Title

Estimation of Prediction Error and Stochastic Volatility Model for Functional Time Series

Author

Bhattacharjee, Samayita

Publication Date

2024

Peer reviewed|Thesis/dissertation

Estimation of Prediction Error and Stochastic Volatility Model for Functional Time Series

By

SAMAYITA BHATTACHARJEE

DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Statistics

in the

OFFICE OF GRADUATE STUDIES

of the

UNIVERSITY OF CALIFORNIA

DAVIS

Approved:

Alexander Aue - Chair

Prabir Burman - Co-chair

Debashis Paul

Committee in Charge

2024

To Sensei, Dr. Daisaku Ikeda (1928-2023), for instilling wisdom, courage and compassion.

Contents

Abstract

With the emergence of modern technology and availability of high frequency data, the study of functional time series has become popular in recent years. One of the main goals of time series is to predict the outcome of the future observations. Though methods of predicting a functional time series have been explored in the literature, not much work has been done to obtain prediction error estimates. The first part of the dissertation proposes several estimates of prediction errors in the functional time series context. Prediction errors are necessary inputs for construction of prediction bands. This dissertation introduces methods of getting prediction bands using the different functional prediction error estimates. The proposed methods are evaluated based on simulation studies as well as real data applications.

A second important application of functional time series can be found in high-frequency finance, viewing the intra-day price movements as functions. Financial time series are characterized by volatility clustering, implying that large price movements tend to be followed by further large price movements and small movements by small movements. The stochastic volatility model is a widely used multiplicative financial model originally introduced to capture this form of heteroscedastic behavior for univariate financial time series. Unlike the competitor GARCH models that depend on past volatility and past residuals, and also aim at capturing the clustering tendency, the stochastic volatility process depends on the product of an independent noise sequence with a latent volatility sequence. When the observations are scalars or vectors, stochastic volatility estimation is often performed within the state-space modeling framework. The second part of the dissertation introduces the functional stochastic volatility model along with a method to estimate the model parameters using the state-space modeling framework and evaluates the proposed methodology on simulated data.

Acknowledgments

My PhD journey has been a bit different than that of most of my peers, since I had to transition to academia after working in the industry for four years. It was undoubtedly very challenging to bridge that gap and gain momentum. There were numerous moments of self-doubts, anxiety and frustration, and yet, I survived those, as well as a pandemic, thanks to the presence of certain people in my life, whom I'd like to thank here.

First and foremost, I would like to express my deepest gratitude to my advisor, Professor Alexander Aue, whose unwavering support, insightful guidance, and encouragement have been invaluable throughout my PhD journey. His passion for time series methods and his excellent teaching have greatly inspired me. I am also grateful for him to have dialogue with me on multiple occasions when I struggled to have confidence.

I am also profoundly thankful to my co-advisor, Professor Prabir Burman, for his meticulous feedback and continuous support, which have significantly shaped my research. His involvement in my research after my third year had been monumental in ensuring a steady progress of both my projects, and I cannot thank him enough for his exceptional guidance. My sincere appreciation extends to Professor Debashis Paul, for his valuable insights and constant encouragement. Professor Paul had been a mentor right from the beginning and I am forever grateful for his support throughout this PhD journey, especially during COVID. I am also grateful to Professor Thomas Lee and Professor Aaron Smith for their support during my qualifying exam and defense presentation. I also wanted to extend my gratitude to Professor Jiming Jiang for kindly agreeing to be my presenter at the commencement ceremony.

I would like to extend my heartfelt thanks to the faculty and staff members of the Statistics Department, especially Andi Carr, Nehad Ismail and Olga Rodriguez, for their constant help and support, making my graduate life at UC Davis smoother.

I owe a great deal of gratitude to my 12th grade Statistics teacher, Kalyan Sir, for igniting my interest in the field of statistics. His teachings laid a strong foundation for my academic pursuits.

I am also very grateful to my professors from my undergraduate studies at St. Xavier's College and my Master's program at Indian Statistical Institute. Their excellent teaching and guidance have been instrumental in shaping my academic path.

A special thanks to my dear friends Sneha Chakraborty, Poorbita Kundu, Emily Chang, Russell Okino, Paromita Dubey, Abhishek Roy, Satarupa Guha, Yuanyuan Li, Xiaoliu Wu, Junwen Yao, Shreyan Ganguly,

Indranil Sahoo, Entejar Alam, Anushree Barjatya, Swati Kumari, and Rebeka Sengupta, whose companionship, support, and encouragement have been a source of great strength and joy throughout this journey. I'd also like to thank my Buddhist group of friends and mentors, especially Eleanor Calkin Roosevelt-Reis, Sedona Tuss, Stephanie Morgan and Nallamai Lakshmanan for encouraging me and being compassionate listeners.

Lastly, but most importantly, I would like to express my deepest appreciation to my parents, Sibnath and Sarmistha Bhattacharjee and my husband, Sunil Bandaru, for their unconditional love, support, and unwavering belief in me. They trusted my abilities even when I could not believe in myself. Their sacrifices and encouragement have been the groundwork of my success, and I am forever grateful for their presence in my life.

CHAPTER 1

Introduction

This dissertation covers research on some advanced topics in functional time series. It consists of two parts: the first part introduces a novel method for estimating the prediction error when forecasting the future (function) value of a functional time series. The second part introduces the functional stochastic volatility model which is an important quantity in the finance literature.

1.1. Introduction to estimation of prediction error for functional time series

Efron (2004) addressed the problem of estimation of prediction error in a signal plus independent and identically distributed noise setting. Suppose the following holds:

$$y = \mu + \epsilon$$

where ϵ represents i.i.d. noise and consider a model $m(\cdot)$ fitted to observations $\mathbf{y} = (y_1, \ldots, y_n)$ which produces the estimate $\hat{\boldsymbol{\mu}} = m(\mathbf{y}) = (\hat{\mu}_1, \ldots, \hat{\mu}_n)$. The prediction error tries to estimate how well $\hat{\boldsymbol{\mu}}$ will predict a future observation. Efron's paper considers squared error $\mathrm{err}_i = (y_i - \hat{\mu}_i)^2$ and then discusses ways of estimating the prediction error which is given by $\Sigma_{i=1}^n \mathrm{err}_i +$ covariance penalty. The covariance penalty can be estimated parametrically under a Gaussian distribution of $\mathbf{y}$ giving rise to Stein's Unbiased Risk Estimate (SURE) or under a general parametric model giving rise to Mallow's C_p estimate. Non-parametric estimation of the covariance penalty has also been discussed using bootstrap methods.

When it comes to time series data, this method of estimating prediction error fails because the observations in time series have serial correlation and are not independent. A simple approach is to consider the empirical estimate which attempts to get an estimate of $\Sigma_{i=1}^n \mathrm{err}_i$. More formally, let $X_1, \ldots, X_n$ be observations from a stationary univariate time series of the form

$$X_t = \mu_t + \epsilon_t$$

where μ_t is the conditional mean of X_t given the past, where the past is the sigma-algebra generated by $X_1, \ldots, X_{t-1}$ or the infinite past, and ϵ_t is mean zero, i.i.d. and independent of μ_t. Suppose a model $\{\mu_t(\theta)\}$ is fitted to the data. If $\hat{\theta}_s$ is the estimate of θ based on the first s observations, then the estimate of μ_t is given by $\mu_t(\hat{\theta}_s)$ and the residual is given by $\epsilon_t(\hat{\theta}_s) = X_t - \mu_t(\hat{\theta}_s)$. The variance for predicting X_{n+1} based on $X_1, \ldots, X_n$ is the prediction error given by

$$\mathrm{PE}_n = E[\epsilon_{n+1}(\hat{\theta}_n)]^2$$

This can be estimated empirically from the sample without the covariance penalties. When $n - k$ residuals are available, the empirical estimate is given by

$$\widehat{\mathrm{PE}}_n^{emp} = \frac{1}{n-k} \sum_{t=k}^{n-1} \epsilon_{t+1}^2(\hat{\theta}_n)$$

However, Efron (2004) pointed out that this estimate is not good enough, because it uses the residuals based on the dataset at hand but it does not measure how well $\mu_t(\hat{\theta}_s)$ estimates a future observation. Rissanen (1986), on the other hand, proposed an accumulated measure of errors which is given by Rissanen's Approximate Prediction Error (APE)

$$\widehat{\mathrm{PE}}_n^{R} = \frac{1}{n-m} \sum_{t=m}^{n-1} \epsilon_{t+1}^2(\hat{\theta}_t)$$

where $m = \lfloor \delta n \rfloor$ with $0 \leq \delta \leq 1$. Note that the parameters are re-estimated sequentially for each t in the sum. This estimate uses the given dataset both to estimate and validate the parameters after each observation is received. Since it computes the residuals for predicting observations at $t+1$ based on $X_1, \ldots, X_t$, this estimate is expected to perform better than the empirical estimate in terms of lowering the bias of the empirical estimate. However, APE is not always a good estimate of the prediction error. If m is small, the bias in estimating PE_n might be significant. This is because for smaller m (δ close to 0), θ is estimated based on a smaller history, so it is more biased. Whereas, when m is close to n (δ close to 1), the bias might be small because θ is estimated based on a longer history, but the variance of the estimate might be high because the APE estimator is obtained by summing over fewer observations.

Aue and Burman (2024) addressed these issues by designing modified versions of the empirical and modified empirical estimates of the prediction error for univariate and multivariate time series. Their proposed estimators are based on minimizing the expected bias $E[\mathrm{PE}_n - \widehat{\mathrm{PE}}_n]$ for the two estimators above.

For the empirical estimate, an estimate of the expected bias was set up as

$$C_n(w) = \frac{1}{n-m}\sum_{t=m}^{n-1} w_t \left(\epsilon_{t+1}^2(\hat{\theta}_t) - \widehat{\mathrm{PE}}_t^{emp}\right)$$

where the weights $\{w_t\}$ are so chosen that the expected bias is minimized. The modified empirical estimate was then given by

$$\widehat{\mathrm{PE}}_n^{ME}(w) = \widehat{\mathrm{PE}}_n^{emp} + C_n(w)$$

Similarly, Rissanen's estimate is modified by using weighted averages of $\epsilon_{t+1}^2(\hat{\theta}_t)$ instead of simple averages, and is given by

$$\widehat{\mathrm{PE}}_n^{MR}(v) = \frac{1}{n-m}\sum_{t=m}^{n-1} v_t \epsilon_{t+1}^2(\hat{\theta}_t)$$

where the weights $\{v_t\}$ are again chosen so that the bias can be minimized. More specifically, the weights are chosen by minimizing $\sum_{t=m}^{n-1} v_t^2$ subject to the constraints $f_0(v) = 1$ and $f_1(v) = 1$ where

$$f_k(v) = \frac{1}{n-m}\sum_{t=m}^{n-1} \left(\frac{n}{t}\right)^k v_t, \qquad k = 0, 1$$

and this term appears in the expected value of the Rissanen's estimate. Simulation studies were conducted based on different choices of weights for the modified estimates. These showed that the modified Rissanen's estimate had the smallest bias but the highest variance. The modified empirical estimate performed the best among all the prediction error estimates maintaining a balance between the bias and the variance in the sense that it is not significantly more biased than the other estimates but has the smallest variance.

With the emergence of modern technology and the availability of high frequency data, the study of functional time series has become popular in recent years. The evolution of intra-day pollution curves, for instance, is a classic example of a functional time series (see Figure (1.1) for reference).

Each curve $X_k(t)$ has a discrete time index k referencing the day it was recorded and a continuous time index t referencing intra-day time, rescaled to the unit interval $[0, 1]$. The temporal evolution across k gives a functional time series object $X_k(t)$ (see Section 4.7 for details). It is imperative now to not only develop methods to model functional time series but also to predict them. This is because the quantification of the uncertainty in prediction is important to assess the quality of the forecast. However, this problem is not trivial since functions are infinite-dimensional and model estimation and prediction requires the estimation of complex operators. Aue et al. (2015) provided an intuitive solution to this. They proposed that instead

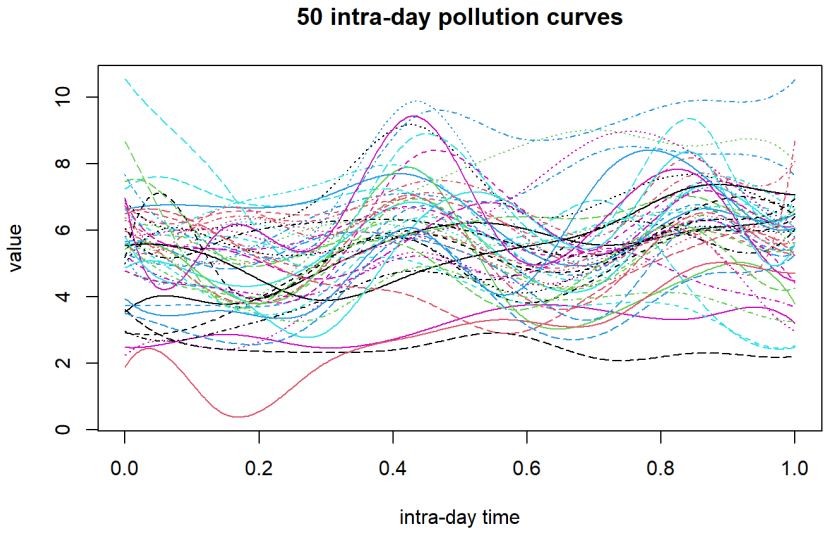

FIGURE 1.1. PM_{10} observations in Graz observed for 50 days

of dealing with infinite-dimensional functions, one can use the data to get the corresponding vectors of functional principal components scores. With this step, the data is transformed to a multivariate time series and any multivariate time series forecasting algorithm can be applied to get predictions. More formally and to illustrate, consider a Functional AutoRegressive (FAR) model of order 1 given by

$$X_k = \Psi(X_{k-1}) + \epsilon_k, \qquad k \in \mathbb{Z}$$

where $X_k = (X_k(t) \colon t \in [0,1])$ are functional time series observations and $\epsilon_k = (\epsilon_k(t) \colon t \in [0,1])$ are centered, independent and identically distributed innovations functions, Ψ is a bounded linear operator to ensure a causal and stationary solution, and $\mathbb{Z}$ is the set of integers. Bosq (2000) proposed the one-step ahead prediction of this series as $\tilde{X}_{n+1} = \tilde{\Psi} X_n$ where $\tilde{\Psi}$ is an estimator of Ψ. The alternative methodology proposed by Aue et al. (2015) uses the sample eigenfunctions to convert the infinite-dimensional functions to finite-dimensional scores. It then uses any multivariate model to predict the scores and the predicted functions can be obtained using the truncated Karhunen–Loève representation. The proposed algorithm is conceptually simple and is not bound by an assumed underlying FAR structure as used for the illustration above. The paper also shows that the one-step ahead predictors $\hat{X}_{n+1}$ from the above algorithm are asymptotically equivalent to the predictors $\tilde{X}_{n+1}$ as obtained from Bosq (2000) for an underlying $\mathrm{FAR}(p)$ process.

It also discussed how to select the optimal values of d, the reduced dimension, and the order p of the VAR model, based on a functional version of Akaike's final prediction error criterion.

Even though Aue et al. (2015) discuss how to get predictions for a functional time series, they did not provide details on quantifying the prediction error. The first part of this dissertation defines prediction errors for functional time series expanding on the ideas of the previous two papers. This work also defines the corresponding estimates. Prediction errors are used to determine which models are reasonable to fit to the data when the end goal is prediction of future observations. The estimation of prediction errors is also important when constructing prediction bands. This research not only provides the estimates of prediction error in the functional time series setting, but it also outlines the construction of point-wise prediction bands for the functions.

1.2. Introduction to functional stochastic volatility

Volatility, a conditional standard deviation, is an important quantity in finance. It is useful for capturing uncertainty in financial markets. Modeling volatility is important because it helps in forecasting the absolute magnitude of returns and such forecasts are useful in risk management, derivative pricing and hedging, trading strategies like market making and other financial activities (Engle and Patton (2000)).

Traditionally, volatility models are based on daily returns of an underlying asset. If P_t is the price of an asset at time t, then the relative daily returns are calculated as

$$y_t = \frac{P_t - P_{t-1}}{P_t}$$

To model volatility, Engle (1982) proposed AutoRegressive Conditional Heteroscedastic (ARCH) models. An ARCH(p) model is given by

$$y_t = w_t h_t^{1/2}$$

$$h_t = \alpha_0 + \alpha_1 y_{t-1}^2 + \cdots + \alpha_p y_{t-p}^2$$

where $\alpha_0 > 0, \alpha_i \geq 0$ for $i > 0$, and w_t is often assumed to be $N(0,1)$. Hence, if we denote by $\mathcal{F}_{t-1}$ the information set available until time $t-1$, then,

$$y_t | \mathcal{F}_{t-1} \sim N(0, h_t)$$

Here, the conditional standard deviation or volatility, h_t, depends on the past p returns. Bollerslev (1986) generalized this model by including both lagged returns and lagged volatility to explain the current volatility which was the Generalized ARCH or GARCH model. A GARCH(p, q) model is given by

$$y_t = w_t h_t^{1/2}$$

$$h_t = \alpha_0 + \sum_{i=1}^{p} \alpha_i y_{t-i}^2 + \sum_{j=1}^{q} \beta_j h_{t-j}$$

where $p \geq 0, q > 0$ and $\alpha_0 > 0, \alpha_i \geq 0, i = 1, \cdots, p$ and $\beta_j \geq 0, j = 1, \cdots q$. Both of these models are endogenous, meaning they depend on their past values. There are various subvariants of GARCH processes, some of which are discussed in Duan (1997) and Aue et al. (2006).

On the other hand, Taylor (1982) introduced an exogenous way of modeling volatility, given by what is known as a stochastic volatility model. In its simplest terms, a Stochastic Volatility (SV) model of order 1 is given by

$$y_t = e^{(\frac{1}{2}h_t)}\varepsilon_t$$

$$h_t = \phi h_{t-1} + \eta_t$$

where it is often assumed that $\varepsilon_t \sim$NID$(0, 1); \eta_t \sim$NID$(0, \sigma_\eta^2)$ for $t = 1, \ldots, n$ and $\{\varepsilon_t\}$ and $\{\eta_t\}$ are independent for all time points. Further, due to the log-volatility structure, positivity of the volatility is ensured in this model.

Both (G)ARCH and SV models as introduced above are based on univariate time series of daily returns of financial assets. With the advent of modern technology, the evolution of an asset price throughout the day can be recorded and stored. Each price curve can be assumed to be a function. When observed over different days, these curves form a functional time series. Hence, volatility curves can also be regarded as functions and hence modeling techniques are required to model functional volatility. Hörmann et al. (2013) and Aue et al. (2017) modeled functional volatility endogenously by proposing functional ARCH (fARCH) and functional GARCH (fGARCH) models, respectively. Jang et al. (2021) proposed a functional stochastic volatility model which was based on Bayesian estimation. The second part of the dissertation aims to propose a new method to estimate a functional version of the stochastic volatility model based on the state-space modeling framework which is an exogenous way of quantifying the volatility curves.

1.3. Dissertation outline

The outline of the dissertation is as follows. Chapter 2 introduces the notions of functional data and their important components. Chapter 3 introduces the important aspects of functional time series. Chapter 4 introduces the prediction error and its estimates for functional time series. This chapter also presents the results for simulation studies as well as applications to real data sets, namely the annual temperature profiles measured at different meteorological stations in Australia and daily pollution curves measured in Graz, Austria. Chapter 5 introduces structure and estimation methods for a functional stochastic volatility model. This chapter provides conditions for the existence of stationary solutions to the defining functional SV equations, sets up an estimation procedure for the model parameters, and evaluates the proposed estimation procedure on simulated data.

CHAPTER 2

Functional data and its properties

A functional observation is a realization of a typically smooth random object that takes values in an abstract function space. Functional data can arise in varied fields of study ranging from geophysics to intraday financial data and climatology. The following figure shows daily maximum temperatures at a meteorological station in Sydney Observatory Hill, Australia for three consecutive years. For each year, it can be thought that an underlying smooth function is driving the observed data points. Each such curve represents a single functional observation.

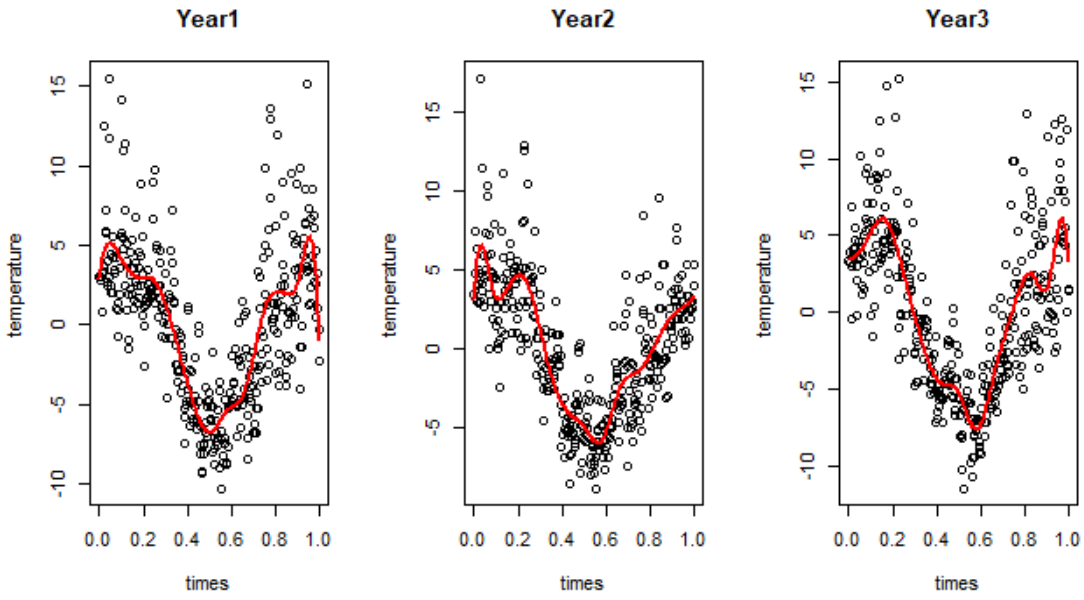

FIGURE 2.1. Example of functional data

Formally, a random element X is a functional variable if it takes values in a function space F, denoted by $X = (X(t) \colon t \in T)$ where, typically, the set T represents the unit interval $[0, 1]$. Notice that functional data is frequently observed in a time series context, where observations can be viewed as sampled from some underlying continuous "time" process. However, T may not always represent time, for example, $X(t)$

could be the concentration of a pollutant at altitude t. The argument of functions, t, can also be bivariate, for example, $X(t)$ could represent the gray level of an image at a spatial location $t \in T \subset \mathbb{R}^2$.

There are many classes of function spaces. For example, it can be $C[0,1]$, the space of continuous functions defined on the unit interval (see for example Dette et al. (2020)). Or it can be the more standard choice $L^2[0,1]$, the space of square-integrable functions on the unit interval. The convention here is to consider $F = L^2[0,1] = L^2$. This means that there is a probability space $(\Omega, \mathcal{A}, P)$ such that $X\colon \Omega \to L^2$ is $\mathcal{A}$-$\mathcal{B}$-measurable, where $\mathcal{B}$ is the Borel σ-algebra generated by the open sets in L^2. These technical aspects of functional data are suppressed in the following.

2.1. Functional observations

A collection of n functional observations $X_1, \ldots, X_n$ is called a functional data set. If the functions are independent and identically distributed, they are called a functional random sample. The observations are denoted by $X_k(t)$ which corresponds to the k-th function at "time" t. Even though functions are assumed to be continuously measured over t, practically, there are no continuous measurements. Hence, any realization of a functional observation X is observed at discrete points $t_1, \ldots, t_K$ only, giving rise to $X(t_1), \ldots, X(t_K)$ for some K. These discrete point measurements can be exact or contaminated with measurement errors. If the sampling frequency is low, sparse functional data is obtained. On the other hand, if the sampling frequency is high, Li and Hsing (2010) showed that densely sampled functional data gives the same theoretical results as in the idealized continuous measurement case. Hence, it is important to imagine a continuous time process $(X(t)\colon t \in [0,1])$ in the background and it is possible to recover $X(t)$ from $(X(t_k)\colon k = 1, \ldots, K)$.

One might wonder about the difference between functional and multivariate data. Multivariate data are considered as concurrently recorded observations involving more than one (type of) measurement, so inherently they are discrete in nature. On the other hand, functions are typically observations of the same variable but over a continuous "time" span. This, however, does not mean the observations are recorded for every value of t, because that would result in storing an uncountable number of values. Rather, it is assumed that there exists a function X that gives rise to the observed data. They are often assumed to be continuous and differentiable which leads to the "smoothness" of the observations. This means that neighboring discrete observations tend to be highly correlated which is a major difference between functional and multivariate data. Functions also allow for the use of derivatives, a concept which is not available for multivariate data.

The other aspect where functional data differ from multivariate data is the sampling scheme. For multivariate data, the data is observed at equidistant time intervals, otherwise it might lead to missing data related problems for standard time series methodology. However, for functional (time series) data, irregular observation times are allowed. Overall, the usual multivariate approach ignores the information about the smooth functional nature of the underlying data generating process. Functional data analysis on the other hand expresses discrete observations from time series in functional form, representing the entire measured function as a single observation. This representation aids in effective noise reduction of data through curve smoothing and also has applications when observations are not recorded at regular time intervals. Functional data analysis also helps to study important patterns and sources of variation in the data and develop appropriate inference procedures.

2.2. Representation and smoothing

Suppose the interest is in analyzing the functional realization $x = (x(t)\colon t \in [0,1])$. However, x is not observable, but observations are only available for a noisy discrete version $y(t_k), k = 1, \ldots, K$, where $t_1, \ldots, t_K \in [0,1]$. Thus, the following model is postulated:

$$y(t_k) = x(t_k) + e(t_k), \quad k = 1, \ldots, K$$

The measurement errors $e(t_1), \ldots, e(t_k)$ can be zero. Recovering functional realizations $x_1, \ldots, x_n$ from the discrete observations are done on an individual basis. To aid readability, from now on, we will represent both the function X and its realization x as X. It will be clear from the context if X represents the function or its realization.

There are some simple methods that might work in transforming discrete observations to functions. The simplest of them is linear interpolation, which is joining adjacent points using straight lines. This method can be useful if there is no measurement error because that guarantees that the observed values are true. However, they are not differentiable at the sampling times, hence they do not lead to smooth functions and retain a lot of redundant information if the underlying data generating process is smooth. One might try to use polynomial interpolation instead but that usually results in oscillatory functions.

The most popular approach is to represent functions by basis functions. A basis function system in the function space F is a set of known functions ϕ_l which are orthogonal. Further, a linear combination of a

sufficiently large number L of these functions should be able to approximate any function defined in F. In other words, basis function procedures represent a function x by a linear expansion

$$X(t) \approx \sum_{l=1}^{L} c_l \phi_l(t)$$

for L known basis functions. If the norm on F is denoted by $\|\cdot\|_F$, then the idea is to make the norm-difference between the function and its approximation using basis functions as close to zero as for sufficiently large L. Thus, the basis is so chosen that

$$\|X - \sum_{l=1}^{L} c_{l,L} \phi_l\|_F \to 0 \quad (L \to \infty)$$

Notice that the coefficients $c_{l,L}$ are also a function of L, the number of basis functions used to represent the function X, hence, it is represented by double indices. However, for a fixed L, they can just be written as c_l. The idea is to fix L and then minimize

$$\sum_{k=1}^{K} \left(X(t_k) - \sum_{l=1}^{L} c_l \phi_l(t_k) \right)^2$$

with respect to the coefficients $c_1, \ldots, c_L$. If $K = L$, then perfect fit is achieved. So, the goal is to have $L << K$. It is also desirable that the basis functions are so chosen that they represent the characteristics of the data. This will ensure that a sufficiently small L can reasonably well approximate the functional observation $X(t_1), \ldots, X(t_K)$. It also ensures that the coefficients c_l are easier to interpret and faster to compute. However, it should be kept in mind that L is not pre-specified but chosen based on the data. If the data is highly variable, a larger L might be needed in order to capture the entire information through the basis system.

There are different choices for the set of basis functions, the simplest being the collection of monomials that are used to construct power series,

$$1, t, t^2, t^3, \ldots, t^k, \ldots$$

Other popular choices of bases include Fourier and B-Spline basis.

2.2.1. Fourier basis. A Fourier basis is often chosen to represent functional data when the data shows some periodicity. They are based on trigonometric functions given by

$$\psi_{2l}(t) = \cos(2\pi lt) \quad \text{and} \quad \psi_{2l+1}(t) = \sin(2\pi lt), \quad l \in \mathbb{N}_0$$

where $\mathbb{N}_0$ is the set of non-negative integers. Notice that $(\psi_l \colon l \in \mathbb{N}_0)$ form an orthogonal basis, however they are not orthonormal. These can be easily made orthonormal by defining $(\phi_l \colon l \in \mathbb{N})$ where

$$\phi_0 = 1, \quad \text{and} \quad \phi_l = \sqrt{2}\psi_l, \quad l \in \mathbb{N}_0$$

The Fast Fourier Transform (FFT) makes it possible to compute the coefficients very efficiently when the time points are equally spaced. The Fourier basis system is very useful for stable functions where there are no strong local features and when the curvature of the functions remains consistent across the domain of the function (Ramsay and Silverman (2005)).

2.2.2. B-Spline basis. The B-Spline basis is known for its computational efficiency and hence it is perhaps the most popular method of approximating non-periodic functional data (Ramsay and Silverman (2005)). Splines are polynomials of specified order, defined over sub-intervals, which are created by breaking the entire interval over which the function is defined into sub-parts. Adjacent polynomials join up smoothly at the breakpoints, so that the function values are constrained to be equal at the junctions. A B-Spline basis is based on convex combinations of spline functions of specific order. The degree of smoothness of the splines at the knots or breakpoints can be chosen based on the data. This system of basis is flexible and can be applied to a variety of data.

2.3. Estimation of the basis coefficients

Previously, we have seen that using the basis representation, any function x can be written as an approximate linear combination of basis functions as

$$X(t) = \sum_{l=1}^{L} c_l \phi_l(t) = \mathbf{c}'\boldsymbol{\phi}$$

where $\mathbf{c} = (c_1, \ldots, c_L)^T$ and $\boldsymbol{\phi} = (\phi_1(t), \ldots, \phi_L(t))^T$. There are multiple ways to compute the coefficients $\{c_l\}$, some of them being listed here. The easiest way to do it is to get the ordinary least squares fit, obtained

from minimizing the least square criterion, given by

$$SSE(\mathbf{y}|\mathbf{c}) = \sum_{k=1}^{K} \left(y(t_k) - \sum_{l=1}^{L} c_l \phi_l(t_k) \right)^2$$

which can be written in matrix notation as

$$SSE(\mathbf{y}|\mathbf{c}) = (\mathbf{y} - \mathbf{\Phi c})'(\mathbf{y} - \mathbf{\Phi c})$$

where $\mathbf{\Phi}$ is the $K \times L$ matrix containing the values $\phi_l(t_k)$, that is $\mathbf{\Phi} = (\Phi_1 : \cdots : \Phi_L)$ with $\Phi_l = (\phi_l(t_1), \ldots, \phi_l(t_K))^T$. The value of $\mathbf{c}$ that minimizes the above SSE is given by

$$\mathbf{c} = (\mathbf{\Phi}'\mathbf{\Phi})^{-1}\mathbf{\Phi}'\mathbf{y}$$

So, the above describes an easy way to represent discrete observations as functions. Next we will look into a few basic objects pertaining to functional data.

