with certain combinations of the variables x_m and the coefficient functions, the proposed model in (4.2) is flexible enough to cover many existing seasonal models.

4.2 Estimation Procedure

For technical reasons, we change the time unit in the coefficient functions to $s_t = t/n$. Then, we can express (4.2) in a matrix notation,

$$\mathbf{Y}_t = \mathcal{X}_t\, \boldsymbol{\theta}(s_t) + \mathbf{e}_t,$$

where

$$\mathbf{Y}_t = \begin{pmatrix} y_{t1} \\ \vdots \\ y_{td} \end{pmatrix}, \mathcal{X}_t = \begin{pmatrix} \mathbf{X}'_{t1} & \mathbf{X}'_{t1} & \mathbf{0} & \mathbf{0} \\ \vdots & \mathbf{0} & \ddots & \mathbf{0} \\ \mathbf{X}'_{t,d-1} & \mathbf{0} & \mathbf{0} & \mathbf{X}'_{t,d-1} \\ \mathbf{X}'_{td} & -\mathbf{X}'_{td} & \dots & -\mathbf{X}'_{td} \end{pmatrix},$$

$$\mathbf{e}_t = \begin{pmatrix} e_{t1} \\ \vdots \\ e_{td} \end{pmatrix}, \boldsymbol{\theta}(s_t) = \begin{pmatrix} \boldsymbol{\alpha}(s_t) \\ \boldsymbol{\beta}_1(s_t) \\ \vdots \\ \boldsymbol{\beta}_{d-1}(s_t) \end{pmatrix},$$

with $\boldsymbol{\alpha}(s_t) = (\alpha_1(s_t) \dots \alpha_p(s_t))'$ and $\boldsymbol{\beta}_j(s_t) = (\beta_{1j}(s_t) \dots \beta_{pj}(s_t))'$. Again, the error term $\{\mathbf{e}_t\}$ is assumed to be stationary with $E(\mathbf{e}_t) = \mathbf{0}$ and $\mathrm{Var}(\mathbf{e}_t) = \boldsymbol{\Sigma}_e$.

For estimating $\alpha(\cdot)$ and $\{\beta_j(\cdot)\}$, a local linear method is employed, although a general local polynomial method is also applicable. Local linear (polynomial) methods have been widely used in nonparametric regression due to their attractive mathematical efficiency, bias reduction and adaptation of edge effects (see Fan and Gijbels 1996). We assume throughout that the trend functions $\{\alpha_k(\cdot)\}$ and the seasonal effect functions $\{\beta_{kj}(\cdot)\}$ have a continuous second derivative. Then, based on the local linear fitting

scheme of Fan and Gijbels (1996), the locally weighted least squares is given by

$$\sum_{t=1}^{n} [\mathbf{Y}_t - \mathcal{X}_t\,\boldsymbol{\theta}_0 - (s_t - s)\,\mathcal{X}_t\,\boldsymbol{\theta}_1]'\,[\mathbf{Y}_t - \mathcal{X}_t\,\boldsymbol{\theta}_0 - (s_t - s)\,\mathcal{X}_t\,\boldsymbol{\theta}_1]\;\; K_h(s_t - s), \tag{4.8}$$

where $K_h(u) = K(u/h)/h$, $K(\cdot)$ is a kernel function and h is the bandwidth satisfying $h \to 0$ and $n\,h \to \infty$ as $n \to \infty$. Let $\widehat{\boldsymbol{\theta}}_0$ and $\widehat{\boldsymbol{\theta}}_1$ be the minimizer of (4.8). Then,

$$\begin{pmatrix} \widehat{\boldsymbol{\theta}}_0 \\ \widehat{\boldsymbol{\theta}}_1 \end{pmatrix} = \begin{pmatrix} \mathbf{G}_0 & \mathbf{G}_1 \\ \mathbf{G}_1 & \mathbf{G}_2 \end{pmatrix}^{-1} \begin{pmatrix} \mathbf{M}_0 \\ \mathbf{M}_1 \end{pmatrix}, \tag{4.9}$$

where

$$\mathbf{G}_k = \frac{1}{n}\sum_{t=1}^{n} \mathcal{X}_t'\,\mathcal{X}_t\,(s_t - s)^k\,K_h(s_t - s) \quad \text{and} \quad \mathbf{M}_k = \frac{1}{n}\sum_{t=1}^{n} \mathcal{X}_t'\,\mathbf{Y}_t\,(s_t - s)^k\,K_h(s_t - s).$$

Therefore, the local linear estimates of $\boldsymbol{\theta}(s)$ and $\boldsymbol{\theta}'(s)$ (the first order derivative of $\boldsymbol{\theta}(s)$) are $\widehat{\boldsymbol{\theta}}(s) = \widehat{\boldsymbol{\theta}}_0$ and $\widehat{\boldsymbol{\theta}}'(s) = \widehat{\boldsymbol{\theta}}_1$, respectively.

Remark 4.2. Note that many other nonparametric smoothing methods can be used here. The locally weighted least square method is just one of the choices. There is a vast amount of literature in theory and empirical study on the comparison of different methods (see Fan and Gijbels 1996).

Remark 4.3. The restriction to the locally weighted least square method suggests that the normality is at least being considered as a baseline. However, when the non-normality is clearly present, a robust approach would be considered. Cai and Ould-Said (2003) considered this aspect in nonparametric regression estimation for time series.

Remark 4.4. The bandwidth selection is always one of the most important parts of any nonparametric procedure. There are several bandwidth selectors in the literature, including the leave-one-out cross validation of Härdle and Marron (1985), the generalized cross-validation of Wahba (1977), the plug-in method of Jones, Marron, and Sheather (1996), and the empirical bias method of Ruppert (1997), among others. They all can be used here. A comparison of different procedures can be found in Jones, Marron,

and Sheather (1996). In this article we use a procedure proposed in Fan, Yao, and Cai (2003), which combines the generalized cross-validation and the empirical bias method.

Remark 4.5. In the above estimation procedure, one bandwidth is used for all functions. It is possible to use different bandwidths to different seasons by using a more computational intensive two-step method (see Cai 2002). We can also incorporate the covariance structure of $\mathbf{e}_t$ in the estimation.

Remark 4.6. Since data are observed in time order as in Burman and Shumway (1998), we assume that $s_t = t/n$ for simplicity although the theoretical results developed later still hold for non-equally spaced design points.

4.3 Simulated Examples

In this section, a Monte Carlo simulation study is conducted to examine the finite sample performance of the proposed procedures. Throughout this section, we use the Epanechnikov kernel, $K(u) = 0.75\,(1-u^2)\,I(|u| \le 1)$ and the bandwidth selector mentioned in Remark 4.4. For simulated examples, the performance of the estimators is evaluated by the mean absolute deviation error (MADE):

$$\mathcal{E}_k = n_0^{-1} \sum_{j=1}^{n_0} |\widehat{\alpha}_k(v_j) - \alpha_k(v_j)| \quad \text{and} \quad \mathcal{E}_{kj} = n_0^{-1} \sum_{j=1}^{n_0} \left|\widehat{\beta}_{kj}(v_j) - \beta_{kj}(v_j)\right|$$

for $\alpha_k(\cdot)$ and $\beta_{kj}(\cdot)$, respectively, where $k = 1, \ldots, p$, $j = 1, \ldots, d$, and $\{v_j,\, j = 1, \ldots, n_0\}$ are the grid points from $(0,\ 1]$. When $p = 1$, the subscript k can be omitted. Simulation is repeated 500 times for each model with different sample sizes. For demonstration purposes, when showing results of a particular simulated series, we use the series with median total MADE value (sum of all MADE values) equals among the 500 MADE values. Such a sample is referred to as *a typical sample*.

Example 4.1. We begin with a simple additive trend and seasonal component model

$$y_{tj} = \alpha(s_t) + \beta_j(s_t) + e_{tj}, \quad t = 1, \ldots, n, \quad j = 1, \ldots, 4,$$

where $s_t = t/n$, $\alpha(x) = \exp(-0.7 + 3.5\,x)$, $\beta_1(x) = -3.1\,x^2 + 17.1\,x^4 - 28.1\,x^5 + 15.7\,x^6$,

$\beta_2(x) = -0.5\,x^2 + 15.7\,x^6 - 15.2\,x^7$, $\beta_3(x) = -0.2 + 4.8\,x^2 - 7.7\,x^3$, and $\beta_4(x) = -\beta_1(x) - \beta_2(x) - \beta_3(x)$, for $0 < x \leq 1$. Here, the error $\{e_m\}$ are generated from the following AR(1) model:

$$e_m = 0.9\,e_{m-1} + \varepsilon_m,$$

and ε_t is generated from $N(0,\ 0.1^2)$.

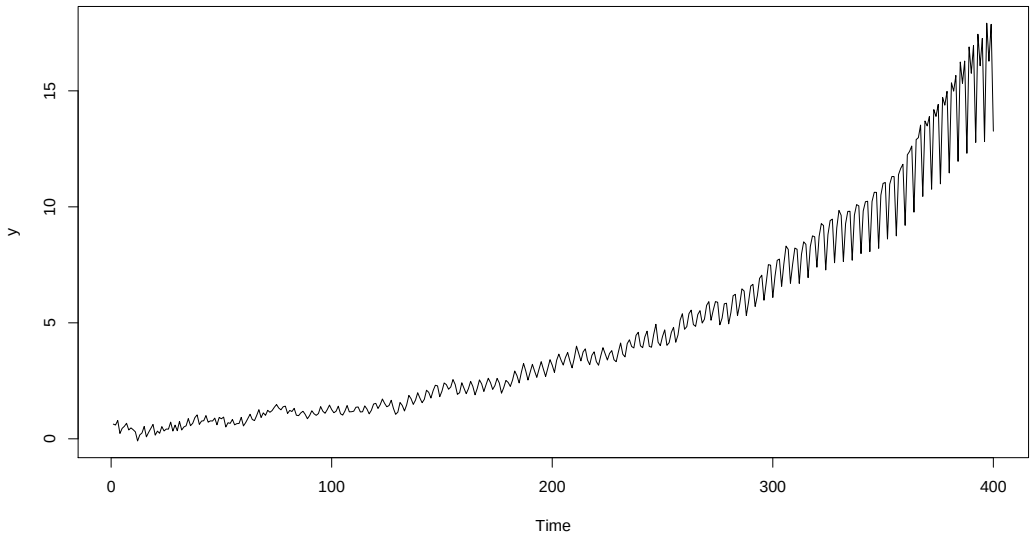


Figure 4.1: Time series plot of a typical sample from Example 4.1 with $n = 100$.

