Technische Universität München

Zentrum Mathematik

Time Series in Functional Data Analysis

Master's Thesis

by

Taoran Wei

Themenstellerin:	Prof. Dr. Claudia Klüppelberg
Betreuer:	M.Sc. Johannes Klepsch
Abgabetermin:	31.December 2015

Abstract

A common way to obtain the *functional time series* is to break the records made over a continuous time interval into several natural consecutive time intervals, e.g. days. Providing reliable predictions for functional time series is an important goal. Since we still lack advanced functional time series methodology, we often assume the functional time series to follow the first-order functional autoregressive model (FAR(1)).

In recent years, *functional data analysis (FDA)* has been widely applied in functional time series analysis. Aue et al. [2015] proposed a prediction algorithm which combines the idea of FDA and functional time series analysis. And the prediction algorithm is not restricted to FAR structure. They reduced the dimension of the data and tranformed the issue of predicting functional time series to predicting multivariate time series.

In this thesis, we will first give an overview of the basic theory of functional data analysis, functional time series analysis and the prediction algorithm by Aue et al. [2015]. Then we will focus on studying the functional ARMA$(p, 1)$ process and its corresponding truncated vector process. We will figure out the structure of the truncated vector process and try to find out a condition under which both the functional and vector processes are stationary. Then based on the truncated vector process, we will compute the one-step (functional) predictor for functional ARMA$(1, 1)$ process. Furthermore, we will compare (functional) predictor with the *functional best linear predictor* proposed by Bosq [2000]. To verify the results of our study, we conduct a simulation study in Chapter 5.

We finish this thesis by applying the theory of FDA in analysing the highway traffic data provided by Autobahndirektion Südbayern. Our goal is to model and predict the traffic data.

Hiermit erkläre ich, dass ich die Diplomarbeit selbstständig angefertigt und nur die angegebenen Quellen verwendet habe.

München, den 22. December 2015

Acknowledgments

First of all, I would like to thank Prof. Dr. Claudia Klüppelberg for offering me the precious opportunity to pursue this quite interesting topic, and for her guidance since last year. I shall extend my thanks to Prof. Dr. Peter J. Brockwell, who during his stay at TU Munich gave me a lot of helpful advices on my work of data analysis.

I am particularly grateful to my supervisor Johannes Klepsch. He gave me valuable guidance and advices in every stage of my writing. Without his kindness and patient instruction, I would have never completed my thesis. I enjoy and cherish the memory of working with him.

Furthermore, I want to express my deep gratitude for my parents, not only during the writing of my thesis, but also about my complete studies, nearly 20 years.

Finally, but most importantly, I would like to thank my wife for her constant love and moral support. I sincerely appreciate her companion, encouragement and the wonderful time we had together in the past ten years.

Contents

Chapter 1

Introduction

Suppose we have observations $X_1, \ldots, X_N$, where $X_1, \ldots, X_N$ can be scalars, vectors or other objects. *Functional data analysis (FDA)* is concerned with observations which are functions. We assume these functional observations $X_1, \ldots, X_N$ are in some Hilbert function space H, e.g. the space of square-integrable functions $L^2([0,1])$, then we call $X_1, \ldots, X_N$ *H-valued data* or H-valued *functional observations*. $X_1, \ldots, X_N$ are realizations of H-valued random functions defined on some common probability space $(\Omega, \mathcal{A}, P)$. Functional data analysis (FDA) focuses on analysing such H-valued random functions and observations.

Functional principal component analysis (FPCA) plays a central role in the FDA. The basic idea of FPCA is to represent the H-valued random function X with the eigenfunctions of the *covariace operator* of X, which is known as *Karhunen-Loéve representation*. Then X is truncated just with a fixed number d of eigenfunctions, where these d eigenfunctions can explain most of the variability of X.

At an early stage of development, FDA focused mainly on i.i.d functional data. In recent years, FDA has also been widely applied in functional time series analysis (e.g. see Hörmann and Kokoszka [2010] and Hörmann and Kokoszka [2012]).

In Hörmann and Kokoszka [2010], they proposed the notion of *L^p-m-approximability*, which can be used to quantify the *temporal dependence* of functional time series. It is an extension of *m-dependence* in scalar and multivariate time series analysis. Based on the work of Hörmann and Kokoszka [2010], the ideas of time series analysis and FDA have been merged and many results in FDA under the assumption of i.i.d have been extented to *L^p-m-approximable functional time series* (e.g. see Hörmann and Kokoszka [2010], Hörmann and Kokoszka [2012]) and Horváth et al. [2013a]).

Providing reliable predictions is one of the most important goals of functional time series analysis. Bosq [2000] has studied the *functional best linear predictor* for (stationary) functional linear process. But the problem is, we do not know the exact math formula of the functional best linear predictor, so in fact it is difficult to implement. Since we still lack advanced functional time series methodology, we often assume the functional time series to follow the first-order functional autoregressive model (FAR(1)). And the prediction is also based on the assumption of FAR(1) structure (e.g. see Chapter 3 of Bosq [2000]). Aue et al. [2015] proposed a prediction algorithm which combines the idea of FDA and functional time series analysis. And the prediction algorithm is not restricted

to FAR structure. The basic idea is to use FPCA to reduce the infinite-dimensional data to finite-dimensional data. Then the issue of predicting functional time series is transformed to the prediction of multivariate time series.

In Chapter 2 we will give an overview of the existing results of the studies on FDA and functional time series analysis. In Chapter 3, we will introduce two hypothesis tests for functional data. One is the Portmanteau test of independence proposed by Gabrys and Kokoszka [2007]. The other one is the test of stationarity of functional time series, which was proposed by Horváth et al. [2013b]. We will apply these two tests in the real data analysis in Chapter 6, to check whether our dataset is properly transformed.

In Chapter 4, after the brief overview of the work by Bosq [2000] (functional best linear predictor) and Aue et al. [2015] (the prediction algorithm for functional time series), we will focus on the prediction of functional ARMA$(p, 1)$ process. In Section 4.3, we will first propose a sufficient condition for stationarity of functional ARMA$(p, 1)$ process. Then we will have a closer look on the vector process truncated from the functional process (by FPCA). We will show that, the vector process "approximately" follows the vector ARMA$(p, 1)$ structure. And under some further constraints, the vector process is rigorously a stationary vector ARMA$(p, 1)$ process. In Section 4.4, based on the truncated vector observations, we will compute the one-step predictor for the functional ARMA$(p, 1)$ process and compare it with the functional best linear predictor (by Bosq [2000]). In Section 4.5, we will bound the prediction error when applying the prediction algorithm by Aue et al. [2015] to functional ARMA$(p, 1)$ process.

To verify the results of our study in Chapter 4, we conduct simulation studies in Chapter 5. Firstly we will simulate FARMA$(p, 1)$ processes. Then we fit different vector ARMA models (with different orders) to the truncatd vector process and compare the goodness of fit. Finally we compute the functional predictors based on these vector ARMA models and compare the prediction errors.

In Chapter 6, we apply the theory of FDA in the real data analysis. We obtained the dataset from Autobahndirektion Südbayern which describes the traffic conditions from 01/01/2014 to 30/06/2014 on a highway in Southern Bavaria, Germany. We will first try to describe and transform the data properly, then we will try to model and predict the traffic data.

1.1 Examples of functional time series

A functional time series is a sequence of (random) functions. These functions can arise from successive measurements made over a time interval, which is divided into several consecutive time intervals with equal length, e.g. days. In our real data analysis (details see Chapter 6), our observations are highway traffic records provided by Autobahndirektion Südbayern. The car velocity and the traffic volume (or flow) are recorded every one minute (1440 minutes in one day) during the period 1/1/2014 0:00 to 30/6/2014 23:59. We use

$$S_n = (S_n(t_1), \ldots, S_n(t_{1440}))^T \tag{1.1}$$

and

$$C_n = (C_n(t_1), \ldots, C_n(t_{1440}))^T \tag{1.2}$$

to represent the velocity and the traffic volume recorded on day n. S_n and C_n belong to the class of *high-dimensional data*.

A macroscopic traffic model involves *velocity*, *flow* and *density*. *Density* is defined as

$$\text{Density} := \frac{\text{Flow}}{\text{Velocity}}, \tag{1.3}$$

and it can reveal the number of vehicles in a unit of length.

The relation among these three variables can be shown with the diagrams of "Velocity-Flow relation" and "Flow-Density relation". The diagram of "Flow-Density relation" is also called *fundamental diagram of traffic flow*. It can be used to predict the capacity of a road system and give guidance for inflow regulations or speed limits.

Figure 1.1 and 1.2 depict the "velocity-flow relation" and "flow-density relation". At a *critical traffic density*, the state of flow will change from stable to unstable. In Figure 1.2, the *critical density* is about 0.45. Combining Figure 1.1 and 1.2, the corresponding flow and velocity of the critical density is around 285 $veh/3$ $minutes$ and 80 km/h.

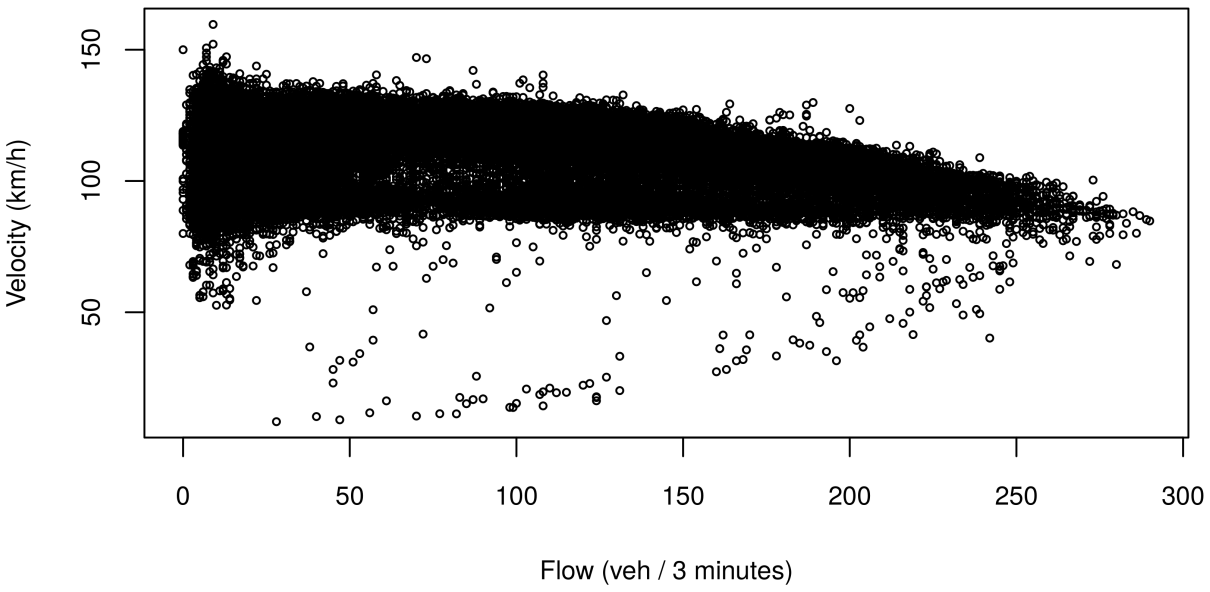

Figure 1.1: Velocity-flow relation on a highway in Southern Bavaria. Depicted are average velocities per 3 minutes versus the number of vehicles (flow) within these 3 minutes during the period 01/01/2014 0:00 to 30/06/2014 23:59. **Source:** Autobahndirektion Südbayern

Now let us view the velocity and the volume records directly. Figure 1.3 and Figure 1.4 depict the velocity and the volume (1440-dimensional) record on several selected weeks. In these two figures, the initial day in each row (week) is a Sunday.

As can be seen from Figure 1.3 and Figure 1.4, both the velocity and the traffic volume records show weekly periodicity. And the periodicity of the traffic volume data is even

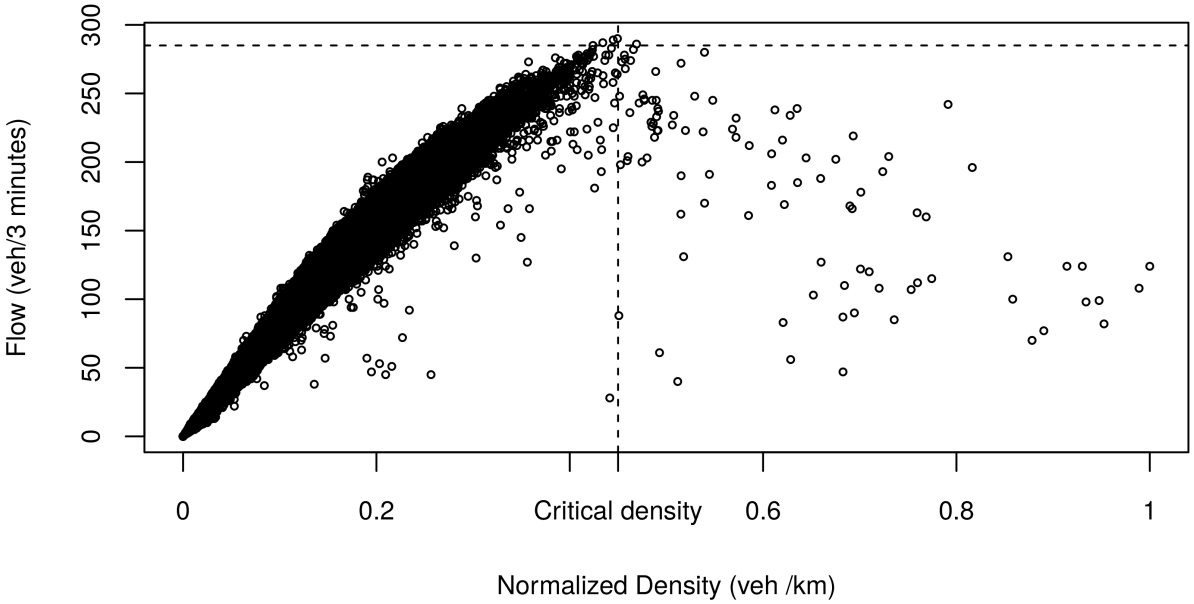

Figure 1.2: Flow-density for the data from Figure 1.1. **Source:** Autobahndirektion Südbayern

more obvious. But note that, e.g see the curves on 21/4 and 9/6 in Figure 1.3 and 1.4, the shape of the curves are apparently different from the curves on the other "normal" Mondays. We checked the calander and found that these two days are public holidays in Germany in 2014. In fact, during 1/1/2014 to 30/6/2014, there are 10 public holidays in Germany: 1/1(Wednesday), 6/1(Monday), 18/4(Friday), 20/4(Sunday), 21/4(Monday), 1/5(Thursday), 29/5(Thursday), 8/6(Sunday), 9/6(Monday) and 19/6(Thursday). Due to the limit of the size of the page, Figure 1.3 and Figure 1.4 just contain the following 7 holidays: 6/1(Monday), 18/4(Friday), 20/4(Sunday), 21/4(Monday), 8/6(Sunday), 9/6(Monday), 19/6(Thursday).

As can be seen from Figure 1.3 and 1.4, besides the curves on 21/4(Monday) and 9/6(Monday) mentioned above, the curves on 6/1(Monday), 18/4(Friday), 1/5(Thursday) and 19/6(Thursday), i.e the holidays falling on a weekday, differ apparently from the other corresponding "normal" weekday curves. Furthermore, the volume curves on these days seem to be lower than those on the other normal weekdays. In Chapter 6, we will exclude these days out of the group **Workingdays** and categorize them into the group **Holidays**. In other words, when we model the traffic data on workingdays, the data on these days will not be considered.

In contrast, on 20/4(Sunday) and 8/6(Sunday), both the velocity and the volume curves do not differ obviously from those on the other normal Sundays. But on 8/6(Sunday), the velocity curve seems to be lower than those on the other normal Sundays. In Chapter 6, we will treat these two days as normal Sundays instead of categorizing them into the group **Holidays**.

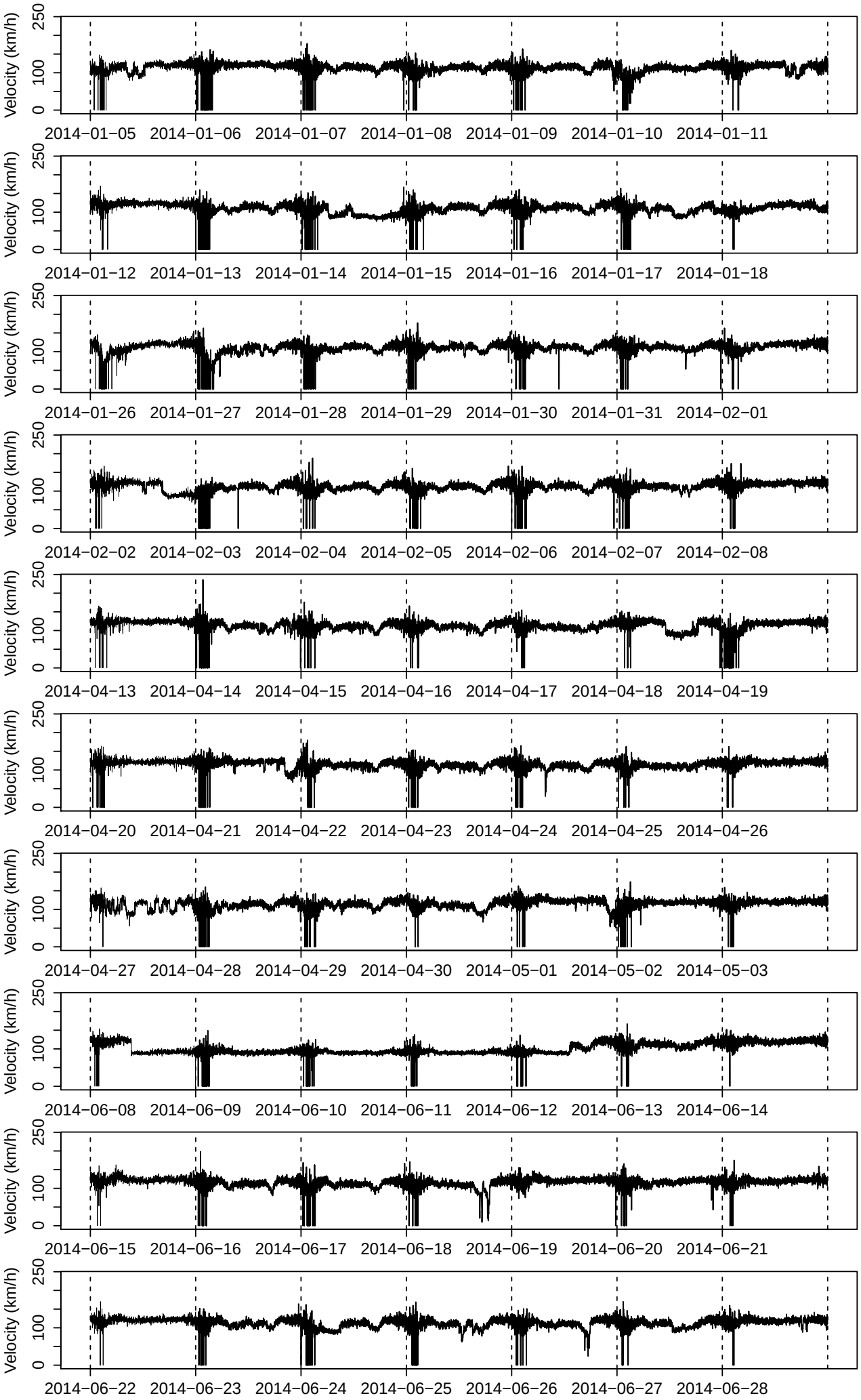

Figure 1.3: Highway velocity records on several weeks. In this figure, 01-06, 04-18, 04-20, 04-21, 06-08, 06-09, 06-19 are public holidays. **Source:** Autobahndirektion Südbayern

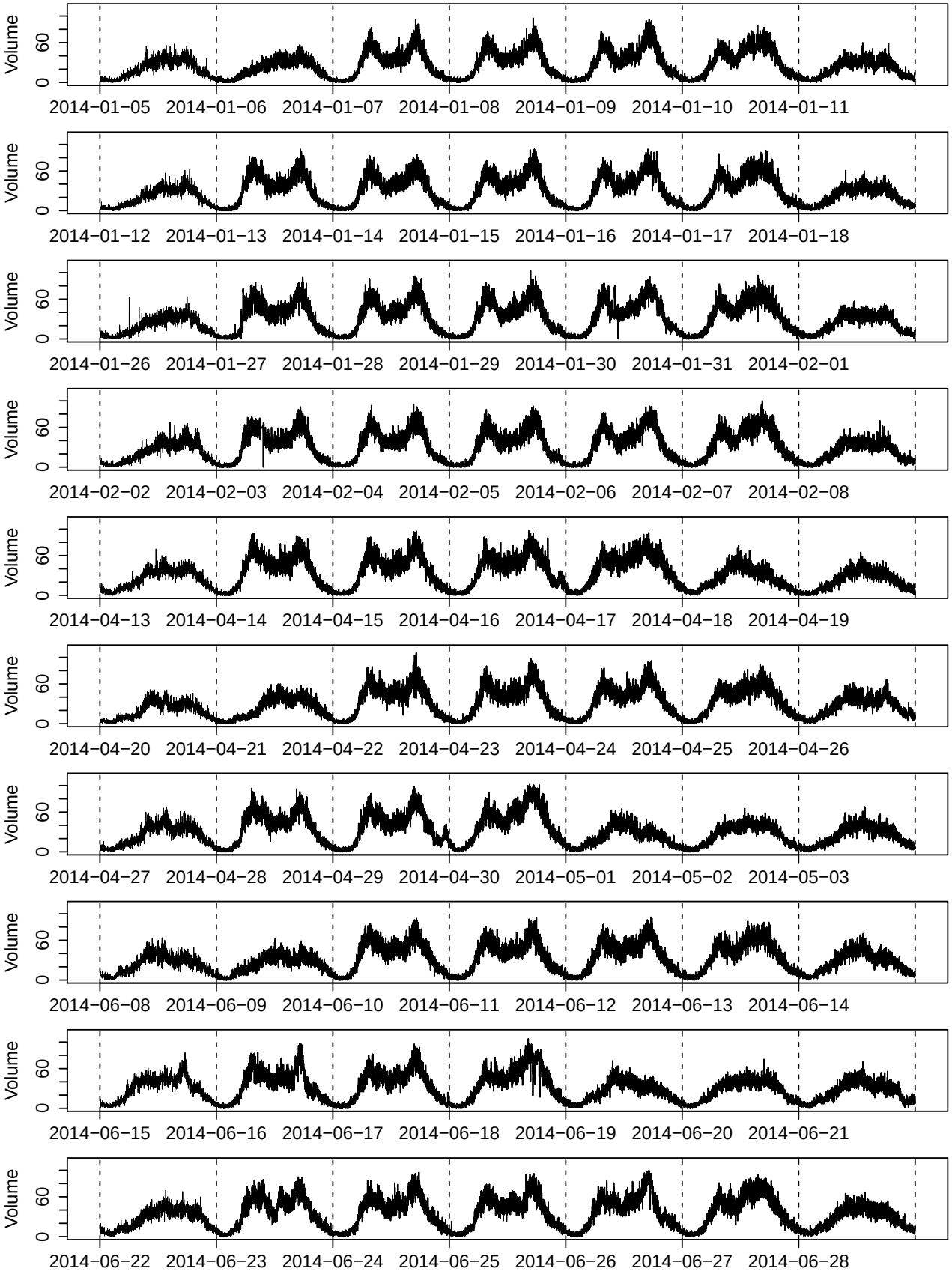

Figure 1.4: Highway traffic volume records on several weeks. In this figure, 01-06, 04-18, 04-20, 04-21, 06-08, 06-09, 06-19 are public holidays. **Source:** Autobahndirektion Südbayern

In functional data analysis, our observations are functions. Especially, we prefer smooth functions. Note that $S_1, \ldots, S_N$ in (1.1) are high-diemsional data. When we have such kind of data, we need to *smooth* the data, i.e. to transform the vector data to smooth functions.

The first step of smoothing is to specify a series of basis functions, e.g. Fourier basis functions and B-spline basis functions. Usually Fourier basis functions are utilized when the data are periodic or without strong local features, while B-splines basis functions are fit for the data with strong local features (more details see Ramsay and Silverman [2002] and Ramsay and Silverman [2005]).

Once we determine the basis functions(we denote them by $F_1, \ldots, F_M$), the vector data can be approximated by

$$S_n(t_j) \approx \sum_{m=1}^{M} c_{nm} F_m(t_j), \quad j = 1, \ldots, 1440, \quad n = 1, \ldots, N.$$

The coefficients c_{nm} are determined by the least square criterion (details see Chapter 5).

After the coefficients are determined, we can smooth the vector data by

$$S_n(t) := \sum_{m=1}^{M} c_{nm} F_m(t), \quad n = 1 \ldots, N.$$

In Chapter 6, we choose $M = 29$ Fourier basis functions to smooth the 178 vector velocity observations (we will show why we choose Fourier basis functions instead of Bspline basis, and why to choose $M = 29$ basis functions in details in Section 6.1.3). Figure 1.5 depicts 21 vector observations (in 3 weeks) and their corresponding functional data. As can be seen from Figure 1.5, the functional velocity curves do not lose the main "shape" of the vector velocity curves.

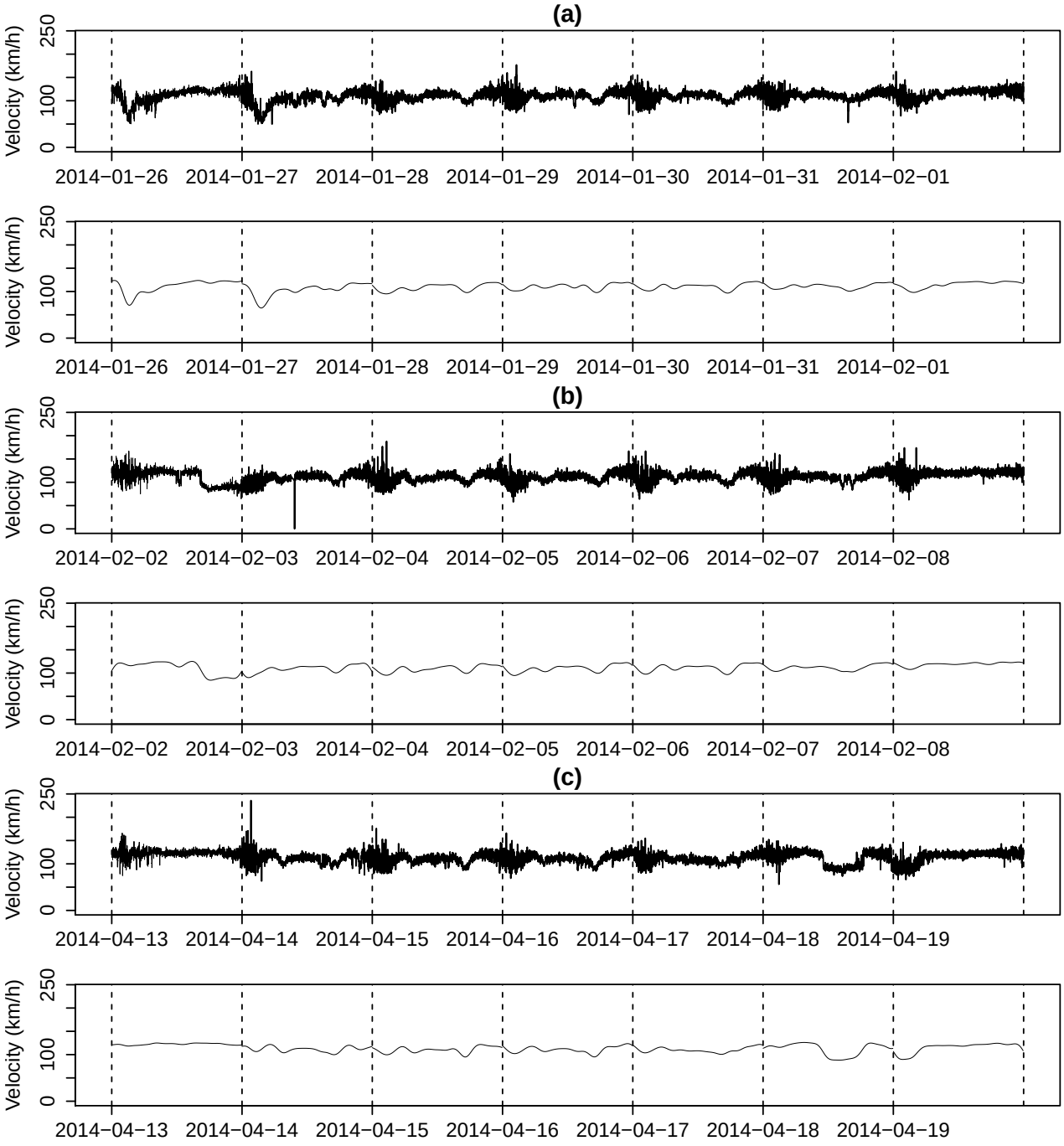

Figure 1.5: $N = 178$ vector observations are smoothed by $M = 29$ Fourier basis functions. Depicted are the vector velocity data versus the corresponding functional data on three weeks. **Source:** Autobahndirektion Südbayern

Chapter 2

Some basics

In this chapter we give an overview of the existing results of the studies on FDA and functional time series analysis. To make this part more understandable, we will start with the mathematical framework required in FDA.

This chapter is mainly based on Chapter 2 and 3 of Horváth and Kokoszka [2012], Hörmann and Kokoszka [2010] and Horváth et al. [2013a].

2.1 Mathematical framework required in FDA

2.1.1 Operators in Hilbert space

We consider a separable Hilbert space H (with a countable orthonormal basis), e.g. $L^2([0,1])$. H is equipped with the inner product $\langle \cdot, \cdot \rangle$ (and the corresponding norm $\| \cdot \|$). We denote by $\mathcal{L}$ the space of bounded operators acting on H, then the norm of a bounded opearator $\Psi \in \mathcal{L}$ is defined by

$$\|\Psi\|_{\mathcal{L}} := \sup\{\|\Psi(x)\| : \|x\| \leq 1, x \in H\}.$$

A bounded operator Ψ is said to be *compact* if there exist two orthonormal bases $\{\nu_j\}$ and $\{f_j\}$ of H, and a real sequence $\{\lambda_j\}$ converging to zero, such that

$$\Psi(x) = \sum_{j=1}^{\infty} \lambda_j \langle x, \nu_j \rangle f_j, \quad x \in H. \tag{2.1}$$

We always assume λ_j are positive, because f_j can be replaced by $-f_j$ if needed.

A compact operator Ψ is said to be a *Hilbert-Schmidt* operator, if

$$\sum_{i=1}^{\infty} \|\Psi(e_i)\|^2 < \infty,$$

where $\{e_i\}$ is an arbitrary orthonormal basis of H. We denote by $\mathcal{S}$ the space of Hilbert-Schmidt operators acting on H. $\mathcal{S}$ is a separable Hilbert space equipped with the following

inner product and the corresponding *Hilbert-Schmidt norm*,

$$\langle \Psi_1, \Psi_2 \rangle_{\mathcal{S}} := \sum_{i=1}^{\infty} \langle \Psi_1(e_i), \Psi_2(e_i) \rangle ,$$

$$\|\Psi\|_{\mathcal{S}} := \sqrt{\langle \Psi, \Psi \rangle_{\mathcal{S}}} = \sqrt{\sum_{i=1}^{\infty} \|\Psi(e_i)\|^2} < \infty. \tag{2.2}$$

If Ψ is a Hilbert-Schmidt operator, then $\|\Psi\|_{\mathcal{L}} \leq \|\Psi\|_{\mathcal{S}}$. Furthermore, (2.2) implies $\sum_{i=1}^{\infty} \lambda_i^2 < \infty$ in (2.1). This holds since

$$\begin{aligned} \|\Psi\|_{\mathcal{S}}^2 &= \sum_{i=1}^{\infty} \langle \Psi(e_i), \Psi(e_i) \rangle \\ &\overset{\text{by(2.1)}}{=} \sum_{i=1}^{\infty} \left\langle \sum_{j=1}^{\infty} \lambda_j \langle e_i, \nu_j \rangle f_j, \sum_{k=1}^{\infty} \lambda_k \langle e_i, \nu_k \rangle f_k \right\rangle \\ &= \sum_{j=1}^{\infty} \sum_{i=1}^{\infty} \lambda_j^2 \langle e_i, \nu_j \rangle^2 = \sum_{j=1}^{\infty} \lambda_j^2 < \infty. \end{aligned} \tag{2.3}$$

We call an operator $\Psi \in \mathcal{L}$ *symmetric* if

$$\langle \Psi(x), y \rangle = \langle x, \Psi(y) \rangle , \quad \forall x, y \in H, \tag{2.4}$$

and *non-negative definite* if

$$\langle \Psi(x), x \rangle \geq 0, \quad \forall x \in H. \tag{2.5}$$

A symmetric non-negative definite Hilbert Schmidt operator Ψ admits the decomposition

$$\Psi(x) = \sum_{j=1}^{\infty} \lambda_j \langle x, \nu_j \rangle \nu_j, \quad \forall x \in H, \tag{2.6}$$

where ν_j are the orthonormal eigenfunctions of Ψ, i.e $\Psi(\nu_j) = \lambda_j \nu_j, \|\nu_j\| = 1, \forall j \in \mathbb{Z}$. From the theory of linear algebra, $(\nu_j, j \in \mathbb{Z})$ in (2.6) forms a basis of H.

2.1.2 The space of square integrable functions

A real-valued measurable function f defined on $[0, 1]$ is said to be *square-integrable*, if

$$\int_0^1 f^2(t) \mathrm{d}t < \infty.$$

We denote by $L^2([0, 1])$ the space of square integrable functions on $[0, 1]$. $L^2([0, 1])$ is a separable Hilbert space with the inner product

$$\langle x, y \rangle := \int_0^1 x(t) y(t) dt, \quad x, y \in L^2([0, 1]) . \tag{2.7}$$

For $x, y \in L^2([0, 1])$, $x = y$ means $\int_0^1 [x(t) - y(t)]^2 = 0$.

2.1.3 Hilbert-Schmidt integral operators

Now we introduce an important class of Hilbert Schmidt operators: *Hilbert-Schmidt integral operators*, which we will utilize in our simulation studies in Chapter 5.

Φ is said to be an *integral operator* acting on $L^2([0,1])$, if Φ has the form of

$$\Phi(x)(t) = \int_0^1 \phi(t,s)x(s)ds, \quad t \in [0,1], \quad x \in L^2([0,1]). \tag{2.8}$$

The $\phi(\cdot,\cdot)$ in (2.8) is called the *integral kernel* of Φ.

An integral operator Φ in (2.8) is Hilbert-Schmidt if and only if

$$\iint \phi^2(t,s)dtds < \infty. \tag{2.9}$$

In this case,

$$\|\Phi\|_{\mathcal{S}}^2 = \iint \phi^2(t,s)dtds < \infty. \tag{2.10}$$

We call the integral operators which satisfy (2.10) *Hilbert-Schmidt integral operators.*

Remark 1. *If $\phi(\cdot,\cdot)$ in (2.8) is a symmetric kernel, i.e $\phi(t,s) = \phi(s,t)$ for $\forall t,s \in [0,1]$, then the corresponding integral operator Φ is symmetric (see (2.4)).*

Proof.

$$\begin{aligned}
\langle \Phi(x), y\rangle &= \iint [\phi(t,s)x(s)\mathrm{d}s]\, y(t)\mathrm{d}t \\
&= \iint [\phi(s,t)y(t)\mathrm{d}t]\, x(s)\mathrm{d}s \\
&= \langle\, x, \Phi(y)\rangle\,, \quad \forall x, y \in L^2([0,1]).
\end{aligned}$$

□

The following theorem, known as Mercer's theorem, list some properties of the integral kernel $\phi(t,s)$.

Theorem 2.1 (Mercer's theorem)**.** *Suppose $\phi(t,s)$ is a continuous, symmetric and non-negative definite integral kernel, and Φ is its corresponding integral operator defined in (2.8). Then there is an orthonormal basis $(\nu_i, i \in \mathbb{Z})$ of $L^2([0,1])$ consisting of eigenfunctions of Φ such that the corresponding eigenvalues $(\lambda_i, i \in \mathbb{Z})$ are non-negative. $(\nu_i, i \in \mathbb{Z})$ and the corresponding $(\lambda_i, i \in \mathbb{Z})$ are defined by*

$$\lambda_i\nu_i(t) = \Phi(\nu_i)(t) = \int_0^1 \phi(t,s)\nu_i(s)\mathrm{d}s, \ \forall t \in [0,1], \ \forall i \in \mathbb{Z}.$$

Furthermore, $(\nu_i, i \in \mathbb{Z})$ are continuous on $[0,1]$ and $\phi(t,s)$ has the representation

$$\phi(t,s) = \sum_{i=1}^{\infty} \lambda_i\nu_i(t)\nu_j(s),$$

where the convergence is absolute and uniform.

In the following we introduce a special integral kernel, *Gausian kernel*. In Chapter 5 we will use Gaussian kernel to simulate functional ARMA$(p, 1)$ process.