2.4. Basic objects

So far, we have stated a model for discrete and noisy observation of a functional variable X and covered techniques to estimate a realization of X. Now, it is assumed that the function X can be observed directly. It is assumed that the underlying function space is the Hilbert space L^2, the details of which are now provided (Gohberg et al. (2013)).

2.4.1. Hilbert space. Before defining what a Hilbert space is, we introduce what an inner product space is. A vector space H is an inner product space such that there is a real number $\langle x, y \rangle$ satisfying the following:

- $\langle x, y \rangle = \langle y, x \rangle$
- $\langle \alpha x + \beta y, z \rangle = \alpha \langle x, z \rangle + \beta \langle y, z \rangle$
- $\langle x, x \rangle \geq 0$ and $\langle x, x \rangle = 0$ iff $x = 0$

Note that here x, y and z are not random. This space induces a norm $\|x\| = \sqrt{\langle x, x \rangle}$. Both the norm and the inner product are continuous. This means that if there are elements $(x_n : n \in \mathbb{N})$ and $(y_n : n \in \mathbb{N})$ such that $\|x_n - x\| \to 0$ and $\|y_n - y\| \to 0$, then $\|x_n\| \to \|x\|$ and $\langle x_n, y_n \rangle \to \langle x, y \rangle$. There can be

different types of norms defined for inner product spaces. For example, H can be $\mathbb{R}^d$ with the Euclidean norm $\langle x, y\rangle = x_1y_1 + \cdots + x_dy_d$. When $H = L^2[0,1]$, the L^2-norm is defined as $\langle x, y\rangle = \int_0^1 x(t)y(t)dt$.

If $(x_n \colon n \in \mathbb{N})$ is an element in the inner product space H, then x_n converges in norm to x if $\|x_n - x\| \to 0$. Further, x_n is a Cauchy sequence, if $\|x_n - x_m\| \to 0$ as $n, m \to 0$. A Hilbert space is an inner product space H such that every Cauchy sequence converges in norm to a limit in H. Notice that, $\mathbb{R}^d$ and $L^2[0,1]$ are both examples of Hilbert spaces. A linear subspace S of H is closed if every converging sequence in S has its limit in S.

Projection Theorem: Let S be a closed subspace of Hilbert space H and $y \in H$. Then there is unique $\hat{y} \in S$ such that

$$\|y - \hat{y}\| \leq \|y - s\|$$

for all $s \in S$. Then $\hat{y}$ is called projection of y onto S. $\hat{y}$ is characterized by $\langle y - \hat{y}, s\rangle = 0$ for all $s \in S$.

Let $e_1, \ldots, e_n$ be elements of Hilbert space H such that $\|e_i\| = 1$ and $\langle e_i, e_j\rangle = 0$ for $i \neq j$, then $e_1, \ldots, e_n$ are called orthonormal elements of H. If further S is the space spanned by $e_1, \ldots, e_n$, that is $S = \overline{\text{sp}}\{e_1, \ldots, e_n\}$, then the projection of $y \in H$ onto S is given by

$$\hat{y} = \hat{\alpha}_1 e_1 + \cdots + \hat{\alpha}_n e_n$$

where $\hat{\alpha}_i = \langle y, e_i\rangle$ are the Fourier coefficients.

A set $S = (e_i \colon i \in I)$ is an orthonormal set if e_is are orthonormal, that is $\|e_i\| = 1$ and $\langle e_i, e_j\rangle = 0$ for $i \neq j, i, j \in I$. An orthonormal set S is the basis of a Hilbert space H if $H = \overline{\text{sp}}\, S$. A Hilbert space H is separable if it has a countable orthonormal basis. Next comes an important result related to a separable Hilbert space. If H is a separable Hilbert space with basis $(e_i : i \in \mathbb{N})$, then $x = \sum_{i=1}^{\infty} \langle x, e_i\rangle e_i$ for all $x \in H$, that is,

$$\|x - \sum_{i=1}^{N} \langle x, e_i\rangle e_i\| \to 0 \quad (N \to \infty)$$

2.4.2. Mean function. Suppose X is a random element defined on common probability space $(\Omega, \mathcal{A}, P)$ taking values in arbitrary separable Hilbert space H assumed to be $\mathcal{A}$-$\mathcal{B}(H)$-measurable, where $\mathcal{B}(H)$ is Borel σ-algebra in H.

The mean function μ in L^2 is defined as $\mu = E[X] = ((E[X])(t) \colon t \in [0,1]) = (E[X(t)] \colon t \in [0,1])$. For any Hilbert space H and $X \in H$, X is weakly integrable if there exists $\mu \in H$ such that $E[\langle X, y\rangle] =$

$\langle \mu, y \rangle$ for all $y \in H$, and μ is the expectation (mean function) of X. In general, it is said that $X \in L_H^p$ if $E[\|X\|^p] < \infty$. So, if $X \in L_H^1$ and $H = L^2$, then $(E[X])(t) = E[X(t)]$ almost everywhere on [0,1]. It is natural to define the sample mean function based on observations $X_1, \ldots, X_n$ as $\hat{\mu}_n(t) = \frac{1}{n} \sum_{k=1}^n X_k(t)$.

2.4.3. Covariance operator. Let $X \in L_H^2$. The covariance operator $C \colon H \to H$ is defined as

$$C(y) = E[\langle X - E[X], y \rangle (X - E[X])], \quad y \in H$$

If $H = L^2$, then C is a kernel operator given by

$$C(y)(t) = \int_0^1 c(t,s) y(s) ds, \qquad y \in L^2$$

where $c(t,s)$ is the covariance kernel given by

$$c(t,s) = E[\{X(t) - \mu(t)\}\{X(s) - \mu(s)\}]$$

If $\mu(t) = 0$ for all $t \in [0,1]$, then $c(t,s)$ simplifies to $c(t,s) = E[X(t)X(s)]$ and $C(y) = E[\langle X, y \rangle, X]$. The covariance kernel is symmetric, positive definite and describes all cross-covariances of the random function $X \in L^2$. The operator C is symmetric, because $\langle C(y), x \rangle = \langle y, C(x) \rangle$ for all $x, y \in H$. It is also non-negative definite because $\langle C(y), y \rangle \geq 0$ for all $y \in H$. Both properties follow from the fact that $\langle C(y), x \rangle = \langle E[\langle X, y \rangle], x \rangle = E[\langle \langle X, y \rangle X, x \rangle] = E[\langle X, y \rangle \langle X, x \rangle]$. The sample covariance operator $\hat{C}_n \colon L^2 \to L^2$ is given by

$$\hat{C}_n(y) = \frac{1}{n} \sum_{k=1}^n \langle X_k - \hat{\mu}_n, y \rangle (X_k - \hat{\mu}_n) = \int_0^1 \hat{c}_n(., s) y(s) ds, \; y \in L^2$$

where the sample covariance kernel is given by

$$\hat{c}_n(t,s) = \frac{1}{n} \sum_{k=1}^n \{X_k(t) - \hat{\mu}_n(t)\}\{X_k(s) - \hat{\mu}_n(s)\}$$

2.5. Functional Principal Components

Functions are in principle infinite-dimensional objects. Hence, it is necessary to develop methods to reduce dimensionality. Functional Principal Component Analysis (FPCA) is the most important of these methods. Suppose H is a separable Hilbert space and $X \in L_H^2$ with $E[X] = 0$. Let $(\lambda_l \colon l \in \mathbb{N})$ be the decreasing eigenvalues of the covariance operator C of X and $(\phi_l \colon l \in \mathbb{N})$ be the corresponding orthonormal

eigenfunctions that is $\|\phi_l\| = 1$. Then, $\xi_l = \langle X, \phi_l \rangle$ is defined as the lth functional principal component score of X. When $H = L^2[0,1]$, the covariance operator admits the spectral decomposition as

$$C(y) = \sum_{l=1}^{\infty} \lambda_l \langle y, \phi_l \rangle \phi_l, \quad y \in L^2$$

Notice that $C(\phi_l) = \lambda_l \phi_l$. Then, any function X in L^2 allows for the Karhunen–Loève representation

$$X = \sum_{l=1}^{\infty} \langle X, \phi_l \rangle \phi_l$$

Similar to multivariate case, $\text{Cov}(\langle X, \phi_l \rangle, \langle X, \phi_{l'} \rangle) = E[\langle \langle X, \phi_{l'} \rangle X, \phi_l \rangle] = \langle C(\phi_{l'}), \phi_l \rangle$ which is equal to λ_l if $l = l'$ or 0 if $l \neq l'$. Notice that ϕ_1 can be obtained from solving

$$\phi_1 = \arg\max \{ \text{Var} \langle X, y \rangle : y \in H, \|y\| = 1 \}$$

Similarly, ϕ_l can be obtained from solving

$$\phi_l = \arg\max \{ \text{Var} \langle X, y \rangle : y \in H, \|y\| = 1, y \perp \phi_1, \ldots, \phi_{l-1} \}$$

In practice, C is unknown, so are its eigenvalues λ_l and eigenfunctions ϕ_l. So, we need an estimator $\hat{C}$ of the covariance operator C. The natural choice for $\hat{C}$ is given by

$$\hat{C}_n(y) = \frac{1}{n} \sum_{k=1}^{n} \langle X_k, y \rangle X_k$$

From this estimate, sample eigenvalues $\hat{\lambda}_l$ and sample eigenfunctions $\hat{\phi}_l$ are computed which serve as proxies for λ_l and ϕ_l. Notice that $\hat{C}_n$ has at most n non-zero eigenvalues, so only a limited number of eigenvalues and eigenfunctions can be estimated from a sample of size n.

The next question arises: how to compute the eigendecomposition of an arbitrary symmetric operator? To answer this, let us assume that the functional observations have the form

$$X_k(t) = \sum_{l=1}^{L} d_{k,l} \phi_l(t)$$

where we note that $(\phi_l : l \in \mathbb{N})$, the eigenfunctions also form an orthonormal basis system. Let us denote $\mathbf{X} = (X_1, \ldots, X_n)^T, \phi = (\phi_1, \ldots, \phi_L)^T$ and $D = (d_{k,l}) \in \mathbb{R}^{n \times L}$. Then, we can write $\mathbf{X} = D\phi$. Also,

$$\hat{c}_n(t,s) = \frac{1}{n}\mathbf{X}^T(t)\mathbf{X}(s) = \frac{1}{n}\phi^T(t)D^T D\phi(s)$$

Now, eigenfunctions of $\hat{C}_n$ must be in the span of $\phi_1, \ldots, \phi_L$. Let η be an eigenfunction of $\hat{C}_n$ with eigenvalue ρ. Then we can write $\eta(s) = \phi^T(s)\mathbf{b}$ for some $\mathbf{b} \in \mathbb{R}^L$. Following this, we have,

$$\begin{aligned}
\hat{C}_n(\eta)(t) &= \int_0^1 \hat{c}_n(t,s)\eta(s)ds \\
&= \int_0^1 \frac{1}{n}\phi^T(t)D^T D\phi(s)\phi^T(s)\mathbf{b}ds \\
&= \frac{1}{n}\phi^T(t)D^T D\mathbf{b} \\
&= \rho\eta(t) \\
&= \rho\phi^T(t)\mathbf{b}
\end{aligned}$$

noting that $\int_0^1 \phi(s)\phi^T(s)ds = I$ by doing component wise integration. This leads to the matrix eigenvalue problem

$$\left(\frac{1}{n}D^T D\right)\mathbf{b} = \rho\mathbf{b}$$

Thus, we see that getting the eigenvalues of the sample covariance operator boils down to getting eigenvalues of a matrix in the multivariate domain.

CHAPTER 3

Functional Time Series

Until now, functional data $X_k(t)$ was discussed in a general context, where,the "intra-day" argument t need not necessarily be time. When functions are indexed in discrete time k, they represent functional time series. The study of univariate and multivariate linear time series has been done extensively, with the availability of extensive theory of ARMA models, its extensions and ready-to-use computer packages. However, when observations are functions, there is an increased complexity as functions are infinite-dimensional, and the available theory and tools are more limited. In the literature, the focus was originally on special cases like first-order functional Auto-Regressive (FAR(1)) models. For the FAR(1) model and for other models of greater complexity, dimension reduction techniques are utilized through FPCA and using the auxiliary FPC score vector time series for modeling using multivariate time series approach.

Linear dependence is the most important concept in univariate and multivariate time series. In the functional context, this dependence is captured by autocovariance operators $C_h(\cdot) = E[\langle X_0, \cdot \rangle X_h], h \in \mathbb{Z}$. However, for $h \neq 0$, these are more complicated objects because they are not symmetric. Some common linear functional time series models are listed below:

- Functional linear process:
$$X_k = \sum_{j=0}^{\infty} \Psi_j \epsilon_{k-j}$$
- Functional moving average process:
$$X_k = \sum_{j=0}^{q} \Theta_j \epsilon_{k-j}$$
- Functional autoregressive process:
$$X_k = \sum_{j=1}^{p} \Phi_j X_{k-j} + \epsilon_k$$

Instead of dealing with prediction algorithms directly on the operator level, an easier way is to project the functions onto their principal components and use the FPC scores for predictions, because vectors of FPC scores will constitute a multivariate time series. But what happens to the linear dependence after projection? Let us illustrate the effect with an example. Suppose we have a first-order functional autoregression model $X_k = \Phi X_{k-1} + \epsilon_k$ with

$$\Phi(x) = a\left(\langle x, \phi_1\rangle + \langle x, \phi_2\rangle\right)\phi_1 + a\langle x, \phi_1\rangle\phi_2, \quad x \in H$$

where $a \in (0, 1)$ and $\phi_1, \phi_2 \in H$ are orthonormal basis functions. Further, assume that $E[\langle \epsilon_k, \phi_1\rangle^2] > 0$ but $E[\langle \epsilon_k, \phi_2\rangle^2] = 0$. Then it can be shown that the first FPC score time series satisfies

$$\langle X_k, \phi_1\rangle = a\langle X_{k-1}, \phi_1\rangle + a^2\langle X_{k-2}, \phi_1\rangle + \langle \epsilon_k, \phi_1\rangle$$

So, if we denote our new time series as $\xi_k = \langle X_k, \phi_1\rangle$ and $e_k = \langle \epsilon_k, \phi_1\rangle$, then it satisfies

$$\xi_k = a\xi_{k-1} + a^2\xi_{k-2} + e_k$$

This shows that the projection of this FAR(1) process is a VAR(2) process. In general, if we assume that all eigenvalues of C_ϵ, the covariance operator of the innovation functions, are positive, then the following relations hold between functional to vector time series dynamics:

- Projection of FMA(q) is VMA(q') with $q' \leq q$
- Projection of FAR(p) is in general not VAR(p') nor FMA(q')
- Projection of FARMA(p, q) is in general not VARMA(p', q')

However, invertibility is preserved under projections onto FPCs.

Now, predictions in functional time series typically rely on estimation of the covariance operator of the observations in the first step. The most often applied functional time series model is the FAR(1) model given by

$$X_k = \Psi(X_{k-1}) + \epsilon_k, \qquad k \in \mathbb{Z}$$

where $X_k = X_k(t)$ are functional time series observations and $\epsilon_k = \epsilon_k(t)$ are centered, independent and identically distributed innovations functions, Ψ is a bounded linear operator satisfying $\|\Psi\|_{\mathcal{L}} < 1$ to ensure

a causal and stationary solution, where the operator norm $\|\cdot\|_{\mathcal{L}}$ for any operator A is given by

$$\|A\|_{\mathcal{L}} = \sup_{\|x\|\leq 1} \|A(x)\|$$

For an appropriately chosen d, X_k can be approximated as $X_k = \sum_{l=1}^{d} \langle X_k, \hat{\phi}_l \rangle \hat{\phi}_l$, where, $\hat{\lambda}_1, \ldots, \hat{\lambda}_d$ are the first d sample eigenvalues and $\hat{\phi}_1, \ldots, \hat{\phi}_d$ are the corresponding sample eigenfunctions of the sample covariance operator. In that case the estimator of Ψ is given by

$$\tilde{\Psi}_n(y) \approx \frac{1}{n-1} \sum_{k=2}^{n} \sum_{l=1}^{d} \sum_{l'=1}^{d} \hat{\lambda}_l^{-1} \langle y, \hat{\phi}_l \rangle \langle X_{k-1}, \hat{\phi}_l \rangle \langle X_k, \hat{\phi}_{l'} \rangle \hat{\phi}_{l'}$$

This leads to the functional predictor $\tilde{X}_{n+1} = \tilde{\Psi} X_n$ introduced in Bosq (2000). Beyond FAR(1), higher order FAR processes can be studied, however, that involves estimation of a number of operators which can be a complex process. On the other hand, Aue et al. (2015) proposed an alternative algorithm to get predictions using methods based on FPC scores. This prediction technique utilizes univariate and multivariate prediction methods and avoids estimating operators of functional time series directly. The three-step algorithm is as follows:

- **Step 1:** Fix d and denote the empirical FPCs as $x_{k,l}^e = \langle X_k, \hat{\phi}_l \rangle$ where $\hat{\phi}_l$ represents the sample eigenfunctions, $l = 1, \ldots, d$. For $k = 1, \ldots, n$, use the data $X_1, X_2, \ldots, X_n$ to compute the vectors containing the first d FPC scores

$$\mathbf{X}_k^e = (x_{k,1}^e, \ldots, x_{k,d}^e)'$$

- **Step 2:** Fix h. Use $\mathbf{X}_1^e, \ldots, \mathbf{X}_n^e$ to determine the h-step ahead prediction for $\mathbf{X}_{n+h}^e$ with an appropriate multivariate algorithm:

$$\hat{\mathbf{X}}_{n+h}^e = (\hat{x}_{n+h,1}^e, \ldots, \hat{x}_{n+h,d}^e)'$$

- **Step 3:** Multivariate predictions are retransformed to functional object using the truncated Karhunen–

  Loève representation

$$\hat{X}_{n+h} = \hat{x}_{n+h,1}^e \hat{\phi}_1 + \cdots + \hat{x}_{n+h,d}^e \hat{\phi}_d$$

This algorithm gives the best linear predictor (in mean square sense) of the population FPC scores. Further, it does not assume any underlying FAR(p) structure or any other functional time series specification. The method is also flexible and any standard prediction algorithms like Durbin–Levinson or innovations algorithms can be applied to get the predicted FPC scores. One can even explore alternative prediction algorithms such as exponential smoothing and non-parametric predictions or can even incorporate covariates in the prediction process. The accuracy and validity of the estimators obtained using FPC scores can be summarized in the following. Let $(X_k \colon k \in \mathbb{Z})$ be an FAR(p) process and denote by $\hat{X}_{n+1}$ the FPC score one-step predictor and by $\tilde{X}_{n+1}$ the standard one-step ahead predictor. Assume that a VAR(p) model is fit to $\mathbf{X}_k^e = (x_{k,1}^e, \ldots, x_{k,d}^e)'$ by means of ordinary least squares, where $x_{k,l}^e = \langle X_k, \hat{\phi}_l \rangle$. Then, the resulting predictor is asymptotically equivalent to the standard predictor:

$$\|\hat{X}_{n+1} - \tilde{X}_{n+1}\| = O_P\left(\frac{1}{\sqrt{n}}\right)$$

Even though there has been some study on predicting a functional time series(such as Besse et al. (2000); Antoniadis et al. (2006); Kargin and Onatski (2008); Hyndman and Shang (2009); Aue et al. (2015); Jiao et al. (2023)), not much work has been done in estimating the prediction error. Studying prediction error is important to assess how good the prediction is and also to provide a confidence statement regarding the predictions. Further, in case the predictions are done using FPC scores, the prediction error can also give some idea of the number of principal components required to transform the infinite-dimensional functions to a finite-dimensional multivariate object. Chapter 4 aims to develop a methodology to estimate the prediction error that builds on Aue and Burman (2024) in the functional time series context.

CHAPTER 4

Prediction error in functional time series

As discussed in Chapter 1, the main idea of this chapter is to devise a strategy for the estimation of prediction errors in functional time series data. The estimation entails both classical and modified estimates extending the prediction issues for univariate and multivariate time series as addressed in Aue and Burman (2024). The idea is to convert infinite-dimensional functions to finite-dimensional multivariate objects using the idea used in Aue et al. (2015) because it is not possible to estimate infinite-dimensional objects nonparametrically without reducing the dimensions.

Suppose we have n functions, $X_k, k = 1, \ldots, n$, that have been obtained from a functional stationary time series of the form

$$X_k = \mu_k + \epsilon_k$$

where μ_k is the conditional mean function of X_k given the past and ϵ_k are the innovation functions with zero mean, independent of μ_k. One can fit a functional time series model, such as an FAR process of order p to this data to get the estimate of μ_k and thus the estimates of $\epsilon_k(t)$ that can be used to get the functional prediction error estimates. However, this method is complex since it deals with infinite-dimensional functions. An easier way is to get the FPC scores which are finite-dimensional. Thus, for a fixed D, we compute the FPC scores $\langle X_k, \hat{\phi}_j \rangle$ where $\hat{\phi}_1, \ldots, \hat{\phi}_D$ are the estimated eigenfunctions. The proportion of variance explained by the FPCs can be computed for each $j, j = 1, \ldots, D$, and an appropriate $d << D$ can be chosen that represents the multivariate data of the d-dimensional scores

$$(\langle X_k, \hat{\phi}_1 \rangle, \ldots, \langle X_k, \hat{\phi}_d \rangle)^T = (X^*_{k,1}, \ldots, X^*_{k,d})^T$$

Now, the new observations can be written in the form

$$X^*_k = \mu^*_k + \epsilon^*_k$$

Suppose a multivariate time series model $\{\mu_k^*(\theta)\}$ is employed to estimate $\{\mu_k^*\}$, where θ is the vector of all the model parameters. For example a VAR of order 2 can be considered for μ_k^* given by $\mu_k^*(\theta) = \alpha_1 X_{k-1}^* + \alpha_2 X_{n-2}^*, \alpha_1$ and α_2 being $d \times d$ matrices. It is to be noted that the underlying data generating process might be different from the one fitted, so when it comes to prediction error, there are two sources. Essentially, it is a bias-variance tradeoff, where the bias comes from model-misspecification and the variance from the estimation procedure. If $\hat{\theta}_s$ is the estimate of θ based on the first s observations, then the estimate of μ_k^* is denoted by $\mu_k^*(\hat{\theta}_s)$ and the corresponding residual is denoted by

$$\epsilon_k^*(\hat{\theta}_s) = X_k^* - \mu_k^*(\hat{\theta}_s)$$

These residuals can now be used to estimate functional prediction error. The following section introduces some notations and describes the above in greater formality.

4.1. Notations

All stationary functional time series that are stationary and in L^2 allow for the Karhunen–Loève (KL) representation

$$X_k = \sum_{j=1}^{\infty} \langle X_k, \phi_j \rangle \phi_j, \qquad k = 1, \ldots, n$$

where $(\phi_j : j \in \mathbb{N})$ represents the underlying basis system of eigenfunctions of the covariance operator associated with the functional time series.

For a fixed d, define $\langle X_k, \phi_j \rangle = X_k(\phi_j)$, $X_k(\phi_{1:d}) = (\langle X_k, \phi_1 \rangle, \ldots, \langle X_k, \phi_d \rangle)^T$ and $\phi_{1:d} = (\phi_1(t), \ldots, \phi_d)^T$. Then, we can write the KL representation as

$$\begin{aligned} X_k &= \sum_{j=1}^{d} \langle X_k, \phi_j \rangle \phi_j + \sum_{j=d+1}^{\infty} \langle X_k, \phi_j \rangle \phi_j \\ &= X_k(\phi_{1:d})' \phi_{1:d} + X_k(\phi_{d+1:\infty})' \phi_{d+1:\infty} \end{aligned}$$

Now, we can fit a VAR model of order p to the d-dimensional scores $X_k(\phi_{1:d})$. Suppose, we fit a $\mathrm{VAR}(2)$ model. Then the one step ahead forecast X_{n+1} is given by

$$\hat{X}_{n+1} = [\hat{\alpha}_1 X_n(\phi_{1:d}) + \hat{\alpha}_2 X_{n-1}(\phi_{1:d})]' \phi_{1:d}$$

where $\hat{\alpha}_1$ and $\hat{\alpha}_2$ are $d \times d$ coefficient matrices of $\mathrm{VAR}(2)$ estimated based on n observations. Some calculations show that,

$$X_{n+1} - \hat{X}_{n+1} = [X_{n+1}(\phi_{1:d}) - \hat{\alpha}_1 X_n(\phi_{1:d}) - \hat{\alpha}_2 X_{n-1}(\phi_{1:d})]^T \phi_{1:d} + X_{n+1}(\phi_{d+1:\infty})^T \phi_{d+1:\infty}$$

and then, taking the norm, we get,

$$\|X_{n+1} - \hat{X}_{n+1}\|^2 = \|X_{n+1}(\phi_{1:d}) - \hat{\alpha}_1 X_n(\phi_{1:d}) - \hat{\alpha}_2 X_{n-1}(\phi_{1:d})\|^2 + \|X_{n+1}(\phi_{d+1:\infty})\|^2$$

Finally, the functional prediction error based on n observations can be defined as

$$\mathrm{FPE}_n = E\|X_{n+1} - \hat{X}_{n+1}\|^2 \tag{4.1}$$

Now, the value of d is so chosen that it explains a significant proportion of variance in the data. If chosen appropriately, the term $E\|X_{n+1}(\phi_{d+1:\infty})\|^2$ should be close to 0, and more precisely, it depends on the bias variance trade-off. Also for d fixed, this part can be ignored because it does not change for different d-dimensional model fits. However, it plays a role when comparing prediction errors over a range of d. Further, since the underlying basis system is unknown, this quantity will be difficult to estimate. So, for computational purposes, the residual functions will be given by

$$\epsilon_{n+1}(\hat{\theta}_n) = [X_{n+1}(\phi_{1:d}) - \hat{\alpha}_1 X_n(\phi_{1:d}) - \hat{\alpha}_2 X_{n-1}(\phi_{1:d})]^T \phi_{1:d}$$

where $\hat{\theta}_n$ is the estimate of θ from the data. Thus, the functional prediction error can be redefined as

$$\mathrm{FPE}_n = E\|\epsilon_{n+1}(\hat{\theta}_n)\|^2 \tag{4.2}$$

$$= E\|X_{n+1}(\phi_{1:d}) - \hat{\alpha}_1 X_n(\phi_{1:d}) - \hat{\alpha}_2 X_{n-1}(\phi_{1:d})\|^2 \tag{4.3}$$

4.2. Functional prediction error estimates

If we fit a VAR(p) model to the data, then $n-p$ residuals will be available. Then, an empirical estimate of FPE_n is given by

$$\widehat{\mathrm{FPE}}_n^{emp} = \frac{1}{n-p}\sum_{k=p}^{n-1}\|\epsilon_{k+1}(\hat{\theta}_n)\|^2 \tag{4.4}$$

where $p \leq k \leq n-1$ and $\epsilon_{k+1}(\hat{\theta}_n)$ are the in-sample residual functions. In particular, when a VAR(2) model is fitted, the residual takes the form

$$\epsilon_{k+1}(\hat{\theta}_n) = [X_{k+1}(\phi_{1:d}) - \hat{\alpha}_1 X_k(\phi_{1:d}) - \hat{\alpha}_2 X_{k-1}(\phi_{1:d})]^T \phi_{1:d}$$

Efron (2004) showed that the empirical estimate may not be a very good estimate of the prediction error in regression models. Aue and Burman (2024) confirmed the same in univariate and multivariate time series. We will analyze the functional time series setting in the following. For $m = \lfloor \delta n \rfloor$ with $0 < \delta < 1$, the functional version of Rissanen's APE estimate of FPE_n is defined as

$$\widehat{\mathrm{FPE}}_n^{R} = \frac{1}{n-m}\sum_{k=m}^{n-1}\|\epsilon_{k+1}(\hat{\theta}_k)\|^2 \tag{4.5}$$

When m is small (δ close to 0), then the predictions are based on fewer observations, so bias in estimating FPE_n using $\widehat{\mathrm{FPE}}_n^{R}$ may not be small. On the other hand, if m is close to n (δ close to 1), it may be unbiased for FPE_n but its variance might be high since the estimate is based on fewer residuals.

4.3. Proposed modified estimates of functional prediction error

Similar to the univariate case, the empirical estimate defined above is biased (Efron [2004]). It is also explained in Chapter 1 and the previous section how the functional Rissanen estimate defined above is also a biased estimate of prediction error. So, modified estimates are proposed to see if the biases can be reduced. An estimate of the bias in estimating FPE_n by $\widehat{\mathrm{FPE}}_n^{emp}$ is used to "correct" the empirical estimate. The correction factor is given by

$$C_n(w) = \frac{1}{n-m}\sum_{k=m}^{n-1} w_k \left(\|\epsilon_{k+1}(\hat{\theta}_k)\|^2 - \widehat{\mathrm{FPE}}_k^{emp}\right)$$

Note that the correction factor is a weighted average of the bias when only k observations are used to predict the $(k+1)^{th}$ observation, $m \leq k \leq n-1$. The weights are so chosen to estimate the expected bias $E(\text{FPE}_n - \widehat{\text{FPE}}_n^{emp})$ well. Here, $\widehat{\text{FPE}}_k^{emp}$ are the in-sample residual functions where the parameters are based on the first k observations and are defined as

$$\widehat{\text{FPE}}_k^{emp} = \frac{1}{k-p}\sum_{s=p}^{k-1}\|\epsilon_{s+1}(\hat{\theta}_k)\|^2$$

Note that $m = \lfloor \delta n \rfloor$ with $0 \leq \delta \leq 1$ and $\epsilon_{k+1}(\hat{\theta}_k)$ are out of sample residual functions. Following similar arguments as discussed in Aue and Burman (2024), the first-order bias correction is achieved by choosing w_k as $w_{1k} = n^{-1}k$. Consequently, the correction factor is given by

$$C_n(w_1) = \frac{1}{n(n-m)}\sum_{k=m}^{n-1} k\left(\|\epsilon_{k+1}(\hat{\theta}_k)\|^2 - \widehat{\text{FPE}}_k^{emp}\right)$$

and the modified empirical estimate is given by

$$\widehat{\text{FPE}}_n^{ME}(w_1) = \widehat{\text{FPE}}_n^{emp} + C_n(w_1) \tag{4.6}$$

The original Rissanen estimate uses the simple average of $\|\epsilon_{k+1}(\hat{\theta}_k)\|^2$ to compute the prediction error. The modified Risannen estimate uses the weighted average instead, which is given by

$$\widehat{\text{FPE}}_n^{MR}(v) = \frac{1}{n-m}\sum_{k=m}^{n-1} v_k\|\epsilon_{k+1}(\hat{\theta}_k)\|^2 \tag{4.7}$$

where the weights are chosen following the approach discussed in Chapter 1, so that the estimator is first-order unbiased. Here, the focus will be on the particular weights v_1 as $\lambda_0 + \lambda_1 n^{-1}k$ where

$$\lambda_1 = \frac{\rho_1 - 1}{\rho_{-1}\rho_1 - 1}, \qquad \lambda_0 = 1 - \rho_{-1}\lambda_1$$

and

$$\rho_z = \frac{1}{n-m}\sum_{k=m}^{n-1}\left(\frac{n}{k}\right)^z$$

Simulation studies in Section 4.5 show that the functional modified Rissanen estimate has lower bias but higher variability compared to the corresponding modified empirical estimates.

