Technische Universität München

Zentrum Mathematik

# Moving Average Processes in Hilbert Spaces

Master's Thesis

by

Pablo Moreno Um

| | |
|---|---|
| Supervisor: | Prof. Dr. Claudia Klüppelberg |
| Advisor: | MSc. Johannes Klepsch |
| Submission date: | September 5th, 2016 |

I hereby declare that this thesis is my own work and that no other sources have been used except those clearly indicated and referenced.

Garching, September 2nd, 2016

# Abstract

Functional data analysis (FDA) deals with time series consisting of functional observations. We estimate a functional moving average process of order 1 by the projection method and by the iterative method. In addition we test the dependence order of a linear strictly stationary functional time series by two hypothesis tests, based on the concepts of the Box-Pierce test and of sample-splitting. The estimation approaches are used for forecasting and the tests help us choose an appropriate time series model, provided that the dependence structure is known. Finally we apply the former and the latter to Bavarian highway traffic data.

# Zusammenfassung

In dieser Arbeit geht es um funktionale Datenanalyse (FDA). Der Begriff umfasst den Umgang mit Zeitreihen, in denen jede Beobachtung eine Funktion ist. Es werden funktionale Moving Average Prozesse der Ordnung 1 mithilfe der Projektionsmethode und der iterativen Methode geschätzt. Darüber hinaus werden zwei Hypothesentests behandelt, welche die Abhängigkeitsordnung einer linearen streng stationären funktionalen Zeitreihe überprüfen. Die Tests beruhen auf dem Box-Pierce Test und auf Sample-Splitting. Mittels der Schätzverfahren können Vorhersagewerte berechnet werden, während die Tests die Auswahl eines geeigneten Zeitreihenmodells unter der gegebenen Abhängigkeitsstruktur erleichtern. All dies ist schließlich in Anwendung auf Verkehrsdaten einer bayerischen Autobahn zu sehen.

# Acknowledgments

First of all I want to thank Prof. Dr. Claudia Klüppelberg. It is not only because people usually thank their professors at first, but also because I would not have had the possibility to focus on time series analysis in a master's thesis at TUM without her. She was so kind to offer me this opportunity after I had asked for it. Moreover, her friendly nature meant a lot to me while working at the chair of statistics.

Secondly, I owe my advisor M.Sc. Johannes Klepsch very much. He spent plenty of time with me and supported me more than I could have imagined before starting with my thesis, although he had a lot to do for his own work. I learnt a great deal from him, not only regarding FDA but also about working as a PhD student. Furthermore, his nice, helpful, polite, honest and complaisant behavior motivated me decisively. I appreciate his personality.

Finally, I need to thank my friends for giving some good life balance between leisure time and working on my thesis as well as for finding plenty of typos. They made me enjoy those six-seven months.

# Contents

# Chapter 1

# Introduction

This thesis is about *functional data analysis* (*FDA*). FDA deals with time series of random functions in contrast to classical time series analysis (TSA) which deals with time series of scalar random variables (univariate TSA) or time series of random vectors (multivariate TSA). We use functional data because we want to model the (e.g. time) progress for each observation and handle high-dimensionality. There are plenty of application areas: daily stock prices, daily traffic flow, spatial temperature profile per hour, etc. Some examples can be found in [Hor, ch. 1].

In this thesis, we focus on two main topics. The first part of the thesis estimates functional moving average processes of order 1 from given functional data. Two estimation approaches, based on [Turb1], are derived and applied to simulated data. The second part is about two hypothesis tests concerning the dependence structure of a functional time series. We derive and prove the first test thoroughly and briefly explain the second test, which is a generalization and improvement of the first one.

The estimation approaches establish a time series model and can be used to forecast functional observations. To ensure that the forecasts are realistic, we have to ensure that an appropriate model is chosen. The hypothesis tests play a vital role in this matter. Since they test the dependence order, they indicate which linear and strictly stationary functional time series model is recommended.

In the following, we briefly survey the literature on FDA. [Rams] discusses numerical issues in converting multivariate data to functional data by interpolation as well as classical statistical procedures for FDA. The `R` package `fda`, based on [Rams], is used in this thesis for data analysis. The more probabilistic topics like linear stochastic processes and estimations including convergence rates are explained in [Bosq]. It focuses on estimating operators of functional autoregressive processes. In the context of predictions, functional autoregressive processes are also covered in [Hor], which discusses several aspects of FDA in Hilbert spaces. The fundamentals of FDA in this thesis are mostly based on [Hor] and partly on [Bosq].

Functional moving average processes are less researched than functional autoregressive processes, because the latter are easier to handle. Functional moving average processes are covered in [Turb1], [Turb2] and [Turb3] and [TurbThese]. The first part of this thesis is based on [TurbThese], but simulation studies play a bigger role here than in [TurbThese], which is designed rather theoretically. Beyond the moving average framework, prediction

of functional data is discussed in [BB].

In addition to discussing properties of estimators and forecasting autoregressive data, [Hor] also discusses an independence test for functional data. Moreover, this is also covered in [GaKo]. [Moon] focuses on dependence tests for univariate data. Apart from these, there is not much literature about (in)dependence tests. Thus the second part of this thesis can be seen as a new way of testing time series. The mathematical background of these tests mainly consists of Kronecker products from [Steeb] and convergence of random variables from [Vaart]. The proofs of the tests are inspired largely by [Brck], the fundamental book about classical, i.e. finite dimensional, time series analysis.

The thesis is organized as follows: Chapter 2 is a brief introduction into FDA. Chapter 3 introduces functional moving average processes including most of the important tools for the following chapters. Both Chapter 2 and Chapter 3 are essential for understanding the main topics of the thesis. In Chapter 4, the two estimation approaches mentioned above are explained. They are analyzed in simulation studies in Chapter 5. Chapter 6 describes the two hypothesis tests and compares them in simulation studies, but the focus is more on the first test. Finally, in Chapter 7 we discuss a real data example involving highway traffic volume data and highway traffic speed data from Bavaria in Germany, both converted to functional datasets. Two clustering algorithms are explained and applied to the speed dataset. At the end, the main methods of this thesis are used for the speed dataset.

# Chapter 2

# Fundamentals of Functional Data Analysis

This chapter gives a brief introduction to some theoretical aspects. Only a few claims will be proven, because the following focusses on what will be required in the main part. The first section deals with some necessary background theory from functional analysis and probability theory, whereas the second one is about functional principal component analysis, an approach about projecting functional observations onto a finite dimensional space in an optimal way.

## 2.1 Mathematical Background

All the statistical models and claims rely on Hilbert space theory. Therefore a brief summary of fundamental Hilbert space and operator theory is given. This will be specified on the space of square-integrable functions $L^2$. Some important terms as covariance operator and eigenvalues/-vectors are described. The last subsection deals with how to estimate them. Most of what is written in this chapter is based on [Hor, ch. 2,3 & 13] and on [Bosq].

### 2.1.1 Separable Hilbert spaces $H$ and Operators on $H$

A *Hilbert space* $H$ is a complete unitary vector space. *Completeness* means that each Cauchy sequence within that vector space converges. A *unitary* vector space is defined to have an *inner product* $\langle \cdot, \cdot \rangle$. Consequently, $H$ is a normed space, equipped with the *norm*

$$\|\cdot\| : \ H \to \mathbb{K}, \ x \mapsto \sqrt{\langle x, x \rangle},$$

where $\mathbb{K}$ is the field corresponding to $H$. $H$ being *separable* means that there exist a dense countable subset of $H$. A Hilbert space $H$ is separable if and only if it contains a countable orthonormal basis $E = \{e_n\}_{n \in \mathbb{N}}$ such that the span of $E$ is dense in $H$. If a real vector space $H$ is a separable Hilbert space, then the following claims hold:

- If $\dim(H) = K \in \mathbb{N}$, then $H$ is isomorphic to $\mathbb{R}^K$.
- If $\dim(H) = \infty$, then $H$ is isomorphic to $l_2(\mathbb{R})$ (square-summable sequences).

Let $\mathcal{L}$ be the space of *bounded linear operators* on $H$, i.e. each $\Psi \in \mathcal{L}$ maps from $H$ to $H$. Note that boundedness and continuity are equivalent for linear operators. $\mathcal{L}$ is a Banach space equipped with the *operator norm*

$$\|\Psi\|_{\mathcal{L}} := \sup_{\|x\| \leq 1} \|\Psi(x)\| .$$

An operator $\Psi \in \mathcal{L}$ is defined to be *compact* (also called *completely continuous*) if one of the following (equivalent) conditions holds:

- There exist two orthonormal bases $(v_j)_{j\in\mathbb{N}}$ and $(f_j)_{j\in\mathbb{N}}$ as well as a real sequence $(\lambda_j)_{j\in\mathbb{N}}$, $\lim_{j\to 0} \lambda_j = 0$, such that $\Psi$ can be represented in the *singular value decomposition*
$$\Psi(x) = \sum_{j=1}^{\infty} \lambda_j \langle x, v_j \rangle f_j, \qquad \forall x \in H. \tag{2.1}$$
All $\lambda_j$ are chosen to be positive because otherwise take $-f_j$ instead of $f_j$.
- $\Psi$ maps every bounded set into a compact set.
- $(\langle y, x_n \rangle \overset{n\to\infty}{\longrightarrow} \langle y, x \rangle \ \ \forall y \in H)$ implies $\|\Psi(x_n) - \Psi(x)\| \overset{n\to\infty}{\longrightarrow} 0$.

A compact operator is a *Hilbert-Schmidt operator* if $\sum_{j=1}^{\infty} \lambda_j^2 < \infty$ with $\lambda_j$ from (2.1). Those operators form the *Hilbert-Schmidt space* $\mathcal{S}$ of Hilbert-Schmidt operators. The Hilbert-Schmidt space is a separable Hilbert space endowed with the inner product

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

where the orthonormal basis $(e_i)_{i\in\mathbb{N}}$ is arbitrary, because this inner product does not depend on the exact choice of $(e_i)_{i\in\mathbb{N}}$. It turns out that $\|\Psi\|_{\mathcal{S}}^2 = \sum_{j=1}^{\infty} \lambda_j^2$ and $\|\Psi\|_{\mathcal{L}} \leq \|\Psi\|_{\mathcal{S}}$. $\Psi \in \mathcal{L}$ is *symmetric* if $\langle \Psi(x), y \rangle = \langle x, \Psi(y) \rangle$ holds for all $x, y \in H$. It is *positive-definitive* if $\langle \Psi(x), x \rangle \geq 0$ holds for all $x \in H$. Hence, a symmetric positive-definite Hilbert-Schmidt operator $\Psi$ can be represented as

$$\Psi(x) = \sum_{j=1}^{\infty} \lambda_j \langle x, v_j \rangle v_j, \qquad \forall x \in H, \tag{2.2}$$

with orthonormal eigenfunctions $v_j$, $j \in \mathbb{N}$, and eigenvalues $\lambda_1 > \lambda_2 > \ldots \geq 0$. $(v_j)_{j\in\mathbb{N}}$ can be assumed to form a basis (where some $\lambda_j$ may be zero), because otherwise a complete orthogonal system can be added in the othogonal complement of $\operatorname{span}\{v_j,\ j \in \mathbb{N}\}$, so that $(v_j)_{j\in\mathbb{N}}$ is extended to a basis thereby. Unfortunately in the infinite dimensional case there is a property of $\mathcal{S}$ that we have to cope with later:

**Lemma 2.1** *If* $\dim(H) = \infty$, *every symmetric positive-definite Hilbert-Schmidt operator* $\Psi$ *is not invertible in* $\mathcal{S}$.

Proof: If $\Psi$ were invertible, then $\Psi^{-1}$ would look like

$$\Psi^{-1}(x) = \sum_{j=1}^{\infty} \frac{1}{\lambda_j} \langle x, v_j \rangle v_j \qquad \forall x \in H,$$

because an inverse function is unique and

$$\Psi^{-1}(\Psi(x)) = \Psi\left(\Psi^{-1}(x)\right) = \sum_{j=1}^{\infty}\sum_{l=1}^{\infty} \lambda_j \frac{1}{\lambda_l} \langle x, v_l \rangle \langle v_l, v_j \rangle v_j \overset{(v_j) \text{ orthon.}}{=} \sum_{j=1}^{\infty} \langle x, v_j \rangle v_j \overset{(v_j) \text{ basis}}{=} x.$$

However,

$$\Psi \in \mathcal{S} \;\Rightarrow\; \sum_{j=1}^{\infty} \lambda_j^2 < \infty \;\Rightarrow\; \lambda_j^2 \overset{j\to\infty}{\longrightarrow} 0 \;\Rightarrow\; \frac{1}{\lambda_j^2} \overset{j\to\infty}{\longrightarrow} \infty \;\Rightarrow\; \sum_{j=1}^{\infty} \frac{1}{\lambda_j^2} = \infty.$$

Hence, $\Psi^{-1}$ is not well-defined in $\mathcal{S}$. □

The *trace* of a positive-definite operator $A \in \mathcal{L}$ is defined by

$$tr(A) := \sum_{n=1}^{\infty} \langle e_n, A(e_n) \rangle$$

for any choice of the orthonormal basis $(e_n)_{n\in\mathbb{N}}$. $A \in \mathcal{L}$ is called *trace class* or *nuclear* if $tr[(A^T A)^{\frac{1}{2}}] < \infty$. $\mathcal{T}$ denotes the set of all trace class operators. It is a vector space endowed with the norm

$$\|A\|_{\mathcal{T}} := tr[(A^T A)^{\frac{1}{2}}]. \tag{2.3}$$

The latter can be expressed as $\sum_{j=1}^{\infty} \lambda_j$ for symmetric positive-definite Hilbert-Schmidt operators. Therefore, the following relations concerning normed vector spaces of operators hold, where $\mathcal{C}$ denotes the set of compact operators:

$$\mathcal{L} \supset \mathcal{C} \supset \mathcal{S} \supset \mathcal{T}, \qquad \|\cdot\|_{\mathcal{L}} \leq \|\cdot\|_{\mathcal{S}} \leq \|\cdot\|_{\mathcal{T}}. \tag{2.4}$$

### 2.1.2 Random Functions of $L^2$ and Covariance Operator

In the following we will focus on the special case where $H = L^2([0,1])$, the space of square-integrable functions (abbr. $L^2$). Its inner product is defined by

$$\langle x, y \rangle := \int x(t)y(t)dt := \int_0^1 x(t)y(t)dt, \qquad \forall x, y \in L^2.$$

For $p \in \mathbb{N} \cup \{\infty\}$, the $L^2$ space is the only $L^p$ space to be endowed with a scalar product. In general it is difficult to give an explicit closed term for an $L^2$ function because it is uniquely defined up to null sets. Two $L^2$ functions $x, y$ are said to be *equal* if

$$\|x - y\|^2 = \int [x(t) - y(t)]^2 dt = 0.$$

A *(real) kernel* is a measurable function $\psi : [0,1]^2 \to \mathbb{R}$. An *integral operator* is a function $\Psi$ on $L^2$ which possesses a kernel $\psi$ such that

$$\Psi(x)(\cdot) = \int \psi(\cdot, s)x(s)ds, \qquad \forall x \in L^2.$$

It is an Hilbert-Schmidt operator if and only if

$$\int\int \psi^2(t,s)dtds < \infty. \qquad \text{Then } \|\Psi\|_{\mathcal{S}}^2 = \int\int \psi^2(t,s)dtds.$$

If $\psi(t,s) = \psi(s,t)$ for all $t, s \in [0,1]$, then $\Psi$ is *symmetric*; if $\int\int \psi(t,s)x(t)x(s)dtds \geq 0$, then $\Psi$ is *positive-definite*. The kernel of a symmetric positive-definite Hilbert-Schmidt integral operator is of the form

$$\psi(t,s) = \sum_{j=1}^{\infty} \lambda_j v_j(t) v_j(s).$$

**Mean, Covariance Operator:** Let $X = \{X(t),\ t \in [0,1]\}$ be a random element of $L^2$ equipped with the Borel $\sigma$-algebra. $X$ is said to be *integrable* if

$$\mathbb{E}\,\|X\| = \mathbb{E}\Big[\big(\int X^2(t)dt\big)^{\frac{1}{2}}\Big] < \infty.$$

An integrable random function $X$ has a unique *mean* $\mu \in L^2$ that satisfies

$$\mathbb{E}\,\langle y, X\rangle = \langle y, \mu\rangle\,, \qquad \forall y \in L^2 \qquad (\Rightarrow \mu(\cdot) = \mathbb{E}[X(\cdot)] \text{ a.e.})\,. \tag{2.5}$$

The expectation commutes with bounded operators ($\mathbb{E}\Psi(X) = \Psi(\mathbb{E}X)$ for all integrable $X$ and $\Psi \in \mathcal{L}$).

**Lemma 2.2 (c.f. [Hor, Lemma 2.1])** *If $X_1, X_2 \in L^2$ are independent, square integrable and $\mathbb{E}[X_1] = 0$, then $\mathbb{E}\,[\langle X_1, X_2\rangle] = 0$.*

**Definition 2.3 (Covariance Operator, c.f. [Hor, ch. 2.3])** *Let $X$ be a square-integrable (i.e. $\mathbb{E}\left[\|X\|^2\right] < \infty$) and centered (i.e. $\mathbb{E}X = 0$) random function. The* covariance operator *of $X$ is*

$$C(y)(t) := \mathbb{E}\,[\langle X, y\rangle\, X(t)] = \int c(t,s)y(s)ds, \qquad \forall y \in L^2$$

*with the* covariance function $c(t,s) = \mathbb{E}\,[X(t)X(s)]$.

The last equality holds because Fubini's theorem yields the interchangeability of the expectation and the inner product as two integrals. (The last expression is finite, because $X$ is square integrable and $y \in L^2$, so Cauchy-Schwarz inequality gives the result.) Using the notation

$$X \otimes X := \langle X, \cdot\rangle\, X,$$

one can write $C(\cdot) = \mathbb{E}[X \otimes X]$. $C$ is a symmetric positive-definite integral operator. $C \in \mathcal{L}(L^2)$ is a covariance operator if and only if it is symmetric positive-definite and its eigenvalues (which are non-negative as a consequence of positive-definiteness of $C$) satisfy $\sum_{j=1}^{\infty} \lambda_j < \infty$. The latter is due to the equalities (where $v_j$ denote the orthonormal eigenvectors of $C$)

$$\sum_{j=1}^{\infty} \lambda_j = \sum_{j=1}^{\infty} \langle Cv_j, v_j \rangle = \sum_{j=1}^{\infty} \langle \mathbb{E}[\langle X, v_j \rangle X], v_j \rangle = \sum_{j=1}^{\infty} \mathbb{E}\left[\langle X, v_j \rangle^2\right] = \mathbb{E}\left[\|X\|^2\right] < \infty,$$

where the last equality results from Parseval's equality

$$\sum_{j=1}^{\infty} \langle X, v_j \rangle^2 = \|X\|^2 . \tag{2.6}$$

Hence, every covariance operator is nuclear and the previous calculation demonstrates

$$\|C\|_{\mathcal{T}} = \mathbb{E}\left[\|X\|^2\right] . \tag{2.7}$$

In particular, a covariance operator as a nuclear operator is Hilbert-Schmidt. Symmetry and positive-definiteness yield the singular value decomposition

$$C(x) = \sum_{j=1}^{\infty} \lambda_j \langle x, v_j \rangle v_j, \qquad \forall x \in H.$$

In the following, $(\lambda_j)_{j\in\mathbb{N}}$ and $(v_j)_{j\in\mathbb{N}}$ denote the eigenvalues and -vectors of $C$.

**Lemma 2.4 (Mercer, c.f. [Bosq, Lemma 1.3])** *Let $c$ be a covariance function which is continuous over $[0,1]^2$. Then there exists a sequence $(v_i)_{i\in\mathbb{N}}$ of* orthonormal *continuous functions*

$$\forall i, j \in \mathbb{N} : \int_0^1 v_i(s) v_j(s) ds = \delta_{i,j} := \begin{cases} 1 & , i = j \\ 0 & , i \neq j \end{cases}$$

*and a decreasing sequence $(\lambda_i)_{i\in\mathbb{N}}$ of positive numbers such that*

$$\int_0^1 c(t,s) v_i(s) ds = \lambda_i v_i(t), \qquad \forall t \in [0,1] \ \forall i \in \mathbb{N}.$$

*Moreover $c$ can be written as*

$$c(t,s) = \sum_{i=1}^{\infty} \lambda_i v_i(t) v_i(s), \qquad \forall s, t \in [0,1],$$

*where the series converges uniformly on $[0,1]^2$. Hence,*

$$\sum_{i=1}^{\infty} \lambda_i = \int_0^1 c(t,t) dt < \infty.$$

This lemma implies a very important regularity criterion for continuous-time processes:

**Theorem 2.5 (Karhunen-Loève, c.f. [Hsing, Theorem 7.3.5])** *Let $X$ be a random function in $L^2([0,1])$. Assume $\mathbb{E}\left[\|X\|^2\right] < \infty$. Let $\mu(\cdot) = \mathbb{E}[X(\cdot)]$. Then*

$$(\xi_i)_{i\in\mathbb{N}} := (\langle (X-\mu), v_i\rangle)_{i\in\mathbb{N}}$$

*is an uncorrelated well-defined collection of mean-zero random variables with*

$$\mathbb{V}\mathrm{ar}[\xi_i] = \lambda_i,\ \forall i \in \mathbb{N}. \qquad (\Rightarrow\ \mathbb{E}[\xi_i\xi_j] = \lambda_i\delta_{i,j},\ \forall i,j \in \mathbb{N}.)$$

*Furthermore for any basis $(e_i)_{i\in\mathbb{N}} \subset L^2$*

$$\mathbb{E}\Big[\big\|(X-\mu) - \sum_{i=1}^{K} \xi_i v_i\big\|^2\Big] \le \mathbb{E}\Big[\big\|(X-\mu) - \sum_{i=1}^{K} \langle (X-\mu), e_i\rangle\, e_i\big\|^2\Big] \overset{K\to\infty}{\longrightarrow} 0. \tag{2.8}$$

*Moreover if the integral kernel $c$ of the covariance operator $C$ is continuous, then*

$$\sup_{t\in[0,1]} \mathbb{E}\Big[\big|(X(t)-\mu(t)) - \sum_{i=1}^{K} \xi_i v_i(t)\big|^2\Big] \overset{K\to\infty}{\longrightarrow} 0.$$

*(To be more precise,*

$$\mathbb{E}\Big[\big\|(X-\mu) - \sum_{i=1}^{K} \xi_i v_i\big\|^2\Big] = \sum_{i>K} \lambda_i$$

*for all $K \in \mathbb{N}$.)*

**Remark:** $X$ is called *Gaussian* if and only if $(\xi_n)_{n\in\mathbb{N}}$ are Gaussian and independent.

**Definition 2.6 (Cross-covariance operator, c.f. [Bosq, ch. 1.5])** *Let $X, Y$ be centered and square integrable random functions. A* cross-covariance operator *of $X$ and $Y$ is defined by*

$$C_{X,Y}(x)(t) = \mathbb{E}\left[\langle X, x\rangle\, Y(t)\right] = \int \mathbb{E}[Y(t)X(s)]x(s)ds, \qquad \forall x \in L^2. \tag{2.9}$$

The order of $X$ and $Y$ matters ($X \otimes Y \neq Y \otimes X$). As well as the covariance operator, the cross-covariance operators are nuclear:

$$\|C_{X,Y}\|_{\mathcal{T}} = \|C_{Y,X}\|_{\mathcal{T}} \le \mathbb{E}\left[\|X\|\,\|Y\|\right] < \infty.$$

Henceforth we go to the more statistical part. In general operators as the covariance operator are unknown. Imagine that $N \in \mathbb{N}$ functions $X_1(\cdot), \ldots, X_N(\cdot)$ (functional data) are given. The goal is to analyze the data. Thus the mean, the covariance operator and the eigenvalues and -functions have to be estimated empirically.

### 2.1.3 Estimation of mean, covariance functions, eigenvalues and -functions

Let $X_1, \ldots, X_N$ be a sample of $N \in \mathbb{N}$ random functions of $L^2$. We consider each curve $X_i$ as a realization of a random function $X$ of $L^2$. Although $X_1, \ldots, X_N$ will not be independent in the main chapter, we assume independence here. In other words we assume in this section

$$X, X_1, \ldots, X_N \text{ } i.i.d. \text{ in } L^2, \ X \text{ square} - \text{integrable}. \tag{2.10}$$

Based on independence we will show that the functional empirical estimators behave similarly to the finite dimensional empirical ones.

**Sample Mean Function:** The empirical counterpart of the mean function $\mu(\cdot) = \mathbb{E}[X(\cdot)]$ is defined by

$$\widehat{\mu}(t) := \frac{1}{N} \sum_{i=1}^{N} X_i(t).$$

For this, some versions of the *central limit theorem* and some versions of the *law of large numbers* exist. In addition, consistency holds.