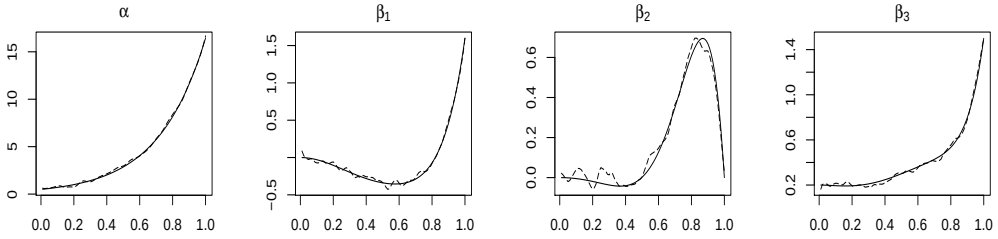


Figure 4.2: Estimation results for a typical sample from Example 4.1 with $n = 100$. The local linear estimator (dashed line) of the trend function $\{\alpha(\cdot)\}$ and seasonal effect functions $\{\beta_j(\cdot)\}$ (solid line).

The sample sizes are $n = 50$, 100, and 300, respectively. Figure 4.1 gives the time plot of a typical sample with the sample size $n = 100$. Figure 4.2 shows the estimated $\alpha(\cdot)$ and $\{\beta_j(\cdot)\}$ (dashed lines) from the typical sample, together with their true values (solid lines), and it can be seen that estimated values are very close to the true values. The median and standard deviation (in parentheses) of the 500 MADE

values are summarized in Table 4.1, which confirms that all the MADE values decrease as n increases, as dictated by the asymptotic theory. Clearly, the proposed modeling procedure performs fairly well.

Table 4.1: The median and standard deviation of 500 MADE values for Example 4.1

n	$\mathcal{E}$	$\mathcal{E}_1$	$\mathcal{E}_2$	$\mathcal{E}_3$	$\mathcal{E}_4$
50	0.151(0.031)	0.041(0.006)	0.030(0.005)	0.041(0.007)	0.030(0.005)
100	0.132(0.024)	0.026(0.004)	0.021(0.003)	0.026(0.004)	0.021(0.003)
300	0.093(0.014)	0.013(0.002)	0.013(0.002)	0.013(0.002)	0.012(0.002)

Example 4.2. In this example, a seasonal AR model with functional coefficients is considered.

$$y_{tj} = (\alpha_1(s_t)+\beta_{1j}(s_t))y_{t,j-1}+(\alpha_2(s_t))+\beta_{2j}(s_t))y_{t-1,j}+e_{tj}, \quad t = 1, \ldots, n, \quad j = 1, \ldots, 4,$$

where $s_t = t/n$, $y_{t,0} = y_{t-1,4}$, $\alpha_1(x) = 0.5\,x^2 + 0.5\,x + 0.13$, $\beta_{11}(x) = -0.8\,x^2 + 0.5$, $\beta_{12}(x) = 0.2\,x^3+0.8\,x^2-0.4\,x$, $\beta_{13}(x) = 0.7\,x^4-0.1\,x^3-0.15\,x$, $\alpha_2(x) = 0.17\sin(2\pi\,x)-0.2$, $\beta_{21}(x) = -0.5\cos(\pi\,x)+0.1$, $\beta_{22}(x) = -0.5\sin(0.5\pi\,x)+0.3$, $\beta_{23}(x) = -0.5\cos(0.5\pi\,x)$, and $\beta_{k4}(x) = -\beta_{k1}(x) - \beta_{k2}(x) - \beta_{k3}(x)$, $k = 1, 2$, for $0 < x \leq 1$. The errors, $\{e_{tj}\}$, are i.i.d. distributed as $N(0, 1)$. The seasonal AR coefficients at lag 1 are polynomial functions, and the seasonal AR coefficients at lag 4 are a combination of trigonometric functions plus some constants.

The sample sizes used are $n = 300$, 500, and 1000, respectively. For a typical sample with the sample size $n = 300$, Figure 4.3 and Figure 4.4 give the time plot of $\{y_t\}$ and the subseries $\{y_{tj}\}$ for each season. The seasonal pattern of the time series is not revealed here. However, the ACF and PACF of the time series (Figure 4.5) demonstrate a clear indication of seasonality. Figure 4.6 plots the estimated $\alpha_k(\cdot)$ and $\{\beta_{kj}(\cdot)\}$ (dashed lines) from a typical sample with $n = 300$, together with their true values (solid lines). It is seen that the estimation is reasonable, considering the small sample size. Note that the main function $\alpha_k(\cdot)$ has a much smaller scale than the rest of the functions. The median and standard deviation (in parentheses) of the 500 MADE values are summarized in Table 4.2.

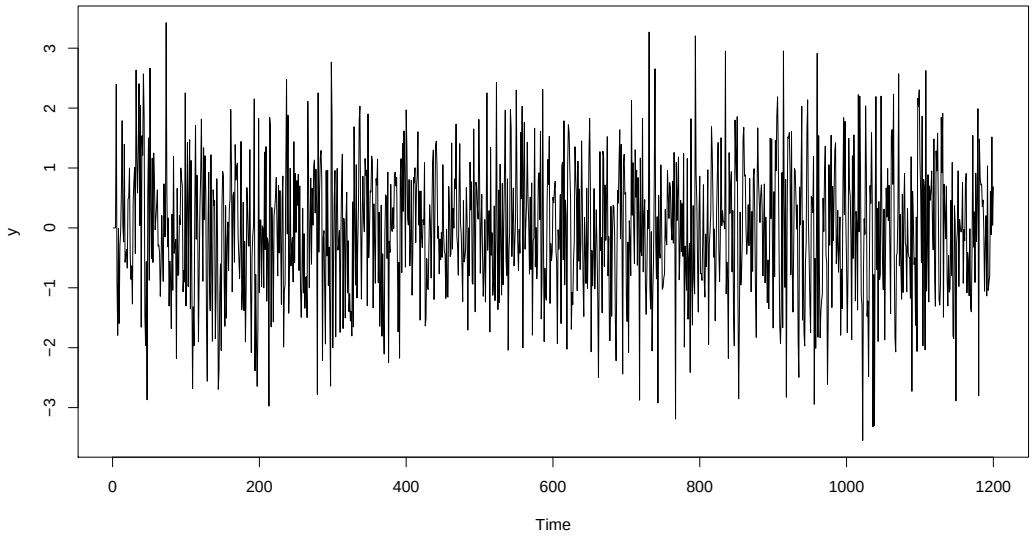


Figure 4.3: Time plot of a typical sample from Example 4.2, with $n = 300$.

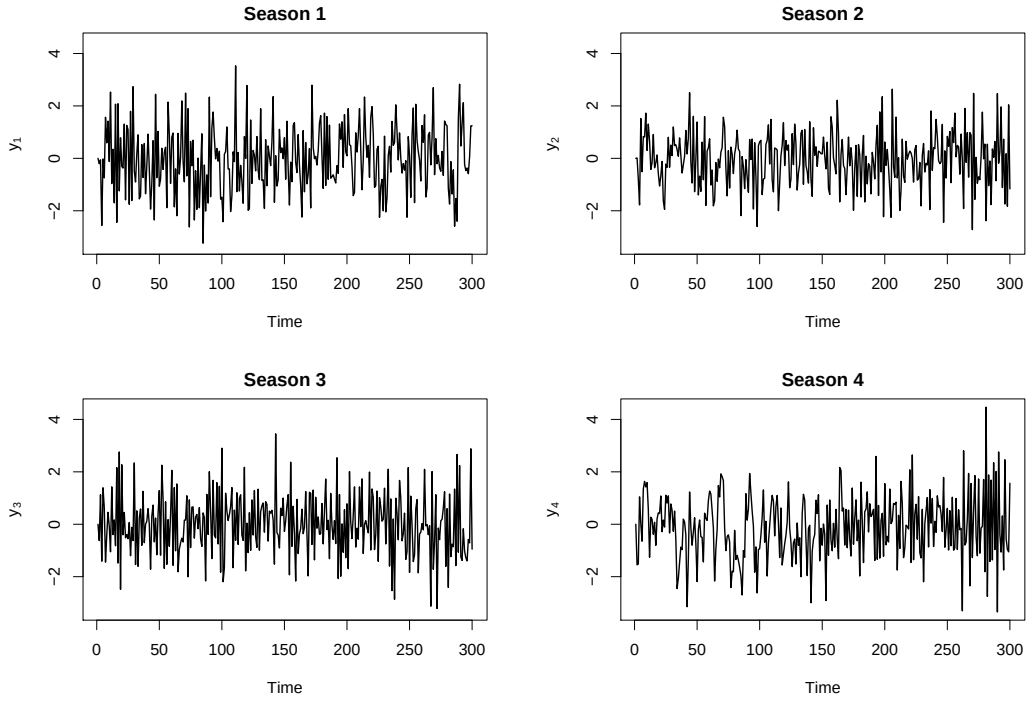


Figure 4.4: Time plots of subseries y_{tj} for each season of a typical sample from Example 4.2 shown in Figure 4.3.

4.4 An Analysis of the Hawaiian Tourism Data

As a major international tourist site, Hawaii's economy relies heavily on tourism. For planning, marketing and pricing purposes, a deep understanding of the dynamics and

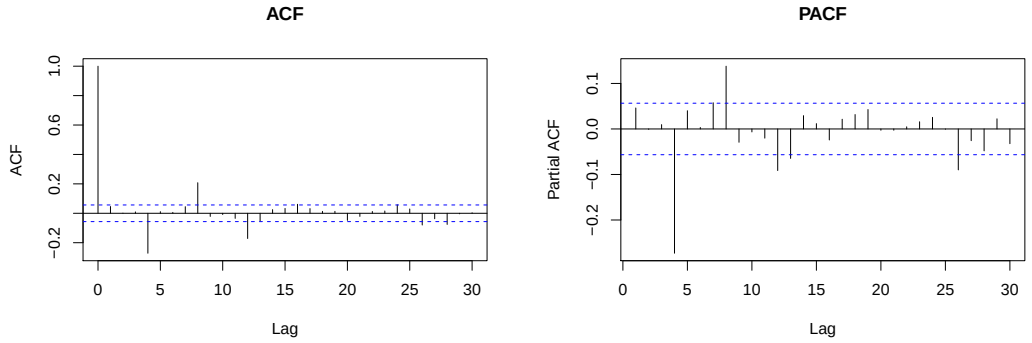


Figure 4.5: ACF and PACF for a typical sample from Example 4.2 shown in Figure 4.3.