4.4. Extension to multiple step predictions

Until now the focus was on one step ahead prediction errors. Of course, sometimes it is useful to predict observations for multiple steps. For example, one might be interested in seeing how the daily temperature curves behave for the next seven days. This problem pertains to multi-step ahead predictions. Of course, the quality of predictions is expected to deteriorate as we try to predict too far into the future. Hence, it may be of interest to analyze the performance of longer term predictions, say, h-steps ahead. Following a similar approach as for the 1-step ahead prediction error described above, multi-step prediction errors can be defined. Notice that the optimal forecast of X_{n+h} based on observations $X_s, s \leq n$, is given by the conditional mean function

$$\mu_{n+h}^{(h)} = E[X_{n+h}|X_s, s \leq n]$$

Let us denote $\hat{\mu}_k^{(h)}(\hat{\theta}_s)$ as the estimated value of $\mu_k^{(h)}$ when the model is fitted based on the first s observations, where θ denote the vector of parameters of the model fitted to the scores of the functional data. Then, the h-step ahead residual function is given by

$$\epsilon_k^{(h)}(\hat{\theta}_s) = X_k - \hat{\mu}_k^{(h)}(\hat{\theta}_s)$$

where $\mu_k^{(h)} = E(X_k|X_s, s \leq k-h)$. The multi-step true functional prediction error based on n observations is then given by

$$\mathrm{FPE}_n(h) = E\|\epsilon_{n+h}^{(h)}(\hat{\theta}_n)\|^2 \tag{4.8}$$

4.4.1. Multi-step estimates of prediction error. If a $\mathrm{VAR}(p)$ model is fitted to the functional scores, then the h-step ahead functional empirical prediction error estimate is given by

$$\widehat{\mathrm{FPE}}_n^{emp}(h) = \frac{1}{n-h-p+1}\sum_{k=p}^{n-h}\|\epsilon_{k+h}^{(h)}(\hat{\theta}_n)\|^2 \tag{4.9}$$

Similarly, the functional multi-step Rissanen estimate is given by

$$\widehat{\mathrm{FPE}}_n^{R}(h) = \frac{1}{n-h-m+1}\sum_{k=m}^{n-h}\|\epsilon_{k+h}^{(h)}(\hat{\theta}_k)\|^2 \tag{4.10}$$

where $m = \lfloor \delta n \rfloor$ with $0 < \delta < 1$. Also note that the parameters of the underlying model are updated for each k. All the discussions for 1-step ahead predictions are valid for h-steps ahead predictions too and do not require any special treatment. One needs to be just mindful when calculating the predictions and residual functions because they depend on the underlying model fitted and the value of h.

4.4.2. Multi-step modified estimates. The modified h-step ahead empirical estimate is given by

$$\widehat{\mathrm{FPE}}_n^{ME}(w,h) = \widehat{\mathrm{FPE}}_n^{emp}(h) + C_n(w,h) \tag{4.11}$$

where the form of the correction factor is similar to that of the one-step case. The weights $\{w_k\}$ satisfy the condition

$$g_1(w,h) = \frac{1}{n-h-m+1} \sum_{k=m}^{n-h} k^{-1} n w_k = 1$$

to correct for the first-order bias. Clearly, choosing $w_{1k} = k/n$ satisfy the above condition and thus the correction factor is given by

$$C_n(w_1,h) = \frac{1}{n(n-h-m+1)} \sum_{k=m}^{n-h} k \left(\|\epsilon_{k+h}^{(h)}(\hat{\theta}_k)\|^2 - \widehat{\mathrm{FPE}}_k^{emp}(h) \right)$$

where $\widehat{\mathrm{FPE}}_k^{emp}(h)$ is defined as

$$\widehat{\mathrm{FPE}}_k^{emp}(h) = \frac{1}{k-h-p+1} \sum_{s=p}^{k-h} \|\epsilon_{s+h}^{(h)}(\hat{\theta}_k)\|^2$$

The modified h-step ahead Rissanen estimate is given by

$$\widehat{\mathrm{FPE}}_n^{MR}(v,h) = \frac{1}{n-h-m+1} \sum_{k=m}^{n-h} v_{k,h} \|\epsilon_{k+h}^{(h)}(\hat{\theta}_k)\|^2 \tag{4.12}$$

where the first-order bias correction is achieved by selecting weights $\{v_{k,h}\}$ such that $f_0(v,h) = 1, f_1(v,h) = 1$, where

$$f_z(v,h) = \frac{1}{n-h-m+1} \sum_{k=m}^{n-h} \left(\frac{n}{k}\right)^z v_{k,h}, \qquad z = 0, 1.$$

Recall that

$$\rho_z(h) = \frac{1}{n-h-m+1} \sum_{k=m}^{n-h} \left(\frac{n}{k}\right)^z$$

By choosing the weights $v_{1,h}$ as $\lambda_0(h) + \lambda_1(h)n^{-1}k$, simple calculations show first-order unbiasedness can be achieved by choosing

$$\lambda_1(h) = \frac{\rho_1(h) - 1}{\rho_{-1}(h)\rho_1(h) - 1}, \qquad \lambda_0(h) = 1 - \rho_{-1}(h)\lambda_1(h)$$

All the discussions on choosing weights are similar to the one-step ahead case.

4.5. Simulation results

4.5.1. Simulations for one-step ahead predictions. Simulation studies were carried out in order to assess the performance of the modified estimates of the functional prediction error for the one-step ahead predictions. For any generic estimate $\widehat{\mathrm{FPE}}_n$ of the true prediction error FPE_n, the mean and standard deviations of the difference $\widehat{\mathrm{FPE}}_n - \mathrm{FPE}_n$ were obtained. The results presented here are for 500 simulation runs. However, the true prediction error has been estimated based on 1500 simulation runs to achieve greater accuracy. The following section gives the set-up for data generation.

4.5.1.1. *Set-up.* Suppose ϵ_k represents a function generated by D Fourier basis, $\phi_1, \ldots, \phi_D$, for some k. The ϵ_k serve as innovations to generate a functional MA(1) model given by

$$X_k = \epsilon_k + \Theta\epsilon_{k-1} \tag{4.13}$$

The k-th innovation function is given by

$$\epsilon_k = \sum_{j=1}^{D} z_{jk}\phi_j$$

where the coefficients of the innovation functions are given by the D x n matrix with elements (z_{jk}) where $z_{jk} \sim N(0, \sigma_j), j = 1, \ldots, D, n$ being the sample size and σ_j is the j-th element of the vector $\sigma = (\sigma_1, \ldots, \sigma_D)'$ where σ_j declines exponentially with j.

The kernel of the MA operator Θ is given by

$$\theta(t, s) = \sum_{j=1}^{D} a_j\phi_j(s)\phi_j(t)$$

where the coefficients a_j are so chosen that they decline rapidly with increasing j. This, along with the structure of σ, ensures that the first few eigenvalues represent a significant proportion of variation of the data generated.

The scores $\langle X_k, \phi_j \rangle$ in this set up can be directly computed as

$$\langle X_k, \phi_j \rangle = z_{jk} + a_j z_{j,k-1} \tag{4.14}$$

for $j = 1, \ldots, D, k = 1, \ldots, n$, and the eigenvalues of the covariance kernel of X are given by $\lambda_j = \sigma_j^2(1 + a_j^2), j = 1, \ldots, D$. Thus, it is evident that in this set up, it is enough to generate the z_{jk}'s and choose a_j and σ_j appropriately.

4.5.1.2. *The choice of parameters and model.* Together with the Fourier basis, it is enough to generate the σ_j and $a_j, j = 1, \ldots, D$, for simulation. A large value of D was considered in order to mimic an approximation of infinite-dimensional functions. Here, D was chosen to be 21 for the simulations. Higher values of D were also considered, such as $D = 31$ or $D = 51$. Since the results were comparable, only those for $D = 21$ are reported in the following. The σ vector was chosen as $\sigma = (1.5^{-j} : j = 1, \ldots, D)'$ and the kernel coefficients are set as

$$a_j = \frac{1}{(j+1)^{1.5}}$$

Here, the sample size n was set as 100. Once the z's are generated, the scores can be computed easily using (4.14). These scores now represent a multivariate time series with D dimensions. The proportion of variation explained by the j-th principal component can be calculated as

$$\frac{\lambda_j}{\sum_{j=1}^{D} \lambda_j}$$

It was found that the first $d = 5$ eigenvalues explained more than 95% of variation in the data. So, the scores corresponding to the first 5 eigenvalues were now considered as the multivariate time series and a VAR(p) process was fitted to the data. In general, if the MA(1) model is invertible, it can be represented as an infinite AR model. Hence, a higher order VAR model was deemed appropriate in this case, and a $\mathrm{VAR}(2)$ model is fitted to the d-dimensional scores for computational simplicity. However, a $\mathrm{VAR}(2)$ might not be the correct model and some higher values of p could have been considered as better approximations. It is to be noted then that the prediction error has sources of error coming from model misspecification as well.

4.5.1.3. *Results.* Once the predictions are made using a $\mathrm{VAR}(p)$ model with $p = 2$, the prediction error and its estimates were calculated. The mean and standard deviations of the deviations of the estimates

from the true prediction error were plotted for different choices of δ varying between 0.2 and 0.95. More specifically, they were computed based on the following:

- *Functional empirical estimates*: $\widehat{\mathrm{FPE}}_n^{emp} - \mathrm{FPE}_n$ which does not depend on δ
- *Functional Rissanen estimates*: $\widehat{\mathrm{FPE}}_n^{R} - \mathrm{FPE}_n$
- *Functional modified empirical estimates*: $\widehat{\mathrm{FPE}}_n^{ME} - \mathrm{FPE}_n$ with weights $w_{1k} = n^{-1}k$
- *Functional modified Rissanen estimates*: $\widehat{\mathrm{FPE}}_n^{MR} - \mathrm{FPE}_n$ with weights $v_{1k} = \lambda_0 + \lambda_1 n^{-1}k$

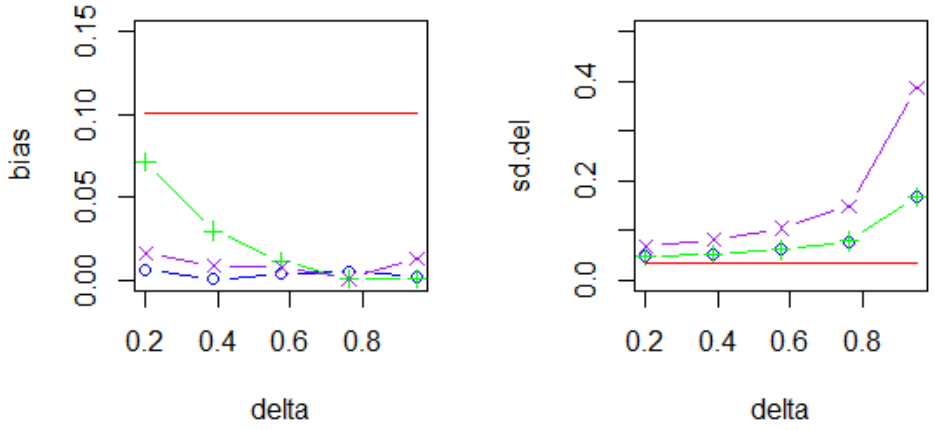

FIGURE 4.1. Simulation results with basis assumed to be known: the left panel showing bias in estimating the true prediction error and the right panel showing the SD of the deviations for Empirical Estimate (——), Modified Empirical Estimate (− ∘ − ∘ −), Rissanen's Estimate (− + − + −), Modified Rissanen's Estimate (− × − × −)

From Figure 4.1, it is evident that the empirical estimate has the largest bias followed by Rissanen's estimate, as expected. The modified estimates are better at lowering the bias of the prediction error estimates. The modified empirical estimates perform consistently well in terms of the lowest bias for all δ and their standard deviations are also not significantly different from those of the empirical estimates, which have the lowest standard deviation. Overall, considering the bias-variance trade-off, the modified empirical estimate seems to perform the best in this case.

4.5.2. When the underlying basis is assumed to be unknown. Even though in simulations all the parameters and basis system for functional data is known, one can assume that only the data is provided for

analysis and the underlying basis that generated the functional data is not known. In that case, FPCA can be performed on the data and the eigenfunctions thus obtained serve as the basis. The analysis remains exactly the same, including the choice of the order of the VAR model. The choice of d is now based on the estimated eigenvalues. The only difference lies in the residual functions which, if a VAR(2) is fitted, are now defined as

$$\epsilon_{n+1}(\hat{\theta}_n) = [X_{n+1}(\hat{\phi}_{1:d}) - \hat{\alpha}_1 X_n(\hat{\phi}_{1:d}) - \hat{\alpha}_2 X_{n-1}(\hat{\phi}_{1:d})]' \hat{\phi}_{1:d}$$

where $\hat{\phi}_1, \ldots, \hat{\phi}_d$ are eigenfunctions from FPCA on mean centered observations.

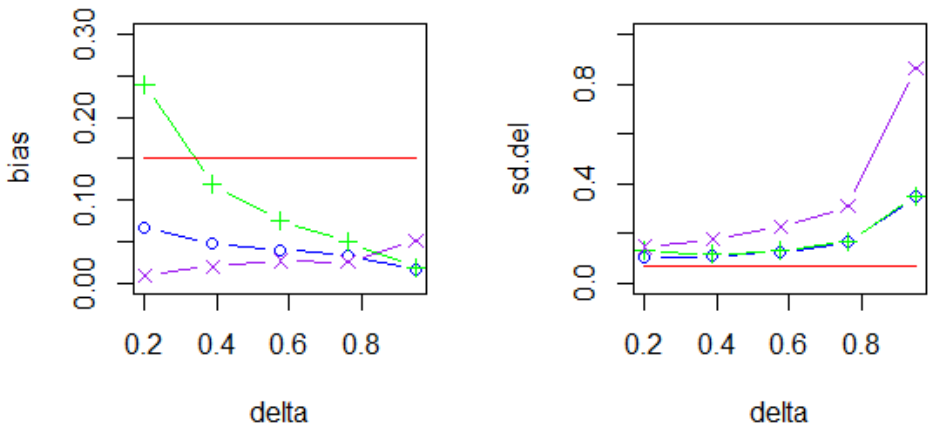

FIGURE 4.2. Simulation results with basis assumed to be unknown: the left panel showing bias in estimating the true prediction error and the right panel showing the SD of the deviations for Empirical Estimate (——), Modified Empirical Estimate (− ∘ − ∘ −), Rissanen's Estimate (− + − + −), Modified Rissanen's Estimate (− × − × −)

As seen in Figure 4.2, here also, an evident improvement in terms of lowering the bias is observed for the modified estimates when compared to the empirical and Rissanen's estimate. We observe a higher bias and lower variance for δ close to 0 and the bias reduces and variance increases for δ close to 1. The modified Rissanen's estimate seem to have the lowest bias until $\delta \approx 0.8$, but its variance is consistently the highest and increases further for $\delta > 0.8$. Overall, it appears that the modified empirical estimate is consistently performing the best in terms of estimating the true one-step ahead functional prediction error.

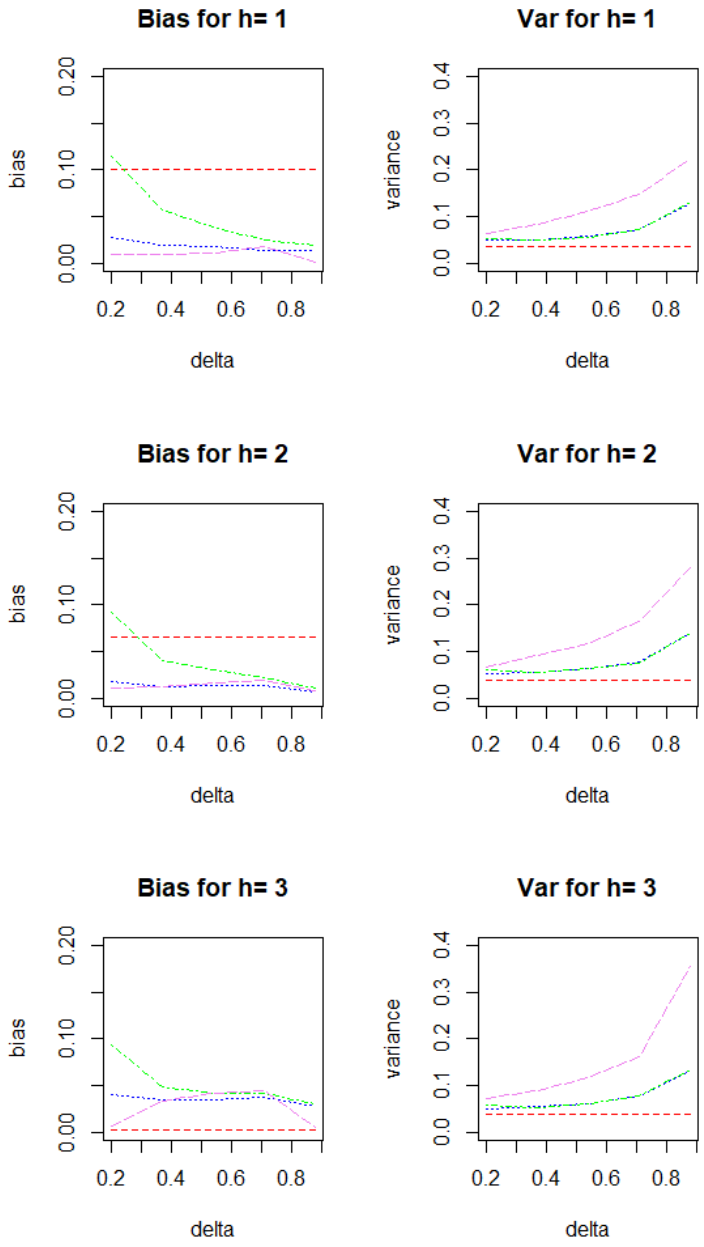

FIGURE 4.3. Multi-Step Prediction Error bias (left) and standard deviations (right) for h = 1, 2, 3: Empirical (- - - - -), Modified Empirical (· · · · · ·), Rissanen (— · — · —), Modified Rissanen (— — —)

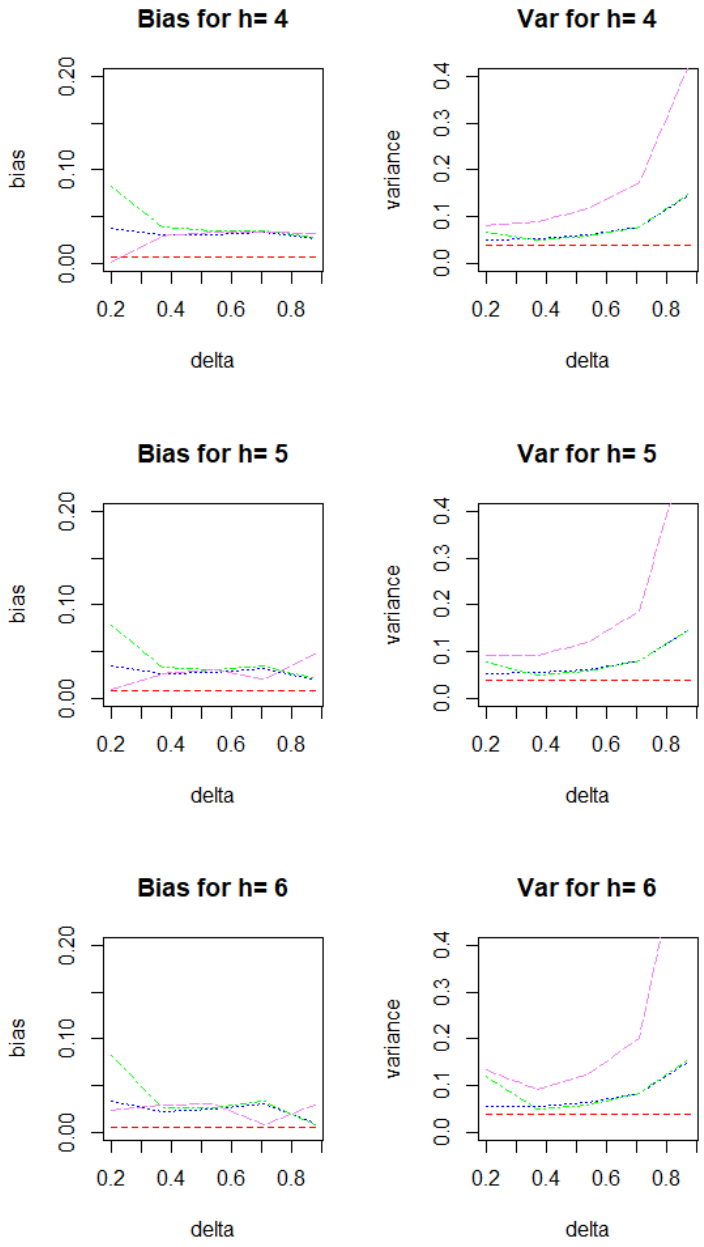

FIGURE 4.4. Multi-Step Prediction Error bias (left) and standard deviations (right) for h = 4, 5, 6: Empirical (- - - -), Modified Empirical (· · · · · ·), Rissanen (— · — · —), Modified Rissanen (— — —)

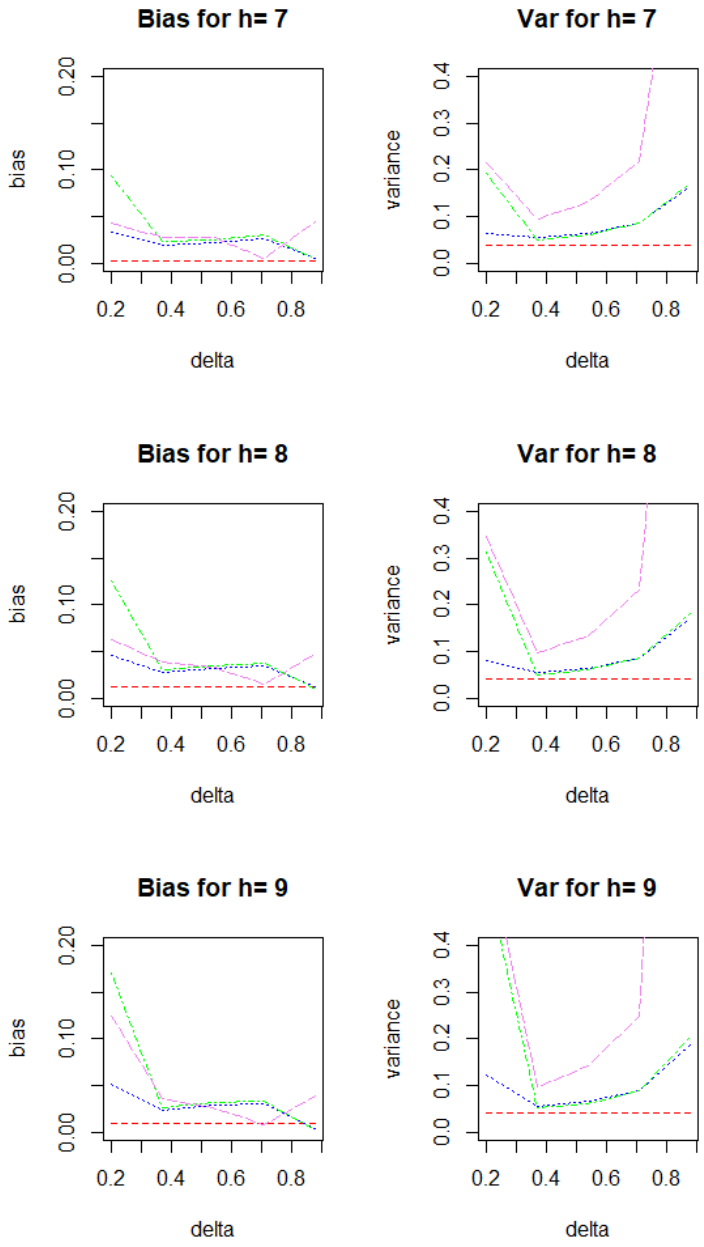

FIGURE 4.5. Multi-Step Prediction Error bias (left) and standard deviations (right) for h = 7, 8, 9: Empirical (- - - - -), Modified Empirical (· · · · · ·), Rissanen (— · — · —), Modified Rissanen (— — —)

4.5.3. Simulations for multi-step ahead predictions. While estimating the multi-step prediction error, the last 10 observations were set aside so that the true prediction error can be computed based on 10-step

ahead predictions. FPCA was implemented on $n_1 = n - 10$ observations and a similar approach was followed as described above to get the prediction error estimates.

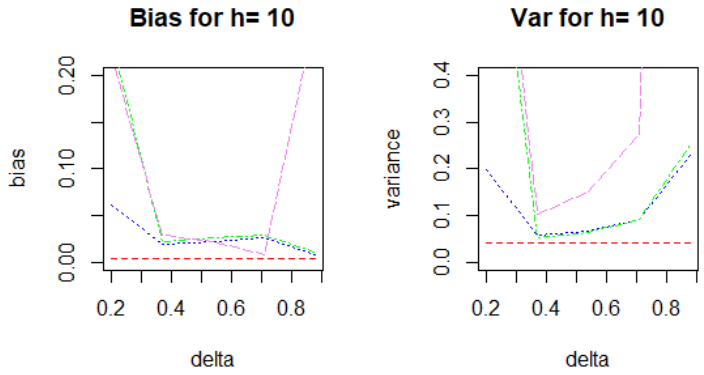

FIGURE 4.6. Multi-Step Prediction Error bias (left) and standard deviations (right) for h = 10: Empirical (- - - - -), Modified Empirical (· · · · · ·), Rissanen (— · — · —), Modified Rissanen (— — —)

Figures 4.3-4.6 are displayed for $h = 1, \ldots, 10$, where h is the number of steps ahead the forecasts are computed. The modified Rissanen estimates become highly variable and highly biased for longer term predictions and for too small or too big delta.

4.6. Application to real data: Australia temperature

To demonstrate the practical utility of prediction error estimators for functional data, the new methods are applied to temperature data. Specifically, daily maximum temperature data measured on a number of meteorological stations in Australia were considered to implement the functional prediction error estimates, where the data corresponding to one year represents a function. The modeling of temperature dynamics plays an important role in understanding the degree of temperature variations during every year. The accurate prediction of future temperatures and the assessment of the corresponding prediction errors has thus direct impact on policy and decision-making processes. Here, detailed results will be provided for the Sydney Observatory Hill meteorological station and major results will be provided for Gayndah Post Office to avoid repetitiveness. The first analysis is provided for the Sydney Observatory Hill meteorological station. Data of this type was considered in other contexts in Aue et al. (2018), Aue and van Delft (2020) and Dette et al. (2020).

4.6.1. Sydney Observatory Hill. The observations are daily maximum temperatures (measured in degree Celsius) recorded at Sydney Observatory Hill from 01 January 1859 to 31 December 2019. The data was obtained from the Bureau of Meteorology, Government of Australia. Since the temperature curve for each year represents a function, there are $n = 161$ functional observations, each constructed from discrete daily observations observed at 365 points. For simplicity, for the leap years, the average of February 28 and 29 was calculated as the temperature for February 28 and February 29 temperature was removed. There were 152 days where the maximum temperature data were unavailable of the Sydney Observatory Hill station. These missing values were treated by mean imputation. For example, if the temperature was missing for January 04 for the year 1960, then, the mean temperature for January 04 for all the other years was imputed for the missing value. The data was then arranged in a 365×161 matrix form, where each column corresponded to an observation with 365 points. The 161 column means corresponded to the trend over the years, while the 365 row means corresponded to the seasonality effect.

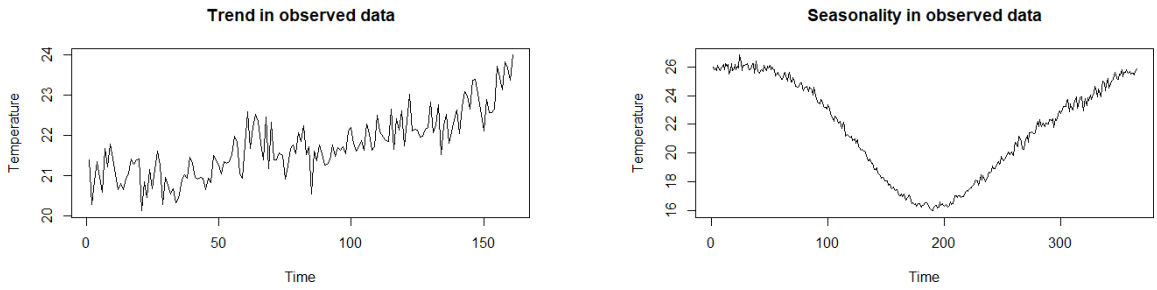

FIGURE 4.7. Trend (column means) and Seasonality (row means) in Sydney Observatory Hill Temperatures when arranged in a matrix form where each column corresponds to a functional observation for a year.

There is an increasing trend over the years as seen in the left panel of Figure 4.7. This upward trend over a significant time period might suggests a broader trend of climate change and global warming. Seasonality is captured in the right panel of Figure 4.7. Since seasonality is the main driver for variation in temperature within a year, the trend was removed from the data so that the data only captures the seasonal pattern. The data was transformed into functions using 15 Fourier basis functions and least squares fitting using the `fda` package available in the `R` statistical software was applied. The functions and their mean are shown in Figure 4.8. The 161 functions $X_1, \ldots, X_{161}$ represent the annual temperature curves for the Sydney Observatory Hill meteorological station. The mean function shows that the temperatures are higher at the beginning and

end of the year whereas it is lower in the middle of the year. It essentially captures the seasonal effect of all the functions and is the dominating component in the temperature functions. This sample mean or average is an estimator of the population seasonal effect. Hence, we subtract the average sample seasonal effect and prediction errors are calculated for the mean centered functions, which represent the random component in the data.

The plot shows considerable variation of the functions, specially at the beginning and the end of the year which represents summer months in Australia. This results in very wiggly sample eigenfunctions estimated using FPCA. The eigenfunctions convey information on deviations from the mean function, which describe the average seasonal behavior for any given year. Therefore, to examine the effect of the first three FPCs on the mean curve of the centered functions, Figure 4.9 is plotted where a multiple chosen as the l-th empirical eigenvalue $\hat{\lambda}_l$ of the l-th empirical eigenfunction $\hat{\phi}_l$ was added to and subtracted from the estimated mean curve of the centered observations for $l = 1, 2, 3$.

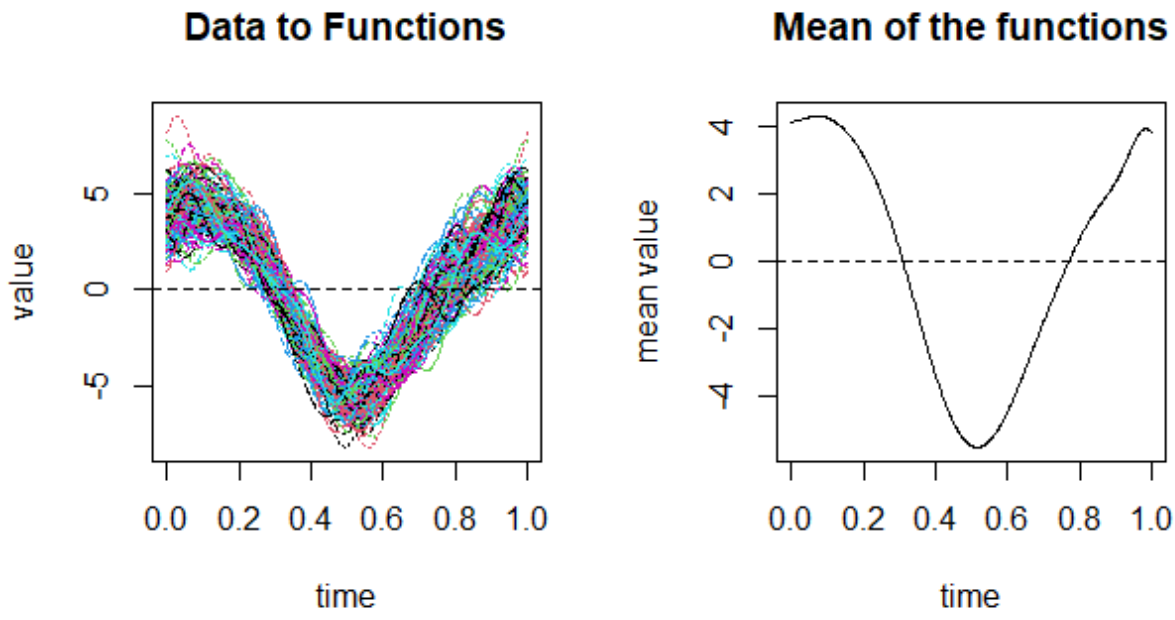

FIGURE 4.8. Sydney Observatory Hill Temperature Functions and their mean

Figure 4.9 shows that the eigenfunctions resemble the data and are more variable in the beginning and end of the year. Upon investigating further, we see that there are two effects overlapping here. The first is

that summer months are somewhat more variable than winter month, but not to such a large extent. The second, and more important effect is the way functions were initially registered which resulted in increased variability in the data extremes (the beginning and end of each year) and the first 5 empirical eigenvalues only explaining 65% of the total variations present in the data. This motivates a new method of function registration developed in this thesis which is described in the next subsection.