**Theorem 2.7 (Central Limit Theorem (CLT), c.f. [Hor, Thm. 2.1])** *Suppose* $(X_i)_{i\in\mathbb{N}}$ *is a sequence of i.i.d. random elements in a separable Hilbert space with* $\mathbb{E}\left[\|X_1\|^2\right] < \infty$ *and* $\mathbb{E}[X_1] = \mu$. *Then*

$$\sqrt{N}(\widehat{\mu} - \mu) \xrightarrow{\mathscr{D}} Z \sim \mathcal{N}(0, C),$$

*i.e. Z is a Gaussian random element with the covariance operator*

$$C(x) = \mathbb{E}\left[\langle Z, x\rangle Z\right] = \mathbb{E}\left[\langle X_1, x\rangle X_1\right].$$

**Theorem 2.8 (Strong Law of Large Numbers (SLLN), c.f. [Hor, Thm. 2.2])** *Let* $(X_i)_{i\in\mathbb{N}}$ *be a sequence of i.i.d. random elements in a separable Hilbert space such that* $\mathbb{E}\left[\|X_1\|^2\right] < \infty$. *Then*

$$\widehat{\mu} \xrightarrow{\text{a.s.}} \mu.$$

**Theorem 2.9 (Consistency of $\widehat{\mu}$, c.f. [Hor, Thm. 2.3])** *Under (2.10)* $\mathbb{E}[\widehat{\mu}] = \mu$ *and* $\mathbb{E}\left[\|\widehat{\mu} - \mu\|^2\right] = O(\frac{1}{N})$. *(Hence,* $\|\widehat{\mu} - \mu\| \xrightarrow{P} 0$.*)*

Concerning estimators for the covariance function and the covariance operator, we will discuss both simultaneously, because a covariance operator is defined by its integral kernel, the covariance function.

**Sample Covariance Function/Operator:** The intuitive way to define empirical estimators for this is to replace the expectation operator by the empirical average and to center the functional observations by using the empirical mean

$$\begin{aligned}\widehat{c}(t,s) &:= \frac{1}{N}\sum_{i=1}^{N}(X_i(t)-\widehat{\mu}(t))(X_i(s)-\widehat{\mu}(s)), \qquad \forall s,t\in[0,1],\\ \widehat{C}(x) &:= \frac{1}{N}\sum_{i=1}^{N}\langle X_i-\widehat{\mu},x\rangle\,(X_i-\widehat{\mu}), \qquad \forall x\in L^2.\end{aligned}$$

They satisfy the integral condition as well as their theoretical counterparts

$$\begin{aligned}\widehat{C}(x)(t) &= \frac{1}{N}\sum_{i=1}^{N}\int(X_i(s)-\widehat{\mu}(s))x(s)ds(X_i(t)-\widehat{\mu}(t))\\ &= \int\frac{1}{N}\sum_{i=1}^{N}(X_i(t)-\widehat{\mu}(t))(X_i(s)-\widehat{\mu}(s))x(s)ds\\ &= \int\widehat{c}(t,s)x(s)ds, \qquad \forall t\in[0,1],\ \forall x\in L^2.\end{aligned}$$

However, one can show that they are biased, similarly to the scalar case

$$\mathbb{E}\left[\widehat{c}(t,s)\right]=\frac{N}{N-1}c(t,s)\ \left(\text{in } L^2([0,1]^2)\right).$$

Nevertheless we neglect this bias. Furthermore, we assume mean-zero random functions. Then

$$\widehat{c}(t,s)=\frac{1}{N}\sum_{i=1}^{N}X_i(t)X_i(s),\ \widehat{C}(x)=\frac{1}{N}\sum_{i=1}^{N}\langle X_i,x\rangle\,X_i.$$

The following theorem implies

$$\mathbb{E}\Big[\int\int\widehat{c}^2(t,s)dtds\Big]<\infty.\ \left(\Rightarrow\ \widehat{c}(\cdot,\cdot)\in L^2([0,1]^2)\text{ a.s.}\right)$$

**Theorem 2.10 (Boundedness of $\widehat{C}$, c.f. [Hor, Thm. 2.4])** *Assume* $\mathbb{E}\left[\|X\|^4\right]<\infty$, $\mathbb{E}[X]=0$ *and (2.10). Then*

$$\mathbb{E}\Big[\big\|\widehat{C}\big\|_{\mathcal{S}}^2\Big]\le\mathbb{E}\left[\|X\|^4\right].$$

Theorem 2.10 even holds for non-i.i.d. data.

**Theorem 2.11 (Consistency of $\widehat{C}$, c.f. [Hor, Thm. 2.5])** *Assume* $\mathbb{E}\left[\|X\|^4\right]<\infty$, $\mathbb{E}[X]=0$ *and (2.10). Then*

$$\mathbb{E}\Big[\big\|\widehat{C}-C\big\|_{\mathcal{S}}^2\Big]\le\frac{1}{N}\mathbb{E}\left[\|X\|^4\right].$$

**Theorem 2.12 (Central Limit Theorem for $\widehat{c}$, c.f. [Hor, Thm. 2.9])** *Assume* $\mathbb{E}\left[\|X\|^4\right] < \infty$, $\mathbb{E}[X] = 0$ *and (2.10). Then*

$$Z_N(t,s) := \sqrt{N}(\widehat{c}(t,s) - c(t,s))$$

*converges weakly in* $L^2([0,1]^2)$ *to a Gaussian process* $\Gamma(t,s)$ *with* $\mathbb{E}[\Gamma(t,s)] = 0$ *and*

$$\mathbb{E}[\Gamma(t,s)\Gamma(t',s')] = \mathbb{E}[X(t)X(s)X(t')X(s')] - c(t,s)c(t',s').$$

**Empirical Eigenvalues/-functions:** Compared to the previous estimators one has to be more careful with defining appropriate estimators for the eigenvectors and eigenfunctions of $C$. Here the intuitive method is finding the tuples

$$\left(\widehat{\lambda}_j, \widehat{v}_j\right) \text{ such that } \widehat{C}\widehat{v}_j = \widehat{\lambda}_j\widehat{v}_j$$

w.r.t. the sample covariance operator $\widehat{C}$. However:

- Since in theory a covariance operator $C$ has infinitely many eigenelements, we only estimate the $p \in \mathbb{N}$ largest eigenvalues (and their corresponding eigenfunctions) such that
$$\lambda_1 > \lambda_2 > \ldots > \lambda_p > \lambda_{p+1} \geq 0.$$
- Of course, the empirical eigenfunctions $\widehat{v}_j$ are designed to be close to the true eigenfunctions $v_j$. Since $v_j$ are chosen to be normalized, i.e. $\|v_j\| = 1$, the same is intended to hold for the empirical ones. However, as we do not know the exact eigenfunctions, we are not sure whether to take $\widehat{v}_j$ or $-\widehat{v}_j$. This is why we have to work with
$$\widehat{s}_j := \operatorname{sign}\left(\langle \widehat{v}_j, v_j \rangle\right)$$
in order to formulate convergence criteria.
- In practice people often work with discrete functions (i.e. finite dimensional vectors), which can be smoothed to functional data elements afterwards. For discrete data the sample covariance operator is a matrix. Therefore it has eigenvectors of finite length. This aspect is neglected here.

Those remarks give us the assumptions for the following convergence criterion.

**Theorem 2.13 (Consistency of Empirical Eigenelements, c.f. [Hor, Thm. 2.7])** *Assume* $\mathbb{E}\left[\|X\|^4\right] < \infty$, $\mathbb{E}[X] = 0$, *(2.10) and*

$$\lambda_1 > \lambda_2 > \ldots > \lambda_p > \lambda_{p+1}. \tag{2.11}$$

*Then for each* $j \in \{1, \ldots, p\}$

$$\limsup_{N\to\infty} N\mathbb{E}\left[\|\widehat{s}_j\widehat{v}_j - v_j\|^2\right] < \infty \qquad \text{and} \qquad \limsup_{N\to\infty} N\mathbb{E}\left[\left|\widehat{\lambda}_j - \lambda_j\right|^2\right] < \infty.$$

For the proof of this theorem, we need to state two additional lemmata. Both of them work with singular value decompositions

$$K_1(x) = \sum_{j=1}^{\infty} \lambda_j \langle y, v_j \rangle f_j, \qquad K_2(x) = \sum_{j=1}^{\infty} \gamma_j \langle x, u_j \rangle g_j \tag{2.12}$$

of two compact operators $K_1, K_2 \in \mathcal{L}$.

**Lemma 2.14 (c.f. [Hor, Lemma 2.3])** *Suppose $K_1, K_2 \in \mathcal{L}$ are two compact operators. Let $K_1$ be symmetric, i.e. $f_j = v_j$ in (2.12) and suppose that its eigenvalues satisfy (2.11). Define*

$$\begin{aligned} v_j' &:= s_j v_j,\ s_j := \operatorname{sign}\left(\langle u_j, v_j \rangle\right), & \alpha_1 &:= \lambda_1 - \lambda_2, \\ \alpha_j &:= \min\{\lambda_{j-1} - \lambda_j, \lambda_j - \lambda_{j+1}\} & \forall j &\in \{2, \dots, p\}. \end{aligned}$$

*Then, for all $j \in \{1, \dots, p\}$*

$$\|u_j - v_j'\| \leq \frac{2\sqrt{2}}{\alpha_j} \|K_2 - K_1\|_{\mathcal{L}}.$$

**Lemma 2.15 (c.f. [Hor, Lemma 2.2])** *Suppose $K_1, K_2 \in \mathcal{L}$ are two compact operators with (2.12). Then, for all $n \in \mathbb{N}$, $|\gamma_j - \lambda_j| \leq \|K_2 - K_1\|_{\mathcal{L}}$.*

Proof of Theorem 2.13: Use Lemma 2.14 and Lemma 2.15 by taking $K_1 = C$ and $K_2 = \widehat{C}$, remember (2.4) (here: $\|\cdot\|_{\mathcal{L}} \leq \|\cdot\|_{\mathcal{S}}$) use the expectation operator on all the expressions and apply Theorem 2.11, i.e.

$$\mathbb{E}\left[\left\|\widehat{C} - C\right\|_{\mathcal{S}}^2\right] \leq \frac{1}{N} \mathbb{E}\left[\|X\|^4\right].$$

Then the result immediately follows. □

## 2.2 Functional Principal Component Analysis (FPCA)

Suppose that we are given a dataset (i.e. non-random elements) of $N \in \mathbb{N}$ observed functions $x_1, \dots, x_N \in L^2$. The idea of FPCA is to reduce the data dimension by projecting the data onto a finite dimensional subspace in an optimal way (i.e. by minimizing the loss of information). Given $p < N$, how can we find an optimal finite dimensional subspace? In other words we want to find an *orthonormal basis* $u_1, \dots, u_p$ that minimizes

$$\widehat{S}^2 := \sum_{i=1}^{N} \left\| x_i - \sum_{k=1}^{p} \langle x_i, u_k \rangle u_k \right\|^2.$$

Having determined such an appropriate orthonormal basis to perform an orthogonal projection, we will be able to work with $p$-dimensional vectors

$$\mathbf{x_i} := \begin{pmatrix} \langle x_i, u_1 \rangle \\ \vdots \\ \langle x_i, u_p \rangle \end{pmatrix} \tag{2.13}$$

instead of infinite dimensional curves $x_j$. The optimal functions $u_1, \ldots, u_p$ are called *optimal empirical orthonormal basis* or *natural orthonormal components*. The words "empirical" and "natural" emphasize that they are computed directly from the functional data. We start with $p = 1$: Find $u \in L^2$ with $\|u\| = 1$ that minimizes

$$\sum_{i=1}^{N} \|x_i - \langle x_i, u\rangle u\|^2 = \sum_{i=1}^{N} \|x_i\|^2 - 2\sum_{i=1}^{N}\langle x_i, u\rangle^2 + \sum_{i=1}^{N}\langle x_i, u\rangle^2\|u\|^2 = \sum_{i=1}^{N} \|x_i\|^2 - \sum_{i=1}^{N}\langle x_i, u\rangle^2,$$

which is equivalent to maximizing

$$\sum_{i=1}^{N}\langle x_i, u\rangle^2 = \left\langle \widehat{C}u, u \right\rangle.$$

This maximization problem is easy to solve. Recall the singular value decomposition of a symmetric positive-definite Hilbert-Schmidt operator $\Psi$ (2.2) such that

$$\langle \Psi(z), z\rangle = \sum_{j=1}^{\infty} \lambda_j \langle z, v_j\rangle^2 \qquad \forall z \in L^2. \tag{2.14}$$

Suppose $\lambda_1 > \lambda_2 > \ldots$ and Parseval's equality (2.6) with normalization $\|z\|^2 = 1$ as constraints for this optimization problem. Then the maximal value of $\langle \Psi(\cdot), \cdot\rangle$ is $\lambda_1$. Thus take $\langle z, v_1\rangle^2 = 1$ and $\langle z, v_j\rangle = 0$ for $j > 1$. This leads to the choice $z = \pm v_1$, but we prefer to choose the positive sign. Back to our original notation, we found $u = \widehat{v}_1$. Here, we finally used an approach which we already know from linear algebra. Recall:

**Theorem 2.16 (Principal Axis Theorem, c.f. [Hor, Thm. 3.1])** *Suppose* $\mathbf{A}$ *is a symmetric* $p \times p$ *matrix. Then there exists an orthogonal matrix* $\mathbf{U} = [\mathbf{u_1}, \ldots, \mathbf{u_p}]$ *whose columns are the eigenvectors of* $\mathbf{A}$*, i.e.*

$$\mathbf{U^T U} = \mathbf{I} \qquad \text{and} \qquad \mathbf{A u_j} = \lambda_\mathbf{j} \mathbf{u_j}.$$

*Moreover,*

$$\mathbf{U^T A U} = \mathbf{\Lambda} = \mathrm{diag}[\lambda_\mathbf{1}, \ldots, \lambda_\mathbf{p}].$$

This implies $\mathbf{A} = \mathbf{U\Lambda U^T}$. Again we assume $\mathbf{A}$ to be symmetric and positive-definite with $\lambda_1 > \ldots > \lambda_p$. The problem is of the form

$$\text{Find} \qquad \widehat{\mathbf{z}} = \arg_\mathbf{z} \max \mathbf{z^T A z},$$

which, by $\mathbf{y} = \mathbf{U^T z}$, can be transformed to

$$\text{Find} \qquad \widehat{\mathbf{y}} = \arg_\mathbf{y} \max \mathbf{y^T \Lambda y}.$$

The latter problem has the same structure as (2.14) because of

$$\mathbf{y^T \Lambda y} = \sum_{\mathbf{j=1}}^{\mathbf{p}} \lambda_\mathbf{j} \mathbf{y_j^2}.$$

Hence the maximal value is $\lambda_1$ which is attained at

$$\mathbf{y} = \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \Rightarrow \mathbf{z} = \mathbf{u_1}.$$

Now we continue to minimize $\widehat{S}$ and deal with $p > 1$. In light of the case $p = 1$, we obtain

$$\widehat{S}^2 = \sum_{i=1}^{N} \|x_i\|^2 - \sum_{i=1}^{N} \sum_{k=1}^{p} \langle x_i, u_k \rangle^2.$$

Therefore maximize

$$\sum_{k=1}^{p} \sum_{i=1}^{N} \langle x_i, u_k \rangle^2 = \sum_{k=1}^{p} \left\langle \widehat{C}(u_k), u_k \right\rangle = \sum_{j=1}^{\infty} \widehat{\lambda}_j \langle u_1, \widehat{v}_j \rangle^2 + \sum_{j=1}^{\infty} \widehat{\lambda}_j \langle u_2, \widehat{v}_j \rangle^2 + \ldots + \sum_{j=1}^{\infty} \widehat{\lambda}_j \langle u_p, \widehat{v}_j \rangle^2$$

w.r.t. $u_1, \ldots, u_p$. Note that $u_1, \ldots, u_p$ is intended to become an orthonormal basis. Thus solve the first term by setting $u_1 = \widehat{v}_1$ as before. However, solve the second term by using the additional constraint $\langle u_2, \widehat{v}_1 \rangle = 0$ (because we know that $(\widehat{v}_j)_{j \in \mathbb{N}}$ is an orthonormal basis, too). Thus, the maximal value of the second term is $\lambda_2$, which is attained at $u_2 = \widehat{v}_2$. Inductively for the $l$-th term with $l \leq p$, consider the constraints $\langle u_l, \widehat{v}_j \rangle = 0$ for all $j \in \{1, \ldots, l-1\}$, which results in $u_l = \widehat{v}_l$. All in all the optimal basis for the target subspace consists of exactly the first $p$ eigenfunctions of the sample covariance operator $\widehat{C}$

$$u_1 = \widehat{v}_1, \ u_2 = \widehat{v}_2, \ \ldots, \ u_p = \widehat{v}_p. \tag{2.15}$$

This derivation proves (2.8) in the Karhunen-Loève Theorem (Theorem 2.5).
The functions in (2.15) are called *empirical functional principal components* (*EFPCs*, abbr. *principal components*) or *harmonics*. The *functional principal components* (*FPCs*) are the eigenfunctions of the covariance operator $C$ of a square integrable $L^2$-valued random function $X$, if the functional observations $X_1, \ldots, X_n$ have the same distribution as $X$. Under the assumptions of Theorem 2.13, the EFPCs estimate the FPCss (up to a sign).
For $j \in \{1, \ldots, p\}$, the inner product $\langle x_i, u_j \rangle$ is called $j$th *score*. According to (2.13), the scores can be interpreted as the weight of the contribution of the FPC $\widehat{v}_j$ to the curve $X_i$. The scores determine the variance of the data in the direction of the principal components: Remember from the Karhunen-Loève Theorem (Theorem 2.5)

$$\mathbb{E}\left[\langle X, v_j \rangle^2\right] = \mathbb{V}\mathrm{ar}\left[\langle X, v_j \rangle\right] = \lambda_j, \qquad \forall j \in \mathbb{N},$$

for a centered random function $X$ of $L^2$. Modify this formula by replacing all expressions of the left hand side with the empirical counterparts

$$\frac{1}{N} \sum_{i=1}^{N} \langle X_i, \widehat{v}_j \rangle^2 = \left\langle \frac{1}{N} \sum_{i=1}^{N} \langle X_i, \widehat{v}_j \rangle X_i, \widehat{v}_j \right\rangle = \langle \widehat{C} \widehat{v}_j, \widehat{v}_j \rangle = \widehat{\lambda}_j, \qquad \forall j \in \{1, \ldots, p\}.$$

Consequently, big (empirical) variances of the scores correspond to big (empirical) eigenvalues, which again correspond to a small projection error $\widehat{S}$ as deduced before. This is why we are interested in big variances of the scores. This makes sense, because big variances imply large variations. The bigger the variance is, the more probable it is to describe the data observations (even the ones which are farther away from the mean) by the principal components.
Seeking the biggest empirical eigenvalues gives us a rule of thumb for the choice of the number of principal components $p \in \mathbb{N}$: Calculate the *cumulative percentage of total variance* $(CPV)$

$$\mathrm{CPV}(p) := \frac{\sum_{k=1}^{p} \widehat{\lambda}_k}{\sum_{k=1}^{N} \widehat{\lambda}_k}$$

and choose $p_0$ such that $CPV(p_0) \geq 0.85$. This requires to calculate the first $N$ eigenelements $(\widehat{\lambda}_1, \widehat{v}_1), \ldots, (\widehat{\lambda}_N, \widehat{v}_N)$ and afterwards to select the appropriate $p$. However, in most of the cases the first eigenvalues are much bigger than the rest, which in plenty of applications leads to $p < N$, typically a single digit number.

# Chapter 3

# Moving Average Process of order 1

This chapter contains the fundamentals about moving average processes. The reader is required to possess some knowledge of (univariate) time series analysis. Most of what is written here refers to [Bosq], [Turb1], [Turb2], [Turb3] and especially [TurbThese].

## 3.1 Definition and Properties

Recall the definition of the cross-covariance operator in Definition 2.6.

**Definition 3.1 (Stationarity, c.f. [Bosq, Def. 2.4])** *Let $(X_n)_{n\in\mathbb{Z}}$ be a sequence of random functions of $L^2$. It is called* (weakly) stationary*, if*

- $\mathbb{E}\left[\|X_n\|^2\right] < \infty, \qquad \forall n \in \mathbb{Z},$
- $\mathbb{E}[X_n] = \mu, \qquad \forall n \in \mathbb{Z},$
- $C_{X_n,X_r} = C_{X_{n+h},X_{r+h}}, \qquad \forall n, r, h \in \mathbb{Z}.$

*It is called* strictly/strongly stationary *if*

$$(X_{n_1}, \ldots, X_{n_m}) \overset{\mathscr{D}}{=} (X_{n_1+h}, \ldots, X_{n_m+h}), \qquad \forall n_1, \ldots, n_m, h \in \mathbb{Z},\ \forall m \in \mathbb{N}.$$

*(Clearly, strict stationarity implies weak stationarity, provided that $\mathbb{E}\left[\|X_n\|^2\right] < \infty$ holds for all $n \in \mathbb{Z}$.)*

For a stationary process we denote $C_h := C_{X_n,X_{n+h}}$ for any $n \in \mathbb{Z}$.

**Definition 3.2 (White Noise, c.f. [Bosq, Def. 3.1])** *A sequence $(\varepsilon_i)_{i\in\mathbb{Z}}$ of random functions of a separable Hilbert space $H$ is called an $H$-*white noise *(WN) if*

*(i)* $C_{\varepsilon_i} := \mathbb{E}\left[\varepsilon_i \otimes \varepsilon_i\right]$ *is independent from $i \in \mathbb{Z}$ and*

$$0 < \mathbb{E}\left[\|\varepsilon_i\|^2\right] = \sigma^2 < \infty,\ \mathbb{E}[\varepsilon_i] = 0, \qquad \forall i \in \mathbb{Z},$$

*(ii)* $\varepsilon_i$ *is* orthogonal *to $\varepsilon_j$ for all $i, j \in \mathbb{Z},\ i \neq j$, in the sense*

$$\mathbb{E}\left(\langle \varepsilon_i, x\rangle \langle \varepsilon_j, y\rangle\right) = 0, \qquad \forall x, y \in H. \tag{3.1}$$

*If instead of the more general condition (3.1) $(\varepsilon_i)_{i\in\mathbb{Z}}$ is a sequence of i.i.d. random functions of $H$, $(\varepsilon_i)_{i\in\mathbb{Z}}$ is said to be an $H$* strong white noise *(SWN).*

Due to independence from $i \in \mathbb{Z}$ one denotes $C_\varepsilon := C_{\varepsilon_i}$.

**Definition 3.3 (Functional Moving Average Model, c.f. [TurbThese, Def. 2.1.7])** *Let $(X_n)_{n\in\mathbb{Z}}$ be a stationary sequence of mean zero random functions of $H = L^2$. If $(X_n)_{n\in\mathbb{Z}}$ is not centered, use the shifted process $(X_n - \mu)_{n\in\mathbb{Z}}$. Moreover let $(\varepsilon_n)_{n\in\mathbb{Z}}$ be an $H$-white noise. Given $l_1, \ldots, l_q \in \mathcal{L}$, $(X_n)_{n\in\mathbb{Z}}$ follows a* functional moving average model of order $q$ *($MAH(q)$) with $q \in \mathbb{N}$ if*

$$X_n = \varepsilon_n + l_1(\varepsilon_{n-1}) + \ldots + l_q(\varepsilon_{n-q}), \qquad \forall n \in \mathbb{Z}.$$

We will only consider $MAH(1)$ processes with $l = l_1$. Here we assume

$$\mathbb{E}\left[\|l(\varepsilon_{n-1})\|\right] > 0, \qquad \forall n \in \mathbb{Z}, \tag{3.2}$$

which means that $(X_n)_{n\in\mathbb{Z}}$ cannot be a "simple" $H$-white noise. Furthermore, $\|l\|$ always denotes $\|l\|_{\mathcal{L}}$.

**Lemma 3.4 (c.f. [Bosq, Lemma 3.1] and [Hor, Lemma 13.1])** *For any $l \in \mathcal{L}$, the following two conditions are equivalent:*

- $\exists j_0 \in \mathbb{N} : \ \|l^{j_0}\| < 1$
- $\exists a > 0,\ b \in (0,1) : \ \forall j \in \mathbb{N}_0 : \ \|l^j\| \le a b^j$

*(Note that $\|l\| < 1$ is a special case of $\|l^{j_0}\| < 1$ for some $j_0 \in \mathbb{N}$.)*

Proof: "⇐" $|b| < 1$ implies that $j \in \mathbb{N}$ can be chosen such that $b^j$ becomes arbitrarily small. Hence there exists a $j_0 \in \mathbb{N}$ such that $b^{j_0} < \frac{1}{a}$. This implies

$$\left\|l^{j_0}\right\| \le a b^{j_0} < 1.$$

"⇒" Let $j_0$ be given. For $\|l^{j_0}\| = 0$ there are only finitely many $j < j_0$ such that $\|l^j\| > 0$ holds. They can be covered by an exponential function as an envelope function subject to $j$ that exceeds $\|l^j\|$ for each $j \in \mathbb{N}$ (choose $a$ sufficiently big). Therefore we assume $\|l^{j_0}\| > 0$. For $j < j_0$ an exponential function as an envelope function can be found by the same argument as before. Thus take any $j > j_0$. There are some $q, r \in \mathbb{N}$, $0 \le r < j_0$, such that $j = j_0 q + r$ (euclidean division). Therefore

$$\left\|l^j\right\| = \left\|l^{j_0 q} l^r\right\| \le \left\|l^{j_0}\right\|^q \left\|l^r\right\|.$$

We know $\|l^{j_0}\| < 1$ and $q > \frac{j}{j_0} - 1$ (by the choice of $r$). Thus

$$\left\|l^j\right\| \le \left\|l^{j_0}\right\|^{\frac{j}{j_0}-1} \left\|l^r\right\| \le \left(\left\|l^{j_0}\right\|^{\frac{1}{j_0}}\right)^j \left\|l^{j_0}\right\|^{-1} \max_{0\le r<j_0} \left\|l^j\right\|$$

Hence, $a = \|l^{j_0}\|^{-1} \max\limits_{0\le r<j_0} \|l^r\|$ and $b = \|l^{j_0}\|^{\frac{1}{j_0}} < 1$ form the envelope function exceeding $\|l^j\|$ for $j > j_0$. Finally the maximum of $a$ from $j < j_0$ and from $j > j_0$ and the maximum of $b$ from $j < j_0$ and from $j > j_0$ yield the result. □

**Definition 3.5 ($m$-Dependence, [TurbThese, Def. 2.1.2])** *Let $(X_n)_{n\in\mathbb{Z}}$ be a strictly stationary sequence of random functions of $L^2$. Given $m \in \mathbb{N}$, it is called $m$-dependent, if $(X_j)_{j\le n}$ and $(X_j)_{j\ge n+m+1}$ are independent for all $n \in \mathbb{Z}$.*

It is easy to see that the $MAH(1)$ process

$$X_n = \varepsilon_n + l(\varepsilon_{n-1}) \tag{3.3}$$

is 1-dependent, if $(\varepsilon_n)_{n\in\mathbb{Z}}$ is an $H$ strong white noise.