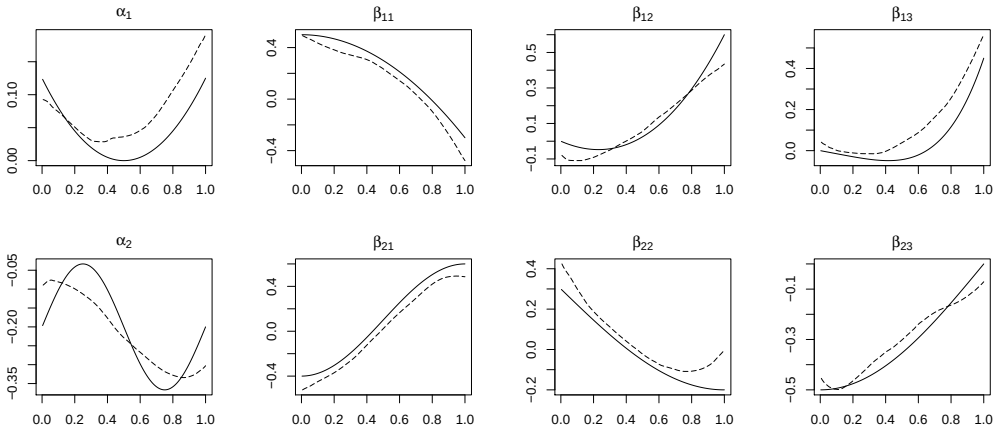


Figure 4.6: Estimation results for a typical sample from Example 2 with $n = 300$. The local linear estimator (dashed line) of the trend function $\{\alpha_k(\cdot)\}$ and seasonal effect functions $\{\beta_{kj}(\cdot)\}$ (solid line).

a capability of accurate prediction of the number of tourists visiting Hawaii are very important to the tourist business and local economy in Hawaii. Due to weather, school

Table 4.2: The median and standard deviation of 500 MADE values for Example 4.2

n	$\mathcal{E}_1$	$\mathcal{E}_{11}$	$\mathcal{E}_{12}$	$\mathcal{E}_{13}$	$\mathcal{E}_{14}$
300	0.035(0.015)	0.058(0.025)	0.061(0.027)	0.061(0.027)	0.064(0.026)
500	0.028(0.011)	0.045(0.018)	0.049(0.021)	0.052(0.019)	0.051(0.021)
1000	0.021(0.007)	0.034(0.013)	0.038(0.014)	0.039(0.013)	0.038(0.013)

n	$\mathcal{E}_2$	$\mathcal{E}_{21}$	$\mathcal{E}_{22}$	$\mathcal{E}_{23}$	$\mathcal{E}_{24}$
300	0.053(0.013)	0.058(0.026)	0.059(0.025)	0.054(0.025)	0.059(0.023)
500	0.046(0.012)	0.047(0.019)	0.047(0.020)	0.041(0.017)	0.048(0.018)
1000	0.031(0.010)	0.037(0.013)	0.036(0.014)	0.034(0.013)	0.036(0.013)

schedule and other factors, numbers of tourists often shows seasonality. Chen and Fomby (1999) used the stable seasonal pattern model to fit the monthly time series of number of tourists visiting Hawaii. Here we apply the proposed functional-coefficient seasonal time series model to analyze an updated version of Hawaiian tourism data (1970-2012), obtained from the Hawaii Visitors Bureau. Hence $n = 43$, $d = 12$ and $T = 516$.

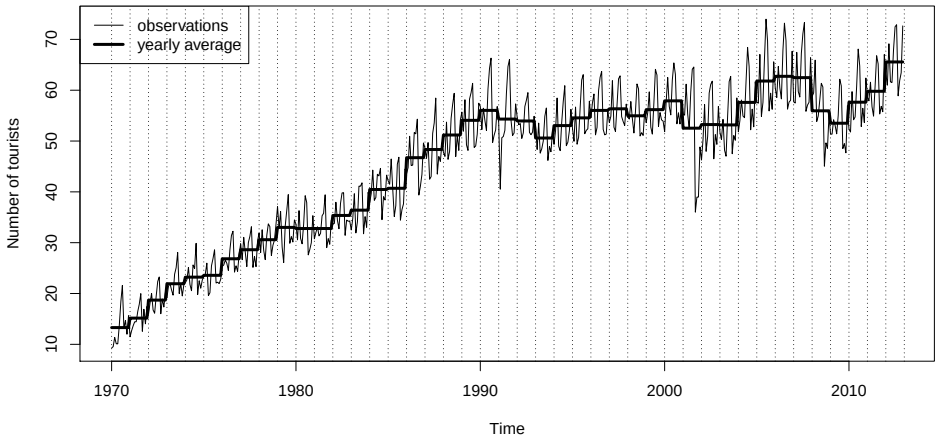
For expositional convenience, we re-scale the data by dividing 10^5. Figure 4.7(a) presents the monthly observations from January 1970 through December 2012 with the yearly averages (thick line). It demonstrates that the number of tourists visiting Hawaii experienced two growing stages. In the first stage, it increased rapidly from 1970 to 1990. In the second stage, the number of tourists still rose steadily from 1991 to 2012 although there were three down turns, which happened in the early 1990s (the economy recession), September 2001 (the 9/11 tragedy), and 2007-2010 (after the financial crisis), respectively. Figure 4.7(b) plots the monthly subseries $\{y_{tj}\}$ for each month over the years. To see more clearly the seasonality, Figure 4.8 gives the boxplot of deviations from the yearly average for each month. It shows that the heaviest travelled months in Hawaii are March, December and the summer.

We first use the nonparametric seasonal model

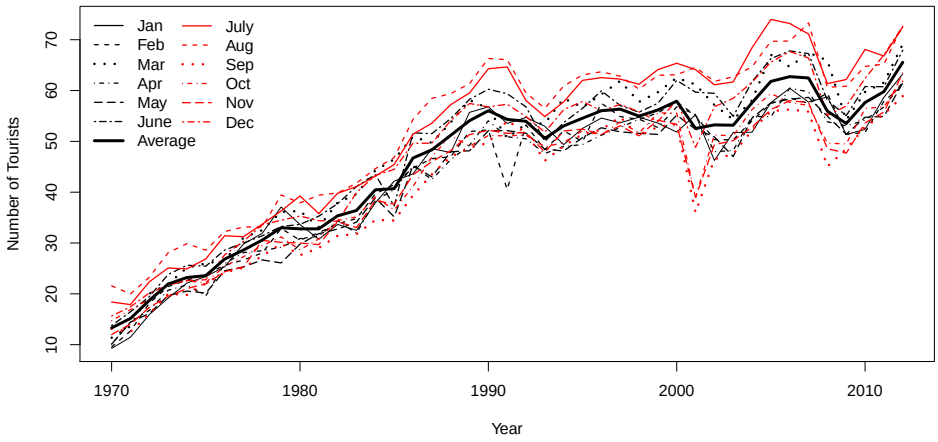
$$y_{tj} = \alpha(s_t) + \beta_j(s_t) + e_{tj}, \quad t = 1, \ldots, 43, \quad j = 1, \ldots, 12, \tag{4.10}$$

to fit the series, with the constraint $\sum_{j=1}^{12} \beta_j(s) = 0$ for all $s \in (0, 1]$. Figure 4.9(a) plots the estimated trend function (solid line) plus/minus twice estimated pointwise standard errors (dashed lines) with the bias ignored. The yearly average (thick line) is also included. We can see that the 95% confidence interval covers most of the observed yearly averages except these in 1990-1992, in 2001 and around 2008 due to the economy recessions and the terrorist attack. Such sudden changes may cause additional bias in the estimation.

Figure 4.10 shows the estimated seasonal effect functions, and it can be seen that the seasonal effect functions of March, December and the summer months are all positive,



(a)



(b)

Figure 4.7: Hawaiian tourism data from 1970 to 2012. (a): Time series plot of number of visitors (solid line) with yearly average (thick line); (b): time series plot of number of visitors for each month with yearly average (thick line).

and for the rest of them are negative. Also, the range of the seasonal effect functions increases over time, as the yearly average. Such dynamics are expected. In addition, economy downturn in 1990 has the largest negative impact on February and March; the 9/11 tragedy decreases the tourists severely on September, October, November and December in 2001; and financial crisis after 2007 does not make some very sharp turning points for seasonal functions, because its influence lasts for a few years (yearly average reduces greatly in 2008-2010). It is also interesting to see that December is becoming more and more popular to visit Hawaii in the recent years.

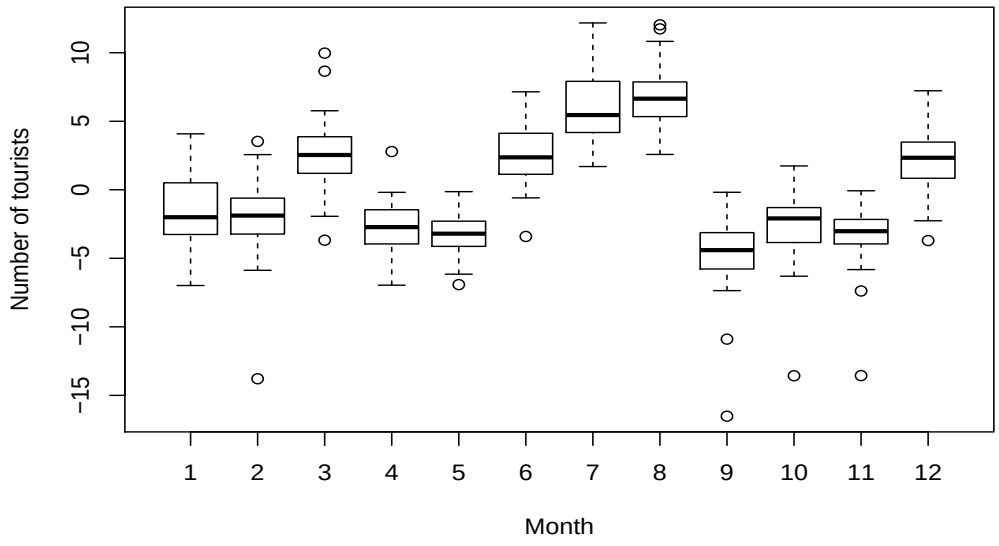
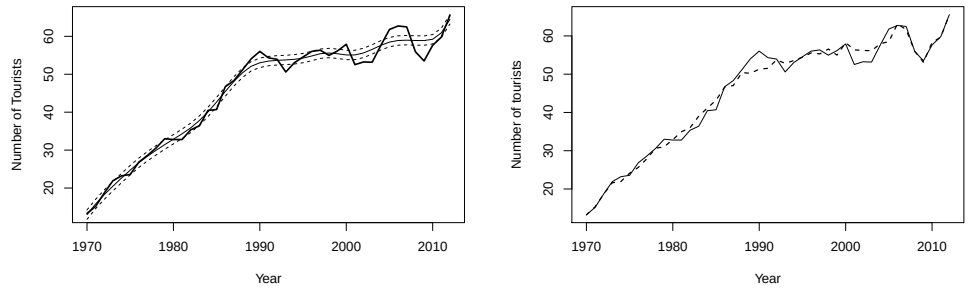


Figure 4.8: Hawaiian tourism data from 1970 to 2012. Boxplot of deviations from the yearly average for each month.