Example 1. *An integral kernel $\phi(t, s)$ is said to be a Gaussian kernel, if it has the form of*

$$\phi(t, s) = C \exp\left(\frac{t^2 + s^2}{2}\right), \quad t, s \in [0, 1], \quad C < \infty \text{ is a constant.} \tag{2.11}$$

2.2 Functional principal components analysis

From now on, if there is no other particular illustration, when we mention a Hilbert space H, we mean $H = L^2([0, 1])$.

Suppose our observations $X_1, \ldots, X_N$ are square integrable functions, i.e. $X_1, \ldots, X_N \in H = L^2([0, 1])$. $X_1, \ldots, X_N$ are called *H-valued observations*. Suppose these observations are realizations of some random function $X = \{X(t, \omega) : t \in [0, 1], \omega \in \Omega\}$ which is defined on some common probability space $(\Omega, \mathcal{A}, P)$. It means, for $\forall \omega \in \Omega$, $X(\cdot, \omega) \in H = L^2([0, 1])$. And X is called an *H-valued random function*.

In this section, Section 2.2, we will introduce functional principal components analysis (FPCA). FPCA plays a central role in FDA. The basic idea is to represent the H-valued random function X with the eigenfunctions of the *covariace operator* of X, which is known as *Karhunen-Loéve representation*. Then X is truncated just with a fixed number d of eigenfunctions, where these d eigenfunctions can explain most of the variability of X.

2.2.1 Functional mean and covariance operator

Suppose X is an H-valued random function. We say $X \in L_H^2$, if

$$\begin{aligned} E\|X\|^2 &= E\left[\langle X, X\rangle\right] \\ &= E\left[\int_0^1 X^2(t)dt\right] < \infty. \end{aligned} \tag{2.12}$$

Generally, we say $X \in L_H^p$, if

$$\begin{aligned} E\|X\|^p &= E\left[\langle X, X\rangle^{\frac{p}{2}}\right] \\ &= E\left[\left(\int X^2(t)dt\right)^{\frac{p}{2}}\right] < \infty. \end{aligned} \tag{2.13}$$

Remark 2. *$\|\cdot\|$ is the norm equipped in $H = L^2([0, 1])$, thus $\|X\|^p$ means "$\|X\|$ to the power p". Do not confuses with the norm $\|.\|_p$ in $L^p([0, 1])$, i.e.*

$$\|y\|_p := \left(\int_0^1 \mid y(t) \mid^p \mathrm{d}t\right)^{\frac{1}{p}}, \quad y \in L^p([0, 1]).$$

For $X \in L_H^2$, we define the *functional mean* and the *covariance operator* of X in the following.

- The *functional mean* of X is defined as

$$\mu(t) := E[X(t)], \quad t \in [0,1]. \tag{2.14}$$

The expectation commutes with bounded operators, i.e if $\Psi \in \mathcal{L}$, $E[\Psi(X)] = \Psi(EX)$.

- The *covariance operator* C of X is defined as

$$\begin{aligned} C : H &\to H \\ x &\mapsto E\left[\langle X-\mu, x\rangle (X-\mu)\right], \quad x \in H. \end{aligned} \tag{2.15}$$

Thus,

$$\begin{aligned} C(x)(t) &= E\left[\int_0^1 \left[(X(s)-\mu(s))\, x(s)\mathrm{d}s\right] \cdot (X(t)-\mu(t))\right] \\ &= \int_0^1 \underbrace{E\left[(X(t)-\mu(t))\,(X(s)-\mu(s))\right]}_{:=c(t,s)} x(s)\mathrm{d}s \\ &:= \int_0^1 c(t,s)x(s)\mathrm{d}s, \quad t \in [0,1], \quad x \in H. \end{aligned} \tag{2.16}$$

Note that C is an integral operator defined in (2.8) and $c(t,s)$ is its corresponding integral kernel.

In the following we gather the properties of the covariance operator C defined in (2.16) into a theorem. Without loss of generity, we assume $\mu = EX = 0$.

Theorem 2.2. *Suppose $X \in L_H^2$ and $EX = 0$, then the covariance operator C of X defined in (2.16) is a symmetric non-negative definite Hilbert-Schmidt integral operator.*

Proof. (i) Symmetric:

$$c(t,s) = E[X(t)X(s)] = E[X(s)X(t)] = c(s,t), \quad \forall t,s \in [0,1],$$

Thus C is symmetric by Remark 1.

(ii) Non-negative definite:

$$\begin{aligned} \langle C(x), x\rangle &\overset{\text{by } (2.16)}{=} \iint c(t,s)x(t)x(s)dtds \\ &= \iint E[X(t)X(s)]x(t)x(s)dtds \\ &= E\left[\left(\int X(t)x(t)dt\right)^2\right] \geq 0, \quad \forall x \in H. \end{aligned}$$

(iii) Since

$$\begin{aligned}\iint c^2(t,s)dtds &= \iint \left(E[X(t)X(s)]\right)^2 dtds \\ &\leq \iint E\left[X(t)\right]^2 E\left[X(s)\right]^2 dtds \\ &= \left(\int_0^1 E\left[X(t)\right]^2 dt\right)^2 = \left(E\|X\|^2\right)^2 < \infty,\end{aligned}$$

then by (2.9) and (2.10), the integral operator C is Hilbert Schimdt.

□

2.2.2 Eigenfunctions and eigenvalues of the covariance operator

We denote $(\nu_j, j \in \mathbb{Z})$ and $(\lambda_j, j \in \mathbb{Z})$ by the (orthonormal) eigenfunctions and the corresponding eigenvalues of the covariance operator C, i.e.

$$C(\nu_j) = \lambda_j \nu_j, \quad j = 1, 2, \ldots. \tag{2.17}$$

By Theorem 2.2, the integral kernel $c(t,s)$ is symmetric and non-negative definite. Furthermore, since $X \in L_H^2$, then $c(t,s)$ (see the definition in (2.16)) is continuous. By Mercer's theorem (Theorem 2.1), the eigenvalues $(\lambda_j, j \in \mathbb{Z})$ are non-negative and the eigenfunctions $(\nu_j, j \in \mathbb{Z})$ form an orthonormal basis of H.

2.2.3 Karhunen-Loéve representation

Without loss of generity, we assume the random function $X \in L_H^2$ is with $EX = 0$. X can be represented with the eigenfunctions $(\nu_i, i \in \mathbb{Z})$ of the covariance operator C, which is known as the *Karhunen-Loéve representation.*

Theorem 2.3 (Karhunen-Loéve Theorem)**.** *Suppose $X \in L_H^2$ with $EX = 0$, then X can be represented by*

$$X = \sum_{i=1}^{\infty} \langle X, \nu_i \rangle \nu_i := \sum_{i=1}^{\infty} x_i \nu_i, \tag{2.18}$$

where $(\nu_i, i \in \mathbb{Z})$ are the orthonormal eigenfunctions of the covariance operator C defined in (2.17). $(x_i, i \in \mathbb{Z})$ defined in (2.18) are called the scores *of X. The scores $(x_i, i \in \mathbb{Z})$ are mean-zero, uncorrelated and with variance λ_i, i.e.*

$$\begin{aligned}&Ex_i = 0, \quad \forall i \in \mathbb{Z}, \\ &E(x_i x_j) = 0, \quad i \neq j, \\ &E(x_i)^2 = E\langle X, \nu_j \rangle^2 = \lambda_i, \quad \forall i \in \mathbb{Z},\end{aligned} \tag{2.19}$$

where $(\lambda_i, i \in \mathbb{Z})$ are the eigenvalues of the covariance operator C of X.

Remark 3. *For an arbitrary orthonormal basis* $(e_i, i \in \mathbb{Z})$ *of* H*, the representation*

$$X = \sum_{i=1}^{\infty} \langle X, e_i \rangle \, e_i$$

still holds by Parseval's identity. But the scores $(\langle X, e_i \rangle, i \in \mathbb{N})$ *may not be uncorrelated.*

By the last equation in (2.19), we have

$$\sum_{j=1}^{\infty} \lambda_j = \sum_{j=1}^{\infty} E \langle X, \nu_j \rangle^2 = E\|X\|^2 < \infty, \quad X \in L_H^2. \tag{2.20}$$

Combining (2.19) and (2.20), we can know that each λ_j can represent some proportion of the tatal variability of X. This is a very useful property. For any integer $d > 0$, we suppose $\lambda_1, \ldots, \lambda_d$ are the largest d eigenvalues of C. Then notion of *cumulative percentage of total variance (CPV(d))* is defined as

$$CPV(d) := \frac{\sum_{j=1}^{d} \lambda_j}{\sum_{j=1}^{\infty} \lambda_j}. \tag{2.21}$$

If we choose $d > 0$ such that the $CPV(d)$ exceeds a predetermined value, e.g. 90%, then $\lambda_1, \ldots, \lambda_d$ or the corresponding $\nu_1, \ldots, \nu_d$ explain most of the variability of X. Here $\nu_1, \ldots, \nu_d$ are also called the *functional principal components (FPC's)*. We can truncate the infinite-dimensional random function X by

$$X_{trunc} := \sum_{i=1}^{d} \langle X, \nu_i \rangle \, \nu_i. \tag{2.22}$$

The trucation X_{trunc} in (2.22) contains most of the information (variability) of the random function X.

The procedure mentioned above to find d such that $CPV(d)$ exceeds the predetermined value, is called the *CPV method*. Of course, in real data analysis, we do not know the exact value of $(\lambda_i, i \in \mathbb{Z})$ and $(\nu_i \in \mathbb{Z})$. Thus the $(\lambda_i, i \in \mathbb{Z})$ in (2.21) and the $(\nu_i, i \in \mathbb{Z})$ in (2.22) will be replaced by their corresponding empirical forms in real data analysis. The asymptotic properties of the empirical eigenvalues and eigenfunctions will be referred in Section 2.4.

In Figure 2.1 we show the application of the CPV method on highway functional velocity data. We compute the empirical eigenvalues $\lambda_1^e, \ldots, \lambda_N^e$ and the empirical eigenfunctions $\nu_1^e, \ldots, \nu_N^e$ from the dataset. Then we made the " $CPV(d)$ vs. d " plot in Figure 2.1. We can see from the plot that, $CPV(3) < 0.8$ and $CPV(4) > 0.8$. Thus if we set the criterion to 80%, then $d = 4$. Then for each day $n \in \{1, \ldots, N\}$, we truncate the daily functional velocity curve S_n by $S_{n,trunc} := \sum_{j=1}^{d} \left\langle S_n, \nu_j^e \right\rangle \nu_j^e$. In Figure 2.2 we show the (centered) functional velocity data and the corresponding truncation. One can see that the error made is limited.

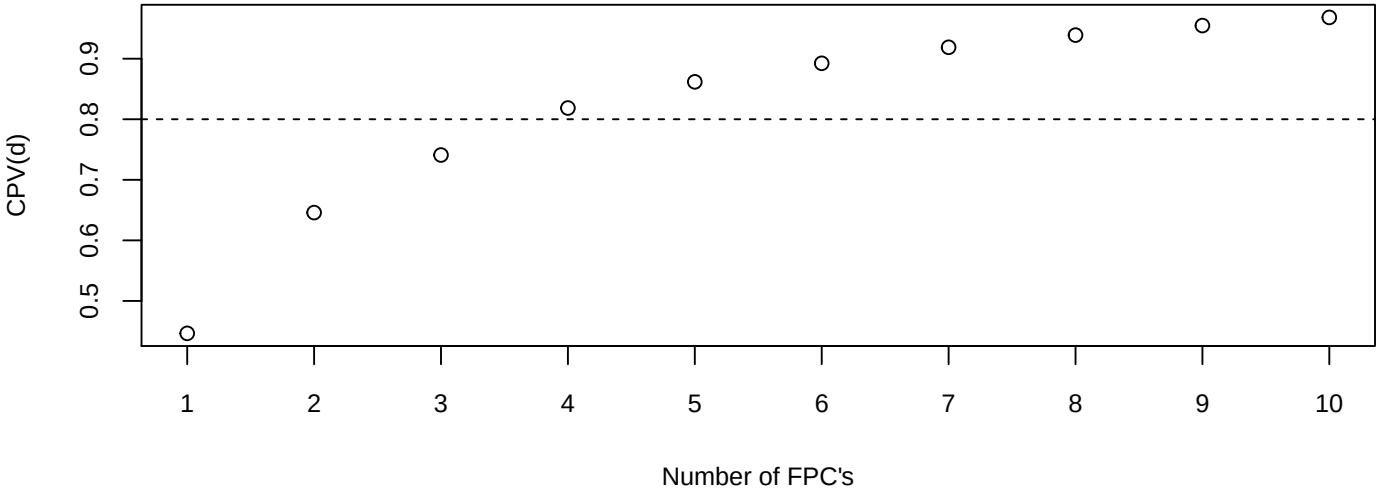

Figure 2.1: Application of CPV method to functional velocity data on 178 days. The CPV criterion is 80%, i.e. $\nu_1, \nu_2, \nu_3, \nu_4$ explain 80% of the total data variability. **Source:** Autobahndirektion Südbayern

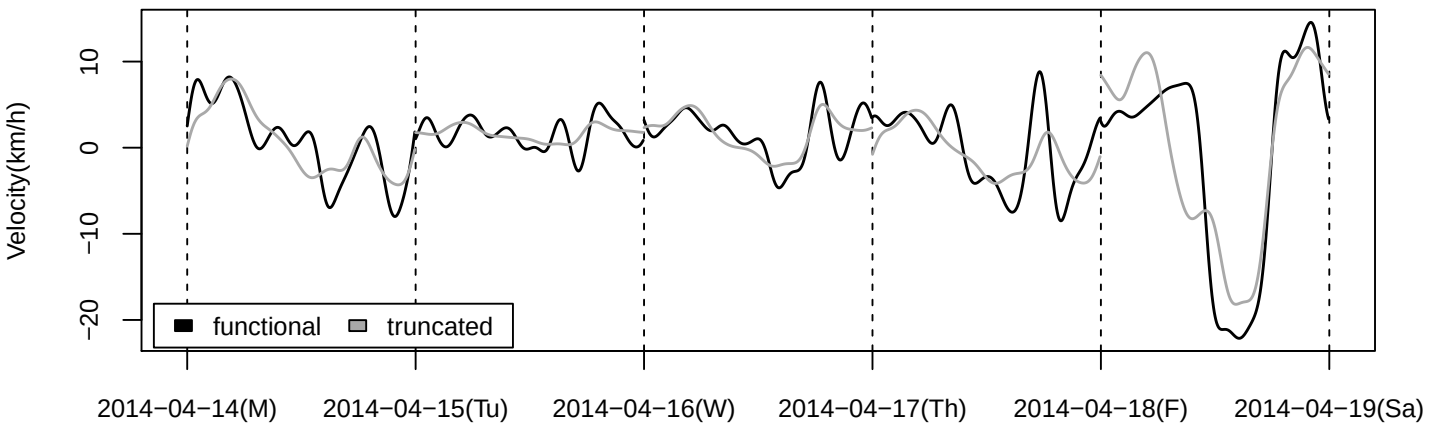

Figure 2.2: Highway (centered) functional velocity data in one week vs. the corresponding truncated data by Karhunen-Loéve representation. The CPV criterion is 80% and the number d of FPC's is 4. **Source:** Autobahndirektion Südbayern

2.3 Functional time series

At an early stage of development, FDA focused mainly on i.i.d functional data. But in many applications, the assumption of i.i.d is too strong, especially when data are collected sequentially over time, e.g the daily records of traffic conditions. It is natural to expect that the current traffic condition more or less depends on the previous conditions.

Hörmann and Kokoszka [2010] proposed the notion of *L^p-m-approximability* to quantify the *temporal dependence* of functional time series. Based on the work of Hörmann and Kokoszka [2010], the ideas of time series analysis and FDA have been merged and many results in FDA under the assumption of i.i.d have been extented to *L^p-m-approximable functional time series*.

In this section we will introduce some basic but important theories of functional time series, which include H-white noise, *L^p-m-approximability* and *Bernoulli shifts*.

2.3.1 White noise in general separable Hilbert space

In this subsection, Section 2.3.1, we do not restrict $H = L^2(0,1)$, i.e. H can be any separable Hilbert space (which includes $L^2([0,1])$).

Definition 2.4 (Definition 3.1, Bosq [2000])**.** *A sequence $(\varepsilon_n, n \in \mathbb{Z})$ of H-random varibles is said to be an H-*white noise *if*

(1) For each $n \in \mathbb{Z}$, $0 < E\|\varepsilon_n\|^2 = \sigma_\varepsilon^2 < \infty$, $E\varepsilon_n = 0$, and the covariance operator C_{ε_n} of ε_n, where

$$C_{\varepsilon_n}(x) := E\left[\langle \varepsilon_n, x\rangle \varepsilon_n\right], \quad x \in H, \tag{2.23}$$

does not depend on n. Thus we denote $C_{\varepsilon_n} := C_\varepsilon$.

(2) For $\forall n, m \in \mathbb{Z}, n \neq m$,

$$E\left[\langle \varepsilon_n, x\rangle \langle \varepsilon_m, y\rangle\right] = 0, \quad \forall x, y \in H. \tag{2.24}$$

*Furthermore, $(\varepsilon_n, n \in \mathbb{Z})$ is said to be H-*strong white noise (SWN)*, if it satiesfies* (1) *in Definition 2.4 and*

(2') $(\varepsilon_n, n \in \mathbb{Z})$ is i.i.d.

Similar to scalar and multivariate time series, H-white noise is often used as the innovations in the functional ARMA process. In the previous sections we have mentioned that, dimension reduction is the basic idea of FPCA. If we reduce the dimension of the H-white noise $(\varepsilon_i, i \in \mathbb{Z})$, will its truncated vector process be a multivariate white noise? The following theorem answers this question.

Theorem 2.5. *Suppose $(\varepsilon_n, n \in \mathbb{Z})$ is an H-valued white noise and $(e_i, i \in \mathbb{Z})$ is an arbitrary orthonormal basis of H. We define the d-dimensional vector process $(\mathbf{E}_n, n \in \mathbb{Z})$ truncated from $(\varepsilon_i, i \in \mathbb{Z})$, where*

$$\mathbf{E}_n := \left(\langle \varepsilon_n, e_1\rangle, \ldots, \langle \varepsilon_n, e_d\rangle\right)^T, \quad \forall n \in \mathbb{Z}. \tag{2.25}$$

Then $(\mathbf{E}_n, n \in \mathbb{N})$ in (2.25) is multivariate white noise.

Proof. To prove $(\mathbf{E}_n, n \in \mathbb{N})$ is a multivariate white noise, we need to show

$$\begin{aligned} &E\left[\mathbf{E}_n\right] = \mathbf{0}, \quad \forall n \in \mathbb{Z}, \\ &E\left[\mathbf{E}_n \mathbf{E}_m^T\right] = \mathbf{0}, \quad n \neq m, \\ &E\left[\mathbf{E}_n \mathbf{E}_n^T\right] \text{ does not depend on } n. \end{aligned}$$

- Since $E\varepsilon_n = 0$, we have

$$E\left[\mathbf{E}_n\right] = E\left(\langle \varepsilon_n, e_1\rangle, \ldots, \langle \varepsilon_n, e_d\rangle\right)^T = \mathbf{0}, \quad \forall n \in \mathbb{Z}.$$

- For $n \neq m$, by (2.25), we know $\mathbf{E}_n \mathbf{E}_m^T$ is a $d \times d$ matrix with ijth entry $\langle \varepsilon_n, e_i\rangle \langle \varepsilon_m, e_j\rangle$. Then by (2.24) in Definition 2.4, we have

$$E\left[\langle \varepsilon_n, e_i\rangle \langle \varepsilon_m, e_j\rangle\right] = 0, \quad \forall i, j \in \{1, \ldots, d\}.$$

  Thus $E\left[\mathbf{E}_n \mathbf{E}_m^T\right] = \mathbf{0}$ for $n \neq m$.

- For each n, $\mathbf{E}_n \mathbf{E}_n^T$ is a $d \times d$ matrix with ijth entry $\langle \varepsilon_n, e_i\rangle \langle \varepsilon_n, e_j\rangle$. Then by (2.23),

$$\begin{aligned} E\left[\langle \varepsilon_n, e_i\rangle \langle \varepsilon_n, e_j\rangle\right] &= \langle C_{\varepsilon_n}(e_i), e_j\rangle \\ &= \langle C_{\varepsilon}(e_i), e_j\rangle, \quad \forall i, j \in \{1, \ldots, d\}. \end{aligned}$$

  It implies that $E\left[\mathbf{E}_n \mathbf{E}_n^T\right]$ does not depend on n.

□

2.3.2 L^p-m-approximable functional time series

Now we introduce the notion *L^p-m-approximability* proposed by Hörmann and Kokoszka [2010] and this subsection, Section 2.3.2, is also mainly based on their paper.

Suppose $(X_n, n \in \mathbb{Z})$ is an H-valued sequence. We denote by $\mathcal{F}_t^- := \sigma\{\ldots, X_{t-1}, X_t\}$ and $\mathcal{F}_t^+ := \sigma\{X_t, X_{t+1}, \ldots\}$ the $\sigma-$algebras generated by the observations up to time t and after t.

Definition 2.6. *The H-valued sequence $(X_n, n \in \mathbb{Z})$ is said to be m-dependent if for $\forall t \in \mathbb{R}$, the σ-algebras $\mathcal{F}_t^-$ and $\mathcal{F}_{m+t}^+$ are independent.*

One idea is to approximate $(X_n, n \in \mathbb{Z})$ by m-dependent process $\left(X_n^{(m)}, n \in \mathbb{Z}\right)$, $m \geq 1$, where for each n, $\left(X_n^{(m)}, n \in \mathbb{Z}\right)$ converges in some sense (e.g in distribution) to X_n, as $m \to \infty$. If the convergence is fast enough, then we can obtain the limiting behavior of $(X_n, n \in \mathbb{Z})$ from $\left(X_n^{(m)}, n \in \mathbb{Z}\right)$. The following definition formalizes this idea.

Definition 2.7 (Definition 2.1, Hörmann and Kokoszka [2010])**.** *For $X \in L_H^p$, i.e. $E\|X\|^p < \infty$, we define*

$$\nu_p(X) := \left(E\|X\|^p\right)^{1/p} < \infty. \tag{2.26}$$

A sequence $(X_n, n \in \mathbb{Z}) \in L_H^p$ *is called* L^p*-m-*approximable *if each* X_n *admits the representation*

$$X_n = f(\varepsilon_n, \varepsilon_{n-1}, \dots), \tag{2.27}$$

where the $(\varepsilon_n, n \in \mathbb{Z})$ *is an i.i.d* H*-valued sequence, and* f *is a measurable function*

$$f : H^\infty \to H.$$

Moreover we assume that if $(\varepsilon'_n, n \in \mathbb{Z})$ *is an independent copy of* $(\varepsilon_n, n \in \mathbb{Z})$ *defined on the same probability space, then letting*

$$X_n^{(m)} = f(\varepsilon_n, \varepsilon_{n-1}, \dots, \varepsilon_{n-m+1}, \varepsilon'_{n-m}, \varepsilon'_{n-m-1}, \dots), \tag{2.28}$$

we have

$$\sum_{m=1}^{\infty} \nu_p(X_m - X_m^{(m)}) < \infty. \tag{2.29}$$

Since $(\varepsilon_n, n \in \mathbb{Z})$ is i.i.d, the L^p-m-approximable sequence $(X_n, n \in \mathbb{Z})$ in (2.27) is strictly stationary.

But we must note that $\left(X_n^{(m)}, n \in \mathbb{Z}\right)$ defined in (2.28) is *not* m-dependent. By (2.28),

$$X_{n+m}^{(m)} = f(\varepsilon_{n+m}, \varepsilon_{n+m-1}, \dots, \varepsilon_{n+1}, \varepsilon'_n, \dots, \varepsilon'_{n-m}, \dots). \tag{2.30}$$

Combining (2.28) and (2.30), we can see that ε'_{n-m} is included both in $X_n^{(m)}$ and $X_{n+m}^{(m)}$. It implies $X_n^{(m)}$ and $X_{n+m}^{(m)}$ are dependent, thus the $\left(X_n^{(m)}, n \in \mathbb{Z}\right)$ defined in (2.28) is *not* m-dependent.

To solve this problem, Hörmann and Kokoszka [2010] proposed a so-called *coupling construction method*. They defined for each n an independent copy $\{\varepsilon_k^{(n)}\}$ of $\{\varepsilon_k\}$, then $\left(X_n^{(m)}, n \in \mathbb{Z}\right)$ is re-defined as

$$X_n^{(m)} := f(\varepsilon_n, \varepsilon_{n-1}, \dots, \varepsilon_{n-m+1}, \varepsilon_{n-m}^{(n)}, \varepsilon_{n-m-1}^{(n)}, \dots). \tag{2.31}$$

Then for each $m \geq 1$, $\left(X_n^{(m)}, n \in \mathbb{Z}\right)$ is strictly stationary and m-dependent, and for each $n \in \mathbb{Z}$, $X_n^{(m)}$ is equal in distribution to X_n.

The L^p*-m-approximable* functional time series $(X_n, n \in \mathbb{Z})$ is *weakly dependent*, i.e. the correlation between X_n and X_{n+h} tends to zero sufficiently quickly as $h \to \infty$. Many results in FDA under the assumption of i.i.d can be extended to weakly-dependent functional time series (see Section 2.4).

In the following we introduce the *Bernoulli shifts*, which also belong to the class of L^p*-m-approximable* functional time series.

Definition 2.8. *Suppose* $(\varepsilon_n, n \in \mathbb{Z})$ *is an i.i.d* H*-valued sequence. Then* $(\eta_n, n \in \mathbb{Z})$ *forms a sequence of Bernoulli shifts, if for each* n, η_n *is* L^p*-m-approximable (see Definition 2.7), and*

$$\varepsilon_j(t) = \varepsilon_j(t, \omega) \quad \textit{is jointly measurable in } (t, \omega), \quad -\infty < j < \infty, \tag{2.32}$$

$$E\eta_0 = 0,\ E\|\eta_0\|^{2+\delta} < \infty, \quad \textit{for some } 0 < \delta < 1. \tag{2.33}$$

The sequence $(\eta_n, n \in \mathbb{Z})$ *can be approximated by m-dependent sequences* $\{\eta_n^{(m)}\}$ *(defined in the way of (2.31)) in the sense that*

$$\sum_{m=1}^{\infty} \left(E\|\eta_n - \eta_n^{(m)}\|^{2+\delta}\right)^{1/\kappa}, \quad \textit{for some} \;\; \kappa > 2+\delta. \tag{2.34}$$

Each causal stationary funtional time series can be represented in the form of functional moving average process, which is similar to the scalar and multivariate case. In fact, the L^p-m-approximable Bernoulli shifts defined in Definition2.8 include all (linear or nonlinear) stationary functional processes used in practice, and we will refer them again in Chapter 3 for the stationarity test of functional observations.

2.4 Estimation

Up to now, we have introduced the most important notions used in FDA. In this section, we will define their corresponding emprical forms (estimators) and show the asymptotic properties of the estimators under the assumption of i.i.d and weak dependence (L^p-m-approximability).

2.4.1 Estimation under the assumption of i.i.d

Assumption 2.1. *The H-valued observations* $X_1, \dots X_N$ *are i.i.d and have the same distribution as X.*

The *sample mean (or empirical mean)* of the observations $X_1, \dots X_N$ is defined as

$$\mu^e := \frac{1}{N}\sum_{k=1}^{N} X_k. \tag{2.35}$$

The superscript "e" stands for "empirical".

Recall the representation of the covariance operator C and the covariance kernel $c(t,s)$ in (2.15) and (2.16), it is natural to define the *empirical covariance operator* C^e and the *empirical covariance kernel* $c^e(t,s)$ as

$$C^e(x) := \frac{1}{N}\sum_{k=1}^{N} \langle X_k - \mu^e, x\rangle \, (X_k - \mu^e), \quad x \in H = L^2([0,1]), \tag{2.36}$$

$$c^e(t,s) := \frac{1}{N}\sum_{k=1}^{N} (X_k(t) - \mu^e(t))(X_k(s) - \mu^e(s)), \quad t,s \in [0,1]. \tag{2.37}$$

Figure 2.3 shows the empirical covariance kernel of the highway functional velocity data on working days. As indicated by the arrows, the $t = 0, s = 0$ point is in the bottom right corner. It estimates $E[(X(0) - \mu(0))(X(0) - \mu(0))]$. The empirical variance of the time series is represented along the diagonal from the bottom right to the top left corner. The peaks along the diagonal represent phases of transition between stable

and unstable traffic states: for instance, the first peak represents the transition at around 05:00 a.m., where traffic becomes denser due to commuting. Peaks away from the diagonal represent high dependencies between different time points. For instance, on working days high traffic density in the morning correlates high traffic density in the evening, again due to commuting.

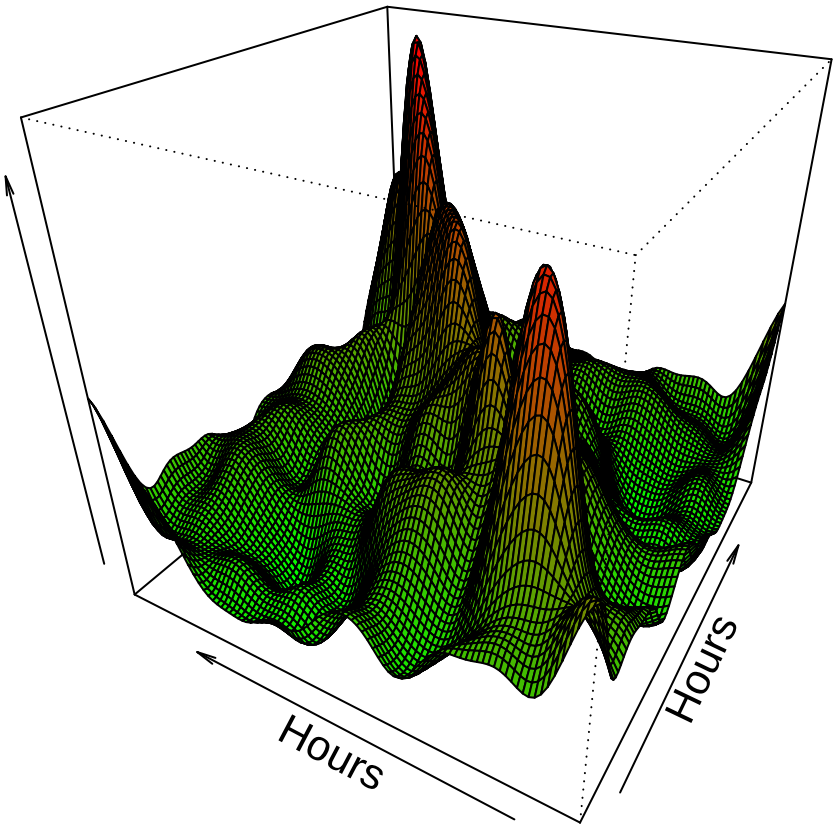

Figure 2.3: Empirical covariance kernel of highway functional velocity data on working days. **Source:** Autobahndirektion Südbayern

The next three theorem show the asymptotic properties of μ^e, C^e and $c^e(t,s)$.

Theorem 2.9 (Theorem 2.3, Horváth and Kokoszka [2012])**.** *If Assumption 2.1 holds, then* $E\mu^e = \mu$ *and* $E\|\mu^e - \mu\|^2 = O(N^{-1})$.

Theorem 2.9 implies that μ^e is an unbiased estimator of μ and consistent in the sense that $\|\mu^e - \mu\| \xrightarrow{P} 0$.

Theorem 2.10 (Theorem 2.4, Horváth and Kokoszka [2012]). *If $X \in L_H^4$, i.e. $E\|X\|^4 < \infty$, $EX = 0$ and Assumption 2.1 holds, then*

$$E\|C^e\|_{\mathcal{S}}^2 \leq E\|X\|^4, \tag{2.38}$$

where $\|\cdot\|_{\mathcal{S}}$ is the Hilbert-Schmidt operator norm.

Theorem 2.11 (Theorem 2.5, Horváth and Kokoszka [2012]). *If $E\|X\|^4 < \infty$, $EX = 0$ and Assumption 2.1 holds, then*

$$E\|C^e - C\|_{\mathcal{S}}^2 = E\iint \left[c^e(t,s) - c(t,s)\right]^2 \mathrm{d}t\mathrm{d}s \leq N^{-1}E\|X\|^4. \tag{2.39}$$

Theorem 2.11 implies that $c^e(t,s)$ is a mean squared consistent estimator of $c(t,s)$.

By the representation of C^e in (2.36), we know that C^e maps $H = L^2([0,1])$ into an N-dimensional subspace spanned by $X_1, \ldots, X_N$. Thus C^e has N eigenvalues and N eigenfunctions. It is natural to take the eigenvalues and the eigenfunctions of C^e as the empirical eigenvalues and eigenfunctions, i.e. the *empirical eigenvalues* and the corresponding *empirical eigenfunctions* are defined as

$$\lambda_j^e \nu_j^e = C^e(\nu_j^e) = \int_0^1 c^e(t,s)\nu_j^e(s)ds, \quad j = 1, \ldots, N. \tag{2.40}$$

The following theorem shows the asymptotic properties of the empirical eigenvalues and eigenfunctions defined in (2.40).

Theorem 2.12 (Theorem 2.7, Horváth and Kokoszka [2012]). *Suppose $E\|X\|^4 < \infty$, $EX = 0$, Assumption 2.1 holds, and*

$$\lambda_1 > \lambda_2 > \ldots \lambda_d > \lambda_{d+1}. \tag{2.41}$$

Then, for each $1 \leq j \leq d$ $(d \leq N)$,

$$E\left[\|\nu_j^e - \nu_j\|^2\right] = O(N^{-1}), \quad E\left[|\lambda_j^e - \lambda_j|^2\right] = O(N^{-1}). \tag{2.42}$$

Theorem 2.12 implies that, under regularity conditions, the empirical eigenvalues and eigenfunctions are consistent estimators.

We call $\nu_1^e, \ldots, \nu_d^e$ the *empirical functional principal components* (EFPC's). $\nu_1^e, \ldots, \nu_d^e$ form an orthonormal basis of the N-dimensional subspace spanned by $X_1, \ldots, X_N$, then we have

$$\frac{1}{N}\sum_{i=1}^N \|X_i\|^2 = \frac{1}{N}\sum_{i=1}^N \sum_{j=1}^N \left\langle X_i, \nu_j^e \right\rangle^2 = \sum_{j=1}^N \left\langle C^e(\nu_j^e), \nu_j^e \right\rangle = \sum_{j=1}^N \lambda_j^e, \tag{2.43}$$

i.e. each λ_j^e represents some proportion of the total variability of the observations. Furthermore, $\nu_1^e, \ldots, \nu_d^e$ can be extended to an orthonormal basis of H.

2.4.2 Estimation under the assumption of weak dependence

In this subsetion we show the asymptotic properties of the estimators mentioned in the last subsection, when $(X_n, n \in \mathbb{Z}) \in L_H^2$ is an L^p-m-approximable (see Definition 2.7) functional time series.

Theorem 2.13 (Theorem 4.1, Hörmann and Kokoszka [2012]). *Suppose $(X_n, n \in \mathbb{Z})$ is an H-valued L^2-m-approximable process with mean μ. Then $E\|\mu_N^e - \mu\|^2 = O(N^{-1})$.*

Theorem 2.14 (Theorem 3.1, Hörmann and Kokoszka [2010]). *Suppose $(X_n, n \in \mathbb{Z})$ is an H-valued L^4-m-approximable process with covariance operator C, and $E\|X_n\|^4 < \infty$. Then there exists some constant $U_X < \infty$, which does not depend on N, such that*

$$E\|C^e - C\|_{\mathcal{S}}^2 \leq U_X N^{-1}. \tag{2.44}$$

Theorem 2.15 (Theorem 3.2, Hörmann and Kokoszka [2010]). *Suppose $(X_n, n \in \mathbb{Z})$ is an H-valued L^4-m-approximable process with covariance operator C, $E\|X_n\|^4 < \infty$, and*

$$\lambda_1 > \lambda_2 > \ldots \lambda_d > \lambda_{d+1}. \tag{2.45}$$

Then for each $1 \leq j \leq d$ $(d \leq N)$,

$$E\left[\|\nu_j^e - \nu_j\|^2\right] = O(N^{-1}), \quad E\left[|\lambda_j^e - \lambda_j|^2\right] = O(N^{-1}). \tag{2.46}$$

2.4.3 Long-run variance kernel for functional time series

Let $(y_n, n \in \mathbb{Z})$ be a scalar stationary sequence with $Ey_n = \mu_y$. Its *long-run variance* is defined as

$$\sigma^2 := \sum_{i \in \mathbb{Z}} \mathrm{Cov}(y_0, y_i), \tag{2.47}$$

provided the sequence (2.47) is absolutely convergent. The long-run variance is used to measure the standard error of the sample mean when there is serial dependence, i.e. $\mathrm{Cov}(y_0, y_i) \neq 0$.