**Proposition 3.6 (cf. [TurbThese, Prop. 2.2.1])** *Let $(X_n)_{n\in\mathbb{Z}}$ be an $MAH(1)$ process with operator $l$ and let $l^*$ be its adjoint operator. Then*

$$C := C_0 = C_\varepsilon + lC_\varepsilon l^* \neq 0, \tag{3.4}$$
$$D := C_1 = lC_\varepsilon \neq 0, \tag{3.5}$$
$$C_h \equiv 0, \qquad \forall |h| > 1. \tag{3.6}$$

Proof: Use the bilinearity of $\otimes$, the white noise property (Definition 3.2) and the definition of an adjoint operator. Thus

$$\begin{aligned} C &= \mathbb{E}[X_n \otimes X_n] = \mathbb{E}[\varepsilon_n \otimes \varepsilon_n] + \mathbb{E}[l(\varepsilon_{n-1}) \otimes l(\varepsilon_{n-1})] + \mathbb{E}[l(\varepsilon_{n-1}) \otimes \varepsilon_n] + \mathbb{E}[\varepsilon_n \otimes l(\varepsilon_{n-1})] \\ &= \mathbb{E}[\varepsilon_n \otimes \varepsilon_n] + \mathbb{E}[l(\varepsilon_{n-1}) \otimes l(\varepsilon_{n-1})] = C_\varepsilon + \mathbb{E}[\langle \varepsilon_{n-1}, l^*(\cdot)\rangle l(\varepsilon_{n-1})] = C_\varepsilon + lC_\varepsilon l^* \end{aligned}$$

and

$$D = \mathbb{E}[X_n \otimes X_{n+1}] = \mathbb{E}[\varepsilon_n \otimes l(\varepsilon_n)] = l\left(\mathbb{E}[\varepsilon_n \otimes \varepsilon_n]\right) = lC_\varepsilon$$

follow. In contrast to $C$ and $D$

$$\begin{aligned} C_h = \mathbb{E}[X_n \otimes X_{n+h}] &= \mathbb{E}[\varepsilon_n \otimes \varepsilon_{n+h}] + \mathbb{E}[l(\varepsilon_{n-1}) \otimes l(\varepsilon_{n+h-1})] + \mathbb{E}[l(\varepsilon_{n-1}) \otimes \varepsilon_{n+h}] + \\ &\quad \mathbb{E}[\varepsilon_n \otimes l(\varepsilon_{n+h-1})] \overset{(3.1)}{=} 0 + 0 + 0 + 0 = 0 \end{aligned}$$

holds. It remains to show that $C$ and $D$ are non-zero operators. As a covariance matrix of a non-constant stochastic process, $\|C\| > 0$ w.r.t. every operator norm holds. Concerning $D$ we conclude from

$$\|lC_\varepsilon l^*\|_{\mathcal{T}} = \|C_{l(\varepsilon_{n-1})}\|_{\mathcal{T}} \overset{(2.7)}{=} \mathbb{E}\big[\|l(\varepsilon_{n-1})\|^2\big],$$

that $\|lC_\varepsilon l^*\|_{\mathcal{T}} > 0$ holds because of (3.2) (use Cauchy-Schwarz inequality). Consequently $\|lC_\varepsilon\|_{\mathcal{T}} > 0$ holds. □

**Proposition 3.7 (c.f. [TurbThese, Prop. 2.2.2])** *If $(X_n)_{n\in\mathbb{Z}}$ is a weakly stationary, regular and mean zero linear process and if*

$$\begin{aligned} \forall |h| > 1 : \quad C_h &\equiv 0, \quad \text{and} \\ \exists l \in \mathcal{L} : \quad C_1 &= lC_\varepsilon \neq 0 \end{aligned}$$

*hold, then $(X_n)_{n\in\mathbb{Z}}$ is an $MAH(1)$ process.*

Proof: See [TurbThese, subsection 2.2.5]. □

**Proposition 3.8** *Let $(X_n)_{n\in\mathbb{Z}}$ be a linear and strongly stationary process of second order. Then it is an $MAH(m)$ process with an $H$ strong white noise $(\varepsilon_n)_{n\in\mathbb{Z}}$ if and only if it is $m$-dependent with $m \in \mathbb{N}$ such that $X_n$ and $X_{n+m}$ are dependent for all $n \in \mathbb{Z}$.*

Proof: "⇒" Take any $n \in \mathbb{Z}$. Since $X_n$ consists of $\varepsilon_n, \ldots, \varepsilon_{n-m}$ and since $(\varepsilon_n)_{n\in\mathbb{Z}}$ are independent, $(X_j)_{j\leq n}$ and $(X_j)_{j\geq n+m+1}$ are independent. This is the definition of $m$-dependence. Moreover both $X_n$ and $X_{n+m}$ contain $\varepsilon_n$ as a summand. Thus they are dependent.

"⇐" Since $X_n$ and $X_{n+m}$ are dependent for all $n \in \mathbb{Z}$ and since independence implies orthogonality,

$$\begin{aligned} C_m &\neq 0 \qquad \text{and} \\ C_h(\cdot)(\cdot) &\overset{\text{ind.}}{=} \mathbb{E}[\langle X_n, \cdot\rangle]\mathbb{E}[\langle X_{n+m}, \cdot\rangle] \equiv 0 \qquad \forall |h| > m \end{aligned}$$

hold. By means of Proposition 3.7 one can show that these (in-)equations imply an $MAH(m)$ process. Now assume that $(X_n)_{n\in\mathbb{Z}}$ does not consist of an $H$ strong white noise $(\varepsilon_n)_{n\in\mathbb{Z}}$. This means that $(\varepsilon_n)_{n\in\mathbb{Z}}$ are not identically distributed or not independent. However, in the former case $(X_n)_{n\in\mathbb{Z}}$ would not be strongly stationary and in the latter one $(X_n)_{n\in\mathbb{Z}}$ would not be $m$-dependent. Hence the equivalence follows. □

**Proposition 3.9** *Let $(X_n)_{n\in\mathbb{Z}}$ be an $MAH(1)$ process and let $K \in \mathbb{N}$. Take any orthonormal functions $e_1, \ldots, e_K \in L^2$ which span the set $A_K = \operatorname{span}\{e_1, \ldots, e_K\}$. Let $P_{A_K}$ be the projection operator*

$$P_{A_K}(f) := \sum_{i=1}^{K} \langle f, e_i\rangle e_i \qquad \forall f \in L^2.$$

*Then $(P_{A_K}(X_n))_{n\in\mathbb{Z}}$ is still an $MAH(1)$ process.*

Proof: As a projection operator $P_{A_K}$ is symmetric and idempotent. Since $P_{A_K}$ is a linear operator, $(P_{A_K}(X_n))_{n\in\mathbb{Z}}$ is still a linear mean-zero process. Furthermore due to

$$\begin{aligned} C_{P_{A_K}(X_n),P_{A_K}(X_{n+h})} &= \mathbb{E}\left[\langle P_{A_K}(X_n), \cdot\rangle\, P_{A_K}(X_{n+h})\right] \\ &= P_{A_K}\Big(\mathbb{E}\left[\langle X_n, P_{A_K}(\cdot)\rangle\, X_{n+h}\right]\Big) = P_{A_K} C_h P_{A_K} \qquad \forall h \in \mathbb{Z} \end{aligned}$$

the cross-covariance operators of $(P_{A_K}(X_n))_{n\in\mathbb{Z}}$ only depend on the difference of the indices. Consequently we denote $(C_h^{P_{A_K}(X)})_{h\in\mathbb{Z}}$ as the cross-covariance operators of $(P_{A_K}(X_n))_{n\in\mathbb{Z}}$ and $(C_h^X)_{h\in\mathbb{Z}}$ as the cross-covariance operators of $(X_n)_{n\in\mathbb{Z}}$. Hence

$$\begin{aligned} \mathbb{E}\Big[\|P_{A_K}(X_n)\|^2\Big] &\overset{(2.7)}{=} \|C_0^{P_{A_K}(X)}\|_{\mathcal{T}} = \|P_{A_K} C_h P_{A_K}\|_{\mathcal{T}} \leq \|P_{A_K}\|_{\mathcal{T}} \|C_h\|_{\mathcal{T}} \|P_{A_K}\|_{\mathcal{T}} \\ &= K\|C_h\|_{\mathcal{T}} K \overset{(2.7)}{=} K^2 \mathbb{E}\Big[\|X_n\|^2\Big] < \infty \end{aligned}$$

holds for all $h \in \mathbb{Z}$, because 1 is the only non-zero eigenvalue of $P_{A_K}$ which appears $K$ times. All in all $(P_{A_K}(X_n))_{n\in\mathbb{Z}}$ is stationary. Thus

$$\begin{aligned} C_h^{P_{A_K}(X)} &= P_{A_K} C_h^X P_{A_K} \overset{\text{Lemma 3.6}}{=} P_{A_K} 0 P_{A_K} = 0 \qquad \forall |h| > 1 \\ C_1^{P_{A_K}(X)} &= P_{A_K} C_1^X P_{A_K} \overset{\text{Lemma 3.6}}{=} P_{A_K} l \; Id \; C_\varepsilon P_{A_K} = \left(P_{A_K} l P_{A_K}\right)\left(P_{A_K} C_\varepsilon P_{A_K}\right) \overset{l, C_\varepsilon \neq 0}{\neq} 0. \end{aligned}$$

Finally by Lemma 3.7 $(P_{A_K}(X_n))_{n\in\mathbb{Z}}$ is an $MAH(1)$ process with $l^{P_{A_K}(X)} = P_{A_K} l P_{A_K}$. □

**Theorem 3.10 (Invertibility of an $MAH(1)$ process, c.f. [TurbThese, Lemma 2.2.1])** *Let $(X_n)_{n\in\mathbb{Z}}$ be an $MAH(1)$ process on $L^2$. Assume that $(\varepsilon_n)_{n\in\mathbb{Z}}$ is an $H$-white noise. If there exists a $j_0 \in \mathbb{N}$ such that $\|l^{j_0}\| < 1$, then the $H$-white noise $(\varepsilon_n)_{n\in\mathbb{Z}}$ can be written as*

$$\varepsilon_n = \sum_{j=0}^{\infty} (-1)^j l^j (X_{n-j}), \qquad \forall n \in \mathbb{Z}. \tag{3.7}$$

*The series converges in $L^2$, endowed with the inner product $\mathbb{E}[\langle \cdot, \cdot \rangle]$.*

Proof: Fix any $n \in \mathbb{Z}$.

$$\begin{aligned} \varepsilon_n &= X_n - l(\varepsilon_{n-1}),\ \varepsilon_{n-1} = X_{n-1} - l(\varepsilon_{n-2}), \dots \\ \Rightarrow\ \varepsilon_n &= \sum_{j=0}^{k} (-1)^j l^j (X_{n-j}) + (-1)^{k+1} l^{k+1}(\varepsilon_{n-k-1}), \qquad \forall k \in \mathbb{N}. \end{aligned}$$

Use Lemma 3.4 to get

$$\begin{aligned} \mathbb{E}\left[\left\| \varepsilon_n - \sum_{j=0}^{k} (-1)^j l^j (X_{n-j}) \right\|^2\right] &= \mathbb{E}\left[\left\| (-1)^{k+1} l^{k+1}(\varepsilon_{n-k-1}) \right\|^2\right] \\ &\leq \|l^{k+1}\|^2 \mathbb{E}\left[\|\varepsilon_{n-k-1}\|^2\right] \leq a^2 b^{2(k+1)} \sigma^2 \overset{k\to\infty}{\longrightarrow} 0. \end{aligned}$$

This means that the $L^2$-limit is $\varepsilon_n$. □

**Example 3.11 (Trunc. Ornstein-Uhlenbeck Process, c.f. [TurbThese, Ex. 2.2.2])** *Let $\left(W_t^{(1)}\right)_{t\in\mathbb{R}}$ and $\left(W_t^{(2)}\right)_{t\in\mathbb{R}}$ two independent standard Wiener processes. Define*

$$W_t = W_t^{(1)} \mathbb{1}_{\mathbb{R}_+}(t) + W_t^{(2)} \mathbb{1}_{\mathbb{R}_-}(t) \qquad \forall t \in \mathbb{R}.$$

*Then the Langevin stochastic differential equation*

$$d\xi_t = -\theta \xi_t dt + \sigma dW_t, \qquad \theta > 0,\ \sigma > 0$$

*can be solved by the stationary zero-mean Gaussian process*

$$\xi_t := \sigma \int_{-\infty}^{t} e^{-\theta(t-s)} dW_s \qquad \forall t \in \mathbb{R}.$$

*This process is called an* Ornstein-Uhlenbeck (O.U.) process. *We work with a modified (truncated) version*

$$\xi_t := \sigma \int_{\{t\}-1}^{t} e^{-\theta(t-s)} dW(s) \qquad \forall t \in \mathbb{R},$$

*with* $\{t\} := \min\{s \in \mathbb{N} \mid s \geq t-1\}$. *Now let* $\sigma = 1$ *and define*

$$X_n(t) := \xi_{n+t}, \qquad \forall t \in [0,1],\ \forall n \in \mathbb{Z}.$$

*Therefore* $(X_n)_{n\in\mathbb{Z}}$ *is a random sequence with components in the Hilbert space* $L^2\left([0,1], \mathcal{B}_{[0,1]}, \lambda + \delta_{(1)}\right)$, *where* $\lambda$ *denotes the Lebesgue measure and* $\delta_{(1)}$ *the Dirac measure at* 1. *(Other choices for a Hilbert spaces are possible too, such that* $(X_n)_{n\in\mathbb{N}}$ *is well-defined.) Moreover, define the operator*

$$l_\theta :\ H \to H,\ l_\theta(x)(t) := e^{-\theta t} x(1) \qquad \forall t \in [0,1],\ \forall x \in H$$

*and the random functions*

$$\varepsilon_n(t) := \int_{n}^{n+t} e^{-\theta(n+t-s)} dW(s).$$

*For* $n \neq m$, $\varepsilon_n$ *and* $\varepsilon_m$ *are independent, as integrals over disjoint intervals. They are identically distributed, because the increments of a Brownian motion are shift-invariant. Therefore,* $(\varepsilon_n)_{n\in\mathbb{Z}}$ *is an* $H$ *strong white noise. Note*

$$\begin{aligned} l_\theta(\varepsilon_{n-1})(t) + \varepsilon_n(t) &= e^{-\theta t} \int_{n-1}^{n} e^{-\theta(n-s)} dW(s) + \int_{n}^{n+t} e^{-\theta(n+t-s)} dW(s) \\ &= \int_{n-1}^{n+t} e^{-\theta(n+t-s)} dW(s) = X_n(t). \end{aligned}$$

*Thus* $(X_n)_{n\in\mathbb{Z}}$ *is an* $MAH(1)$ *process. In addition,*

$$\|l_\theta\|_{\mathcal{L}}^2 = \int_0^1 e^{-2\theta t} d(\lambda + \delta_{(1)})(t) = \int_0^1 e^{-2\theta t} dt + \int_0^1 e^{-2\theta} dt = \frac{1 - e^{-2\theta}}{2\theta} + e^{-2\theta} =: \alpha(\theta),$$

*which means that for* $\theta > \frac{1}{2}$ *we obtain* $\|l_\theta\|_{\mathcal{L}}^2 < 1$, *so that* $(X_n)_{n\in\mathbb{Z}}$ *is invertible. For* $0 < \theta \leq \frac{1}{2}$, $(X_n)_{n\in\mathbb{Z}}$ *being invertible still holds, because one can show*

$$\|l_\theta^j\|_{\mathcal{L}}^2 = e^{-2\theta(j-1)} \alpha(\theta), \qquad \forall j \geq 1.$$

*Hence, choosing* $j_0 > 1$ *sufficiently large yields invertibility of* $(X_n)_{n\in\mathbb{Z}}$.

The next subsection is about estimating the mean function and the covariance function empirically, which was already discussed in the previous chapter. However, here the independence assumption for $(X_n)_{n\in\mathbb{Z}}$ does not hold any more.

## 3.2 Estimation of Mean and Cross-Covariance Operator

The formulas for these estimators are still the same as in the case of i.i.d. observations, but we have to adjust the argumentations when proving the limit theorems, especially concerning the sample covariance operator.

**Mean Estimator:** Recall the sample mean function:

$$\widehat{\mu}(t) := \frac{1}{N}\sum_{i=1}^{N} X_i(t) \qquad \forall t \in [0,1]$$

$\widehat{\mu}$ is consistent:

**Proposition 3.12 (c.f. [TurbThese, Prop. 2.2.4])** *Let $(X_n)_{n\in\mathbb{Z}}$ an $MAH(1)$ process and define $C_{\widehat{\mu}} := \mathbb{E}\left[\widehat{\mu}\otimes\widehat{\mu}\right]$. Then*

$$\left\| NC_{\widehat{\mu}} - (D + C + D^*) \right\|_{\mathcal{T}} \overset{N\to\infty}{\longrightarrow} 0.$$

*Consequently*

$$N\mathbb{E}\left[\|\widehat{\mu}-\mu\|^2\right] \overset{N\to\infty}{\longrightarrow} \|D + C + D^*\|_{\mathcal{T}}.$$

*Here recall the trace norm from (2.3).*

**Covariance Estimator:** Recall the sample covariance operator:

$$\widehat{C} := \frac{1}{N}\sum_{i=1}^{N} X_i \otimes X_i$$

Proving consistency of $\widehat{C}$ is much more tedious than working with $\widehat{\mu}$. Since we do not want to get lost into details of the long proofs, we only consider some sketches. The idea is to handle $(X_n \otimes X_n)_{n\in\mathbb{Z}}$, which itself turns out to be a moving average process on the Hilbert Schmidt space $\mathcal{S}$ (which itself is a Hilbert space, too), i.e. $MA\mathcal{S}(1)$. Then we can apply the same principle as for $\widehat{\mu}$.

**Lemma 3.13 (c.f. [TurbThese, Lemma 2.2.2])** *Let $(X_n)_{n\in\mathbb{Z}}$ be an $MAH(1)$ process equipped with an $H$ strong white noise $(\varepsilon_n)_{n\in\mathbb{Z}}$. Let $j_0 \in \mathbb{N}$ such that $\|l^{j_0}\| < 1$. Assume $\mathbb{E}\|X_0\|^4 < \infty$ and the cross-covariance operator $C_{X_0\otimes X_0, X_1\otimes X_1} \neq 0$. Then $(X_n \otimes X_n)_{n\in\mathbb{Z}}$ is an $MA\mathcal{S}(1)$ process with mean $C$.*

Sketch of proof: One can show that every regular, centered and (weakly) stationary process $(Y_n)_{n\in\mathbb{Z}}$ with

$$C_{Y_n,Y_{n+h}} = 0 \ \forall h > 1 \qquad \text{and} \qquad \exists l \in \mathcal{L},\ H-\text{white noise } (Z_n)_{n\in\mathbb{Z}} :\ C_{Y_n,Y_{n+1}} = lC_Z$$

is an $MAH(1)$ process with an operator $l$ and a white noise $(Z_n)_{n\in\mathbb{Z}}$ (c.f. [TurbThese, Proposition 2.2.2]). Furthermore, due to $\mathbb{E}\left[\|X_n\otimes X_n\|_{\mathcal{S}}^2\right] = \mathbb{E}\left[\|X_n\|^4\right] < \infty$ for all $n \in \mathbb{Z}$

the process $(X_n \otimes X_n - C)_{n\in\mathbb{Z}}$ is of second-order. Some tedious calculations reveal that $(X_n \otimes X_n - C)_{n\in\mathbb{Z}}$ is stationary. Therefore, define $\Gamma_h := C_{X_n\otimes X_n, X_{n+h}\otimes X_{n+h}}$. One can verify $\Gamma_h \equiv 0$ for all $h > 1$. After this, the rest follows from the claim at the beginning of this sketch of proof. □

**Proposition 3.14 (c.f. [TurbThese, Prop. 2.2.7])** *Let $(X_n)_{n\in\mathbb{Z}}$ be an $MAH(1)$ process equipped with an $H$ strong white noise $(\varepsilon_n)_{n\in\mathbb{Z}}$. Let $j_0 \in \mathbb{N}$ such that $\|l^{j_0}\| < 1$. Assume $\mathbb{E}\|X_0\|^4 < \infty$. Define for all $n \in \mathbb{Z}$*

$$Z_n := \langle X_n, \cdot\rangle X_n - C$$

*($\overline{Z}_N = \widehat{C} - C$) and $\Gamma_{\cdot,\cdot}$ to be the cross-covariance operator of two random operators of $\mathcal{S}$. Then*

$$\left\| N\Gamma_{\widehat{C}-C} - (\Gamma_{Z_0,Z_{-1}} + \Gamma_{Z_0,Z_0} + \Gamma_{Z_0,Z_1}) \right\| \overset{N\to\infty}{\longrightarrow} 0$$

*and, analogously to Proposition 3.12,*

$$N\mathbb{E}\left[\|\widehat{C} - C\|_{\mathcal{S}}^2\right] \overset{N\to\infty}{\longrightarrow} \left\|\Gamma_{Z_0,Z_{-1}} + \Gamma_{Z_0,Z_0} + \Gamma_{Z_0,Z_1}\right\|_{\mathcal{T}},$$

*where here $\mathcal{T}$ denotes the trace class of operators on $\mathcal{S}$.*

Proof: The claim follows directly from Proposition 3.12 and from Lemma 3.13. □

**Cross-covariance Estimator:** The general form of a sample cross-covariance operator looks like

$$\frac{1}{N-h}\sum_{i=1}^{N-h} X_i \otimes X_{i+h}.$$

However, we are only interested in the special case $h = 1$, i.e.

$$\widehat{D} := \frac{1}{N-1}\sum_{i=1}^{N-1} X_i \otimes X_{i+1}.$$

The idea to prove consistency of $\widehat{D}$ is very similar to the arguments concerning $\widehat{C}$. In fact, we consider the random operators

$$W_i := \langle X_i, \cdot\rangle X_{i+1} - D$$

for $i \in \{1, \ldots, N-1\}$, given $MAH(1)$ realizations $X_1, \ldots, X_N$. In contrast to $(X_n \otimes X_n)_{n\in\mathbb{Z}}$ (which is 1-dependent), $(X_n \otimes X_{n+1})_{n\in\mathbb{Z}}$ is 2-dependent. So is $(W_i)_{i\in\mathbb{Z}}$. Anyway the same limit theorems that were used for $\widehat{\mu}$ and for $\widehat{C}$ can be applied here to gain the convergence behavior

$$\mathbb{E}\left[\|\widehat{D} - D\|_{\mathcal{S}}^2\right] = \mathcal{O}\left(\frac{1}{N}\right).$$

This part is skipped here.

# Chapter 4

# Estimation of the Coefficient Operator $l$

From now on we consider an $MAH(1)$ process $(X_n)_{n\in\mathbb{Z}}$ equipped with an $H$ strong white noise $(\varepsilon_n)_{n\in\mathbb{Z}}$. Again we denote $\|\cdot\| := \|\cdot\|_{\mathcal{L}}$. Recall the equations (3.4) and (3.5). They indicate how to eliminate $C_\varepsilon$, which is intended because in practice $(\varepsilon_n)_{n\in\mathbb{Z}}$ is observed in contrast to $(X_n)_{n\in\mathbb{Z}}$. From

$$lC \overset{(3.4)}{=} lC_\varepsilon + l^2 C_\varepsilon l^* \overset{(3.5)}{=} D + l^2 D^*$$

we get

$$l^2 D^* - lC + D = 0. \tag{4.1}$$

It is an equation for $l$ depending on $C$ and $D$ (which can be estimated from the data). Unfortunately it is quadratic, which makes it hard to solve. In [TurbThese] two approaches are introduced to estimate $l$: The first one is based on projecting onto a subspace generated by the eigenvectors of $C$ and deals with solving that quadratic equation for the eigenvalues. The second one relies on regarding (4.1) as a fixed-point equation and therefore establishing a fixed-point iteration. Both of them require certain assumptions, which are explained at the beginning.

## 4.1 Assumptions

Recall Proposition 3.14. It guarantees consistency of the sample covariance matrix provided that there exists a $j_0 \in \mathbb{N}$ such that $\|l^{j_0}\| < 1$ holds and provided that $\mathbb{E}\|X_0\|^4 < \infty$ holds. The former implies in addition that $(X_n)_{n\in\mathbb{Z}}$ is invertible according to Theorem 3.10 and the latter implies in addition that $\widehat{C}$ is a Hilbert-Schmidt operator according to Theorem 2.10. Beyond those conditions we will elaborate another three.

- Note that (4.1) contains both $D$ and $D^*$, which makes it hard to solve. Thus we intend to simplify (4.1) by assuming that $D$ is self-adjoint. The property that $l$ and $C_\varepsilon$ commute leads to this.