4.6.1.1. *New method of registration of functions.* The functions depicted in Figure 4.8 are wiggly especially around intraday time $t = 0$ and $t = 1$. While they are still continuous on $[0, 1]$, if we are to collate them year by year, we would introduce discontinuities, each time the old year ends and the new year starts. This means that there was a jump in between December temperatures of the old year and the January temperatures of the next year. When using functional data in this context, how the function behaves per year is more important, because the major variations in the data is captured by the seasonal pattern each year. However, such an approach implies that we lose the interpretation of temperature as a stochastic process evolving over the years. The proposed new approach of function registration reconciles the two views by ensuring an almost continuous evolution of the trajectories from one year to the next in the stochastic process representation.

To deal with this, for $k = 2, \ldots, n-1$, the daily observations for the k-th year were concatenated with the December observations of the $(k-1)$-th year and the January observations of the $(k+1)$-th year. For $k = 1$, only the January temperatures of year 2 were concatenated. For $k = n$, only the December temperatures of year $n-1$ were concatenated. After registering extended functions based on concatenated discrete daily observations using least square smoothing methods, a pruning step was applied so that each function starts in January and ends in December for a given year. This was done using the `funData` package available in the `R` statistical software. So the `fda` objects were converted to `funData` objects, then the `extractObs` function was used to prune the functions and then they were transformed to an `fda` object using 15 Fourier basis functions where each function has the domain $[0, 1]$. The resulting functions are plotted in Figure 4.10 against the old functions to compare the performance of the new registration. The functions obtained from the new registration method are less volatile and more smooth overall than the functions obtained from the traditional registration method as shown in Figure 4.10.

Plotting the functions as a stochastic process in Figure (4.11), shows that continuity has been ensured to a reasonable degree. Even though this is not pursued in the context of this dissertation, the new way

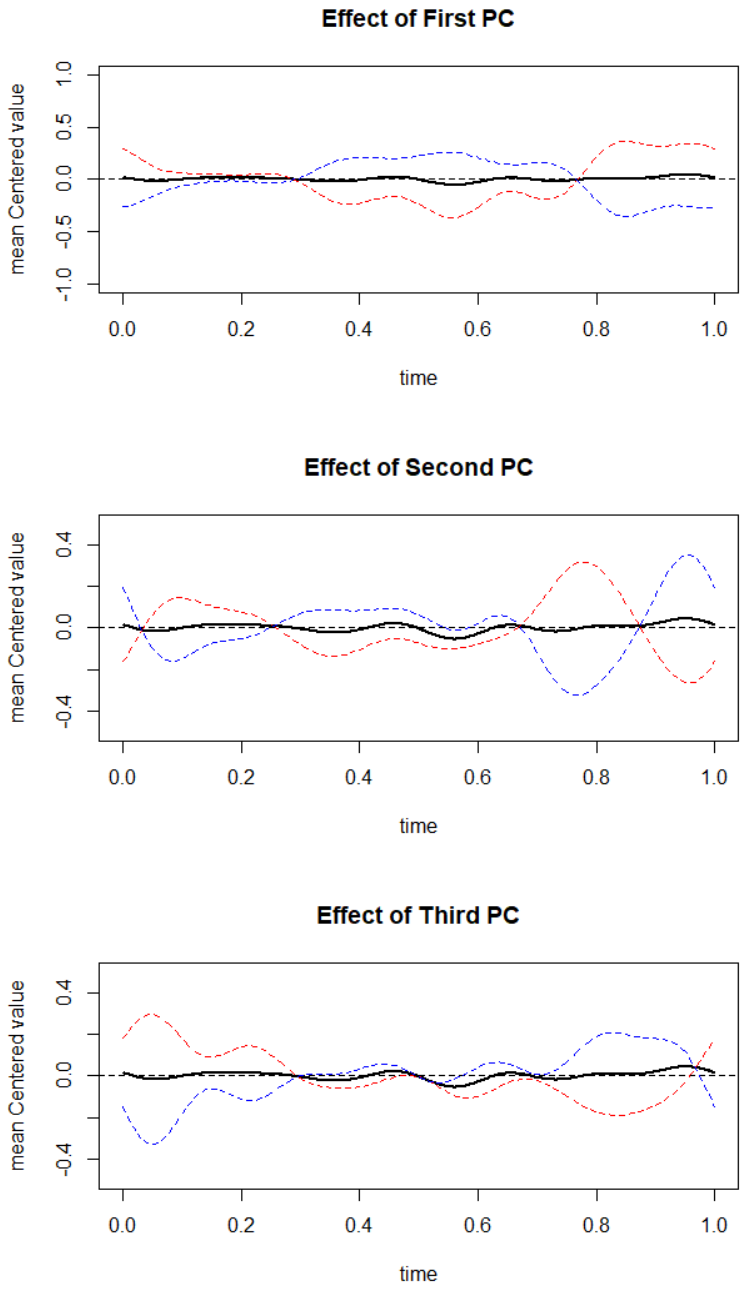

FIGURE 4.9. Effect of First Three PC on the Mean of the centered functions: $\hat{\mu} + \hat{\lambda}_l \hat{\phi}_l$ represented as (- - - -) and $\hat{\mu} - \hat{\lambda}_l \hat{\phi}_l$ represented as (- - - -) along with the mean $\hat{\mu}$ (———) for $l = 1, 2, 3$

of registering functional time series data should also have benefits if interest is in predicting when partial knowledge of the future curve is available; see Jiao et al. (2023).

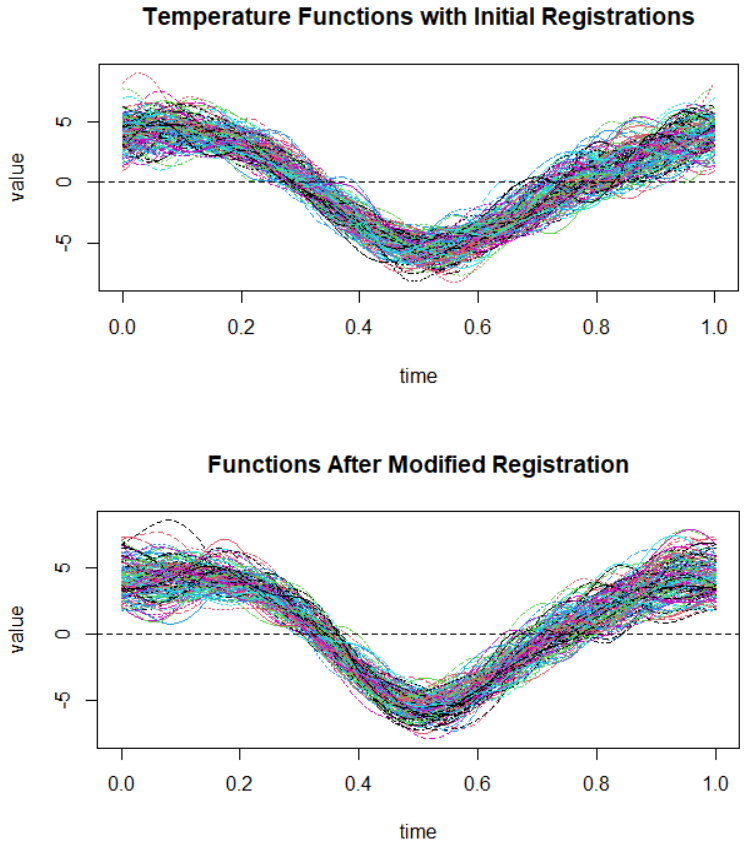

FIGURE 4.10. Initial Functions and Modified Functions of Sydney Observatory Hill

The first 5 empirical eigenvalues explain around 75% variation in the data, which is clearly higher than for the previous registration. Figure 4.12 shows the effect of the first 3 FPCs on the mean on the newly registered functions. The eigenfunctions appear to be much more stable than before and hence, less wiggly.

The first principal component shows that if the temperatures are higher than normal in December (end of the observation window), it will be higher than normal in January (start of the observation window) and vice versa , indicating preservation of continuity. The second and third FPCs are compensating for the volatile summer months. The second FPC shows that if the fall temperatures are higher than normal, so will be the spring temperatures. Overall, it seems that the new method of function registration works well, hence, we will proceed with the newly generated functions as our observations.

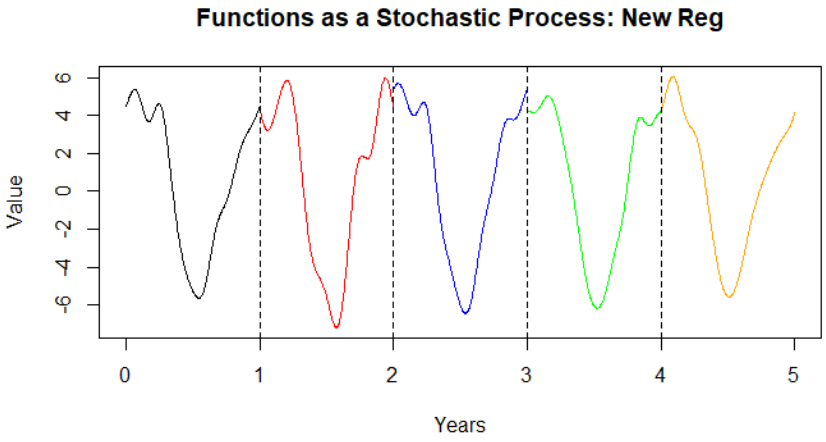

FIGURE 4.11. 5 years of temperature curves after the new registration

4.6.1.2. *Choosing the 'best' p and d.* Unlike for simulated data, the optimal VAR order p to be fitted to the scores of the functions is unknown. A small value of p might lead to a biased model. Similarly, a small value of d might not explain a significant proportion of variance in the data. So, the choice of both p and d depends on a bias-variance trade-off. Based on this principle, the optimal values of the VAR order p and the reduced dimension d can be chosen based on the algorithm described in Aue et al. (2015). Their criterion is based on minimizing the mean squared error (MSE) when a VAR(p) model is fitted to the d-dimensional scores. It was shown in the paper that an approximate expression of one-step prediction error is given by

$$E\|X_{n+1} - \hat{X}_{n+1}\|^2 \approx \frac{n - pd}{n + pd}\text{tr}(\hat{\Sigma}_Z) + \sum_{l>d} \hat{\lambda}_l$$

where Z's are the residuals of the VAR(p) model fitted and $\Sigma_Z = E[Z_1 Z_1']$ and $\hat{\Sigma}_Z$ is its estimate. $\sum_{l>d} \hat{\lambda}_l$ represents the proportion of variance unexplained by the first d eigenvectors.

Using this algorithm, the 'best' values of p and d came out to be 2 and 3 respectively. This means that the functional prediction error estimate is minimum when a VAR(2) model is fitted to three dimensional scores. However, it is to be noted that unlike for simulated data, here, the true functional prediction error is unknown, and hence the functional prediction error estimates could not be compared to the true prediction error.

4.6.1.3. *Prediction bands using FPE estimates.* In the introduction in Section 1, it was mentioned that a very useful application of prediction errors is to construct prediction bands. So, instead of comparing the

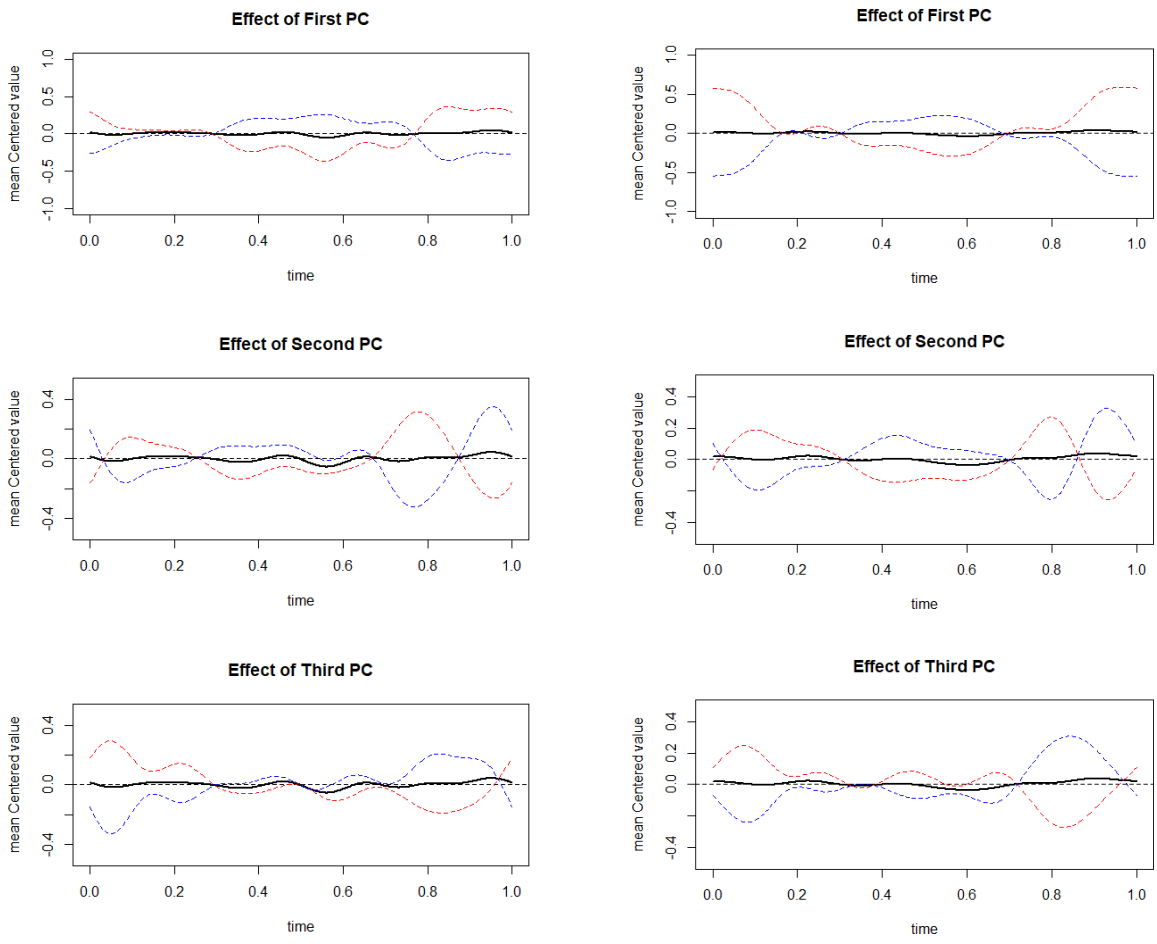

FIGURE 4.12. Effect of First Three PC on the Mean of New Functions (right) as compared to the Old Functions (left): $\hat{\mu} + \hat{\lambda}_l \hat{\phi}_l$ represented as (- - - -) and $\hat{\mu} - \hat{\lambda}_l \hat{\phi}_l$ as (- - - -) along with the mean $\hat{\mu}$ (———) for $l = 1, 2, 3$

estimates to the true prediction error, one can get h step ahead prediction error estimates defined in Section 4.4 and use them to construct point-wise prediction bands.

Recall that for a fixed d, we defined $X_k(\phi_{1:d}) = (\langle X_k, \phi_1 \rangle, \ldots, \langle X_k, \phi_d \rangle)^T$ and $\phi_{1:d} = (\phi_1, \ldots, \phi_d)^T$. Now, the h-step ahead future observation, using the truncated KL representation, can be written as

$$X_{n+h} = \sum_{j=1}^{d} \langle X_{n+h}, \phi_j \rangle \phi_j$$

$$= X_{n+h}(\phi_{1:d})^T \phi_{1:d}$$

Similarly, the h-step ahead predicted functions can be represented as

$$\hat{X}_{n+h} = \hat{X}_{n+h}(\phi_{1:d})^T \phi_{1:d}$$

Prediction bands were constructed based on Scheffe's method (Scheffe (1969)) of simultaneous confidence intervals. Let $\hat{\theta} \sim N_d(\theta, \Sigma)$ be the estimator of the parameter $\theta \in \mathbb{R}^d$, then a $100(1-\alpha)\%$ approximate simultaneous confidence interval for a linear combinations of θ, given by $a^T\theta, a \in \mathbb{R}^d$ is given by

$$a^T\hat{\theta} \pm \sqrt{\chi^2_d(1-\alpha)}\sqrt{a^T\Sigma a} \tag{4.15}$$

For a prediction interval, instead of Σ, $\mathrm{Cov}(\hat{\theta}-\theta)$ is used. Taking $\theta = X_{n+h}(\phi_{1:d})$ and $\hat{\theta} = \hat{X}_{n+h}(\phi_{1:d})$, we can define the pointwise prediction band for h-step ahead predictions using (4.15). In this case, $a = \phi_{1:d}(t)$, so the prediction band has to be evaluated for each $t \in [0,1]$. Thus, the approximate simultaneous prediction band for $t \in [0,1]$ can be written as

$$\hat{X}_{n+h}(\phi_{1:d})^T \phi_{1:d}(t) \pm \sqrt{\chi^2_d(1-\alpha)}\sqrt{\phi^T_{1:d}(t)\Sigma\phi_{1:d}(t)} \tag{4.16}$$

Now, Σ here is given by

$$\begin{aligned}\Sigma = \mathrm{Cov}(\hat{X}_{n+h}(\phi_{1:d}) - X_{n+h}(\phi_{1:d})) &= \mathrm{Cov}(\epsilon^{(h)}_{n+h}(\phi_{1:d})) \\ &= E\left([\epsilon^{(h)}_{n+h}(\hat{\theta}_n)(\phi_{1:d})][\epsilon^{(h)}_{n+h}(\hat{\theta}_n)(\phi_{1:d})]^T\right)\end{aligned}$$

This is exactly the same $\epsilon^{(h)}_{n+h}(\hat{\theta}_n)(t)$ that was used to define the h-step ahead functional prediction error, but instead of the norm of the scores, the covariance matrix of the scores is considered here.

However, Σ is unknown and has to be estimated by $\widehat{\Sigma}$ to get the prediction bands, where $\widehat{\Sigma}$ is based on different FPE estimates as defined in Section 4.4. For example, based on $\widehat{\mathrm{FPE}}^{emp}_n(h)$, as given by equation (4.9), it is given by

$$\widehat{\Sigma}_{emp}(h) = \frac{1}{n-h-p+1}\sum_{k=p}^{n-h}[\epsilon^{(h)}_{k+h}(\hat{\theta}_n)(\phi_{1:d})]\,[\epsilon^{(h)}_{k+h}(\hat{\theta}_n)(\phi_{1:d})]^T$$

When it is based on $\widehat{\mathrm{FPE}}_n^R(h)$ (equation (4.10)), it is given by

$$\widehat{\Sigma}_R(h) = \frac{1}{n-h-m+1} \sum_{k=m}^{n-h} [\epsilon_{k+h}^{(h)}(\hat{\theta}_k)(\phi_{1:d})] \, [\epsilon_{k+h}^{(h)}(\hat{\theta}_k)(\phi_{1:d})]^T$$

Similarly, $\widehat{\Sigma}_{ME}(h)$ and $\widehat{\Sigma}_{MR}(h)$ can be defined. For illustration purpose, $h = 1, 2$ and 5 are chosen to see the performance of 1-step, 2-steps and 5-steps ahead prediction errors in terms of their prediction bands. The reason to choose such h values is to see the performance of the prediction error estimates for short as well as medium term future. Here, $\alpha = 0.05$ was chosen, so the bands are pointwise 95% prediction bands. If the prediction bands contain the true observed function with certain confidence, then it can be concluded that the functional prediction errors estimates are able to capture the uncertainty of predicting future observations. In general, it is expected that the immediately next time point prediction for $h = 1$ will be better, and hence the prediction error will be smaller but the prediction errors increase with increase in h.

While computing the modified empirical, Rissanen and modified Rissanen's estimate, a choice of δ is required. Here, $\delta = 0.6$ is chosen since from the simulation studies, we have observed that the modified empirical and modified Rissanen's estimates have the smallest bias around such a δ and for a higher δ, the variance of the estimates increases.

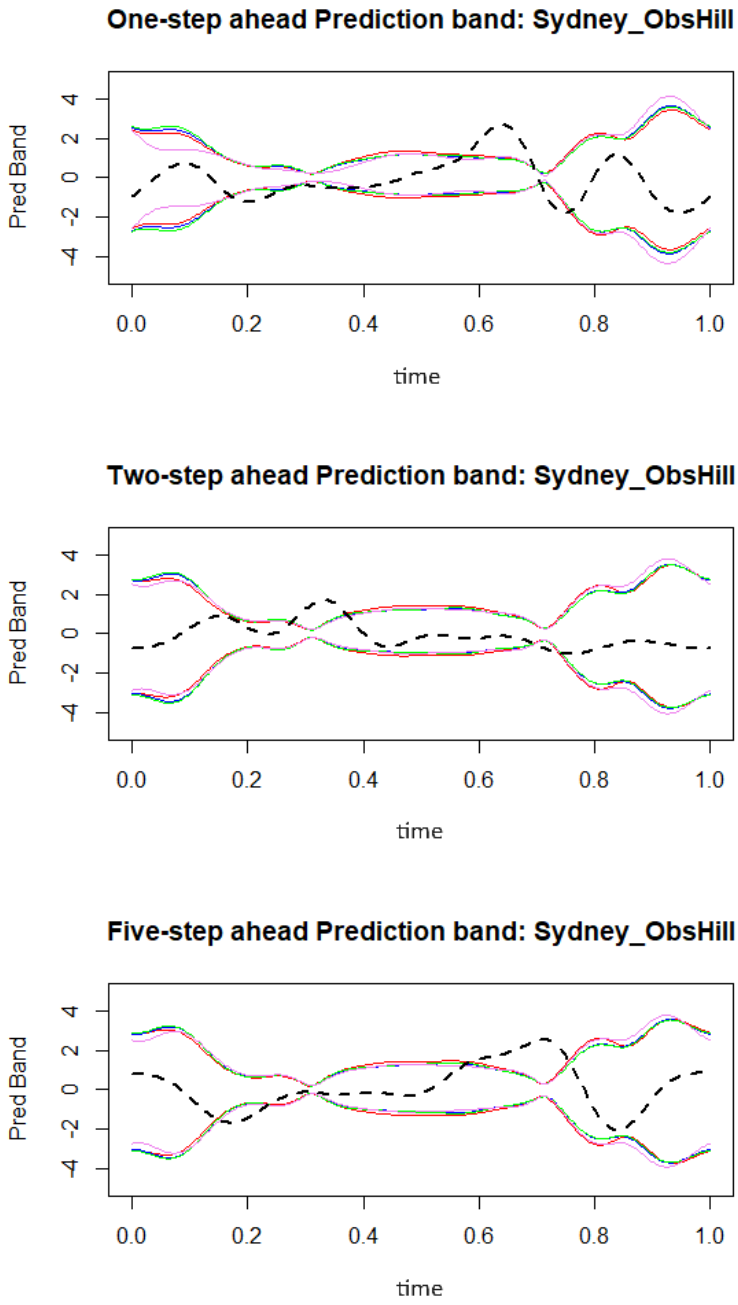

FIGURE 4.13. One, two and five steps Prediction bands for mean centered Sydney Observatory Hill data: Empirical(—), Modified empirical (—), Rissanen (—) and Modified Rissanen (—) along with the true functions (- - -)

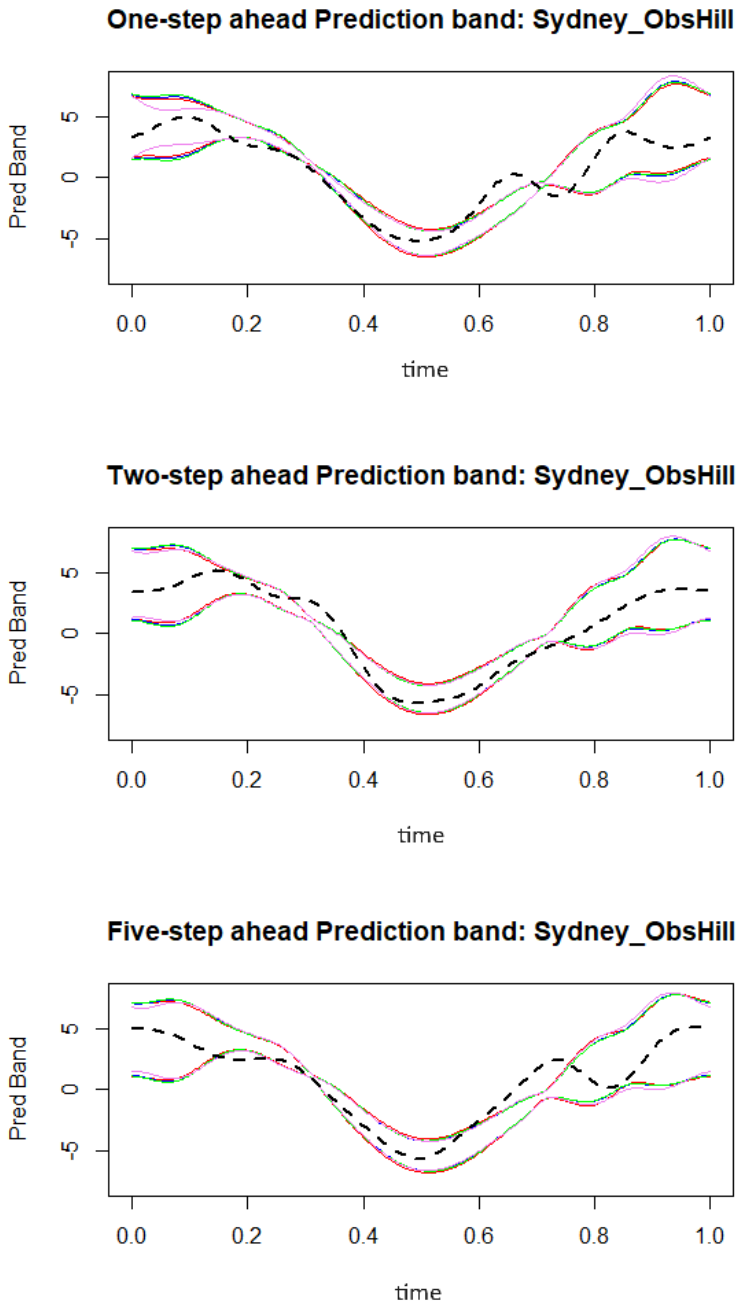

FIGURE 4.14. One, two and five steps Prediction bands for annual temperature profiles of Sydney Observatory Hill: Empirical(—), Modified empirical (—), Rissanen (—) and Modified Rissanen (—) along with the true functions (**- - -**)

Figure 4.13 shows the prediction bands for 1, 2 and 5 steps along with the true function. The bands are able to capture the variations in the mean centered data, since most of the observed functions are within the bands. There is not much difference between the bands produced by the different FPE estimates. But all the bands are wider at the ends, since the volatility, even though decreased after the new registration of the functions, is still high at the beginning and the end of the year. The prediction bands can also be generated for the annual temperature profiles instead of the mean centered functions by adding back the mean of the functions that represents the seasonal component.

It is evident from Figure 4.14 that the temperature profile variations are well captured by the prediction bands. Overall, we can conclude that for this dataset, there is no significant difference in the performances of the different functional prediction error estimates, and the prediction bands produced by the estimates produce reasonable ranges of future temperature profiles.

4.6.2. Gayndah Post Office. Gayndah was the second meteorological station whose data was selected for analysis. The daily maximum temperatures in degree Celsius were recorded from January 1, 1894 to December 31, 2008. There were 749 missing observations which were treated using mean imputation. Temperature functions were obtained from the observations using the methods described in Section 4.6.1. There were a total of 115 observations in this dataset.

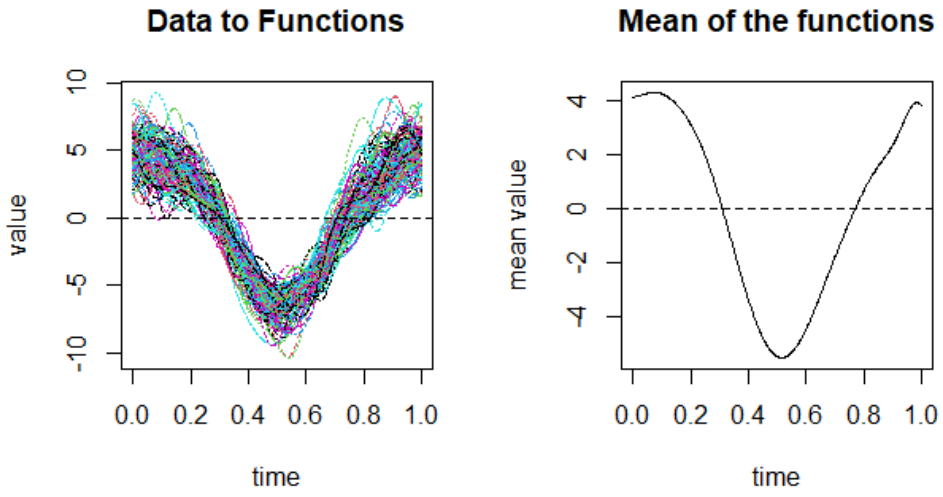

FIGURE 4.15. Gayndah Post Office Temperature Functions and their mean

Figure (4.15) shows that these functions are very volatile throughout the year. The first 5 FPCs explain about 71% of variations in the data. The new method of registrations reduces this variation to some extent by making the functions less wiggly and more smooth as shown in Figure (4.16) and the first 5 FPCs explain about 80% of the variations in the newly registered functions.