For an H-valued stationary functional time series $(X_n, n \in \mathbb{Z})$ with $EX_n = 0$, the *long-run variance kernel* of $(X_n, n \in \mathbb{Z})$ is defined in the similar way to (2.47) as

$$c_L(t,s) := E\left[X_0(t)X_0(s)\right] + \sum_{i \geq 1}\left[X_0(t)X_i(s)\right] + \sum_{i \geq 1}\left[X_0(s)X_i(t)\right]. \tag{2.48}$$

If the serial covariance is 0 (e.g. $(X_n, n \in \mathbb{Z})$ is i.i.d), then the long-run covariance kernel $c_L(t,s)$ is exactly the covariance kernel $c(t,s)$ defined in (2.16).

Up to now, the covariance kernel $c(t,s)$ is still applied much more widely than $c_L(t,s)$ in FDA (or FPCA), even if the serial covariance of the data is not 0. Maybe it is an outlook to try to apply the long-run covariance kernel $c_L(t,s)$ in FPCA when the data have dependence structure. Horváth et al. [2013b] have applied the long-run covariance kernel in the hypothesis test for stationarity of functional observations (see Chapter 3).

In this subsection we will first introduce the central limit theorem for the sample mean of an L^2-m-approximable functional time series, which is helpful to understand the long-run variance kernel. Then we give a form of the *emprical long-run covariance kernel* and show the consistency of the estimator. This subsection is mainly based on Horváth et al. [2013a].

Theorem 2.16 (Theorem 2.1, Horváth et al. [2013a]). *Suppose the functional time series $(X_n, n \in \mathbb{Z})$ is L^2-m-approximable (thus stationary, see (2.27)) and $EX_n = 0$, $E\|X_n\|^2 < \infty$. Then*

$$N^{-\frac{1}{2}} \sum_{i=1}^{N} X_i \xrightarrow{D} Z \quad \text{in } L^2, \tag{2.49}$$

where Z is a Gaussian process with

$$EZ(t) = 0, \quad \text{and} \quad E\left[Z(t)Z(s)\right] = c_L(t,s), \tag{2.50}$$

where $c_L(t,s)$ is defined in (2.48). Furthermore, $c_L(t,s)$ converges in $L^2\left([0,1]\times[0,1]\right)$, i.e. c_L is a square integrable function on the unit square.

By Theorem 2.16, $c_L(t,s)$ converges in $L^2\left([0,1]\times[0,1]\right)$, then by (2.9), the corresponding integral operator C_L with the integral kernel $c_L(t,s)$ is Hilbert Schmidt, where

$$C_L(x)(t) := \int_0^1 c_L(t,s)x(s)\mathrm{d}s, \quad x \in H = L^2([0,1]). \tag{2.51}$$

We denote $(\lambda_{nL}, n \in \mathbb{Z})$ and $(\nu_{nL}, n \in \mathbb{Z})$ by the eigenvalues and the corresponding eigenfunctions of of C_L defined in (2.51).

In the following we will define the empirical form $c_L^e(t,s)$ of $c_L(t,s)$. The definition of the *empirical long-run covariance kernel* $c_L^e(t,s)$ is not so "natural".

We consider the L^2-m-approximable sequence $(X_n, n \in \mathbb{Z})$ with mean μ and $E\|X_n\|^2 < \infty$, then for each n, $X_n(t)$ can be written as

$$X_n(t) = \mu(t) + \eta_n(t). \tag{2.52}$$

$(\eta_n, n \in \mathbb{Z})$ is thus L^2-m-approximable with zero mean and can be approximated by the m-dependent sequence $\left(\eta_n^{(m)}, n \in \mathbb{Z}\right)$ (see Definition 2.7). In addition, we assume

$$\lim_{m\to\infty} m\left[E\int\left(\eta_n(t) - \eta_n^{(m)}(t)\right)^2 \mathrm{d}t\right]^{1/2} = 0. \tag{2.53}$$

Let $K(\cdot)$ be a kernel (weight) function defined on the line and satisfying the following conditions:

$$K(0) = 1, \tag{2.54}$$
$$K(u) = 0, \text{ if } \mid u \mid > c, \text{ for some } c > 0, \tag{2.55}$$
$$\text{K is continuous,} \tag{2.56}$$
$$\text{K is bounded.} \tag{2.57}$$

We define the empirical (sample) correlation functions:

$$\gamma_i^e(t,s) := \frac{1}{N}\sum_{j=i+1}^{N}\left(X_j(t) - \bar{X}_N(t)\right)\left(X_{j-i}(s) - \bar{X}_N(s)\right), \tag{2.58}$$

where $\bar{X}_N(t) = \frac{1}{N}\sum_{i=1}^{N} X_i(t)$.

Now the *empirical long-run covariance kernel* $c_L^e(t,s)$ is defined as

$$c_L^e(t,s) := \gamma_0^e(t,s) + \sum_{i=1}^{N-1} K\left(\frac{i}{h}\right)\left(\gamma_i^e(t,s) + \gamma_i^e(s,t)\right), \tag{2.59}$$

where $h = h(N)$ is the smoothing bandwidth satisfying

$$h(N) \to \infty \quad \text{and} \quad \frac{h(N)}{N} \to 0, \quad \text{as } N \to \infty. \tag{2.60}$$

Then the empirical long-run covariance kernel $c_L^e(t,s)$ defined in (2.59) is a consistent estimator of $c_L(t,s)$, which is stated in the following theorem.

Theorem 2.17 (Theorem 2.2, Horváth et al. [2013a])**.** *Suppose* $(X_n, n \in \mathbb{Z})$ *is* L^2*-m approximable satisfying (2.52) and (2.53). Under conditions (2.54)-(2.57) and (2.60),*

$$\iint \left(c_L^e(t,s) - c_L(t,s)\right)^2 \mathrm{d}t\mathrm{d}s \xrightarrow{P} 0, \tag{2.61}$$

with $c_L(t,s)$ *defined in (2.48) and* $c_L^e(t,s)$ *defined in (2.59).*

Chapter 3

Hypothesis test for functional data

In the real data analysis, before we attempt to choose a model to fit to the data, we need to transform the data properly.

For example, if we want to fit ARMA model to the data, we need to make the data more stationary by e.g. deducting the trend and seasonal part. Thus, to test whether the data have been properly transformed is an important issue.

In this chapter we will introduce two hypothesis tests for functional data. One is the Portmanteau test of independence proposed by Gabrys and Kokoszka [2007]. The other one is the test of stationarity of functional time series, which was proposed by Horváth et al. [2013b]. We will apply these two tests to check whether our dataset is properly transformed in the real data analysis in Chapter 6.

3.1 Portmanteau test of independence for functional observations

This section is mainly based on Gabrys and Kokoszka [2007].

Suppose we have random functional observations $X_1(t), \ldots, X_N(t),\ t \in [0,1]$. We want to test

$$H_0 : X_1, \ldots, X_N \text{ are i.i.d,}$$
$$H_1 : H_0 \text{ does not hold.}$$

By FPCA, we can approximate the $X_n(t)$ with truncated Karhunen-Loéve representation

$$X_n(t) \approx \sum_{k=1}^{d} x_{kn}^e \nu_k^e(t), \tag{3.1}$$

where $\nu_1^e, \ldots, \nu_d^e$ are empirical eigenfuntions, and x_{kn}^e is the kth score of X_n:

$$x_{kn}^e := \int_0^1 X_n(t)\nu_k^e(t)\mathrm{d}t. \tag{3.2}$$

Note that here all the notations are in the empirical form.

The d in (3.1) is choosen with the CPV method (see Section 2.2.3). To establish the null distribution of the test statistic, the following assumption is required.

Assumption 3.1. *The H-valued functional observations $X_1, \ldots, X_N$ are i.i.d. For each $n \in \{1, \ldots, N\}$, $E\|X_n\|^4 < \infty$ and $EX_n = 0$. Furthermore, we assume that the population eigenvalues satisfy*

$$\lambda_1 > \lambda_2 > \cdots > \lambda_d > \lambda_{d+1}. \tag{3.3}$$

We work with the vector consisting of scores (see (3.2))

$$\mathbf{X}_n^e := \left[x_{1n}^e, x_{2n}^e, \ldots, x_{dn}^e\right]^T, \tag{3.4}$$

and the unobservable vectors in population form

$$\mathbf{X}_n := \left[x_{1n}, x_{2n}, \ldots, x_{dn}\right], \tag{3.5}$$

where

$$x_{kn} = \int_0^1 X_n(t)\nu_k(t)\mathrm{d}t, \quad k = 1, \ldots, d. \tag{3.6}$$

Under H_0, the $\mathbf{X}_n$ in (3.5) are i.i.d zero mean d-dimensional random vectors. We denote the cross-covariance matrix by $\mathbf{V} := E\left[\mathbf{X}_n, \mathbf{X}_n\right]$. Then the ijth entry $v(i,j)$ of $\mathbf{V}$ is

$$v(i,j) = E[x_{in}x_{jn}], \quad i, j = 1, \ldots, d.$$

And we define $\mathbf{C}_h$ as the sample autocovariance with klth entry

$$c_h(k,l) := \frac{1}{N}\sum_{n=1}^{n-h} x_{kn}x_{l,n+h}, \quad k, l = 1, \ldots, d. \tag{3.7}$$

Remark 4. *Since $\mathbf{X}_n$ in (3.5) are unobservable, $\mathbf{C}_h$ in (3.7) can not be computed from the data. So we do not put the superscript "e" on $\mathbf{C}_h$, even though we call it "sample" autocovariance.*

Finally, the test statistic is constructed as

$$Q_N^e = N\sum_{h=1}^{H}\sum_{i,j=1}^{d} r_{f,h}^e(i,j)r_{b,h}^e(i,j), \tag{3.8}$$

where $r_{f,h}^e(i,j)$ and $r_{b,h}^e(i,j)$ are the ijth entry of $\left(\mathbf{C}_0^e\right)^{-1}\mathbf{C}_h^e$ and $\mathbf{C}_h^e\left(\mathbf{C}_0^e\right)^{-1}$ respectively, where the ijth entry of $\mathbf{C}_h^e$ is

$$c_h^e(k,l) := \frac{1}{N}\sum_{n=1}^{n-h} x_{kn}^e x_{l,n+h}^e, \quad k, l = 1, \ldots, d. \tag{3.9}$$

The following lemma provides us with an equivaluent form of the test statistic Q_N^e in (3.8).

Lemma 3.1 (Lemma 7.1, Horváth and Kokoszka [2012])**.** *The statistic Q_N^e in (3.8) has an equivalent form*

$$Q_N^e = N\sum_{h=1}^{H}\sum_{i,j=1}^{d}\left(c_h^e(i,j)\right)^2\left(\lambda_i^e\right)^{-1}\left(\lambda_j^e\right)^{-1}, \tag{3.10}$$

where $\lambda_1^e, \ldots, \lambda_d^e$ are the empirical eigenvalues.

The following theorem establishes the limit null distribution of the test statistic Q_N^e.

Theorem 3.2 (Theorem 1, Gabrys and Kokoszka [2007]). *Under H_0, if Assumption 3.1 holds, then $Q_N^e \xrightarrow{D} \chi^2_{d^2H}$ (Chi-square distribution with d^2H degrees of freedom).*

3.2 Testing stationarity for functional time series

This section is mainly based on Horváth et al. [2013b].

Suppose $(\eta_n, n \in \mathbb{Z})$ are Bernoulli shifts defined in Definition 2.8. We wish to test

$$\begin{aligned} &H_0 : X_i(t) = \mu(t) + \eta_i(t), \quad 1 \le i \le N,\ \mu \in H = L^2([0,1]), \\ &H_A : H_0 \text{ does not hold.} \end{aligned} \tag{3.11}$$

Remark 5. *The H_0 in (3.11) implies that $(X_n, n \in \mathbb{Z})$ is stationary and can be represented in the form of a functional moving average process.*

Remark 6. *The mean function $\mu(t)$ in (3.11) is unknown. Furthermore, the alternative hypothesis H_A is very general. In Horváth et al. [2013b], the asymptotic behaviors of the test statistics under the following alternatives have also been studied.*

- *Change point alternative:*

$$H_{A,1} : X_i(t) = \mu(t) + \delta(t) I\{i > k^*\} + \eta_i(t), \quad 1 \le i \le N,\ 1 \le k^* = \lfloor N\tau \rfloor < N,$$

  where $0 < \tau < 1$. The size of the change $\delta(t)$ and the time of the change k^* are all unknown parameters.

- *Integrated alternative:*

$$H_{A,2} :\ X_i(t) = \mu(t) + \sum_{l=1}^{i} \eta_l(t), \quad 1 \le i \le N.$$

- *Deterministic trend alternative:*

$$H_{A,3} :\ X_i(t) = \mu(t) + g(i/N)\delta(t) + \eta_i(t), \quad 1 \le i \le N,$$

  where $g(t)$ is a piecewise Lipschitz continuous function on $[0,1]$.

But in our thesis, we will not list the asymptotic behaviors of the test statistics under the alternatives above. More details see Horváth et al. [2013b]. In our real data analysis, if we can not reject H_0 in (3.11), then we will attempt to fit ARMA models to the data.

Before we introduce the test statistics, let us first get familiar with some notations.

The sample mean of the observations is defined as

$$\bar{X}(t) := \frac{1}{N}\sum_{i=1}^{N} X_i(t). \tag{3.12}$$

The partial sum process of the curves $X_1(t), \ldots, X_N(t)$ is defined as

$$S_N(x,t) := \frac{1}{\sqrt{N}}\sum_{i=1}^{\lfloor Nx \rfloor} X_i(t), \quad 0 \le x,t \le 1, \tag{3.13}$$

where $\lfloor \cdot \rfloor$ denotes the floor function.

We define

$$Z_N(x,t) := S_N(x,t) - xS_N(1,t), \quad 0 \le x,t \le 1. \tag{3.14}$$

Thus $Z_N(x,t)$ has the form of a functional Brownian bridge.

With these notations, we can introduce the test statistics. In Horváth et al. [2013b], they introduced two classes of tests. One is based on the curves themselves, i.e. on population level. The other one is based on the finite dimensional projections of the curves on the EFPC's.

- **Fully functional tests**

  The test statistics T_N and M_N are defined as

$$T_N := \iint Z_N^2(t,s)\mathrm{d}t\mathrm{d}x, \tag{3.15}$$

$$\begin{aligned} M_N &:= T_N - \int_0^1 \left(\int_0^1 Z_N(x,t)\mathrm{d}x\right)^2 \mathrm{d}t \\ &= \iint \left(Z_N(x,t) - \int_0^1 Z_N(y,t)\mathrm{d}y\right)^2 \mathrm{d}x\mathrm{d}t. \end{aligned} \tag{3.16}$$

- **Tests based on finite-diemensional projections**

  There are four test statistics, which are defined respectively as

$$T_N^0(d) := \sum_{i=1}^{d} \frac{1}{\lambda_{iL}^e} \int_0^1 \langle Z_N(x,\cdot), \nu_{iL}^e\rangle^2 \,\mathrm{d}x, \tag{3.17}$$

$$T_N^*(d) := \sum_{i=1}^{d} \int_0^1 \langle Z_N(x,\cdot), \nu_{iL}^e\rangle^2 \,\mathrm{d}x, \tag{3.18}$$

$$M_N^0(d) := \sum_{i=1}^{d} \frac{1}{\lambda_{iL}^e} \int_0^1 \left(\langle Z_N(x,\cdot), \nu_{iL}^e\rangle - \int_0^1 \langle Z_N(u,\cdot), \nu_{iL}^e\rangle \,\mathrm{d}u\right)^2 \mathrm{d}x, \tag{3.19}$$

$$M_N^*(d) := \sum_{i=1}^{d} \int_0^1 \left(\langle Z_N(x,\cdot), \nu_{iL}^e\rangle - \int_0^1 \langle Z_N(u,\cdot), \nu_{iL}^e\rangle \,\mathrm{d}u\right)^2 \mathrm{d}x. \tag{3.20}$$

where $\lambda^e_{1L}, \ldots, \lambda^e_{dL}$ and $\nu^e_{1L}, \ldots, \nu^e_{dL}$ are the eigenvalues and eigenfunctions of the empirical long-run covariance kernel $c^e_L(t,s)$ defined in (2.59).

Under H_0 in (3.11), the asymptotic behaviour of the test statistics above are stated in the following two theorems. We assume $B_1, B_2, \ldots$ are i.i.d Brownian bridges on $[0,1]$.

Theorem 3.3 (Theorem 2.1, Horváth et al. [2013b]). *Under H_0,*

$$T_N \xrightarrow{D} \sum_{i=1}^{\infty} \lambda_{iL} \int_0^1 B_i^2(x)\mathrm{d}x, \tag{3.21}$$

and

$$M_N \xrightarrow{D} \sum_{i=1}^{\infty} \lambda_{iL} \int_0^1 \left(B_i(x) - \int_0^1 B_i(y)\mathrm{d}y \right)^2 \mathrm{d}x, \tag{3.22}$$

where $(\lambda^e_{iL}, i \in \mathbb{Z})$ and $(\nu^e_{iL}, i \in \mathbb{Z})$ are the eigenvalues and eigenfunctions of the long-run covariance kernel $c_L(t,s)$ defined in (2.48).

Theorem 3.4 (Theorem 2.2, Horváth et al. [2013b]). *Under H_0 and assume $\lambda_{1L} > \lambda_{2L} > \ldots \lambda_{dL} > \lambda_{d+1,L} > 0$, then*

$$T^0_N(d) \xrightarrow{D} \sum_{i=1}^{d} \int_0^1 B_i^2(x)\mathrm{d}x, \tag{3.23}$$

$$T^*_N(d) \xrightarrow{D} \sum_{i=1}^{d} \lambda_{iL} \int_0^1 B_i^2(x)\mathrm{d}x, \tag{3.24}$$

$$M^*_0(d) \xrightarrow{D} \sum_{i=1}^{d} \int_0^1 \left(B_i(x) - \int_0^1 B_i(y)\mathrm{d}y \right)^2 \mathrm{d}x, \tag{3.25}$$

$$M^*_N(d) \xrightarrow{D} \sum_{i=1}^{d} \lambda_{iL} \int_0^1 \left(B_i(x) - \int_0^1 B_i(y)\mathrm{d}y \right)^2 \mathrm{d}x. \tag{3.26}$$

The proof of Theorem 3.3 and Theorem 3.4 see Horváth et al. [2013b].

Chapter 4

Prediction of functional ARMA process

This chapter addresses the study of functional ARMA$(p, 1)$ process. Providing reliable predictions is one of the most important goals of functional time series analysis. In scalar and multivariate time series analysis, there already exist many prediction methods which can be easily implemented, e.g. Durbin-Levinson and innovations algorithms (e.g see Brockwell and Davis [1991]). In functional case, Bosq [2000] has proposed the *functional best linear predictor* for general linear process. But it is difficult to implement in practice, because we do not know the exact math formula of the predictor. Since we still lack advanced prediction methodology for functional time series, the functional time series is often assumed to follow the functional AR(1) (FAR(1)) models. And the prediction is also based on the assumption of FAR(1) structure (see Chapter 3 of Bosq [2000]).

Aue et al. [2015] proposed a prediction algorithm which combines the idea of FPCA and time series analysis, and it is not restricted to FAR structure. The basic idea is to reduce the infinite-dimensional functional data to finite-dimensional vector data (by FPCA). Then the issue of predicting functional time series is transformed to the prediction of multivariate time series. Furthermore, Aue et al. [2015] applied the algorithm to predict the functional AR(1) process and bounded the prediction error.

In this chapter we will try to extend the work by Aue et al. [2015]. In Section 4.1 and 4.2 we will quickly review the work by Bosq (functional best linear predictor) and Aue *et al.* (the prediction algorithm mentioned above). Then we will focus on the functional ARMA$(p, 1)$ process and its corresponding truncated vector process. In Section 4.3, we will first seek a sufficient condition for stationarity of functional ARMA$(p, 1)$ process. Then we will have a closer look on the vector process truncated from the functional process. We will show that, the vector process "approximately" follows the vector ARMA$(p, 1)$ structure. Furthermore, under some assumptions, the vector process is rigorously a stationary vector ARMA$(p, 1)$ process. In Section 4.4, we will show the relation between the functional best linear predictor (based on functional observations) and the vector best linear predictor (based on the truncated vector observations). At last, in Section 4.5, we will bound the prediction error when we apply the prediction algorithm by Aue et al. [2015] on the prediction of functional ARMA$(p, 1)$ process.

In this chapter we still assume $H := L^2([0, 1])$, and all our random functions are defined

on some common probability space $(\Omega, \mathcal{A}, P)$. H is equipped with the inner product $\langle x, y\rangle := \int_0^1 x(t)y(t)\mathrm{d}t$ and the norm $\|x\| := \sqrt{\langle x, x\rangle}$, $\forall x, y \in H$. We say an H-valued random function X is in L_H^p if $E\|X\|^p < \infty$.

4.1 Functional best linear predictor

Suppose we have a d-dimensional stationary time series $(\mathbf{Y}_n, n \in \mathbb{Z})$ with $E\mathbf{Y}_n = \mathbf{0}$. We denote the "matrix linear span" of the observations $\mathbf{Y}_1, \ldots, \mathbf{Y}_n$ by $\mathbf{M}_1$, where

$$\mathbf{M}_1 := \left\{ \sum_{i=1}^{n} \mathbf{A}_{ni}\mathbf{Y}_i,\ \mathbf{A}_{ni} \text{ is real } d \times d \text{ matrix},\ i = 1, \ldots, n \right\}.$$

Then the *vector best linear predictor* $\hat{\mathbf{Y}}_{n+1}$ of $\mathbf{Y}_{n+1}$ based on the observations $\mathbf{Y}_1, \ldots, \mathbf{Y}_n$ is defined as the orthogonal projection of $\mathbf{Y}_{n+1}$ on $\mathbf{M}_1$, i.e.

$$\hat{\mathbf{Y}}_{n+1} := P_{\{\mathbf{M}_1\}}\mathbf{Y}_{n+1}.$$

Equivalently, we have

$$E\left[\left(\mathbf{Y}_{n+1} - \hat{\mathbf{Y}}_{n+1}\right)\mathbf{y}^T\right] = \mathbf{0}, \quad \forall\ \mathbf{y} \in \mathbf{M}_1.$$

By the projection theorem (see Theorem 2.3.1 of Brockwell and Davis [1991]), $\hat{\mathbf{Y}}_{n+1}$ is the unique elmement in $\mathbf{M}_1$ such that

$$E\|\mathbf{Y}_{n+1} - \hat{\mathbf{Y}}_{n+1}\|_2^2 = \inf_{\mathbf{y}\in\mathbf{M}_1} E\|\mathbf{Y}_{n+1} - \mathbf{y}\|_2^2,$$

where $\|\cdot\|_2$ denotes the Euclidean vector norm.

In the following we will introduce the notion of *functional best linear predictor* proposed by Bosq [2000].

Suppose we have an H-valued stationary functional time series $(X_n, n \in \mathbb{Z})$ with $EX_n = 0$ and $E\|X_n\|^2 < \infty$, i.e. $X_n \in L_H^2$. The *functional best linear predictor* $\hat{X}_{n+1}$ of X_{n+1} based on the observations $X_1, \ldots, X_n$, is defined as the orthogonal projection of X_{n+1} on a specific subspace of L_H^2 containing $X_1, X_2, \ldots X_n$. This specific subspace of L_H^2 is introduced in the following definition.

Definition 4.1 (Definition 1.1, Bosq [2000]). *$\mathcal{G}$ is said to be an **$\mathcal{L}$-closed subspace**(LCS) or **hermetically closed subspace** of L_H^2, if*

(1) $\mathcal{G}$ is a Hilbertian subspace of L_H^2.

(2) If $X \in \mathcal{G}$ and $l(\cdot) \in \mathcal{L}$, then $l(X) \in \mathcal{G}$, where $\mathcal{L}$ denotes the space of bounded linear operators acting on H.

$\mathcal{G}$ is said to be a zero-mean LCS if it contains only zero-mean H-random variables.

A property of the LCS is listed in the following theorem.

Theorem 4.2 (Theorem 1.8, Bosq [2000]). *Let F be a subset of L_H^2. Then the LCS $\mathcal{G}_F$ generated by F is the closure of $\mathcal{G}'_F$, where*

$$\mathcal{G}'_F := \left\{ \sum_{i=1}^{n} l_i(X_i), l_i \in \mathcal{L}, X_i \in F,\ i = 1, \dots n,\ n \geq 1 \right\}.$$

Now we define

$$F_n := \{X_1, \dots, X_n\},$$

by Theorem 4.2, the LCS $\mathcal{G} := \mathcal{G}_{F_n}$ generated by F_n , is the closure of $\mathcal{G}' := \mathcal{G}'_{F_n}$, where

$$\mathcal{G}'_{F_n} := \left\{ \sum_{i=1}^{n} l_i(X_i), l_i \in \mathcal{L}, X_i \in F_n,\ i = 1, \dots n,\ n \geq 1 \right\}. \tag{4.1}$$

Note that $\mathcal{G}$ contains full information of $X_1, X_2, \dots, X_n$.

The *functional best linear predictor* $\hat{X}_{n+1}^{\mathcal{G}}$ of X_{n+1} is defined as the orthogonal projection of X_{n+1} on $\mathcal{G}$, i.e.

$$\hat{X}_{n+1}^{\mathcal{G}} := P_{\{\mathcal{G}\}} X_{n+1} \in \mathcal{G}. \tag{4.2}$$

Equivalently, we have

$$E \left\langle X_{n+1} - \hat{X}_{n+1}^{\mathcal{G}}, y \right\rangle = 0, \quad \forall y \in \mathcal{G}. \tag{4.3}$$

Since $\hat{X}_{n+1}^{\mathcal{G}} \in \mathcal{G}$, $\hat{X}_{n+1}^{\mathcal{G}}$ has the form of

$$\hat{X}_{n+1}^{\mathcal{G}} = \sum_{i=1}^{n} g_{n,i}(X_i) \ \in \mathcal{G}, \quad g_{n,i} \in \mathcal{L},\ i = 1, \dots, n.$$

Again, by the projection theorem, $\hat{X}_{n+1}^{\mathcal{G}}$ is the unique element in $\mathcal{G}$ such that

$$E\|X_{n+1} - \hat{X}_{n+1}^{\mathcal{G}}\|^2 = \inf_{y \in \mathcal{G}} E\|X_{n+1} - y\|^2.$$

We denote the mean square error of the predictor $\hat{X}_{n+1}^{\mathcal{G}}$ by

$$\sigma_n^2 := E\|X_{n+1} - \hat{X}_{n+1}^{\mathcal{G}}\|^2. \tag{4.4}$$

4.2 Prediction algorithm

Suppose the H-valued stationary functional time series $(X_n, n \in \mathbb{Z})$ is with $EX_n = 0$ and $E\|X_n\|^2 < \infty$. And we have N observations $X_1, X_2, \dots X_N$.

For simplicity, we call the prediction algorithm by Aue et al. [2015] **Algorithm I**. It can be organized into three steps.

Algorithm I:

(1) Select d, the number of FPC's, with CPV method (see Section 2.2.3) such that most of the total data variability can be explained by $\nu_1, \dots, \nu_d$. We compute the FPC scores $x_{k,l} := \langle X_k, \nu_l \rangle$ by projecting each observation on $\nu_1, \dots, \nu_d$. We write the scores into the vector form:

$$\mathbf{X}_k := (x_{k,1}, \dots x_{k,d})', \quad k = 1, \dots N, \ l = 1, \dots, d. \tag{4.5}$$

(2) We view the d-dimensional vectors $\mathbf{X}_1, \dots \mathbf{X}_N$ as observations. We fix the prediction lag h (in this thesis we choose $h = 1$), then we choose an appropriate multivariate prediction algorithm, e.g. innovations algorithm, to produce the one-step ahead (vector) predictor

$$\hat{\mathbf{X}}_{N+1} = (\hat{x}_{N+1,1}, \dots \hat{x}_{N+1,d})'.$$

(3) At last, we re-transform the vector predictor $\hat{\mathbf{X}}_{N+1}$ into the functional form $\hat{X}_{N+1}$ by the truncated Karhunen-Loéve representation:

$$\begin{aligned} \hat{X}_{N+1} &:= \hat{x}_{N+1,1}\nu_1 + \cdots + \hat{x}_{N+1,d}\nu_d \\ &= (\nu_1 \ \dots \ \nu_d)\, \hat{\mathbf{X}}_{N+1}. \end{aligned}$$

Note that, in **Algorithm I**, no specific data struture is required. Thus it can be applied on the prediction of any stationary functional time series. This is very important, since there is still no well developed theory of the prediction of functional ARMA process.

The first and the third step in **Algorithm I** can be implemented in R with the package **FDA**, and the second step can be achieved with the R package **MTS**.

4.3 Sufficient conditions for stationarity

Algorithm I can be applied on the prediction of the stationray functional ARMA$(p, 1)$ process. But we must note that, if the vector process in (4.5) is *not* stationary, things become more complicated. It is more difficult to predict a non-stationary process than a stationary process. Thus, our motivation is to seek conditions under which the vector process in (4.5) is stationary.

Before that, in the beginning of this section, we will first try to find out a condition for stationarity of the functional ARMA$(p, 1)$ process. Secondly, we will study the structure

of the truncated vector process in (4.5). We will show that, the vector process "approximately" follows the vector ARMA$(p,1)$ structure. And under some further constraints, it is rigorously a vector ARMA$(p,1)$ process. At the end of this section, we will try to seek conditions under which the vector process is stationary as well.

4.3.1 Sufficient conditions for stationarity of functional ARMA$(p,1)$ process

Definition 4.3. *A zero mean, H-valued sequence $(X_n, n \in \mathbb{Z})$ with $E\|X_n\|^2 < \infty$, $\forall n \in \mathbb{Z}$, is called a* ***functional ARMA*** $(\mathbf{p},\mathbf{1})$ ***process****, if it is* ***stationary*** *and for every $n \in \mathbb{Z}$,*

$$X_n = \sum_{i=1}^{p} \phi_i(X_{n-i}) + \varepsilon_n + \theta(\varepsilon_{n-1}), \tag{4.6}$$

where the H-valued i.i.d sequence $(\varepsilon_n, n \in \mathbb{Z})$ is with $E\varepsilon_n = 0$ and satisfies $E\|\varepsilon_n\|^2 = \sigma_\varepsilon^2 < \infty$. And $\phi_1, \ldots, \phi_p, \theta \in \mathcal{L}$.

In this section, we will seek a sufficient condition for stationarity of functional ARMA$(p,1)$ process in (4.6). To make the this part more understandable, we start with functional ARMA$(1,1)$ process. Then we will use state space equation to extend the result of ARMA(1,1) to ARMA$(p,1)$. First of all, we need the next two lemmas.

Lemma 4.4 (Lemma 3.1, Bosq [2000])**.** *For any $\phi \in \mathcal{L}$, the following two conditions are equivalent:*

C_0 : *There exists an integer j_0 such that $\|\phi^{j_0}\|_{\mathcal{L}} < 1$.*

C_1 : *There exist $a > 0$ and $0 < b < 1$ such that for every $j \geq 0$, $\|\phi^j\|_{\mathcal{L}} < ab^j$.*

If $\phi \in \mathcal{L}$ satisfies C_1, then it follows

$$\sum_{j=0}^{\infty} \|\phi^j\|_{\mathcal{L}}^2 < \sum_{j=0}^{\infty} a^2 b^{2j} = \frac{a^2}{1-b^2} < \infty. \tag{4.7}$$

Remark 7. *$\|\phi\|_{\mathcal{L}} < 1$ is the special case of C_0 (by choosing $j_0 = 1$). But C_0 (or C_1) does not imply $\|\phi\|_{\mathcal{L}} < 1$ (see Example 3.4 in Bosq [2000]). Thus C_0 (or C_1) is weaker than $\|\phi\|_{\mathcal{L}} < 1$.*

Lemma 4.5. *The H-valued i.i.d sequence $(\varepsilon_n, n \in \mathbb{Z})$ is with zero mean and satisfies $E\|\varepsilon_n\|^2 = \sigma_\varepsilon^2 < \infty$. If C_0 holds, then*

$$\sum_{j=0}^{\infty} \left\|\phi^j(\varepsilon_{n-j})\right\| < \infty \quad a.s. \tag{4.8}$$

implies the almost sure convergence of

$$\sum_{j=0}^{\infty} \phi^j(\varepsilon_{n-j}). \tag{4.9}$$

Proof. We define $C_m := \sum_{j=0}^{m} \phi^j(\varepsilon_{n-j})$. Since H is a Hilbert space, if C_m converges in H, then the limit of C_m (defined in (4.9)) is still in H. For simplicity, we denote the limit of C_m defined in (4.9) by $C := \sum_{j=0}^{\infty} \phi^j(\varepsilon_{n-j})$.

Similarly, we define $D_m := \sum_{j=0}^{m} \|\phi^j(\varepsilon_{n-j})\|$, and the limit of D_m defined in (4.8) is simplified to $D := \sum_{j=0}^{\infty} \|\phi^j(\varepsilon_{n-j})\|$. Then

$$\begin{aligned} P\left(\lim_{m\to\infty} D_m = D\right) &= P\left(\lim_{m\to\infty} (D - D_m) = 0\right) \\ &= P\left(\lim_{m\to\infty} \sum_{j=m}^{\infty} \|\phi^j(\varepsilon_{n-j})\| = 0\right) \end{aligned}$$

Since

$$\left\| \sum_{j=m}^{\infty} \phi^j(\varepsilon_{n-j}) \right\| \leq \sum_{j=m}^{\infty} \|\phi^j(\varepsilon_{n-j})\|,$$

we have

$$P\left(\lim_{m\to\infty} \sum_{j=m}^{\infty} \|\phi^j(\varepsilon_{n-j})\| = 0\right) \leq P\left(\lim_{m\to\infty} \left\| \sum_{j=m}^{\infty} \phi^j(\varepsilon_{n-j}) \right\| = 0\right).$$

In a Hilbert space H, $x = 0$ if and only if $\|x\| = 0$, $x \in H$. Thus,

$$\begin{aligned} P\left(\lim_{m\to\infty} \left\| \sum_{j=m}^{\infty} \phi^j(\varepsilon_{n-j}) \right\| = 0\right) &= P\left(\lim_{m\to\infty} \sum_{j=m}^{\infty} \phi^j(\varepsilon_{n-j}) = 0\right) \\ &= P\left(\lim_{m\to\infty} (C - C_m) = 0\right) \\ &= P\left(\lim_{m\to\infty} C_m = C\right). \end{aligned}$$

Thus, $P\left(\lim_{m\to\infty} D_m = D\right) = 1$ implies $P\left(\lim_{m\to\infty} C_m = C\right) = 1$. □

In the following theorem we give a sufficient condition for stationarity of functional ARMA$(1,1)$ process.

Theorem 4.6. *If C_0 holds, then there is a unique stationary causal solution to (4.6) (p=1). And the unique solution is given by*

$$X_n = \varepsilon_n + \sum_{j=1}^{\infty} \phi^{j-1}(\phi + \theta)(\varepsilon_{n-j}), \tag{4.10}$$

where ϕ^0 denotes the identity operator in H. Furthermore, the solution in (4.10) converges almost surely and in mean square, i.e

$$E\left\|X_n - \varepsilon_n - \sum_{j=1}^{m} \phi^{j-1}(\phi+\theta)(\varepsilon_{n-j})\right\|^2 \to 0, \quad as\ m \to \infty.$$

Proof. • First of all we prove the mean square convergence of (4.10).

We define $X_n^{(m)} := \varepsilon_n + \sum_{j=1}^{m} \phi^{j-1}(\phi+\theta)(\varepsilon_{n-j})$. Then for $\forall m' > m \geq 1$,

$$\begin{aligned}
E\|X_n^{(m')} - X_n^{(m)}\|^2 &= E\left\|\sum_{j=m}^{m'} \phi^{j-1}(\phi+\theta)(\varepsilon_{n-j})\right\|^2 \\
&= \sum_{j=m}^{m'}\sum_{k=m}^{m'} E\left\langle \phi^{j-1}(\phi+\theta)(\varepsilon_{n-j}), \phi^{k-1}(\phi+\theta)(\varepsilon_{n-k})\right\rangle \\
&= \sum_{j=m}^{m'} E\left\|\phi^{j-1}(\phi+\theta)(\varepsilon_{n-j})\right\|^2 \\
&\leq \left(\sum_{j=m}^{m'} \left\|\phi^{j-1}(\phi+\theta)\right\|_{\mathcal{L}}^2\right) E\|\varepsilon_0\|^2 \\
&\leq \left(\sum_{j=m}^{m'} \left\|\phi^{j-1}\right\|_{\mathcal{L}}^2\right) \|\phi+\theta\|_{\mathcal{L}}^2\ \sigma_\varepsilon^2.
\end{aligned}$$

By (4.7), we have

$$\left(\sum_{j=m}^{m'} \left\|\phi^{j-1}\right\|_{\mathcal{L}}^2\right) \|\phi+\theta\|_{\mathcal{L}}^2\ \sigma_\varepsilon^2 \leq \|\phi+\theta\|_{\mathcal{L}}^2\ \sigma_\varepsilon^2 a^2 \sum_{j=m}^{m'} b^{2(j-1)} \to 0, \quad \text{as } m, m' \to \infty.$$

By the Cauchy criterion, it follows that the series (4.10) converges in mean square.