(a) (b)

Figure 4.9: Hawaiian tourism data from 1970 to 2012. (a) Estimated trend function (solid line) plus/minus twice estimated standard errors (dashed lines) with bias ignored and the yearly average (thick line) for model (4.10); (b) estimated trend function (dashed line) with the yearly average for model (4.11).

To model more accurately the negative impacts of tourism around 1991-1992 and 2008-2010, partially due to the economy recessions in the U.S. around these two periods, we incorporate some economic indices as exogenous variables. Since U.S. and Japan are the major regions that contribute about 85% of the tourists to visit Hawaii, we add the growth rate of annual personal disposable income (PDI) of both countries to the

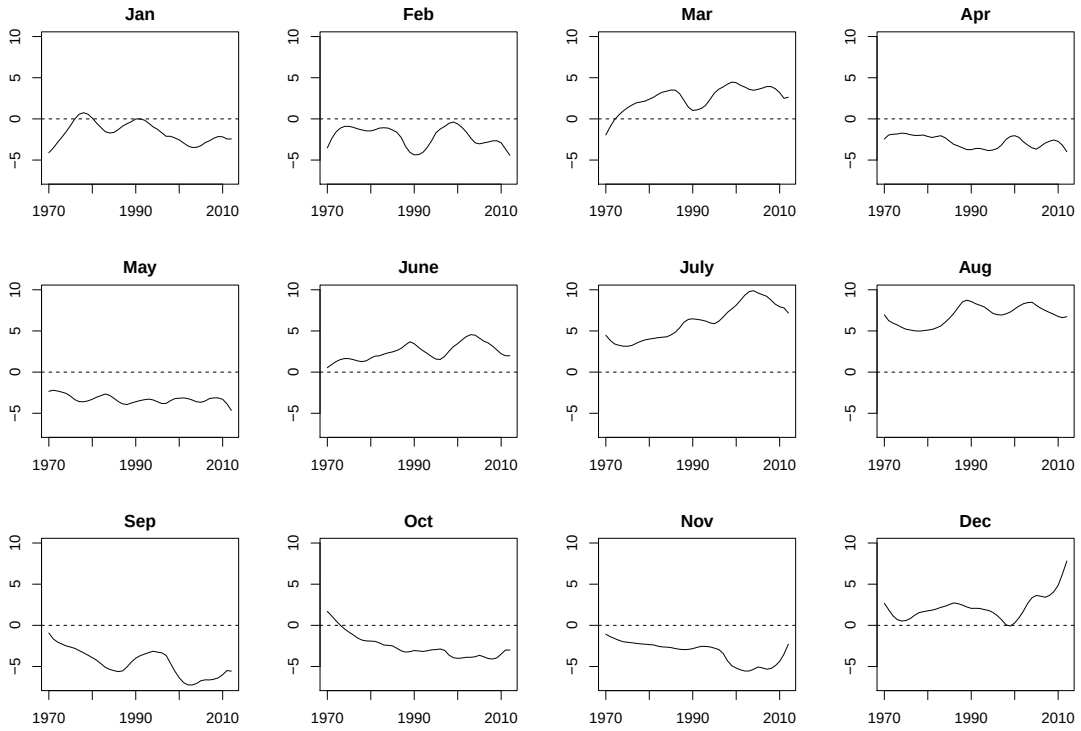


Figure 4.10: Hawaiian tourism data from 1970 to 2012. Estimated seasonal functions (solid line) with the zero line (dashed line) for model (4.10).

model, as in Chen and Fomby (1999). They are denoted by x_1 and x_2 for U.S. and Japan, respectively.

Specifically, we consider the following seasonal functional-coefficient model

$$y_{tj} = [\alpha_0(s_t) + \beta_{0j}(s_t)] + [\alpha_1(s_t) + \beta_{1j}(s_t)]x_{1t} + [\alpha_2(s_t) + \beta_{2j}(s_t)]x_{2t} + e_{tj}, \qquad (4.11)$$

$t = 1, \ldots, 43$, $j = 1, \ldots, 12$, subject to the constraints

$$\sum_{j=1}^{12} \beta_{kj}(s) = 0 \quad \text{for each } k = 0,\ 1,\ 2 \ \text{ and all } \ s \in (0, 1].$$

Comparing to model (4.10), the two extra terms in model (4.11) try to make adjustments using the economic variables. In Figure 4.9(b), the dash line shows the estimated overall trend function $\widehat{\alpha}_0(s_t) + \widehat{\alpha}_1(s_t)\, x_{1t} + \widehat{\alpha}_2(s_t)\, x_{2t}$ against t, calculated with the observed values of x_{1t} and x_{2t}. The solid line shows observed yearly average. It is roughly the same as that using the simpler model (4.10) before 2005, but the adjustment to the

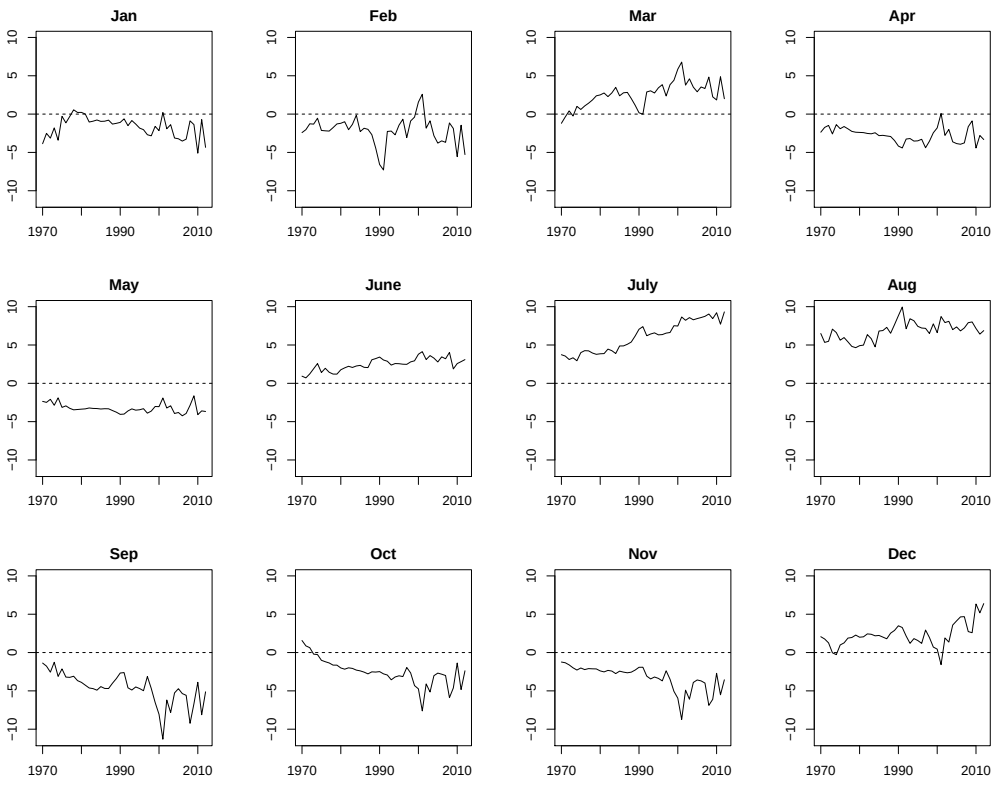


Figure 4.11: Hawaiian tourism data from 1970 to 2012. Estimated seasonal functions (solid line) with the zero line (dashed line) for model (4.11).

overall annual trend improves significantly the estimation for years after 2005. The estimated seasonal functions $\widehat{\beta}_{0j}(s_t) + \widehat{\beta}_{1j}(s_t)\, x_{1t} + \widehat{\beta}_{2j}(s_t)\, x_{2t}$, plotted against time, again calculated with the observed values of x_{1t} and x_{2t}, are depicted in Figure 4.11. The basic shapes of the seasonal functions remain similar as those shown in Figure 4.10, but the extra terms using economic indices make the seasonal functions less smooth and reflect the significant influence of the financial crisis.

In Figure 4.9(b), the dash line shows the estimated overall trend function $\widehat{\alpha}_0(s_t) + \widehat{\alpha}_1(s_t)\, x_{1t} + \widehat{\alpha}_2(s_t)\, x_{2t}$ against t, calculated with the observed values of x_{1t} and x_{2t}. The solid line shows observed yearly average. It is roughly the same as that using the simpler model (4.10) before 2005, but the adjustment to the overall annual trend improves significantly the estimation for years after 2005. The estimated seasonal functions $\widehat{\beta}_{0j}(s_t) + \widehat{\beta}_{1j}(s_t)\, x_{1t} + \widehat{\beta}_{2j}(s_t)\, x_{2t}$, plotted against time, again calculated with the observed values of x_{1t} and x_{2t}, are depicted in Figure 4.11. The basic shapes of the seasonal

functions remain similar as those shown in Figure 4.10, but the extra terms using economic indices make the seasonal functions less smooth and reflect the significant influence of the financial crisis.