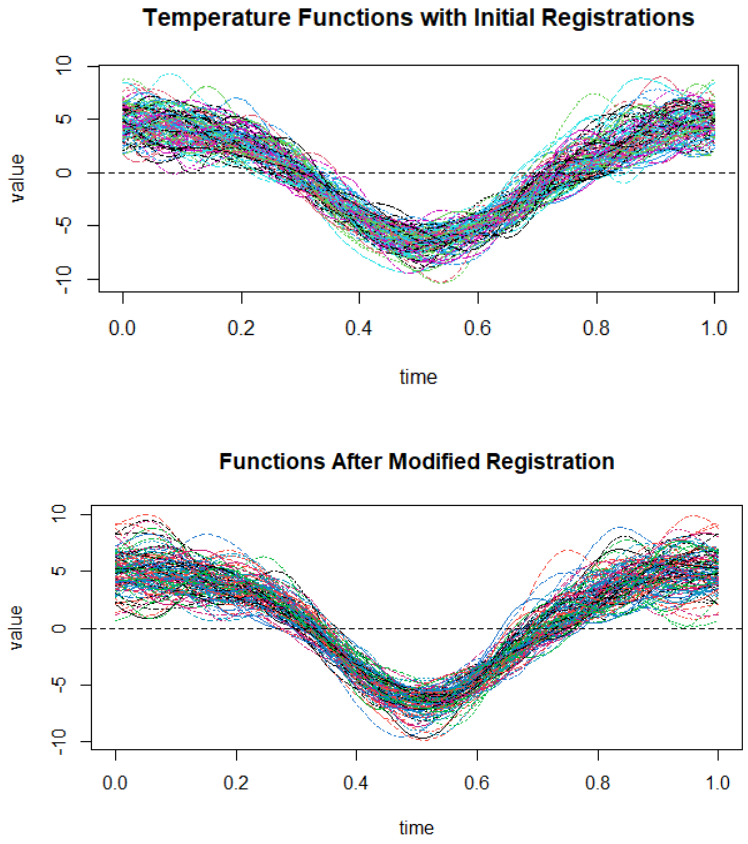

FIGURE 4.16. Initial Functions and Modified Functions of Gayndah Post Office

With the new registered data, the optimal values of both p and d came out to be 3. It means that we can fit a $\mathrm{VAR}(3)$ model to 3-dimensional scores to get the optimal values of prediction error. Using those values of p and d, FPE estimates are calculated. Using the FPE estimates, prediction bands are obtained like before for $h = 1, 2$ and 5 steps ahead for the mean centered functions which are shown below. Figure (4.17) shows the prediction bands for 1, 2 and 5 steps along with the true function. It shows that even though the bands are able to capture the variations in the mean centered data, the prediction bands are wider to accommodate

more volatile curves. Like before, the bands are more wider at the beginning and the end of the year, since the functions are more volatile at those times.

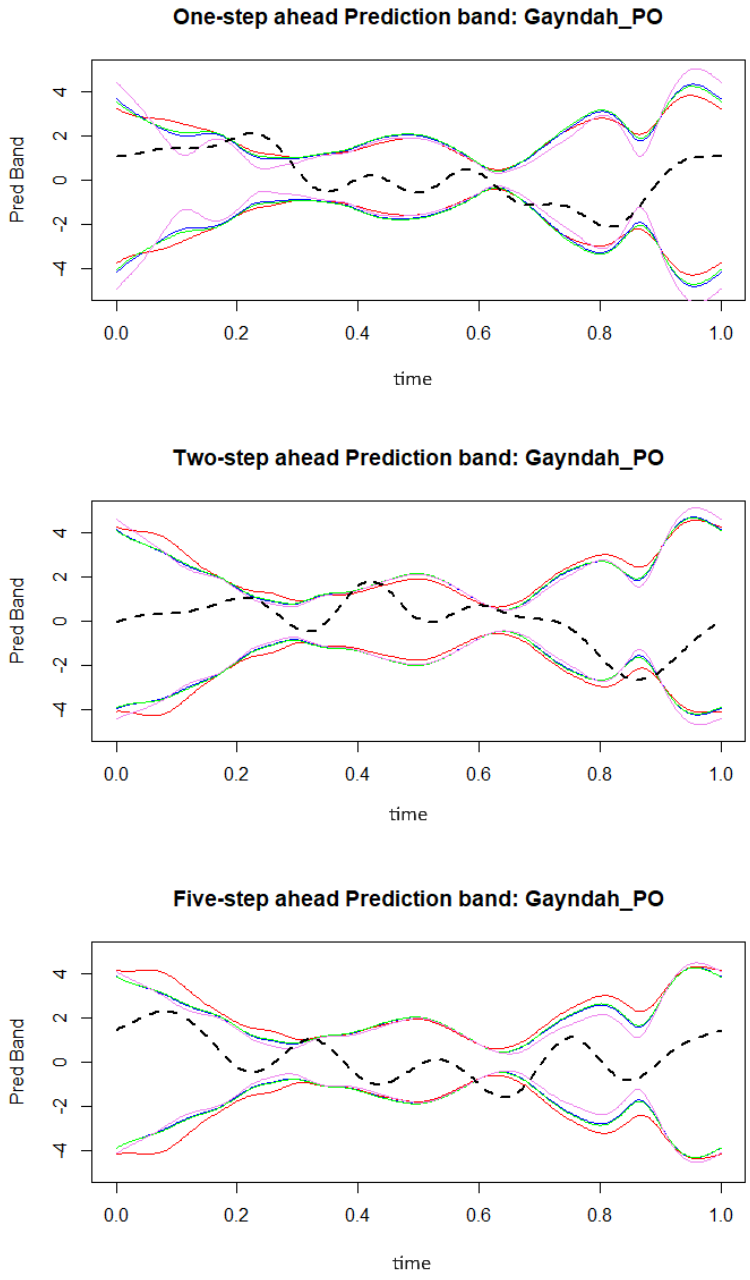

FIGURE 4.17. One, two and five steps Prediction bands for mean centered Gayndah Post Office data: Empirical(—), Modified empirical (—), Rissanen (—) and Modified Rissanen (—) along with the true functions (**- - -**)

The prediction bands are also generated for the annual temperature profiles.

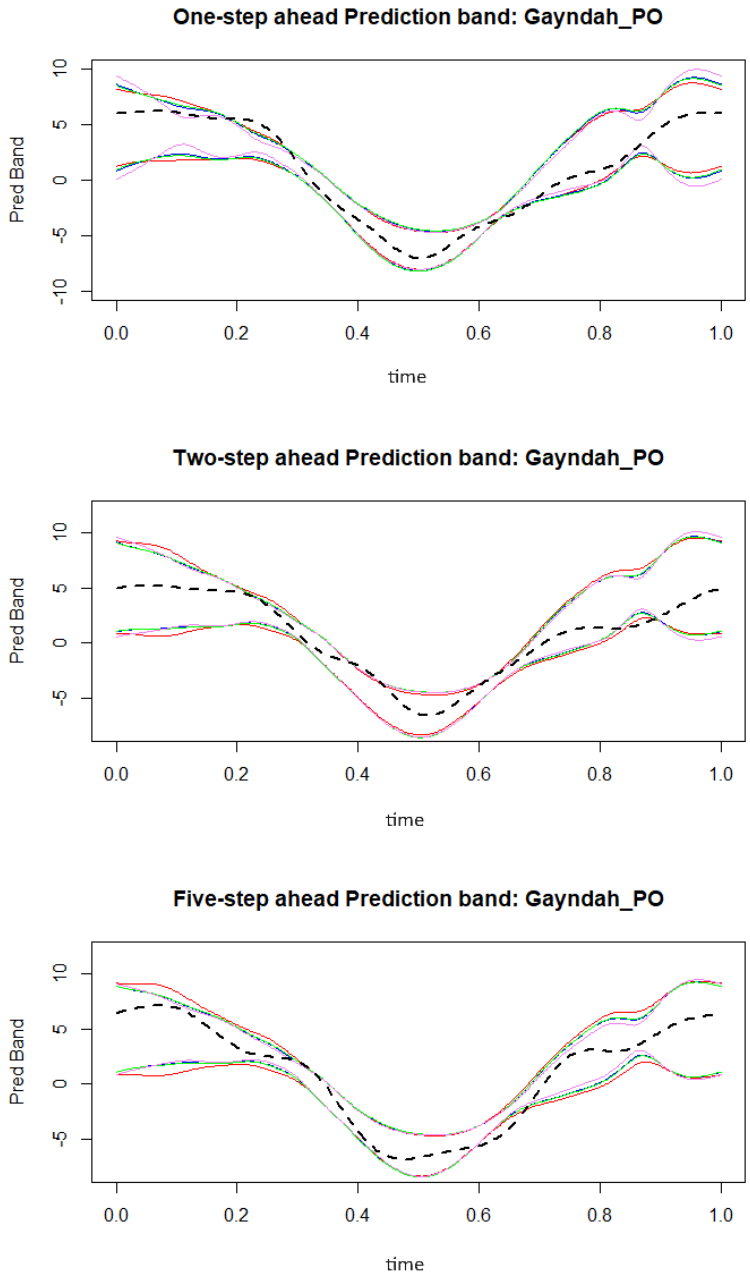

FIGURE 4.18. One, two and five steps Prediction bands for annual temperature profiles of Gayndah Post Office: Empirical(—), Modified empirical (—), Rissanen (—) and Modified Rissanen (—) along with the true functions (**- - -**)

It is evident that even though the temperature profiles for this meteorological station were much more volatile, the prediction bands are able to capture the true data. There is also not much variation in the prediction bands produced by different FPE estimates. Overall, we can conclude that this method works for these annual temperature profiles from Australia.

4.7. Application to real data: Austria particulate matter concentrations

As a second application of the prediction error estimates on real data, daily curves of particulate matter were considered with an aerodynamic diameter of less than $10\mu m$, known as PM_{10}. PM_{10} is routinely measured in most of the major cities in the northern hemisphere, because its high concentration affects health negatively and can cause respiratory and cardiovascular diseases.

There are many causes for high concentrations of particulate matter. The primary cause of high pollution in urban environments is road traffic volume. In addition, strong winter temperature inversions magnify these effects in the cold season. As a result, the limits set by authorities, for example, EU regulation can frequently be violated. In order to meet regulations, prediction of particulate matter concentration levels is an important tool, because it helps to judge whether measures, such as partial traffic regulation, have to be implemented. But in order to accurately implement regulations, not only prediction is required but also a measure is required that can determine how good the predictions are. As a result, measurement of prediction error is important in this context. Data of this type has been considered by a number of authors, including Stadlober et al. (2008), and Dienes and Aue (2014).

Here, the observations are recorded on a half-hourly basis in Graz, Austria, over one winter season, more specifically from October, 2010 to March 2011. Thus, every day's data can be considered as a function measured at 48 discrete points throughout the day. A square root transformation was applied to the data to stabilize the variance. Exploratory data analysis showed that the PM_{10} values were exceedingly high around the New Year's Eve, due to firework activities. The corresponding week's data was removed from the sample. Another adjustment was made due to lower volume of traffic during the weekends than on weekdays and hence PM_{10} is expected to be lower on weekends. Thus, the data was centered and adjusted for weekly seasonality by subtracting the corresponding day of the week average from each observation.

After the initial treatment, 48 observations from a single day were stacked into a vector and then transformed into functional data using 10 B-spline basis functions and least squares fitting. Thus, 175 daily

curves $X_1, X_2, \ldots, X_{175}$ were obtained, which are displayed in the upper left panel of Figure 4.19. The figure also shows the effect of the first three FPCs on the mean curve obtained by adding to and subtracting from the mean curve a multiple (here square root of the l-th eigenvalue) of the l-th empirical eigenfunction.

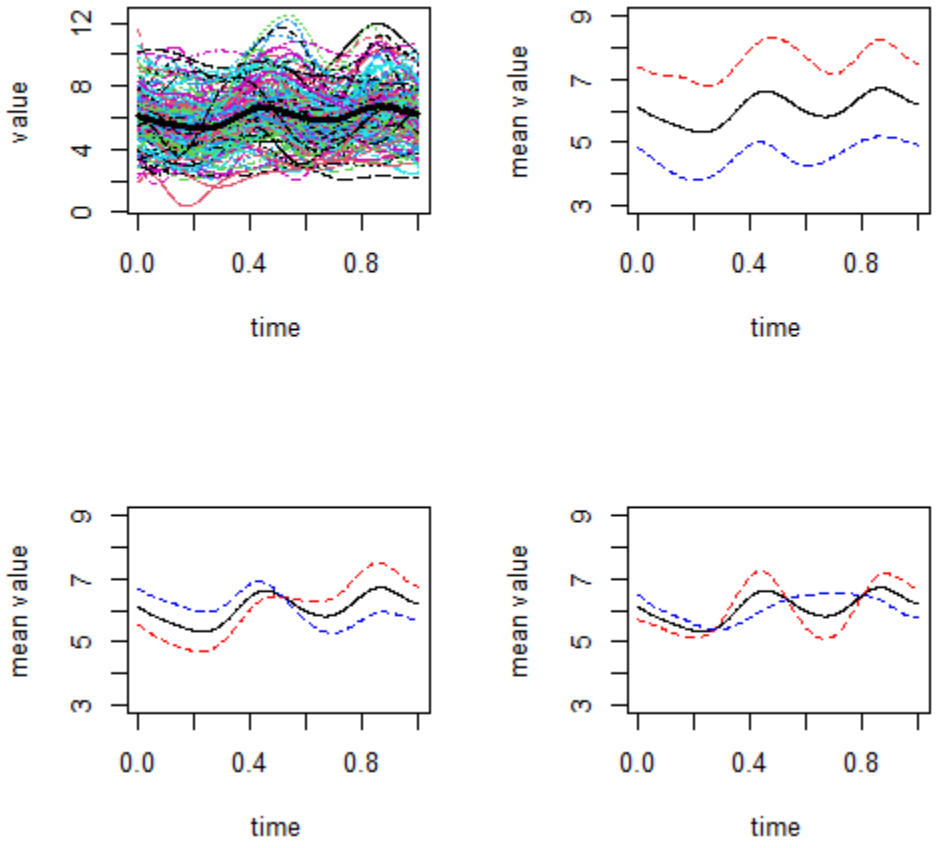

FIGURE 4.19. Transformed PM_{10} observations with overall mean function (top left panel), effect of the first FPC (top right panel), effect of the second FPC (bottom left panel) and effect of the third FPC (bottom right panel): $\hat{\mu} + \hat{\lambda}_l \hat{\phi}_l$ represented as (- - - -) and $\hat{\mu} - \hat{\lambda}_l \hat{\phi}_l$ represented as (- - - -) along with the mean $\hat{\mu}$ (——) for $l = 1, 2, 3$

The top right panel shows that when the first FPC score is large (small) then a positive (negative) mean shift occurs. The bottom left panel shows the effect of the second FPC which describes the intraday trend. The third FPC in the bottom right panel whether the diurnal peaks are more or less pronounced. The data (upper left panel) also shows that the variation in the daily pollution curves are high.

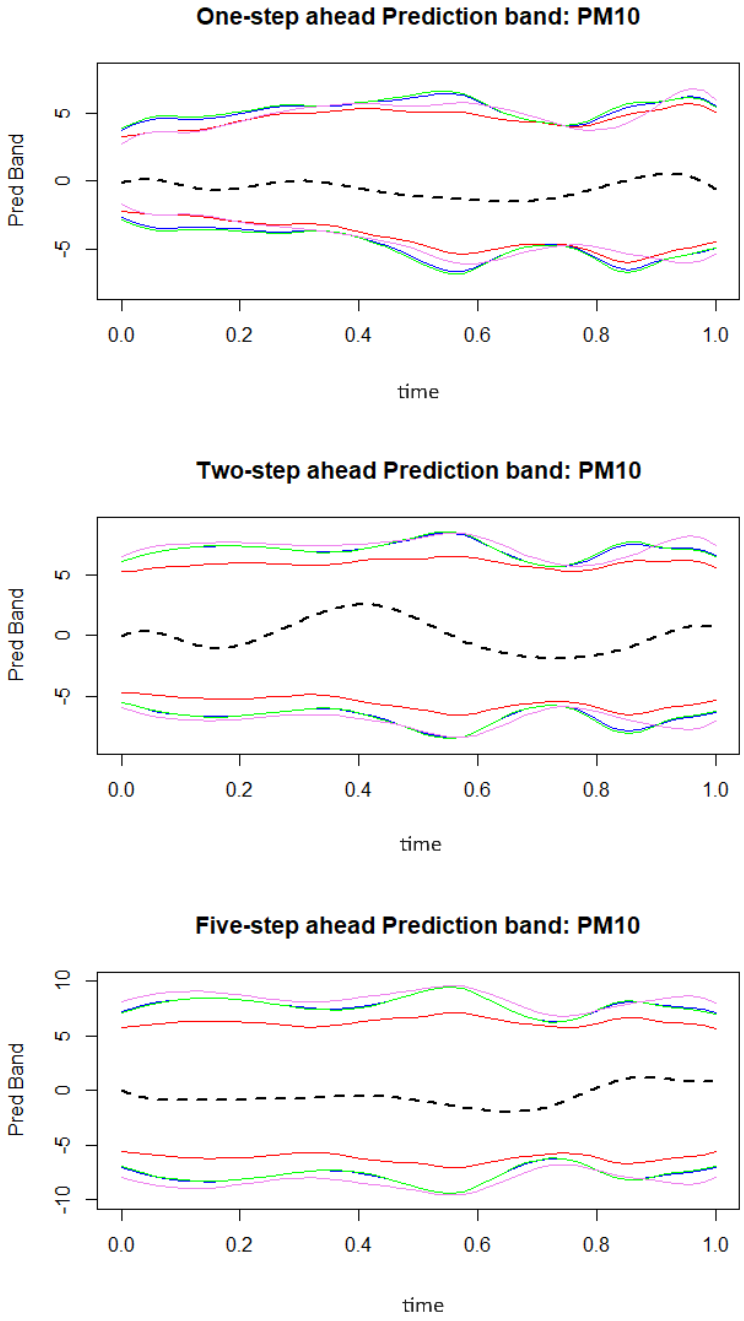

FIGURE 4.20. One, two and five steps Prediction bands for PM_{10} data: Empirical(—), Modified empirical (—), Rissanen (—) and Modified Rissanen (—) along with the true functions (**- - -**)

Here, the optimal p and d as described in Aue et al. (2015) came out to be $p = 1, d = 7$. The higher value of d is attributed to the fact that since the variation is high in the dataset, higher number of eigenvalues are required to explain such high variability. Hence, we would expect the prediction bands to be wide. Prediction bands are computed based on the four different FPE estimates and $\delta = 0.6$ due to the reason mentioned in the Australia datasets. The bands are plotted along with the true function for $h = 1, 2$ and 5 steps.

Indeed, as seen in Figure 4.20, the 95% prediction bands are so wide for all the functional prediction error estimates, that they contain the true curve with 100% certainty. The most narrow band is given by the empirical estimate while the prediction bands produced by the other three estimates are similar for $h = 1, 2$ and 5. And with increase in h, one becomes less certain about the future, hence the prediction band widens further.

One might be curious here to see if the new method of registration of functions as described in the previous section is useful here to reduce the variability in the data and tightening the prediction bands thereafter. Here, the pollution curve for the k-th day is concatenated with last 2 hours' observations (4 observed data) of the previous day and first 2 hours' observations of the next day, $k = 2, \ldots, n-1$. For $k = 1$ and $k = n$, only one sided concatenation is done. The following figure shows the functions after the new registration. There is only a slight reduction in variation in the data, which is not very apparent from the plots.

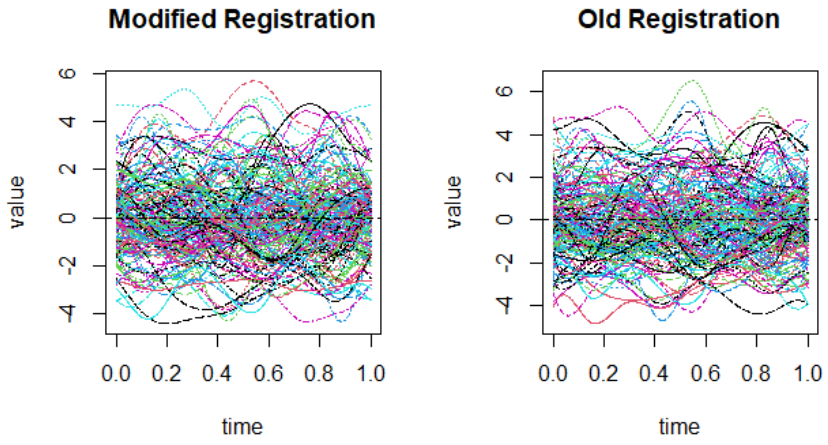

FIGURE 4.21. New and Old registration of functions from PM_{10} observations

Here, $p = 4, d = 3$ came out as the optimal values. It means a VAR(4) model needs to be fitted to the 3-dimensional scores. Using the new p and d, the prediction bands (Figure 4.22) are plotted.

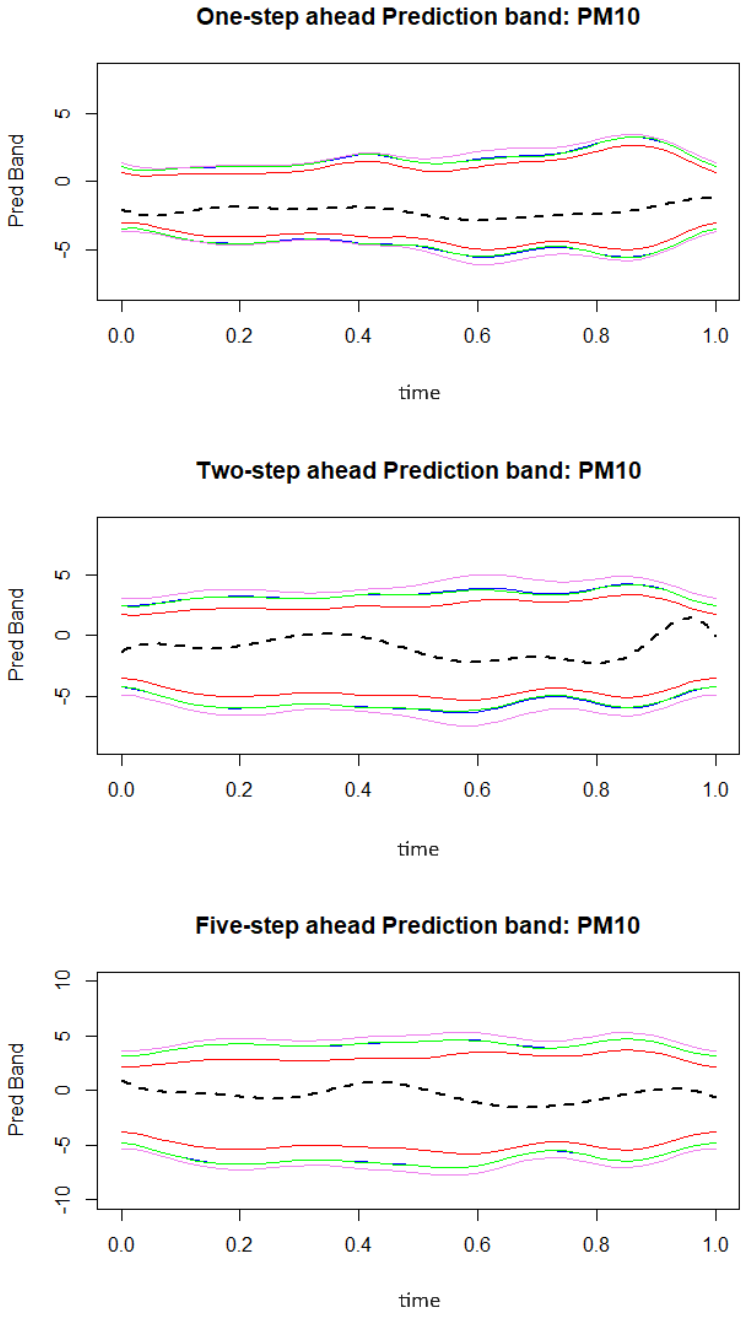

FIGURE 4.22. One, two and five steps Prediction bands for newly registered PM_{10} data: Empirical(—), Modified empirical (—), Rissanen (—) and Modified Rissanen (—) along with the true functions (**- - -**)

Figure 4.23 shows that the new prediction bands are narrower than before, but yet, it is wide enough to contain the true pollution curves 100% of the times. Here also, for all h, the empirical estimate is giving the most narrow prediction band. For $h = 1$, the performances of the other three estimates are similar. However, for $h = 2$ and 5, Rissanen's estimate and modified empirical estimates perform similarly, while whereas the widest prediction band is produced by the modified Rissanen estimate.

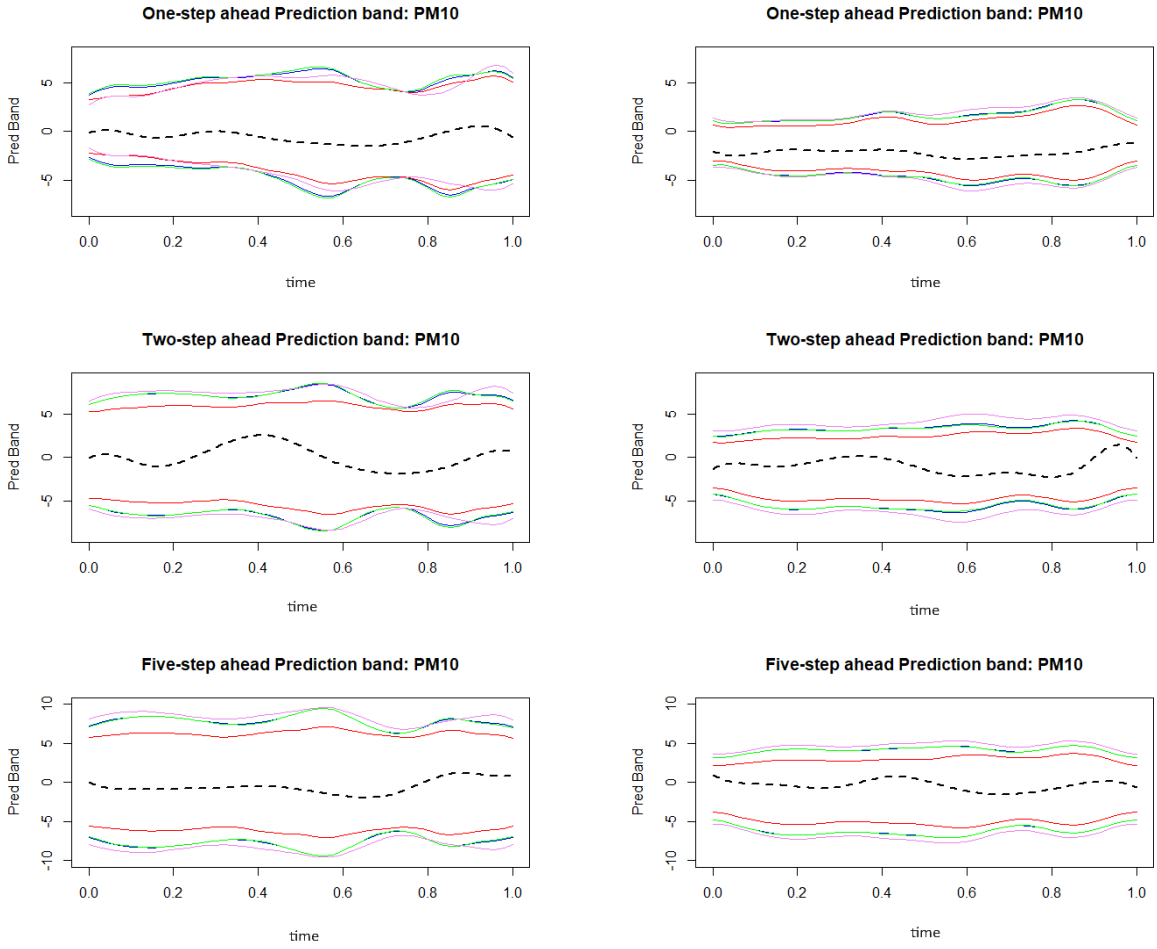

FIGURE 4.23. Comparison of one, two and five steps Prediction bands for old (left) and new (right) registered PM_{10} data: Empirical(—), Modified empirical (—), Rissanen (—) and Modified Rissanen (—) along with the true functions (- - -)

4.8. Conclusion

This chapter introduced prediction error estimates for functional time series. It compared the performances of the different estimates for one-step ahead predictions with the help of simulation studies. As expected, the bias of the estimates decreases with increase in δ, whereas the variance increases. For practical purposes, a δ somewhere around 0.5 will have the optimal bias and variance. Overall, modified empirical estimate seemed to perform the best in terms of bias-variance trade-off. Multi-step prediction error estimates were also defined, which depended on the number of steps ahead the predictions were made.

The prediction error estimates were applied to annual temperature profiles for two meteorological stations in Australia as well as daily pollution curves in Graz, Austria. For the Australia temperatures, the summer temperatures were more volatile, so there were a lot of variations at the beginning and end of each year. A new method of function registration was thus introduced to handle such volatility as well as to ensure continuity of functions from one year to the other to a reasonable extent.

This chapter also introduced methods to construct prediction bands utilizing the multi-step ahead prediction error estimates. Unlike simulation, for real data, the true prediction error was unknown, hence the performances of the different prediction error estimates were compared by constructing prediction bands. We saw that there was not much difference in the prediction bands derived from different prediction error estimates, and the bands were able to capture the different variations of future functions. However, the bands tend to get wider for longer term predictions, because one would be more uncertain when trying to predict more into the future. Similar behavior was observed if the data was more volatile, because the bands would be wider to account for more volatile curves.

Overall, this chapter produced comprehensive analyses based on both simulation and real data on the performances of the proposed prediction error estimates and it can be concluded that they do a satisfactory job to estimate the prediction error for functional time series.

CHAPTER 5

Structure and estimation of functional stochastic volatility

5.1. Financial time series and volatility

Financial time series analysis involves the study and interpretation of data points collected at successive time intervals within the realm of finance. As understood from Taylor (2008) and Andersen et al. (2009), these time series datasets play a crucial role in financial analysis, providing insights into the past performance, trends, and fluctuations of various assets. Analyzing financial time series data is a fundamental aspect of quantitative finance, enabling investors, analysts, and researchers to make informed decisions, model market behaviors, and develop predictive strategies. The dynamic nature of financial markets and the interdependencies among various economic factors make the study of financial time series both challenging and essential for understanding the complexities of the global financial landscape.

However, there are some features distinguishing financial time series from other time series. For example, financial markets experience volatility, representing the magnitude of price fluctuations within a given period. Volatility clustering, where periods of high volatility tend to cluster together, is a prevalent feature in financial time series (Shumway and Stoffer (2000)). This fluctuating volatility poses challenges in predicting future price movements accurately. For a daily stock return series, this volatility is not directly observable (Tsay (2005)), which adds another layer of uncertainty in modeling and estimating volatility. There are other factors such as economic indicators, geopolitical events, investor sentiment and presence of irregularities such as outliers and sudden spikes which require specialized statistical techniques and methods to capture such complexities of financial time series. In this chapter, we are focusing on a particular method of modeling volatility.

Volatility of an asset return is the conditional standard deviation of the underlying asset return. It measures the fluctuation in the price of a financial instrument over time. Volatility plays an important role in risk management. It helps investors and portfolio managers understand the potential range of price movements for an asset or a portfolio and calculate metrics like Value at Risk (VaR) to estimate potential losses

in various market conditions (Bams et al. (2017)). It also plays a pivotal role in options pricing models like the Black–Scholes model (Gong et al. (2010)). Investors consider volatility when constructing portfolios. Modern portfolio theory emphasizes diversification to minimize risk. Volatility helps in selecting assets with different risk profiles to achieve an optimal balance between risk and return (Bouchey et al. (2012)). Furthermore, the volatility of past price series might have a significant impact on investors' forecasting behaviors (Lawrence and Makridakis (1989)).