**Lemma 4.1 (c.f. [TurbThese, ch. 2.2.4])** *Suppose all eigenspaces of $C_\varepsilon$ are of dimension 1. Then, if $l$ and $C_\varepsilon$ commute, $l$ and $D$ are symmetric operators. Moreover, $C_\varepsilon$, the eigenvectors of $C$, $D$ and $l$ coincide.*

Proof: As a covariance operator, $C_\varepsilon$ is symmetric and can be denoted by

$$C_\varepsilon = \sum_{j=1}^{\infty} \gamma_j v_j \otimes v_j.$$

Since $C_\varepsilon$ and $l$ commute, for each $j \in \mathbb{N}$ $l(v_j)$ is an eigenvector of $C_\varepsilon$ with eigenvalue $\gamma_j$

$$C_\varepsilon[l(v_j)] = l[C_\varepsilon(v_j)] = l(\gamma_j v_j) = \gamma_j l(v_j).$$

Since all eigenspaces of $C_\varepsilon$ are one-dimensional, we obtain

$$l(v_j) = \pm \|l(v_j)\| v_j,$$

i.e. the eigenvectors of $C_\varepsilon$ are collinear to their images w.r.t. $l$ and therefore are eigenvectors of $l$ too. All the other eigenvectors of $l$ can be expressed by a (possibly infinite) linear combination of $(v_j)_{j\in\mathbb{N}}$. Thus the eigenvectors of $l$ and $C_\varepsilon$ coincide. Consequently

$$l(x) = l\left(\sum_{j=1}^{\infty} \langle x, v_j\rangle v_j\right) = \sum_{j=1}^{\infty} \langle x, v_j\rangle l(v_j) = \sum_{j=1}^{\infty} \pm\|l(v_j)\| \langle x, v_j\rangle v_j$$

follows for all $x \in H$. Since the right hand side describes a symmetric operator, $l$ is symmetric. (3.5) yields that, as a combination of symmetric operators with the same eigenvectors, $D$ is symmetric and has the same eigenvectors. Due to (3.4) the same holds for $C$. □

- Recall Lemma 2.1. This states that in our case $C$ is not invertible. However, we can make $C$ invertible by restricting its domain to its image $C(H)$. We will prove

**Lemma 4.2** *If the eigenvalues of $C$ are strictly positive, then*

$$C(H) = \left\{ y \in H : \sum_{j=1}^{\infty} \frac{\langle y, v_j\rangle^2}{c_j^2} < \infty \right\} =: B.$$

Proof: "$\subseteq$" Let $y \in C(H)$. Then, there exists an $x \in H$ such that $C(x) = y$. Then

$$y = \sum_{j=1}^{\infty} c_j \langle x, v_j\rangle v_j.$$

Plugging this into fraction on the right hand side of (4.2) yields

$$\sum_{j=1}^{\infty} \frac{\langle y, v_j\rangle^2}{c_j^2} = \sum_{j=1}^{\infty} \langle x, v_j\rangle^2 = \|x\|^2 < \infty,$$

which proves $C(H) \subseteq B$.

"$\supseteq$" Let $y \in B$. Then define $x = \sum_{j=1}^{\infty} \frac{\langle y, v_j \rangle}{c_j} v_j$. Since $C$ maps $x$ to $y$

$$C(x) = \sum_{j=1}^{\infty} \langle y, v_j \rangle v_j = y$$

and $x \in H$ holds because of

$$\|x\|^2 = \sum_{j=1}^{\infty} \frac{\langle y, v_j \rangle^2}{c_j^2} < \infty \ (\Rightarrow \ x \in H),$$

$C(H) \supseteq B$ holds and thus the result follows. □

This lemma motivates us to enforce $C$ possessing strictly positive eigenvalues in order to get an intuition for $C(H)$ according to (4.2). Due to the relation

$$c_j = \langle C(v_j), v_j \rangle = \mathbb{E}\left[\langle X_0, v_j \rangle^2\right]$$

for all $j \in \mathbb{N}$, we can ensure $c_j > 0$ by

$$\mathbb{P}(\langle X_0, v_j \rangle = 0) = 0$$

for all $j \in \mathbb{N}$.

- Recall the autocorrelation function of a univariate moving average process with parameter $\theta$ such that $|\theta| < 1$ and white noise variance $\sigma_\varepsilon^2$

$$\rho(h) = \begin{cases} 1 & , h = 0, \\ \dfrac{\theta}{1+\theta^2} & , |h| = 1, \\ 0 & , |h| > 1. \end{cases}$$

Due to

$$1 \pm 2\theta + \theta^2 = (1 \pm \theta)^2 > 0$$

we get

$$2|\theta| < 1 + \theta^2.$$

This is equivalent to $|\rho(\pm 1)| < 0.5$ and thus $|\rho(h)| < 0.5$ for all $h \in \mathbb{Z}\backslash\{0\}$ as a necessary condition. Since $(1+\theta^2)\sigma_\varepsilon^2$ corresponds to (3.4) and $\theta\sigma^2$ to (3.4), we analogously require $\|\rho\| < 0.5$ for $\rho := DC^{-1}$ in the $MAH(1)$ case.

All in all we investigated the assumptions

$$\exists j_0 \geq 1 : \ \left\|l^{j_0}\right\| < 1 \tag{H1}$$

$$\mathbb{E}\|X_0\|^4 < \infty \tag{H2}$$

$$lC_\varepsilon = C_\varepsilon l \tag{H3}$$

$$\forall j \in \mathbb{N} : \ \mathbb{P}(\langle X_0, v_j \rangle = 0) = 0 \tag{H4}$$

$$\|\rho\| < \frac{1}{2}. \tag{H5}$$

Those assumptions allow us to apply the two estimation approaches for $l$ onto quadratic equation

$$l^2 D - lC + D = 0, \tag{4.2}$$

simplified by Lemma 4.1.

## 4.2 Projection method

The idea is to determine the eigenvalues of $l$ and thus to construct an approximation for $l$. For this we need the eigenfunctions $(v_j)_{j\in\mathbb{N}}$ of $C$ (and thus of $D$, $C_\varepsilon$ and $l$). Here we assume that they are known.

### 4.2.1 Methodology

Due to (H3), we are allowed to denote

$$C = \sum_{j=1}^{\infty} c_j v_j \otimes v_j,\ D = \sum_{j=1}^{\infty} d_j v_j \otimes v_j,\ l = \sum_{j=1}^{\infty} \lambda_j v_j \otimes v_j.$$

Therefore for each $k \in \mathbb{N}$, plugging an orthogonal basis element $v_k$ into each operator in (4.2) leads to

$$\lambda_k^2 d_k - \lambda_k c_k + d_k = 0, \qquad \forall k \in \mathbb{N}. \tag{4.3}$$

Thus we transform the operational equation (4.2) to infinitely many scalar quadratic equations. Since the eigenfunctions of $l$ are assumed to be known, all the $\lambda_k$ determine $l$ uniquely. Solving an equation of the form (4.3) leads to

$$\lambda_k = \frac{c_k \pm \sqrt{c_k^2 - 4d_k^2}}{2d_k}, \qquad \forall k \in \mathbb{N}. \tag{4.4}$$

The discriminant is always positive because of (H5): Consider that the operator

$$\rho := DC^{-1} = \sum_{j=1}^{\infty} \frac{d_j}{c_j} v_j \otimes v_j$$

is well defined on $C(H)$ according to (H4). Since here the $\mathcal{L}$-norm $\|\cdot\|$ takes the absolute maximal eigenvalue (because its corresponding eigenfunction maximizes the expression which defines $\|\cdot\|$) we obtain

$$\frac{|d_k|}{c_k} \leq \max_{j\in\mathbb{N}} \frac{|d_j|}{c_j} = \|\rho\| \overset{(H5)}{<} \frac{1}{2} \tag{4.5}$$

for all $k \in \mathbb{N}$, which states that the disciminant is always positive.
Now recall (H1). It only holds if $\sup_{j\in\mathbb{N}} |\lambda_j^{j_0}| < 1$ for some $j_0 \in \mathbb{N}$ and hence $\sup_{j\in\mathbb{N}} |\lambda_j| < 1$. In

other words, we require $|\lambda_k| < 1$ for all $k \in \mathbb{N}$. By contrast $\frac{c_k}{2|d_k|} > 1$ holds for all $k \in \mathbb{N}$ because of (4.5). Hence we need to force the numerator of (4.4) to become a difference

$$\lambda_k = \frac{c_k - \sqrt{c_k^2 - 4d_k^2}}{2d_k}, \qquad \forall k \in \mathbb{N}.$$

This suffices $|\lambda_k| \in (0;1)$ for all $k \in \mathbb{N}$: One can show that (H4) implies $d_k \neq 0$. Therefore $c_k - \sqrt{c_k^2 - 4d_k^2} > 0$ and $|\lambda_k| > 0$. The rest follows from

$$\begin{aligned}
(4.5) \quad &\Leftrightarrow \quad 2 < \frac{c_k}{|d_k|} \quad \Leftrightarrow \quad -\frac{c_k}{|d_k|} + 1 < -1 \quad \Leftrightarrow \quad \frac{c_k^2}{4d_k^2} - \frac{c_k}{|d_k|} + 1 < \frac{c_k^2}{4d_k^2} - 1 \\
&\Leftrightarrow \quad \left(\frac{c_k}{2|d_k|} - 1\right)^2 < \frac{c_k^2 - 4d_k^2}{4d_k^2} \quad \Leftrightarrow \quad \frac{c_k}{2|d_k|} - 1 < \frac{\sqrt{c_k^2 - 4d_k^2}}{2|d_k|} \quad \Leftrightarrow \quad \frac{c_k - \sqrt{c_k^2 - 4d_k^2}}{2|d_k|} < 1 \\
&\Leftrightarrow \quad |\lambda_k| < 1 \qquad (\text{because } c_k - \sqrt{c_k^2 - 4d_k^2} > 0).
\end{aligned}$$

In this section we assumed that the eigenfunctions $(v_j)_{j\in\mathbb{N}}$ are known, but not the eigenvalues of $C$ and $D$. Therefore we need to estimate them in order to estimate the ones of $l$. Given $n \in \mathbb{N}$ and functional observations $X_1, \ldots, X_n$ in $L^2$, the intuitive estimators are

$$c_{kn} = \frac{1}{n}\sum_{i=1}^{n} \langle X_i, v_k\rangle^2, \;\; d_{kn} = \frac{1}{n-1}\sum_{i=1}^{n-1} \langle X_i, v_k\rangle\langle X_{i+1}, v_k\rangle,$$

gained by replacing the expectation operator by the empirical average. Then $(c_{kn})_{k\in\mathbb{N}}$ and $(d_{kn})_{k\in\mathbb{N}}$ satisfy the condition of nuclear operators

$$\sum_{k=1}^{\infty} |c_{kn}| = \frac{1}{n}\sum_{i=1}^{n} \|X_i\|^2 < \infty, \;\; \sum_{k=1}^{\infty} |d_{kn}| \leq \frac{1}{n-1}\sum_{i=1}^{n-1} \|X_i\|\|X_{i+1}\| < \infty.$$

However, we do not know if, for all $k \in \mathbb{N}$, $c_{kn} > 2|d_{kn}| > 0$ holds according to (4.5). Therefore we define $\delta > 0$ sufficienty small and adapt $(c_{kn})_{k\in\mathbb{N}}$ to

$$c'_{kn} = \max\left(c_{kn}, (2+\delta)|d_{kn}|\right).$$

Then $\lambda_{kn}$ is defined to be the absolute lower solution of

$$\lambda_{kn}^2 d_{kn} - \lambda_{kn} c'_{kn} + d_{kn} = 0$$

for all $k \in \mathbb{N}$. Finally choose an appropriate $p_n \in \mathbb{N}$ (e.g. the optimal number of principal components when using PCA) and estimate $l$ by

$$l_{n,p_n} := \sum_{k=1}^{p_n} \lambda_{kn} v_k \otimes v_k.$$

This is the *estimator by projection*. The mean squared error of $l_{n,p_n}$ can be expressed by the eigenvalues:

$$\begin{aligned}
\mathbb{E}\left[\|l_{n,p_n} - l\|_{\mathcal{S}}^2\right] &= \mathbb{E}\left[\sum_{i=1}^{\infty} \langle l_{n,p_n}(v_i) - l(v_i), l_{n,p_n}(v_i) - l(v_i)\rangle\right] \\
&= \sum_{i=1}^{p_n} \mathbb{E}\left[(\lambda_{kn} - \lambda_k)^2\right] + \sum_{i=p_n+1}^{\infty} \lambda_k^2.
\end{aligned}$$

**Remark:** We require the eigenvectors to be known and the interchangeability of $l$ and $C_\varepsilon$ to hold, which unfortunately is unlikely for real data and therefore a hard restriction. Nevertheless the projection approach gives an intuitive and comprehensive attempt to estimate $l$ directly.

### 4.2.2 Convergence Criteria

We will prove that $(c_{kn})_{n\in\mathbb{N}}$ and $(d_{kn})_{n\in\mathbb{N}}$ converge to $c_k$ and respectively $d_k$. Consequently $(c'_{kn})_{n\in\mathbb{N}}$ and $(\lambda_{kn})_{n\in\mathbb{N}}$ are consistent too under certain assumptions. Eventually this will yield the consistence of the estimator by projection $l_{n,p_n}$.

**Lemma 4.3 (c.f. [TurbThese, Lemma 3.2.1])** *Suppose (H1)-(H5). Then*

$$\mathbb{E}\left[\sup_{k\in\mathbb{N}}|c_{kn}-c_k|^2\right]=\mathcal{O}\left(\frac{1}{n}\right)\qquad\text{and}\qquad\sup_{k\in\mathbb{N}}|c_{kn}-c_k|\overset{\text{a.s.}}{\to}0.$$

**Lemma 4.4 (c.f. [TurbThese, Lemma 3.2.2])** *Suppose (H1)-(H5). Then*

$$\mathbb{E}\left[\sup_{k\in\mathbb{N}}|d_{kn}-d_k|^2\right]=\mathcal{O}\left(\frac{1}{n}\right)\qquad\text{and}\qquad\sup_{k\in\mathbb{N}}|d_{kn}-d_k|\overset{\text{a.s.}}{\to}0.$$

**Lemma 4.5 (c.f. [TurbThese, Lemma 3.2.5], slightly modified)** *Suppose that (H1)-(H5) hold. Then,*

$$\mathbb{E}\left[(c'_{kn}-c_k)^2\right]\le 2\mathbb{E}\left[(c_{kn}-c_k)^2\right]+\mathcal{O}\left(\frac{1}{n}\right).$$

Sketch of proof: For any $a,b\in\mathbb{R}$ consider the fact

$$a^2-2ab+b^2=(a-b)^2\ge 0\ \Rightarrow 2ab\le a^2+b^2\ \Rightarrow (a+b)^2\le 2a^2+2b^2- \tag{4.6}$$

From the equation (as a direct consequence of the definition of $c'_{kn}$)

$$c'_{kn}-c_k=((2+\delta)|d_{kn}|-c_k)\mathbb{1}_{c_{kn}\le(2+\delta)|d_{kn}|}+(c_{kn}-c_k)\mathbb{1}_{c_{kn}\le 2d_{kn}}$$

it follows from Cauchy-Schwarz inequality and (4.6) that

$$\begin{aligned}\mathbb{E}\left[(c'_{kn}-c_k)^2\right] &\le 2\mathbb{E}\left[(c_{kn}-c_k)^2\right]+2\mathbb{E}\left[((2+\delta)|d_{kn}|-c_k)^2\mathbb{1}_{c_{kn}\le(2+\delta)|d_{kn}|}\right]\\ &\le 2\mathbb{E}\left[(c_{kn}-c_k)^2\right]+2\sqrt{\mathbb{E}\left[((2+\delta)|d_{kn}|-c_k)^4\right]}\sqrt{\mathbb{P}(c_{kn}\le(2+\delta)|d_{kn}|)}\end{aligned}$$

holds. Lemma 4.3 guarantees the convergence of the first summand. Thus we only consider the second one. Due to (H5) we know $(2|d_k|-c_k)<0$. Hence $((2+\delta)|d_k|-c_k)<0$ follows, because $\delta>0$ is supposed to be chosen sufficiently small. Consequently

$$(2+\delta)|d_{kn}|-c_k=(2+\delta)(|d_{kn}|-|d_k|)+((2+\delta)|d_k|-c_k)<(2+\delta)(|d_{kn}|-|d_k|)$$

follows as well as

$$((2+\delta)|d_{kn}| - c_k)^4 < (2+\delta)^4(|d_{kn}| - |d_k|)^4 \leq (2+\delta)^4(d_{kn} - d_k)^4.$$

All in all it remains to handle $\sqrt{\mathbb{E}\left[(d_{kn} - d_k)^4\right]}$ and $\sqrt{\mathbb{P}(c_{kn} \leq (2+\delta)|d_{kn}|)}$. One can show

$$\mathbb{E}\left[(d_{kn} - d_k)^4\right] = \mathcal{O}\left(\frac{1}{n^2}\right).$$

Some arguments from probability theory yield that $\sqrt{\mathbb{P}(c_{kn} \leq (2+\delta)|d'_{kn}|)}$ has an upper bound consisting of $\mathbb{E}\left[(c_{kn} - c_k)^2\right]$ and $\mathbb{E}\left[(d_{kn} - d_k)^2\right]$, so that

$$\sqrt{\mathbb{E}\left[(d'_{kn} - d_k)^4\right]}\sqrt{\mathbb{P}(c_{kn} \leq (2+\delta)|d'_{kn}|)} = \mathcal{O}\left(\frac{1}{n}\right)$$

gives the result. □

**Lemma 4.6 (c.f. [TurbThese, Lemma 3.2.6])** *Suppose (H1)-(H5). Then,*

$$\mathbb{E}\left[(\lambda_{kn} - \lambda_k)^2\right] = \mathcal{O}\left(\frac{1}{n}\right) \qquad \forall k \in \mathbb{N}.$$

Sketch of proof: The idea is to calculate

$$\mathbb{E}\left[(\lambda_{kn} - \lambda_k)^2\right] < \left(\frac{\lambda_k}{1-\lambda_k}\right)^2 \mathbb{E}\left[\left(\frac{c'_{kn}}{d_{kn}} - \frac{c_k}{d_k}\right)^2\right].$$

The latter term can be estimated by second and fourth moments of $(c_{kn} - c_k)$ and $(d_{kn} - d_k)$. The calculations are skipped because they are very tedious (see [TurbThese, ch. 3.4]). It remains to check that the fourth moment of $(c_{kn} - c_k)$ is of $\mathcal{O}\left(\frac{1}{n^2}\right)$ too. Then the result follows. □

**Proposition 4.7 (c.f. [TurbThese, Proposition 3.3.1])** *Suppose (H1)-(H5). Let $p_n \in \mathbb{N}$ such that $\alpha_n \leq \frac{d_{p_n}}{2}$. Then,*

$$\mathbb{E}\left[\|l_{n,p_n} - l\|_{\mathcal{S}}^2\right] \leq \frac{1}{n}\sum_{k=1}^{p_n}\frac{2}{d_k^2} + \sum_{k>p_n}\frac{d_k^2}{c_k^2}$$

*holds as an as an asymptotical (not exact) inequality.*

Sketch of proof: We already showed

$$\mathbb{E}\left[\|l_{n,p_n} - l\|_{\mathcal{S}}^2\right] = \sum_{i=1}^{p_n}\mathbb{E}\left[(\lambda_{kn} - \lambda_k)^2\right] + \sum_{i=p_n+1}^{\infty}\lambda_k^2.$$

One result of the tedious calculations which were mentioned in the proof of Lemma 4.6 is

$$\sum_{i=1}^{p_n}\mathbb{E}\left[(\lambda_{kn} - \lambda_k)^2\right] \leq \frac{1}{n}\sum_{i=1}^{p_n}\left(\frac{\lambda_k^2}{d_k^2} + \frac{\lambda_k^2 c_k^2}{d_k^4}\right)$$

as an asymptotic inequality. Furthermore the equation $\lambda_k^2 d_k - \lambda_k c_k + d_k = 0$ is equivalent to

$$\lambda_k c_k = \left(\lambda_k^2 + 1\right) d_k.$$

Since $\lambda_k \in (0;1)$, we obtain (if $d_k > 0$)

$$d_k < \left(\lambda_k^2 + 1\right) d_k < 2d_k.$$

Altogether this leads to the inequalities

$$\frac{d_k}{c_k} < \lambda_k < 2\frac{d_k}{c_k},$$

which states the asymptotic equivalence $\lambda_k \simeq \frac{d_k}{c_k}$. The other case $d_k < 0$ results in

$$2\frac{d_k}{c_k} < \lambda_k < \frac{d_k}{c_k},$$

so that asymptotic equivalence does not change. Hence

$$\mathbb{E}\left[\|l_{n,p_n} - l\|_{\mathcal{S}}^2\right] \leq \frac{1}{n}\sum_{i=1}^{p_n}\left(\frac{1}{c_k^2} + \frac{1}{d_k^2}\right) + \sum_{i=p_n+1}^{\infty}\frac{d_k^2}{c_k^2}$$

follows. Since $c_k > 2|d_k|$ holds, the inequality

$$\frac{1}{c_k^2} + \frac{1}{d_k^2} < \frac{2}{d_k^2}$$

holds and thus gives the result. □

**Theorem 4.8 (c.f. [TurbThese, Theorem 3.3.2])** *Suppose (H1)-(H5). Moreover assume that there exists some numbers $r, r'$ with $0 < r' < r < 1$ such that for all $k \in \mathbb{N}$*

$$c_k = r^k,\ d_k = r'^k \tag{4.7}$$

*holds. Furthermore define $b = \left(2\log(r) - 4\log(r')\right)^{-1}$ and choose $p_n = b\log(n)$. Then*

$$\mathbb{E}\left[\|l_{n,p_n} - l\|_{\mathcal{S}}^2\right] = \mathcal{O}\left(\frac{1}{\sqrt{n}}\right)$$

Sketch of proof: Combining (4.7) and Proposition 4.7 yields the asymptotic inequality

$$\mathbb{E}\left[\|l_{n,p_n} - l\|_{\mathcal{S}}^2\right] \leq \frac{2}{n}\sum_{k=1}^{p_n}\left(\frac{1}{r'}\right)^{2k} + \sum_{k>p_n}\left(\frac{r'}{r}\right)^{2k}.$$

On the one hand we have

$$\frac{2}{n}\sum_{k=1}^{p_n}\left(\frac{1}{r'}\right)^{2k} \leq \frac{p_n}{n}\left(\frac{1}{r'}\right)^{2p_n},$$

which leads to

$$\frac{2}{n}\sum_{k=1}^{p_n}\left(\frac{1}{r'}\right)^{2k} \overset{n \text{ big}}{\simeq} \frac{1}{n}\left(\frac{1}{r'}\right)^{2p_n}.$$

On the other hand we have

$$\sum_{k>p_n}\left(\frac{r'}{r}\right)^{2k} \simeq \left(\frac{r'}{r}\right)^{2p_n}.$$

Equating the first asymptotical result with the second one

$$\frac{1}{n}\left(\frac{1}{r'}\right)^{2p_n} = \left(\frac{r'}{r}\right)^{2p_n}$$

in order to get one single expression for the asymptotic behavior of both sums results in

$$n = \left(\frac{r^2}{r'^4}\right)^{p_n}$$

and thus in

$$p_n = b\log n \text{ with } b = \frac{1}{2\log r - 4\log r'}.$$

This is the reason for that specific choice of $p_n$. Furthermore define $\gamma := \frac{\log r'}{\log r}$. The result

$$\left(\frac{r'}{r}\right)^{2p_n} = \exp\left(\frac{2(\log r' - \log r)}{2\log r - 4\log r'}\log n\right) = n^{-\frac{1-\gamma}{1-2\gamma}} \overset{\gamma\to\infty}{\longrightarrow} \frac{1}{\sqrt{n}}$$

follows, because due to $0 < r' < r < 1$ we have $\gamma \to \infty$. □

## 4.3 Iterative method

The idea is to solve (4.2) as a fixed-point equation, c.f. [Turb2]. It is based on a recursion that was designed to output the square root of a symmetric positive definite operator. According to [Turb1] it even solves (4.1), but we keep on working with (4.2) for simplicity.

### 4.3.1 Inspiration by Riesz-Nagy

Suppose we want to find a symmetric operator $A$ such that

$$A^2 - 2A + R = 0 \tag{4.8}$$

holds. Here $R$ is assumed to be a positive definite and symmetric operator such that $\langle x, Rx\rangle \leq \|x\|^2$ for all $x \in H$. Rewrite (4.8) as

$$A = \frac{1}{2}(R + A^2),$$

such that the recursive formulas

$$A_0 := 0, \qquad A_{r+1} := \frac{1}{2}(R + A_r^2), \qquad \forall r \geq 0,$$

can be established.

**Lemma 4.9 (Riesz-Nagy, c.f. [Riesz-Nagy, ch. 104])** *If $R$ is as above, this recursion converges to the unique solution of (4.8).*

This claim was stated and proven in order to show the existence and uniqueness of a square root $X$ for any positive-definite symmetric operator $B$. Defining $A$ and $R$ such that $X = Id - A$ and $B = Id - R$ hold implies

$$X^2 = B \qquad \Longleftrightarrow \qquad A^2 - 2A + R = 0.$$

Since this recursion is not the main focus of this thesis, we skip the proof of Lemma 4.9, although it is not too hard and long. It is based on the next theorem. Here *monotonicity* of a sequence of operators $(A_k)_{k\in\mathbb{N}}$ means that all increments $A_k - A_{k-1}$ are positive definite for $k \in \mathbb{N}_0$.