- If we want to prove the almost sure convergence of (4.10), by Lemma 4.5, it suffices to verify that

$$\sum_{j=1}^{\infty} \left\|\phi^{j-1}(\phi+\theta)(\varepsilon_{n-j})\right\| < \infty \quad \text{a.s.}$$

Since

$$\begin{aligned}
E\left(\sum_{j=1}^{\infty} \left\|\phi^{j-1}(\phi+\theta)(\varepsilon_{n-j})\right\|\right)^2 &\leq \left(\sum_{j=1}^{\infty} \left\|\phi^{j-1}\right\|_{\mathcal{L}} \|\phi+\theta\|_{\mathcal{L}}\right)^2 E\|\varepsilon_0\|^2 \\
&\leq \sigma_\varepsilon^2 \|\phi+\theta\|_{\mathcal{L}}^2 \left(\sum_{j=1}^{\infty} \left\|\phi^{j-1}\right\|_{\mathcal{L}}^2\right)^2,
\end{aligned}$$

then by (4.7), we have

$$\sigma_\varepsilon^2 \|\phi+\theta\|_{\mathcal{L}}^2 \left(\sum_{j=1}^{\infty} \left\|\phi^{j-1}\right\|_{\mathcal{L}}^2\right)^2 = \sigma_\varepsilon^2 \|\phi+\theta\|_{\mathcal{L}}^2 \left(\sum_{j=1}^{\infty} ab^{j-1}\right)^2 = \sigma_\varepsilon^2 \|\phi+\theta\|_{\mathcal{L}}^2 \frac{a^2}{(1-b)^2} < \infty.$$

It implies

$$E\left(\sum_{j=1}^{\infty} \left\|\phi^{j-1}(\phi+\theta)(\varepsilon_{n-j})\right\|\right)^2 < \infty.$$

Thus we obtain the a.s. convergence of (4.10).

- (4.10) is clearly stationary, now we prove it is a solution of (4.6) (p=1). We plug (4.10) into (4.6), then we have

$$\begin{aligned} X_n - \phi(X_{n-1}) &= \varepsilon_n + \sum_{j=1}^{\infty} \phi^{j-1}(\phi+\theta)(\varepsilon_{n-j}) - \phi\left(\varepsilon_{n-1} + \sum_{j=1}^{\infty} \phi^{j-1}(\phi+\theta)(\varepsilon_{n-1-j})\right) \\ &= \varepsilon_n - \phi(\varepsilon_{n-1}) + \sum_{j=1}^{\infty} \phi^{j-1}(\phi+\theta)(\varepsilon_{n-j}) - \sum_{j=1}^{\infty} \phi^{j}(\phi+\theta)(\varepsilon_{n-1-j}) \\ &= \varepsilon_n - \phi(\varepsilon_{n-1}) + \sum_{j=1}^{\infty} \phi^{j-1}(\phi+\theta)(\varepsilon_{n-j}) - \sum_{j=2}^{\infty} \phi^{j-1}(\phi+\theta)(\varepsilon_{n-j}) \\ &= \varepsilon_n - \phi(\varepsilon_{n-1}) + (\phi+\theta)(\varepsilon_{n-1}) \\ &= \varepsilon_n + \theta(\varepsilon_{n-1}). \end{aligned}$$

It implies that (4.10) is a solution of equation (4.6).

- Finally we prove the uniqueness of the solution. Let (X'_n) be another stationary solution of (4.6)(p=1). A straightforward induction gives

$$X'_n = \phi^{k+1}(X'_{n-k-1}) + \varepsilon_n + \sum_{j=1}^{k} \phi^{j-1}(\phi+\theta)(\varepsilon_{t-j}) + \phi^k\theta(\varepsilon_{t-k-1}), \quad k \geq 1.$$

Therefore,

$$\begin{aligned} E\left\|X'_n - \varepsilon_n - \sum_{j=1}^{k} \phi^{j-1}(\phi+\theta)(\varepsilon_{t-j})\right\|^2 &= E\left\|\phi^{k+1}(X'_{n-k-1}) + \phi^k\theta(\varepsilon_{t-k-1})\right\|^2 \\ &\leq E\left\|\phi^{k+1}(X'_{n-k-1})\right\|^2 + E\left\|\phi^k\theta(\varepsilon_{t-k-1})\right\|^2 \\ &\leq \|\phi^{k+1}\|_{\mathcal{L}}^2 E\|X'_{n-k-1}\|^2 + \sigma_\varepsilon^2 \|\phi^k\|_{\mathcal{L}}^2 \|\theta\|_{\mathcal{L}}^2. \end{aligned}$$

Due to the stationarity, $E\|X'_{n-k-1}\|^2$ remains constant. By the condition C_1 in Lemma 4.4,

$$\|\phi^k\|_{\mathcal{L}}^2 < a^2 b^{2k} \to 0, \quad \text{as } k \to 0.$$

Thus

$$X'_n = \varepsilon_n + \sum_{j=1}^{\infty} \phi^{j-1}(\phi+\theta)(\varepsilon_{n-j}).$$

This proves the uniqueness.

□

Now we turn to the functional ARMA$(p,1)$ case. In the proof of Theorem 4.6, we did not use the fact that $(X_n, n \in \mathbb{Z})$ and $(\varepsilon_n, n \in \mathbb{Z})$ take values in $H = L^2([0,1])$, i.e. $L^2([0,1])$ can be replaced by a general separable Hilbert space. And the proof remains literally unchanged. Using this fact, we write the functional ARMA$(p,1)$ process with state space equation

$$\underbrace{\begin{pmatrix} X_n \\ X_{n-1} \\ \vdots \\ X_{n-p+1} \end{pmatrix}}_{:=Y_n} = \underbrace{\begin{pmatrix} \phi_1 & \cdots & \phi_{p-1} & \phi_p \\ I & & & 0 \\ & \ddots & & \vdots \\ & & I & 0 \end{pmatrix}}_{:=\phi^*} \underbrace{\begin{pmatrix} X_{n-1} \\ X_{n-2} \\ \vdots \\ X_{n-p} \end{pmatrix}}_{:=Y_{n-1}} + \underbrace{\begin{pmatrix} \varepsilon_n \\ 0 \\ \vdots \\ 0 \end{pmatrix}}_{:=\delta_n} + \underbrace{\begin{pmatrix} \theta & \cdots & 0 & 0 \\ & & & 0 \\ & \ddots & & \vdots \\ & & 0 & 0 \end{pmatrix}}_{\theta^*} \underbrace{\begin{pmatrix} \varepsilon_{n-1} \\ 0 \\ \vdots \\ 0 \end{pmatrix}}_{\delta_{n-1}}, \tag{4.11}$$

i.e.

$$Y_n = \phi^*(Y_{n-1}) + \delta_n + \theta^*(\delta_{n-1}), \tag{4.12}$$

where I and 0 in (4.11) denote the indentity operator and the zero operator respectively. Y_n and δ_n in (4.12) take values in the space $H_p := (L^2([0,1]))^p$. H_p is a Hilbert space equipped with the inner product

$$\langle x, y \rangle_p := \sum_{j=1}^{p} \langle x_j, y_j \rangle, \tag{4.13}$$

and the corresponding norm

$$\|x\|_p := \sqrt{\langle x, y \rangle_p} \quad . \tag{4.14}$$

ϕ^* and θ^* in (4.12) are bounded opeartors acting on H_p, where the norm (of the bounded operators) of ϕ^* and θ^* are defined as

$$\|\phi^*\|_{\mathcal{L}} := \sup \{\|\phi^*(x)\|_p : \ \|x\|_p \le 1\}. \tag{4.15}$$

$(\delta_n, n \in \mathbb{Z})$ is still i.i.d in H_p. The following theorem is immediate.

Theorem 4.7. *If there exists an integer j_0 such that $\|(\phi^*)^{j_0}\|_{\mathcal{L}} < 1$, where ϕ^* defined in (4.11) is a bounded operator acting on H_p. Then there is a unique stationary causal solution to the functional ARMA$(p,1)$ process (4.6). And the solution can be written as*

$$Y_n = \delta_n + \sum_{j=1}^{\infty} (\phi^*)^{j-1}(\phi^* + \theta^*)(\delta_{n-j}), \tag{4.16}$$

where Y_n, δ_n, ϕ^* and θ^* have the forms defined in (4.11).

4.3.2 The vector ARMA$(p, 1)$ structure

Now we will study the structure of the vector process truncated from the stationary functional ARMA$(p, 1)$ process. Similar to the last subsection, we start with the functional ARMA$(1, 1)$ process:

$$X_n = \phi(X_{n-1}) + \varepsilon_n + \theta(\varepsilon_{n-1}). \tag{4.17}$$

Furthermore, we need the constraints below and we call them **FARMA(1,1)**:

- $\phi, \theta \in \mathcal{S}$, i.e ϕ, θ are Hilbert-Schmidt operators and $\|\phi\|_{\mathcal{S}} < 1$.
- The H-valued sequence $(\varepsilon_n, n \in \mathbb{Z})$ is i.i.d with zero mean and satisfies $E\|\varepsilon_n\|^2 = \sigma_\varepsilon^2 < \infty$.

Since $\|\phi\|_{\mathcal{L}} \leq \|\phi\|_{\mathcal{S}} < 1$, then by Lemma 4.4 and Theorem 4.6, if the condition **FARMA(1,1)** holds, the functional ARMA(1,1) process in (4.17) has a unique stationary solution.

Suppose the functional ARMA(1,1) process $(X_n, n \in \mathbb{Z})$ in (4.17) satisfies **FARMA(1,1)**. Now we implement the first step in **Algorithm I**, i.e. to project observations $X_1, \cdots, X_n$ on $(\nu_l, l \in \mathbb{Z})$, then we have

$$\langle X_n, \nu_l \rangle = \langle \phi(X_{n-1}), \nu_l \rangle + \langle \varepsilon_n, \nu_l \rangle + \langle \theta(\varepsilon_{n-1}), \nu_l \rangle, \quad \forall l \in \mathbb{Z}. \tag{4.18}$$

For every l, we expand $\langle \phi(X_{n-1}), \nu_l \rangle$ and $\langle \theta(\varepsilon_{n-1}), \nu_l \rangle$ by

$$\begin{aligned} \langle \phi(X_{n-1}), \nu_l \rangle &= \left\langle \phi\left(\sum_{l'=1}^{\infty} \langle X_{n-1}, \nu_{l'} \rangle \nu_{l'} \right), \nu_l \right\rangle \\ &= \sum_{l'=1}^{\infty} \langle \phi(\nu_{l'}), \nu_l \rangle \langle X_{n-1}, \nu_{l'} \rangle, \end{aligned} \tag{4.19}$$

and

$$\begin{aligned} \langle \theta(\varepsilon_{n-1}), \nu_l \rangle &= \left\langle \theta\left(\sum_{l'=1}^{\infty} \langle \varepsilon_{n-1}, \nu_{l'} \rangle \nu_{l'} \right), \nu_l \right\rangle \\ &= \sum_{l'=1}^{\infty} \langle \theta(\nu_{l'}), \nu_l \rangle \langle \varepsilon_{n-1}, \nu_{l'} \rangle. \end{aligned} \tag{4.20}$$

In (4.19) and (4.20) we used Karhunen-Loéve representation. Then with CPV method (see Section 2.2.3), we choose d such that most of the entire variability can be explained by $\nu_1, \ldots, \nu_d$. With the help of (4.19) and (4.20), we can write (4.18) into the matrix form:

$$\left(\begin{array}{c} \langle X_n,\nu_1\rangle \\ \vdots \\ \langle X_n,\nu_d\rangle \\ \hline \langle X_n,\nu_{d+1}\rangle \\ \vdots \end{array}\right) = \left[\begin{array}{ccc|c} \langle \phi(\nu_1),\nu_1\rangle & \dots & \langle \phi(\nu_d),\nu_1\rangle & \langle \phi(\nu_{d+1}),\nu_1\rangle & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \langle \phi(\nu_1),\nu_d\rangle & \dots & \langle \phi(\nu_d),\nu_d\rangle & \langle \phi(\nu_{d+1}),\nu_d\rangle & \dots \\ \hline \langle \phi(\nu_1),\nu_{d+1}\rangle & \dots & \langle \phi(\nu_d),\nu_{d+1}\rangle & \langle \phi(\nu_{d+1}),\nu_{d+1}\rangle & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots \end{array}\right] \left(\begin{array}{c} \langle X_{n-1},\nu_1\rangle \\ \vdots \\ \langle X_{n-1},\nu_d\rangle \\ \hline \langle X_{n-1},\nu_{d+1}\rangle \\ \vdots \end{array}\right)$$

$$+ \left(\begin{array}{c} \langle \varepsilon_n,\nu_1\rangle \\ \vdots \\ \langle \varepsilon_n,\nu_d\rangle \\ \hline \langle \varepsilon_n,\nu_{d+1}\rangle \\ \vdots \end{array}\right) + \left[\begin{array}{ccc|c} \langle \theta(\nu_1),\nu_1\rangle & \dots & \langle \theta(\nu_d),\nu_1\rangle & \langle \theta(\nu_{d+1}),\nu_1\rangle & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \langle \theta(\nu_1),\nu_d\rangle & \dots & \langle \theta(\nu_d),\nu_d\rangle & \langle \theta(\nu_{d+1}),\nu_d\rangle & \dots \\ \hline \langle \theta(\nu_1),\nu_{d+1}\rangle & \dots & \langle \theta(\nu_d),\nu_{d+1}\rangle & \langle \theta(\nu_{d+1}),\nu_{d+1}\rangle & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots \end{array}\right] \left(\begin{array}{c} \langle \varepsilon_{n-1},\nu_1\rangle \\ \vdots \\ \langle \varepsilon_{n-1},\nu_d\rangle \\ \hline \langle \varepsilon_{n-1},\nu_{d+1}\rangle \\ \vdots \end{array}\right) \tag{4.21}$$

We simplify the notations in (4.21) to

$$\left(\begin{array}{c} \mathbf{X}_n \\ \hline \mathbf{X}_n^\infty \end{array}\right) = \left[\begin{array}{c|c} \mathbf{\Phi} & \mathbf{\Phi}^\infty \\ \hline \vdots & \vdots \end{array}\right] \left(\begin{array}{c} \mathbf{X}_{n-1} \\ \hline \mathbf{X}_{n-1}^\infty \end{array}\right) + \left(\begin{array}{c} \mathbf{E}_n \\ \hline \mathbf{E}_n^\infty \end{array}\right) + \left[\begin{array}{c|c} \mathbf{\Theta} & \mathbf{\Theta}^\infty \\ \hline \vdots & \vdots \end{array}\right] \left(\begin{array}{c} \mathbf{E}_{n-1} \\ \hline \mathbf{E}_{n-1}^\infty \end{array}\right), \tag{4.22}$$

where

$$\begin{aligned} \mathbf{X}_n &:= \left(\langle X_n,\nu_1\rangle,\dots,\langle X_n,\nu_d\rangle\right)^T, \\ \mathbf{E}_n &:= \left(\langle \varepsilon_n,\nu_1\rangle,\dots,\langle \varepsilon_n,\nu_d\rangle\right)^T, \\ \mathbf{X}_n^\infty &:= \left(\langle X_n,\nu_{d+1}\rangle,\dots\right)^T, \\ \mathbf{E}_n^\infty &:= \left(\langle \varepsilon_n,\nu_{d+1}\rangle,\dots\right)^T. \end{aligned} \tag{4.23}$$

$\mathbf{\Phi}$ and $\mathbf{\Theta}$ are $d\times d$ matrices with entries $\langle\phi(\nu_{l'}),\nu_l\rangle$ and $\langle\theta(\nu_{l'}),\nu_l\rangle$ in the l'th column and lth row respectively. $\mathbf{\Phi}^\infty$ and $\mathbf{\Theta}^\infty$ are $d\times\infty$ matrices with ll'th entries $\langle\phi(\nu_{l'+d}),\nu_l\rangle$ and $\langle\theta(\nu_{l'+d}),\nu_l\rangle$ respectively.

In practice, we can only focus on the d-dimensional vector observations $(\mathbf{X}_n, n\in\mathbb{Z})$. Then by (4.22), for each n,

$$\begin{aligned} \mathbf{X}_n &= \mathbf{\Phi}\mathbf{X}_{n-1} + \mathbf{E}_n + \mathbf{\Theta}\mathbf{E}_{n-1} + \underbrace{\mathbf{\Phi}^\infty\mathbf{X}_{n-1}^\infty + \mathbf{\Theta}^\infty\mathbf{E}_{n-1}^\infty}_{:=\mathbf{\Delta}_{n-1}} \\ &:= \mathbf{\Phi}\mathbf{X}_{n-1} + \mathbf{E}_n + \mathbf{\Theta}\mathbf{E}_{n-1} + \mathbf{\Delta}_{n-1}. \end{aligned} \tag{4.24}$$

By Theorem 2.5, the d-dimensional vector process $(\mathbf{E}_n, n\in\mathbb{Z})$ truncated from $(\varepsilon_n, n\in\mathbb{Z})$ is multivariate white noise. Note that, $\mathbf{\Delta}_{n-1}$ in (4.24) is a d-dimensional vector and its lth element $(\mathbf{\Delta}_{n-1})_l$ is

$$\begin{aligned} (\mathbf{\Delta}_{n-1})_l &= \left(\mathbf{\Phi}^\infty\mathbf{X}_{n-1}^\infty + \mathbf{\Theta}^\infty\mathbf{E}_{n-1}^\infty\right)_l \\ &= \sum_{l'=d+1}^{\infty} \langle\phi(\nu_{l'}),\nu_l\rangle\langle X_{n-1},\nu_{l'}\rangle + \sum_{l'=d+1}^{\infty} \langle\theta(\nu_{l'}),\nu_l\rangle\langle\varepsilon_{n-1},\nu_{l'}\rangle. \end{aligned} \tag{4.25}$$

Thus, $\mathbf{\Delta}_{n-1}$ depends in a complex way on X_{n-1}, i.e. the error term $\mathbf{\Delta}_{n-1}$ is not uncorrelated with past observations. Thus, rigorously speaking, the vector process $(\mathbf{X}_n, i \in \mathbb{Z})$ in (4.24) is **not** a vector ARMA$(1,1)$ process.

But if we can show that $\mathbf{\Delta}_{n-1}$ is a quite small error term, i.e. its impact can be negelected, then we can *approximately* treat $(\mathbf{X}_n, i \in \mathbb{Z})$ in (4.24) as a vector ARMA$(1,1)$ process. We will show this in Lemma 4.9. Before that we need the following technical lemma.

Lemma 4.8. *Suppose* $\mathbf{a}$ *and* $\mathbf{b}$ *are arbitrary d-dimensional vectors, i.e.* $\mathbf{a}, \mathbf{b} \in \mathbb{R}^d$. *Then the squared Euclidean norm* $\|\cdot\|_2^2$ *is convex, i.e.*

$$\|\alpha\mathbf{a} + (1-\alpha)\mathbf{b}\|_2^2 \leq \alpha\|\mathbf{a}\|_2^2 + (1-\alpha)\|\mathbf{b}\|_2^2, \quad \forall\alpha \in [0,1].$$

Proof. Since $f(x) := x^2$ is a convex function, by Jensen's inequality, we have

$$\begin{aligned}
\|\alpha\mathbf{a} + (1-\alpha)\mathbf{b}\|_2^2 &\leq \left(\|\alpha\mathbf{a}\|_2 + \|(1-\alpha)\mathbf{b}\|_2\right)^2 \\
&= f\left(\alpha\|\mathbf{a}\|_2 + (1-\alpha)\|\mathbf{b}\|_2\right) \\
&\leq \alpha f\left(\|\mathbf{a}\|_2\right) + (1-\alpha) f\left(\|\mathbf{b}\|_2\right) \\
&= \alpha\|\mathbf{a}\|_2^2 + (1-\alpha)\|\mathbf{b}\|_2^2
\end{aligned}$$

□

Remark 8. *Lemma (4.8) holds not only for the Euclidean norm* $\|\cdot\|_2$. *For an arbitrary norm* $\|\cdot\|$, *the convexity of* $\|\cdot\|^2$ *still holds.*

Now we will show that the $\mathbf{\Delta}_{n-1}$ defined in (4.24) is bounded and tends to 0 as $d \to \infty$.

Lemma 4.9. *Suppose* $\|\cdot\|_2$ *denotes the Euclidean norm of vector and the d-dimensional vector* $\mathbf{\Delta}_{n-1}$ *is defined in (4.24). Then* $E\|\mathbf{\Delta}_{n-1}\|_2^2$ *is bounded and tends to 0 as* $d \to \infty$.

Proof. By the convexity of $\|\cdot\|_2^2$ (Lemma 4.8),

$$\begin{aligned}
E\|\mathbf{\Delta}_{n-1}\|_2^2 &= E\|\mathbf{\Phi}^\infty \mathbf{X}_{n-1}^\infty + \mathbf{\Theta}^\infty \mathbf{E}_{n-1}^\infty\|_2^2 \\
&\leq 2\left(E\|\mathbf{\Phi}^\infty \mathbf{X}_{n-1}^\infty\|_2^2 + \|\mathbf{\Theta}^\infty \mathbf{E}_{n-1}^\infty\|_2^2\right). \qquad (4.26)
\end{aligned}$$

Now we calculate the two parts $E\|\mathbf{\Phi}^\infty \mathbf{X}_{n-1}^\infty\|_2^2$ and $E\|\mathbf{\Theta}^\infty \mathbf{E}_{n-1}^\infty\|_2^2$ respectively. By

(4.25), we have

$$
\begin{aligned}
E\|\mathbf{\Phi}^{\infty}\mathbf{X}_{n-1}^{\infty}\|_2^2 &= E\left[\sum_{l=1}^{d}\left(\sum_{l'=d+1}^{\infty}\langle\phi(\nu_{l'}),\nu_l\rangle\underbrace{\langle X_{n-1},\nu_{l'}\rangle}_{:=x_{n-1,l'}}\right)^2\right] \\
&:= E\left[\sum_{l=1}^{d}\left(\sum_{l'=d+1}^{\infty}\langle\phi(\nu_{l'})x_{n-1,l'},\nu_l\rangle\right)^2\right] \\
&\leq E\left[\sum_{l=1}^{\infty}\left(\sum_{l'=d+1}^{\infty}\langle\phi(\nu_{l'})x_{n-1,l'},\nu_l\rangle\right)^2\right] \\
&= E\left[\sum_{l=1}^{\infty}\left\langle\sum_{l'=d+1}^{\infty}\phi(\nu_{l'})x_{n-1,l'},\nu_l\right\rangle^2\right]. \qquad (4.27)
\end{aligned}
$$

By Parseval's identity, we continue with the computation of (4.27),

$$
\begin{aligned}
E\left[\sum_{l=1}^{\infty}\left\langle\sum_{l'=d+1}^{\infty}x_{n-1,l'}\phi(\nu_{l'}),\nu_l\right\rangle^2\right] &= E\left\|\sum_{l'=d+1}^{\infty}x_{n-1,l'}\phi(\nu_{l'})\right\|^2 \\
&= E\left\langle\sum_{l=d+1}^{\infty}x_{n-1,l}\phi(\nu_l),\sum_{l'=d+1}^{\infty}x_{n-1,l'}\phi(\nu_{l'})\right\rangle. \qquad (4.28)
\end{aligned}
$$

By Karhunen-Loéve theorem (see Theorem 2.3), the scores $(x_{n-1,l}, l\in\mathbb{Z})$ are uncorrelated. Thus (4.28) is equal to

$$
\begin{aligned}
E\left\langle\sum_{l=d+1}^{\infty}x_{n-1,l}\phi(\nu_l),\sum_{l'=d+1}^{\infty}x_{n-1,l'}\phi(\nu_{l'})\right\rangle &= E\left[\sum_{l'=d+1}^{\infty}x_{n-1,l'}^2\left\|\phi(\nu_{l'})\right\|^2\right] \\
&= \sum_{l'=d+1}^{\infty}E\left(x_{n-1,l'}\right)^2\left\|\phi(\nu_{l'})\right\|^2.
\end{aligned}
$$

Recall (2.19) in Theorem 2.3, i.e. $E\left(x_{n-1,l'}\right)^2=\lambda_{l'}$, then we have

$$
\begin{aligned}
\sum_{l'=d+1}^{\infty}E\left(x_{n-1,l'}\right)^2\left\|\phi(\nu_{l'})\right\|^2 &= \sum_{l'=d+1}^{\infty}\lambda_{l'}\left\|\phi(\nu_{l'})\right\|^2 \\
&\leq \lambda_1\sum_{l'=d+1}^{\infty}\left\|\phi(\nu_{l'})\right\|^2, \qquad (4.29)
\end{aligned}
$$

where λ_1 is the largest eigenvalue of the covariance operator C. Combining (4.27)-(4.29), we have

$$
E\|\mathbf{\Phi}^{\infty}\mathbf{X}_{n-1}^{\infty}\|_2^2 \leq \lambda_1\sum_{l'=d+1}^{\infty}\left\|\phi(\nu_{l'})\right\|^2. \qquad (4.30)
$$

Note that $\sum_{l'=d+1}^{\infty} \|\phi(\nu_{l'})\|^2 \leq \|\phi\|_{\mathcal{S}}^2 < \infty$, thus $E\|\mathbf{\Phi}^{\infty}\mathbf{X}_{n-1}^{\infty}\|_2^2$ is bounded and tends to 0 as $d \to \infty$. The proof to bound for $E\|\mathbf{\Theta}^{\infty}\mathbf{E}_{n-1}^{\infty}\|_2^2$ is exactly the same. So we just list the result in the followling,

$$\begin{aligned} E\|\mathbf{\Theta}^{\infty}\mathbf{E}_{n-1}^{\infty}\|_2^2 &\leq \sum_{l'=d+1}^{\infty} E\left(\varepsilon_{n-1,l'}\right)^2 \|\theta(\nu_{l'})\|^2 \\ &\leq \sigma_{\varepsilon}^2 \sum_{l'=d+1}^{\infty} \|\theta(\nu_{l'})\|^2 < \infty. \end{aligned} \tag{4.31}$$

Combining (4.26), (4.30) and (4.31), we get that $E\|\mathbf{\Delta}_{n-1}\|_2^2$ is bounded and tends to 0 as $d \to \infty$. □

Thus, even though the error term $\mathbf{\Delta}_{n-1}$ is correlated with the past observations, we can still treat the $(\mathbf{X}_n, i \in \mathbb{Z})$ in (4.24) as a vector ARMA$(1,1)$ process by neglecting $\mathbf{\Delta}_{n-1}$, i.e.

$$\mathbf{X}_n \approx \mathbf{\Phi}\mathbf{X}_{n-1} + \mathbf{E}_n + \mathbf{\Theta}\mathbf{E}_{n-1}. \tag{4.32}$$

Now the next question is, whether the vector ARMA$(p,1)$ process in (4.32) is stationary? The following theorem answers this question.

Theorem 4.10. *Consider the functional ARMA(1,1) process defined (4.17), and the condition* ***FARMA(1,1)*** *holds. Then for* $\forall d \geq 1$, *the vector process in* (4.32) *has a unique stationary causal solution.*

Proof. Since $\|\phi\|_{\mathcal{L}} \leq \|\phi\|_{\mathcal{S}} < 1$, then by Theorem 4.6, the functional ARMA$(1,1)$ process (4.17) has a unique stationary solution (4.10).

Let us first recall the form of the $d \times d$ matrix $\mathbf{\Phi}$ in the vector process (4.32) (see (4.21) and (4.22)), i.e.

$$\mathbf{\Phi} = \begin{pmatrix} \langle \phi(\nu_1), \nu_1 \rangle & \dots & \langle \phi(\nu_d), \nu_1 \rangle \\ \vdots & \vdots & \vdots \\ \langle \phi(\nu_1), \nu_d \rangle & \dots & \langle \phi(\nu_d), \nu_d \rangle \end{pmatrix}.$$

To prove the vector process (4.32) has a unique stationary causal solution, it suffices to prove that, for an arbitrary eigenvalue λ of $\mathbf{\Phi}$, $|\lambda| < 1$ (see Theorem 11.3.1 of Brockwell and Davis [1991]).

For an arbitrary eigenvalue λ of $\mathbf{\Phi}$, we suppose $\mathbf{a}$ is the corresponding eigenvector, i.e

$$\mathbf{\Phi}\mathbf{a} = \lambda\mathbf{a}, \quad \|\mathbf{a}\|_2 = 1.$$

Then we have

$$\begin{aligned}
|\lambda| = \|\lambda\mathbf{a}\|_2 &= \|\mathbf{\Phi a}\|_2 \\
&= \langle \mathbf{\Phi a}, \mathbf{\Phi a}\rangle^{\frac{1}{2}} = \left\langle \mathbf{a}, \mathbf{\Phi}^T\mathbf{\Phi a}\right\rangle^{\frac{1}{2}} \\
&\leq \|\mathbf{a}\|_2^{1/2}\|\mathbf{\Phi}^T\mathbf{\Phi a}\|_2^{1/2} = \|\mathbf{\Phi}^T\mathbf{\Phi a}\|_2^{1/2} \\
&\leq \left(\max_{\|\mathbf{b}\|_2=1} \|\mathbf{\Phi}^T\mathbf{\Phi b}\|_2\right)^{\frac{1}{2}} \\
&= \left(\max_{\|\mathbf{b}\|_2\neq 0} \frac{\|\mathbf{\Phi}^T\mathbf{\Phi b}\|_2}{\|\mathbf{b}\|_2}\right)^{\frac{1}{2}}. \qquad (4.33)
\end{aligned}$$

In linear algebra, the *spectral norm* $\|\cdot\|_{spectral}$ of a square matrix $\mathbf{\Phi}$ is defined as

$$\begin{aligned}
\|\mathbf{\Phi}\|_{spectral} &:= \big(\text{maximum eigenvalue of } \mathbf{\Phi}^T\mathbf{\Phi}\big) \\
&= \max_{\|\mathbf{b}\|_2\neq 0} \frac{\|\mathbf{\Phi b}\|_2}{\|\mathbf{b}\|_2}. \qquad (4.34)
\end{aligned}$$

Thus by (4.33) and (4.34), we have

$$|\ \lambda\ | \leq \|\mathbf{\Phi}^T\mathbf{\Phi}\|_{spectral}^{\frac{1}{2}} = \|\mathbf{\Phi}\|_{spectral}. \qquad (4.35)$$

Furthermore, by the property of the spectral matrix norm $\|\cdot\|_{spectral}$, we have

$$\|\mathbf{\Phi}\|_{spectral} \leq \|\mathbf{\Phi}\|_F, \qquad (4.36)$$

where $\|\cdot\|_F$ is the Frobenius matrix norm:

$$\begin{aligned}
\|\mathbf{\Phi}\|_F &:= \left(\sum_{i=1}^{d}\sum_{j=1}^{d} \langle \phi(v_i), \nu_j\rangle^2\right)^{1/2} \\
&\leq \left(\sum_{i=1}^{\infty}\sum_{j=1}^{\infty} \langle \phi(v_i), \nu_j\rangle^2\right)^{1/2} \\
&= \left(\sum_{i=1}^{\infty} \|\phi(\nu_i)\|^2\right)^{1/2} \\
&= \|\phi\|_{\mathcal{S}} < 1. \qquad (4.37)
\end{aligned}$$

Then combining (4.35)-(4.37), we have

$$|\lambda| \leq \|\mathbf{\Phi}\|_F \leq \|\phi\|_{\mathcal{S}} < 1. \qquad (4.38)$$

□

Up to now we can conclude that, if the condition **FARMA(1,1)** holds, then both the functional ARMA(1,1) process in (4.17) and its truncated vector process (by negelecting $\mathbf{\Delta}_{n-1}$) in (4.32) are stationary.

Now we will apply the state space equation in (4.11) again to extend the ARMA(1,1,) case to ARMA$(p,1)$ case. We rewrite the functional ARMA$(p,1)$ process as

$$Y_n = \phi^*(Y_{n-1}) + \delta_n + \theta^*(\delta_{n-1}), \tag{4.39}$$

where Y_n, ϕ^*, θ^* and δ_n are defined in (4.11) and take values in $H_p = (L^2([0,1]))^p$ (see (4.13) and (4.14)). Here we assume ϕ^* and θ^* in (4.39) are Hilbert-Schmidt opeartors acting on H_p, where the Hilbert-Schmidt norm of ϕ^* and θ^* are defined as

$$\|\phi^*\|_{\mathcal{S}} := \left(\sum_{i=1}^{\infty} \|\phi^*(\nu_i^*)\|_p^2\right)^{\frac{1}{2}}, \tag{4.40}$$

where $\|\cdot\|_p$ is the norm in H_p (see (4.14)) and $(\nu_i^*, i \in \mathbb{Z})$ is an arbitrary orthonormal basis in H_p.

Analogously, we summarize the condition under which both the functional ARMA$(p,1)$ process and its truncated vector process are stationary. We call the condition **FARMA**$(\mathbf{p},\mathbf{1})$:

- $\phi_1, \ldots, \phi_p, \theta$ defined in (4.6) are all Hilbert-Schmidt operators, and $\|\phi^*\|_{\mathcal{S}} < 1$ (see (4.40)).
- The H-valued sequence $(\varepsilon_n, n \in \mathbb{Z})$ in (4.6) is i.i.d with zero mean and satisfies $E\|\varepsilon_n\|^2 = \sigma_\varepsilon^2 < \infty$.

If the condition **FARMA**$(\mathbf{p},\mathbf{1})$ holds, both the functional ARMA$(p,1)$ process and its truncated vector process are stationary.

4.3.3 Some further notes on the vector process

We consider the vector process

$$\mathbf{X}_n = \sum_{i=1}^{p} \mathbf{\Phi}_i \mathbf{X}_{n-i} + \mathbf{E}_n + \mathbf{\Theta}\mathbf{E}_{n-1} + \mathbf{\Delta}_{n-1}, \tag{4.41}$$

which is truncated from the stationary functional ARMA$(p,1)$ process. By (4.25), we can derive form of $\mathbf{\Delta}_{n-1}$ in (4.41) directly,

$$\begin{aligned}(\mathbf{\Delta}_{n-1})_l &= \left(\sum_{i=1}^{p} \mathbf{\Phi}_i^{\infty} \mathbf{X}_{n-i}^{\infty} + \mathbf{\Theta}^{\infty} \mathbf{E}_{n-1}^{\infty} \right)_l \\ &= \sum_{i=1}^{p} \sum_{l'=d+1}^{\infty} \langle \phi_i(\nu_{l'}), \nu_l \rangle \langle X_{n-i}, \nu_{l'} \rangle + \sum_{l'=d+1}^{\infty} \langle \theta(\nu_{l'}), \nu_l \rangle \langle \varepsilon_{n-1}, \nu_{l'} \rangle, \end{aligned} \tag{4.42}$$

where the $(\mathbf{\Delta}_{n-1})_l$ in (4.42) is the lth element of $\mathbf{\Delta}_{n-1}$ in (4.41).

We have shown in Lemma 4.9 that $E\|\mathbf{\Delta}_{n-1}\|_2^2$ is bounded and tends to 0 as $d \to \infty$ (for ARMA(1,1)). Analogously this also holds for ARMA$(p,1)$.

So if we neglect the $\mathbf{\Delta}_{n-1}$ in (4.41), we can treat the vector process (4.41) as a vector ARMA$(p,1)$ process.

In this section we will show that, under some further constraints, even if we do not neglect $\mathbf{\Delta}_{n-1}$ in (4.41), the vector process (4.41) is still **rigorously** a vector ARMA$(p,1)$ process. Before that, we first introduce a lemma, which will be helpful to our proof later.

Lemma 4.11 (Lemma 1, Lütkepohl [1984])**.** *Suppose the k-dimensional process $(\mathbf{M}_n, n \in \mathbb{Z})$ is k-dimensional MA(q). Let $\mathbf{F} \neq \mathbf{0}$ be a real $l \times k$ matrix. Then the l-dimensional process $(\mathbf{L}_n, n \in \mathbb{Z})$, $\mathbf{L}_n := \mathbf{F}\mathbf{M}_n$, is l-dimensional MA(q^*), where $q^* \leq q$.*

In the following we provide two conditions under which the vector process $(\mathbf{X}_n, n \in \mathbb{Z})$ in (4.41) **rigorously** follows vector ARMA$(p,1)$ structure.

- **Condition I:** *For $i = 1, \ldots, p$, $\|\phi_i\|_{\mathcal{S}}^2 = \sum_{j=1}^{d} \|\phi_i(\nu_j)\|^2$ and $\|\theta\|_{\mathcal{S}}^2 = \sum_{j=1}^{d} \|\theta(\nu_j)\|^2$.*

In this case,

$$\begin{aligned}\|\phi_i\|_{\mathcal{S}}^2 = \sum_{j=1}^{d} \|\phi_i(\nu_j)\|^2 \implies \sum_{j=d+1}^{\infty} \|\phi_i(\nu_j)\|^2 = 0 \\ \implies \phi_i(\nu_j) = 0, \quad \forall j \geq d+1,\end{aligned}$$

and

$$\|\theta\|_{\mathcal{S}}^2 = \sum_{j=1}^{d} \|\theta(\nu_j)\|^2 \implies \theta(\nu_j) = 0, \quad \forall j \geq d+1.$$

From (4.42) we know that, $\mathbf{\Delta}_{n-1} = \mathbf{0}$. Thus, if **Condition I** holds, the vector process (4.41) is **rigorously** a vector ARMA$(p,1)$ process.

- **Condition II:** *For* $i = 1, \ldots, p$, $\|\phi_i\|_{\mathcal{S}}^2 = \sum_{j=1}^{d} \|\phi_i(\nu_j)\|^2$.