Tsay (2005) listed a few characteristics of volatility that are crucial in developing models for volatility. For example, volatility clusters are commonly seen in financial time series, that is, volatility may be high for certain periods and low for some other periods. Another important feature is that volatility jumps are rare. Volatility varies within some fixed range and does not diverge to infinity. Tsay (2005) also observed that the log-return series may be serially uncorrelated but dependent. Volatility models attempt to capture such dependence in the return series. One of the most popular methods of modeling the volatility was proposed by Engle (1982) and is called the Autoregressive Conditional Heteroscedastic (ARCH) model which assumes that the dependence in the log returns, can be described by a simple quadratic function of its lagged values. Even though this is a simple and effective model, sometimes it takes higher-order lags to capture the dependence structure. Bollerslev (1986) proposed an extension called the Generalized ARCH (GARCH) model which incorporates not only the past squared returns but also past volatilities to provide accurate volatility predictions. In a separate strand of related research, Taylor (1982) introduced the Stochastic Volatility (SV) model. The main difference of this model from the previous ones is that (G)ARCH models are endogenous, that is, they depend on past observations, whereas stochastic volatility models are exogenous, that is, there is a random term that is not based on the past observations.

5.1.1. Univariate stochastic volatility model. As mentioned, stochastic volatility models add a stochastic noise term to the equation of the volatility. To ensure positivity of the conditional standard deviation, it uses a log-volatility structure. Let P_t be the price of an asset at time index t. Then the (relative) returns y_t are given by $y_t = \frac{P_t - P_{t-1}}{P_{t-1}}$. A stochastic volatility model of order 1 is given by a set of two equations:

$$\begin{aligned} y_t &= \mu_y + \exp(\frac{1}{2}h_t)\varepsilon_t \\ h_t &= \gamma + \phi h_{t-1} + \eta_t \end{aligned} \tag{5.1}$$

where y_t is the time series of asset returns with constant mean μ_y and time-varying variance $\exp(h_t)$, and it is typically assumed that $\varepsilon_t \sim$NID$(0,1)$; $\eta_t \sim$NID$(0,\sigma_\eta^2)$ for $t = 1, \ldots, n$, where NID stands for Normal and independent. Additionally, $\{\varepsilon_t\}$ and $\{\eta_t\}$ are independent for all time points. It is usually assumed that $0 < \phi < 1$. It is to be noted that adding the innovation term η_t significantly increases the flexibility of the model in describing the evolution of the volatility $\sigma_t = \exp(\frac{1}{2}h_t)$. The stochastic time-varying variance of y_t conditional on the past $\mathcal{F}_{t-1}$ is $\sigma_t^2 = E[(y_t - \mu_y)^2|\mathcal{F}_{t-1}] = \exp(h_t)$, where $\mathcal{F}_{t-1}$ is the set of past observations as described in Chapter 1 . We note that y_t is observable but h_t is not. The first equation in (5.1) involving y_t is called the observation equation and the second equation in (5.1) involving h_t is called the state equation.

The above model is an SV(1) model, since the volatility is described by an AR(1) model. Similarly, we can also define an SV(p) model where the state equation is represented by an AR(p) process. The SV(p) model is given by:

$$\begin{aligned} y_t &= \mu_y + e^{(\frac{1}{2}h_t)}\varepsilon_t \\ (1 - \phi_1 B - \cdots - \phi_p B^p)h_t &= \gamma + \eta_t \end{aligned} \tag{5.2}$$

Here, it is also assumed that all zeros of the polynomial $1 - \sum_{i=1}^{p} \phi_i B^i$ are greater than 1 in modulus and B is the back-shift operator. We will focus on the SV(1) model.

The basic SV model is multiplicative due to the product of two stochastic variables, that is $y_t - \mu_y = \exp(\frac{1}{2}h_t)\varepsilon_t$. Estimating μ_y by its consistent estimator, the sample mean of y_t, we define for $t = 1, \ldots, n$,

$$\tilde{y}_t = \log(y_t - \bar{y})^2 \quad \text{where} \quad \bar{y} = n^{-1}\sum_{t=1}^{n} y_t \tag{5.3}$$

Given (5.1), y_t can be modeled by

$$\begin{aligned} \tilde{y}_t &= \kappa_1 + h_t + u_t \\ h_t &= \gamma + \phi h_{t-1} + \eta_t \end{aligned} \tag{5.4}$$

where $u_t = \log(\varepsilon_t^2) - \kappa_1$ is distributed according to the centered logχ^2 density with one degree of freedom. The mean and variance of logε_t^2 are given by κ_1 and κ_2 where $\kappa_1 \approx -1.27$ and $\kappa_2 = \pi^2/2$. The model in (5.4) is linear and the observation disturbance has a non-Gaussian density. However, we may consider u_t

to be a sequence of independent noise terms with mean zero and variance κ_2 and then apply linear methods to obtain estimators of h_t that belong to the class of minimum mean squares linear estimators. If the metric for estimation is chosen to be a Gaussian likelihood, then the approach is called quasi-maximum likelihood analysis. Thus, model (5.4) remains valid with $u_t \sim$ i.i.d. $(0, \kappa_2)$ and it falls under the framework of State-Space models (Shumway and Stoffer (2000)).

5.1.2. State-space models. State space models provide a powerful framework for modeling and analyzing time series data by separating observed measurements from underlying unobserved (hidden) states that evolve over time. They are widely used in various fields such as economics, engineering, finance, and biology. This framework allows for efficient inference and forecasting by estimating the latent states given the observed data.

State-space models or dynamic linear models were introduced by Kalman (Kalman, 1960; Kalman and Bucy, 1961) as an application to primarily aerospace related research. The basic model is given by

$$\begin{aligned} \boldsymbol{y}_t &= A_t \boldsymbol{x}_t + \boldsymbol{v}_t \\ \boldsymbol{x}_t &= \Phi \boldsymbol{x}_{t-1} + \boldsymbol{w}_t \end{aligned} \tag{5.5}$$

where the first equation is the observation equation and the second equation is the state equation which is unobservable. The state equation determines the underlying process to generate the $p \times 1$ state vector $\boldsymbol{x}_t$ from its past for time points $t = 1, \ldots, n$. The observed data vector $\boldsymbol{y}_t$ is $q \times 1$ where q can be larger or smaller than p. The additive observation noise $\boldsymbol{v}_t$ is assumed to be a white noise and Gaussian with $q \times q$ covariance matrix R. It is also assumed that the noise $\boldsymbol{w}_t$ in the state equation are $p \times 1$ independent and identically distributed, zero-mean normal vectors with covariance matrix Q. In addition, it is assumed that the process starts with a normal vector $\boldsymbol{x}_0$ that has mean μ_0 and covariance matrix Σ_0. Further it is also assumed that $\boldsymbol{x}_0, \{\boldsymbol{w}_t\}, \{\boldsymbol{v}_t\}$ are uncorrelated. Thus under this set up, we do not observe the state vector but only a linear transformed version of it which is randomized by adding a noise. The model arose originally in the space tracking setting, where the state equation defines the motion equations for the position or state of a spacecraft with location $\boldsymbol{x}_t$ and $\boldsymbol{y}_t$ reflects information that can be observed from a tracking device such as velocity. However, even though this framework adds flexibility, it also increases the difficulty in estimating the parameters of the model, especially because the underlying state is hidden and unobserved. One common

way to estimate the parameters is using the quasi-likelihood method via Kalman filtering which is explained in the next subsection.

5.1.3. The Kalman filter. The goal of the state space modeling framework given in (5.5) is to get estimators of the underlying unobserved state $\boldsymbol{x}_t$ using the data available until time s given by $Y_s = \{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_s\}$. When $s < t$, it is the problem of prediction, when $s = t$, the problem is called filtering and when $s < t$, the problem of smoothing. Let us introduce the notations following Shumway and Stoffer (2010):

$$\boldsymbol{x}_t^s = E(\boldsymbol{x}_t | \, Y_s) \tag{5.6}$$

and

$$P_{t_1,t_2}^s = E\{(\boldsymbol{x}_{t_1} - \boldsymbol{x}_{t_1}^s)(\boldsymbol{x}_{t_2} - \boldsymbol{x}_{t_2}^s)'\} \tag{5.7}$$

When $t_1 = t_2 = t$ (say) in (5.7), then we can write P_t^s for convenience. Here, we are focusing on the filtering equations. These equations are derived based on Gaussian assumptions of the processes. Under such assumptions, (5.7) is also the conditional error variance, that is,

$$P_{t_1,t_2}^s = E\{(\boldsymbol{x}_{t_1} - \boldsymbol{x}_{t_1}^s)(\boldsymbol{x}_{t_2} - \boldsymbol{x}_{t_2}^s)' | Y_s\}$$

This is because for any t and s, $(\boldsymbol{x}_t - \boldsymbol{x}_t^s)$ and Y_s are orthogonal and uncorrelated. Under Gaussianity, uncorrelated implies independent, hence the conditional distribution of $(\boldsymbol{x}_t - \boldsymbol{x}_t^s)$ given Y_s is same as the unconditional distribution of $(\boldsymbol{x}_t - \boldsymbol{x}_t^s)$.

This method is called filtering because $\boldsymbol{x}_t^t$ is a linear filter of the observations $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_t$ that is

$$\boldsymbol{x}_t^t = \sum_{s=1}^{t} B^s \boldsymbol{y}_s$$

for appropriately chosen $p \times q$ matrices B_s. The Kalman filter helps in specifying how to update the filter from $\boldsymbol{x}_{t-1}^{t-1}$ to $\boldsymbol{x}_t^t$ when a new observation $\boldsymbol{y}_t$ is added to the data set without having to reprocess the entire data.

Given the state space model in equation (5.5) with initial estimates $\boldsymbol{x}_0^0 = \mu_0$ and $P_0^0 = \Sigma_0$, we have for $t = 1, \ldots, n$,

$$\boldsymbol{x}_t^{t-1} = \Phi \boldsymbol{x}_{t-1}^{t-1} \tag{5.8}$$

$$P_t^{t-1} = \Phi P_{t-1}^{t-1} \Phi' + Q \tag{5.9}$$

with

$$\boldsymbol{x}_t^t = \boldsymbol{x}_t^{t-1} + K_t \left(\boldsymbol{y}_t - A_t \boldsymbol{x}_t^{t-1} \right), \tag{5.10}$$

$$P_t^t = \left[I - K_t A_t \right] P_t^{t-1}, \tag{5.11}$$

where

$$K_t = P_t^{t-1} A_t' \left[A_t P_t^{t-1} A_t' + R \right]^{-1} \tag{5.12}$$

is called the Kalman gain. Thus, starting at $\boldsymbol{x}_0^0$ and P_0^0, one can arrive, for any $1 \leq t \leq n$, at $\boldsymbol{x}_{t-1}^{t-1}$ and P_{t-1}^{t-1}. Then one can use (5.8) and (5.9) to do the prediction of the state for step t. Then, the prediction of the observation $\boldsymbol{y}_t$ is obtained as $E\left(\boldsymbol{y}_t \mid Y_{t-1}\right) = A_t \boldsymbol{x}_t^{t-1}$. With this, using equation (5.10), the current state $\boldsymbol{x}_t^t$ is updated along with its covariance matrix P_t^t. Predictions for $t > n$ can be similarly obtained using (5.8) and (5.9) with initial conditions $\boldsymbol{x}_0^0$ and P_n^n.

Thus, we saw that from this method, we also get the innovations or prediction errors given by

$$\boldsymbol{\epsilon}_t = \boldsymbol{y}_t - E\left(\boldsymbol{y}_t \mid Y_{t-1}\right) = \boldsymbol{y}_t - A_t \boldsymbol{x}_t^{t-1} \tag{5.13}$$

with the corresponding variance-covariance matrix

$$\Sigma_t \stackrel{\text{def}}{=} \operatorname{Var}\left(\boldsymbol{\epsilon}_t\right) = \operatorname{Var}\left[A_t \left(\boldsymbol{x}_t - \boldsymbol{x}_t^{t-1} \right) + \boldsymbol{v}_t \right] = A_t P_t^{t-1} A_t' + R \tag{5.14}$$

Under the assumption of Gaussian processes, the innovations are also independent, Gaussian with mean 0 and covariance Σ_t. Hence, a likelihood based method can be used to estimated the model parameters based on (5.13).

5.1.4. Maximum likelihood estimation. Let us denote the vector of parameters of the state space model defined in (5.5) by $\Theta = \{\boldsymbol{\mu}_0, \Sigma_0, \Phi, Q, R\}$, consisting of the initial mean $\boldsymbol{\mu}_0$ and covariance matrix

Σ_0, the transition matrix Φ and the error covariance matrices Q and R. The maximum likelihood method is derived based on the assumption that the initial state is normal, that is, $\boldsymbol{x}_0 \sim N(\boldsymbol{\mu}_0, \Sigma_0)$ and the errors $\{\boldsymbol{v}_t\}$ and $\{\boldsymbol{w}_t\}$ are jointly normal and uncorrelated vector variables.

The likelihood is computed based on the innovations $\boldsymbol{\epsilon}_t$ given in (5.13) by noting the fact that $\boldsymbol{\epsilon}_1, \ldots, \boldsymbol{\epsilon}_n$ are independent mean zero Gaussian random vectors, with covariance matrix given by

$$\Sigma_t = A_t P_t^{t-1} A_t' + R$$

which can be denoted as $\Sigma_t(\Theta)$ to emphasize the dependence of the innovations on the model parameters. Thus, ignoring the constants, the negative log-likelihood can be written as

$$-\ln L_Y(\Theta) = \frac{1}{2}\sum_{t=1}^{n} \log|\Sigma_t(\Theta)| + \frac{1}{2}\sum_{t=1} \boldsymbol{\epsilon}_t' \Sigma_t(\Theta)^{-1} \boldsymbol{\epsilon}_t \tag{5.15}$$

This likelihood is highly nonlinear and a complicated function of the model parameters. Hence iterative optimization techniques like Newton–Raphson methods are implemented after fixing the initial state $\boldsymbol{x}_0$ to minimize the negative log-likelihood and the estimates of the parameters Φ, Q, R are obtained by minimizing the negative log likelihood in an iterative way.

5.1.5. State-space model and stochastic volatility. We will now describe how the SV model described in (5.4) is akin to the state-space framework described by (5.5). For that we note from (5.4), $E(h_t) = \mu_h$ (say) is $\mu_h = E(\tilde{y}_t) - \kappa_1$ and $\gamma = (1-\phi)\mu_h$. Let us define, $h_t^* = h_t - \mu_h$ and $y_t^* = \tilde{y}_t - E(\tilde{y}_t)$. Then, we can reformulate (5.4) as

$$\begin{aligned} y_t^* &= h_t^* + u_t \\ h_t^* &= \phi h_{t-1}^* + \eta_t \end{aligned} \tag{5.16}$$

Thus, for all practical purposes, we can take $\tilde{y}_t - \bar{\tilde{y}}_t$ as the mean centered observations and denote them as our new observations y_t^*, where $\bar{\tilde{y}}_t = n^{-1}\sum_{t=1}^{n} \tilde{y}_t$ and the new mean centered state h_t^*. Under this SV model, the parameter vector is given by $\Theta = (\phi, \sigma_\eta)'$. However, real life data will not always have a Gaussian noise (ε_t) for the observation equation. Hence, the variance of $\log(\varepsilon_t^2)$ might not be κ_2. Then, we can have σ_ε also as a parameter and the parameter vector becomes $\Theta = (\phi, \sigma_\eta, \sigma_\varepsilon)'$. This now corresponds to the state space modeling framework described in (5.5), but instead of vectors, we have scalar observations

and state. That means, here, $p = q = 1, A_t = 1, \boldsymbol{y}_t = y_t^*, \boldsymbol{x}_t = h_t^*, \boldsymbol{v}_t = u_t, \boldsymbol{w}_t = \eta_t$ and Φ, Q, R are not matrices but scalar valued parameters $\Phi = \phi, Q = \sigma_\eta^2, R = \sigma_u^2$. The only difference is the Gaussian assumption of the observation noise u_t which is not Gaussian, hence the likelihood method of estimation is called the Quasi Maximum Likelihood Estimation. We can now show how Kalman filtering methods can be used to estimate the model parameters given by Θ.

Let us assume the model set up given in (5.16). Let us further assume that for $t = 1, \ldots, n; u_t \sim$ i.i.d. $N(0, \sigma_u^2)$ and $\eta_t \sim$ i.i.d. $N(0, \sigma_\eta^2)$. The initial state $h_0^* \sim N(0, \sigma_\eta^2/(1-\phi^2))$, $\{u_t\}, \{\eta_t\}, h_0^*$ are all independent. Since h_t^* is a stationary AR(1) process, we can exploit the properties of AR(1) process to obtain the initial estimates required for the Kalman filtering method. We know that the autocovariance function of h_t^* for lag j is

$$\gamma_h(j) = \frac{\sigma_\eta^2}{1-\phi^2}\phi^j, \qquad j = 0, 1, 2, \ldots \tag{5.17}$$

Here, the goal is to investigate how the presence of observation noise u_t affects the dynamics of the AR(1) model of the state. Note that we have assumed $0 < \phi < 1$, hence both h_t^* and y_t^* are stationary, because the observations are the sum of two independent stationary components. We then have,

$$\gamma_y(0) = \mathrm{Var}(y_t^*) = \mathrm{Var}(h_t^* + u_t) = \frac{\sigma_\eta^2}{1-\phi^2} + \sigma_u^2 \tag{5.18}$$

and when $j \geq 1$,

$$\gamma_y(h) = \mathrm{Cov}(y_t^*, y_{t-j}^*) = \mathrm{Cov}(h_t^* + u_t, h_{t-j}^* + u_{t-j}) = \gamma_h(j) \tag{5.19}$$

Consequently, the ACF of the observations for $j \geq 1$ is given by

$$\rho_y(j) = \frac{\gamma_y(h)}{\gamma_y(0)} = \left(1 + \frac{\sigma_u^2}{\sigma_\eta^2}(1-\phi^2)\right)^{-1} \phi^j \tag{5.20}$$

For applying Kalman filtering, we need the initial estimates of the parameter vector $\Theta = (\phi, \sigma_\eta, \sigma_\varepsilon)'$ from the observations, which we can now get by utilizing the ACF structure of y_t^* given in (5.19). Thus, we have the initial estimate of $\phi^{(0)}$ given by

$$\phi^{(0)} = \hat{\rho}_y(2)/\hat{\rho}_y(1) \tag{5.21}$$

and from (5.19) and (5.18), we get,

$$\sigma_{\eta}^{2^{(0)}} = \left(1 - \phi^{(0)^2}\right) \widehat{\gamma}_y(1)/\phi^{(0)} \tag{5.22}$$

$$\sigma_{u}^{2^{(0)}} = \widehat{\gamma}_y(0) - \frac{\sigma_{\eta}^{2^{(0)}}}{\left(1 - \phi^{(0)^2}\right)} \tag{5.23}$$

Once we have the initial estimates, we can apply the Kalman filtering with Newton–Raphson methods to get the model parameter estimates by minimizing the log-likelihood specified in (5.15). With this background, we can now turn our attention to a functional version of the stochastic volatitlity model.

5.2. Functional stochastic volatility

With the advent of improved tools, modern technology enables the tracking of high frequency intra-day price movements at tick-by-tick level. It is convenient to view the underlying stochastic process and its volatility as a daily function. In such a set-up, where intra-day volatility movements are considered functions, models are needed to capture the heteroskedasticity, as mentioned in Section 5.1, exhibited through clustering tendency of the volatility. Modeling time-varying volatility is essential for accurate uncertainty quantification in forecasting problems. Hörmann et al. (2013) proposed functional ARCH processes to capture heteroskedasticity, whereas Aue et al. (2017) approached the same problem by proposing a generalized framework with a functional GARCH process. Both of these processes rely on modeling the conditional volatility as a deterministic function of past data. Jang et al. (2021) proposed a functional version of the stochastic volatility model, where the volatility functions are driven by their own stochastic process. However, this process was based on a Bayesian hierarchical time series framework. The subsequent sections of this chapter propose an alternative approach to estimate a functional stochastic volatility model based on Kalman filtering methods.

In this work, it is assumed that the observations are elements of a Hilbert Space $H = L^2[0,1]$, which is the set of measurable real-valued functions x defined on [0,1] satisfying $\int_0^1 x^2(t)dt < \infty$. It is a separable Hilbert space with the inner product $\langle x, y\rangle = \int x(t)y(t)dt$ where for future reference the integral sign without limits will denote integration over [0,1]. $\mathcal{L}(H)$ denotes the space of bounded linear operators on H.

5.2.1. Model. Let $\{\eta_i\}_{i\in Z}, \{\epsilon_i\}_{i\in Z}$ be two sequences of i.i.d. random functions defined on a Hilbert Space H, independent of each other. A Functional Stochastic Volatility process $\{y_i\}_{i\in Z}$ of order p, denoted

by $\mathrm{FSV}(p)$ can be defined in terms of point-wise multiplication of two functions as

$$y_i = \sigma_i \eta_i \tag{5.24}$$

$$\log\sigma_i^2 = \delta + \sum_{j=1}^{p} \alpha_j(\log\sigma_{i-j}^2) + \epsilon_i \tag{5.25}$$

where δ is a function of $t \in [0,1]$ and the integral operators α_j for $t \in [0,1]$ are defined as

$$(\alpha_j x)(t) = \int_0^1 \alpha_j(t,s)x(s)ds \tag{5.26}$$

where x is an arbitrary element in H. There are two time variables noted in (5.24) and (5.25). The one labeled by the integer i often refers to trading day i, even though other time units are possible. The second time variable is labeled by real valued t which without loss of generality takes values between [0,1]. This variable is latent in (5.24) and (5.25) and it refers to intra-day time.

5.2.2. Existence of stationary solutions. Note that for $i \in Z$, if $\log\sigma_i^2$ is a strictly stationary process, so are σ_i and y_i. Let $Z_i = \log\sigma_i^2$, then (5.25) can be written in a state space form as

$$\begin{cases} Z_i = \delta + \alpha_1 (Z_{i-1}) + \ldots + \alpha_p (Z_{i-p}) + \epsilon_i \\ Z_{i-j} = Z_{i-j}, \qquad j = 1, \ldots, p-1 \end{cases} \tag{5.27}$$

Further, (5.27) can be written in vector form as

$$\mathbf{Z_i} = \boldsymbol{\delta} + \boldsymbol{\Psi}(\mathbf{Z_{i-1}}) + \boldsymbol{\epsilon_i} \tag{5.28}$$

where, $\mathbf{Z_i} = (Z_i, \ldots, Z_{i-p+1})^T$, $\boldsymbol{\delta} = (\delta, 0, \ldots, 0)^T$ and $\boldsymbol{\epsilon_i} = (\epsilon_i, 0, \ldots, 0)^T$. Further, $\boldsymbol{\Psi} \in H \times \cdots \times H$ is defined as:

$$\begin{bmatrix} \alpha_1 & \alpha_2 & \ldots & \ldots & \alpha_p \\ I_H & 0 & \ldots & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \ldots & I_H & 0 \end{bmatrix}$$

All 0's in the above matrix denote zero-operators. Equation (5.28) is a FAR(1) process. We have the following theorem.

THEOREM 5.2.1. *There is a unique strictly stationary causal solution to (5.28) if there exists an integer j_0 such that $\|\mathbf{\Psi}^{j_0}\| < 1$. The solution is given by*

$$\mathbf{Z_i} = \sum_{j=0}^{\infty} \mathbf{\Psi}^j(\delta + \epsilon_{i-j}) \tag{5.29}$$

The series converges almost surely and in L^2.

Proof: On subsequent iterations, we get,

$$\begin{aligned}
\mathbf{Z}_i &= \boldsymbol{\delta} + \mathbf{\Psi}(\boldsymbol{Z}_{i-1}) + \boldsymbol{\epsilon}_i \\
&= \boldsymbol{\delta} + \mathbf{\Psi}[\boldsymbol{\delta} + \mathbf{\Psi}(\boldsymbol{Z}_{i-2}) + \boldsymbol{\epsilon}_{i-1}] + \boldsymbol{\epsilon}_i \\
&= \boldsymbol{\delta} + \mathbf{\Psi}(\boldsymbol{\delta}) + \mathbf{\Psi}^2(\mathbf{Z}_{i-2}) + \boldsymbol{\epsilon}_i + \mathbf{\Psi}(\boldsymbol{\epsilon}_{i-1}) \\
&= \dots \\
&= \sum_{j=0}^{N-1} \mathbf{\Psi}^j(\boldsymbol{\delta}) + \sum_{j=0}^{N-1} \mathbf{\Psi}^j(\boldsymbol{\epsilon}_{i-j}) + \mathbf{\Psi}^N(\mathbf{Z}_{i-N})
\end{aligned}$$

for some $N \geq j_0$. Now, as $N \to \infty$, we have from Equation (5.29),

$$\begin{aligned}
E\|\mathbf{Z}_i - \sum_{j=0}^{N-1} \mathbf{\Psi}^j(\boldsymbol{\delta} + \boldsymbol{\epsilon}_{i-j})\|^2 &= E\|\mathbf{\Psi}^N(\mathbf{Z}_{i-N})\|^2 \\
&\leq \|\mathbf{\Psi}^N\|_{\mathcal{L}}^2 E\|\mathbf{Z}_{i-N}\|^2 \\
&\leq \|\mathbf{\Psi}^{j_0}\|_{\mathcal{L}}^{\frac{2N}{j_0}} E\|\mathbf{Z}_{i-N}\|^2 \to 0
\end{aligned}$$

since $\|\mathbf{\Psi}^{j_0}\|_{\mathcal{L}} < 1$ and $\mathbf{Z}_{i-N} \in L^2$.

Now, for almost sure convergence, we have

$$\begin{aligned}
\|\mathbf{Z}_i - \sum_{j=0}^{\infty} \mathbf{\Psi}^j(\boldsymbol{\delta} + \boldsymbol{\epsilon}_{i-j})\| &= \| - \sum_{j=N}^{\infty} \mathbf{\Psi}^j(\boldsymbol{\delta} + \boldsymbol{\epsilon}_{i-j}) + \mathbf{\Psi}^N(\mathbf{Z}_{i-N})\| \\
&\leq \| \sum_{j=N}^{\infty} \mathbf{\Psi}^j(\boldsymbol{\delta} + \boldsymbol{\epsilon}_{i-j})\| + \|\mathbf{\Psi}^N(\mathbf{Z}_{i-N})\|
\end{aligned}$$

From the definition of operator norm, we have,

$$\|\mathbf{\Psi}^N(\mathbf{Z}_{i-N})\| \leq \|\mathbf{\Psi}^N\|_{\mathcal{L}}\|\mathbf{Z}_{i-N}\| \xrightarrow{a.s.} 0 \tag{5.30}$$

since, due to stationarity, we have the sequence $\{\|\mathbf{Z}_i\|\}_{i\in\mathbb{Z}}$ bounded almost surely. Further,

$$\begin{aligned}\|\sum_{j=N}^{\infty}\mathbf{\Psi}^j(\boldsymbol{\delta}+\boldsymbol{\epsilon}_{i-j})\| &\leq \sum_{j=N}^{\infty}\|\mathbf{\Psi}^j(\boldsymbol{\delta}+\boldsymbol{\epsilon}_{i-j})\| \\ &\leq \sum_{j=N}^{\infty}\|\mathbf{\Psi}^j\|_{\mathcal{L}}\|\boldsymbol{\delta}+\boldsymbol{\epsilon}_{i-j}\|\end{aligned}$$

Now, $\|\mathbf{\Psi}^j\|_{\mathcal{L}} < 1$ after some j_0. Also, $\|\boldsymbol{\delta}+\boldsymbol{\epsilon}_{i-j}\|$ is bounded since $\boldsymbol{\delta}$ and $\boldsymbol{\epsilon}_i$ are elements of L^2. Since $N \geq j_0$, we have from Markov's inequality, for all $\xi > 0$ and $\forall i$,

$$\begin{aligned}P(\sum_{j=N}^{\infty}\|\mathbf{\Psi}^j\|_{\mathcal{L}}\|\delta+\boldsymbol{\epsilon}_{i-j}\| > \xi) &\leq \frac{1}{\epsilon}E(\sum_{j=N}^{\infty}\|\mathbf{\Psi}^j\|_{\mathcal{L}}\|\delta+\boldsymbol{\epsilon}_{i-j}\|) \\ &= \frac{1}{\epsilon}\sum_{j=N}^{\infty}\|\mathbf{\Psi}^j\|_{\mathcal{L}}E(\|\delta+\boldsymbol{\epsilon}_{i-j}\|)\end{aligned}$$

Since, $\|\mathbf{\Psi}^j\|_{\mathcal{L}} \to 0$ exponentially fast, we have

$$P(\sum_{j=N}^{\infty}\|\mathbf{\Psi}^j\|_{\mathcal{L}}\|\boldsymbol{\delta}+\boldsymbol{\epsilon}_{i-j}\| > \xi) \to 0$$

which implies

$$P(\|\sum_{j=N}^{\infty}\mathbf{\Psi}^j(\boldsymbol{\delta}+\boldsymbol{\epsilon}_{i-j})\| > \xi) \to 0$$

Therefore,

$$\sum_{i=1}^{\infty}P(\|\sum_{j=N}^{\infty}\mathbf{\Psi}^j(\boldsymbol{\delta}+\boldsymbol{\epsilon}_{i-j})\| > \xi) < \infty$$

By the Borel–Cantelli Lemma, we have,

$$\|\sum_{j=N}^{\infty}\mathbf{\Psi}^j(\boldsymbol{\delta}+\boldsymbol{\epsilon}_{i-j})\| \xrightarrow{a.s.} 0 \tag{5.31}$$

From (5.30) and (5.31), therefore, we get,

$$P(\|\mathbf{Z}_i - \sum_{j=0}^{\infty} \mathbf{\Psi}^j(\boldsymbol{\delta} + \boldsymbol{\epsilon}_{i-j})\| = 0) = 1$$

which proves almost sure convergence.

5.3. Quasi-maximum likelihood estimation (QML)

For simplicity, we will focus on lag-order 1 of the $\mathrm{FSV}(p)$ model defined in equation (5.25), which is the $\mathrm{FSV}(1)$ model given by

$$y_i = \sigma_i \eta_i \tag{5.32}$$

$$\log\sigma_i^2 = \delta + \alpha(\log\sigma_{i-1}^2) + \epsilon_i \tag{5.33}$$

where it is typically assumed that $\|\alpha\|_{\mathcal{L}} < 1$ to assume stationarity. Higher-order lags can also be considered but would involve additional computational challenges, for example the likelihood might be more difficult to optimize. Hence the following method focuses on the lag-1 FSV model.

To further facilitate the estimation procedure, it is assumed that E(ϵ_i) = 0 and E(η_i) = 0. Since Equation (5.32) is multiplicative, it can be redefined as

$$\log y_i^2 = E(\log\eta_i^2) + h_i + \xi_i \tag{5.34}$$

$$h_i = \delta + \alpha(h_{i-1}) + \epsilon_i \tag{5.35}$$

where $\xi_i = \log\eta_i^2 - E(\log\eta_i^2)$ and $h_i = \log\sigma_i^2$. The statistical properties of ξ_i will depend on the distribution of η_i. It can be illustrated in the following example.

EXAMPLE 5.3.1. *If we assume that* $\eta_i = W_i(t) = \sqrt{t}X_i, X_i \sim N(0,1)$, *extending the work of Ruiz (1994) to the continuous case. Then,* $\log\eta_i^2(t) = \log(t) + \log(X_i^2)$. *The mean and variance of* $\log(X_i^2)$ *are known to be* $\psi(\frac{1}{2}) - \log(\frac{1}{2}) \approx -1.27$ *and* $\pi^2/2$ *respectively, where,* $\psi(\cdot)$ *is the Digamma function, see Abramowitz and Stegun (1968). Under this set-up, the mean and variance of* $\log\eta_i^2(t)$ *is* $\log(t) - 1.27$

and $\pi^2/2$ and the QML can be carried out by treating $\xi_i(t) \sim N(0, \pi^2/2)$. The actual form of η_i will be explicitly mentioned in the simulation set up Section 5.4.