**Theorem 4.10 (Vigier, c.f. [Vigier])** *Every monotone bounded (w.r.t. $\|\cdot\|_{\mathcal{L}}$) sequence $(A_k)_{k\in\mathbb{N}}$ of symmetric operators converges strongly to a symmetric operator $A$, i.e.*

$$\|A_k h - Ah\| \overset{k\to\infty}{\longrightarrow} 0, \qquad \forall h \in E,$$

*where $E$ is a linear subspace out of the Hilbert space $H$.*

In light of Riesz-Nagy and according to (4.2) we want to derive a solution of

$$A^2 R - A + R = 0 \tag{4.9}$$

with a positive-definite symmetric operator $R$ such that $\|R\| < \frac{1}{2}$ by recursion:

$$A_0 := 0, \qquad A_{r+1} := A_r^2 R + R, \qquad \forall r \geq 0.$$

It is easy to see that $\|A_r\| < 1$ for all $r \in \mathbb{N}_0$ follows from $\|R\| < \frac{1}{2}$: Trivially, this holds for $r \in \{0, 1\}$. The rest follows by induction by seeing that

$$\|A_{r+1}\| = \|A_r^2 R + R\| \leq \|A_r^2 + Id\| \cdot \|R\| \leq \frac{1}{2}(\|A_r^2\| + 1) < 1.$$

The next proposition states that this sequence converges.

**Proposition 4.11 (c.f. [TurbThese, Prop. 4.2.1])** *Let $(A_r)_{r\in\mathbb{N}_0}$ and $R$ as above. Then there is a symmetric operator $A$ that suffices (4.9) and*

$$\|A_r - A\| \leq (2\|R\|)^r \overset{r\to\infty}{\longrightarrow} 0.$$

Proof: The idea is to apply Vigier's Theorem. Therefore, it suffices to show

- $A_r$ is symmetric for all $r \in \mathbb{N}_0$
- $(A_r)_{r\in\mathbb{N}_0}$ is monotonically increasing, i.e. $\langle h, A_1 h\rangle \leq \langle h, A_1 h\rangle \leq \ldots$ for all $h \in H$

analogously to the Riesz-Nagy recursion.

(i) $A_r$ symmetric for all $r \in \mathbb{N}_0$: Note that the combination of two commutative symmetric operators $M$,$P$ is symmetric, too:

$$\langle MPx, y\rangle = \langle Px, My\rangle = \langle x, PMy\rangle = \langle x, MPy\rangle \qquad \forall x, y \in H$$

Therefore, it suffices to show that $A_r$ and $R$ commute for all $r \in \mathbb{N}_0$. Obviously this holds for $r = 0, 1$. Assume it holds for some $r$. Then it trivially holds for $r+1$ because of

$$A_{r+1}R = (A_r^2R + R)R = A_r^2R^2 + R^2 = A_rRA_rR + R^2 = RA_r^2R + R^2 = RA_{r+1}.$$

Commutativity and symmetry of $R$ lead to symmetry of $A_r$ for all $r \in \mathbb{N}_0$, which again can be shown by induction: Clearly $A_0$ and $A_1$ are symmetric. Under the assumption that $A_r$ is symmetric up to some $r > 0$ we have for all $x, y \in H$

$$\langle A_{r+1}x, y\rangle = \langle A_r^2Rx, y\rangle + \langle Rx, y\rangle = \langle x, RA_r^2y\rangle + \langle x, Ry\rangle = \langle x, A_{r+1}y\rangle.$$

Hence all $A_r$ are symmetric.

(ii) $(A_r)_{r\in\mathbb{N}_0}$ monotonically increasing: As a polynomial of the positive-definite operator $R$, $A_r$ is positive definite for each $r \in \mathbb{N}$. The same holds for the increments: Clearly, $A_1 - A_0 = R$ is positive definite. Given $r > 0$,

$$A_{r+1} - A_r = (A_r^2 - A_{r-1}^2)R = \big(A_r(A_r - A_{r-1}) + (A_r - A_{r-1})A_{r-1}\big)R$$

states by induction that all increments are positive definite. Hence $(A_r)_{r\in\mathbb{N}_0}$ is monotonically increasing.

Furthermore recall that $\|A_r\| < 1$ for all $r \in \mathbb{N}_0$ because of $\|R\| < \frac{1}{2}$. Hence Vigier's theorem can be applied on $(A_r)_{r\in\mathbb{N}_0}$, which means that there exists a symmetric operator $A$ such that

$$A = \lim_{r\to\infty} A_{r+1} = \lim_{r\to\infty} A_r^2R + R = A^2R + R.$$

Since $\|A_r\| < 1$ for all $r \in \mathbb{N}$, $\|A\| \leq 1$ follows. Given that limit, we can conclude

$$\begin{aligned}\|A_{r+1} - A\| &\leq \|R\|\|A_r^2 - A^2\| \leq \|R\|\big(\|A_r\| + \|A\|\big)\|A_r - A\| \\ &\leq 2\|R\|\|A_r - A\| \leq \big(2\|R\|\big)^2\|A_{r-1} - A\| \leq \ldots \leq \big(2\|R\|\big)^{r+1} \qquad \square\end{aligned}$$

This proposition indicates how to calculate the solution of

$$l^2\rho - l + \rho = 0 \tag{4.10}$$

with $\rho := DC^{-1}$. (4.10) is equivalent to (4.2) under (H4). Since $\rho$ is not known, we need to estimate $\rho$.

### 4.3.2 Estimation of $\rho$

If the eigenfunctions $v_1, v_2, \ldots$ of $C$ are known, we can work analogously to the derivation of the projection method. Due to (H3) and Lemma 4.1, we can restrict our calculations on the eigenvalues of $D$ and $C$. Therefore for a fixed $n \in \mathbb{N}$ we define the estimator

$$\widehat{\rho}_{n,p_n}(x) := \sum_{j=1}^{p_n} \frac{d_{jn}}{c_{jn}} \langle x, v_j \rangle v_j, \qquad \forall x \in H \text{ a.e.}, \tag{4.11}$$

for some $p_n \in \{1, \ldots, n\}$ where $(d_{jn})_{j\in\mathbb{N}}$ and $(c_{jn})_{j\in\mathbb{N}}$ again denote the estimated eigenvalues of $D$ and respectively $C$

$$c_{kn} = \frac{1}{n} \sum_{i=1}^{n} \langle X_i, v_k \rangle^2, \; d_{kn} = \frac{1}{n-1} \sum_{i=1}^{n-1} \langle X_i, v_k \rangle \langle X_{i+1}, v_k \rangle.$$

**Lemma 4.12 (Consistency of $\widehat{\rho}_{n,p_n}$, c.f. [Bosq, Prop. 4.3.1])** *Given an $MAH(1)$ process $(X_n)_{n\in\mathbb{Z}}$, suppose that $\rho$ is a Hilbert-Schmidt operator. Then under (H1)-(H5) $\|\widehat{\rho}_{n,p_n} - \rho\|$ converges to 0 almost surely, provided that*

$$\liminf_{n\to\infty} \frac{n c_{p_n}^8}{\log(n)^\alpha} > 0$$

*holds for all $\alpha > 2$.*

Sketch of proof: The idea is to verify that a term containing $\|\widehat{C} - C\|$ and $\|\widehat{D} - D\|$ converges to 0 almost surely and is an upper bound for $\|\widehat{\rho}_{n,p_n} - \rho\|$ (see [Bosq, Sec. 4.4]). □

According to [Bosq, ch. 8.3] a generalization of (4.11) for the case that the eigenfunctions of $C$ are not known or that (H3) does not hold is

$$\widehat{\rho}^{\text{gen.}}_{n,p_n}(x) := \frac{1}{n-1} \sum_{i=1}^{n-1} \sum_{l=1}^{p_n} \sum_{j=1}^{p_n} \frac{1}{c_{jn}} \langle x, \widehat{v}_j \rangle \langle X_i, \widehat{v}_j \rangle \langle X_{i+1}, \widehat{v}_l \rangle \widehat{v}_l$$

with some estimators $\widehat{v}_1, \widehat{v}_2, \ldots$ for the theoretical eigenfunctions of $C$ (for example by principal component analysis in `R`).

**Lemma 4.13 (Consistency of $\widehat{\rho}^{\text{gen.}}_{n,p_n}$, c.f. [Bosq, Thm. 8.7] and [Turb1, Section 3])** *Given an $MAH(1)$ process $(X_n)_{n\in\mathbb{Z}}$, suppose that the eigenvalues $(c_n)_{n\in\mathbb{N}}$ of the theoretical covariance operator $C$ suffice*

$$c_1 > c_2 > \ldots > 0$$

*and the eigenvalues $c_{p_n n}$ of the empirical covariance operator $\widehat{C}$ (depending on $n$) suffice*

$$c_{p_n n} > 0, \qquad \forall n \in \mathbb{N} \text{ (a.s.)}.$$

*Furthermore define the sequence $(b_n)_{n\in\mathbb{N}}$ such that*

$$b_1 = 2\sqrt{2}(c_1 - c_2)^{-1}, \; b_j = 2\sqrt{2} \max\left[(c_{j-1} - c_j)^{-1}, (c_j - c_{j+1})^{-1}\right] \; \forall j \geq 2.$$

*If there exists a number $\beta > 1$ such that*

$$c_p n^{-1} \sum_{j=1}^{p_n} b_j = \mathcal{O}(n^{\frac{1}{4}} (\log n)^{-\beta}),$$

*then $\left\| \widehat{\rho}^{\text{gen.}}_{n,p_n} - \rho \right\|$ converges to zero almost surely.*

Sketch of proof: The idea is to use consistency of $\widehat{C}$ and $\widehat{D}$ (see [Bosq, Section 8.6]). □

**Remark:** The following analogy is useful for practical implementation: Suppose we work in a finite dimensional space $\mathbb{R}^n$ instead of $H$. Denote $\widetilde{V} \in \mathbb{R}^{n \times p_n}$ the matrix with $\widehat{v}_1, \ldots, \widehat{v}_{p_n}$ in the columns. Furthermore, define $\widetilde{D}_n$ to be the matrix with

$$(\widetilde{D}_n)_{l,j} = \frac{1}{n-1} \sum_{i=1}^{n-1} \langle X_i, \widehat{v}_j \rangle \langle X_{i+1}, \widehat{v}_l \rangle$$

and still $\widetilde{C}_n^{-1} = \operatorname{diag}(\frac{1}{c_{1n}}, \ldots, \frac{1}{c_{p_n n}})$. Then the $p_n \times p_n$ matrix $\widehat{\rho}^{\text{gen.}}_{n,p_n}$ can be written as

$$\widehat{\rho}^{\text{gen.}}_{n,p_n} = \widetilde{V} \widetilde{D}_n \widetilde{C}_n^{-1} \widetilde{V}^T,$$

which is a useful notation for programming.

### 4.3.3 Recursive Approach

At the end we are ready to establish the iterative method for $l$. As written in [Turb1, Section 3], (H3) is not necessary for this approach. This means that the iterative method finds a solution even for the generalized problem (4.1), where $D$ is not symmetric. In that case we have to use $\widehat{\rho}^{\text{gen.}}_{n,p_n}$. However, those claims are not proven here. We will argue the convergence of the iterative method given $\widehat{\rho}_{n,p_n}$ in (4.11). Moreover in addition to those five assumptions, which were already required for the projection method, we need

$$\langle x, \rho x \rangle \geq 0, \qquad \forall x \in C(H), \tag{H6}$$

for convergence of the recursion according to Proposition 4.11.
For the next Proposition we need the set

$$\Omega_0 := \left\{ \omega_0 \mid \exists N(\omega_0) : \ \forall n \geq N(\omega_0) : \ \|\widehat{\rho}_{n,p_n}\| < \frac{1}{2} \right\}.$$

Due to (H5) and due to consistency of $\widehat{\rho}_{n,p_n}$ this set suffices $\mathbb{P}(\Omega_0) = 1$.

**Proposition 4.14 (c.f. [TurbThese, Prop. 4.3.2])** *Fix $n \in \mathbb{N}$ and $p_n \in \{1, \ldots, n\}$. Under (H1) – (H6) the sequence*

$$\widehat{l}_{0,n,p_n} \equiv 0, \qquad \widehat{l}_{r+1,n,p_n} = \widehat{l}^2_{r,n,p_n} \widehat{\rho}_{n,p_n} + \widehat{\rho}_{n,p_n}, \qquad \forall r \in \mathbb{N},$$

*defined for all* $\omega_0 \in \Omega_0$ *and for all* $n \geq N(\omega_0)$*, converges to the solution* $\widehat{l}_{n,p_n}$ *of*

$$\widehat{l}^2_{n,p_n}\widehat{\rho}_{n,p_n} - \widehat{l}_{n,p_n} + \widehat{\rho}_{n,p_n} = 0 \tag{4.12}$$

*for* $r \to \infty$ *with* $\|\widehat{l}_{n,p_n}\| < 1$ *and with convergence rate*

$$\left\|\widehat{l}_{r,n,p_n} - \widehat{l}_{n,p_n}\right\| \leq (2\,\|\widehat{\rho}_{n,p_n}\|)^{r+1}.$$

This immediately follows from Proposition 4.11.

**Theorem 4.15 (c.f. [TurbThese, Thm. 4.3.1])** *If* $\rho$ *is a Hilbert-Schmidt operator, then under (H1) – (H6)*

$$\left\|\widehat{l}_{r,n,p_n} - l\right\| \overset{r,n\to\infty}{\longrightarrow} 0 \qquad \text{a.s.}$$

*(first* $r \to \infty$*, then* $n \to \infty$*) follows, provided* $\liminf_{n\to\infty} \frac{nc^8_{p_n}}{\log(n)^\alpha} > 0$ *for all* $\alpha > 2$*.*

Proof: Proposition 4.14 yields $\widehat{l}_{r,n,p_n} \overset{r\to\infty}{\longrightarrow} \widehat{l}_{n,p_n}$. Thus it remains to show $\widehat{l}_{n,p_n} \overset{n\to\infty}{\longrightarrow} l$:

$$\begin{aligned}
\left\|\widehat{l}_{n,p_n} - l\right\| &= \left\|\widehat{l}^2_{n,p_n}\widehat{\rho}_{n,p_n} + \widehat{\rho}_{n,p_n} - l^2\rho - \rho\right\| \\
&\leq \left\|\widehat{l}^2_{n,p_n}\widehat{\rho}_{n,p_n} - l^2\rho\right\| + \|\widehat{\rho}_{n,p_n} - \rho\| \\
&\leq \left\|\widehat{l}^2_{n,p_n}\left(\widehat{\rho}_{n,p_n} - \rho\right) + \left(\widehat{l}^2_{n,p_n} - l^2\right)\rho\right\| + \|\widehat{\rho}_{n,p_n} - \rho\| \\
&\leq \left\|\widehat{l}^2_{n,p_n}\right\| \|\widehat{\rho}_{n,p_n} - \rho\| + \left\|\widehat{l}^2_{n,p_n} - l^2\right\| \|\rho\| + \|\widehat{\rho}_{n,p_n} - \rho\| \\
&\leq \left(\left\|\widehat{l}^2_{n,p_n}\right\| + 1\right) \|\widehat{\rho}_{n,p_n} - \rho\| + \left(\left\|\widehat{l}_{n,p_n}\right\| + \|l\|\right) \left\|\widehat{l}_{n,p_n} - l\right\| \|\rho\|
\end{aligned}$$

Consequently, $\left\|\widehat{l}_{n,p_n} - l\right\|$ depends on $\|\widehat{\rho}_{n,p_n} - \rho\|$, because $\left\|\widehat{l}_{n,p_n}\right\| < 1$ according to Proposition 4.14 and hence $\|l\| \leq 1$

$$\left\|\widehat{l}_{n,p_n} - l\right\| \leq \frac{\left\|\widehat{l}^2_{n,p_n}\right\| + 1}{1 - \left(\left\|\widehat{l}_{n,p_n}\right\| + \|l\|\right)\|\rho\|} \|\widehat{\rho}_{n,p_n} - \rho\| \leq \frac{2}{1 - 2\,\|\rho\|} \|\widehat{\rho}_{n,p_n} - \rho\|\,.$$

The rest follows from Proposition 4.12. □

# Chapter 5

# Implementation of the Estimation Approaches

This chapter is the link from theory to practice. Here we apply the estimation approaches derived in the previous chapter on simulated data and compare them with two other estimation methods. At first, the structure of the implementation is described. Then several examples are discussed.

## 5.1 Structure of the implementation

The following describes how to simulate $MAH(1)$ data and how to implement the estimation errors. It does not go into details of coding in `R`, but gives some useful hints and remarks.

(i) We **simulate a** $MAH(1)$ **process** in the multivariate way with vector length $m$. This can be converted in functional data by using the `fda` package in `R`. Then, we apply the estimation procedures either on the multivariate or on the functional data. We generate an multivariate strong white noise matrix $E \in \mathbb{R}^{(b+n+1)\times m}$, where the $i$-th row denotes the (discrete) white noise element $\varepsilon_i$. Depending on $l$ and the white noise, we simulate our process $\mathtt{X} \in \mathbb{R}^{(n+b)\times m}$ according to the moving average formula. (Apparently, we need one white noise observation more than the wished number of moving average elements.) Since a simulation needs a warm-up period to approximately follow the given model formula, the initial observations of `X` might be bad estimates for an $MA$ process. Thus, we delete the first $b$ observations, i.e. the number of observations of our simulated moving average process `X` is actually $n$. It is hard to determine an appropriate white noise. There are plenty of possibilities, e.g. i.i.d. multivariate normally distributed vectors with mean vector 0 and covariance matrix $\Sigma$, Brownian motion/bridge and trigonometric linear combination with i.i.d. standard normally distributed random variables. The choice of the white noise affects the estimators decisively!
Having determined the white noise, we know the theoretical covariance matrix $C_\varepsilon$. Consequently, we are able to calculate $D = lC_\varepsilon$ and $C = lC_\varepsilon l^* + C_\varepsilon$. Then we can check whether $\|DC^{-1}\| < \frac{1}{2}$ holds.

If $m$ is big, it makes sense to convert our simulated multivariate process to a functional one. Note that the formula for the functional $MAH(1)$ process is of the form

$$X_{n+1}(t) = \int_0^1 l(s,t)\varepsilon_n(s)ds + \varepsilon_{n+1}(t) \approx \frac{1}{m}\sum_{j=0}^{m} l(s_j,t)\varepsilon_n(s_j) + \varepsilon_{n+1}(t),$$

which means that for the functional process we have to divide the matrix-vector multiplication by $m$ before transforming into functional data. We use $B \approx 51$ B-Spline basis functions on $[0,1]$ and choose $t \in \mathbb{R}^m$ as argument values for the functional data. However, only the first $n-1$ observations are converted. It is because we will apply our estimation procedures only on the first $n-1$ observations. Having estimated $l$, we will calculate a one-step predictor estimate and compare it with the remaining observation that will not have been used for the estimation approaches (i.e. we will analyze prediction errors). This holds for the multivariate implementation too. The following is implemented both for the multivariate and for the functional case in the same way.

(ii) We perform the **projection method** for each simulation. Firstly, we perform principal component analysis on the simulated process in order to gain the eigenelements of the sample covariance matrix. Given some $k_n \in \mathbb{N}$, the eigenvalues are the estimates $(c_{kn})_{k\in\{1,\dots,k_n\}}$ and the eigenvectors/-functions are $(\widehat{v}_{kn})_{k\in\{1,\dots,k_n\}}$; the latter will be needed for backtransformation from singular value representation to the "original" one. By looking at CPV$(\cdot)$ and comparing it with the threshold 85% (see the end of subsection 2.2), we can determine the optimal number $p_n$ of principal components. Depending on the situation we will either use $\min\{k_n, m, p_n\}$ or $\min\{k_n, m\}$ as the number of principal components, again denoted as $k_n$. For small $m$ (e.g. 3) using the optimal number of principal components does not make sense, because reducing the dimension is not necessary and leads to loss of information. However, for big $m$ using $p_n$ principal components is strongly recommended in order to decrease the computation duration.
PCA in `R` yields the scores $\big(\langle X_i, v_j\rangle\big)_{i\in[n],j\in[m]}$. Thus we can calculate $(d_{kn})_{k\in\{1,\dots,k_n\}}$ directly. Another option is to compute the matrix $\widetilde{D}_n$ (see subsection about the iterative method) given by

$$(\widetilde{D}_n)_{kj} := \frac{1}{n-1}\sum_{i=1}^{n-1}\langle \widehat{v}_k, X_{i+1}\rangle\langle X_i, \widehat{v}_j\rangle, \qquad \forall k,j \in \{1,\dots,k_n\},$$

which we will need in the iterative approach, and to take its diagonal entries. After this we calculate

$$c'_{kn} = \max\left(c_{kn}, (2+\delta)|d_{kn}|\right), \qquad \forall k \in \{1,\dots,k_n\}.$$

The result of the projection procedure is the vector $(\lambda_{kn})_{k\in\{1,\dots,k_n\}}$. After converting it into a diagonal matrix, it is backtransformed via $V\Lambda V^T = L_{\text{pr}}$ by using the first $k_n$ principal components in the columns of $V \in \mathbb{R}^{n\times k_n}$.

(iii) We perform the **iterative method** for each simulation. For this we need the estimates $(c_{kn})_{k\in\{1,\dots,k_n\}}$, $(\widehat{v}_{kn})_{k\in\{1,\dots,k_n\}}$ and $\widetilde{D}_n$ from the projection approach. We will use the matrix $\widetilde{\rho}_{n,k_n} = \widetilde{D}_n\widetilde{C}_n^{-1} \in \mathbb{R}^{k_n\times k_n}$ for the iterations – without the principal components (as columns of the matrix $V$) in contrast to the theory. It is because it does not matter whether we use the transformation $Q \mapsto VQV^T$ before or after the recursions.
In the projection approach we worked with modified eigenvalues $(c'_{kn})_{k\in\{1,\dots,k_n\}}$ to ensure $\frac{d_{kn}}{c'_{kn}} < \frac{1}{2}$. Analogously here we have to make sure $\|\widetilde{\rho}_{n,k_n}\| < \frac{1}{2}$ specifically. If $\|\widetilde{\rho}_{n,k_n}\| \geq \frac{1}{2}$ holds, then we determine a "modifier" $d_{\text{mod}} := \frac{0.5-\delta}{\|\widetilde{\rho}_{n,k_n}\|}$ and redefine $\widetilde{\rho}_{n,k_n} = d_{\text{mod}}\widetilde{D}_n\widetilde{C}_n^{-1}$ with some small $\delta > 0$ such that $\|\widetilde{\rho}_{n,k_n}\| < \frac{1}{2}$ is made sure. If we did not ensure that, the iterative method would output a matrix containing `Inf` or `NA` values provided that $\|\widetilde{\rho}_{n,k_n}\| \geq \frac{1}{2}$ holds.
Now we apply the iterative procedure by iterating $M = 100$ times and gain a matrix $It$ that has to be backtransformed via $V(It)V^T = L_{\text{it}}$ (or $\frac{1}{d_{\text{mod}}}V(It)V^T = L_{\text{it}}$ if $\widetilde{\rho}_{n,k_n}$ has been modified) as well as in the projection procedure. Then $L_{\text{it}}$ can be compared with the original matrix $l$.

(iv) We compare the results of the projection and iterative approach with other estimation procedures: the `R`-function `VMA` based on "using the conditional multivariate Gaussian likelihood function" (see `R` documentation) and the innovation algorithm. The former (**V**ector **M**oving **A**verage) is computationally much heavier than the other methods, which results in long computation duration. Note that the `MTS` package (which contains `VMA`) assumes the (multivariate) moving average formula to be

$$X_n = \varepsilon_n - \sum_{j=1}^{q} l_j\varepsilon_{n-j}$$

for $q \in \mathbb{N}$, so the result of the `VMA` estimation has to be multiplied by $-1$ so that we obtain the right estimate.
The basic idea of the innovation algorithm is to compute appropriate estimates $\theta_{nj}$ with $n \in \mathbb{N}$ and $j \in \{1,\dots,n\}$ such that

$$\forall n \in \mathbb{N}:\ \widehat{X}_{n+1} := \sum_{j=1}^{n}\theta_{nj}(X_{n+1-j} - \widehat{X}_{n+1-j}),\ \widehat{X}_0 \equiv 0$$

for a multivariate time series holds. Unfortunately there is no functional innovation algorithm, but if we work with functional data, we can restrict ourselves to the finite dimensional case by using PCA. That is why we are fine with the multivariate innovation algorithm. We omit the details of it. (One can show that the innovation algorithm and the likelihood-based routine are equivalent.) For the multivariate $MA(1)$ process with matrix $l$ it turns out that the innovation algorithm reduces to

$$V_0 := C,\ \forall n \in \mathbb{N}:\ \theta_{n1} = DV_{n-1}^{-1},\ \theta_{nj} \equiv 0\ \forall j \in \{2,\dots,n\},\ V_n := C - DV_{n-1}^{-1}D^T$$

and that $\theta_{n1} \overset{n\to\infty}{\longrightarrow} l$ holds. In the functional case one can prove $\theta_{n1}^{k_n} \overset{n\to\infty}{\longrightarrow} l$, where one has to analyze this convergence for the number of observations $n$ and the number of principal components $k_n$ simultaneously.

The whole implementation will be applied throughout the next examples – from multivariate to functional moving average processes. In the following section we start with a very simple $MA(1)$ process and then analyze some requirements of the procedures.

## 5.2 Multivariate Time Series Examples

Actually we are interested in functional time series, but for simplicity we start with multivariate examples in order to get an idea of what is happening in the simulations. We here consider the three-dimensional $MA(1)$ process

$$X_{i+1} = l\varepsilon_i + \varepsilon_{i+1},\ l \in \mathbb{R}^{3\times 3},\ \varepsilon_i \sim \mathcal{N}_3(0,\Sigma) \text{ i.i.d.},\ \Sigma = \begin{pmatrix} 0.5 & 0 & 0 \\ 0 & 0.3 & 0 \\ 0 & 0 & 0.1 \end{pmatrix} \tag{5.1}$$

for $i \in \mathbb{Z}$ and analyze the estimation procedures depending on the choice of $l$. Due to the low dimension neither the `fda` package nor the optimal number of principal components are used in this section.