Under **Condition II**, the first part of $\mathbf{\Delta}_{n-1}$, $\sum_{i=1}^{p} \mathbf{\Phi}_i^{\infty} \mathbf{X}_{n-i}^{\infty}$ is $\mathbf{0}$, which is the as **Condition I**. We denote the second part of $\mathbf{\Delta}_{n-1}$ by $\mathbf{K}_{n-1}$, i.e.

$$\mathbf{K}_{n-1} := \mathbf{\Theta}^{\infty} \mathbf{E}_{n-1}^{\infty}. \tag{4.43}$$

If we can show that

$$\mathbf{E}_n + \mathbf{\Theta}\mathbf{E}_{n-1} + \mathbf{K}_{n-1} \tag{4.44}$$

is a vector MA(1) process, then the vector process

$$\mathbf{X}_n = \sum_{i=1}^{p} \mathbf{\Phi}_i \mathbf{X}_{n-i} + \mathbf{E}_n + \mathbf{\Theta}\mathbf{E}_{n-1} + \mathbf{K}_{n-1}. \tag{4.45}$$

is really a vector ARMA$(p, 1)$ process. Now we begin the proof.

Proof. We will organize our proof into three steps. In the first step we will show $(\mathbf{K}_n, n \in \mathbb{Z})$ defined in (4.43) is white noise. In the second step we will show the $2d$-dimensional vector process $\left(\begin{pmatrix} \mathbf{E}_n \\ \mathbf{K}_n \end{pmatrix}, n \in \mathbb{Z} \right)$ is white noise. Finally we will use Lemma 4.11 to prove that $(\mathbf{E}_n + \mathbf{\Theta}\mathbf{E}_{n-1} + \mathbf{K}_{n-1}, n \in \mathbb{Z})$ defined in (4.44) is a vector MA(1)process.

(1) For $\forall l \in \{1, \ldots, d\}$, by the form of the lth element of $\mathbf{K}_n$ (see (4.42) and (4.43)), we have

$$E[(\mathbf{K}_n)_l] = \sum_{l'=d+1}^{\infty} \langle \theta(\nu_{l'}), \nu_l \rangle E[\langle \varepsilon_n, \nu_{l'} \rangle] = 0.$$

Thus

$$E[\mathbf{K}_n] = \mathbf{0}, \quad \forall n \in \mathbb{Z}. \tag{4.46}$$

For $n \neq m$, we compute $E\left[\mathbf{K}_n \mathbf{K}_m^T\right]$ and show it is $\mathbf{0}$. For $\forall k, l \in \{1, \ldots, d\}$,

$$\begin{aligned} E\left[\left(\mathbf{K}_n \mathbf{K}_m^T\right)_{k,l}\right] &= E\left[\left(\sum_{i=d+1}^{\infty} \langle \theta(\nu_i), \nu_k \rangle \langle \varepsilon_n, \nu_i \rangle\right)\left(\sum_{j=d+1}^{\infty} \langle \theta(\nu_j), \nu_l \rangle \langle \varepsilon_m, \nu_j \rangle\right)\right] \\ &= E\left[\sum_{i,j=d+1}^{\infty} \langle \theta(\nu_i), \nu_k \rangle \langle \theta(\nu_j), \nu_l \rangle \langle \varepsilon_n, \nu_i \rangle \langle \varepsilon_m, \nu_j \rangle\right]. \end{aligned} \tag{4.47}$$

By the definition of H-white noise in Definition 2.4, i.e.

$$E[\langle \varepsilon_n, x \rangle \langle \varepsilon_m, y \rangle] = 0, \quad \forall x, y \in H, \quad n \neq m,$$

then (4.47) is equal to

$$\sum_{i,j=d+1}^{\infty} \langle \theta(\nu_i), \nu_k \rangle \langle \theta(\nu_j), \nu_l \rangle E[\langle \varepsilon_{n-1}, \nu_i \rangle \langle \varepsilon_{m-1}, \nu_j \rangle] = 0.$$

Thus

$$E\left[\mathbf{K}_n\mathbf{K}_m^T\right] = \mathbf{0}, \quad n \neq m. \tag{4.48}$$

Now we show $E\left[\mathbf{K}_n\mathbf{K}_n^T\right]$ does not depend on n.

$$\begin{aligned}
E\left[\left(\mathbf{K}_n\mathbf{K}_n^T\right)_{k,l}\right] &= E\left[\left(\sum_{i=d+1}^{\infty}\langle\theta\left(\nu_i\right),\nu_k\rangle\langle\varepsilon_n,\nu_i\rangle\right)\left(\sum_{j=d+1}^{\infty}\langle\theta\left(\nu_j\right),\nu_l\rangle\langle\varepsilon_n,\nu_j\rangle\right)\right] \\
&= E\left[\sum_{i,j=d+1}^{\infty}\langle\theta\left(\nu_i\right),\nu_k\rangle\langle\theta\left(\nu_j\right),\nu_l\rangle\langle\varepsilon_n,\nu_i\rangle\langle\varepsilon_n,\nu_j\rangle\right] \\
&= E\left[\sum_{i,j=d+1}^{\infty}\langle\theta\left(\nu_i\right),\nu_k\rangle\langle\theta\left(\nu_j\right),\nu_l\rangle\langle C_{\varepsilon_n}(\nu_i),\nu_j\rangle\right] \quad \text{(see (2.23))} \\
&= E\left[\sum_{i,j=d+1}^{\infty}\langle\theta\left(\nu_i\right),\nu_k\rangle\langle\theta\left(\nu_j\right),\nu_l\rangle\langle C_{\varepsilon}(\nu_i),\nu_j\rangle\right]. \qquad (4.49)
\end{aligned}$$

Thus $E\left[\mathbf{K}_n\mathbf{K}_n^T\right]$ does not depend on n. Combining (4.46), (4.48) and (4.49), we get that $(\mathbf{K}_n, n \in \mathbb{Z})$ defined in (4.43) is d-dimensional white noise.

(2) Now we will show the $2d$-dimensional vector process

$$\left(\begin{pmatrix}\mathbf{E}_n\\ \mathbf{K}_n\end{pmatrix}, n \in \mathbb{Z}\right) \tag{4.50}$$

is white noise. Since both $(\mathbf{E}_n, n \in \mathbb{Z})$ and $(\mathbf{K}_n, n \in \mathbb{Z})$ are white noise, it suffices to show $E\left[\mathbf{E}_n\mathbf{K}_m^T\right] = 0$ for $n \neq m$ and $E\left[\mathbf{E}_n\mathbf{K}_n^T\right]$ does not depend on n.

For $n \neq m$, $\forall k, l \in \{1, \ldots, d\}$,

$$\begin{aligned}
E\left[\left(\mathbf{E}_n\mathbf{K}_m^T\right)_{k,l}\right] &= E\left[\langle\varepsilon_n,\nu_k\rangle\left(\sum_{j=d+1}^{\infty}\langle\theta\left(\nu_j\right),\nu_l\rangle\langle\varepsilon_m,\nu_j\rangle\right)\right] \\
&= E\left[\sum_{j=d+1}^{\infty}\langle\theta\left(\nu_j\right),\nu_l\rangle\langle\varepsilon_n,\nu_k\rangle\langle\varepsilon_m,\nu_j\rangle\right] = 0. \qquad (4.51)
\end{aligned}$$

And

$$\begin{aligned}
E\left[\left(\mathbf{E}_n\mathbf{K}_n^T\right)_{k,l}\right] &= E\left[\langle\varepsilon_n,\nu_k\rangle\left(\sum_{j=d+1}^{\infty}\langle\theta\left(\nu_j\right),\nu_l\rangle\langle\varepsilon_n,\nu_j\rangle\right)\right] \\
&= E\left[\sum_{j=d+1}^{\infty}\langle\theta\left(\nu_j\right),\nu_l\rangle\langle\varepsilon_n,\nu_k\rangle\langle\varepsilon_n,\nu_j\rangle\right] = 0. \qquad (4.52)
\end{aligned}$$

The last step in (4.52) holds due to the orthogonality of the two spaces spanned by $\{\nu_1, \ldots, \nu_d\}$ and $\{\nu_{d+1}, \ldots\}$ (note that $j \geq d+1$ and $1 \leq k \leq d$ in (4.52)).

Anyway, $E\left[\mathbf{E}_n\mathbf{K}_n^T\right]$ does not depend on n. Combining (4.51) and (4.52), the $2d$-dimensional vector process (4.50) is white noise.

(3) We define a $2d$-dimensional vector process $(\mathbf{Z}_n, n \in \mathbb{Z})$, where

$$\mathbf{Z}_n := \begin{pmatrix} \mathbf{Z}_{n,1} \\ \mathbf{Z}_{n,2} \end{pmatrix} := \begin{pmatrix} \mathbf{I}_d & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \begin{pmatrix} \mathbf{E}_n \\ \mathbf{K}_n \end{pmatrix} + \begin{pmatrix} \mathbf{\Theta} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \begin{pmatrix} \mathbf{E}_{n-1} \\ \mathbf{K}_{n-1} \end{pmatrix}. \tag{4.53}$$

In (4.53), $\mathbf{Z}_{n,1}$ and $\mathbf{Z}_{n,1}$ are all d-dimensional, $\mathbf{I}_d$ is $d \times d$ identity matrix and $\mathbf{0}$ is $d \times d$ zero matrix. We have proved that (4.50) is $2d$-dimensional white noise, thus $(\mathbf{Z}_n, n \in \mathbb{Z})$ in (4.53) is a $2d$-dimensional MA(1) process. Then we follow the Lemma 4.11 to define the non-zero $d \times 2d$ matrix

$$\mathbf{F} := (\mathbf{I}_d, \mathbf{I}_d). \tag{4.54}$$

Since $(\mathbf{Z}_n, n \in \mathbb{Z})$ in (4.53) is $2d$-dimensional MA(1) process, then by Lemma 4.11,

$$(\mathbf{F}\mathbf{Z}_n, n \in \mathbb{Z}) \tag{4.55}$$

is a d-dimensional MA(1) or MA(0) (i.e. white noise) process. Now we compute $\mathbf{F}\mathbf{Z}_n$ in (4.55) by

$$\begin{aligned} \mathbf{F}\mathbf{Z}_n &= (\mathbf{I}_d, \mathbf{I}_d) \begin{pmatrix} \mathbf{I}_d & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \begin{pmatrix} \mathbf{E}_n \\ \mathbf{K}_n \end{pmatrix} + (\mathbf{I}_d, \mathbf{I}_d) \begin{pmatrix} \mathbf{\Theta} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \begin{pmatrix} \mathbf{E}_{n-1} \\ \mathbf{K}_{n-1} \end{pmatrix} \\ &= \mathbf{E}_n + \mathbf{\Theta}\mathbf{E}_{n-1} + \mathbf{K}_{n-1}. \end{aligned}$$

Thus, $(\mathbf{E}_n + \mathbf{\Theta}\mathbf{E}_{n-1} + \mathbf{K}_{n-1}, n \in \mathbb{Z})$ (see (4.44)) is a d-dimensional MA(1) or MA(0) (i.e. white noise) process. We need to prove that (4.44) is not a white noise, and it suffices to show

$$E\left[\left(\mathbf{E}_{n+1} + \mathbf{\Theta}\mathbf{E}_n + \mathbf{K}_n \right) \left(\mathbf{E}_n + \mathbf{\Theta}\mathbf{E}_{n-1} + \mathbf{K}_{n-1} \right)^T \right] \neq \mathbf{0}. \tag{4.56}$$

Since $(\mathbf{E}_n, n \in \mathbb{Z})$ and $(\mathbf{K}_n, n \in \mathbb{Z})$ are white noise, $E\left[\mathbf{E}_n \mathbf{K}_m^T\right] = \mathbf{0}$ for $n \neq m$ (see (4.51)) and $E\left[\mathbf{E}_n \mathbf{K}_n^T\right] = \mathbf{0}$ (see (4.52)), then we have

$$\begin{aligned} E\left[\left(\mathbf{E}_{n+1} + \mathbf{\Theta}\mathbf{E}_n + \mathbf{K}_n \right) \left(\mathbf{E}_n + \mathbf{\Theta}\mathbf{E}_{n-1} + \mathbf{K}_{n-1} \right)^T \right] &= E\left[\mathbf{\Theta}\mathbf{E}_n \mathbf{E}_n^T \right] \\ &= \mathbf{\Theta} E\left[\mathbf{E}_n \mathbf{E}_n^T \right]. \end{aligned} \tag{4.57}$$

By the definition of $\mathbf{\Theta}$ and $\mathbf{E}_n$ (see (4.21)-(4.23)), (4.57) is equal to

$$\begin{aligned} &\begin{pmatrix} \langle \theta(\nu_1), \nu_1 \rangle & \dots & \langle \theta(\nu_d), \nu_1 \rangle \\ \vdots & \vdots & \vdots \\ \langle \theta(\nu_1), \nu_d \rangle & \dots & \langle \theta(\nu_d), \nu_d \rangle \end{pmatrix} \begin{pmatrix} E\left[\langle \varepsilon_n, \nu_1 \rangle^2 \right] & 0 & \cdots & 0 \\ 0 & E\left[\langle \varepsilon_n, \nu_2 \rangle^2 \right] & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & E\left[\langle \varepsilon_n, \nu_d \rangle^2 \right] \end{pmatrix} \\ &= \begin{pmatrix} \lambda_1 \langle \theta(\nu_1), \nu_1 \rangle & 0 & \cdots & 0 \\ 0 & \lambda_2 \langle \theta(\nu_2), \nu_2 \rangle & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_d \langle \theta(\nu_d), \nu_d \rangle \end{pmatrix} \neq \mathbf{0}. \end{aligned} \tag{4.58}$$

Thus the d-dimensional process $(\mathbf{E}_n + \mathbf{\Theta}\mathbf{E}_{n-1} + \mathbf{K}_{n-1}, n \in \mathbb{Z})$ in (4.44) is not white noise, so it is MA(1). It implies that $(\mathbf{X}_n, n \in \mathbb{Z})$ in (4.41), which is truncated from the functional ARMA$(p, 1)$ process, is **rigorously** a vector ARMA$(p, 1)$ process.

□

4.4 Relation between the functional and vector best linear predictor

Suppose $(X_n, n \in \mathbb{Z})$ is a functional ARMA$(p,1)$ process defined in (4.6) and the condition **FARMA**$(\mathbf{p},\mathbf{1})$ holds. Then both the functional process $(X_n, n \in \mathbb{Z})$ and the vector process $(\mathbf{X}_n, n \in \mathbb{Z})$ are stationary.

Suppose $\hat{X}_{n+1}^{\mathcal{G}}$ is the functional best linear predictor of X_{n+1}, and $\hat{\mathbf{X}}_{n+1}$ is the d-dimensional vector best linear predictor of $\mathbf{X}_{n+1}$ (see Section 4.1).

If we project $\hat{X}_{n+1}^{\mathcal{G}}$ on $\nu_1, \dots \nu_d$, then the projection $\hat{\mathbf{X}}_{n+1}^{\mathcal{G}}$, where

$$\hat{\mathbf{X}}_{n+1}^{\mathcal{G}} := \left(\left\langle \hat{X}_{n+1}^{\mathcal{G}}, \nu_1 \right\rangle, \dots, \left\langle \hat{X}_{n+1}^{\mathcal{G}}, \nu_d \right\rangle \right)^T, \tag{4.59}$$

is d-dimensional as well. Is there a relation between $\hat{\mathbf{X}}_{n+1}^{\mathcal{G}}$ and $\hat{\mathbf{X}}_{n+1}$? In other words, is the distance between these two vectors small enough? Before we list the main result, let us first get familiar with some notions.

- The vector best linear predictor $\hat{\mathbf{X}}_{n+1}$ is defined as the orthogonal projection of $\mathbf{X}_{n+1}$ on the "matrix linear span" of the d-dimensional observations $\mathbf{X}_1, \dots, \mathbf{X}_n$. We denote the "matrix linear span" by $\mathbf{M}_1$, where

$$\mathbf{M}_1 = \left\{ \sum_{i=1}^{n} \mathbf{A}_{ni} \mathbf{X}_i, \ \mathbf{A}_{ni} \text{ is an arbitrary real } d \times d \text{ matrix}, \ i = 1, \dots, n \right\}. \tag{4.60}$$

  $\hat{\mathbf{X}}_{n+1}$ satiesfies

$$E\left[\left(\mathbf{X}_{n+1} - \hat{\mathbf{X}}_{n+1} \right) \mathbf{y}_1^T \right] = \mathbf{0}, \quad \text{for all } \mathbf{y}_1 \in \mathbf{M}_1. \tag{4.61}$$

  $\mathbf{M}_1$ is a linear subspace of $\mathbb{R}^d$.

- Suppose $\mathcal{G}$ is the closure of $\mathcal{G}'$, where

$$\mathcal{G}' = \left\{ \sum_{i=1}^{n} s_{ni}(X_i), \ \forall s_{ni} \in \mathcal{S}, X_i \in F_n = \{X_1, \dots, X_n\}, \ i = 1, \dots n, \ n \geq 1 \right\}. \tag{4.62}$$

  Note that in (4.62) we put a further constraint on $\mathcal{G}$, i.e. we assume $s_{ni}, \ i = 1, \dots, n$ are Hilbert-Schmidt operators (compare it with (4.1)). $\hat{X}_{n+1}^{\mathcal{G}}$ is defined as the orthogonal projection of X_{n+1} on $\mathcal{G}$ and satisfies

$$E\left\langle X_{n+1} - \hat{X}_{n+1}^{\mathcal{G}}, y \right\rangle = 0, \quad \forall y \in \mathcal{G}. \tag{4.63}$$

  $\hat{X}_{n+1}^{\mathcal{G}}$ has the form of

$$\hat{X}_{n+1}^{\mathcal{G}} = \sum_{i=1}^{n} g_{ni}(X_i), \quad g_{ni} \in \mathcal{S}, \quad i = 1, \dots, n. \tag{4.64}$$

- Now let us study the projection of $\hat{X}_{n+1}^{\mathcal{G}}$ on $\nu_1, \ldots, \nu_d$. For $\forall l \in \{1, \ldots, d\}$,

$$\begin{aligned}\left\langle \hat{X}_{n+1}^{\mathcal{G}}, \nu_l \right\rangle &= \left\langle \sum_{i=1}^{n} g_{ni}(X_i), \nu_l \right\rangle \\ &= \left\langle \sum_{i=1}^{n} \sum_{l'=1}^{\infty} \langle X_i, \nu_{l'} \rangle \, g_{ni}\left(\nu_{l'}\right), \nu_l \right\rangle \\ &= \sum_{i=1}^{n} \sum_{l'=1}^{\infty} \langle X_i, \nu_{l'} \rangle \left\langle g_{ni}\left(\nu_{l'}\right), \nu_l \right\rangle . \end{aligned} \tag{4.65}$$

  By (4.65), we can rewrite $\hat{\mathbf{X}}_{n+1}^{\mathcal{G}}$ defined in (4.59) as

$$\hat{\mathbf{X}}_{n+1}^{\mathcal{G}} = \sum_{i=1}^{n} \left[\begin{array}{ccc|ccc} \langle g_{ni}(\nu_1), \nu_1 \rangle & \ldots & \langle g_{ni}(\nu_d), \nu_1 \rangle & \langle g_{ni}(\nu_{d+1}), \nu_1 \rangle & \ldots \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \langle g_{ni}(\nu_1), \nu_d \rangle & \ldots & \langle g_{ni}(\nu_d), \nu_d \rangle & \langle g_{ni}(\nu_{d+1}), \nu_d \rangle & \ldots \end{array}\right] \left(\begin{array}{c} \langle X_i, \nu_1 \rangle \\ \vdots \\ \langle X_i, \nu_d \rangle \\ \hline \langle X_i, \nu_{d+1} \rangle \\ \vdots \end{array}\right)$$

$$:= \sum_{i=1}^{n} \mathbf{G}_{ni} \mathbf{X}_i + \sum_{i=1}^{d} \mathbf{G}_{ni}^{\infty} \mathbf{X}_i^{\infty}, \tag{4.66}$$

  where $\mathbf{G}_{ni}$ is a $d \times d$ matrix with ll'th entry (lth row and l'th column) $\langle g_{ni}(\nu_{l'}), \nu_l \rangle$ and $\mathbf{G}_{ni}^{\infty}$ is a $d \times \infty$ matrix with ll'th entry $\langle g_{ni}(\nu_{d+l'}), \nu_l \rangle$. Similar to (4.37), the Frobenius matrix norm of the $d \times \infty$ matrix $\mathbf{G}_{ni}^{\infty}$ is bounded by

$$\|\mathbf{G}_{ni}^{\infty}\|_F \leq \|g_{ni}\|_{\mathcal{S}} < \infty, \quad i = 1, \ldots, n.$$

- For $\forall y \in \mathcal{G}$, there exist $s_{n1}, \ldots, s_{n,n} \in \mathcal{S}$ such that

$$y = \sum_{i=1}^{n} s_{ni}(X_i). \tag{4.67}$$

  Similar to (4.65), we project $y \in \mathcal{G}$ on $\nu_1, \ldots, \nu_d$, then we have

$$\begin{aligned}\mathbf{y} &:= \left(\langle y, \nu_1 \rangle \ldots \langle y, \nu_d \rangle\right)^T \\ &= \left(\left\langle \sum_{i=1}^{n} s_{ni}(X_i), \nu_1 \right\rangle \ldots \left\langle \sum_{i=1}^{n} s_{ni}(X_i), \nu_d \right\rangle \right)^T \\ &:= \sum_{i=1}^{n} \mathbf{S}_{ni} \mathbf{X}_i + \sum_{i=1}^{n} \mathbf{S}_{ni}^{\infty} \mathbf{X}_i^{\infty}. \end{aligned} \tag{4.68}$$

  The $d \times d$ matrix $\mathbf{S}_{ni}$ and the $d \times \infty$ matrix $\mathbf{S}_{ni}^{\infty}$ in (4.67) are defined in the same way as $\mathbf{G}_{ni}$ and $\mathbf{G}_{ni}^{\infty}$ in (4.66). We denote the set of the projection of $\forall y \in \mathcal{G}$ on $\{\nu_1, \ldots, \nu_d\}$, by $\mathbf{M}$, where

$$\begin{aligned}\mathbf{M} :&= \left\{ \mathbf{y} \;\middle|\; \mathbf{y} = \left(\langle y, \nu_1 \rangle, \ldots, \langle y, \nu_d \rangle\right)^T, \ \forall y \in \mathcal{G} \right\} \\ &= \left\{ \mathbf{y} \;\middle|\; \mathbf{y} = \sum_{i=1}^{n} \mathbf{S}_{ni} \mathbf{X}_i + \sum_{i=1}^{n} \mathbf{S}_{ni}^{\infty} \mathbf{X}_i^{\infty}, \ \forall s_{ni} \in \mathcal{S}, \ i = 1, \ldots, n \right\}. \end{aligned} \tag{4.69}$$

- Now we want to show $\mathbf{M}_1 \subseteq \mathbf{M}$, i.e. if $\mathbf{y}_1 \in \mathbf{M}_1$, then $\mathbf{y}_1 \in \mathbf{M}$.

  For $\forall \mathbf{y}_1 \in \mathbf{M}_1$, by the definition of $\mathbf{M}_1$ in (4.60), there exist $d \times d$ matrices $\mathbf{A}_{n1}, \ldots, \mathbf{A}_{nn}$ such that

$$\mathbf{y}_1 = \sum_{i=1}^{n} \mathbf{A}_{ni} \mathbf{X}_i. \tag{4.70}$$

  For each $i \in \{1, \ldots, n\}$, we denote the jkth entry (jth row and kth column) of $\mathbf{A}_{ni}$ in (4.70) by $A_{ni;jk}$.

  Now we construct n operators $s_{ni}(\cdot), \ldots, s_{nn}(\cdot)$ in the following way. For each $i \in \{1, \ldots, n\}$,

$$s_{ni}(\nu_l) := \begin{cases} \sum_{m=1}^{d} A_{ni;ml} \nu_m & 1 \le l \le d, \\ 0 & l > d. \end{cases} \tag{4.71}$$

  Then for $\forall j, k \in \{1, \ldots, d\}$,

$$\begin{aligned} \langle s_{ni}(\nu_k), \nu_j \rangle &= \left\langle \sum_{m=1}^{d} A_{ni;mk} \nu_m, \nu_j \right\rangle \\ &= \langle A_{ni;jk} \nu_j, \nu_j \rangle \\ &= A_{ni;jk}, \end{aligned} \tag{4.72}$$

  which implies that the $s_{n1}(\cdot), \ldots, s_{nn}(\cdot)$ constructed in (4.71) are such that

$$\mathbf{S}_{ni} = \mathbf{A}_{ni}, \quad i = 1, \ldots, n, \tag{4.73}$$

  where the $d \times d$ matrix $\mathbf{S}_{ni}$ is defined in (4.68). We still need to prove, the $s_{n1}(\cdot), \ldots, s_{nn}(\cdot)$ in (4.71) are Hilbert-Schmidt operators and such that $\mathbf{S}_{ni}^{\infty} = \mathbf{0},\ i = 1, \ldots, n$.

  For $\forall l \in \{1, \ldots, d\}$ and $\forall l' \ge 1$, by (4.71), we have

$$\langle s_{ni}(\nu_{d+l'}), \nu_l \rangle = \langle 0, \nu_j \rangle = 0, \tag{4.74}$$

  which implies $\mathbf{S}_{ni}^{\infty} = \mathbf{0},\ i = 1, \ldots, n$. Furthermore,

$$\begin{aligned} \sum_{l=1}^{\infty} \|s_{ni}(\nu_l)\|^2 &= \sum_{l=1}^{d} \|s_{ni}(\nu_l)\|^2 \\ &= \sum_{l=1}^{d} \| \sum_{m=1}^{d} A_{ni;ml} \nu_m \|^2 \\ &= \sum_{l=1}^{d} \sum_{m=1}^{d} A_{ni;ml}^2 < \infty, \quad i = 1, \ldots, n, \end{aligned} \tag{4.75}$$

  which implies the $s_{n1}(\cdot), \ldots, s_{nn}(\cdot)$ in (4.71) are Hilbert-Schmidt operators. Now we summarize the process above. For $\forall \mathbf{y}_1 \in \mathbf{M}_1$, which can be represented by

$$\mathbf{y}_1 = \sum_{i=1}^{n} \mathbf{A}_{ni} \mathbf{X}_i,$$

we can always construct n Hilbert-Schmidt operators $s_{n1}, \ldots, s_{nn}$ in the way of (4.71) such that (by (4.68), (4.73) and (4.74))

$$\underbrace{\left(\left\langle \sum_{i=1}^{n} s_{ni}(X_i), \nu_1 \right\rangle \ldots \left\langle \sum_{i=1}^{n} s_{ni}(X_i), \nu_d \right\rangle\right)^T}_{\in \mathbf{M}} = \sum_{i=1}^{n} \mathbf{S}_{ni}\mathbf{X}_i + \sum_{i=1}^{n} \mathbf{S}_{ni}^{\infty}\mathbf{X}_i^{\infty}$$

$$= \underbrace{\sum_{i=1}^{n} \mathbf{A}_{ni}\mathbf{X}_i}_{\in \mathbf{M}_1}.$$

Thus we prove $\mathbf{M}_1 \subseteq \mathbf{M}$.

- The mapping

$$\begin{aligned} f : \mathcal{G} &\to \mathbf{M} \\ y &\mapsto \left(\langle y, \nu_1\rangle, \ldots, \langle y, \nu_d\rangle\right)^T \end{aligned} \tag{4.76}$$

is surjective, it implies that for $\forall \mathbf{y} \in \mathbf{M}$, we can always find a $y \in \mathcal{G}$ such that

$$\left(\langle y, \nu_1\rangle, \ldots, \langle y, \nu_d\rangle\right)^T = \mathbf{y}. \tag{4.77}$$

Since $\hat{\mathbf{X}}_{n+1} \in \mathbf{M}_1$, then $\hat{\mathbf{X}}_{n+1} \in \mathbf{M}$. By the surjectivity, the functional form $\hat{X}_{n+1}$ of $\hat{\mathbf{X}}_{n+1}$, where $\hat{X}_{n+1} := (\nu_1 \ldots \nu_d)^T \hat{\mathbf{X}}_{n+1}$ is in $\mathcal{G}$.

Now we will compute the "distance " between $\hat{\mathbf{X}}_{n+1}$ and $\hat{\mathbf{X}}_{n+1}^{\mathcal{G}}$ in (4.59), and the "distance " is defined as $E\left\|\hat{\mathbf{X}}_{n+1} - \hat{\mathbf{X}}_{n+1}^{\mathcal{G}}\right\|_2^2$. We will show that, the distance is bounded and tends to 0 as $d \to \infty$. The result is stated in the following theorem.

Theorem 4.12. *Suppose we have functional ARMA$(p,1)$ process $(X_n, n \in \mathbb{Z})$ and the condition* ***FARMA*****$(\mathbf{p},\mathbf{1})$** *holds. $\hat{X}_{n+1}^{\mathcal{G}}$ is the functional best linear predictor of X_{n+1} defined in (4.62)-(4.63) and $\hat{\mathbf{X}}_{n+1}^{\mathcal{G}}$ is defined in (4.59). $\hat{\mathbf{X}}_{n+1}$ is the best linear predictor of $\mathbf{X}_{n+1}$, which is based on vector observations $\mathbf{X}_1, \ldots, \mathbf{X}_n$. Then the distance between $\hat{\mathbf{X}}_{n+1}$ and $\hat{\mathbf{X}}_{n+1}^{\mathcal{G}}$ is bounded by*

$$E\left\|\hat{\mathbf{X}}_{n+1} - \hat{\mathbf{X}}_{n+1}^{\mathcal{G}}\right\|_2^2 \leq 4\left(\sum_{i=1}^{n}\left(\sum_{l=d+1}^{\infty}\left\|g_{n,i}(\nu_l)\right\|^2\right)^{\frac{1}{2}}\right)^2 \sum_{l=d+1}^{\infty}\lambda_l. \tag{4.78}$$

Furthermore, $E\left\|\hat{\mathbf{X}}_{n+1} - \hat{\mathbf{X}}_{n+1}^{\mathcal{G}}\right\|_2^2$ tends to 0 as $d \to \infty$.

Before we directly prove this theorem, we need a technical lemma.

Lemma 4.13. *Suppose $(X_n, n \in \mathbb{Z})$ is a zero mean stationary functional ARMA$(p,1)$ proecess. $(\nu_l, l \in \mathbb{Z})$ are eigenfunctions of the covariance operator C. Then for $\forall j, l \in \mathbb{Z}$, (4.63) implies*

$$E\left[\left\langle X_{n+1} - \hat{X}_{n+1}^{\mathcal{G}}, \nu_l \right\rangle \langle y, \nu_j\rangle\right] = 0, \quad \forall y \in \mathcal{G}. \tag{4.79}$$

Proof. For $\forall j, l \in \mathbb{Z}$, we define $s_{l,j}(\cdot) := \langle \cdot, \nu_l \rangle \nu_j$.
$s_{l,j}(\cdot)$ is Hilbert Schmidt, since

$$\begin{aligned}\sum_{k=1}^{\infty} \|s_{l,j}(\nu_k)\|^2 &= \sum_{k=1}^{\infty} \| \langle \nu_k, \nu_l \rangle \nu_j\|^2 \\ &= \| \langle \nu_l, \nu_l \rangle \nu_j\|^2 = 1.\end{aligned}$$

Using the fact that

$$E\left\langle X_{n+1} - \hat{X}_{n+1}^{\mathcal{G}}, y \right\rangle = 0, \quad \forall y \in \mathcal{G},$$

and $s_{l,j}(y) \in \mathcal{G}$, we have

$$\begin{aligned}E\left[\left\langle X_{n+1} - \hat{X}_{n+1}^{\mathcal{G}}, s_{l,j}(y) \right\rangle\right] &= E\left[\left\langle X_{n+1} - \hat{X}_{n+1}^{\mathcal{G}}, \langle y, \nu_l \rangle \nu_j \right\rangle\right] \\ &= E\left[\left\langle X_{n+1} - \hat{X}_{n+1}^{\mathcal{G}}, \nu_l \right\rangle \langle y, \nu_j \rangle\right] \\ &= 0, \quad \forall j, l \in \mathbb{Z}.\end{aligned}$$

□

Now we begin to prove Theorem 4.12.
Proof of Theorem 4.12:

Proof. By (4.79), we have

$$\begin{aligned}\sum_{j=1}^{d} E\left[\langle y, \nu_j \rangle \left\langle X_{n+1} - \hat{X}_{n+1}^{\mathcal{G}}, \nu_j \right\rangle\right] &= E\left[\mathbf{y}^T \left(\mathbf{X}_{n+1} - \hat{\mathbf{X}}_{n+1}^{\mathcal{G}}\right)\right] \\ &\overset{(4.66)}{=} E\left[\mathbf{y}^T \left(\mathbf{X}_{n+1} - \sum_{i=1}^{n} \mathbf{G}_{ni}\mathbf{X}_i - \sum_{i=1}^{n} \mathbf{G}_{ni}^{\infty}\mathbf{X}_i^{\infty}\right)\right] \\ &= 0, \quad \forall y \in \mathcal{G}, \end{aligned} \tag{4.80}$$

where

$$\mathbf{y} = \left(\langle y, \nu_1 \rangle, \ldots, \langle y, \nu_d \rangle\right)^T \in \mathbf{M}. \tag{4.81}$$

By (4.76), (4.80) holds for $\forall \mathbf{y} \in \mathbf{M}$. Furthermore, since $\mathbf{M}_1 \subseteq \mathbf{M}$, (4.80) holds for $\forall \mathbf{y} \in \mathbf{M}$ implies (4.80) holds for $\forall \mathbf{y}_1 \in \mathbf{M}_1$, i.e.

$$E\left[\mathbf{y}_1^T \left(\mathbf{X}_{n+1} - \sum_{i=1}^{n} \mathbf{G}_{ni}\mathbf{X}_i - \sum_{i=1}^{n} \mathbf{G}_{ni}^{\infty}\mathbf{X}_i^{\infty}\right)\right] = 0, \quad \forall \mathbf{y}_1 \in \mathbf{M}_1. \tag{4.82}$$

Combining (4.82) and (4.61), we have

$$E\left[\mathbf{y}_1^T \left(\hat{\mathbf{X}}_{n+1} - \sum_{i=1}^{n} \mathbf{G}_{ni}\mathbf{X}_i\right)\right] = E\left[\mathbf{y}_1^T \left(\sum_{i=1}^{n} \mathbf{G}_{ni}^{\infty}\mathbf{X}_i^{\infty}\right)\right], \quad \forall \mathbf{y}_1 \in \mathbf{M}_1. \tag{4.83}$$

Since both $\hat{\mathbf{X}}_{n+1}$ and $\sum_{i=1}^{n} \mathbf{G}_{ni}\mathbf{X}_i$ are in $\mathbf{M}_1$, (4.83) especially holds when

$$\mathbf{y}_1 = \hat{\mathbf{X}}_{n+1} - \sum_{i=1}^{n} \mathbf{G}_{ni}\mathbf{X}_i \in \mathbf{M}. \tag{4.84}$$

We plug the $\mathbf{y}_1$ defined in (4.84) in (4.83), then we have

$$E\left[\left(\hat{\mathbf{X}}_{n+1} - \sum_{i=1}^{n} \mathbf{G}_{ni}\mathbf{X}_i\right)^T \left(\hat{\mathbf{X}}_{n+1} - \sum_{i=1}^{n} \mathbf{G}_{ni}\mathbf{X}_i\right)\right] = E\left[\left(\hat{\mathbf{X}}_{n+1} - \sum_{i=1}^{n} \mathbf{G}_{ni}\mathbf{X}_i\right)^T \left(\sum_{i=1}^{n} \mathbf{G}_{ni}^{\infty}\mathbf{X}_i^{\infty}\right)\right]. \tag{4.85}$$

Now we caculate the left and the right part of (4.85) respectively:

$$E\left[\left(\hat{\mathbf{X}}_{n+1} - \sum_{i=1}^{n} \mathbf{G}_{ni}\mathbf{X}_i\right)^T \left(\hat{\mathbf{X}}_{n+1} - \sum_{i=1}^{n} \mathbf{G}_{ni}\mathbf{X}_i\right)\right] = E\left\|\hat{\mathbf{X}}_{n+1} - \sum_{i=1}^{n} \mathbf{G}_{ni}\mathbf{X}_i\right\|_2^2, \tag{4.86}$$

and

$$\begin{aligned} E\left[\left(\hat{\mathbf{X}}_{n+1} - \sum_{i=1}^{n} \mathbf{G}_{ni}\mathbf{X}_i\right)^T \left(\sum_{i=1}^{n} \mathbf{G}_{ni}^{\infty}\mathbf{X}_i^{\infty}\right)\right] &= E\left[\left\langle \hat{\mathbf{X}}_{n+1} - \sum_{i=1}^{n} \mathbf{G}_{ni}\mathbf{X}_i, \sum_{i=1}^{n} \mathbf{G}_{ni}^{\infty}\mathbf{X}_i^{\infty} \right\rangle_{\mathbb{R}^d}\right] \\ &\le E\left[\left\|\hat{\mathbf{X}}_{n+1} - \sum_{i=1}^{n} \mathbf{G}_{ni}\mathbf{X}_i\right\|_2 \cdot \left\|\sum_{i=1}^{n} \mathbf{G}_{ni}^{\infty}\mathbf{X}_i^{\infty}\right\|_2\right] \\ &\le \left(E\left\|\hat{\mathbf{X}}_{n+1} - \sum_{i=1}^{n} \mathbf{G}_{ni}\mathbf{X}_i\right\|_2^2\right)^{\frac{1}{2}} \left(E\left\|\sum_{i=1}^{n} \mathbf{G}_{ni}^{\infty}\mathbf{X}_i^{\infty}\right\|_2^2\right)^{\frac{1}{2}}, \end{aligned} \tag{4.87}$$

where $\langle\ ,\ \rangle_{\mathbb{R}^d}$ denotes the scalar product of two d-dimensioanl vectors and $\|\cdot\|_2$ denotes the Euclidean vector norm.