Since α is a linear operator, denoting $m_1(t)$ as the mean function of $h_i(t)$, (5.35) can be written as in (5.36). Noting that expectation commutes with bounded operators (Hörmann and Kokoszka (2012)), we obtain (5.37).

$$h_i - m_1 = \delta - m_1 + \alpha(m_1) + \alpha(h_{i-1} - m_1) + \epsilon_i \tag{5.36}$$

$$\delta - m_1 + \alpha(m_1) = 0 \tag{5.37}$$

Consequently, model (5.34) & (5.35) can be generalized as

$$y_i^* = h_i^* + \xi_i \tag{5.38}$$

$$h_i^* = \alpha(h_{i-1}^*) + \epsilon_i \tag{5.39}$$

where, $y_i^* = \log y_i^2 - E(\log y_i^2)$ and $h_i^* = h_i - m_1$. The quantities $E(\log y_i^2)$ and $m_1(t)$ needs to be estimated from the data and will also depend on the form of $\eta_i(t)$. In the above example (5.3.1), $y_i^* = \log y_i^2 - \log(t) + 1.27 - m_1$ and the estimate of $m_1(t)$ is

$$\hat{m}_1(t) = \frac{1}{n}\sum_{i=1}^{n} \log y_i^2(t) - \log(t) + 1.27$$

With those estimates, δ can be estimated as $\hat{\delta} = \hat{m}_1 - \hat{\alpha}(\hat{m}_1)$.Thus, the problem is reduced to estimation of α and the variance of $\epsilon_i(t)$ and $\xi_i(t)$ in equations (5.38) and (5.39).

5.3.1. Parametrization. Following Aue et al. (2017) and Cerovecki et al. (2019), we introduce an m-dimensional class $\Phi_m = \{\phi_1, \phi_2, \ldots, \phi_m\}$ of orthonormal functions on [0,1] to represent α to approximate the infinite-dimensional parameters. It is assumed that the integral kernel $\alpha(t, s)$ is an element of the span of $\Phi_m \times \Phi_m$, that is,

$$\alpha(t, s) = \sum_{k,l=1}^{m} a_{k,l}\phi_k(t)\phi_l(s) \tag{5.40}$$

With this assumption regarding the integral kernel, the problem of estimating α is reduced to estimating the set of real valued parameters $\{a_{k,l}\colon k,l=1,\ldots,m\}$. We then project $y_1^*,\ldots,y_n^*$ and $h_1^*,\ldots,h_n^*$ onto Φ_m and define the m-dimensional vectors $\mathbf{y_i^{(2)}}=(y_{i,1}^{(2)},\ldots,y_{i,m}^{(2)})^T$ and $\mathbf{h_i^{(2)}}=(h_{i,1}^{(2)},\ldots,h_{i,m}^{(2)})^T$ through their entries $y_{i,k}^{(2)}=\langle y_i^*,\phi_k\rangle$ and $h_{i,k}^{(2)}=\langle h_i^*,\phi_k\rangle$ where $\langle\cdot,\cdot\rangle$ denotes the inner product in L^2. Therefore, (5.38) and (5.39) can be reduced to:

$$\mathbf{y_i^{(2)}}=\mathbf{h_i^{(2)}}+\boldsymbol{\xi_i^{(2)}} \tag{5.41}$$

$$\mathbf{h_i^{(2)}}=A\mathbf{h_{i-1}^{(2)}}+\boldsymbol{\epsilon_i^{(2)}} \tag{5.42}$$

where A is defined as

$$\begin{bmatrix} a_{11} & a_{12} & \ldots & a_{1m} \\ a_{21} & a_{22} & \ldots & a_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \ldots & a_{mm} \end{bmatrix}$$

We then use the quasi-likelihood estimation and Kalman filtering methods to get the estimates of A and the covariance matrices of $\boldsymbol{\xi_i^{(2)}}$ and $\boldsymbol{\epsilon_i^{(2)}}$.

5.3.2. Estimation overview. Using the vector and matrix system, we can project the infinite-dimensional functions to a finite-dimensional setting as in equations (5.41) and (5.42) using the basis representation. This is akin to the state-space model for multivariate time series which can be used to develop a procedure for the estimation of this functional time series.

Let $\mathbf{y_i^{(2)}}$ in (5.41) be denoted by $\mathbf{Y}_i$ for simplicity. The innovations are then defined as

$$\zeta_i=\mathbf{Y}_i-P_{i-1}\mathbf{Y}_i \tag{5.43}$$

where P_i is the orthogonal projection onto $\overline{\text{span}}\{\mathbf{Y}_v\colon -\infty<v\leq i\}$. It is to be noted that the infinite past is not available for practical purposess, so the span is defined on observations available until time i. Let the variances of $\boldsymbol{\epsilon_i^{(2)}}$ and $\boldsymbol{\xi_i^{(2)}}$ be Q and S, respectively. Then, based on (5.41) and (5.42), the parameter of interest can be defined as $\boldsymbol{\theta}=\text{vec}(A,Q,S)\in\Theta$. Let V_θ be the innovation variance. Then, the state-space method allows conveniently to compute the quasi-likelihood under which $\boldsymbol{\xi_i^{(2)}}$ are Gaussian with mean 0 and variance S. The QML estimator $\hat{\theta}$ for the parameter θ based on the sample $\boldsymbol{y}^{(2)}=(\boldsymbol{y}_1^{(2)},\ldots,\boldsymbol{y}_n^{(2)})$ is

defined as

$$\hat{\boldsymbol{\theta}} = \text{argmin}_{\theta \in \Theta} \mathcal{L}(\boldsymbol{\theta}, \boldsymbol{y}^{(2)}) \tag{5.44}$$

where

$$\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{y}^{(2)}) = \frac{1}{n} \sum_{i=1}^{n} [m\log(2\pi) + \log|V_\theta| + \zeta_{\theta,i}^T V_\theta^{-1} \zeta_{\theta,i}] \tag{5.45}$$

The steps for optimizing the likelihood are:

- Select initial values for the parameters, say $\theta^{(0)}$
- Use Kalman filtering and the initial parameters, obtain a set of innovations $\{\hat{\zeta}_{\theta,i}^{(0)}\}$ and error variances $\{\hat{V}_{\theta,i-1}^{(0)}\}$, $i = 1, \ldots, n$.
- Run one iteration of an optimization procedure with $\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{y}^{(2)})$ as the objective function to obtain a new set of estimates, say $\theta^{(1)}$.
- At iteration j, where $j = 1, 2, \ldots,$ repeat step 2 using $\theta^{(j)}$ instead of $\theta^{(j-1)}$ to obtain a new set of innovations. Then repeat step 3 to obtain a new estimate $\theta^{(j+1)}$.
- Stop when the estimates or the likelihood stabilize.

5.3.3. Consistency of the parameter estimates. Consistency of the parameter estimates can be proved using the method described in Whittle (1953). Whittle showed that the least squares estimates obtained from a multivariate stationary ARMA process are equivalent to the maximum likelihood estimates under Gaussian innovations. Based on similar arguments, we will prove consistency of the parameter estimates in this state-space framework.

THEOREM 5.3.1. *The parameter estimates of FSV model obtained from transforming the functions into a multivariate stationary state-space process with Gaussian noise is consistent.*

Proof: For simplicity, let us denote equations (5.41) and (5.42) in the following way:

$$\mathbf{Y}_i = \mathbf{X}_i + \mathbf{Z}_i \tag{5.46}$$

$$\mathbf{X}_i = A\mathbf{X}_i + \mathbf{E}_i \tag{5.47}$$

where, $\mathrm{Cov}(\mathbf{Z}) = S = \Sigma_Z$ and $\mathrm{Cov}(\mathbf{E}) = Q = \Sigma_E$. It can be shown that

$$\begin{aligned}\mathrm{Cov}(\mathbf{Y}_i, \mathbf{Y}_{i-h}) &= \mathrm{Cov}(\mathbf{X}_i, \mathbf{X}_{i-h}), \quad h \neq 0 \\ &= \mathrm{Cov}(\mathbf{X}_i, \mathbf{X}_{i-h}) + S, \quad h = 0\end{aligned}$$

Note that each $\mathbf{Y}_i$ is an m-dimensional vector $(Y_{1i}, \ldots, Y_{mi})'$ and each $\mathbf{X}_i$ is an m-dimensional vector $(X_{1i}, \ldots, X_{mi})'$. Let us denote the cross-covariances of the observation and state by

$$\begin{aligned}\gamma_{jk}(h) &= \mathrm{Cov}(Y_{j,i+h}, Y_{ki}) \\ \gamma^X_{jk}(h) &= \mathrm{Cov}(X_{j,i+h}, X_{ki})\end{aligned} \tag{5.48}$$

for $j, k = 1, \ldots, m, i = 1, \ldots, n$. Then,

$$\begin{aligned}\mathrm{Cov}(\mathbf{Y}_i, \mathbf{Y}_{i-h}) &= \begin{bmatrix} \gamma_{11}(h) & \gamma_{12}(h) & \ldots & \gamma_{1m}(h) \\ \vdots & \vdots & \ddots & \vdots \\ \gamma_{m1}(h) & \gamma_{m2}(h) & \ldots & \gamma_{mm}(h) \end{bmatrix} \\ &= \begin{bmatrix} \gamma^X_{11}(h) & \gamma^X_{12}(h) & \ldots & \gamma^X_{1m}(h) \\ \vdots & \vdots & \ddots & \vdots \\ \gamma^X_{m1}(h) & \gamma^X_{m2}(h) & \ldots & \gamma^X_{mm}(h) \end{bmatrix}, \quad h \neq 0 \\ &= \begin{bmatrix} \gamma^X_{11}(0) + S_{11} & \gamma^X_{12}(0) + S_{12} & \ldots & \gamma^X_{1m}(0) + S_{1m} \\ \vdots & \vdots & \ddots & \vdots \\ \gamma^X_{m1}(0) + S_{m1} & \gamma^X_{m2}(0) + S_{m2} & \ldots & \gamma^X_{mm}(0) + S_{mm} \end{bmatrix}, \quad h = 0\end{aligned} \tag{5.49}$$

Thus, the spectral density of the process $\mathbf{Y}$ is given by

$$\begin{aligned}F^Y_{jk}(\omega) &= \sum_{h=-\infty}^{\infty} \gamma_{jk}(h) e^{i\omega h} \\ &= \sum_{h \neq 0} \gamma^X_{jk}(h) e^{i\omega h} + [\gamma^X_{jk}(0) + S_{jk}] \\ &= \sum_{h=-\infty}^{\infty} \gamma^X_{jk}(h) e^{i\omega h} + S_{jk} \\ &= F^X_{jk}(\omega) + S_{jk}\end{aligned} \tag{5.50}$$

and the corresponding spectral density matrix is given by

$$\begin{aligned}\mathbf{F}^{Y}(\omega) &= \left(F_{jk}^{Y}(\omega)\right) \\ &= \left(F_{jk}^{X}(\omega)\right) + (S_{jk}) \\ &= \mathbf{F}^{X}(\omega) + S\end{aligned} \tag{5.51}$$

Let $f_{jk}^{Y}(\omega)$ be the empirical spectral density based on the sample covariances of $\mathbf{Y} = (\mathbf{Y}_1, \ldots, \mathbf{Y}_n)'$ and $\mathbf{f}^{Y}(\omega) = \left(f_{jk}^{Y}(\omega)\right)$. Thus the spectral density matrix for a state-space process can be expressed in terms of the spectral density matrix of an AR process perturbed by a constant matrix. Hence, we showed that even though Whittle's method was developed for a general ARMA process, which in this case is $\mathbf{X}_i$, it can be implemented to the state-space framework as well.

Under the assumption of Gaussian noise, the negative log-likelihood based on the innovations $\zeta_i = \mathbf{Y}_i - \widehat{\mathbf{Y}}_i$ can be written as

$$-2\log l = (nm)\log(2\pi) + \sum_{i=1}^{n} \log|V_{i-1}| + \sum_{i=1}^{n} \zeta_i^T V_{i-1}^{-1} \zeta_i \tag{5.52}$$

where $V_i = E[\zeta_{i+1}\zeta_{i+1}^T]$ is the prediction error covariance matrix for $\mathbf{Y}_{i+1}, i = 0, 1, \ldots, n-1$. Let us also define a matrix D as

$$D = \begin{pmatrix} V_0 & & \\ & \ddots & \\ & & V_{n-1} \end{pmatrix} \tag{5.53}$$

Following the definition of Wilks (1932) of the total variance of a vector as the determinant of the covariance matrix, Whittle termed the quantity $|V_{i-1}|$ the total prediction variance, which gives a measure of the total random variance entering the process at every step, say step i. It measures the random variation injected intro the process since the last instant of time. Whittle expressed this total prediction variance in terms of the spectral density matrix of the process as

$$\log|V_{i-1}| = \frac{1}{2\pi}\int_0^{2\pi} \log|\mathbf{F}^{Y}(\omega)|d\omega \tag{5.54}$$

where $\mathbf{F}^Y(\omega)$ is given by equation (5.51). Therefore,

$$\sum_{i=1}^{n} \log|V_{i-1}| = \frac{n}{2\pi} \int_0^{2\pi} \log|\mathbf{F}^Y(\omega)| d\omega \tag{5.55}$$

Also notice that

$$\sum_{i=1}^{n} \zeta_i^T V_{i-1}^{-1} \zeta_i = \sum_{i=1}^{n} (\zeta_i^*)^T \zeta_i^* = \sum_{i=1}^{n} \sum_{j=1}^{m} (\zeta_{ji}^*)^2 \tag{5.56}$$

where

$$\zeta_i^* = V_{i-1}^{-1/2} \zeta_i$$

Let us also define

$$\mathbf{Y}_i^* = V_{i-1}^{-1/2} \mathbf{Y}_i$$

such that

$$\mathbf{Y}^* = \begin{pmatrix} \mathbf{Y}_1^* \\ \vdots \\ \mathbf{Y}_n^* \end{pmatrix} = \begin{pmatrix} V_0^{-1/2} & & \\ & \ddots & \\ & & V_{n-1}^{-1/2} \end{pmatrix} \begin{pmatrix} \mathbf{Y}_1 \\ \vdots \\ \mathbf{Y}_n \end{pmatrix} = D^{-1/2} \mathbf{Y}$$

Then the corresponding population and sample spectral density matrix of $\mathbf{Y}^*$ are given by

$$\begin{aligned} \mathbf{F}^*(\omega) &= D^{-1/2} \mathbf{F}^Y(\omega) (D^{-1/2})^T \\ \mathbf{f}^*(\omega) &= D^{-1/2} \mathbf{f}^Y(\omega) (D^{-1/2})^T \end{aligned} \tag{5.57}$$

It is also easy to verify that ζ_i^* and ζ_k^* are uncorrelated for $i \neq k$, since, ζ_i and ζ_k are uncorrelated. Whittle showed that

$$\sum_{i=1}^{n} \sum_{j=1}^{m} (\zeta_{ji}^*)^2 = \frac{n}{2\pi} \int_0^{2\pi} \mathrm{tr}[\mathbf{f}^*(\omega) \mathbf{F}^*(\omega)^{-1}] d\omega \tag{5.58}$$

Now, using equation (5.57), we have

$$\begin{aligned} \mathrm{tr}[\mathbf{f}^*(\omega) \mathbf{F}^*(\omega)^{-1}] &= \mathrm{tr}[D^{-1/2} \mathbf{f}^Y(\omega) (D^{-1/2})^T \left((D^{-1/2})^T \right)^{-1} \mathbf{F}^Y(\omega)^{-1} (D^{-1/2})^{-1}] \\ &= \mathrm{tr}[D^{-1/2} \mathbf{f}^Y(\omega) \mathbf{F}^Y(\omega)^{-1} D^{1/2}] \\ &= \mathrm{tr}[D^{1/2} D^{-1/2} \mathbf{f}^Y(\omega) \mathbf{F}^Y(\omega)^{-1}] \qquad (\text{since } \mathrm{tr}(AB) = \mathrm{tr}(BA)) \\ &= \mathrm{tr}[\mathbf{f}(\omega) \mathbf{F}(\omega)^{-1}] \end{aligned} \tag{5.59}$$

Using equations (5.55), (5.58) and (5.59) in the likelihood equation (5.52), we get

$$-2\log l = (nm)\log(2\pi) + \frac{n}{2\pi}\int_0^{2\pi} \left[\log|\mathbf{F}^Y(\omega)| + \mathrm{tr}[\mathbf{f}(\omega)\mathbf{F}(\omega)^{-1}]\right] d\omega \tag{5.60}$$

Whittle (1953) showed that the least square parameter estimates are obtained by minimizing equation (5.60) which is the negative log-likelihood. Hence, the parameter estimates thus obtained are nothing but the MLEs and any MLE is consistent. It is to be noted that even though Gaussian noise assumption might not valid in this context, it may be reasonable to assume that since smooth functions of the whole empirical spectral density estimate is consistent, and hence the estimates obtained of the FSV model are consistent.

5.4. Simulations

The estimation of the above FSV model was evaluated on simulated data. At first, the logarithm of the stochastic volatility functions $h_i(t) = \log\sigma_i^2(t)$ was generated for $0 < t < 1$. As seen in equation (5.35), h_i follows an FAR(1) structure. In order to generate functions, the `fda` package is used. It is available in the `R` statistical software. An underlying basis system of D basis functions $\{\phi_1, \ldots, \phi_D\}$ is chosen where D is sufficiently large so that it can reasonably mimic the infinite-dimensionality of functions. For generating n functions $h_1, \ldots, h_n$, we first note that each h_i is the sum of three functions, $\delta, \alpha(h_{i-1})$ and ϵ_i, each of which can be represented as a linear combination of the underlying basis system. For example, δ can be written as

$$\delta = \sum_{k=1}^{D} d_k\phi_k$$

for some non-random choice of $d_k, k = 1, \ldots, D$. Similarly, one can write for each $i = 1, \ldots, n$,

$$\alpha(h_{i-1}) = \sum_{k=1}^{D} b_k\phi_k$$

and

$$\epsilon_i = \sum_{k=1}^{D} c_k\phi_k$$

so that we can write

$$h_i = \sum_{k=1}^{D} (d_k + b_k + c_k)\phi_k$$

The exact choice of parameters and basis will be provided in Section 5.4.2.

For generating the coefficients $\{b_k\}$ and $\{c_k\}$, firstly a D-dimensional vector is created represented by

$$\sigma = (\sigma_1, \ldots, \sigma_D)'$$

where σ_j is a point-wise decreasing function of $j = 1, \ldots, D$. Then, a $D \times D$ matrix, which represents a D-dimensional operator in the function space, is created based on σ given by $\Psi = (\psi_{ij})_{D\times D}$ where $\psi_{ij} \sim N(0, \sigma_i\sigma_j)$ represents elements from the product basis. This Ψ is then divided by the largest singular value of Ψ to ensure the norm is 1. It is further multiplied by a constant less than 1 to ensure stationarity of the $\mathrm{FAR}(1)$ model. This matrix is then multiplied with the $D \times 1$ coefficient vector of h_{i-1} to obtain the vector $(b_1, \ldots, b_D)'$. The D coefficients of each of the ϵ_i, given by $(c_1, \ldots, c_D)'$ are generated such that $c_k \sim N(0, \sigma_k)$.

The innovation function of the observation equation (5.32), η_i is chosen to be a Brownian motion for each day $i, i = 1, \ldots, n$. At first, the Brownian motion for each intraday point j is generated for $j = 1, \ldots, T$, where T is the number of intraday time points at which the functions are observed. They are given by the following set of equations:

$$\begin{aligned} S_0 &= 0 \\ S_j &= S_{j-1} + Z_j;\ Z_j \sim N(0,1) \\ B_j &= \frac{1}{\sqrt{T}} S_j \end{aligned}$$

Then, these B_js, which are nothing but the point-wise evaluations of the functions η_i are converted to Brownian motion functions η_i for each day by registering them with a B-spline basis system. The resulting functions are shown in Figure 5.1.

The σ_i functions in equation (5.32) are obtained from $h_i = \log\sigma_i^2$ by transforming the functions. However, in `R`, there is no direct way to transform functions. To handle this problem, it was checked if transforming the functions is same as transforming the pointwise evaluations of the the functions. The steps are outlined below.

5.4.1. Transforming functions. First we note that, $\sigma_i = e^{\frac{1}{2}\log\sigma_i^2} = e^{(h_i/2)}$. Here, we are checking whether transforming functions is equivalent to transforming the pointwise evaluations of the functions.

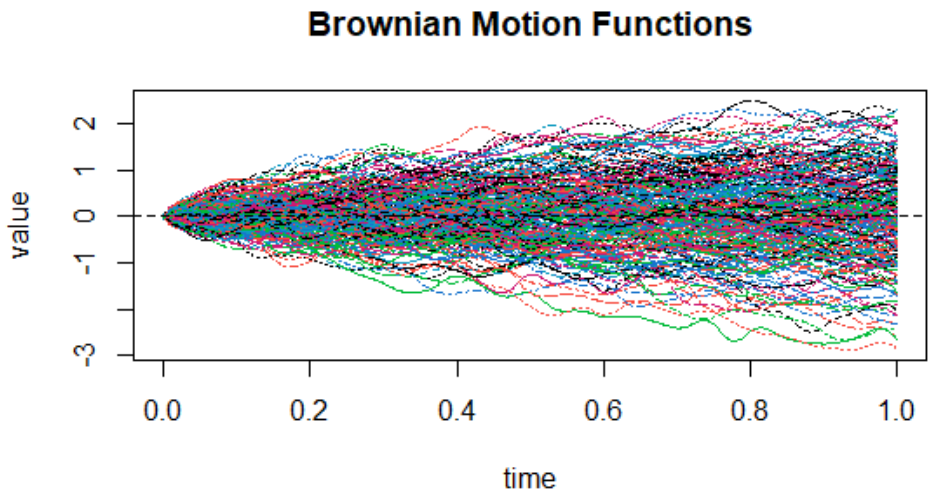

FIGURE 5.1. Brownian Motion Functions

- The $h_i(t) = \log\sigma_i^2(t)$ are generated as functions and evaluated at intraday time points $t_1, \ldots, t_T$ given by $h_i(t_1), \ldots, h_i(t_T)$.
- Then we can transform the pointwise evaluations as $e^{(h_i/2)}$ to get $\sigma_i(t_1), \ldots, \sigma_i(t_T)$.
- We can then convert them back as functions to obtain $\sigma_i(t)$. Note that this is the potential functional form of the stochastic volatility function.
- For a different set of intraday points $s_1, \ldots, s_T$, we evaluate the function $\sigma_i(t)$ to obtain $\sigma_i(s_1), \ldots, \sigma_i(s_T)$.
- We can also get $e^{(h_i/2)}$ evaluated at $s_1, \ldots, s_T$ by evaluating h_i at $s_1, \ldots, s_T$ and taking the transformation.
- We then compute

$$\Delta = \sqrt{\frac{1}{n*T}\sum_{i=1}^{n}\sum_{j=1}^{T}\left(\sigma_i(s_j) - e^{\frac{1}{2}(h_i(s_j))}\right)^2} \tag{5.61}$$

  If this quantity is negligible, we conclude that functions can be transformed using pointwise evaluations.

Once we get the σ_i, we can easily obtain the observation functions y_i as

$$y_i = \sigma_i\eta_i$$

This serves as the functional observations on which estimation procedure needs to be carried out. Once the above functions are obtained, the linearized functions (equation (5.34)) can be easily obtained which subsequently leads to equation (5.38). Notice that in equation (5.38), y_i^* depends on the mean function of h_i given by m_1. So, after plugging in the estimate of m_1, equation (5.38) reduces to

$$\tilde{y}_i = \log y_i^2 - \frac{1}{n}\sum_{i=1}^{n} \log y_i^2 = h_i^* + \xi_i \tag{5.62}$$

So, we use the left-hand side of equation (5.62) as our observations based on which the estimation is done.

5.4.1.1. *Dimension reduction.* As discussed in Section 5.3.1, it can be assumed that a class of orthonormal function serves as the underlying basis system for the observed functions. One common way to obtain such a basis from the data is to consider the eigenfunctions obtained from applying FPCA. It is to be noted that FPCA based estimation is not covered by our theory. We choose the first m eigenfunctions as the basis where the first m eigenvalues explain a significant proportion of variance in the data. Let us denote this basis system as $\Gamma_m = \{\gamma_1, \ldots, \gamma_m\}$. In this system, the integral kernel $\alpha(t, s)$ are elements of $\Gamma_m \times \Gamma_m$ and can be expressed as

$$\alpha(t, s) = \sum_{k,l=1}^{m} a_{k,l}^* \gamma_k(t)\gamma_l(s) \tag{5.63}$$

We then project $\tilde{y}_1, \ldots, \tilde{y}_n$ and $h_1^*, \ldots, h_n^*$ onto Γ_m and define the m-dimensional vectors

$$\mathbf{Y_i} = (y_{i,1}^{(2)}, \ldots, y_{i,m}^{(2)})^T$$

and

$$\mathbf{X_i} = (h_{i,1}^{(2)}, \ldots, h_{i,m}^{(2)})^T$$

through their entries $y_{i,k}^{(2)} = \langle \tilde{y}_i, \gamma_k \rangle$ and $h_{i,k}^{(2)} = \langle h_i^*, \gamma_k \rangle$. $\mathbf{Y_i}$ now represents an m-dimensional multivariate time series which can be placed into the state-space modeling framework. Hence Kalman-filtering methods as described in Lütkepohl, 2005; Shumway and Stoffer, 2000 can be applied to carry out the optimization steps of the likelihood function (equation (5.45)) in the estimation process as mentioned above in Section 2.2.

5.4.2. Choice of parameters for simulations. The logarithm of the stochastic volatility functions h_i are generated using $D = 31$ basis functions. The D dimensional σ vector is chosen as $\sigma = (\sigma_1, \ldots, \sigma_D)'$

where

$$\sigma_j = 1.5^{-j},\ j = 1, \ldots, D$$

which is a decreasing function of j. The FAR(1) operator is multiplied with a coefficient of 0.8 to ensure stationarity of the functions generated. The delta function coefficients were chosen such that the norm of the coefficient vector is smaller than 0.01. The use of the Fourier basis produced a realistic version of the log-volatility functions as seen in Figure 5.2. The Δ defined in Section 5.4.1 came out to be $2.33 * 10^{-14}$ which is almost 0. This proves that we can take transformations of the pointwise evaluations of functions to transform the functions. Thus, the functions σ_i were obtained by point-wise transforming observed h_i.

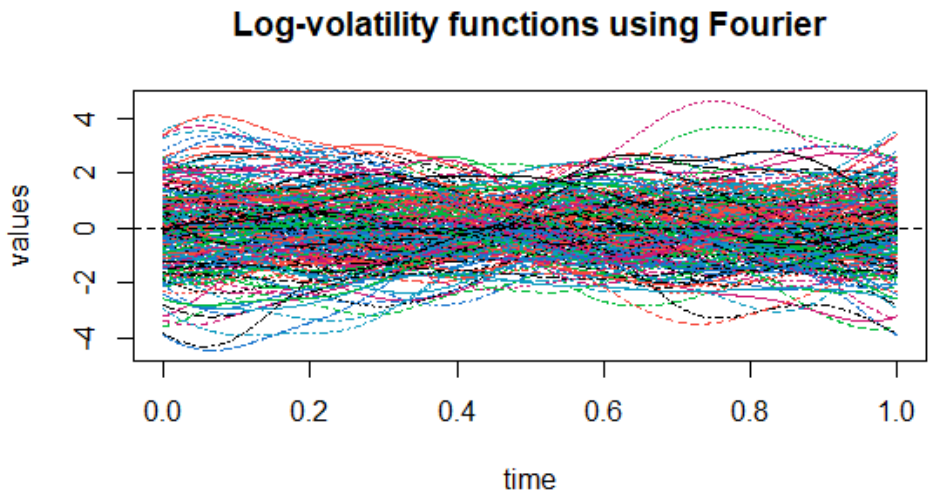

FIGURE 5.2. Log-volatility functions generated using Fourier basis

Once the linearized functions are obtained as given in equation (5.62), FPCA was applied on them. The first five resulting eigenvalues explained around 90% of variations in the data, hence m was chosen to be 5. In other words, Kalman filtering will be applied to 5-dimensional vectors.

5.4.2.1. *Getting initial estimates.* Recall equations (5.46) and (5.47)

$$\mathbf{Y}_i = \mathbf{X}_i + \mathbf{Z}_i$$

$$\mathbf{X}_i = A\mathbf{X}_i + \mathbf{E}_i$$

where $\mathbf{Y_i} = (y_{i,1}^{(2)}, \ldots, y_{i,m}^{(2)})^T$ and $\mathbf{X_i} = (h_{i,1}^{(2)}, \ldots, h_{i,m}^{(2)})^T$ are m-dimensional vectors representing a multivariate time series. Also recall from Section 5.3.2 that the optimization of this state-space likelihood using Kalman filtering requires some initial parameter estimates. The choice of the initial estimates can greatly impact the optimization of the log-likelihood equation (5.45) and good initial estimates will ensure accurate estimates coming out of maximizing the log-likelihood.