### 5.2.1 Multiple Simulation Study with Diagonal Matrix

We vary the number of observations $n \in \{100, 500, 1000, 5000, 10000\}$ and generate $S = 100$ multivariate time series for each $n$. Having sampled from the multivariate normal distribution $\mathcal{N}_3(0,\Sigma)$ (with $\Sigma$ as above) to obtain a strong white noise for each simulation, the $MA(1)$ process is calculated according to (5.1) with

$$l = \begin{pmatrix} 0.3 & 0 & 0 \\ 0 & 0.2 & 0 \\ 0 & 0 & 0.1 \end{pmatrix}. \tag{5.2}$$

The components of $l$ were intended to suffice (H1)-(H5). Let $\|\cdot\|_{\text{sp}}$ denote the spectral norm of a matrix (i.e. maximal eigenvalue w.r.t. its absolute value), which is the multivariate equivalent of $\|\cdot\|_{\mathcal{L}}$. (H1) holds because of $\|l\|_{\text{sp}} = 0.3 < 1$, (H3) holds because $l$ and $\Sigma$ are diagonal matrices and (H5) holds because of

$$\|\rho\|_{\text{sp}} = \| \left(l\Sigma\right) \left(\Sigma + l\Sigma l)\right)^{-1} \|_{\text{sp}} \approx 0.275 < 0.5.$$

Now the goal is to evaluate the errors of the 100 simulations for each $n$. We calculate **estimation errors** as well as **forecasts and their relative prediction errors** for each simulation. For determination of estimation errors we compute $\|\hat{l} - l\|_{\text{sp}}$, where the estimates from the projection method, from the iterative method, from `VMA` and from the innovation algorithm are plugged in for $\hat{l}$. Thus we obtain one single error value per method and per simulation. Finally we take the average over all simulations for each method and for each observation number. The results can be seen in the Table 5.1. The `VMA` routine is run only for $n \in \{100, 500\}$ due to computation duration.

| | Projection method | Iterative method | VMA | Innovation algorithm |
|---|---|---|---|---|
| $n = 100$ | 0.1518 | 0.3729 | 0.3648 | 0.3731 |
| $n = 500$ | 0.0712 | 0.1624 | 0.1480 | 0.1609 |
| $n = 1000$ | 0.0452 | 0.1064 | - | 0.1056 |
| $n = 5000$ | 0.0209 | 0.0471 | - | 0.0469 |
| $n = 10000$ | 0.0152 | 0.0330 | - | 0.0329 |

Table 5.1: Average of absolute estimation errors over 100 simulations for $l$ as in (5.2).

Forecasting turns out to be quite tedious. The moving average formula requires applying $l$ on the white noise, but we wish to use the data (in our case `X` up to the $(n-1)$-st observation) for forecasting. Suppose we wish to predict $\widehat{X}_n$. At first we approximate the $(n-1)$-st white noise $\varepsilon_{n-1}$ according to (3.7) by the approximated inverse formula

$$\widehat{\varepsilon}_{n-1} = \sum_{j=0}^{\min\{z,n-2\}} (-1)^j \hat{l}^j X_{n-1-j} \tag{5.3}$$

with some threshold $z \in \mathbb{N}$. This means that we get one specific white noise estimate for each estimation method!
As a problem, it turns out that calculating $\widehat{\varepsilon}_{n-1}$ would take very long, if we do not design the summands of the inverse formula for $\varepsilon_{n-1}$ to be limited by $z$ (e.g. $z = 20$). This is the reason for the minimum expression. Afterwards we set

$$\widehat{X}_n := \hat{l}\widehat{\varepsilon}_{n-1},$$

i.e. we perform forecasts for each estimator. Those are compared with the true vectors $X_n$ according to the formula

$$\frac{\left\| X_n - \widehat{X}_n - \varepsilon_n \right\|}{\|X_n\|}.$$

Here, we use the euclidean norm. Since we wish to assess the prediction errors with respect to the data values (otherwise it is hard to say if an absolute error is good or bad), all of those mean squared errors are divided by the euclidean norm of the respective true vector. Thus we obtain one relative mean squared error for each observation number, for each approach and for each simulation. Table 5.2 shows the averages over the simulations. Recall that the `VMA` routine is run only for $n \in \{100, 500\}$ due to computation duration.

| | Projection method | Iterative method | VMA | Innovation algorithm |
|---|---|---|---|---|
| $n = 100$ | 0.0252 | 0.1205 | 0.1801 | 0.1160 |
| $n = 500$ | 0.0112 | 0.0281 | 0.0174 | 0.0254 |
| $n = 1000$ | 0.0035 | 0.0109 | - | 0.0107 |
| $n = 5000$ | 0.0015 | 0.0025 | - | 0.0025 |
| $n = 10000$ | 0.0007 | 0.0023 | - | 0.0022 |

Table 5.2: Average of relative forecast errors over 100 simulations for $l$ as in (5.2).

As expected both the absolute estimation errors and the relative forecast errors decrease for $n$ increasing. The iterative method and the innovation algorithm seem to perform similarly (because both are fixed-point algorithms). Surprisingly the projection method is the best for this example – even better than the computationally heavy `VMA` routine.

### 5.2.2 Close-Up: One Simulation with (almost) Diagonal Matrices

We are analyzing one single simulation result for $n = 100$, but we estimate $l$ with the whole data (i.e. the last observation is included) because now we are not interested in predictions. Having run one simulation, we obtain the estimation errors

$$\|l-\hat{l}_{\text{proj.}}\|_{\text{sp}} = 0.1151,\ \|l-\hat{l}_{\text{iter.}}\|_{\text{sp}} = 0.4137,\ \|l-\hat{l}_{\texttt{VMA}}\|_{\text{sp}} = 0.4396,\ \|l-\hat{l}_{\text{innov.}}\|_{\text{sp}} = 0.4168.$$

The estimations errors do not reveal that the estimates are far from the true matrix in fact.

$$\hat{l}_{\text{proj.}} = \begin{pmatrix} 0.1852 & -0.0030 & -0.0002 \\ -0.0030 & 0.1193 & 0.0031 \\ -0.0002 & 0.0031 & 0.1672 \end{pmatrix},\ \hat{l}_{\text{iter.}} = \begin{pmatrix} 0.2321 & -0.1875 & -0.3591 \\ -0.0546 & 0.1287 & 0.0286 \\ 0.0167 & -0.0534 & 0.1778 \end{pmatrix},$$

$$\hat{l}_{\texttt{VMA}} = \begin{pmatrix} 0.2289 & -0.2418 & -0.3596 \\ -0.0568 & 0.1892 & 0.0047 \\ 0.0211 & -0.0525 & 0.1523 \end{pmatrix},\ \hat{l}_{\text{innov.}} = \begin{pmatrix} 0.1803 & -0.1833 & -0.3502 \\ -0.0432 & 0.1279 & 0.0247 \\ 0.0218 & -0.0541 & 0.1804 \end{pmatrix}.$$

The reason for those outputs is the low number of observations $n = 100$. In this case, the simulated white noise process is not sampled well enough because it has too big non-zero-lags, although they are supposed to be around zero. Consequently the simulated $MA(1)$ process possesses $k$-lags, $k \geq 2$, of similar size as the 1-lag, but in theory the 1-lag of a $MA(1)$ process is supposed to be much higher than the others. They are illustrated in Figure 5.1. The y-axis denotes the spectral norm of the lag $k$ sample cross-correlation matrix, $k \in \{0, 1, 2, 3, 4\}$.

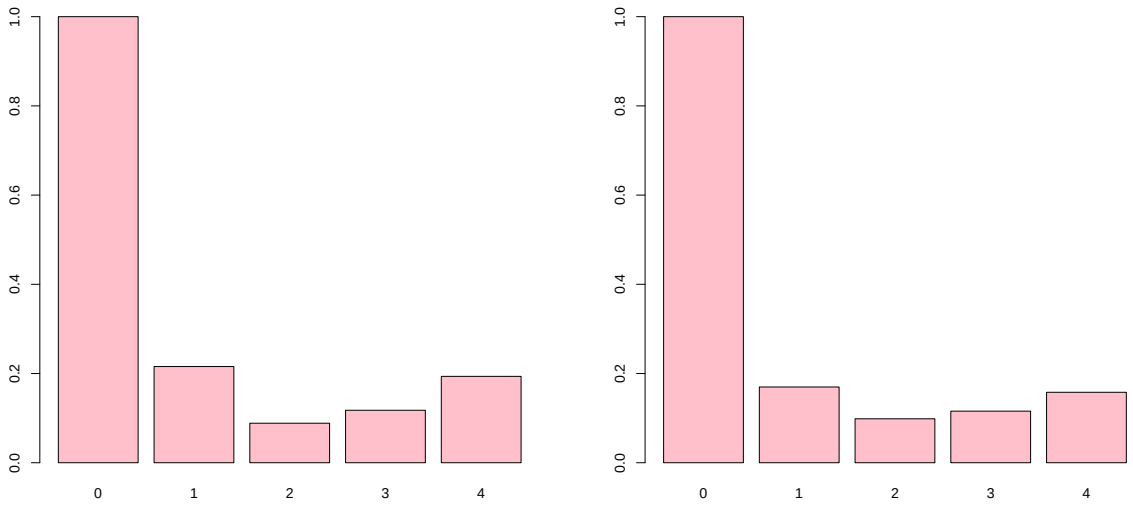


Figure 5.1: Lags for $n = 100$ for $MA(1)$ series (left) and white noise (right).

Since the estimation procedures require clear $MA(1)$ processes, which does not hold here, the estimates are not very good. Therefore we need to increase the number of observations.We set $n = 1000$. Since maximizing the Gaussian likelihood is computationally tedious in general, but even more for a large number of observations, the `VMA` routine is not applied from now on. For the other methods, the errors have decreased decisively:

$$\hat{l}_{\text{proj.}} = \begin{pmatrix} 0.2629 & -0.0028 & 0.0016 \\ -0.0028 & 0.1997 & 0.0011 \\ 0.0016 & 0.0011 & 0.1270 \end{pmatrix}, \ \|l - \hat{l}_{\text{proj.}}\|_{\text{sp}} = 0.0373,$$

$$\hat{l}_{\text{iter.}} = \begin{pmatrix} 0.2680 & -0.0057 & 0.1216 \\ 0.0087 & 0.1996 & 0.0426 \\ -0.0014 & -0.0175 & 0.1287 \end{pmatrix}, \ \|l - \hat{l}_{\text{iter.}}\|_{\text{sp}} = 0.1352,$$

$$\hat{l}_{\text{innov.}} = \begin{pmatrix} 0.2628 & -0.0066 & 0.1203 \\ 0.0072 & 0.1992 & 0.0409 \\ -0.0031 & -0.0180 & 0.1289 \end{pmatrix}, \ \|l - \hat{l}_{\text{innov.}}\|_{\text{sp}} = 0.1348.$$

After rounding the matrices by using one digit, the estimates and the original matrix $l$ coincide. This improvement makes sense, because the sample lags of the simulated $MA(1)$ process and of the simulated white noise rather correspond to the theoretical values, which we expected. See Figure 5.2. Again the y-axis denotes the spectral norm of the lag $k$ sample cross-correlation matrix, $k \in \{0, 1, 2, 3, 4\}$.

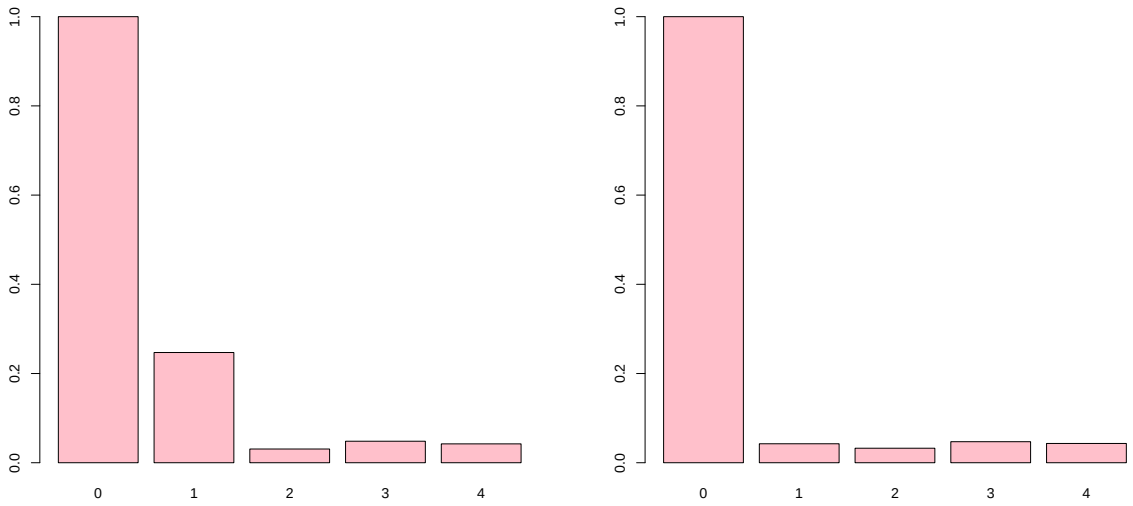


Figure 5.2: Lags for $n = 1000$ for $MA(1)$ series (left) and white noise (right).

We will procede in the same way by working with one simulation and $n = 1000$, but $l$ is modified to

$$l = \begin{pmatrix} 0.3 & 0 & 0 \\ 0.25 & 0.2 & 0 \\ 0 & 0 & 0.1 \end{pmatrix}, \tag{5.4}$$

i.e. $l$ is not a diagonal matrix any more. The resulting estimates are:

$$\begin{aligned}
\hat{l}_{\text{proj.}} &= \begin{pmatrix} 0.2646 & 0.0008 & 0.0046 \\ 0.0008 & 0.2275 & 0.0052 \\ 0.0046 & 0.0052 & 0.1060 \end{pmatrix}, \ \|l - \hat{l}_{\text{proj.}}\|_{\text{sp}} = 0.2533, \\
\hat{l}_{\text{iter.}} &= \begin{pmatrix} 0.2589 & -0.0274 & 0.0781 \\ 0.2245 & 0.2557 & 0.0425 \\ -0.0108 & 0.0098 & 0.1092 \end{pmatrix}, \ \|l - \hat{l}_{\text{iter.}}\|_{\text{sp}} = 0.1022, \\
\hat{l}_{\text{innov.}} &= \begin{pmatrix} 0.2572 & -0.0223 & 0.0722 \\ 0.2237 & 0.2670 & 0.0322 \\ -0.0115 & 0.0093 & 0.1094 \end{pmatrix}, \ \|l - \hat{l}_{\text{innov.}}\|_{\text{sp}} = 0.0959.
\end{aligned}$$

Those outputs completely make sense: The iterative method and the innovation algorithm are proven to converge even for non-symmetric operators. By contrast, the projection method only returns symmetric matrices and requires (H3) to hold, which is violated here. This is why the projection method is very good in the simple example before, but is not recommended to be applied in practice where (H3) cannot be made sure.

### 5.2.3 One Simulation with Orthogonal Matrices

The same proceeding as before is demonstrated with $n = 1000$ for

$$l = \begin{pmatrix} -0.0479 & -0.3660 & -0.1540 \\ -0.3660 & 0.1008 & -0.1259 \\ -0.1540 & -0.1259 & 0.3470 \end{pmatrix},$$

which was generated by 0.4·`rortho(3)`. The `R` function `rortho(k)` from the package `pracma` creates a symmetric orthogonal matrix of size `k`. We chose this because firstly its spectral norm equals 1, which means that we can control the spectral norm by scaling, and secondly it contains only non-zero elements so that it is a more concrete non-diagonal example. It is scaled by 0.4 because then due to

$$\|\rho\|_{\text{sp}} \approx 0.3804 < 0.5$$

(H5) holds. As before since (H3) is not satisfied, the estimate of the projection method is inappropriate in contrast to the other ones, where the innovation algorithm performs better then the iterative method:

$$\hat{l}_{\text{proj.}} = \begin{pmatrix} 0.0371 & -0.0021 & -0.0102 \\ -0.0021 & 0.0208 & -0.0085 \\ -0.0102 & -0.0085 & 0.3593 \end{pmatrix}, \ \|l - \hat{l}_{\text{proj.}}\|_{\text{sp}} = 0.4537,$$

$$\hat{l}_{\text{iter.}} = \begin{pmatrix} -0.0371 & -0.3303 & -0.0100 \\ -0.3056 & 0.1255 & -0.0654 \\ -0.1490 & -0.1137 & 0.3553 \end{pmatrix}, \ \|l - \hat{l}_{\text{iter.}}\|_{\text{sp}} = 0.1662,$$

$$\hat{l}_{\text{innov.}} = \begin{pmatrix} -0.0401 & -0.3436 & -0.0564 \\ -0.2937 & 0.1262 & -0.0789 \\ -0.1519 & -0.1079 & 0.3651 \end{pmatrix}, \ \|l - \hat{l}_{\text{innov.}}\|_{\text{sp}} = 0.1239.$$

However, if we generate another orthogonal matrix and scale it by 0.8 so that we obtain

$$l = \begin{pmatrix} -0.7101 & -0.2755 & -0.2447 \\ -0.2755 & 0.7497 & -0.0447 \\ -0.2447 & -0.0447 & 0.7603 \end{pmatrix},$$

the innovation algorithm still performs well, but both the projection and the iterative method output bad estimates because (H5) is not satisfied according to $\|\rho\|_{\text{sp}} \approx 0.5470 > 0.5$.

$$\hat{l}_{\text{proj.}} = \begin{pmatrix} -0.7942 & -0.1002 & -0.1918 \\ -0.1002 & 0.7232 & -0.0115 \\ -0.1918 & -0.0115 & 0.6713 \end{pmatrix}, \ \|l - \hat{l}_{\text{proj.}}\|_{\text{sp}} = 0.2356,$$

$$\hat{l}_{\text{iter.}} = \begin{pmatrix} -0.5830 & -0.2660 & 0.0197 \\ -0.1907 & 0.5594 & -0.0406 \\ -0.1827 & -0.0368 & 0.5542 \end{pmatrix}, \ \|l - \hat{l}_{\text{iter.}}\|_{\text{sp}} = 0.3428,$$

$$\hat{l}_{\text{innov.}} = \begin{pmatrix} -0.9176 & -0.3812 & -0.3246 \\ -0.2627 & 0.7549 & -0.0889 \\ 0.2713 & -0.0384 & 0.6379 \end{pmatrix}, \ \|l - \hat{l}_{\text{innov.}}\|_{\text{sp}} = 0.255.$$

Again the projection method performs better than the iterative method as in the beginning examples.

### 5.2.4 Evaluation from Several Simulations

To make sure that there is no equivalence between the iterative method and the innovation algorithm, we simulate 100 $MA(1)$ time series – again with $l$ as in (5.4), as a random orthogonal matrix multiplied by 0.4 and as a random orthogonal matrix multiplied by 0.8. The problem that occurs in all three examples is that some estimates $\widehat{l}$ are of spectral norm greater than (or equal) 1, so that the estimated $MAH(1)$ process is not invertible and so that (5.3) returns very large estimates $\widehat{\varepsilon}_{n-1}$. (We did not have this issue in the case where $l$ is diagonal.) Thus we display three tables for each example. The first table indicates the number of simulations where the outcome is sufficiently small for inversion, i.e. $\|\widehat{l}\|_{\text{sp.}} < 1$. The second table indicates the average estimation errors of the simulations where $\|\widehat{l}\|_{\text{sp.}} < 1$ holds and the third table indicates the average relative forecast errors of the simulations where $\|\widehat{l}\|_{\text{sp.}} < 1$ holds.

| | Projection method | Iterative method | VMA | Innovation algorithm |
|---|---|---|---|---|
| $n = 100$ | 100 | 100 | 100 | 99 |
| $n = 500$ | 100 | 100 | 100 | 100 |
| $n = 1000$ | 100 | 100 | - | 100 |
| $n = 5000$ | 100 | 100 | - | 100 |
| $n = 10000$ | 100 | 100 | - | 100 |

Table 5.3: Number of simulations where $\|\widehat{l}\|_{\text{sp.}} < 1$ for $l$ as in (5.4).

For $l$ as in (5.4) there is only one replicate such that $\|\widehat{l}\|_{\text{sp.}} \geq 1$, see Table 5.3. We ignore this simulation in Table 5.4 and Table 5.5; otherwise the forecast error for $n = 100$ for the innovation algorithm would be much bigger.

| | Projection method | Iterative method | VMA | Innovation algorithm |
|---|---|---|---|---|
| $n = 100$ | 0.2809 | 0.4032 | 0.3597 | 0.4119 |
| $n = 500$ | 0.2377 | 0.1542 | 0.1361 | 0.1545 |
| $n = 1000$ | 0.2315 | 0.1108 | - | 0.1118 |
| $n = 5000$ | 0.2297 | 0.0517 | - | 0.0484 |
| $n = 10000$ | 0.2276 | 0.0392 | - | 0.0353 |

Table 5.4: Average of absolute estimation errors over 100 simulations for $l$ as in (5.4).

| | Projection method | Iterative method | VMA | Innovation algorithm |
|---|---|---|---|---|
| $n = 100$ | 0.1852 | 0.2015 | 0.2034 | 0.2663 |
| $n = 500$ | 0.0647 | 0.0121 | 0.0097 | 0.0124 |
| $n = 1000$ | 0.0690 | 0.0146 | - | 0.0138 |
| $n = 5000$ | 0.0486 | 0.0028 | - | 0.0021 |
| $n = 10000$ | 0.0815 | 0.0020 | - | 0.0013 |

Table 5.5: Average of relative forecast errors over 100 simulations for $l$ as in (5.4).

For some orthogonal matrix multiplied by 0.4

$$l = \begin{pmatrix} -0.0215 & -0.3674 & -0.1568 \\ -0.3674 & 0.0798 & -0.1367 \\ -0.1568 & -0.1367 & 0.3417 \end{pmatrix} \tag{5.5}$$

again all but some outcomes from the innovation algorithm for $n = 100$ are of $\|\widehat{l}\|_{\text{sp.}} < 1$.

| | Projection method | Iterative method | VMA | Innovation algorithm |
|---|---|---|---|---|
| $n = 100$ | 100 | 100 | 100 | 83 |
| $n = 500$ | 100 | 100 | 100 | 100 |
| $n = 1000$ | 100 | 100 | - | 100 |
| $n = 5000$ | 100 | 100 | - | 100 |
| $n = 10000$ | 100 | 100 | - | 100 |

Table 5.6: Number of simulations where $\|\widehat{l}\|_{\text{sp.}} < 1$ for $l$ as in (5.5).

However, for $n = 100$ for the innovation algorithm we have more outcomes such that $\|\widehat{l}\|_{\text{sp.}} \geq 1$ holds than before.

| | Projection method | Iterative method | VMA | Innovation algorithm |
|---|---|---|---|---|
| $n = 100$ | 0.4260 | 0.3987 | 0.3342 | 0.3972 |
| $n = 500$ | 0.4271 | 0.1820 | 0.1299 | 0.1784 |
| $n = 1000$ | 0.4283 | 0.1344 | - | 0.1145 |
| $n = 5000$ | 0.4329 | 0.0909 | - | 0.0552 |
| $n = 10000$ | 0.4336 | 0.0801 | - | 0.0389 |

Table 5.7: Average of absolute estimation errors over 100 simulations for $l$ as in (5.5).

Since the averave errors of the iterative method and the ones of the innovation algorithm are different for large $n$, we conclude that those two estimation approaches are not equivalent.

| | Projection method | Iterative method | VMA | Innovation algorithm |
|---|---|---|---|---|
| $n = 100$ | 0.2612 | 0.1427 | 0.0822 | 0.1147 |
| $n = 500$ | 0.2349 | 0.0233 | 0.0137 | 0.0210 |
| $n = 1000$ | 0.2613 | 0.0118 | - | 0.0092 |
| $n = 5000$ | 0.2381 | 0.0029 | - | 0.0016 |
| $n = 10000$ | 0.3059 | 0.0036 | - | 0.0017 |

Table 5.8: Average of relative forecast errors over 100 simulations for $l$ as in (5.5).

For some orthogonal matrix multiplied by 0.8

$$l = \begin{pmatrix} -0.5413 & -0.5717 & -0.1420 \\ -0.5717 & 0.5564 & -0.0605 \\ -0.1420 & -0.0605 & 0.7850 \end{pmatrix} \tag{5.6}$$

we obtain much more outcomes $\widehat{l}$ with $\|\widehat{l}\|_{\text{sp.}} \geq 1$, which have to be excluded.

| | Projection method | Iterative method | VMA | Innovation algorithm |
|---|---|---|---|---|
| $n = 100$ | 100 | 90 | 59 | 35 |
| $n = 500$ | 100 | 100 | 100 | 52 |
| $n = 1000$ | 100 | 100 | - | 54 |
| $n = 5000$ | 100 | 100 | - | 70 |
| $n = 10000$ | 100 | 100 | - | 91 |

Table 5.9: Number of simulations where $\|\widehat{l}\|_{\text{sp.}} < 1$ for $l$ as in (5.6).

Remember the modifier $d_{\text{mod}}$ of the iterative method as already described in Section 5.1. $d_{\text{mod}}$ makes sure that the iterative method is able to run, even if in fact $\|\widehat{\rho}_{n,p_n}\| \geq \frac{1}{2}$ holds. Therefore the iterative method is less likely to produce outcomes $\widehat{l}$ with $\|\widehat{l}\|_{\text{sp.}} \geq 1$ than the innovation algorithm, see Table 5.9.

| | Projection method | Iterative method | VMA | Innovation algorithm |
|---|---|---|---|---|
| $n = 100$ | 0.4904 | 0.4808 | 0.2236 | 0.6128 |
| $n = 500$ | 0.3072 | 0.3421 | 0.0874 | 0.3406 |
| $n = 1000$ | 0.2919 | 0.2997 | - | 0.2834 |
| $n = 5000$ | 0.2379 | 0.2832 | - | 0.1382 |
| $n = 10000$ | 0.2211 | 0.2761 | - | 0.1016 |

Table 5.10: Average of absolute estimation errors over 100 simulations for $l$ as in (5.6).