Combining (4.83), (4.86) and (4.87), we have

$$E\left\|\hat{\mathbf{X}}_{n+1} - \sum_{i=1}^{n} \mathbf{G}_{ni}\mathbf{X}_i\right\|_2^2 \le E\left\|\sum_{i=1}^{n} \mathbf{G}_{ni}^{\infty}\mathbf{X}_i^{\infty}\right\|_2^2. \tag{4.88}$$

Recall that our task is to compute the distance $E\left\|\hat{\mathbf{X}}_{n+1}-\hat{\mathbf{X}}_{n+1}^{\mathcal{G}}\right\|_2^2$. By (4.88), we have

$$\begin{aligned} E\left\|\hat{\mathbf{X}}_{n+1}-\hat{\mathbf{X}}_{n+1}^{\mathcal{G}}\right\|_2^2 &= E\left\|\hat{\mathbf{X}}_{n+1}-\sum_{i=1}^{n}\mathbf{G}_{ni}\mathbf{X}_i-\sum_{i=1}^{n}\mathbf{G}_{ni}^{\infty}\mathbf{X}_i^{\infty}\right\|_2^2 \\ &\leq 2E\left\|\hat{\mathbf{X}}_{n+1}-\sum_{i=1}^{n}\mathbf{G}_{ni}\mathbf{X}_i\right\|_2^2+2E\left\|\sum_{i=1}^{n}\mathbf{G}_{ni}^{\infty}\mathbf{X}_i^{\infty}\right\|_2^2 \\ &\leq 4E\left\|\sum_{i=1}^{n}\mathbf{G}_{ni}^{\infty}\mathbf{X}_i^{\infty}\right\|_2^2. \end{aligned} \tag{4.89}$$

Recall the representation of $\sum_{i=1}^{n}\mathbf{G}_{ni}^{\infty}\mathbf{X}_i^{\infty}$ defined in (4.66), it is a d-dimensional vector with lth element

$$\sum_{i=1}^{n}\sum_{l'=d+1}^{\infty}\langle X_i,\nu_{l'}\rangle\langle g_{n,i}(\nu_l'),\nu_l\rangle=\sum_{i=1}^{n}\sum_{l'=d+1}^{\infty}x_{i,l'}\langle g_{n,i}(\nu_l'),\nu_l\rangle.$$

Then we calculate $E\left\|\sum_{i=1}^{n}\mathbf{G}_{ni}^{\infty}\mathbf{X}_i^{\infty}\right\|_2^2$ in the following part:

$$
\begin{aligned}
E\left\|\sum_{i=1}^{n}\mathbf{G}_{ni}^{\infty}\mathbf{X}_{i}^{\infty}\right\|_{2}^{2} &= E\left[\sum_{l=1}^{d}\left(\sum_{i=1}^{n}\sum_{l'=d+1}^{\infty}x_{i,l'}\left\langle g_{n,i}(\nu_l'),\nu_l\right\rangle\right)^{2}\right] \\
&= E\left\|\sum_{l=1}^{d}\left(\sum_{i=1}^{n}\sum_{l'=d+1}^{\infty}x_{i,l'}\left\langle g_{n,i}(\nu_l'),\nu_l\right\rangle\right)\nu_l\right\|^{2} \\
&\leq E\left\|\sum_{l=1}^{\infty}\left(\sum_{i=1}^{n}\sum_{l'=d+1}^{\infty}x_{i,l'}\left\langle g_{n,i}(\nu_l'),\nu_l\right\rangle\right)\nu_l\right\|^{2} \\
&\overset{\text{Parseval's identity}}{=} E\left\|\sum_{i=1}^{n}\sum_{l'=d+1}^{\infty}x_{i,l'}g_{n,i}(\nu_l')\right\|^{2} \\
&= E\left[\left\langle\sum_{i=1}^{n}\sum_{l=d+1}^{\infty}x_{i,l}g_{n,i}(\nu_l),\sum_{j=1}^{n}\sum_{l'=d+1}^{\infty}x_{j,l'}g_{n,j}(\nu_l')\right\rangle\right] \\
&= \sum_{i,j=1}^{n}\sum_{l,l'=d+1}^{\infty}E\left(x_{i,l}x_{j,l'}\right)\left\langle g_{n,i}(\nu_l),g_{n,j}(\nu_{l'})\right\rangle \\
&\leq \left(\sum_{i=1}^{n}\sum_{l=d+1}^{\infty}\sqrt{E(x_{i,l})^2}\left\|g_{n,i}(\nu_l)\right\|\right)\left(\sum_{j=1}^{n}\sum_{l'=d+1}^{\infty}\sqrt{E(x_{j,l'})^2}\left\|g_{n,j}(\nu_l')\right\|\right) \\
&= \left(\sum_{i=1}^{n}\sum_{l=d+1}^{\infty}\sqrt{\lambda_l}\left\|g_{n,i}(\nu_l)\right\|\right)\left(\sum_{j=1}^{n}\sum_{l'=d+1}^{\infty}\sqrt{\lambda_l'}\left\|g_{n,j}(\nu_l')\right\|\right) \\
&\leq \left(\sum_{i=1}^{n}\left(\sum_{l=d+1}^{\infty}\lambda_l\right)^{\frac{1}{2}}\left(\sum_{l=d+1}^{\infty}\left\|g_{n,i}(\nu_l)\right\|^2\right)^{\frac{1}{2}}\right) \\
&\times \left(\sum_{j=1}^{n}\left(\sum_{l'=d+1}^{\infty}\lambda_l'\right)^{\frac{1}{2}}\left(\sum_{l'=d+1}^{\infty}\left\|g_{n,j}(\nu_l')\right\|^2\right)^{\frac{1}{2}}\right) \\
&= \left(\sum_{i=1}^{n}\left(\sum_{l=d+1}^{\infty}\left\|g_{n,i}(\nu_l)\right\|^2\right)^{\frac{1}{2}}\right)^{2}\sum_{l=d+1}^{\infty}\lambda_l \qquad (4.90)
\end{aligned}
$$

Note that

$$\left(\sum_{l=d+1}^{\infty}\left\|g_{n,i}(\nu_l)\right\|^2\right)^{\frac{1}{2}} \leq \left\|g_{n,i}\right\|_{\mathcal{S}}^{\frac{1}{2}} < \infty.$$

Thus (4.90) is bounded by

$$\left(\sum_{i=1}^{n}\left(\sum_{l=d+1}^{\infty}\left\|g_{n,i}(\nu_l)\right\|^2\right)^{\frac{1}{2}}\right)^{2}\sum_{l=d+1}^{\infty}\lambda_l \leq \left(\sum_{i=1}^{n}\left\|g_{n,i}\right\|_{\mathcal{S}}^{\frac{1}{2}}\right)^{2}\sum_{l=d+1}^{\infty}\lambda_l < \infty,$$

and

$$\left(\sum_{i=1}^{n}\left(\sum_{l=d+1}^{\infty}\|g_{n,i}(\nu_l)\|^2\right)^{\frac{1}{2}}\right)^2\sum_{l=d+1}^{\infty}\lambda_l\to 0,\quad \text{as } d\to\infty. \tag{4.91}$$

we finish the proof by combining (4.89), (4.90) and (4.91). □

4.5 Bound for the prediction error

In the **Algorithm I**, once we have the vector predictor

$$\hat{\mathbf{X}}_{n+1}=(\hat{x}_{n+1,1},\ldots,\hat{x}_{n+1,d})^T, \tag{4.92}$$

which is based on the vector observations $\mathbf{X}_1,\ldots,\mathbf{X}_n$, we need to transform it to its functional form

$$\hat{X}_{n+1}:=\sum_{i=1}^{d}\hat{x}_{n+1,i}\nu_i=(\nu_1\ldots\nu_d)^T\hat{\mathbf{X}}_{n+1}. \tag{4.93}$$

In this section we want to bound the mean square prediction error $E\|X_{n+1}-\hat{X}_{n+1}\|^2$. We will show that, the prediction error is composed of two parts. The first part is relevant to $E\|X_{n+1}-\hat{X}^{\mathcal{G}}_{n+1}\|^2$ and thus bounded by σ_n^2. It also implies that $\hat{X}_{n+1}$ can not be better than the functional best linear predictor $\hat{X}^{\mathcal{G}}_{n+1}$. Furthermore, the second part tends to zero as $d\to\infty$.

Theorem 4.14. *Consider the functional ARMA$(p,1)$ process* (4.6) *and suppose the condition* ***FARMA*****(p, 1)** *holds, i.e. both the functional process and the vector process are stationary. Then the mean square prediction error is bounded by*

$$E\left\|X_{n+1}-\hat{X}_{n+1}\right\|^2\le\sigma_n^2+\gamma_d,$$

where

$$\begin{aligned}
&\sigma_n^2:=E\|X_{n+1}-\hat{X}^{\mathcal{G}}_{n+1}\|^2,\\
&\gamma_d=4\left[\sum_{i=1}^{n}g_{n;i,d}\right]^2\sum_{l=d+1}^{\infty}\lambda_l+E\left\|\sum_{l=d+1}^{\infty}\left\langle\hat{X}^{\mathcal{G}}_{n+1},\nu_l\right\rangle\nu_l\right\|^2,\\
&g_{n;i,d}^2=\sum_{l=d+1}^{\infty}\|g_{n,i}(\nu_l)\|^2\le\|g_{n,i}\|_{\mathcal{S}}^2.
\end{aligned}$$

Furthermore, γ_d tends to 0 as $d\to\infty$.

Proof. Since $\left(X_{n+1}-\hat{X}^{\mathcal{G}}_{n+1}\right)$ is orthogonal to $\mathcal{G}$ and $\left(\hat{X}^{\mathcal{G}}_{n+1}-\hat{X}_{n+1}\right)$ is in $\mathcal{G}$, we have

$$\begin{aligned}
E\left\|X_{n+1}-\hat{X}_{n+1}\right\|^2&=E\left\|\left(X_{n+1}-\hat{X}^{\mathcal{G}}_{n+1}\right)+\left(\hat{X}^{\mathcal{G}}_{n+1}-\hat{X}_{n+1}\right)\right\|^2\\
&=E\left\|X_{n+1}-\hat{X}^{\mathcal{G}}_{n+1}\right\|^2+E\left\|\hat{X}^{\mathcal{G}}_{n+1}-\hat{X}_{n+1}\right\|^2.
\end{aligned}\tag{4.94}$$

The first part of (4.94) is bounded by

$$E\left\|X_{n+1}-\hat{X}_{n+1}^{\mathcal{G}}\right\|^2 \leq \sigma_n^2. \tag{4.95}$$

Now we bound the second part of (4.94).

$$\begin{aligned} E\left\|\hat{X}_{n+1}^{\mathcal{G}}-\hat{X}_{n+1}\right\|^2 &= E\left\|\sum_{l=1}^{\infty}\left\langle \hat{X}_{n+1}^{\mathcal{G}}-\hat{X}_{n+1},\nu_l\right\rangle \nu_l\right\|^2 \\ &= E\left\|\sum_{l=1}^{d}\left\langle \hat{X}_{n+1}^{\mathcal{G}}-\hat{X}_{n+1},\nu_l\right\rangle \nu_l\right\|^2 + E\left\|\sum_{l=d+1}^{\infty}\left\langle \hat{X}_{n+1}^{\mathcal{G}},\nu_l\right\rangle \nu_l\right\|^2 \\ &= E\left\|\hat{\mathbf{X}}_{n+1}^{\mathcal{G}}-\hat{\mathbf{X}}_{n+1}\right\|_2^2 + E\left\|\sum_{l=d+1}^{\infty}\left\langle \hat{X}_{n+1}^{\mathcal{G}},\nu_l\right\rangle \nu_l\right\|^2. \end{aligned} \tag{4.96}$$

By Theorem 4.12, $E\left\|\hat{\mathbf{X}}_{n+1}-\hat{\mathbf{X}}_{n+1}^{\mathcal{G}}\right\|_2^2$ is bounded by

$$E\left\|\hat{\mathbf{X}}_{n+1}-\hat{\mathbf{X}}_{n+1}^{\mathcal{G}}\right\|_2^2 \leq 4\left(\sum_{i=1}^{n}\left(\sum_{l=d+1}^{\infty}\|g_{n,i}(\nu_l)\|^2\right)^{\frac{1}{2}}\right)^2\sum_{l=d+1}^{\infty}\lambda_l, \tag{4.97}$$

and tends to 0 as $d\to\infty$.
The second part of (4.96) is bounded by

$$\begin{aligned} E\left\|\sum_{l=d+1}^{\infty}\left\langle \hat{X}_{n+1}^{\mathcal{G}},\nu_l\right\rangle \nu_l\right\|^2 &\leq E\left\|\sum_{l=1}^{\infty}\left\langle \hat{X}_{n+1}^{\mathcal{G}},\nu_l\right\rangle \nu_l\right\|^2 \\ &= E\left\|\hat{X}_{n+1}^{\mathcal{G}}\right\|^2. \end{aligned} \tag{4.98}$$

Furthermore, $E\left\|\sum_{l=d+1}^{\infty}\left\langle \hat{X}_{n+1}^{\mathcal{G}},\nu_l\right\rangle \nu_l\right\|^2$ tends to 0 as $d\to\infty$. We finish the proof by gathering (4.94)-(4.98). □

Chapter 5

Simulation study

To verifty the results of our study on functional ARMA$(p, 1)$ process in Chapter 4, we conduct simulation studies in this chapter. We will do the following things.

- First of all we use Gaussian integral kernel (see Section 2.1.3) to simulate stationary functional ARMA$(p, 1)$ processes.
- To verify whether the truncated vector process still follows ARMA$(p, 1)$ structure, we fit different vector ARMA models (with different orders) to the truncated vector observations. We compare the goodness of fit (AIC and BIC) and check whether the vector ARMA$(p, 1)$ model fits the data better than other models.
- We assume the vector observations follow the models in the last step respectively, and then implement **Algorithm I** introduced in Section 4.2. Then we compute and compare the prediction errors and check whether the predictor based on the assumption of ARMA$(p, 1)$ is better than others.
- Furthermore, we will apply different types of univariate prediction methods, e.g. exponential smoothing, on the prediction of the functional process. And we compare these univariate predictors with the functional predictors (based on vector models).

For the functional data analysis (e.g. the computation of empirical eigenfunctions and eigenvalues), we use the R package **FDA**. For the multivariate time series analysis, e.g. parameter estimation and the computation of multivariate predictors, we use the R package **MTS**.

5.1 Procedure and some notations

5.1.1 Settings in the functional ARMA$(p, 1)$ model

We use Gaussian integral kernel to simulate the functional ARMA$(p, 1)$ process, i.e.

$$X_n(t) = \sum_{i=1}^{p} \Phi_i(X_{n-i})(t) + B_n(t) + \Theta(B_{n-1})(t)$$
$$= \sum_{i=1}^{p} \int_0^1 c_i * \phi_i(t,s) X_{n-i}(s)\mathrm{d}s + B_n(t) + \int_0^1 c_\theta * \theta_i(t,s) B_{n-1}(s)\mathrm{d}s, \quad i = 1, \ldots, p, \tag{5.1}$$

where

$$\phi_i(t,s) = \theta(t,s) = \exp\left(\frac{t^2+s^2}{2}\right), \quad t, s \in [0,1],\ i = 1, \ldots, p, \tag{5.2}$$

and c_i, c_θ are some constants. Note that

$$\iint_0^1 \exp\left(\frac{t^2+s^2}{2}\right) \mathrm{d}t\mathrm{d}s \approx 2.3772.$$

The innovations $(B_n, n \in \mathbb{Z})$ are assumed to be i.i.d Brownian bridges on $[0,1]$.

5.1.2 Generation and transformation of the observations

We divide $[0,1]$ into 1000 intervals with equal length, i.e. $t_0 = 0 < t_1 < \cdots < t_{999} < t_{1000} = 1$ and $\Delta_t := t_j - t_{j-1} = 0.001,\ j = 1, \ldots, 1000$. We set our number of oberservations $N = 200$. Then the $N = 200$ vector observations $\left\{(X_n(t_0), \ldots, X_n(t_{1000}))^T, n = 1, \ldots, N\right\}$ are generated as follows:

- X_1 is initialized to 0, i.e. $X_1(t_j) = 0,\ j = 0, \ldots, 1000$.
- We choose appropriate c_i such that the functional process is stationary. For example, to simulate stationary functional ARMA(1,1) process, c_1 needs to satisfy

$$\left(\iint_0^1 c_1^2 \exp\left(\frac{t^2+s^2}{2}\right) \mathrm{d}t\mathrm{d}s\right)^{\frac{1}{2}} = \|\Phi\|_{\mathcal{S}} < 1.$$

- We approximate and simulate $\left\{(X_n(t_0), \ldots, X_n(t_{1000}))^T, n = 2, \ldots, N\right\}$ by

$$X_n(t_j) \approx \sum_{i=1}^{p} \sum_{k=0}^{1000} c_i \phi_i(t_j, t_k) X_{n-i}(t_k) \Delta_t + B_n(t_j) + \sum_{k=0}^{1000} c_\theta \theta(t_j, t_k) B_{n-1}(t_k) \Delta_t, \tag{5.3}$$

  where $j, k = 0, \ldots, 1000$. Note that, as $\Delta_t \to 0$, (5.3) tends to (5.1).

Up to now, we have generated the N vector observations $\left\{(X_n(t_0), \ldots, X_n(t_{1000}))^T, n = 1, \ldots, N\right\}$. Now we need to transform the vector observations to functional observations (smoothing).

We choose M Fourier basis functions which are defined as

$$F_m(x) = \begin{cases} 1 & m = 1, \\ \cos(\pi m x) & m \text{ is even}, \\ \sin(\pi(m-1)x) & m \text{ is odd}, \end{cases}$$

where $x \in [0, 1]$. The Fourier basis functions can be generated by the command *create.fourier.basis()* in the R package **FDA**. For a **fixed** M and for each $n \in \{1, \ldots, N\}$, we determine the coefficients $c_{n1}, \ldots, c_{nM}$ of the M basis functions by minimizing the least squares criterion

$$\sum_{j=0}^{1000} \left(X_n(t_j) - \sum_{m=1}^{M} c_{nm} F_m(t_j) \right)^2 . \tag{5.4}$$

We define that

$$\mathbf{F} := \begin{pmatrix} F_1(t_0) & \ldots & F_M(t_0) \\ \vdots & \vdots & \vdots \\ F_1(t_{1000}) & \ldots & F_M(t_{1000}) \end{pmatrix},$$

then by the theory of multivariate linear regression, for each $n \in \{1, \ldots, N\}$,

$$\begin{pmatrix} c_{n1} \\ \vdots \\ c_{nM} \end{pmatrix} = \left(\mathbf{F}^T \mathbf{F}\right)^{-1} \mathbf{F}^T \begin{pmatrix} X_n(t_0) \\ \vdots \\ X_n(t_{1000}) \end{pmatrix}.$$

With these coefficients, we can transform the high-dimensional vector observations to functional observations $\{X_n(t),\ t \in [0, 1],\ n = 1, \ldots, N\}$, where

$$X_n(t) := \sum_{m=1}^{M} c_{nm} F_m(t), \quad n = 1 \ldots, N,\ t \in [0, 1]. \tag{5.5}$$

But how to choose the number M of basis functions? In fact, there is no one "gold" standard method to select an optimal M. One way is to compare the mean absolute errors

$$\frac{1}{N} \sum_{n=1}^{N} \frac{1}{1000} \sum_{j=0}^{1000} \mid X_n(t_j) - \sum_{m=1}^{M} c_{nm} F_m(t_j) \mid \tag{5.6}$$

for different M. To show the process of choosing M, we simulate functional ARMA$(1, 1)$ process with $\|\Phi\| = \|\Theta\| = 0.9$ ($N = 200$ observations). From Figure 5.1, we can see that, as the number M of the Fourier basis functions increases, the mean absolute error defined in (5.6) decreases. And the error curve becomes level from about $M = 29$. Furthermore, the error is quite small in comparison to the data. In our simulation study, we choose $M = 29$. More details of chooing the number of basis functions see Ramsay and Silverman [2005].

Remark 9. *In Figure 5.1, we smooth the observations with the R command* Data2fd *instead of* smooth.basis. *We will compare these two commands in details in Section 6.1.3 (see Remark 11 and 12).*

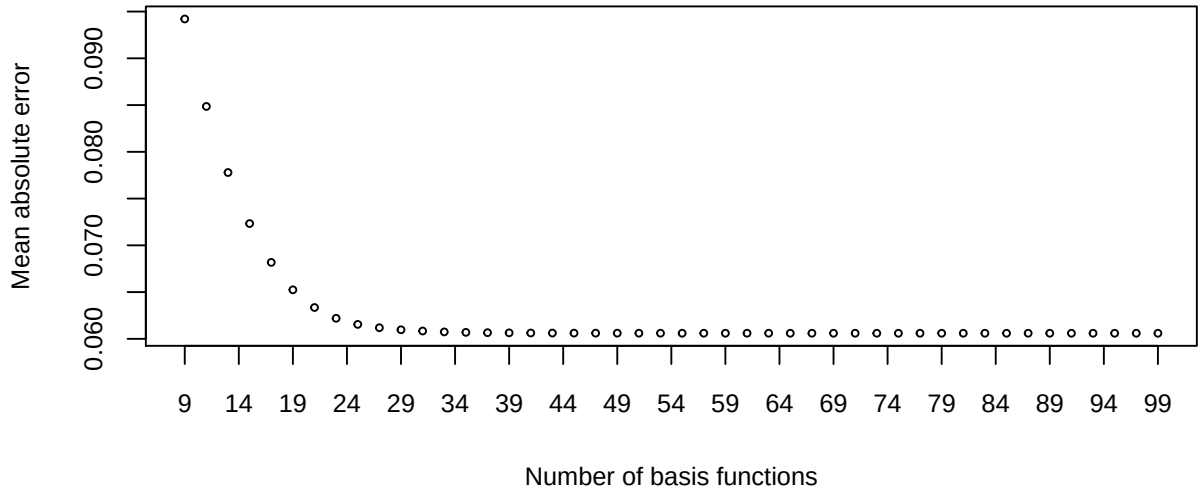

Figure 5.1: $N = 200$ observations of the functional ARMA(1,1) process ($\|\Phi\| = \|\Theta\| = 0.9$) are smoothed with different number of Fourier basis, with the R command *Data2fd*. Depicted are the mean absolute errors by choosing different number of Fourier basis functions. The results are based on 100 loops.

5.1.3 Model fitting and prediction

Up to now, we have $N = 200$ functional observations (by choosing $M = 29$), and the data type of these observations is "fd" in R. With CPV method (see Section 2.2.3), we determine the number d of FPC's by setting the criterion to 90%, i.e. $CPV(d-1) < 0.9$ and $CPV(d) \geq 0.9$. Then we truncate the functional observations $X_1, \ldots X_N$ by

$$X_n^e = \sum_{l=1}^{d} \langle X_n, \nu_l^e \rangle \, \nu_l^e, \quad n = 1, \ldots, N. \tag{5.7}$$

The corresponding vector form of X_n^e in (5.7) is

$$\mathbf{X}_n^e := \left(\langle X_n, \nu_1^e \rangle, \ldots, \langle X_n, \nu_d^e \rangle \right)^T, \quad n = 1, \ldots, N.$$

The empirical eigenfunctions and the corresponding empirical eigenvalues can be computed with the R command *pca.fd()*.

The following steps are executed 100 times, and all the results are the average value in 100 loops.

- We fit different vector ARMA(p, q) models (i.e. different orders, e.g AR(1), AR(2), MA(1), ARMA(1,1)) to the vector data $\mathbf{X}_1^e, \ldots, \mathbf{X}_N^e$. We compare the goodness of fit (AIC and BIC) among different models. The model fitting and the results of the goodness of fit (AIC and BIC) can be achieved with the commands *VARMA()* and *VAR()* installed in the R package **MTS**.

- We implement multivariate prediction algorithm based on the assumtion that $\mathbf{X}_1^e, \ldots, \mathbf{X}_N^e$ follow the models mentioned above respectively. We compute the vector predictor $\hat{\mathbf{X}}_{N-9}^e, \ldots, \hat{\mathbf{X}}_N^e$ for the last 10 observations $X_{N-9}, \ldots, X_N$. The vector predictors can be computed with the commands *VARMApred()* and *VARpred()* installed in the R package **MTS**.

- We transform the 10 vector predictors to functional predictor:

$$\hat{X}_{N-k}^e := (\nu_1^e \ldots \nu_d^e)^T \hat{\mathbf{X}}_{N-k}^e, \quad k = 0, \ldots, 9. \tag{5.8}$$

  For each functional predictor $\hat{X}_{N-k}^e,\ k = 0, \ldots, 9.$, we compare it with the truncated functional observation X_{N-k}^e defined in (5.7) and compute the prediction error. Then we compute the average value of the 10 prediction errors. We use two types of error, which are as stated in the following.

  - **Root mean square error (RMSE)**
    The average root mean square error (RMSE) of $\hat{X}_{N-9}^e, \hat{X}_{N-8}^e, \ldots, \hat{X}_N^e$ is defined as

$$\frac{1}{10} \sum_{k=0}^{9} \sqrt{\sum_{j=0}^{1000} \left(\hat{X}_{N-k}^e(t_j) - X_{N-k}^e(t_j) \right)^2 \Delta_t}. \tag{5.9}$$

    Note that, as $\Delta_t \to 0$, (5.9) tends to

$$\frac{1}{10} \sum_{k=0}^{9} \sqrt{\int_0^1 \left(\hat{X}_{N-k}^e(t) - X_{N-k}^e(t) \right)^2 \mathrm{d}t}.$$

  - **Mean absolute error (MAE)**
    The average mean absolute error (MAE) of $\hat{X}_{N-9}^e, \hat{X}_{N-8}^e, \ldots, \hat{X}_N^e$ is defined as

$$\frac{1}{10} \sum_{k=0}^{9} \sum_{j=0}^{1000} \mid \hat{X}_{N-k}^e(t_j) - X_{N-k}^e(t_j) \mid \Delta_t. \tag{5.10}$$

- Furthermore, we implement three univariate prediction methods . In this case, for each time point t_j, we see $X_1^e(t_j), X_2^e(t_j), \ldots, X_N^e(t_j)$ as a univariate time series, $j = 1, \ldots, 1000$. The three univariate prediction methods are stated in the following.

  - **Mean method**
    For each $k \in \{0, \ldots, 9\}$, we denote the mean method predictor of X_{N-k} by $\hat{X}_{N-k}^{Mean}$ and it is defined as

$$\hat{X}_{N-k}^{Mean}(t_j) := \frac{1}{N-k-1} \sum_{i=1}^{N-k-1} X_i^e(t_j), \quad j = 0, \ldots, 1000. \tag{5.11}$$

– **Drift method**
For each $k \in \{0, \ldots, 9\}$, we denote the drift method predictor of X_{N-k} by $\hat{X}_{N-k}^{Drift}$ and it is defined as

$$\hat{X}_{N-k}^{Drift}(t_j) := X_{N-k-1}^{e}(t_j) + \frac{1}{N-k-2} \sum_{i=2}^{N-k-1} \left(X_i^e(t_j) - X_{i-1}^e(t_j) \right), \quad j = 0, \ldots, 1000. \tag{5.12}$$

– **Exponential smoothing (ES)**
For each $k \in \{0, \ldots, 9\}$, we denote the exponential smoothing predictor of X_{N-k} by $\hat{X}_{N-k}^{ES}$ and it is defined as

$$\hat{X}_{N-k}^{ES}(t_j) := \alpha_j X_{N-k-1}^{e}(t_j) + (1-\alpha_j)\, \hat{X}_{N-k-1}^{ES}(t_j) \quad j = 0, \ldots, 1000. \tag{5.13}$$

α_j is selected such that

$$\sum_{i=1}^{N-k} \left(X_i^e(t_j) - \hat{X}_i^{ES}(t_j) \right)^2$$

is minimized, $j = 0, \ldots, 1000$. This step can be achived with the command *forecast.ets()*.

5.2 Results

5.2.1 Functional ARMA$(1,1)$ process

In the simulation of function ARMA$(1,1)$ process, we choose $c_1 = c_\theta = 0.6153$ such that $\|\Phi\|_{\mathcal{S}} = \|\Theta\|_{\mathcal{S}} \approx 0.9$. We choose the number d of FPC's such that $\nu_1^e, \ldots, \nu_d^e$ explain 95% of the total data variability. In the 100 loops, the average d is 2.25.

We fit vector AR(1), AR(2), MA(1), MA(2), ARMA(1,1) model to the vector data respectively. But we find that MA(1) model does not perform well both in the model fitting and prediction. So we do not list the results of MA(1) in Table 5.1 and 5.2.

In Table 5.1 we list the results of the goodness of fit for vector AR(1), AR(2), MA(2), ARMA(1,1) model. We can see that, both AIC and BIC of ARMA(1,1) are smaller than the others, which means vector ARMA(1,1) model fits the vector observations better than the other vector models.

Model fit	VAR(1)	VAR(2)	VMA(2)	VARMA(1,1)
AIC	-7.0861	-7.3450	-3.8123	**-7.5820**
BIC	-6.9979	-7.1684	-3.6542	**-7.4055**

Table 5.1: Simulation of functional ARMA(1,1) with $\|\Phi\|_{\mathcal{S}} = 0.9$ and $\|\Theta\|_{\mathcal{S}} = 0.9$. Goodness of fit of different models. Results are based on 100 loops.

Then we implement the multivariate prediction algorithm based on the assumption of AR(1), AR(2), MA(2), ARMA(1,1), and the three univariate prediction algorithms mentioned in Section 5.1.3. We can compare the results in Table 5.2. We can see that, the functional predictor based on vector ARMA(1,1) model is better than any other functional and univariate predictors. Furthermore, generally speaking, the functional predictors perform better than the univariate predictors.

We also visualize the four functional predictors in Figure 5.2. The real black line shows the truncated functional data $X_{190}^e, \ldots, X_{200}^e$ defined in (5.7). Other lines with different types and in different colours show the different functional predictors defined in (5.8).

Model fit	VAR(1)	VAR(2)	VMA(2)	VARMA(1,1)	Mean	Drift	ES
RMSE	0.3226	0.3027	0.4161	**0.2867**	0.9070	0.3997	0.4005
MAE	0.2921	0.2714	0.3857	**0.2555**	0.8760	0.3545	0.3535

Table 5.2: Simulation of functional ARMA(1,1) with $\|\Phi\|_{\mathcal{S}} = 0.9$ and $\|\Theta\|_{\mathcal{S}} = 0.9$. Prediction errors from different prediction methods. Results are based on 100 loops.

5.2.2 Functional MA(1) process

In the simulation of function MA(1) process, we choose $c_\theta = 0.3891$ such that $\|\Theta\|_{\mathcal{S}} \approx 0.6$. We choose the number d of FPC's such that $\nu_1^e, \ldots, \nu_d^e$ explain 90% of the total data variability. In the 100 loops, the average d is 4.38.

We still tried to fit vector AR(1), AR(2), MA(1), MA(2), ARMA(1,1) model to the vector data respectively. But we find that ARMA(1,1) model does not perform well both in the model fitting and prediction. So we do not list the results of ARMA(1,1) in Table 5.3 and 5.4.

From Table 5.3, both AIC and BIC of MA(1) are smaller than others, which means vector MA(1) model fits the vector observations better than other vector models.

From Table 5.4, the predictor based on MA(1) is better than any other multivariate and univariate predictors. And generally speaking, the functional predictors still perform better than univariate predictors.

Again, we visualize the four multivariate predictors in Figure 5.3.

Model fit	VAR(1)	VAR(2)	VMA(1)	VMA(2)
AIC	-19.8785	-19.8637	**-20.015**	-19.9158
BIC	-19.4873	-19.0812	**-19.623**	-19.1333

Table 5.3: Simulation of functional MA(1) with $\|\Theta\|_{\mathcal{S}} = 0.6$. Goodness of fit of different models. Results are based on 100 loops.

Model fit	VAR(1)	VAR(2)	VMA(1)	VMA(2)	Mean Method	Drift Method	ES
RMSE	0.3618	0.3630	**0.3536**	0.3612	0.3817	0.4724	0.3855
MAE	0.3087	0.3083	**0.3011**	0.3062	0.3293	0.3993	0.3315

Table 5.4: Simulation functional MA(1) with $\|\Theta\|_{\mathcal{S}} = 0.6$. Prediction errors from different prediction methods. Results are based on 100 loops.

5.3 Summary

We finish this chapter by summarizing the results listed in Section 5.2.1 and 5.2.2.

By the theory in Chapter 4, the vector process $(\mathbf{X}_n, n \in \mathbb{Z})$ obtained from a stationary functional ARMA$(p, 1)$ process $(X_n, n \in \mathbb{Z})$ follows the vector ARMA$(p, 1)$ structure. Based on this fact, the vector ARMA$(p, 1)$ is supposed to perform better than other vector models in the model fitting and prediction. The results listed in Section 5.2.1 and 5.2.2. are in accordance with the theory.

In fact, except the two cases listed in this chapter (ARMA(1,1) with $\|\Phi\|_{\mathcal{S}} = \|\Theta\|_{\mathcal{S}} = 0.9$ and MA(1) with $\|\Theta\|_{\mathcal{S}} = 0.6$), we also executed the whole procedure on the other ARMA process by changing the order p and the operator norm. As expected and without surprise, the results are still in accordance with our theory.

Since both the theory and the computation of the functional prediction are much more complicated than univariate prediction, so we expect that the functional predictors perform better than the univariate predictors, which makes sense to the work on functional predicton. From the results of the prediction errors, we can say the functional predictors are at least not worse than the univariate predictors, on the prediction of functional observations.

In the next chapter, the whole procedure in this chapter is applied on analysing real highway traffic data provided by Autobahndirektion Südbayern.

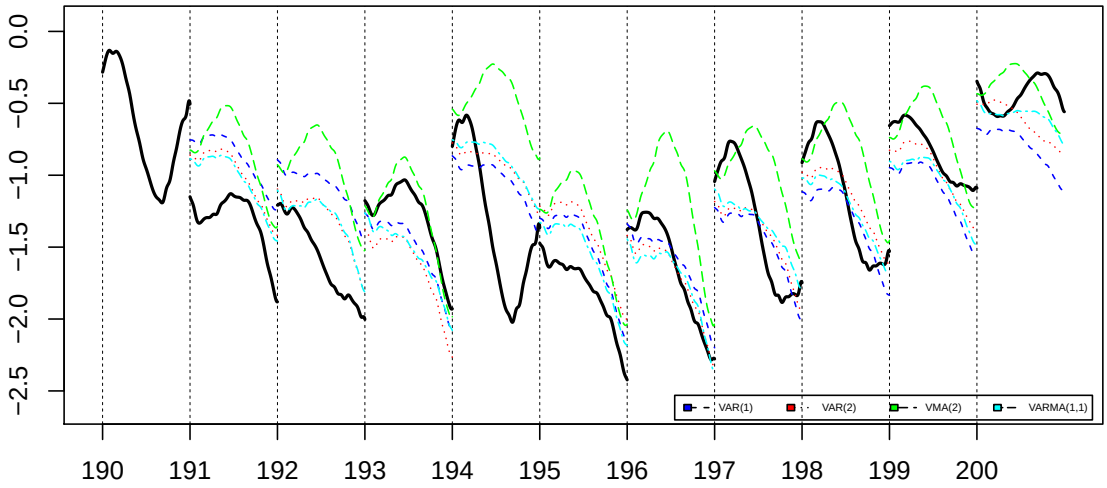

Figure 5.2: Simulation of functional ARMA(1,1) with $\|\Phi\|_{\mathcal{S}} = 0.9$ and $\|\Theta\|_{\mathcal{S}} = 0.9$. Functional predictors based on based on vector ARMA models with different orders.