Computing the covariance at lag 0 of the above equations lead to

$$\Sigma_Y(0) = \Sigma_X(0) + S \tag{5.64}$$

$$\Sigma_X(0) = A\Sigma_X(0)A^T + Q \tag{5.65}$$

It can be shown that the cross-covariances of the observations are functions of cross-covariances of the state. More specifically:

$$\begin{aligned} \Sigma_Y(1) &= \mathrm{Cov}(\mathbf{Y}_i, \mathbf{Y}_{i-1}) \\ &= A\Sigma_X(0) \end{aligned} \tag{5.66}$$

Similarly,

$$\Sigma_Y(2) = A^2\Sigma_X(0)$$

Thus, we have,

$$A\Sigma_Y(1) = \Sigma_Y(2)$$

So, we need an A that minimizes the norm difference

$$\|A\Sigma_Y(1) - \Sigma_Y(2)\|_F^2$$

and the minimizer of this is given by

$$\tilde{A} = \Sigma_Y(2)\Sigma_Y(1)^T \left(\Sigma_Y(1)\Sigma_Y(1)^T\right)^{-1}$$

Finally, the initial estimate of A is given by

$$\widehat{A}_{ini} = k\frac{\tilde{A}}{\lambda_1(\tilde{A}\tilde{A}^T)} \tag{5.67}$$

where $0 < k < 1$ and this ensures that the norm of the initial estimate is less than 1. For our simulation, k is chosen to be 0.8 since that was the norm of the FAR(1) operator for generating the functions. Now, from equation (5.66), we get

$$\Sigma_X(0) = A^{-1}\Sigma_Y(1)$$

and hence

$$\tilde{\Sigma}_X(0) = \widehat{A}_{ini}^{-1}\Sigma_Y(1)$$

Note that, $\Sigma_X(0)$ needs to be symmetric and positive definite, since it is a covariance matrix. However, $\tilde{\Sigma}_X(0)$ need not be symmetric and postive definite. To make it so, the following is chosen as an estimate of $\Sigma_X(0)$:

$$\widehat{\Sigma}_X(0) = \left(\tilde{\Sigma}_X(0)\tilde{\Sigma}_X(0)^T\right)^{1/2}$$

But this estimate had very large eigenvalues which can be a problem in getting estimates of Σ_Z from equation (5.64). Hence, the initial estimate of $\Sigma_X(0)$ was chosen to be

$$\widehat{\Sigma}_{X_{ini}}(0) = \frac{\widehat{\Sigma}_X(0)}{\lambda_1((\widehat{\Sigma}_X(0))} \tag{5.68}$$

where $\lambda_1((\widehat{\Sigma}_X(0))$ is the largest eigenvalue of $\widehat{\Sigma}_X(0)\widehat{\Sigma}_X(0)^T$. Finally, following from equations (5.64) and (5.65), the initial estimates of S and Q are given by

$$\widehat{S_{ini}} = \Sigma_Y(0) - \widehat{\Sigma}_{X_{ini}}(0) \tag{5.69}$$

$$\widehat{Q_{ini}} = \widehat{\Sigma}_{X_{ini}}(0) - \widehat{A}_{ini}\widehat{\Sigma}_{X_{ini}}(0)\widehat{A}_{ini}^T \tag{5.70}$$

5.4.3. Estimation results. Recall that here we are trying to estimate three $m \times m$ matrices, the AR(1) coefficient matrix A and the covariances of $\boldsymbol{\epsilon}_{\boldsymbol{i}}^{(\mathbf{2})}$ and $\boldsymbol{\xi}_{\boldsymbol{i}}^{(\mathbf{2})}$ given by Q & S, respectively. In order to check how close the estimated matrices are to the true matrices, we considered a number of metrics. For any $m \times m$ matrix $B = (b_{ij})$ and its estimate $\hat{B} = (\hat{b}_{ij})$, the following gives an overview of the metrics:

- Mean Squared Error (MSE):

$$\frac{1}{m^2}\sum_{i=1}^{m}\sum_{j=1}^{m}(b_{ij}-\hat{b}_{ij})^2$$

- Root Mean Squared Error (RMSE):

$$\sqrt{MSE}=\sqrt{\frac{1}{m^2}\sum_{i=1}^{m}\sum_{j=1}^{m}(b_{ij}-\hat{b}_{ij})^2}$$

- Mean Absolute Error (MAE):

$$\frac{1}{m^2}\sum_{i=1}^{m}\sum_{j=1}^{m}|b_{ij}-\hat{b}_{ij}|$$

- Frobenius norm:

$$\|B-\hat{B}\|_F$$

  where $\|B\|_F=\sqrt{\sum_{i=1}^{m}\sum_{j=1}^{m}b_{ij}^2}$

- Relative Frobenius norm:

$$\frac{\|B-\hat{B}\|_F}{\|B\|_F}$$

- Max norm:

$$\|B-\hat{B}\|_M$$

  where $\|B\|_M=\max_{ij}|b_{ij}|$

- Relative Max norm:

$$\frac{\|B-\hat{B}\|_M}{\|B\|_M}$$

- Spectral norm:

$$\|B-\hat{B}\|_2$$

  where $\|B\|_2=\sqrt{\lambda_1}$ where λ_1 is the largest singular value of B

- Relative Spectral norm:

$$\frac{\|B-\hat{B}\|_2}{\|B\|_2}$$

These metrics were then computed for different scenarios. It is to be noted that the parameters A and Q are coming from the observation equation and S is from the state equation of the state-space process. If the

amount of variation explained by the state equation relative to the variation in the noise of the observation equation is high, then the signal is considered high. If the reverse is true, it is considered that the signal is low and noise is high. So we would expect A and Q will be estimated well when the signal is high as compared to noise, however, we would expect S to be estimated well if signal is low and noise is high in the data. See below for more details. It was also shown in Section 5.3 that the estimates are consistent. So, we will show how the metrics changes as we increase the sample size. Let us analyze the metrics in these aspects.

5.4.3.1. *Metrics for low signal.* As mentioned in Section 5.4.2, infinite-dimensional functions were reduced to $m = 5$ dimensional vectors using FPCA. For $m = 5$, we had a total of 55 parameters, 25 for A and 15 each for Q and S since they are symmetric.

Since number of parameters was high, a higher sample size was considered so that the parameters can be estimated well. Here, a size of $n = 1300$ functions was used for estimation.

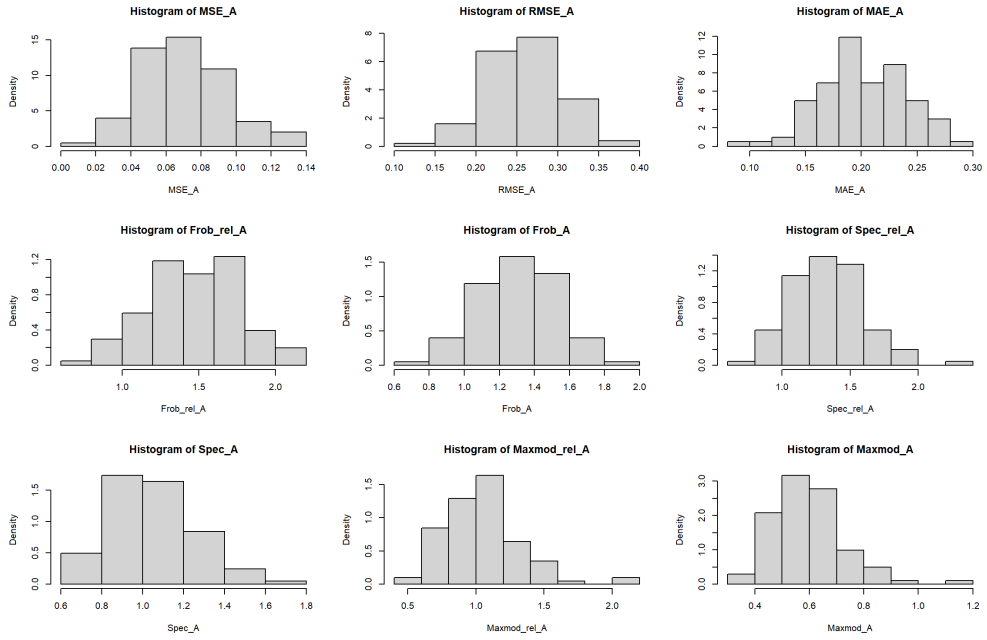

FIGURE 5.3. Histograms of all metrics for A: low signal

Parameter estimates were obtained using methods discussed in Section 5.3.2. In this case, we are minimizing the negative log-likelihood iteratively to get the parameter estimates. The initial estimates for the optimization were obtained as described in Section 5.4.2.1.

We computed the metrics mentioned above for 100 different samples and plotted the histograms of the metrics to see the range of their values across different samples. The following figures show the histograms of the various metrics for the three parameters, A, Q and S.

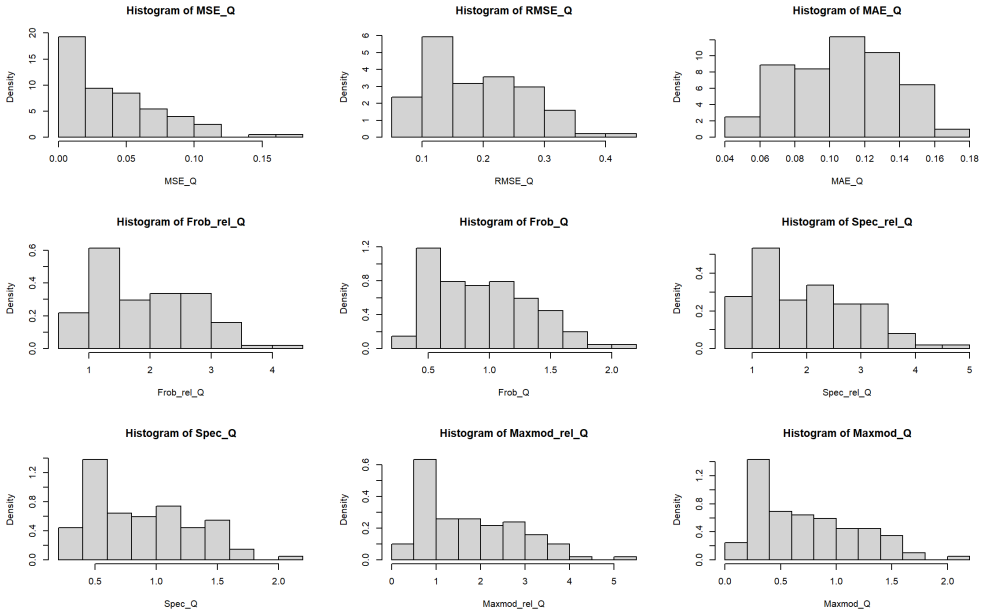

FIGURE 5.4. Histograms of all metrics for Q: low signal

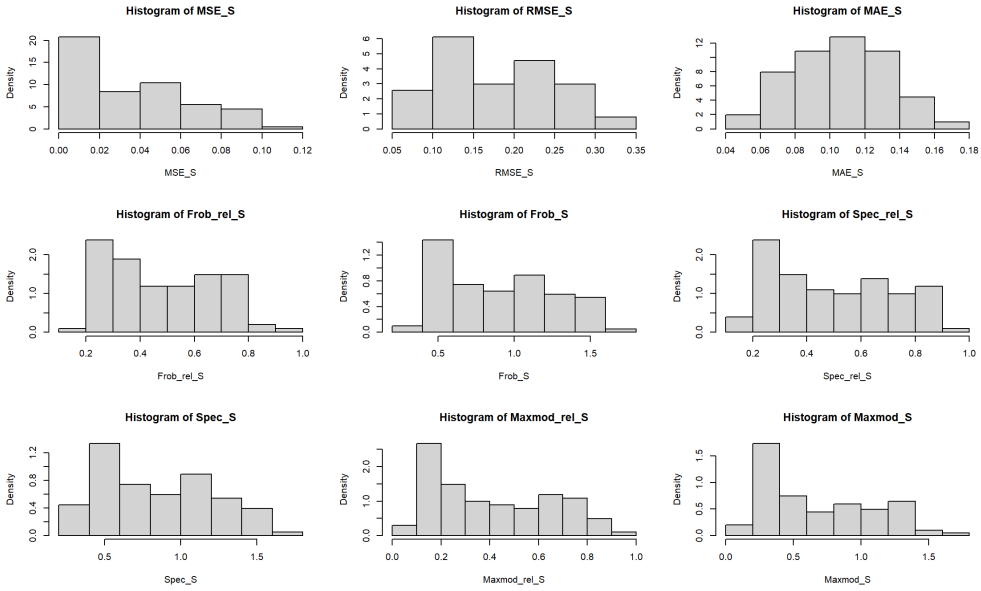

FIGURE 5.5. Histograms of all metrics for S: low signal

From Figures 5.3 to 5.5, it can be seen that the MSE, RMSE and MAE values are small for all three parameters. However, in terms of the relative measures, the parameter S has a smaller value compared to those of A and Q. That means, S is getting estimated better than the other two parameters. In order to consolidate the results across all samples, we decided to calculate the median for all metrics, because the median is not impacted by outliers. These values are given in Table 5.1.

Table 5.1 indicates that A and Q tend to have larger values of the metrics than S. Further, the relative measures of A and Q are larger than the regular measure. For example, relative Frobenius norm is larger than Frobenius norm, which indicates that the denominator of the relative measure is small. Since the denominator is the norm of the true parameter, it also indicates that the true parameter values are also small, providing a reason why estimation is difficult in this setting.

TABLE 5.1. Accuracy of the parameter estimates: low signal

Metric	A	Q	S
MSE	0.0665	0.0325	0.0292
RMSE	0.2579	0.1803	0.1707
MAE	0.1991	0.1072	0.1059
Relative Frobenius norm	1.4329	1.8153	0.4509
Frobenius norm	1.2897	0.9013	0.8537
Relative Spectral norm	1.3419	1.8171	0.4643
Spectral norm	1.0608	0.8144	0.7682
Relative Max norm	1.0226	1.5163	0.3296
Max norm	0.5802	0.6273	0.5567

From equation (5.64), it is evident that the total variation in the observations can be attributed to the variation coming from the underlying state equation and the variation from the observation noise. Upon computing the trace of the covariance matrix of the observations $\mathbf{Y}$, it came out to be around 4.2, whereas, the trace of the covariance matrix of the true state functions reduced to m dimensions was around 1.0, and the trace of the covariance matrix of the observation noise is around 3.2, which explains that the signal in this case is low and most of the variation in the observations are coming from the observation noise. As a

result, the parameters corresponding to the noise, S was getting estimated well, whereas the ones related to the state equation, A and Q were not getting estimated that well. Hence, we looked for ways to increase the signal to noise ratio, so that the majority of the variation in the observations can be explained by the state equation.

5.4.3.2. *Metrics for high signal.* A high signal was achieved by increasing the variability of the state functions. When FPCA was done on the newly simulated observations, it was found that the first three eigenvalues were sufficient in explaining around 90% variations in the data, which is intuitive because increasing the signal will increase the signal to noise ratio. This indicates that the the major source of variation in the observations is the underlying state equation and not the observation noise, so fewer eigenvalues are sufficient to explain the majority of variation in the data. So for this case, we decided to reduce the functions to $m = 3$ dimensional scores. Now each of the three parameters are 3×3, so we have a total of $p = 21$ parameters. As a result, we don't need a sample size as big as $n = 1300$. Keeping a similar n/p ratio which will keep the relative degrees of freedom in estimation at a similar level to that of the low signal case, we chose $n = 500$ for the estimation in this case when $m = 3$.

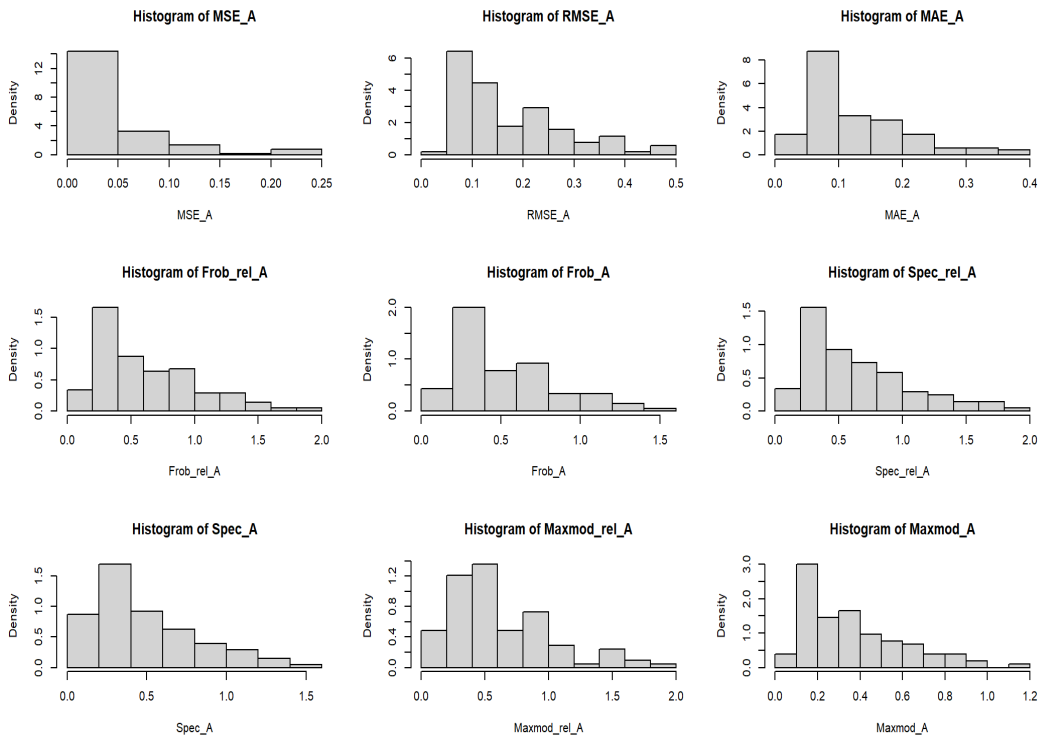

FIGURE 5.6. Histograms of all metrics for A: high signal

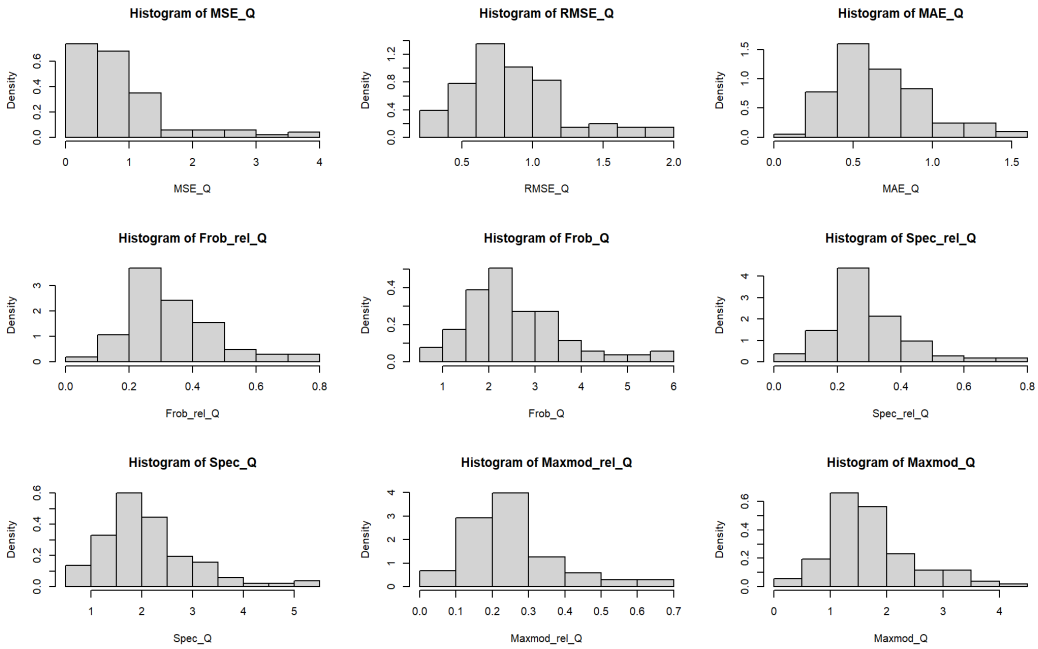

FIGURE 5.7. Histograms of all metrics for Q: high signal

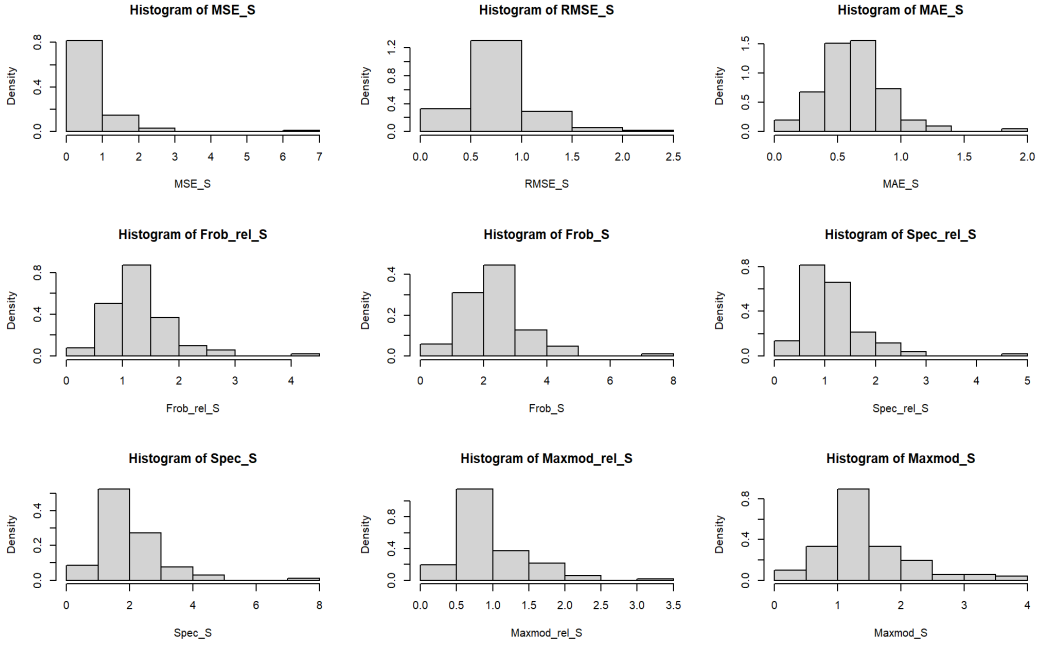

FIGURE 5.8. Histograms of all metrics for S: high signal

Similar to the low signal case, the metrics for the three parameters were obtained for about 100 different samples and the histograms were plotted. From Figures 5.6 – 5.8, we see that MSE, RMSE and MAE are small for all parameters. It is also seen that the parameters related to the signal, A and Q, are more accurately estimated than in the low signal case. In fact, most of the relative measures of A are around 0.5 and that of Q are between 0.2 and 0.3. The relative measures of S are somewhat larger, which is expected, because now the signal to noise ratio is higher.

Like before, the median metrics values are computed which are given in the Table 5.2. From the table, we see that the estimates are close to the true parameters for all three matrices. Specifically the MSE values are very small for all parameters indicating that the error in estimation of parameters in this case is lower.

TABLE 5.2. Accuracy of the parameter estimates: high signal

Metric	A	Q	S
MSE	0.0189	0.6383	0.5557
RMSE	0.1375	0.7989	0.7454
MAE	0.0963	0.6324	0.6121
Relative Frobenius norm	0.4740	0.3028	1.2347
Frobenius norm	0.4124	2.3968	2.2363
Relative Spectral norm	0.5025	0.2659	1.0231
Spectral norm	0.3830	1.9221	1.7541
Relative Max norm	0.4993	0.2381	0.9148
Max norm	0.3121	1.6145	1.3343

The median relative norm of A is around 0.5 and that of Q is around 0.3, which is intuitive, since these are the parameters associated with the underlying state equation which contributes to the majority of variation in the data. Even the relative norms of S are smaller than the absolute norms. However, the norms of S are higher now, because S is the covariance matrix of the observation noise. This analysis indicates the overall effectiveness of the process in estimating the parameters.

5.4.3.3. *Variation with sample size.* We have proved consistency of our parameter estimates in Section 5.3. That means, as the sample size n increases, the estimates will be closer to the true parameter values.

Since the metrics described in Section 5.4.3 are based on the error in estimation, it implies that the metrics will be decreasing as n increases.

Here, we focus on the high signal case. The following table shows the values of the above metrics after running the estimation procedure using the initial estimates defined above for different sample sizes when $m = 3$.

TABLE 5.3. Metrics for Different Parameters and Sample Sizes

	A			Q			S		
	$n = 300$	$n = 750$	$n = 1500$	$n = 300$	$n = 750$	$n = 1500$	$n = 300$	$n = 750$	$n = 1500$
MSE	0.1889	0.0112	0.0083	0.8675	0.5802	0.5596	0.6806	0.6059	0.4998
RMSE	0.4347	0.1059	0.0909	0.9314	0.7617	0.7481	0.8250	0.7784	0.7070
MAE	0.3081	0.0746	0.0678	0.8198	0.6078	0.6346	0.7192	0.6764	0.5820
Relative Frobenius	1.5190	0.3659	0.3134	0.3755	0.2952	0.2896	1.3640	1.3163	1.2000
Frobenius	1.3040	0.3178	0.2727	2.7942	2.2851	2.2443	2.4750	2.3351	2.1209
Relative Spectral	1.9199	0.4043	0.3885	0.3244	0.2652	0.2447	1.2026	1.0527	1.0655
Spectral	1.2926	0.2730	0.2631	2.1424	1.8809	1.7268	1.9766	1.6991	1.7147
Relative Maxnorm	1.9548	0.4713	0.4826	0.2166	0.1628	0.1626	1.0532	0.7488	0.7453
Maxnorm	0.9387	0.2162	0.2214	1.4213	1.1452	1.1427	1.6432	1.1638	1.1318

From the table, we see that indeed the values of the metrics go down as n increases, however, they are not going down too fast. This indicates empirically that the estimates are consistent.

5.5. Conclusion

This chapter provided a structure of a functional stochastic volatility model, where the intra-day volatility could be considered as a function. It proved that there exists a strictly stationary and causal solution to the volatility process. Stochastic volatility models form an exogenous way of modeling volatility in finance, where the process does not depend on the past observation but an underlying stochastic process.

The methods of estimation of model parameters were introduced leveraging the state-space framework and Kalman filtering methods, where the returns were observed, whereas the underlying volatility was unobserved. This method is an easier alternative to the Bayesian Hierarchical time series methods of estimation available in the literature. The main principle is to reduce the infinite-dimensional functions

to finite-dimensional objects using FPCA and implementing the multivariate state-space framework using quasi-likelihood and Kalman filtering techniques. Detailed outline were provided on how to choose the initial estimates required for iteratively maximizing the likelihood numerically.

The effectiveness of the estimation methods were supported through empirical studies. The estimation method was proved to be effective in terms of smaller error when the amount of variation explained by the state equation was high relative to the variation in the noise of the observation equation yielding a high signal to noise ratio. However, one should keep in mind that the methods might be prone to a relatively larger error when the signal to noise ratio is low. Consistency of the parameter estimates were proved theoretically following the methods proposed by Whittle and were also proved empirically.

Thus, in conclusion, as long as the signal to noise ratio is high in the data, this chapter provides an effective method of estimating the parameters of a functional stochastic volatility model. The model parameters pertaining to the observation equation are estimated better than those in the state equation, if the signal to noise ratio is low.

Bibliography

Milton Abramowitz and Irene A Stegun. *Handbook of mathematical functions with formulas, graphs, and mathematical tables*, volume 55. US Government printing office, 1968.

Torben G Andersen, Richard A Davis, Jens-Peter Kreiß, and Thomas V Mikosch. *Handbook of financial time series*. Springer Science & Business Media, 2009.

Anestis Antoniadis, Efstathios Paparoditis, and Theofanis Sapatinas. A functional wavelet–kernel approach for time series prediction. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 68 (5):837–857, 2006.

Alexander Aue and Prabir Burman. Estimation of prediction error in time series. *Biometrika*, 111:643–660, 2024.

Alexander Aue and Anne van Delft. Testing for stationarity of functional time series in the frequency domain. *The Annals of Statistics*, 48:2505–2547, 2020.

Alexander Aue, István Berkes, and Lajos Horváth. Strong approximation for the sums of squares of augmented GARCH sequences. *Bernoulli*, 12(4):583–608, 2006.

Alexander Aue, Diogo Dubart Norinho, and Siegfried Hörmann. On the prediction of stationary functional time series. *Journal of the American Statistical Association*, 110(509):378–392, 2015.

Alexander Aue, Lajos Horváth, and Daniel F. Pellatt. Functional generalized autoregressive conditional heteroskedasticity. *Journal of Time Series Analysis*, 38(1):3–21, 2017.

Alexander Aue, Gregory Rice, and Ozan Sönmez. Detecting and dating structural breaks in functional data without dimension reduction. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 80(3):509–529, 2018.

Dennis Bams, Gildas Blanchard, and Thorsten Lehnert. Volatility measures and value-at-risk. *International Journal of Forecasting*, 33(4):848–863, 2017.

Philippe C Besse, Hervé Cardot, and David B Stephenson. Autoregressive forecasting of some functional climatic variations. *Scandinavian Journal of Statistics*, 27(4):673–687, 2000.

Tim Bollerslev. Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, 31(3): 307–327, 1986.

Denis Bosq. *Linear processes in function spaces: theory and applications*, volume 149. Springer Science & Business Media, 2000.

Paul Bouchey, Vassilii Nemtchinov, Alex Paulsen, and David M Stein. Volatility harvesting: Why does diversifying and rebalancing create portfolio growth? *The Journal of Wealth Management*, 15(2):26–35, 2012.

Clément Cerovecki, Christian Francq, Siegfried Hörmann, and Jean-Michel Zakoïan. Functional GARCH models: The quasi-likelihood approach and its applications. *Journal of Econometrics*, 209(2):353–375, 2019.

Holger Dette, Kevin Kokot, and Alexander Aue. Functional data analysis in the Banach space of continuous functions. *The Annals of Statistics*, 48(2):1168–1192, 2020.

Christopher Dienes and Alexander Aue. On-line monitoring of pollution concentrations with autoregressive moving average time series. *Journal of Time Series Analysis*, 35(3):239–261, 2014.

Jin-Chuan Duan. Augmented GARCH (p, q) process and its diffusion limit. *Journal of Econometrics*, 79 (1):97–127, 1997.

Bradley Efron. The estimation of prediction error: covariance penalties and cross-validation. *Journal of the American Statistical Association*, 99(467):619–632, 2004.

Robert F Engle. Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. *Econometrica*, pages 987–1007, 1982.

Robert F Engle and Andrew J Patton. Error-correction model of quote prices. 2000.

Israel Gohberg, Seymour Goldberg, and Marius A Kaashoek. *Classes of linear operators*, volume 63. Birkhäuser, 2013.

Hanlu Gong, Aera Thavaneswaran, and J Singh. A Black–Scholes model with GARCH volatility. *Mathematical Scientist*, 35(1), 2010.

Siegfried Hörmann and Piotr Kokoszka. Functional time series. In *Handbook of Statistics*, volume 30, pages 157–186. Elsevier, 2012.

Siegfried Hörmann, Lajos Horváth, and Ron Reeder. A functional version of the ARCH model. *Econometric Theory*, 29(2):267–288, 2013.

Rob J Hyndman and Han Lin Shang. Forecasting functional time series. *Journal of the Korean Statistical Society*, 38(3):199–211, 2009.

Phillip A Jang, Michael Jauch, and David S Matteson. Functional stochastic volatility. 2021.

Shuhao Jiao, Alexander Aue, and Hernando Ombao. Functional time series prediction under partial observation of the future curve. *Journal of the American Statistical Association*, 118(541):315–326, 2023.

Rudolph E Kalman. A new approach to linear filtering and prediction problems. 1960.

Rudolph E Kalman and Richard S Bucy. New results in linear filtering and prediction theory. 1961.

Vladislav Kargin and Alexei Onatski. Curve forecasting by functional autoregression. *Journal of Multivariate Analysis*, 99(10):2508–2526, 2008.

Michael Lawrence and Spyros Makridakis. Factors affecting judgmental forecasts and confidence intervals. *Organizational Behavior and Human Decision Processes*, 43(2):172–187, 1989.

Yehua Li and Tailen Hsing. Uniform convergence rates for nonparametric regression and principal component analysis in functional/longitudinal data. 2010.

Helmut Lütkepohl. *New Introduction to Multiple Time Series Analysis*. Springer Science & Business Media, 2005.

Jim O Ramsay and Bernard W Silverman. *Functional Data Analysis*. Springer, 2005.

Jorma Rissanen. Order estimation by accumulated prediction errors. *Journal of Applied Probability*, 23(A): 55–61, 1986.

Esther Ruiz. Quasi-maximum likelihood estimation of stochastic volatility models. *Journal of econometrics*, 63(1):289–306, 1994.

Henry Scheffe. A method for judging all contrasts in the analysis of variance. *Biometrika*, 56(1):229–229, 1969.

Robert H Shumway and David S Stoffer. *Time Series Analysis and its Applications*, volume 3. Springer, 2000.

Ernst Stadlober, Siegfried Hörmann, and Brigitte Pfeiler. Quality and performance of a PM10 daily forecasting model. *Atmospheric Environment*, 42(6):1098–1109, 2008.

Stephen J Taylor. Financial returns modelled by the product of two stochastic processes-a study of the daily sugar prices 1961-75. *Time series analysis: theory and practice*, 1:203–226, 1982.

Stephen J Taylor. *Modelling financial time series*. World Scientific, 2008.

Ruey S Tsay. *Analysis of financial time series*. John Wiley & Sons, 2005.

Peter Whittle. The analysis of multiple stationary time series. *Journal of the Royal Statistical Society: Series B (Methodological)*, 15(1):125–139, 1953.

Samuel S Wilks. Certain generalizations in the analysis of variance. *Biometrika*, pages 471–494, 1932.