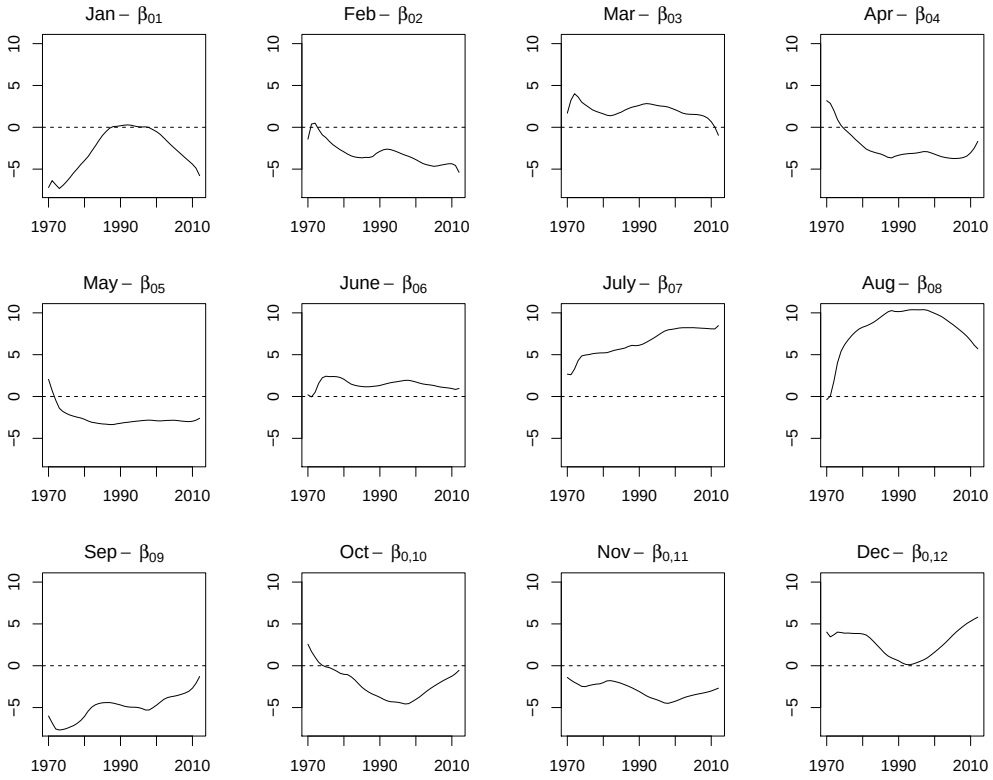


Figure 4.12: Hawaiian tourism data from 1970 to 2012 for model (4.11). Estimated seasonal trend β_{0i} for $i = 1, \ldots, 12$.

Figure 4.12 shows the estimated seasonal trend, β_{0i} for $i = 1, \ldots, 12$. Comparing with Figure 4.10, we can see that for most months, the pattern of the trend remains similar. However, β_{01} and β_{08} estimated for model (4.11) have different trend from these estimated for model (4.10). They increased in 1970s, and decreased after 2000 for model (4.11), while they basically remained at the same level for model (4.10). It partially indicates that the increase of tourism in January and August are due to the growth of PDI from 2000 to 2012 based on model (4.11).

Figure 4.13 and Figure 4.14 display the estimated seasonal income effects of U.S. and Japan, respectively. For most months, the income effect of U.S. was weakened from 1970 to 1985, and then was strengthen until subprime mortgage crisis. For Japan, the

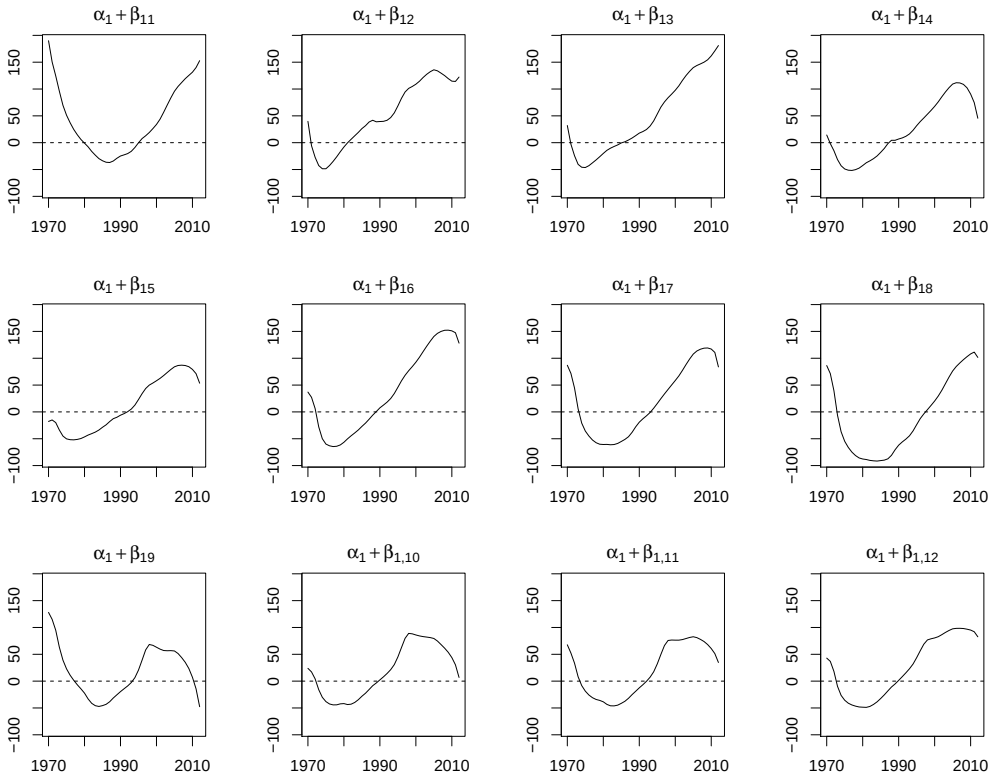


Figure 4.13: Hawaiian tourism data from 1970 to 2012 for model (4.11). Estimated seasonal income effect of U.S. for each month, estimated $\alpha_1 + \beta_{1i}$ for $i = 1, \ldots, 12$.

income effect for the first half of the year was rather weak. However, PDI had a very strong impact on the number of tourists for the second half of the year, especially after 1995, and it had been increasing over the whole period.

We select two sets of estimated functions which have the larger variations among 12 months, and whose estimated seasonal trend functions are different from these for model (4.10), plotted in Figure 4.15. The estimation results give us some detailed explanations of the Hawaiian tourism data. Figure 4.15(a) presents α_1, the overall income effect of U.S. over time. It is seen that the income effect decreased in the period of 1970-1988, then gradually increased until the financial crisis. However, the additional income effect of U.S. in the month of January β_{11} decreased in the period of 1970-1988, and then increased, shown in Figure 4.15(b). The overall income effect on U.S. in January over time is plotted in Figure 4.15(c). The overall income effect for Japan α_2 increases consistently over time, shown in Figure 4.15(d). The income effect

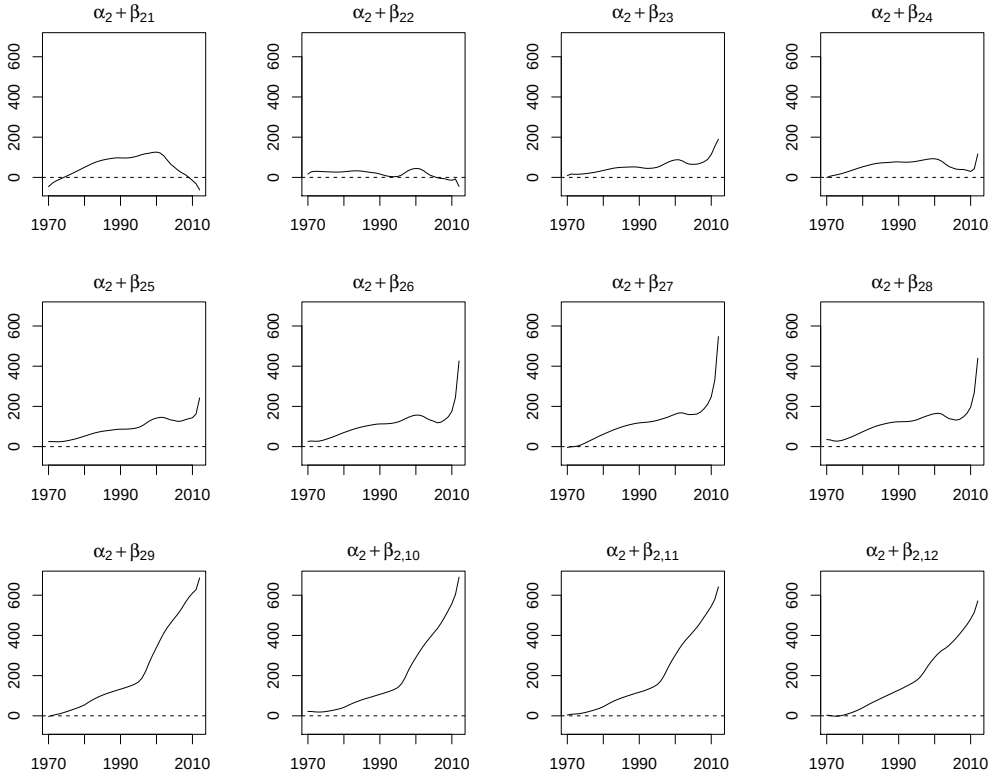


Figure 4.14: Hawaiian tourism data from 1970 to 2012 for model (4.11). Estimated seasonal income effect of Japan for each month, estimated $\alpha_2 + \beta_{2i}$ for $i = 1, \ldots, 12$.

of the month of August β_{28} decreased after 1995, and then rose very sharply after 2007 (Figure 4.15(e)). Overall, the income growth becomes a more and more deciding factor on the number of Japanese tourists visiting Hawaii in August.

We compare model (4.11) with the seasonal ARIMA model by out-sample rolling forecasting. Specifically, for $m_0 = T_0, \ldots, T - \ell$, we use data observed at time m_0, $\{y_j, j = 1, \ldots, m_0\}$ to predict number of tourists visiting Hawaii at time $m_0 + \ell$, $\{y_{m_0+\ell}\}$, where the forecast horizon is ℓ months. Here we set $T_0 = 408$, and let ℓ take values from 1 to 48. For computational convenience, when the forecast origin is m_0, where $m_0 = 12(t_0 - 1) + j_0$, $1 \leq j_0 \leq 12$, data is separated into $(t_0 - 1)$ periods, not by the calendar year, but by the following rule: the months $j_0 + 1, \ldots, 12$ in the year h, and the months $1, \ldots, j_0$ in the year $h + 1$ are defined as the h-th period, for each $h = 1, \ldots, t_0 - 1$. In other words, data is separated into periods as $\{y_{j_0+1}, \ldots, y_{j_0+12}\}, \ldots, \{y_{m_0-11}, \ldots, y_{m_0}\}$, when the forecast origin is m_0.