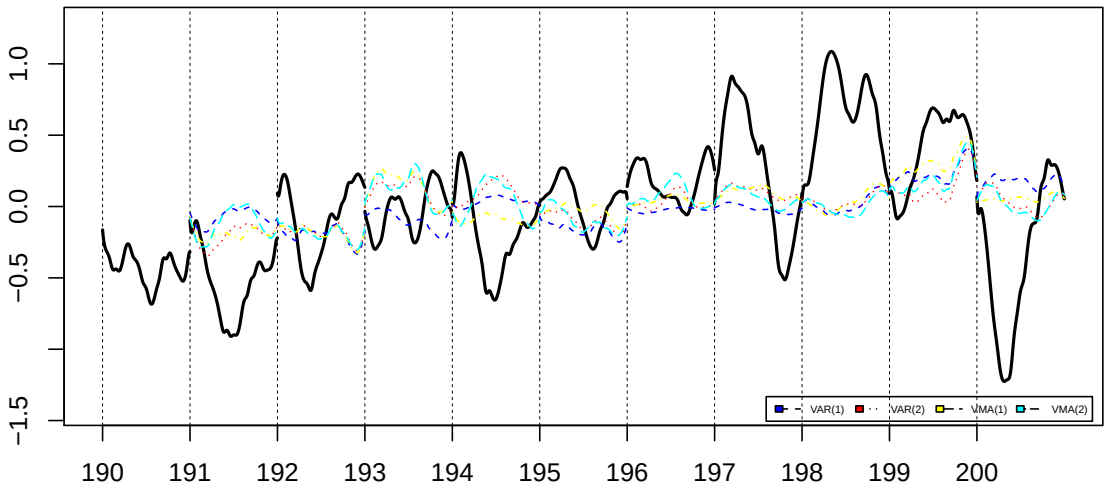

Figure 5.3: Simulation of functional MA(1) with $\|\Theta\|_{\mathcal{S}} = 0.6$. Functional predictors based on based on vector ARMA models with different orders.

Chapter 6

Real data analysis

In this chapter we analyse the highway traffic data provided by Autobahndirektion Südbayern. Our goal will be to forecast the highway velocity.

Firstly we will describe the dataset and process the data, e.g. excluding outliers, smoothing, etc. Then we will conduct Portmanteau test and stationarity test introduced in Chapter 3, to check whether the data have been transformed properly. Finally, under the framework of Chapter 5, we will fit different ARMA$(p, 1)$ models to the data and compute the (one-step) functional predictors based on these models, and then compare them with several univariate predictors.

6.1 Description and transformation of the dataset

6.1.1 Data description

Our dataset describes the traffic conditions (in one direction) which include car types, velocity and traffic volume, etc. The data come from a fixed measure point on a highway in Southern Bavaria, Germany. The traffic conditions on three lanes in one direction are recored every one minute (1440 minutes in one day) during the period 1/1/2014 00:00 to 30/6/2014 23:59.

We are interested in the car velocity and the traffic volume data. We denote these two variables by $S_{ln}(t_j)$ and $C_{ln}(t_j)$, where $S_{ln}(t_j)$ ($C_{ln}(t_j)$) represents the car velocity (traffic volume) measured at time point t_j on the nth day on Lane l, $\quad l = 1, 2, 3,\ n = 1, \ldots, N$.

To simplify the computation, we consider the three lanes as one lane. We take the sum of the traffic volume on three lanes by

$$C_n(t_j) := \sum_{l=1}^{3} C_{ln}(t_j), \quad j = 1, \ldots, 1440,\ n = 1, \ldots, N, \tag{6.1}$$

then we treat $\left\{ (C_n(t_1), \ldots, C_n(t_{1440}))^T, n = 1, \ldots, N \right\}$ as our vector traffic volume observations. And we take the weighted average velcocity on three lanes by

$$S_n(t_j) := \sum_{l=1}^{3} \frac{C_{ln}(t_j) \cdot S_{ln}(t_j)}{C_{1n}(t_j) + C_{2n}(t_j) + C_{3n}(t_j)}, \quad j = 1, \ldots, 1440,\ n = 1, \ldots, N, \tag{6.2}$$

then we treat $\left\{(S_n(t_1),\ldots,S_n(t_{1440}))^T, n=1,\ldots,N\right\}$ as our vector velocity observations.

We have seen the figures of the vector velocity S_n in (6.1) and the traffic volume data C_n in (6.2) on several selected weeks in Section 1.1 (see Figure 1.3 and Figure 1.4).

6.1.2 Data processing

After we check through the whole dataset, we find out the following three kinds of data which need to be further processed.

(1) As can be seen from Figure 6.1, many (in fact, 159) consecutive "0" records of velocity and traffic volume appear simultaneously between about 0:00-4:00 on 13/03. This may be due to the technical fault of the surveillance camera. In this case, both the velocity and traffic volume records need to be processed.

(2) Except on the commuting time (we set it to 8:00-10:00 and 16:00-20:00), we can still find some "0" velocity and traffic volume records which apprear simultaneously. For example, see the "0" velocity and volume records at about 23:00 on 10/04 in Figure 6.1. In this case, it means there are no cars passing the measure point at the measure time. Of course, this kind of "0" records reflect the real traffic condition indeed. But, when we collect the velocity records as observations to support the work of forecasting, are this kind of "0" velocity records supposed to be processed, or just to remain?

From the upper panel of Figure 6.1, one can see that, there is just one "0" velocity record at about 23:00 on 10/04. But all the other velocity records in the neighborhood are very high, which implies that the traffic condition was quite good at that time. If we really treat such "0" velocity records as "0", it give rise to a low-value velocity predictor, which will misguide the driver about the highway condition. Thus we will process this kind of "0" velocity records.

But on the commuting time, if the the velocity record is zero while the corresponding traffic volume record is not "0", it implies a traffic congestion at that time. And in this case, we will not do anything about the "0" velocity records, since a low-value velocity predictor is reasonable and expected.

(3) Except on the commuting time (8:00-10:00 and 16:00-20:00), the velocity records below $50km/h$ appear quite rarely in the whole dataset. Meanwhile, the other velocity records in the neighborhood are far above $50km/h$ (e.g see the volocity records between 12:00 to 14:00 on 24/3 in Figure 6.1). Similar to (2), we treat this kind of data as outliers. Again, we will not do anything about the low velocity records on the commuting time (e.g see the records between 16:00-20:00 on 28/5).

Now we show how we process the three kinds of outliers mentioned above.

We divide $t_1, \ldots, t_{1440}$ into 288 groups, $U_1, \ldots, U_{288}$. For each $i \in \{1, \ldots, 288\}$, $U_i = \{t_{5*i-4}, \ldots, t_{5*i}\}$.

- For the second and the third kind of outliers (non-consecutive "0" and low velocity records except on the commuting time), we do the following thing.

  For day n, if we find $S_n(t_j) < 50$ (it includes $S_n(t_j) = 0$), where $t_j \in U_i$ for some i, then we replace the value of $S_n(t_j)$ with the average velocity records in U_i which are above 50, i.e.

$$S_n(t_j) := \frac{\sum_{k=5*i-4, k\neq j}^{5*i} S_n(t_k) 1_{\{S_n(t_k)\geq 50\}}}{\sum_{k=5*i-4, k\neq j}^{5*i} 1_{\{S_n(t_k)\geq 50\}}}. \tag{6.3}$$

  In comparison to Figure 6.1, Figure 6.2 depict the records after data processing. The low velocity (and non-consecutive 0) records, e.g 12:00-14:00 on 24/03 and at about 23:00 on 10/04, are replaced with higher velocity records (>50 km/h).

- For the first kind of data (consecutive "0" velocity and traffic volume data due to the technical fault), we do the following thing.

  For day n, we suppose U_i is the first interval in which all five velocity and taffic volume records are 0. Then we look for the smallest $m > 5i$ such that $S_n(t_m) \geq 50$ and $C_n(t_m) > 0$. Then we replace the value of $S_n(t_{5i-4}), \ldots, S_n(t_{m-1})$ and $C_n(t_{5i-4}), \ldots, C_n(t_{m-1})$ by linear interpolation, i.e.

$$S_n(t_j) := S_n(t_{5i-5}) + \frac{S_n(t_m) - S_n(t_{5i-5})}{t_m - t_{5i-5}} (t_j - t_{5i-5}), \tag{6.4}$$

$$C_n(t_j) := C_n(t_{5i-5}) + \frac{C_n(t_m) - C_n(t_{5i-5})}{t_m - t_{5i-5}} (t_j - t_{5i-5}), \quad t_j \in U_i. \tag{6.5}$$

  As can be seen from the upper panel of Figure 6.2, the consecutive "0" velocity records (see 0:00-4:00 on 13/03 in Figure 6.1) are replaced with two clear slanting lines, which are derived from the linear interpolation in (6.4).

In Figure 6.3, we plot the velocity data processed in the way mentioned above, on the same weeks as in Figure 1.3. Comparing these two figures, we can see that, many "0" records which are not on the commuting time, e.g see 16/4, 30/4, etc in, disappear. Meanwhile, those low speed records which are on the commuting time, e.g. see 18/6 and 26/6, remain.

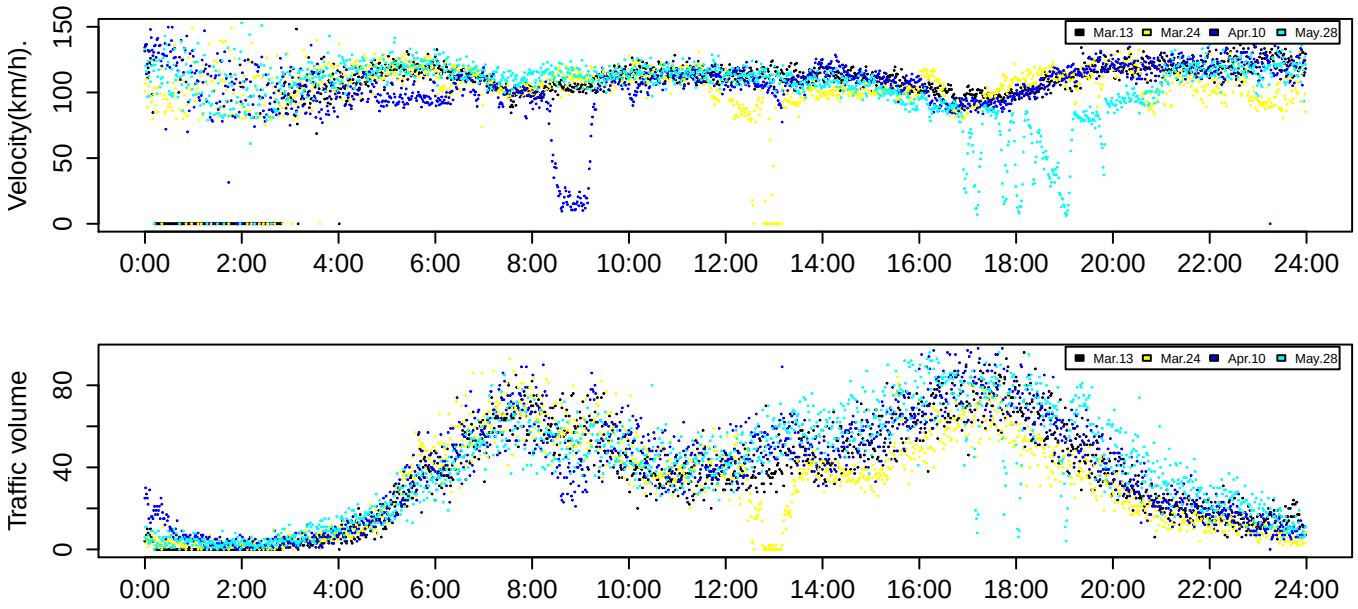

Figure 6.1: Velocity and traffic volume records on 4 selected days before data processing.
Source: Autobahndirektion Südbayern

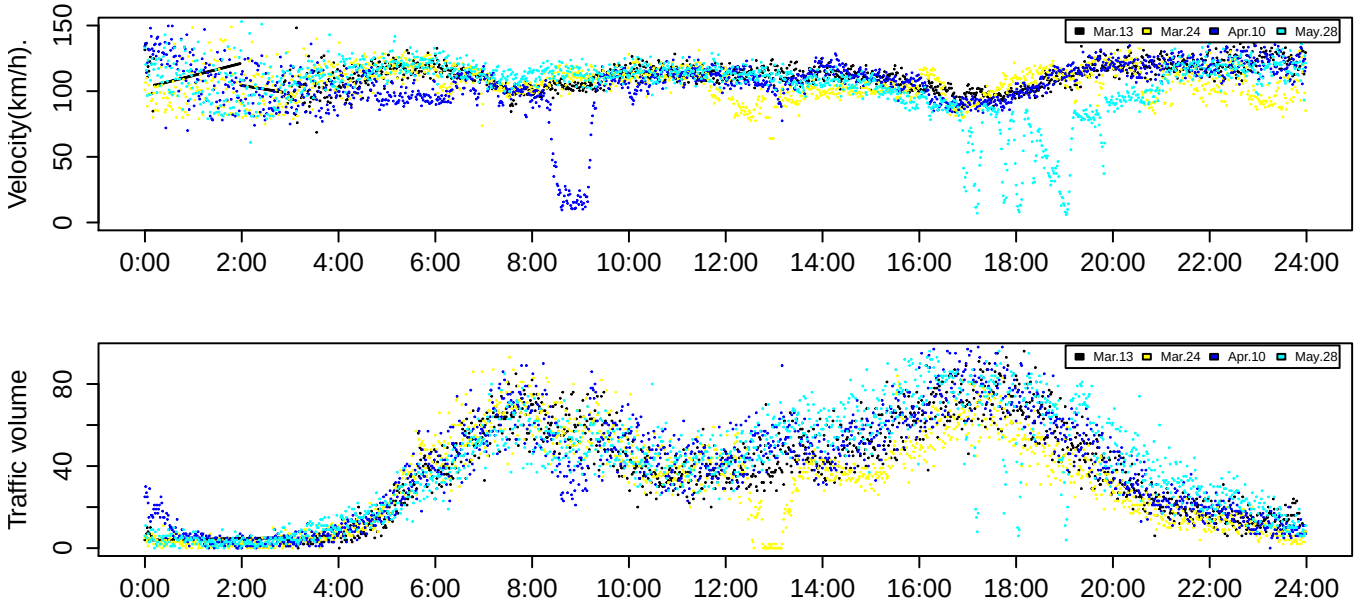

Figure 6.2: Velocity and traffic volume records on 4 selected days after data processing.
Source: Autobahndirektion Südbayern

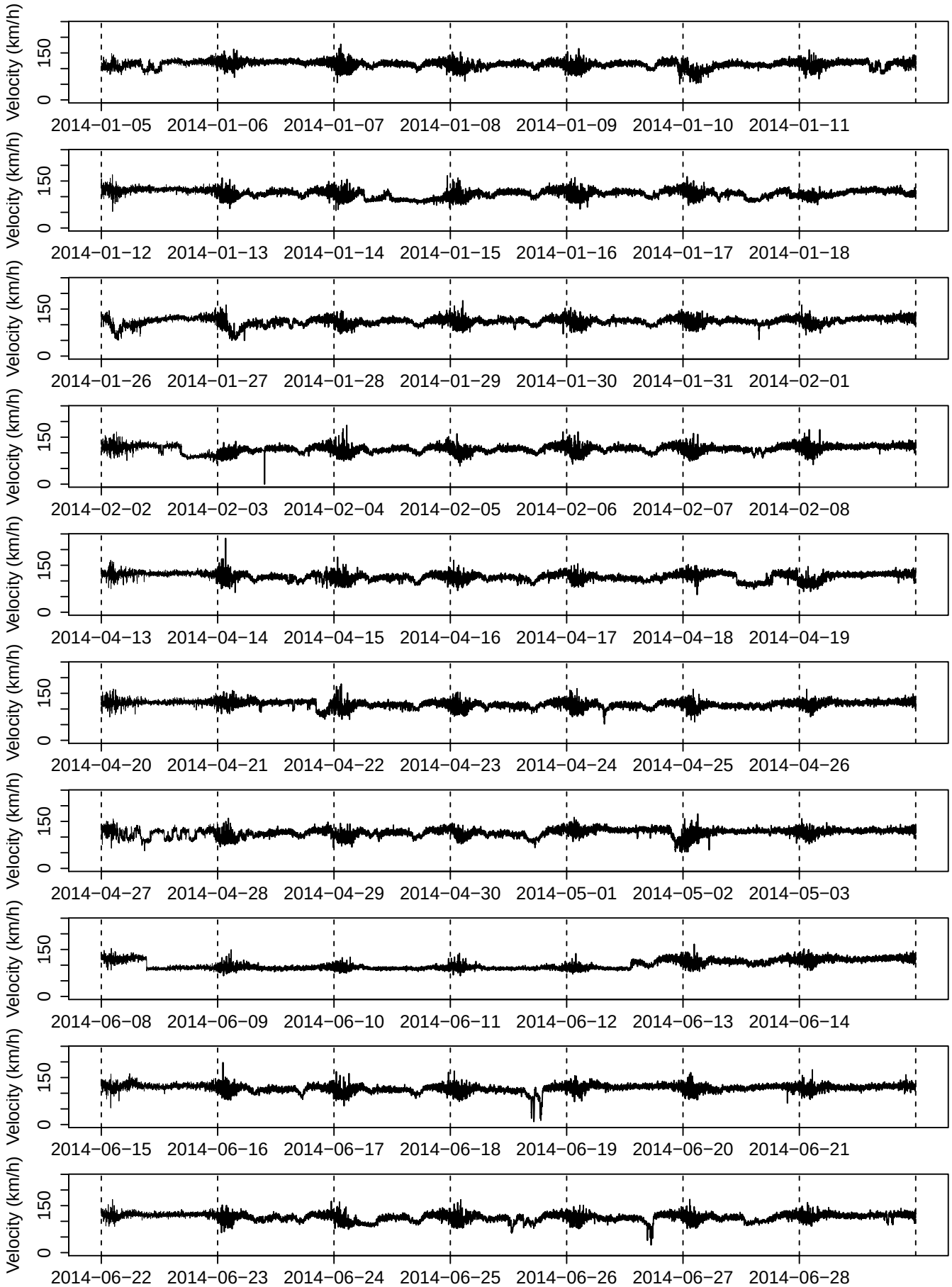

Figure 6.3: Processed vector velocity data on the same weeks as in Figure 1.3. The data are processed in the way introduced in Section 6.1.2. **Source:** Autobahndirektion Südbayern

6.1.3 From discrete to functional data

Remark 10.

- *There are 181 days from 1/1/2014 to 30/6/2014. But in our dataset, many records are lost on 20/1, 30/3 and 7/5. Thus we exclude these three days out of our dataset. Thus the number of the observations in the group* ***All days*** *is* $N = 181 - 3 = 178$.

- *1/1(Wednesday), 6/1(Monday), 18/4(Friday), 20/4(Sunday), 21/4(Monday), 1/5 (Thursday), 29/5(Thursday), 8/6(Sunday), 9/6(Monday) and 19/6 (Thursday) are 10 public holidays during 1/1/2014 to 30/6/2014. In Section 1.1, we have seen from Figure 1.3 and Figure 1.4 that, the velocity and volume curves on the holidays falling on weekdays, e.g. 18/4(Friday) and 9/6(Monday), differ apparently from those on the other corresponding "normal" weekdays. So we exclude these days out of the group* ***Workingdays*** *and categorize them into the group* ***Holidays****. In contrast, we treat the holidays falling on Sundays, e.g. 8/6(Sunday), as normal Sundays instead of categorizing them into the group* ***Holidays****.*

  Thus, we have 8 observations in the group ***Holidays****: 1/1(Wednesday), 6/1(Monday), 18/4(Friday), 21/4(Monday), 1/5(Thursday), 29/5(Thursday), 09/6(Monday), 19/6(Thursday).*

- *Except the 8 holidays mentioned above, there are 51 weekends from 1/1/2014 to 30/6/2014. Thus, we have* $178 - 8 - 51 = 119$ *observations in the group* ***Workingdays****.*

- *In Section 6.2 and 6.3, the hypothesis tests, the model fitting and the predictions algorithm will be conducted to the functional data in the group of* ***All days*** *and* ***Workingdays*** *respectively.*

Up to now, we have N 1440-dimensional velocity records

$$\left\{ (S_n(t_1), \ldots, S_n(t_{1440}))^T, n = 1, \ldots, N \right\},$$

which have been processed (excluding outliers) in Section 6.1.2 ($N = 178$ in **All days** and $N = 119$ in **Workingdays**). They still belong to the class of *high-dimensional* data. In this subsection we will transform these vector data to functional data. We do the following things.

(i) Recall that in Section 6.1.2, we divide $t_1, \ldots, t_{1440}$ into 288 groups $U_1, \ldots, U_{288}$. Now we define $T_1, \ldots, T_{288}$ as the initial time point of $U_1, \ldots, U_{288}$ respectively, i.e.

$$T_i = t_{5*i-4}, \quad i = 1, \ldots, 288.$$

Firstly we reduce the 1440-dimensional data $\left\{ (S_n(t_1), \ldots, S_n(t_{1440}))^T, n = 1, \ldots, N \right\}$ to 288-dimensional by computing the weighted average velocity every 5 minutes in each day n, i.e.

$$X_n(T_i) := \frac{\sum_{k=0}^{4} S_n(t_{5i-k}) C_n(t_{5i-k})}{\sum_{k=0}^{4} C_n(t_{5i-k})}, \quad i = 1, \ldots, 288. \tag{6.6}$$

Then $\left\{(X_n(T_1),\ldots,X_n(T_{288}))^T, n=1,\ldots,N\right\}$ are 288-dimensional velocity data.

(ii) **Centering the data:** Usually speaking, centering the data implies deducting the average value of all the observations from each observation. But we will do it in another way. Figure 6.4 depicts the mean velocity curve of the 178 observations in **All days** (real black line), and Mondays, ..., Sundays and holidays respectively. As can be seen that, the mean curve of Saturdays, Sundays and holidays are significantly different from those on working days. Thus, instead of deducting the mean velocity of all $N = 178$ days from each observation, we deduct individual weekday mean, i.e. we deduct the the mean (velocity) of all Mondays from each Monday observations, etc.

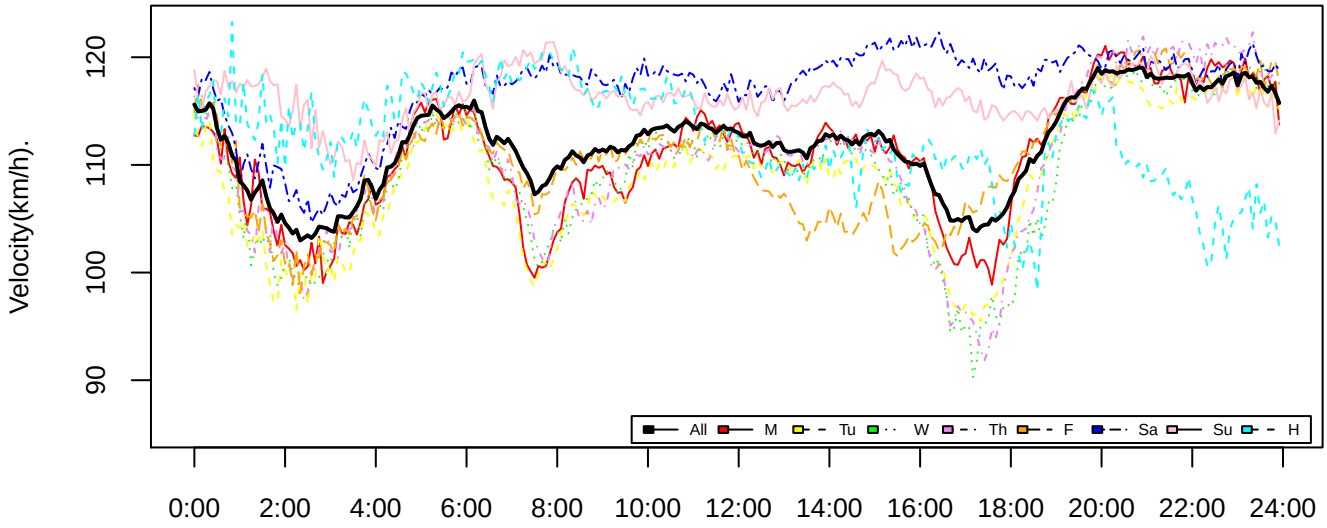

Figure 6.4: Mean velocity curves of the 178 observations in **All days** and Mondays, Tuesday, Wednesdays, Thursdays, Fridays, Saturdays, Sundays and holidays from 01/01/2014 to 30/06/2014. **Source:** Autobahndirektion Südbayern

We center the data since we want to make the data more stationary, i.e. we hope the curves get closer (which implies the mean and the variance more stable). In Figure 6.5, 6.6 and 6.7, we can compare the raw data on 50 workingdays and the corresponding centered data (deducted with the overall mean of all workingdays and the individual weekday mean respectively).

The scales of the y-axis in Figure 6.6 and Figure 6.7 are the same. As can be seen that, both of the centered velocity curves in Figure 6.6 and 6.7 get closer to 0 and look much more "level" than the non-centered curves in Figure 6.5. If we only compare the curves in 6.6 and 6.7, they do not look much different. This is because

the traffic conditions on workingdays are not much different form each other, e.g see Figure 6.4. But we note that, at some extreme value points, the curves in Figure 6.7 (deducted with individual weekday mean) get a bit closer to 0 than those in Figure 6.6, see e.g. the value of 15km/h at 8:00 and the -60km/h at 18:00.

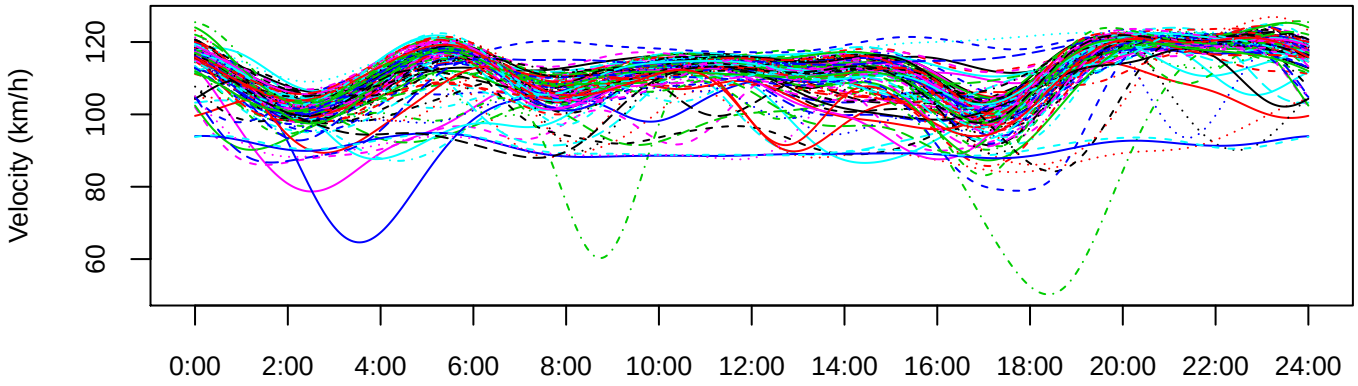

Figure 6.5: 119 observations in the group **Workingdays** are smoothed with 29 Fourier basis functions. Depicted are the functional velocity data on 50 workingdays (11/4/2014 to 30/6/2014). **Source:** Autobahndirektion Südbayern

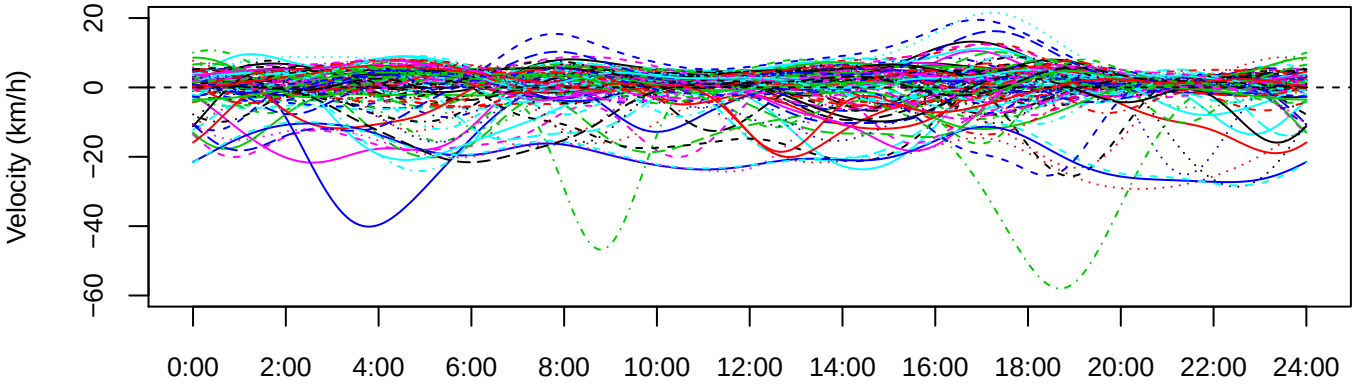

Figure 6.6: Functional velocity data (deducted with the overall mean of all workingdays) on the same days as in Figure 6.5. **Source:** Autobahndirektion Südbayern

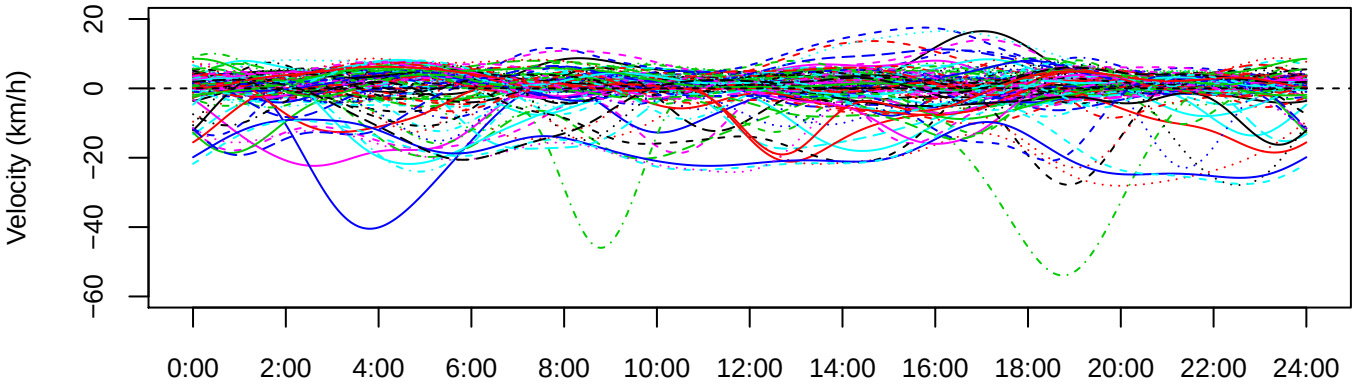

Figure 6.7: Functional velocity data (deducted with the individual weekday mean) on the same days as in Figure 6.5. **Source:** Autobahndirektion Südbayern

(iii) To smooth the vector data $\left\{(X_n(T_1),\ldots,X_n(T_{288}))^T, n=1,\ldots,N\right\}$ in (6.6), we need to specify a series of basis functions. We denote these basis functions by $\phi_1,\ldots,\phi_K$. For each vector observation $(X_n(T_1)\ldots X_n(T_{288}))'$, we hope it can be smoothed by

$$Y_n(t) := \sum_{k=1}^{K} c_{nk}\phi_k(t), \quad \forall t \in [0,1]. \tag{6.7}$$

For fixed number K of the basis functions, the simplest criterion to choose the coefficients (c_{nk}) is to minimize the *SSE*, where

$$\begin{aligned} SSE :&= \sum_{n=1}^{N}\sum_{j=1}^{288}\left(X_n(t_j)-Y_n(t_j)\right)^2 \\ &= \sum_{n=1}^{N}\sum_{j=1}^{288}\left(X_n(t_j)-\sum_{k=1}^{K}c_{nk}\phi_k(t_j)\right)^2. \end{aligned} \tag{6.8}$$

By assuming the residuals (ε_j) in the model

$$X_n(t_j) = Y_n(t_j) + \varepsilon_j \tag{6.9}$$

are independent and normally distributed (with mean 0 and constant mean), the least-squares estimate of the coefficients (c_{nk}) are

$$\begin{pmatrix} c_{11} & \ldots & c_{N1} \\ \vdots & & \vdots \\ c_{1K} & \ldots & c_{NK} \end{pmatrix} := \hat{\mathbf{c}} = \left(\mathbf{\Phi}^T\mathbf{\Phi}\right)^{-1}\mathbf{\Phi}^T\mathbf{X}, \tag{6.10}$$

where

$$\mathbf{X} := \begin{pmatrix} X_1(t_1) & \ldots & X_N(t_1) \\ \vdots & & \vdots \\ X_1(t_{288}) & \ldots & X_N(t_{288}) \end{pmatrix}, \tag{6.11}$$

and

$$\mathbf{\Phi} := \begin{pmatrix} \phi_1(t_1) & \ldots & \phi_K(t_1) \\ \vdots & \vdots & \vdots \\ \phi_1(t_{288}) & \ldots & \phi_K(t_{288}) \end{pmatrix}. \tag{6.12}$$

We call the method of smoothing above *smoothing by regression analysis.*

Remark 11. *Besides the method of* smoothing by regression analysis, smoothing with roughness penalties *is also a usually used smoothing method. The motivation is to obtain a smooth derivative rather than just a smooth function. The* penalty *is often defined as*

$$PEN_2(y_n) := \int [D^2Y_n(t)]^2 \mathrm{d}t, \tag{6.13}$$

where $D^2Y_n(t)$ is the second derivative of Y_n defined in (6.7). Of course the penalty can be defined in various other ways, see Chapter 5 of Ramsay et al. [2009]. The coefficients is determined by minizing

$$PENSSE := \sum_{n=1}^{N}\sum_{j=1}^{288}(X_n(t_j) - Y_n(t_j)^2 + \lambda \int [D^2Y_n(t)]^2 \mathrm{d}t. \tag{6.14}$$

The smoothing parameter λ *specifies the emphasis on the second term in (6.14). If λ is 0, then it is exactly the same as* smoothing by regression analysis*. All in all, with the method of* smoothing with roughness penalties*, we can obtain more smooth functions, but at the cost of losing some information (bigger error). More details see Chapter 5 of Ramsay et al. [2009]. In R, the command smooth.basis() smooths the data by regression analysis, if the argument* fdParobj *is a functional basis object with the class "basisfd". But it can also deal with smoothing with roughness penalties by choosing the class of "fdPar" in the argument* fdParobj*, and setting the value of the argument* dfscale*. Meanwhile, the command* Data2fd*() smooth the data with roughness penalties, and the default value of the* smoothing parameter λ *is* $3 * 10^{-8}$.

Now we go back to the issue of choosing basis functions. In our data anlysis, we used Fourier and Bspline basis functions to smooth the vector data respectively. In the following we will explain why we choose Fourier rather than Bspline from two aspects.

- For both Fourier and Bspline, we choose different number M of the basis functions and compute the mean absolute error (MAE) defined in (5.6), between the functional data and the vector data. By smoothing we used the R command *Data2fd*, i.e. smoothing with roughness penalties and the *smoothing parameter* λ is $3 * 10^{-8}$. In Figure 6.8 (a) and (b) are depicted the MAE between the 178 vector data and the corresponding functional data, by choosing different number M of Fourier and Bspline basis functions respectively. And in Figure 6.8 (c) and (d) are depicted the results of 119 **Workingdays**.

  As can be seen from Figure 6.8 (a) and (c), i.e. by choosing Fourier basis, the error curves become level from about $M = 29$. In comparison, in Figure 6.8 (b) and (d), i.e. by choosing Bspline basis, the error decreases as M increases. Merely from the angle of MAE, Bsplines is a little bit better than Fourier. But we must note that, the errors are relatively small in comparison to the velocity value.

  In the following we evaluate Fourier and Bspline from the angle of dimensional reduction.

- Now we choose 29 Fourier and Bspline basis functions to smooth the vector data in the two datasets (**All days** and **Workingdays**) respectively. We choose $M = 29$ is not because of some specific "gold" standard. One side is that the error curves in Figure 6.8 (a) and (c) become level from $M = 29$, and the value

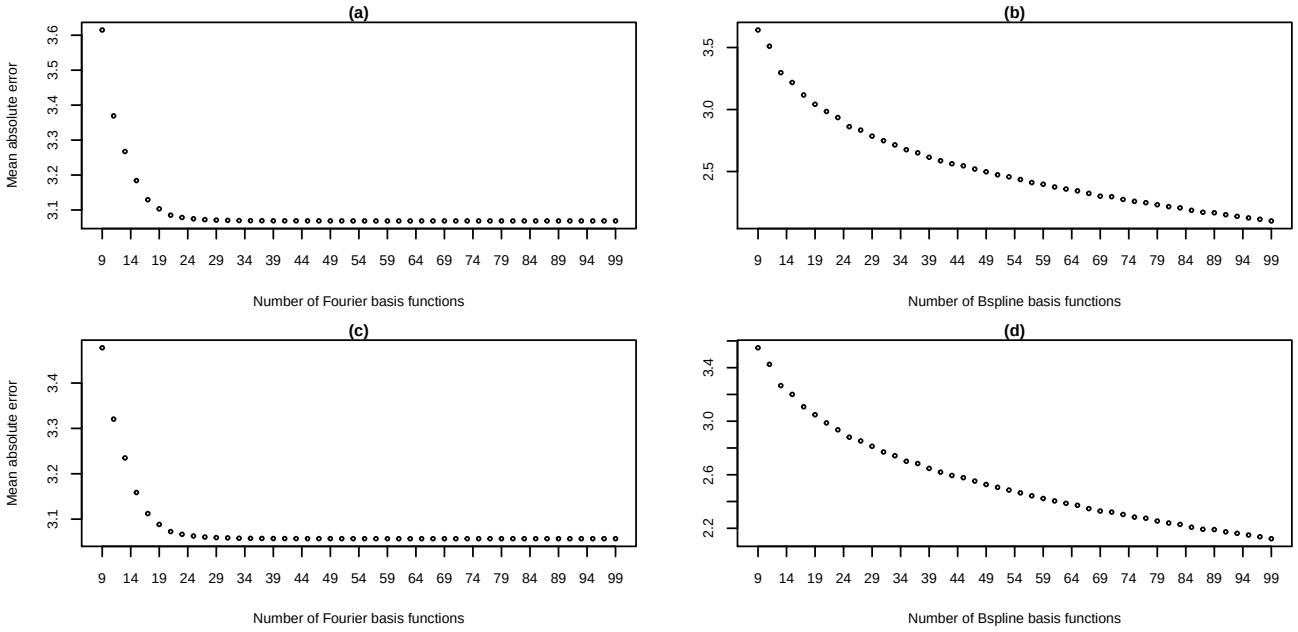

Figure 6.8: (a) and (b) are mean absolute errors (MAE) between the 178 vector data in **All days** and the corresponding functional data, by choosing different number of Fourier and Bspline basis functions respectively (with R command *Data2fd*). (c) and (d) are the results for **Workingdays**.

of errors (including Bspline) are small in comparison to the velocity value, so it does not deserve to choose more basis functions. On the other side, in many real case studies, for about 200 observations, 29 basis functions are usually enough.