However, when $d_{\text{mod}}$ is needed, the estimate $\widehat{l}_{\text{iter.}}$ does not converge to $l$, because it solves the fixed-point equation (4.12) approximately for $\widehat{\rho}_{n,p_n}$ modified. Therefore the iterative method produces the largest estimation errors for large $n$.

| | Projection method | Iterative method | VMA | Innovation algorithm |
|---|---|---|---|---|
| $n = 100$ | 0.7552 | 0.1964 | 0.0810 | 0.5340 |
| $n = 500$ | 0.2354 | 0.1097 | 0.0102 | 0.1288 |
| $n = 1000$ | 0.3631 | 0.1161 | - | 0.1317 |
| $n = 5000$ | 0.1198 | 0.0573 | - | 0.0524 |
| $n = 10000$ | 0.1290 | 0.1257 | - | 0.0476 |

Table 5.11: Average of relative forecast errors over 100 simulations for $l$ as in (5.6).

## 5.3 Functional Time Series Examples

Finally, we look at some functional data examples. The optimal number of principal components plays a very big role in contrast to what we saw before. Again we work with only one simulation. We consider the $MAH(1)$ process of the form

$$X_{i+1} = l(\varepsilon_i) + \varepsilon_{i+1} \qquad \text{where} \qquad l(x)(t) = \int \varphi(s,t)x(s)\mathrm{d}s.$$

The reason for taking an integral operator is that in this case it suffices to consider only the integral kernel $\varphi(\cdot,\cdot)$. We assume all functional data elements to map from $[0,1]$, discretize the input argument $t = (0, \frac{1}{m-1}, \frac{2}{m-1}, \ldots, 1) \in \mathbb{R}^m$ and take $m = 100$. Consequently $\varphi$ is implemented as a $m \times m$ matrix $\widetilde{\varphi}$. Here visualizing the matrices in a 3D plot is more useful than displaying the matrix numbers, what we did before.
Furthermore the relation $\|l\|_{\mathcal{L}} \approx \dfrac{\|\widetilde{\varphi}\|_{\mathrm{sp}}}{m}$ holds. The reason is the integral approximation

$$l(x)(t_i) = \int \varphi(t_i,s)x(s)\mathrm{d}s \approx \frac{1}{m}\sum_{j=1}^{m} \widetilde{\varphi}_{i,j}x(t_j)$$

for $i \in \{1,\ldots,m\}$. Therefore the eigenvalues of $\Phi$ (here e.g. $c_{\text{func.}}$) equal the ones of $\widetilde{\varphi}$ (here e.g. $c_{\text{MV}}$) divided by $\frac{1}{m}$, which leads to $\frac{\|\widetilde{\varphi}\|_{\mathrm{sp}}}{m}$

$$c_{\text{func.}}v = l(v) \approx \frac{1}{m}\widetilde{\varphi}v = \frac{1}{m}c_{\text{MV}}v.$$

In addition, let $\Theta$ and $\Omega$ be to two integral operators with kernels $\theta$ and $\omega$ as well as with matrix equivalents $\widetilde{\theta}$ and $\widetilde{\omega}$. Then due to

$$\begin{aligned}(\Theta \circ \Omega)(x)(t_i) &= \int \theta(t_i,u)\int \omega(u,s)x(s)\mathrm{d}s\mathrm{d}u \\ &\approx \frac{1}{m}\sum_{j=1}^{m}\widetilde{\theta}_{i,j}\frac{1}{m}\sum_{k=1}^{m}\widetilde{\omega}_{j,k}x(t_k) = \frac{1}{m}\sum_{k=1}^{m}\Big(\frac{1}{m}\sum_{j=1}^{m}\widetilde{\theta}_{i,j}\widetilde{\omega}_{j,k}\Big)x(t_k)\end{aligned}$$

the computational counterpart of $\Theta \circ \Omega$ is $\frac{1}{m}\widetilde{\theta}\widetilde{\omega}$.
The iterative method and the innovation algorithm will be assessed by applying them on two different integral kernels. Since in the previous section it turned out that a high number of observations is necessary even for the low-dimensional case $m = 3$, choosing $n = 1000$ here is reasonable.

### 5.3.1 Exponential Integral Kernel

Let the integral kernel be

$$\varphi(s,t) = \alpha \exp\left(-\frac{s^2+t^2}{2}\right), \qquad \forall s,t \in [0,1]$$

and choose $\alpha = 0.5$. We already know that (H3) is certainly not satisfied, therefore performing the projection method is not recommended. Neither is the `VMA` routine (although we could apply it to the PC scores) because of $n = 1000$. $\varphi$ looks as in Figure 5.3.

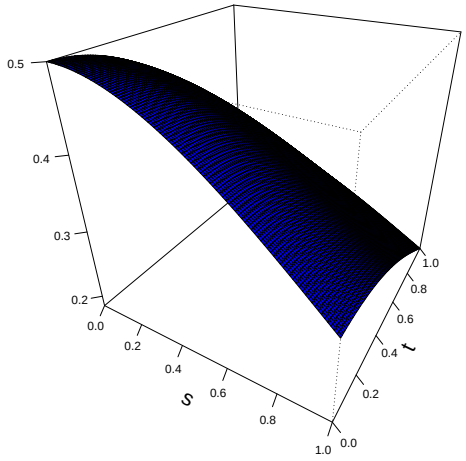


Figure 5.3: 3D-plot of $\varphi(s,t) = \alpha \exp\left(-\frac{s^2+t^2}{2}\right)$ restricted to $[0,1]$.

Before simulating the time series we have to select a certain white noise process. We choose each white noise observation to be a Brownian bridge on $[0,1]$

$$\varepsilon_t := B_t - tB_1, \ \forall t \in [0,1],$$

where $(B_t)_{t\geq 0}$ denotes the standard Brownian motion. Note that in `R` we simulate a Brownian bridge of length $m+1$ and remove the last component which equals zero. Otherwise we would obtain a matrix that represents the white noise process and whose last column only consists of zeros. This would result in a bad sample covariance matrix of the $MAH(1)$ simulation. Having simulated the white noise process we calculate the $MAH(1)$ process and transform it into a functional time series via the `fda` package. It turns out that the optimal number of principal components is 3:

| $p$ | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| CPV($p$) | 0.6048 | 0.7650 | 0.8422 | 0.8778 | 0.9022 |

Table 5.12: $CPV$ for Brownian bridge based $MAH(1)$ process with exponential integral kernel.

Having set $k_n = 3$ we are able to compute $\widetilde{\rho}_{n,k_n}$ for the iterative method as well as $\widetilde{C}_n$ and $\widetilde{D}_n$ for the innovation algorithm. Note that in fact (H5) does not hold because of $\|\rho\|_{\mathcal{L}} \approx 0.5301$. Nevertheless we can apply the iterative method due to dimension reduction to $k_n = 3$. The recursions work well because of $\|\widetilde{\rho}_{n,k_n}\|_{\text{sp}} = 0.2863 < 0.5$. Here, the innovation algorithm gives a slightly better result than the iterative method

$$\tau_\varphi(\widehat{\varphi}) := \frac{\|\varphi - \widehat{\varphi}\|_{\text{sp}}}{\|\varphi\|_{\text{sp}}},\ \tau_\varphi(\widehat{\varphi}_{\text{iter.}}) = 0.4240,\ \tau_\varphi(\widehat{\varphi}_{\text{innov.}}) = 0.4063.$$

Those errors look quite big, but in fact it is hard to compare large matrices (which can be imagined as surfaces) with each other. Both estimates can be seen in Figure 5.4.

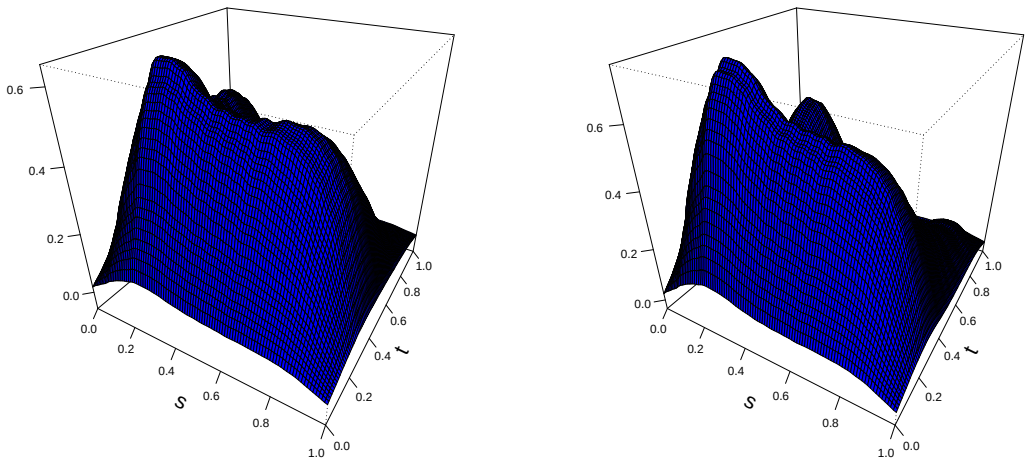


Figure 5.4: Iterative method estimate (left) and innovation algorithm estimate (right) with Brownian bridge.

After running 100 simulations and taking the mean of $\tau_\varphi(\widehat{\varphi}_{\text{iter.}})$ and of $\tau_\varphi(\widehat{\varphi}_{\text{innov.}})$ as well as their forecast errors (see first multivariate example) over the number of simulations Table 5.13 shows that the innovation algorithm tends to perform better.

| | Iterative method | Innovation algorithm |
|---|---|---|
| Mean of $\tau_\varphi(\widehat{\varphi})$ | 0.4608 | 0.4257 |
| Mean of relative forecast errors | 0.1678 | 0.1686 |

Table 5.13: Average errors for the Brownian bridge based $MAH(1)$ process with exponential integral kernel over 100 replications.

As already discussed in the multivariate examples the $k$-lags of the $MAH(1)$ simulation tend to zero for $k \geq 2$, whereas the 0-lag is the biggest and the 1-lag is clearly bigger than zero. The height of the 1-lag column depends on the spectral norm of $\varphi$. If we had chosen $\alpha$ to be small, the 1-lag would attain a small number too. If we had chosen $\alpha$ to be bigger than 1, the 1-lag would still be bigger than zero below the 0-lag, but all the values would be scaled in a different way. This is why $\alpha$ was chosen to be out of $[0.2, 0.8]$, here 0.5. For the white noise process, only the 0-lag does not equal zero. This is illustrated in Figure 5.5. Here the y-axis denotes the spectral norm of the lag $k$ sample cross-covariance matrix divided by the spectral norm of the sample covariance matrix.

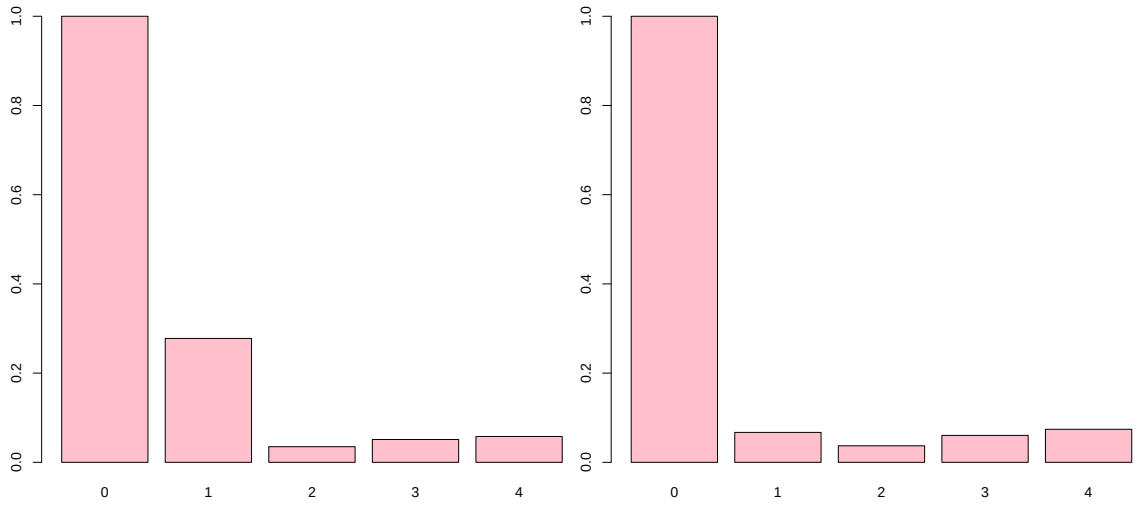


Figure 5.5: Lags for $MAH(1)$ series (left) and white noise (right).

This pattern is what we already know. However, we projected the data onto the finite dimensional linear space spanned by the most influential empirical principal components $\widehat{v}_1$, $\widehat{v}_2$ and $\widehat{v}_3$. How do the processes $\left(\langle X_i, \widehat{v}_j\rangle\right)_{i\in\mathbb{Z}}$ look like for $j \in \{1, 2, 3\}$? We create some barplots for them too. Note that they are univariate time series. Hence we obtain a single number for each series after calculating their correlation coefficients, which means that we do not need to take any matrix norm. Figure 5.6 shows that it is surprisingly only

$\left(\langle X_i, \widehat{v}_1\rangle\right)_{i\in\mathbb{Z}}$ whose 1-lag does not tend to zero. All the others have a similar dependence structure to the white noise. This means that all the information about the $MAH(1)$ structure is carried only on the first principal component. In fact the lags of $\left(\langle X_i, \widehat{v}_1\rangle\right)_{i\in\mathbb{Z}}$ seem to have the same proportions as the lags of $(X_i)_{i\in\mathbb{Z}}$.

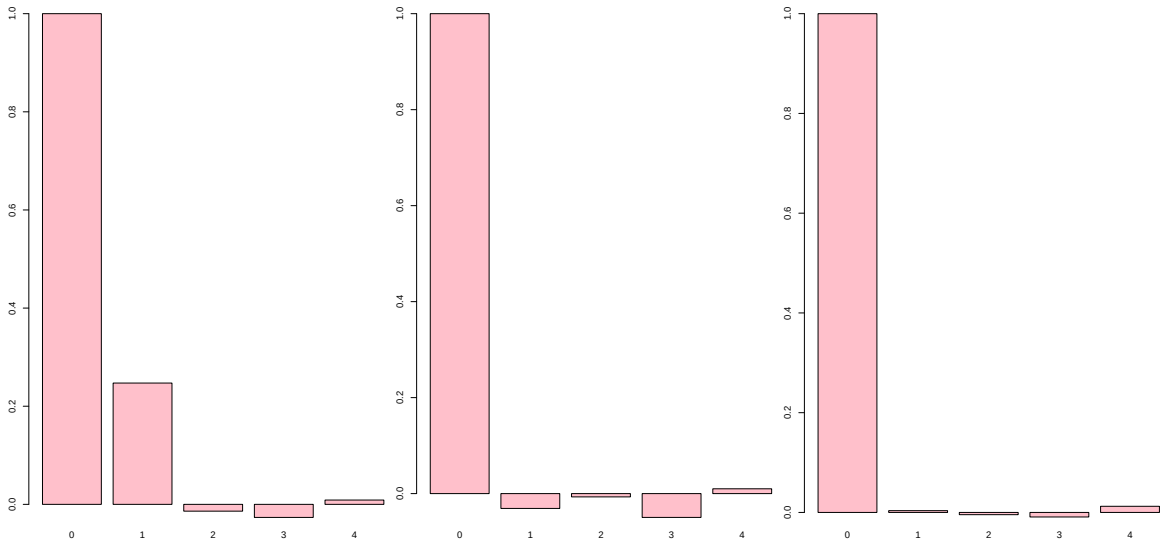

Figure 5.6: Lags for $\left(\langle X_i, \widehat{v}_1\rangle\right)_{i\in\mathbb{Z}}$ (left) $\left(\langle X_i, \widehat{v}_2\rangle\right)_{i\in\mathbb{Z}}$ (middle) and $\left(\langle X_i, \widehat{v}_3\rangle\right)_{i\in\mathbb{Z}}$ (right).

Now the role of the white noise is explained. Suppose for every $i \in \mathbb{Z}$ the random function $\varepsilon_i$ is a Brownian motion instead of a Brownian bridge. In theory the convergence of the estimates does not depend on the choice of the white noise. By contrast when it comes to implementation, the data and hence the estimates depend on the choice of the white noise decisively. As before one simulation and afterwards the two estimation procedures are run for $n = 1000$ again. The optimal number of principal components is only 1:

| $p$ | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| CPV($p$) | 0.8138 | 0.9066 | 0.9377 | 0.9545 | 0.9646 |

Table 5.14: $CPV$ for Brownian motion based $MAH(1)$ process with exponential integral kernel.

The reason for this is that the discretized covariance operator of this specific white noise has one eigenvalue which is much bigger than the others. Consequently both approaches are performed on scalars. Figure 5.7 shows that they do not perform well.

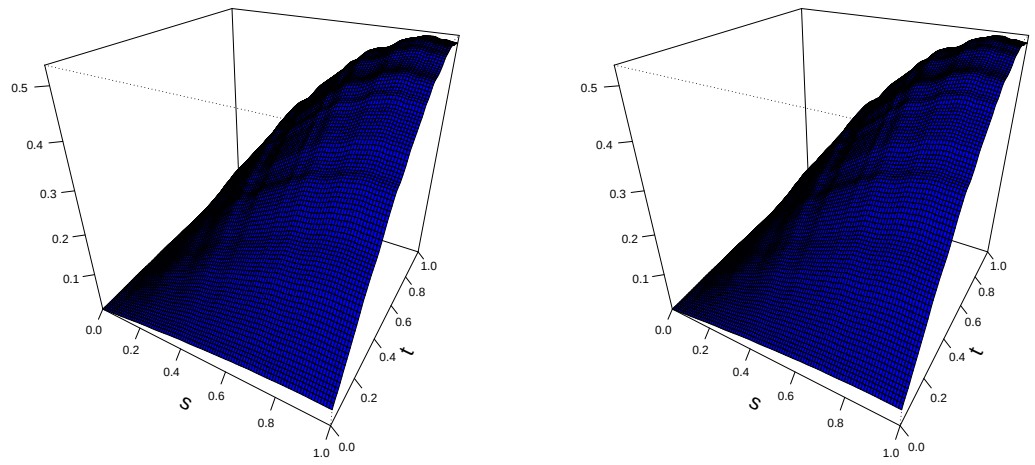


Figure 5.7: Iterative method estimate (left) and innovation algorithm estimate (right) with Brownian motion.

Both surfaces ascend to the wrong direction and this is because of $\widehat{v}_1$, which alone determines the shape of the estimates' surfaces. $\widehat{v}_1$ is supposed to contain most of the information of the data. Since $\widehat{v}_1$ depends on $C$ tremendously, whose structure itself depends on the one of $C_\varepsilon$, we analyze the integral kernels of $C_\varepsilon$ for the white noise based on Brownian bridge and for the one based on Brownian motion. Figure 5.8 reveals that for the Brownian bridge the integral kernel of $C_\varepsilon$ possesses the shape of a hill. It is very flexible in the sense that it can adjust to another surface more easily than the shape of the second integral kernel. The latter is just a straight plane ascending from the origin and hence cannot be deformed easily.

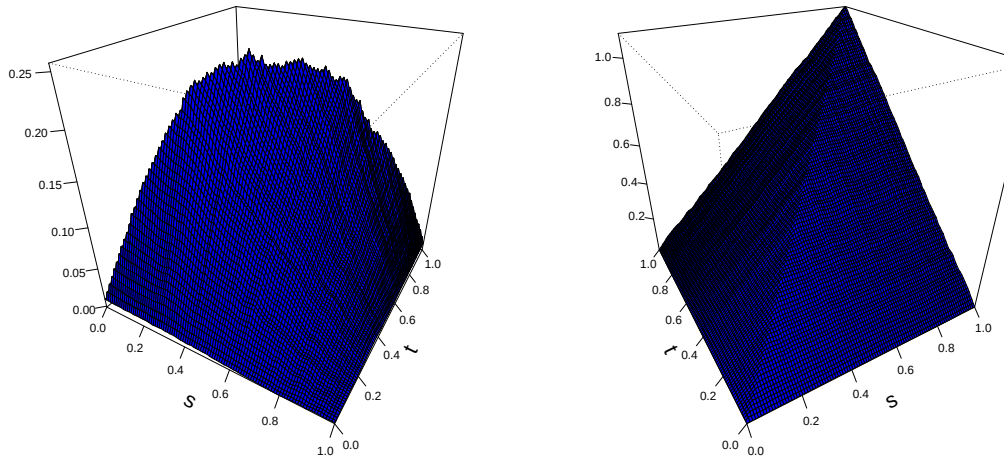


Figure 5.8: Estimated integral kernel of $C_\varepsilon$ based on Brownian bridge (left) and Brownian motion (right).

Using the Brownian motion is appropriate only if $\varphi$ looks like that (which Table 5.15 containing the average errors over 100 Brownian motion based simulations shows as well) and this does not make sense in practice. Given some real data there is no white noise generation and no hint for the structure of the moving average operator.

| | Iterative method | Innovation algorithm |
|---|---|---|
| Mean of $\tau_\varphi(\widehat{\varphi})$ | 0.5966 | 0.5966 |
| Mean of relative forecast errors | 0.1257 | 0.1257 |

Table 5.15: Average errors for the Brownian motion based $MAH(1)$ process with exponential integral kernel over 100 replications.

The same what is done in this subsection will be analyzed for the following integral kernel: Visualizing the true kernel as well as its estimates, investigating the lags and examining the choice of the white noise.

### 5.3.2 Bilinear Integral Kernel

We select the iterative method and the innovation algorithm again to approximate the integral kernel

$$\varphi(s,t) = \alpha st, \ \forall s,t \in [0,1]$$

with $\alpha = 0.5$, illustrated in Figure 5.9. The choice of $\alpha$ is because of the same reason as before: a tiny $\alpha$ would cause a small 1-lag of the $MAH(1)$ series and hence make the estimates inappropriate, whereas a large $\alpha$ would make the $MAH(1)$ series unsuitable.

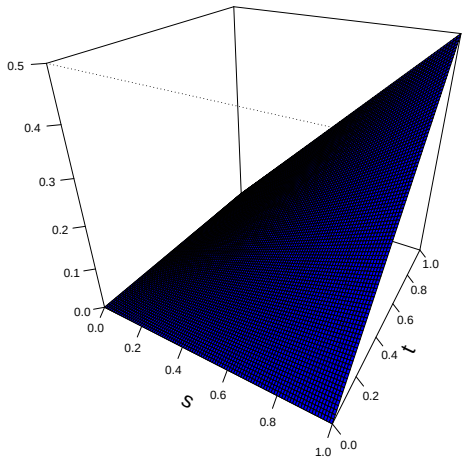


Figure 5.9: 3D-plot of $\varphi(s,t) = \alpha st$ restricted to $[0,1]$.

As before let $n = 1000$ and $m = 100$. Again we take the Brownian bridge in order to simulate the white noise. Having generated the $MAH(1)$ process according to $\varphi$ and having applied FPCA, the optimal number of principal components $k_n$ is 3 again:

| $p$ | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| CPV($p$) | 0.6134 | 0.7643 | 0.8371 | 0.8762 | 0.9006 |

Table 5.16: $CPV$ for Brownian bridge based $MAH(1)$ process with bilinear integral kernel.

Here (H5) is satisfied thanks to $\|\rho\|_{\mathcal{L}} \approx 0.1324 < 0.5$. Compared to the exponential kernel, we obtain worse relative estimation errors, but this is again the problem of comparing surfaces.

$$\tau_\varphi(\widehat{\varphi}_{\text{iter.}}) = 0.4999, \ \tau_\varphi(\widehat{\varphi}_{\text{innov.}}) = 0.4993.$$

Here the innovation algorithm performs slightly better. Having run 100 simulations and calculated the mean of $\tau_\varphi(\widehat{\varphi}_{\text{iter.}})$ and of $\tau_\varphi(\widehat{\varphi}_{\text{innov.}})$ and the mean of the relative forecast errors over the number of simulations, we get very good forecast results, as Table 5.17 shows:

| | Iterative method | Innovation algorithm |
|---|---|---|
| Mean of $\tau_\varphi(\widehat{\varphi})$ | 0.8605 | 0.8678 |
| Mean of relative forecast errors | 0.0239 | 0.0238 |

Table 5.17: Average errors for the Brownian bridge based $MAH(1)$ process with bilinear integral kernel over 100 replications.

Back to one simulation: In spite of the relative estimation error values, the estimates $\widehat{\varphi}_{\text{iter.}}$ and $\widehat{\varphi}_{\text{innov.}}$ look close to $\varphi$ according to Figure 5.10. Here we can see that the hills caused by the Brownian bridge adjust to the shape of $\varphi$.