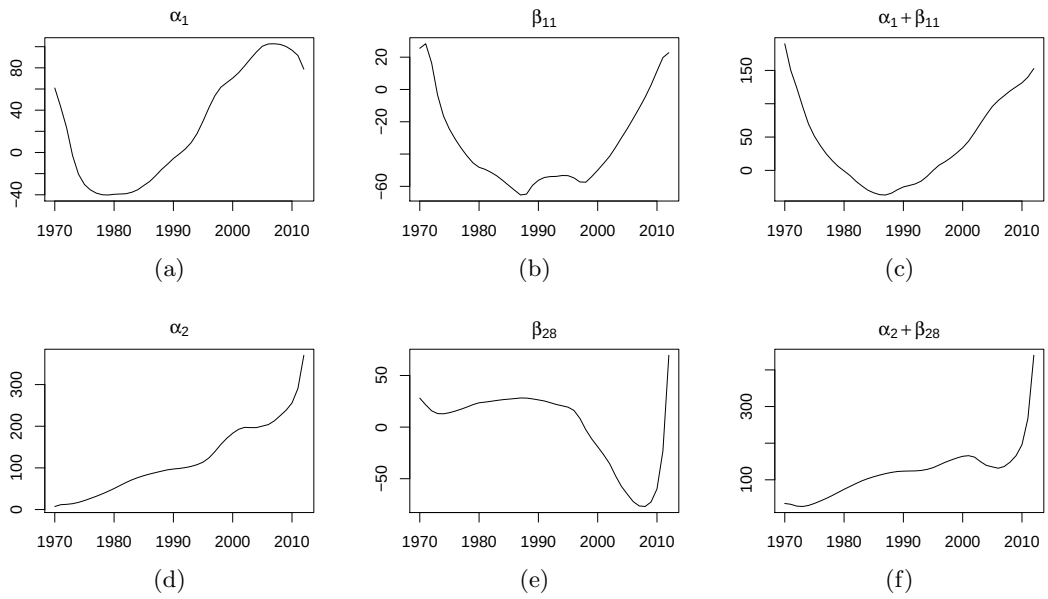


Figure 4.15: Hawaiian tourism data from 1970 to 2012 for model (4.11). Top panel shows the seasonal income effect of U.S. in January: (a) estimated α_1; (b) estimated β_{11}; (c) estimated $\alpha_1 + \beta_{11}$. Bottom panel shows the seasonal income effect of Japan in August: (d) estimated α_2; (e) estimated β_{28}; (f) estimated $\alpha_2 + \beta_{28}$.

Figure 4.16 shows the time series plot and sample autocorrelations of residuals $\{\widehat{e}_m\}$ by model (4.11). There is no significant seasonality but serial dependence in the data after extracting seasonal trend and income effects. Hence, we specify an AR(1) model for the residuals

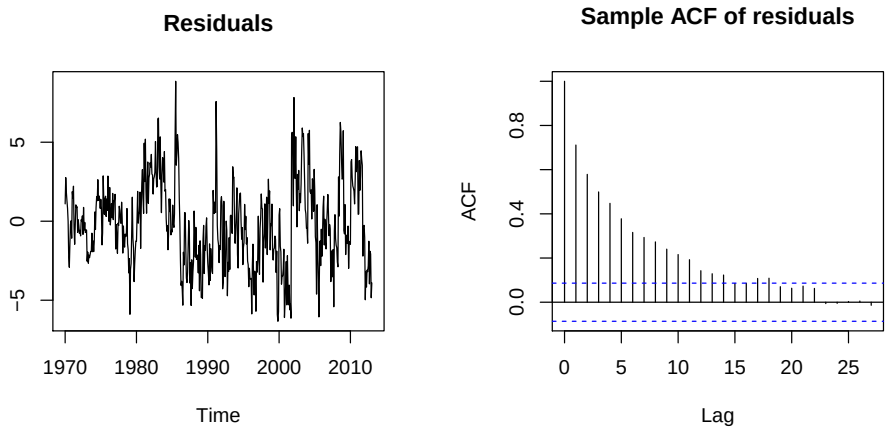
$$\widehat{e}_m = \phi \widehat{e}_{m-1} + \eta_m,$$

where $\{\eta_m\}$ is a white noise process.

When the forecast origin is m_0, ϕ can be estimated by least squares, i.e., $\widehat{\phi} = \arg\min \sum_{m=2}^{m_0} (\widehat{e}_m - \phi \widehat{e}_{m-1})^2$, and $y_{m_0+\ell}$ can be predicted as

$$\widehat{y}_{m_0+\ell} = [\widehat{\alpha}_0(s_t) + \widehat{\beta}_{0j}(s_t)] + [\widehat{\alpha}_1(s_t) + \widehat{\beta}_{1j}(s_t)]x_{1t} + [\widehat{\alpha}_2(s_t) + \widehat{\beta}_{2j}(s_t)]x_{2t} + \widehat{\phi}^{\ell}\, \widehat{e}_{m_0}, \quad (4.12)$$

where the $(m_0 + \ell)$-th month is the j-th month in the t-th period, i.e. $m_0 + \ell - j_0 = 12(t-1) + j$, $1 \le j \le 12$, $\{\widehat{\alpha}_k(s_t), \widehat{\beta}_{kj}(s_t), k = 0, 1, 2\}$ are estimates of trend and seasonal components in j-th month and t-th period based on data observed at time m_0 with equation (9), and $\widehat{e}_{m_0}$ is the residual at time m_0.



(a) (b)

Figure 4.16: Hawaiian tourism data from 1970 to 2012. (a) Time series plot of residuals for model (4.11); (b) sample auto-correlations of residuals for model (4.11).

For the seasonal ARIMA model, we select the following model based on AIC to fit the data

$$(1 - \phi_1 B)(1 - \phi_{12} B^{12})(1 - B^{12}) y_m = a_m, \tag{4.13}$$

where B is the back-shift operator, ϕ_1 and ϕ_{12} are the AR coefficient and the seasonal AR coefficient respectively, and $\{a_m\}$ is a white noise process.

Figure 4.17 plots the mean squared out-sample prediction error against different forecast horizon for two models. Although the seasonal ARIMA model predicts the number of tourists less than 1-year ahead better than our model, it suffers severely from the increases of forecast horizon. Predictions by our model are more stable, and outperform when forecast horizon is longer. The functional-coefficient seasonal model characterizes the long-term trend of the series, and describes the dynamic relationship between growth of PDI and the number of tourists visiting Hawaii.

4.5 Concluding Remarks

We propose a nonparametric seasonal time series model with functional coefficients. By allowing the coefficients to change over time, it describes the time-varying impact

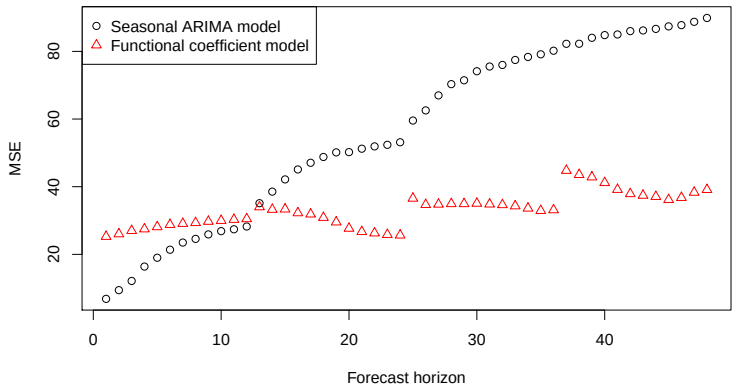


Figure 4.17: Hawaiian tourism data from 1970 to 2012. The mean squared forecasting error for model (4.11) and the seasonal ARIMA model against different forecast horizon.

the trend and possible exogenous variables exert on the process. The seasonal components in the model help to characterize the periodic behaviors of the time series data. This chapter focuses on the nonparametric approach, with its flexibility and minimum subjective assumptions. It should be pointed out that the results from the nonparametric approach can be used as a first step for building a more parsimonious models which may lead to more accurate and stable estimation and better performance. The proposed method is implemented to analyze the Hawaii tourism data, and results show that our model provides easier interpretation and better long term prediction than a linear seasonal ARIMA models.

Bibligraphy

Anderson, T.W. (1963). The use of factor analysis in the statistical analysis of multiple time series. *Psychometrika*, **28**, 1-25.

Andrews, D.W.K. (1991). Heteroskedasticity and autocorrelation consistent covariance matrix estimation. *Econometrica*, **59**, 817-858.

Antoniadis, A., Paparoditis, E., and Sapatinas, T. (2009). Bandwidth selection for functional time series prediction. *Statistics and Probability Letter*, **79**, 733-740.

Aue, A., Noriho, D.D., and Hörmann, S. (2012). On the prediction of functional time series. Technical Report.

Auestad, B., and Tjøstheim, D. (1990). Identification of nonlinear time series: First order characterization and order determination. *Biometrika*, **77**, 669-687.

Bai, J. and Ng, S. (2002). Determining the number of factors in approximate factor models. *Econometrica*, **70**, 191-221.

Bathia, N., Yao, Q., and Ziegelmann, F. (2010). Identifying the finite dimensionality of curve time series. *The Annals of Statistics*, **38**, 3352-3386.

Bauwens, L., Laurent, S., and Rombouts, J.V. (2006). Multivariate GARCH models: a survey. *Journal of Applied Econometrics*, **21**, 79-109.

Berkes, I., Horvath, L., and Rice, G. (2013). Weak invariance principles for sums of dependent random function. *Stochastic Processes and their Application*, **123**, 385-403.

Bernanke, B.S. and Gertler, M. (2000). Monetary policy and asset price volatility. *National Bureau of Economic Research*.

Besse, P., Cardot, H., and Stephenson, D. (2000). Autoregressive forecasting of some functional climatic variations. *Scandinavian Journal of Statistics*, **27**, 673-687.