Now we apply the CPV method to the functional data (smoothed by 29 Fourier and Bspline respectively), to determine the number d of *FPC's* (see (2.21) in Section 2.2.3). In Figure 2.2, we have shown that, for the 178 functional data in **All days**, under the CPV criterion 80% (see (2.21)), the number d of the FPC's should be $d = 4$.

In the following table, Table 6.1, we compare the results (number d of FPC's) by changing the CPV criterion.

As can be seen from Table 6.1, in both datasets, and under each CPV criterion, the number of FPC's by choosing Fourier is always smaller than that by choosing Bspline. It is an important advantage of Fourier. We always want to use fewer FPC's to explain more data variability, which makes the computation easier and faster.

Thus, from the results in Table 6.1, we decide to choose $M = 29$ Fourier basis functions to smooth the vector data in the group of **All days** and **Workingdays**. And the CPV criterion is set to 80%. Thus for the both datasets, the number d of the empirical FPC's should all be 4 according to Table 6.1.

Figure 6.9 depicts the 4 empirical eigenfunctions of the 178 functional data

	CPV criterion	0.8	0.85	0.9
All days ($N = 178$)	Number d of FPC's (with F-basis)	4	5	7
	Number d of FPC's (with B-basis)	6	8	11
Workingdays ($N = 119$)	Number d of FPC's (with F-basis)	4	5	7
	Number d of FPC's (with B-basis)	6	8	11

Table 6.1: Application of CPV method to the functional data in the group of **All days** and **Workingdays**, under different CPV criteria. In both datasets, the functional data are derived from 29 Fourier and Bspline basis functions (with R command *Data2fd*) respectively.

on **All days**, while Figure 6.10 depict the 4 empirical eigenfunctions of the 121 data on **Workingdays**. And the proportion of the total data variability explained by each empirical eigenfunction is listed in the captions of Figure 6.9 and Figure 6.10.

As can be seen from Figure 6.9 and Figure 6.10, the eigenfunctions ν_1, which explains 45% and 40% of the data variability respectively, look quite level. It is in accordance with the fact that, most of the velocity curves look level (see Figure 6.3).

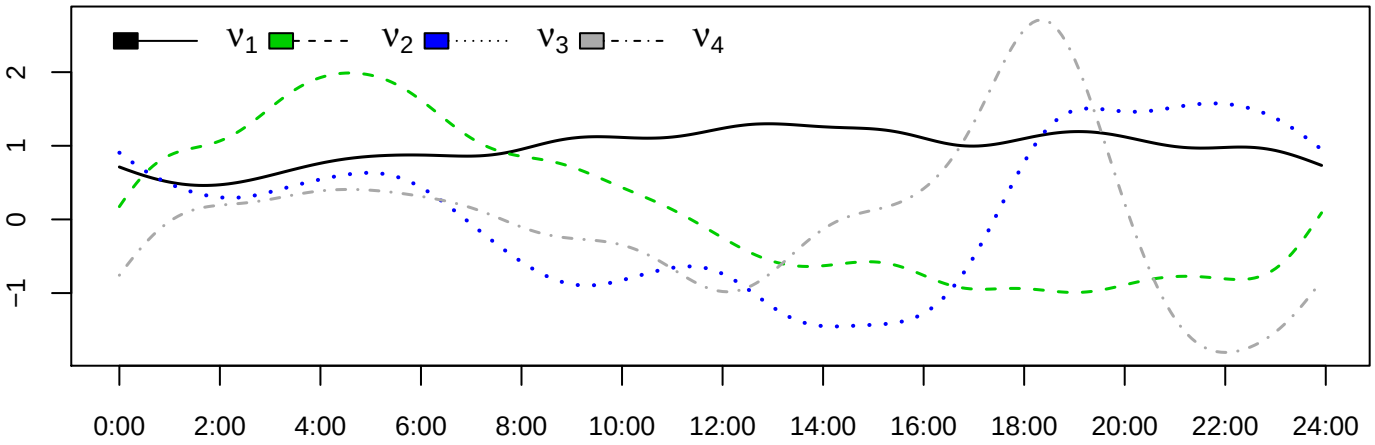

Figure 6.9: Four empirical eigenfunctions of the 178 functional data on **All days**. The CPV criterion is 80%. The proportion of the total data variability explained by $\nu_1, \nu_2, \nu_3, \nu_4$ is 45%, 20%, 9.5% and 8% respectively. **Source:** Autobahndirektion Südbayern

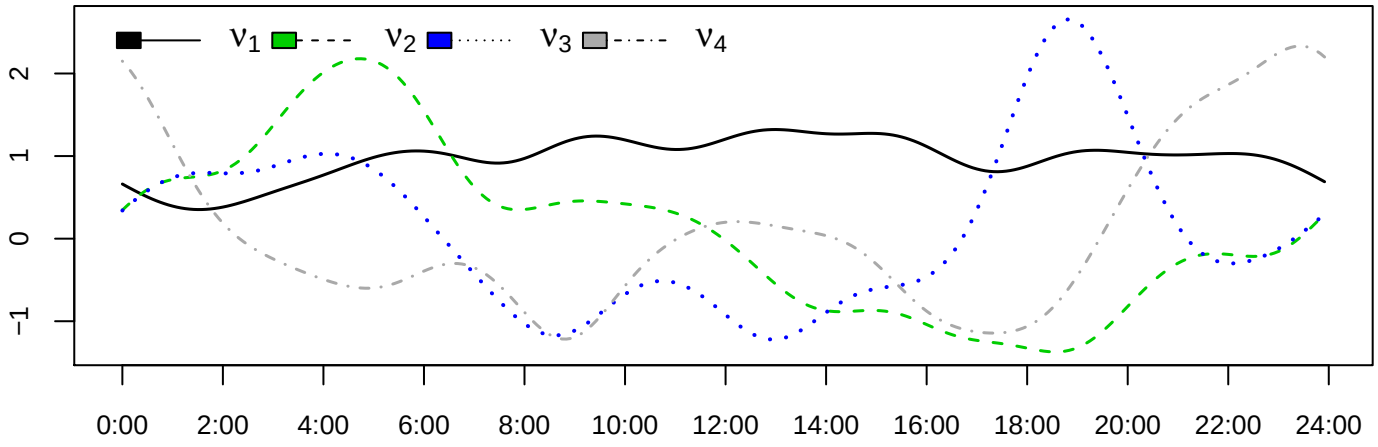

Figure 6.10: Four empirical eigenfunctions of the 119 functional data on **Working-days**. The CPV criterion is 80%. The proportion of the total data variability explained by $\nu_1, \nu_2, \nu_3, \nu_4$ is 40%, 22%, 11% and 8% respectively. **Source:** Autobahndirektion Südbayern

Remark 12. *Recall that in Remark 11, we mentioned that the R command* Data2fd() *smooths the vector data with roughness penalty and the smoothing parameter* $\lambda = 3 * 10^{-8}$ *(by default setting), while* smooth.basis() *smooths the data by regression analysis(if the argument* fdParobj *is a functional basis object with the class "basisfd"). Now we compare these two commands in the application to our dataset.*

Figure 6.11 is a similar plot to Figure 6.8, but the data are smoothed with the R command smooth.basis(). *As expected, with the method of smoothing by regression analysis (*smooth.basis*()), the curves describe the data better (smaller MAE, compare Figure 6.11 with Figure 6.8), and MAE goes to 0 as the number of basis functions is sufficiently large. However, due its precision, the functions are not so smooth as those generated by* Data2fd(), *see Figure 6.12. In addition, more FPC's are needed under the same CPV criterion (compare Table 6.2 with Table 6.1). Due to the two facts of using* Data2fd(), *smoother curves and fewer FPC's, we choose* Data2fd() *instead of* smooth.basis().

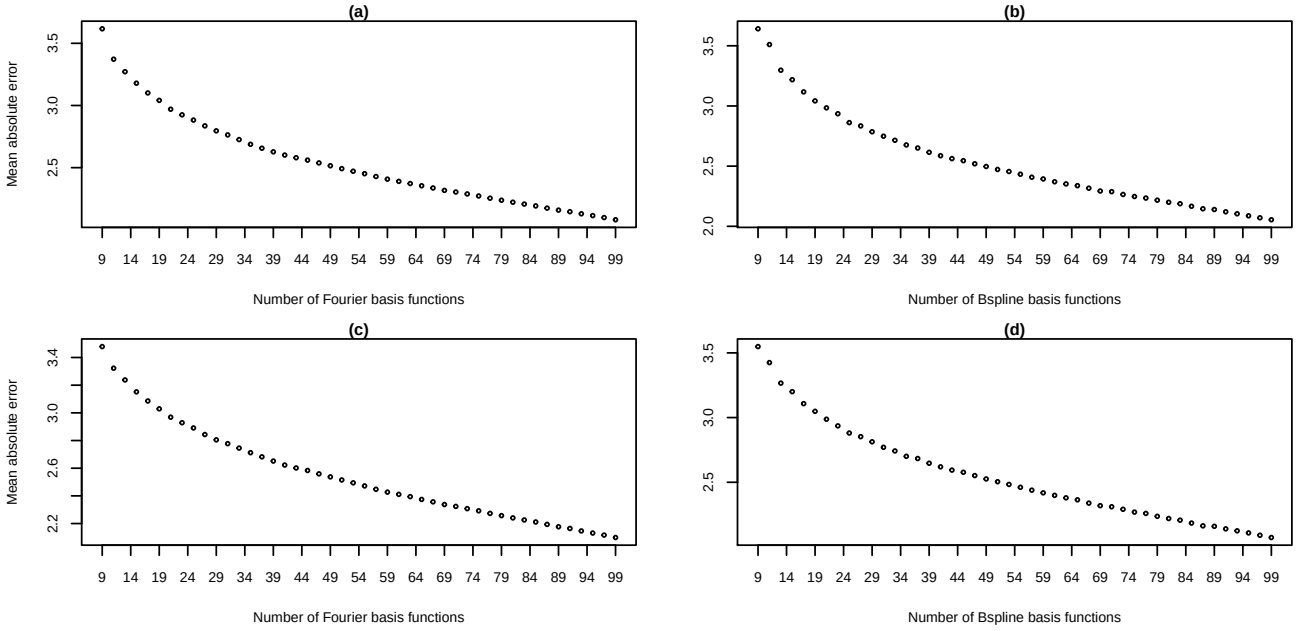

Figure 6.11: Similar plot to Figure 6.8, the data are smoothed with the R command *smooth.basis()*.

	CPV criterion	0.8	0.85	0.9
All days ($N = 178$)	Number d of FPC's (with F-basis)	6	8	11
	Number d of FPC's (with B-basis)	6	8	11
Workingdays ($N = 119$)	Number d of FPC's (with F-basis)	6	8	11
	Number d of FPC's (with B-basis)	6	8	11

Table 6.2: Similar table to Table 6.1, data are smoothed with the R command *smooth.basis()*.

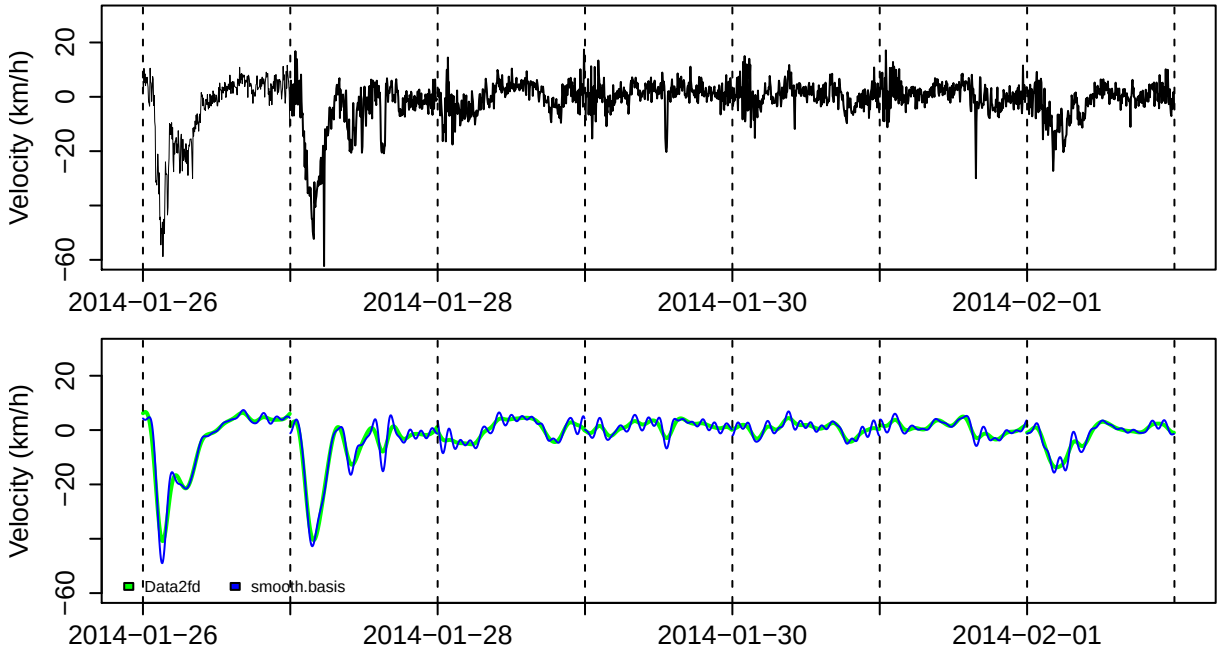

Figure 6.12: In the upper panel are the centered raw velocity data on one week. In the lower panel are the corresponding functional data generated with the R command *Data2fd()* and *smooth.basis()* respectively.

6.2 Hypothesis Test

In this section, we will conduct the Portmanteau test (H_0: data are i.i.d) and the stationarity test (H_0: data have the functional moving average forms) introduced in Chapter 3, to the **centered** functional velocity data in the group of **All days** and **Workingdays** respectively. We want to check, whether the data have been transformed properly. We hope that the independence can be rejected and the stationarity can not be rejected, then it is reasonable to fit ARMA model to the data (see Section 6.3).

For both of the datasets, the CPV criterion is set to 80%, then the number of FPC's is $d = 4$ (see Table 6.1).

- **Test on the data in *All days***

Lag (H)	1	2
P-value	$6.90 * 10^{-11}$	$1.38 * 10^{-7}$

Table 6.3: Portmanteau test on 178 centered functional velocity data in **All days**.

For lags $H = 1$ and $H = 2$, the independence can be rejected ever for a significance level as small as $\alpha = 10^{-6}$.

Test Statistics	T_N	M_N	$T_N^*(d)$	$M_N^*(d)$
P-value	0.0966	0.2239	0.0917	0.2176

Table 6.4: Stationarity test on on 178 centered functional velocity data in **All days**.

The form of the test statistics T_N, M_N, $T_N^*(d)$ and $M_N^*(d)$ can be found in (3.15), (3.16), (3.18) and (3.20). The stationarity can not be rejected ever for a significance level as small as $\alpha = 0.05$.

- **Test on the data in *Workingdays***

Lag (H)	1	2
P-value	$3.72 * 10^{-7}$	$9.85 * 10^{-2}$

Table 6.5: Portmanteau test on 119 centered functional velocity data in **Workingdays**.

For lags $H = 1$ and $H = 2$, the independence can be rejected ever for a significance level as small as $\alpha = 0.01$.

Test Statistics	T_N	M_N	$T_N^*(d)$	$M_N^*(d)$
P-value	0.1038	0.1935	0.0964	0.1766

Table 6.6: Stationarity test on 119 centered functional velocity data in **Workingdays**.

The stationarity can not be rejected ever for a significance level as small as $\alpha = 0.05$.

Combining the results in Table 6.3 to Table 6.6, it is reasonable to fit ARMA model to the data.

6.3 Model fitting and prediction

Since for both of the datasets (**All days** and **Workingdays**), the number of FPC's is chosen by $d = 4$ (under the CPV criterion 80%). Thus each functional observation $X_1, \ldots, X_N$ ($N = 178$ in **All days** and $N = 119$ in **Workingdays**) are truncated to 4-dimensional, i.e.

$$X_n \approx X_n^e := \sum_{l=1}^{d} \langle X_n, \nu_l^e \rangle \, \nu_l^e, \quad n = 1, \ldots, N. \tag{6.15}$$

And the vector form of (6.15) is

$$\mathbf{X}_n^e := \left(\langle X_n, \nu_1^e \rangle \ldots \langle X_n, \nu_4^e \rangle\right)^T, \quad n = 1, \ldots, N. \tag{6.16}$$

In this section, we will fit different vector ARMA$(p, 1)$ models to the truncated vector observations (6.16) in **All days** and **Workingdays** respectively. We will compare the goodness of fit among these vector models. Then we will compute the one-step functional predictors based on these vector models, and compare them with several univariate predictors. The framework of this section is exactly the same as that of the simulation study in Chapter 5.

- **Study of the 178 observations in All days**

  In our data analysis, we found that the vector ARMA$(1, 1)$ does not perform well both in the model fitting and the prediction of the 178 observations in **All days**. Thus in the following two tables, Table 6.9 and Table 6.10, we do not list the results of ARAM(1,1).

  In Table 6.8 and Table 6.10, "Mean", "Drift" and "ES" are the abbreviations of "Mean method", "Drfit method" and "Exponential smoothing" introduced in Section 5.1.3. And the prediction errors in Table 6.8 and Table 6.10 are the "distance" between the predictors and the truncated funtional observations in (6.15). The definitions of *RMSE* and *MAE* see (5.9) and (5.10).

Model fit	VAR(1)	VAR(2)	VMA(1)	VMA(2)
AIC	7.72	7.69	**7.18**	7.72
BIC	**8.02**	8.27	8.09	8.31

Table 6.7: Goodness of fit of different vector models to the 178 observations in **All days**.

Model fit	VAR(1)	VAR(2)	VMA(1)	VMA(2)	Mean	Drift	ES
RMSE	5.67	5.53	5.74	6.05	**5.37**	7.70	5.46
MAE	5.05	4.95	5.12	5.40	**4.73**	6.77	4.76

Table 6.8: Average prediction errors of the predictors for the last 10 observations in the group **All days**.

  As can be seen from Table 6.7, VAR(1) and VMA(1) fit the data better than the other vector models. But from the results of the prediction errors in Table 6.8, the

predictors based on VAR(1) and VMA(1) is not better than the predictor based on VAR(2), and even not better than two univariate predictors (mean method and ES).

In Figure 6.14 and Figure 6.15, we visulized the functional predictor (based on VMA(1), Figure 6.14), and the univariate predictors (Figure 6.15) of the velocity data on the last 10 days of **All days**. One can compare these two figures with Figure 6.13, which depicts the raw velocity data and the corresponding functional data on the same days.

One can see from Figure 6.13 that, the velocity curve on 29/6 deviates highly from the other curves. Meanwhile, all the (functional and univariate) predictors perform not well on this day. Besides on this day, all the predictors perform not well when the velocity curve highy fluctuates, e.g. see 24/6 and 26/6.

- **Study of the 119 observations in Workingdays**

Model fit	VAR(1)	VAR(2)	VMA(1)	VMA(2)	VARMA(1,1)
AIC	6.96	7.12	7.00	6.90	**6.87**
BIC	**7.34**	7.89	7.39	7.67	7.64

Table 6.9: Goodness of fit of different models to the 119 observations in **Workingdays**

Model fit	VAR(1)	VAR(2)	VMA(1)	VMA(2)	VARMA(1,1)	Mean	Drift	ES
RMSE	4.05	**3.87**	3.89	4.78	4.50	4.24	5.17	5.06
MAE	3.19	**3.059**	3.06	3.77	3.59	3.45	3.83	4.41

Table 6.10: Average prediction errors of the predictors for the last 10 observations in the group **Workingdays**.

As can be seen from Table 6.9, VAR(1) and VARMA(1,1) fits the data relatively better than the other vector models. But from the results of the prediction errors in Table 6.10, the predictors based on VAR(1) and VARMA(1,1) are not better than the predictor based on VAR(2).

If we compare the functional predictors and the univariate predictors, we can see from Table 6.8 that the functional predictors are generally better.

Similar to Figure 6.13, 6.14 and 6.15, Figure 6.16 depicts the raw velocity data versus the corresponding functional data on the last 10 days of **Workingdays**, while Figure 6.17 and Figure 6.18 depict the corresponding functional (VMA(1)) predictor and the univariate predictors respectively.

Similarly, we can see from Figure 6.17 and Figure 6.18 that, all the predictors perform not well when the velocity curve highy fluctuates, e.g. see 18/6, 20/6 and 24/6. But on the other "normal" days, all the predictors look not bad.

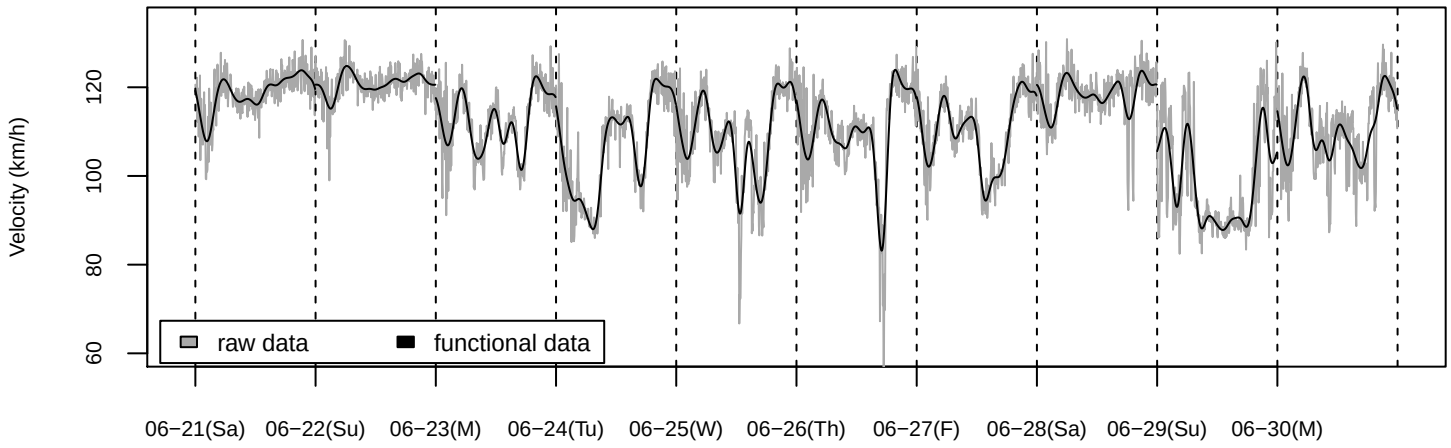

Figure 6.13: Raw velocity data and the corresponding functional data on the last 10 days of **All days**. **Source:** Autobahndirektion Südbayern

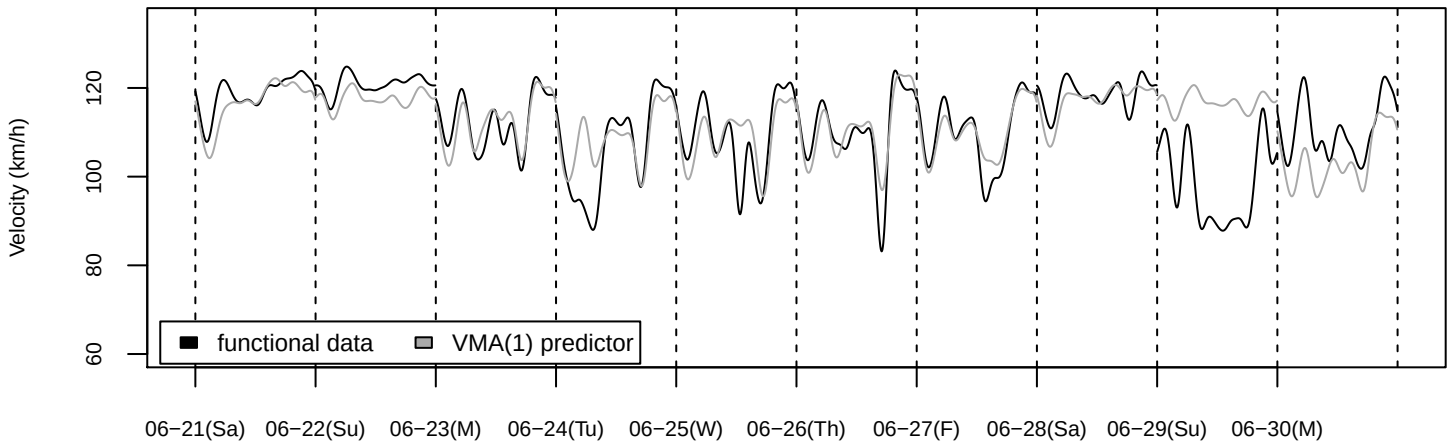

Figure 6.14: Functional velocity data on the last 10 days of **All days** and the corresponding functional predictor based on VMA(1) model. **Source:** Autobahndirektion Südbayern

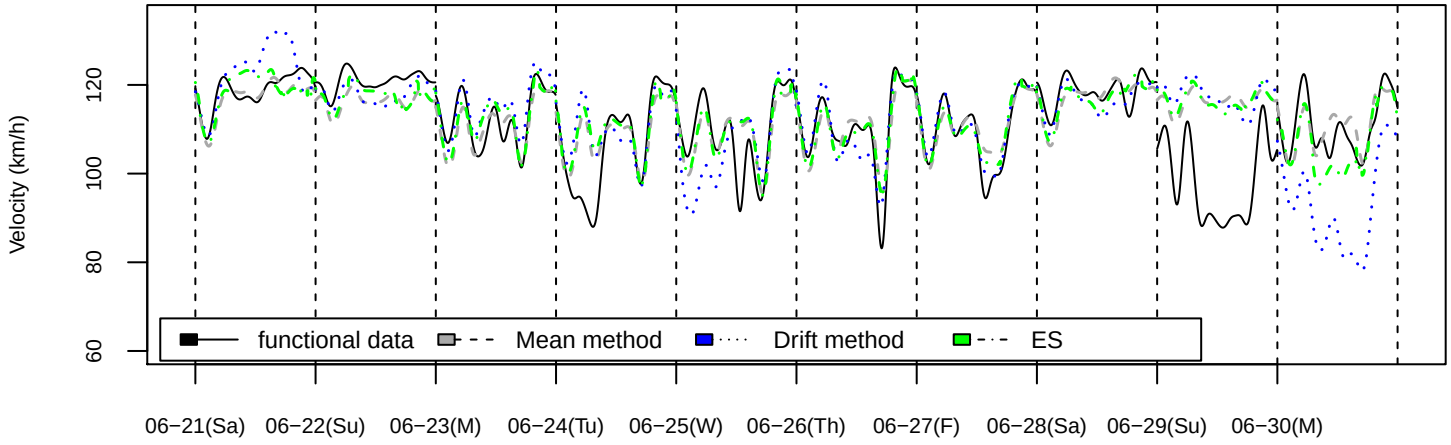

Figure 6.15: Univariate predictors of the velocity data on the last 10 days of **All days**. **Source:** Autobahndirektion Südbayern

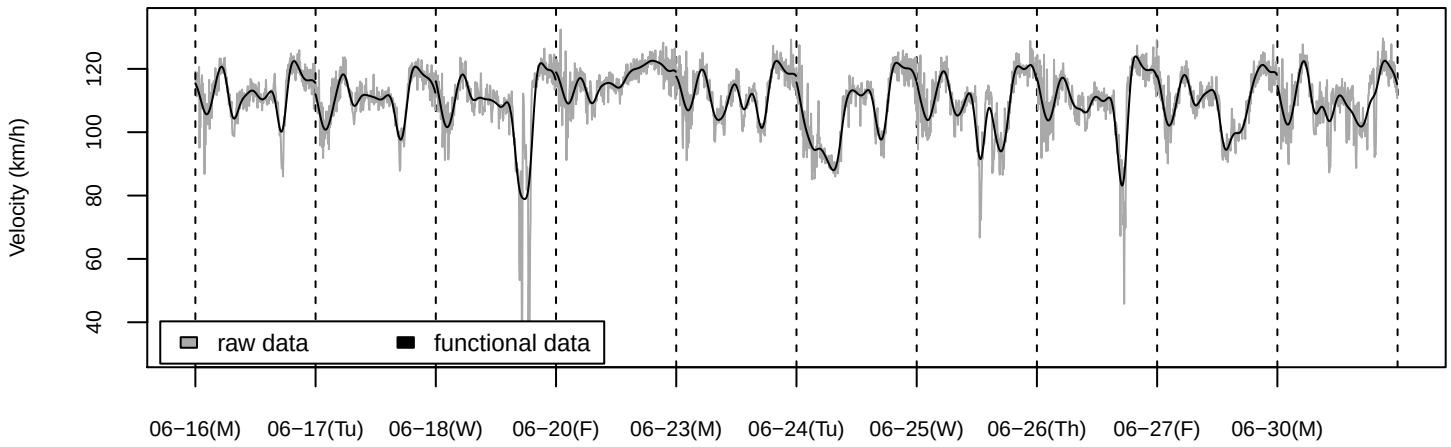

Figure 6.16: Raw velocity data and the corresponding functional data on the last 10 days of **Workingdays**. **Source:** Autobahndirektion Südbayern

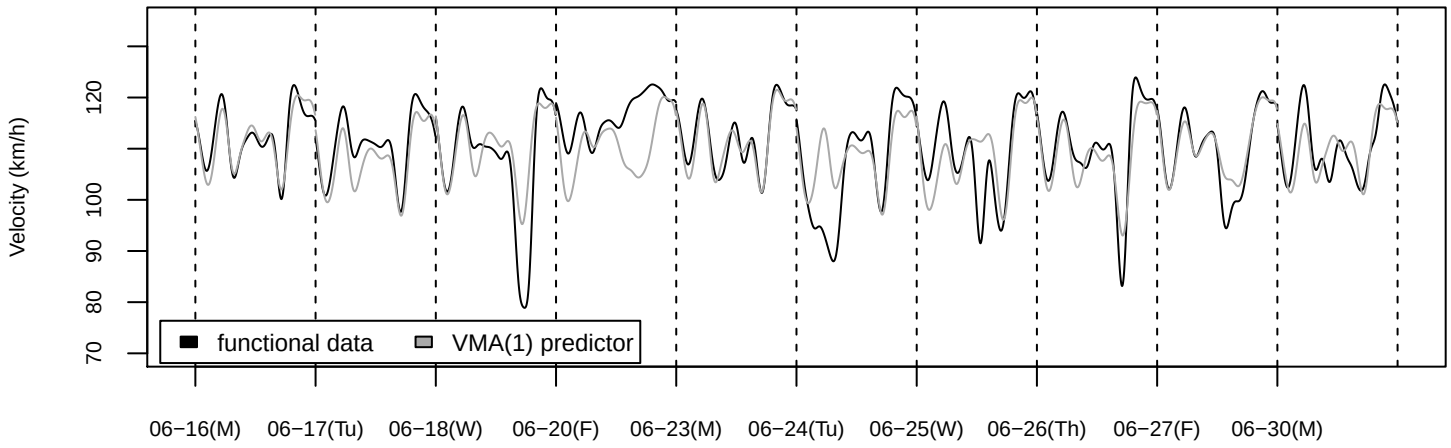

Figure 6.17: Functional velocity data on the last 10 days of **Workingdays** and the corresponding functional predictor based on VMA(1) model. **Source:** Autobahndirektion Südbayern

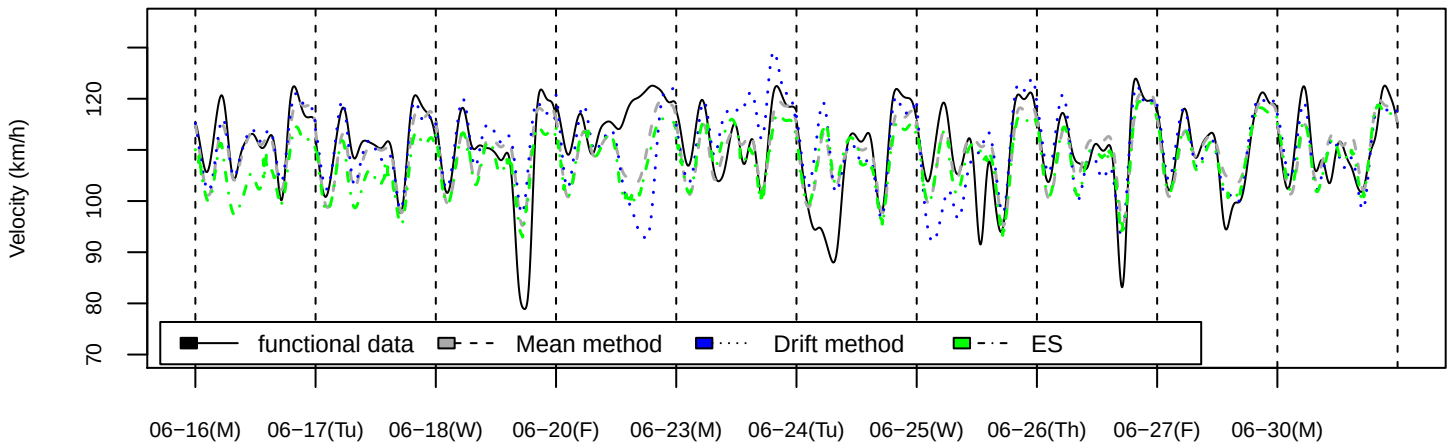

Figure 6.18: Univariate predictors of the velocity data on the last 10 days of **Workingdays**. **Source:** Autobahndirektion Südbayern

Now we finish this section with a brief summary.

Firstly, we talk about the functional predictors based on the different vector ARMA models. As can be seen from the results in Table 6.7 to Table 6.10, it is very hard to conclude which vector ARMA models really fits the data better than other vector models, which is quite different to the case in our simulation study in Chapter 5.

One of the possible reasons is, the real data maybe do not follow any specific ARMA model. Another reason is, the size of our sample is not big enough (we just have 178 days and 119 working days). But in the model fitting, we need to estimate $4 * 4 * 2 = 32$ parameters in the vector ARMA(1,1) and MA(2) model (since $d = 4$).

Then we compare the functional predictors with the univariate predictors. From the results in Table 6.8 and Table 6.10, we can at least say, the functional predictors perform not worse than the univariate predictors, which makes sense to the work on functional data analysis.

Bibliography

A. Aue, D. Norinho, and S. Hörmann. On the prediction of stationary functional time series. *Journal of the American Statistical Association*, 110:378–392, 2015.

D. Bosq. *Linear Processes in Function Spaces: Theory and Applications*. Springer, New York, 2000.

P.J. Brockwell and R.A. Davis. *Time Series: Theory and Methods (2nd Ed.)*. Springer, New York, 1991.

R. Gabrys and P. Kokoszka. Portmanteau test of independence for functional observations. *Journal of the American Statistical Association*, Vol. 102(No. 480), 2007.

S. Hörmann and P. Kokoszka. Weakly dependent functional data. *The Annals of Statistics*, 38(3):1845–1884, 2010.

S. Hörmann and P. Kokoszka. Functional time series. *Handbook of Statistics: Time Series Analysis-Methods and Applications*, 30:157–186, 2012.

L. Horváth and P. Kokoszka. *Inference for Functional Data with Applications*. Springer, New York, 2012.

L. Horváth, P. Kokoszka, and R. Reeder. Estimation of the mean of functional time series and a two sample problem. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 75:103–122, 2013a.

L. Horváth, P. Kokoszka, and G. Rice. Testing stationarity of functional time series. *Journal of Econometrics*, 2013b.

H. Lütkepohl. Linear transformations of vector arma processes. *Journal of Econometrics*, 26:283–293, 1984.

J.O. Ramsay and B.W. Silverman. *Applied Functional Data Analysis*. Springer Series in Statistics, 2002.

J.O. Ramsay and B.W. Silverman. *Functional Data Analysis*. Springer Series in Statistics, 2005.

J.O. Ramsay, G. Hooker, and S. Graves. *Functional Data Analysis with R and Mathlab*. Springer New York, 2009.