Bosq, D. (2000). *Linear Processes in Function Spaces, Theory and Applications. Lecture Notes in Statistics*, New York: Springer-Verlag.

Box, G.E.P. and Jenkins, G.M. (1971). *Time Series Analysis, Forecasting and Control*, San Francisco: Holden-Day, Inc.

Box, G.E.P., Jenkins, G.M., and Reinsel, G.C. (1994). *Time Series Analysis, Forecasting and Control*, (3th ed.), Englewood Cliffs, NJ: Prentice-Hall.

Bradley, R.C. (2005). Basic properties of strong mixing conditions. A survey and some open questions. *Probability Surveys*.

Brockwell, P. and Davis, R. (1990). *Time Series: Theory and Methods*, New York: Springer.

Brumback, B. and Rice, J.A.(1998). Smoothing spline models for the analysis of nested and crossed samples of curves. *Journal of the American Statistical Association*, **93**, 961-994.

Burman, P. and Shumway, R.H. (1998). Semiparametric modeling of seasonal time series. *Journal of Time Series Analysis*, **19**, 127-145.

Cai, Z. (2002). Two-step likelihood estimation procedure for varying-coefficient models. *Journal of Multivariate Analysis*, **82**, 189-209

Cai, Z. and Chen, R. (2006). Flexible seasonal time series models. In *Econometric Analysis of Financial and Economic Time Series*,**20**, 63-87.

Cai, Z., Fan, J., and Yao, Q. (2000). Functional-coefficient regression models for nonlinear time series. *Journal of the American Statistical Association*, **95**, 941-956.

Cai, Z., Li, Q., and Park, J.Y. (2009). Functional-coefficient models for nonstationary time series data", *Journal of Econometrics*, **148**, 101-113.

Cai, Z. and Ould-Said, E. (2001). Local robust regression estimation for time series. *Manuscript*.

Cai, Z. and Tiwari, R.C. (2000). Application of a local linear autoregressive model to BOD time series. *Environmetrics*, **11**, 341-350.

Chamberlain, G. and Rothschild, M. (1983). Arbitrage, factor structure, and mean-variance analysis on large asset markets. *Econometrica*, **51**, 1281-1304.

Chan, K.S. (1993). Consistency and limiting distribution of the least squares estimator of a threshold autoregressive model. *The Annals of Statistics*, 520-533.

Chang, J., Guo, B. and Yao, Q. (2013). High dimensional stochastic regression with latent factors, endogeneity and nonlinearity. Technical report.

Chen, R. (1995). Threshold variable selection in open-loop threshold autoregressive models. *Journal of Time Series*, **16**, 461-481.

Chen, R. and Fomby, T. (1999). Forecasting with stable seasonal pattern models with an application of Hawaiian tourist data. *Journal of Business & Economic Statistics*, **17**, 497-504.

Chen, R., Liu, J.S., and Tsay, R.S. (1995). Additivity tests for nonlinear autoregressive models. *Biometrika*, **82**, 369-383.

Chen, R. and Tsay, R.S. (1993a). Functional coefficient autoregressive models. *Journal of the American Statistical Association*, **88**, 298-308.

Chen, R. and Tsay, R.S. (1993b). Nonlinear additive ARX models. *Journal of the American Statistical Association*, **88**, 955-967.

Chen, X. (2008). Large sample sieve estimation of semi-nonparametric models. *Handbook of Econometrics*, **6**, 5549-5632.

Chen, X. and Shen, X. (1998). Sieve extremum estimates for weakly dependent data. *Econometrica*, **66**, 289-314.

Cleveland, R.B., Cleveland, W.S., McRae, J.E., and Terpenning, I. (1990). STL: A seasonal-trend decomposition procedure based on loess. *Journal of Official Statistics*, **6**, 3-73.

Cressie, H. and Huang, H. (1999). Classes of nonseparable, spatio-temporal stationary covariance function. *Journal of the American Statistical Association*, **94**, 1330-1339.

Diebold, F. X. and Li, C. (2006). Forecasting the term structure of government bond yields. *Journal of Econometrics*, **130**, 337-364.

Diebold, F.X. and Rudebusch, G.D. (1994). Measuring business cycles: a modern perspective. *National Bureau of Economic Research.*

Doukhan, P. and Louhichi, S. (1999). A new weak dependence condition and applications to moment inequalities", *Stochastic Processes and Their Applications*, **84**, 313-342.

Doz, C., Giannone, D., Reichlin, L. (2011). A two-step estimator for large approximate dynamic factor models based on Kalman filtering. *Journal of Econometrics*, **164**, 188-205.

Engle, R.F. and Granger, C.W. (1987). Co-integration and error correction: representation, estimation, and testing. *Econometrica*, **55**, 251-276.

Engle, R.F. and Kroner, K.F. (1995). Multivariate simultaneous generalized ARCH. *Econometric Theory*, **11**, 122-150.

Fan, J. and Gijbels, I. (1996). *Local Polynomial Modelling and Its Applications*, London: Chapman and Hall.

Fan, J. and Yao, Q. (2003). *Nonlinear Time Series: Nonparametric and Parametric Methods*, New York: Springer-Verlag.

Fan, J., Yao, Q., and Cai, Z. (2003). Adaptive varying-coefficient linear models. *Journal of Royal Statistical Society, Series B*, **65**, 57-80.

Ferraty, F. and Vieu, P. (2006). *Nonparametric Functional Data Analysis*, New York: Springer.

Forney Jr, G.D. (1973). The Viterbi algorithm. *Proceedings of the IEEE*, **61**, 268-278.

Forni, M., Hallin, M., Lippi, M., and Reichlin, L. (2000). The generalized dynamic factor model: identification and estimation. *The Review of Economics and Statistics*, **82**, 540-554.

Franses, P.H. (1996). *Periodicity and Stochastic Trends in Economic Time Series*, New York: Cambridge University Press.

Franses, P.H. (1998). *Time Series Models for Business and Economic Forecasting*, New York: Cambridge University Press.

Gasser, T., Müller, H.G., Köhler, W., Molinari, L., and Prader, A. (1984). Nonparametric regression anlaysis of growth curves. *Annals of Statistics*, **12**, 210-229.

Ghysels, E. and Osborn, D.R. (2001). *The Econometric Analysis of Seasonal Time Series*, New York: Cambridge University Press.

Gneiting, T. (2002). Nonseparable, stationary covariance functions for space-time data. *Journal of the American Statistical Association*, **97**, 590-600.

Gorodetskii, V.V. (1977). On the strong mixing property for linear sequences. *Theory of Probability and Its Applications*, **22**, 411-413.

Grenander, U. and Szego, G.(1984). *Toeplitz Forms and Their Applications*, University of California Press.

Grey, S.F. (1996). Modeling the conditional distribution of interest rates as a regime-switching process. *Journal of Economics*, **42**, 27-62.

Halberstam, H. and Richert, H.E. (2013). *Sieve Methods*, Courier Dover Publications.

Hallin, M. and Liška, R. (2007). Determining the number of factors in the general dynamic factor model. *Journal of the American Statistical Association*, **102**, 603-617.

Hamilton, J. (1989). A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica*, **57**, 357-384.

Hamilton, J. (1996). Specification testing in Markov-switching time-series models. *Journal of Econometrics*, **70**, 127-157.

Hamilton, J. and Susmel R. (1994). Autoregressive conditional heteroscedasticity and changes in regime. *Journal of Econometrics*, **64**, 307-333.

Handcock, M.S. and Wallis, J.R. (1994). An approach to statistical spatial-temporal modeling of meteorological fields. *Journal of the American Statistical Association*, **89**, 368-378.

Hansen, B.E. (1992). The likelihood ratio test under nonstandard conditions: testing the Markov switching model of GNP. *Journal of Applied Econometrics*, **7**, S61-S82.

Härdle, W., Chen, R. and Lüetkepohl, H. (1997). A review of nonparametric time series analysis. *International Statistical Review*, **65**, 49-72.

Härdle, W., Hall, P., and Ichimura, H. (1993). Optimal smoothing of single index models. *Annals of Statistics*, **21**, 157-178.

Härdle, W., Lütkepohl, H., and Chen, R.(1997). A review of nonparametric time series analysis. *International Statistical Reviews*, **65**, 45-72.

Härdle, W. and Marron, S. (1985). Optimal bandwidth selection in nonparametric regression function estimation. *Annals of Statistics*, **13**, 1465-1481.

Härdle, W., Müller, M., Sperlich, S., and Werwatz, A.(2004). *Nonparametric and Semiparametric Models*, Heidelberg: Springer-Verlag.

Härdle, W. and Vieu, P. (1992). Kernel regression smoothing of time series. *Journal of Time Series*, **13**, 209-232.

Harvey, A., Ruiz, E. and Shephard, N. (1994). Multivariate stochastic variance models. *The Review of Economic Studies*, **61**, 247-264.

Hastie, T.J. and Tibshirani, R. (1990). *Generalized Additive Models*, London: Chapman and Hall.

Hastie, T.J. and Tibshirani, R. (1993). Varying-coefficient models (with discussion). *Journal of the Royal Statistical Society, Series B*, **55**, 757-796.

Hastie, T., Tibshirani, R. and Friedman, J. (2009). *The Elements of Statistical Learning*, New York: Springer.

Hörmann, S., Horváth, L., and Reeder, R. (2013). A functional version of ARCH model. *Econometric Theory*, **29**, 267-288.

Hörmann, S. and Kokoszka, P. (2010). Weakly dependent functional data. *The Annals of Statistics*, **38(3)**, 1845-1884.

Horváth, L., Huskova, M., and Kokoszka, P. (2010). Testing the stability of the functional autoregressive process. *Journal of Multivariate Analysis*, **101**, 352-367.

Horváth, L. and Kokoszka, P. (2012). *Inference for Functional Data with Applications*, Springer.

Horváth, L., Kokoszka, P., and Reeder, R. (2012). Estimation of the mean of functional time series and a two-sample problem. *Journal of Royal Statistical Society: Series B*, **75**, 103-122.

Houghton, A.N., Flannery, J., and Viola, M.V. (1980). Malignant melanoma in Connecticut and Denmark. *International Journal of Cancer*, **25**, 95-104.

Huang, J.Z. (2003). Local asymptotics for polynomial spline regression. *The Annals of Statistics*, **31(5)**, 1600-1635.

Hylleberg, S. (1992). The historical perspective. In *Modelling Seasonality* (S. Hylleberg, ed.), Oxford: Oxford University Press.

Hyndman, R. and Shahid, Md. U. (2007). Robust forecasting of mortality and fertility rates: a functional data approach. *Computational Statistics & Data Analysis*, **51**, 4942-4956.

James, G.M., Hastie, T.J., and Sugar, C.A. (2000). Principal component models for sparse functional data. *Biometrika*, **87(3)**, 587-602.

James, G.M. and Sugar, C.A. (2003). Clustering for sparsely sampled functional data. *Journal of the American Statistical Association*, **98(462)**, 397-408.

Jones, M.C., Marron, J.S., and Sheather, S.J. (1996). A brief survey of bandwidth selection for density estimation. *Journal of the American Statistical Association*, **91**, 401-407.

Kearns, M., Mansour, Y. and Ng, A.Y. (1998). An information-theoretic analysis of hard and soft assignment methods for clustering. *Learning in Graphical Models*, 495-520.

Kemeny, J.G. and Snell, J.L. (1960). *Finite Markov Chains*, Princeton, NJ: Nostrand.

Keselman, H.J. and Keselman, J.C. (1993). Analysis of repeated measurements. In L.K. Edwards (ed.) *Applied analysis of Variance in Behavioral Science*, New York: Marcel Dekker, 105-145.

Kim, C.J. and Nelson, C.R. (1998). Business cycle turning points, a new coincident index, and tests of duration dependence based on a dynamic factor model with regime switching. *Review of Economics and Statistics*, **80**,188-201.

Lam, C. and Yao, Q. (2012). Factor modeling for high-dimensional time series: inference for the number of factors. *The Annals of Statistics*, **40**, 694-726.

Lam, C., Yao, Q. and Bathia, N. (2011). Estimation of latent factors for high-dimensional time series. *Biometrika*, **98**, 901-918.

Lu, Z. (1998). On the ergodicity of non-linear autoregressive model with an autoregressive conditional heteroscedastic term. *Statistica Sinica*, **8**, 1205-1217.

Lu, Z., Tjøstheim, D., and Yao, Q. (2007). Adaptive varying-coefficient linear models for stochastic processes: asymptotic theory. *Statistica Sinica*, **17**, 177-198.

Lütkepohl, H. (1985). Comparison of criteria for estimating the order of a vector autoregressive process. *Journal of Time Series Analysis*, **6**, 35-52.

Lütkepohl, H. (2005), *New Introduction to Multiple Time Series Analysis*, Berlin: Springer.

Masry, E. and Tjøstheim, D. (1995). Nonparametric estimation and identification of nonlinear ARCH time series: Strong convergence and asymptotic normality. *Econometric Theory*, **11**, 258-289.

Masry, E. and Tjøstheim, D. (1997). Additive nonlinear ARX time series and projection estimates. *Econometric Theory*, **13**, 214-252.

Merikoski, J.K. and Kumar, R. (2004). Inequalities for spreads of matrix sums and products. *Applied Mathematics E-notes*, **4**, 150-159.

Meyne, S.P. and Tweedie, R.L. (2009). *Markov Chains and Stochastic Stability*, Cambridge University Press.

Nadaraya, E.A. (1964). On estimating regression. *Theory of Probability & Its Application*, **9**, 141-142.

Neuman, E. (1981). Moments and Fourier transforms of B-splines. *Journal of Computational and Mathematics*, **7(1)**, 51-62.

Newey, W.K. and West, K.D. (1987). A simple, positive-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica*, **55**, 703-708.

Pan, J. and Yao, Q., (2008). Modelling multiple time series via common factors. *Biometrika*, **95**, 365-379.

Peña, D. and Box, G.E.P. (1987). Identifying a simplifying structure in time series. *Journal of American Statistical Association*, **82**, 836-843.

Peña, D., Tiao, G.C., and Tsay, R.S. (2001). *A Course in Time Series Analysis*, New York: John Wiley & Sons.

Priestley, M.B., Rao, T.S. and Tong, H. (1974). Application of principal component analysis and factor analysis in the identification of multivariable systems. *Automatic Control , IEEE Transcations*, **19**, 730-734.

Quenouille, M.H. (1957). *The Analysis of Multiple Time Series*, London: Griffin.

Ramsay, J.O. and Silverman, B.W. (2005). *Functional Data Analysis*, New York: Springer.

Rao, B.P. (1999). *Semimartingales and Statistical Inference*, London: Chapman and Hall.

Ratcliffe, S.J., Leader, L.R., and Heller, G.Z. (2002a). Functional data analysis with application to periodically stimulated foetal heart rate data. I: Functional regression. *Statistics in Medicine*, **21**, 1103-1114.

Ratcliffe, S.J., Leader, L.R., and Heller, G.Z. (2002b). Functional data analysis with application to periodically stimulated foetal heart rate data. II: Functional logistic regression. *Statistics in Medicine*, **21**, 1115-1127.

Roberts, J. M. (1995). New Keynesian economics and the Phillips curve. *Journal of Money, Credit and Banking*, 975-984.

Roussas, G.G. (1989). Consistent regression estimation with fixed design points under dependence conditions", *Statistics and Probability Letters*, **8**, 41-50.

Roussas, G.G., Tran, L.T., and Ioannides, D.A. (1992). Fixed design regression for time series: Asymptotic normality. *Journal of Multivariate Analysis*, **40**, 262-291.

Ruppert, D. (1997). Empirical-bias bandwidths for local polynomial nonparametric regression and density estimation. *Journal of the American Statistical Association*, **92**, 1049-1062.

Ruppert,D., Sheather, S., and Wand, M.P. (1995). An effective bandwidth selector for local least squares regression. *Journal of the American Statistical Association*, **90(432)**, 1257-1270.

Schumaker, L. (1981). *Spline Functions: Basic Theory*, New York: John Wiley & Sons.

Shumway, R.H. (2000). Dynamic mixed models for irregularly observed time series. *Resenhas IME-USP*, **4**, 433-456.

Shumway, R.H. and Stoffer, D.S. (2000). *Time Series Analysis & Its Applications*, New York: Springer-Verlag.

Silverman, B.W. (1984). Spline smoothing: the equivalent variable kernel method. *The Annals of Statistics*, 898-916.

Sims, C.A. and Zha, T. (2006). Were there regime switches in US monetary policy? *The American Economic Review* , 54-81.

Stock, J.H. and Watson, M.W. (1989). New indexes of coincident and leading economic indicators. *National Bureau of Economic Research, Macroeconomics Annual*, **4**, 351-394.

Stock, J.H. and Watson, M.W. (2002). Macroeconomic forecasting using diffusion indices. *Journal of Business & Economic Statistics*, **20**, 147-162.

Stock, J.H. and Watson, M.W. (2005). Implications of dynamic factor models for VAR analysis. *National Bureau of Economic Research.*

Tao, M., Wang, Y., Yao, Q., and Zou, J. (2011). Large volatility matrix inference via combining low-frequency and high-frequency approaches. *Journal of the American Statistical Association*, **106**, 1025-1040.

Tiao, G.C. and Box, G.E.P. (1981). Modeling multiple time series with applications. *Journal of the American Statistical Association*, **76**, 802-816.

Tiao, G.C. and Tsay, R.S. (1983). Multiple time series modeling and extended sample cross-correlations. *Journal of Business &Economic Statistics*, **1(1)**, 43-56.

Tiao, G.C. and Tsay, R.S. (1989). Model specification in multivariate time series. *Journal of Royal Statistical Society. Series B*, 157-213.

Tjøstheim, D. (1994). Non-linear time series: a selective review. *Scandinavian Journal of Statistics*, **21**, 97-130.

Tong, H. (1983). *Threshold Models in Non-linear Time Series Analysis*, Lecture notes, Springer-Verlag.

Tong, H. (1990), *Nonlinear Time Series: A Dynamic System Approach*, Clarendon Press: Oxford, UK.

Tong, H. and Lim, K.S. (1980). Threshold autoregression, limit cycles and cyclical data. *Journal of Royal Statistical Society: Series B*, 245-292.

Tran, L., Roussas, G., Yakowitz, S., and Van, B.T. (1996). Fixed-design regression for linear time series. *The Annals of Statistics*, **24**, 975-991.

Tsay, R.S. (2010). *Analysis of Financial Time Series*, New Jersey: John Wiley & Sons.

Tsay, R.S. and Tiao, G.C. (1983). Identification of multiplicative ARMA models for seasonal time series. Technical report and research series No. 7; Chicago: University of Chicago Graduate School of Business.

Viterbi, A.J. (1967). Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *Information Theory, IEEE Transaction on*, **13**, 260-269.

Wahba, G. (1977). A survey of some smoothing problems and the method of generalized cross-validation for solving them. In *Applications of Statistics* (P.R. Krisnaiah, ed.), 507-523, Amsterdam, North Holland.

Watson, M.W. and Engle, R.F. (1983). Alternative algorithms for the estimation of dynamic factor, mimic and varying coefficient regression models. *Journal of Econometrics*, **23**, 385-400.

Watson, G.S. (1964). Smooth regression analysis. *Sankhyā: The Indian Journal of Statistics, Series A*, 359-372.

Whitle, P. (1951). Hypothesis testing in time series analysis (Vol. 4). Almqvist &Wiksells.

Withers, C.S. (1981). Conditions for linear processes to be strong mixing. *Zeitschrift fur Wahrs- cheinlichkeitstheorie verwandte Gebiete*, **57**, 477-480.

Wu, W.B. (2005). Nonlinear system theory: Another look at dependence. *Proceedings of the National Academy of Sciences of the United States of America*, **102(40)**, 14150-14154.

Xia, Y. and Li, W.K. (1999a). On the estimation and testing of functional-coefficient linear models. *Statistica Sinica*, **9**, 735-757.

Xia, Y. and Li, W.K. (1999b). On single-index coefficient regression models. *Journal of American Statistical Association*, **94**, 1275-1285.

Yang, L., Park, B., Xue, L., and Härdle, W. (2006). Estimation and testing for varying coefficients in additive models with marginal integration. *Journal of the American Statistical Association*, **101** 1212-1227.

Zhou, S., Shen, X., and Wolfe, D.A. (1998). Local asymptotics for regression splines and confidence regions. *The Annals of Statistics*, **26**, 1760-1782.

Zygmund, A. (2002). *Trigonometric series*, Cambridge University press.