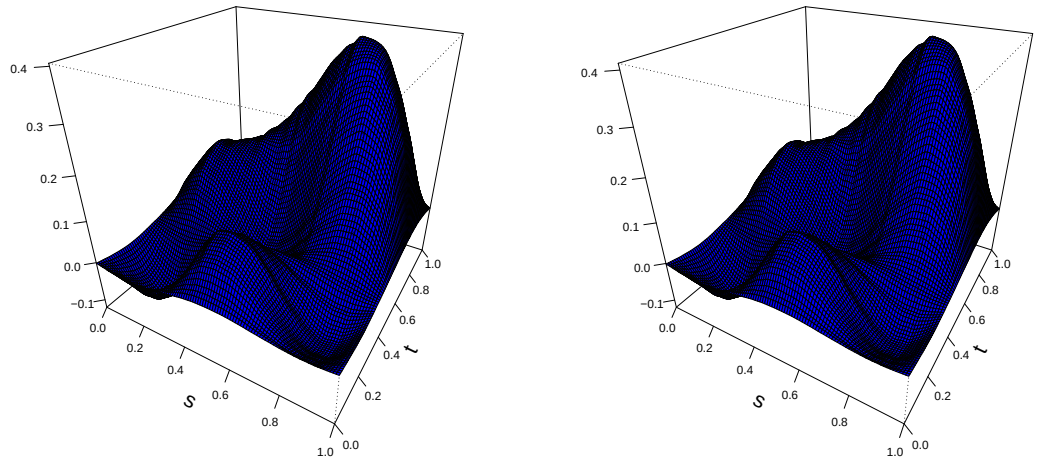


Figure 5.10: Iterative method estimate (left) and innovation algorithm estimate (right) with Brownian bridge.

When we have a look at the lags of this $MAH(1)$ process and the ones of its respective white noise in Figure 5.11, we realize that the 1-lag of the $MAH(1)$ process is lower than the one in the previous example. This is because of the choice of $\varphi$. Nevertheless the barplots of the lags of the $MAH(1)$ process look typical for a moving average process of order 1. As already expected, the non-zero-lags of the white noise are approximately 0. Again the heights of the bars correspond to the spectral norm of the respective cross-covariance matrix divided by the spectral norm of the covariance matrix.

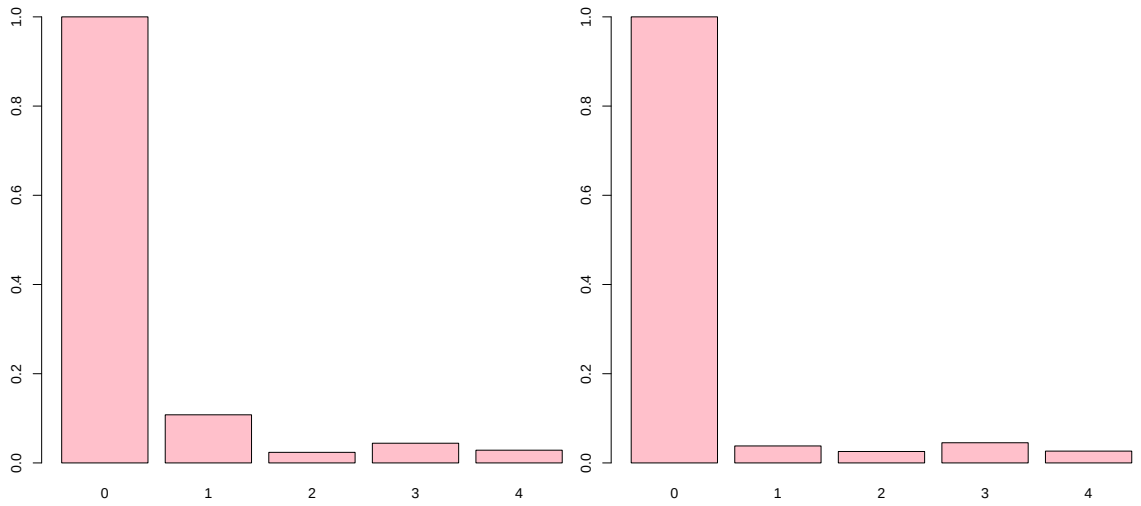


Figure 5.11: Lags for $MAH(1)$ series (left) and white noise (right).

When we consider the projections of the time series onto the principal components, again it is $\left(\langle X_i, \widehat{v}_1\rangle\right)_{i\in\mathbb{Z}}$ which rather possesses the $MA(1)$ structure. Figure 5.12 shows that the $MA(1)$ structure of the univariate time series $\left(\langle X_i, \widehat{v}_j\rangle\right)_{i\in\mathbb{Z}}$ tends to vanish for $j$ increasing.

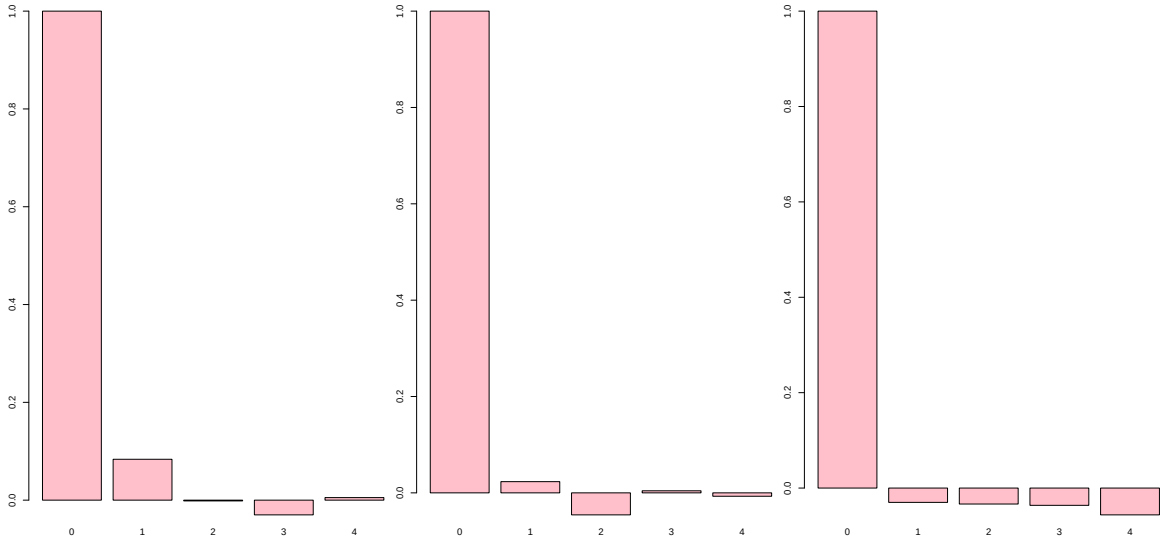


Figure 5.12: Lags for $\left(\langle X_i, \widehat{v}_1\rangle\right)_{i\in\mathbb{Z}}$ (left) $\left(\langle X_i, \widehat{v}_2\rangle\right)_{i\in\mathbb{Z}}$ (middle) and $\left(\langle X_i, \widehat{v}_3\rangle\right)_{i\in\mathbb{Z}}$ (right).

At the end we change the choice of the white noise. Comparing Figure 5.8 and Figure 5.9 we might reckon that the white noise based on Brownian motion is appropriate for this bilinear kernel. As already mentioned, this cannot be investigated in practice because there is no white noise to simulate and $\varphi$ is not known at all. Anyway we conclude from the 3D-plots that sampling the Brownian motion might improve our estimation results. Indeed it does! As in the previous section the optimal number of principal components equals 1 (because of the structure of $C_\varepsilon$):

| $p$ | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| CPV($p$) | 0.8235 | 0.9108 | 0.9410 | 0.9554 | 0.9649 |

Table 5.18: $CPV$ for Brownian motion based $MAH(1)$ process with bilinear integral kernel.

This means that the scalar iterative method and the scalar innovation algorithm are performed. Figure 5.13 shows that they give a very good approximation. In fact they lead to the same estimates! Due to lack of variability with only one principal component we get

$$\|\widehat{\varphi}_{\text{iter.}} - \widehat{\varphi}_{\text{innov.}}\|_{\text{sp}} = 0.$$

Compared to the previous relative estimation errors,

$$\tau_\varphi(\widehat{\varphi}_{\text{iter.}}) = 0.1462 = \tau_\varphi(\widehat{\varphi}_{\text{innov.}})$$

is quite low. Nevertheless the Brownian motion is not recommended in general because of the surface of the covariance function $c_\varepsilon$.

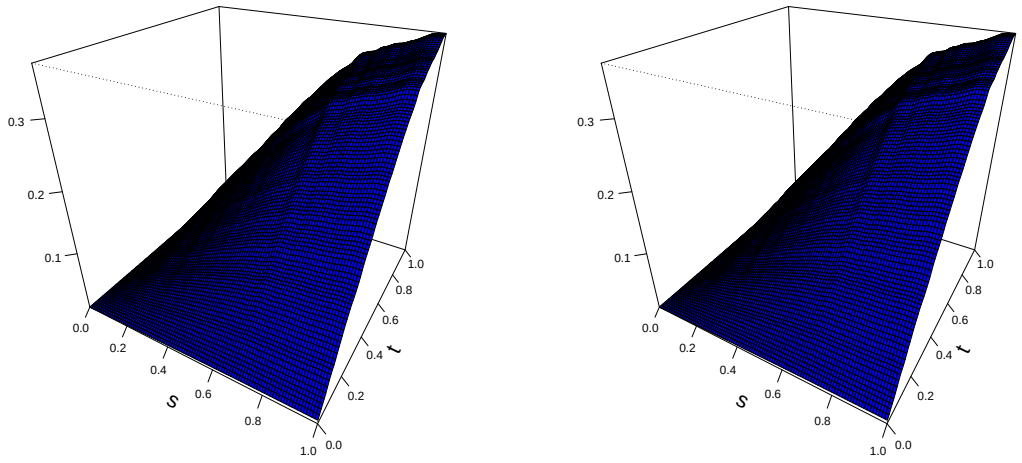


Figure 5.13: Iterative method estimate (left) and innovation algorithm estimate (right) with Brownian motion.

All those one-simulation-results might be suggestive of some equivalence between the iterative method and the innovation algorithm. However, we saw in Subsection 5.2.4 that this is not true.

### 5.3.3 Evaluation from Several Simulations

Tables 5.19-5.22 show the average errors of the four previous functional examples over 100 simulations. The errors of the iterative method and the ones of the innovation algorithm coincide for Brownian motion based $MAH(1)$ processes, because the optimal number of PCs is 1 and thus there is less variability for both approaches. For both approaches the results from the Brownian bridge based $MAH(1)$ processes are still similar, but the innovation algorithm performs better.

| | $\tau_\varphi(\widehat{\varphi}_{\text{iter.}})$ | $\tau_\varphi(\widehat{\varphi}_{\text{innov.}})$ | F.cast. iter. | F.cast. innov. |
|---|---|---|---|---|
| $n = 100$ | 0.8908 | 0.9299 | 0.1614 | 0.1568 |
| $n = 500$ | 0.5149 | 0.5030 | 0.1277 | 0.1231 |
| $n = 1000$ | 0.4799 | 0.4361 | 0.1588 | 0.1560 |
| $n = 5000$ | 0.3800 | 0.2991 | 0.1064 | 0.1089 |
| $n = 10000$ | 0.3613 | 0.2783 | 0.1175 | 0.1177 |

Table 5.19: Average errors for the Brownian bridge based $MAH(1)$ process with exponential integral kernel over 100 simulations for various observation numbers.

| | $\tau_\varphi(\widehat{\varphi}_{\text{iter.}})$ | $\tau_\varphi(\widehat{\varphi}_{\text{innov.}})$ | F.cast. iter. | F.cast. innov. |
|---|---|---|---|---|
| $n = 100$ | 0.6614 | 0.6614 | 0.1430 | 0.1430 |
| $n = 500$ | 0.5979 | 0.5979 | 0.1606 | 0.1606 |
| $n = 1000$ | 0.5887 | 0.5887 | 0.1311 | 0.1311 |
| $n = 5000$ | 0.5896 | 0.5896 | 0.1523 | 0.1523 |
| $n = 10000$ | 0.5873 | 0.5873 | 0.2696 | 0.2696 |

Table 5.20: Average errors for the Brownian motion based $MAH(1)$ process with exponential integral kernel over 100 simulations for various observation numbers.

| | $\tau_\varphi(\widehat{\varphi}_{\text{iter.}})$ | $\tau_\varphi(\widehat{\varphi}_{\text{innov.}})$ | F.cast. iter. | F.cast. innov. |
|---|---|---|---|---|
| $n = 100$ | 2.1233 | 2.1560 | 0.0760 | 0.0750 |
| $n = 500$ | 0.9911 | 1.0022 | 0.0283 | 0.0280 |
| $n = 1000$ | 0.8031 | 0.8052 | 0.0290 | 0.0289 |
| $n = 5000$ | 0.5256 | 0.5224 | 0.0229 | 0.0229 |
| $n = 10000$ | 0.4734 | 0.4693 | 0.0172 | 0.0172 |

Table 5.21: Average errors for the Brownian bridge based $MAH(1)$ process with bilinear integral kernel over 100 simulations for various observation numbers.

| | $\tau_\varphi(\widehat{\varphi}_{\text{iter.}})$ | $\tau_\varphi(\widehat{\varphi}_{\text{innov.}})$ | F.cast. iter. | F.cast. innov. |
|---|---|---|---|---|
| $n = 100$ | 0.5627 | 0.5627 | 0.0983 | 0.0983 |
| $n = 500$ | 0.2767 | 0.2767 | 0.0993 | 0.0993 |
| $n = 1000$ | 0.2564 | 0.2564 | 0.0899 | 0.0899 |
| $n = 5000$ | 0.1603 | 0.1603 | 0.0874 | 0.0874 |
| $n = 10000$ | 0.1533 | 0.1533 | 0.0460 | 0.0460 |

Table 5.22: Average errors for the Brownian motion based $MAH(1)$ process with bilinear integral kernel over 100 simulations for various observation numbers.

Here the first main part of the thesis ends. These approaches will be applied in a similar way in Chapter 7.

# Chapter 6

# Testing $m$-Dependence of Functional Data

The previous chapters deal with moving average processes of order 1. In general we at first do not know which time series model to select and, if moving average, which order to choose. Thus the idea here is to develop a hypothesis test for the order of a moving average process. By Proposition 3.8 this is equivalent to testing the dependence order (recall Definition 3.5), if we assume the functional time series to be linear and strictly stationary. In the light of this fact we will develop two asymptotical $m$-dependence hypothesis tests for $m \in \mathbb{N}_0$. The first one is based on the ideas of [Moon] (which is about $m$-dependence of univariate data) and [GaKo] (which is concerned with independence of functional data) and generalizes their results. It is proven thouroughly at the end of this chapter. The second test is an improvement of the first one in the statistical power. Since many ideas are the same, some details in the derivation are omitted.

## 6.1 First Hypothesis Test

The idea is to construct an asymptotical $\chi^2$ test, where under the null hypothesis a random vector $\widetilde{R} \in \mathbb{R}^v$ is approximately $\mathcal{N}(0, \Xi)$ distributed, such that

$$N\widetilde{R}^T\widetilde{\Xi}^{-1}\widetilde{R} \xrightarrow{\mathscr{D}} \chi^2_v$$

follows. Actually we could derive an asymptotical multivariate normal test, but there is a big choice for confidence regions, which makes it hard to assess when to reject the null hypothesis. Hence we aim at establishing a univariate test, which leads to the $\chi^2$ distribution.
Let $N \in \mathbb{N}$ and $Y_1, \dots, Y_N \in L^2([0,1])$ be some functional observations of an $MAH(m)$ process with finite fourth moments. It suffices to work with the scores of the empirical principal components (see Section 2.2), because by means of Proposition 3.9 an $MAH(m)$ process projected onto the finite dimensional space spanned by the PCs is still a moving average process of order $m$ (which leads to $m$-dependence of the projected process according to Proposition 3.8). Furthermore as a symmetric positive-definite Hilbert-Schmidt

operator the sample covariance operator

$$\widehat{C}(\cdot) = \frac{1}{N}\sum_{n=1}^{N}\langle Y_i, \cdot\rangle Y_i$$

possesses orthonormal eigenfunctions $\widehat{v}_1, \widehat{v}_2, \ldots$ according to (2.2). This means that we do multivariate statistics although this chapter is about testing functional data. Hence let $p \in \{1, \ldots, N\}$ be the number of principal components and let $\mathbf{X}_1, \ldots, \mathbf{X}_N \in \mathbb{R}^p$ be the PC projections, i.e. $\mathbf{X}_n = (\mathbf{X}_{n,1}, \ldots, \mathbf{X}_{n,p})^T$ with $\mathbf{X}_{n,i} := \langle Y_n, \widehat{v}_i\rangle$ for all $n \in \{1, \ldots, N\}$ and $i \in \{1, \ldots, p\}$. In the following the tilde $\widetilde{\cdot}$ denotes a multivariate estimator.
We will need the sample and theoretical cross-covariance matrices of $(\mathbf{X}_n)_{n\in\{1,\ldots,N\}}$ for $h \in \{1, \ldots, N-1\}$

$$\widetilde{C}_h := \frac{1}{N}\sum_{n=1}^{N-h}\mathbf{X}_{n+h}\mathbf{X}_n^T, \qquad C_h := \mathbb{E}\big[\mathbf{X}_{1+h}\mathbf{X}_1^T\big]. \tag{6.1}$$

Since $(\mathbf{X}_n)_{n\in\{1,\ldots,N\}}$ are $m$-dependent, we can divide the whole sample into $k := m + 1$ subsamples such that each subsample consists of independent observations. Here we assume that $N$ is a multiple of $k$. This procedure is called *sample splitting* and looks as follows:

| subsample | 1 | ... | $k$ |
|---|---|---|---|
| | $\mathbf{X}_1$ | ... | $\mathbf{X}_k$ |
| | $\mathbf{X}_{k+1}$ | ... | $\mathbf{X}_{2k}$ |
| | $\vdots$ | | $\vdots$ |
| | $\mathbf{X}_{N-k+1}$ | ... | $\mathbf{X}_N$ |

For every $a \in \{1, \ldots, k\}$ and $t \in \{1, \ldots, \frac{N}{k}\}$ we denote

$$\mathbf{X}_t^{(a)} := \mathbf{X}_{(t-1)k+a} \tag{6.2}$$

the $t$-th observation of the $a$-th subsample. If the dependence order was smaller than $(k-1)$, but we still split $(\mathbf{X}_n)_{n\in\{1,\ldots,N\}}$ into $k$ subsamples, each subsample would still consist of independent elements. This would not hold, if the dependence order was greater than $(k-1)$.
Let $H \in \{1, \ldots, \frac{N}{k} - 1\}$ be the number of non-zero lags that we want to work with. For each $a \in \{1, \ldots, k\}$ and $h \in \{1, \ldots, H\}$ let $\widetilde{C}_h^{(a)}$ denote the $h$-th sample cross-covariance matrix (which corresponds to the $h$-lag) of the $a$-th subsample, i.e.

$$\widetilde{C}_h^{(a)} := \frac{k}{N}\sum_{t=1}^{\frac{N}{k}-h}\mathbf{X}_{t+h}^{(a)}\big(\mathbf{X}_t^{(a)}\big)^T, \tag{6.3}$$

which is supposed to be endowed with small entries. Furthermore, we need the sample covariance matrix $\widetilde{C}_0$ which can be computed from the entire sample. By means of Proposition 3.14 it converges to the theoretical covariance matrix $C_0 \in \mathbb{R}^{p\times p}$ of $(\mathbf{X}_n)_{n\in\{1,\ldots,N\}}$ in probability (because $L^p$ convergence implies convergence in probability). Define

$$\widetilde{R}_h^{(a)} := \widetilde{C}_0^{-1}\widetilde{C}_h^{(a)} \tag{6.4}$$

for every $a \in \{1, \dots, k\}$ and $h \in \{1, \dots, H\}$, which corresponds to the $h$-th sample cross-correlation matrix of the $a$-th subsample. Then we take the averages over the subsamples

$$\widetilde{R}_h := \frac{1}{k} \sum_{a=1}^{k} \widetilde{R}_h^{(a)}. \tag{6.5}$$

Now we aim at deriving the asymptotic behavior of the random vector

$$\widetilde{R} := (\mathrm{vec}(\widetilde{R}_1)^T, \dots, \mathrm{vec}(\widetilde{R}_H)^T)^T$$

of length $Hp^2$ in order to establish the asymptotical hypothesis test. We will show

**Lemma 6.1** *$\widetilde{R}_h = \widetilde{C}_0^{-1} \widetilde{C}_{hk}$ for all $h \in \{1, \dots, H\}$.*

Consequently it suffices to analyze the asymptotic behavior of the random vector

$$(\mathrm{vec}(\widetilde{C}_k)^T, \mathrm{vec}(\widetilde{C}_{2k})^T, \dots, \mathrm{vec}(\widetilde{C}_{Hk})^T)^T$$

of length $Hp^2$ consisting of the overall cross-covariance matrices with multiple lags of $k$. For $h_1, h_2 \in \{1, \dots, H\}$ define the asymptotical covariance matrices

$$\Psi_{h_1,h_2} := \lim_{N \to \infty} N \mathrm{cov}\big(\widetilde{C}_{h_1 k}, \widetilde{C}_{h_2 k}\big) \in \mathbb{R}^{p^2 \times p^2}$$

and assemble them to the block matrix $\Psi \in \mathbb{R}^{Hp^2 \times Hp^2}$

$$\Psi := \begin{pmatrix} \Psi_{1,1} & \cdots & \Psi_{1,H} \\ \vdots & \ddots & \vdots \\ \Psi_{H,1} & \cdots & \Psi_{H,H} \end{pmatrix},$$

which is supposed to equal the asymptotical covariance matrix of $\sqrt{N}(\mathrm{vec}(\widetilde{C}_k)^T, \dots, \mathrm{vec}(\widetilde{C}_{Hk})^T)^T$ by construction.

**Lemma 6.2 (Shape of the Covariance Matrix)** *Under strong stationarity and m-dependence*

$$\Psi_{h_1,h_2} = \begin{cases} \sum_{\nu=1}^{k-1} C_\nu^T \otimes C_{k-\nu}, & h_2 = h_1 + 1 \\ \sum_{\nu=1}^{k-1} C_\nu \otimes C_{k-\nu}^T, & h_1 = h_2 + 1 \\ \sum_{\nu=1}^{k-1} C_\nu^T \otimes C_\nu^T + \sum_{\nu=0}^{k-1} C_\nu \otimes C_\nu, & h_1 = h_2 > 1 \\ \frac{1}{k} \sum_{a<b} \left( \Delta_\alpha^{(a,b)} + \left(\Delta_\alpha^{(a,b)}\right)^T + \Delta_\beta^{(a,b)} + \left(\Delta_\beta^{(a,b)}\right)^T \right) + C_0 \otimes C_0, & h_1 = h_2 = 1 \\ 0, & \text{else} \end{cases}$$

*with fourth moments*

$$\Delta_{\alpha}^{(a,b)} := \mathbb{E}\Big[\big(\boldsymbol{X}_2^{(a)}(\boldsymbol{X}_2^{(b)})^T\big) \otimes \big(\boldsymbol{X}_1^{(a)}(\boldsymbol{X}_1^{(b)})^T\big)\Big], \tag{6.6}$$

$$\Delta_{\beta}^{(a,b)} := \mathbb{E}\Big[\big(\boldsymbol{X}_3^{(a)}(\boldsymbol{X}_2^{(b)})^T\big) \otimes \big(\boldsymbol{X}_2^{(a)}(\boldsymbol{X}_1^{(b)})^T\big)\Big]. \tag{6.7}$$

*This implies that* $\Psi$ *is a block band matrix with block band width* 1*.*

This determines the asymptotical distribution of $\sqrt{N}(\mathrm{vec}(\widetilde{C}_k)^T, \mathrm{vec}(\widetilde{C}_{2k})^T, \ldots, \mathrm{vec}(\widetilde{C}_{Hk})^T)^T$.

**Lemma 6.3 (Asymptotical Normality)** *Under strong stationarity and* $m$*-dependence*

$$\sqrt{N}(\mathrm{vec}(\widetilde{C}_k)^T, \mathrm{vec}(\widetilde{C}_{2k})^T, \ldots, \mathrm{vec}(\widetilde{C}_{Hk})^T)^T \xrightarrow{\mathscr{D}} \mathcal{N}_{Hp^2}(0, \Psi)$$

*holds for* $N \to \infty$*.*

Finally we are ready to derive the asymptotical hypothesis test. From a functional moving average process via multivariate statistics we reach the asymptotical distribution of the test statistic $\widetilde{Q}_N$ (see below), which consequently is likely to attain small values. Conversely if the test statistic attains large values, the asymptotical distribution and thus the moving average assumption is likely not to hold. Note that – as usual in statistical test theory – a small value of the test statistics does not imply the null hypothesis.
All the arguments hold as well if we have an $m$-dependent (functional) stationary time series, which is not a moving average process. It is because the "dependence [of a functional time series] is captured by only a few most important PCs" (c.f. [GaKo, Section 4]), which can be seen in Figure 5.6 (although it is an example of an $MAH(1)$ process). Since we derive the test with the most important PCs, the entire framework still holds.

**Theorem 6.4 ($m$-Dependence Hypothesis Test)** *Suppose strong stationarity and*

$$H_0: \text{ Data } m-\text{dependent}; \qquad H_1: \text{ Data at least } (m+1)-\text{dependent}.$$

*Then under* $H_0$

$$\widetilde{Q}_N := N\widetilde{R}^T\widetilde{\Xi}^{-1}\widetilde{R} \xrightarrow{\mathscr{D}} \chi^2_{p^2H}, \tag{6.8}$$

*where the empirical covariance matrix* $\widetilde{\Xi}$ *is a block band matrix*

$$\widetilde{\Xi} := \begin{pmatrix} \widetilde{M}_0 & \widetilde{M}_1 & & 0 \\ (\widetilde{M}_1)^T & \widetilde{M}_2 & \ddots & \\ & \ddots & \ddots & \widetilde{M}_1 \\ 0 & & (\widetilde{M}_1)^T & \widetilde{M}_2 \end{pmatrix} \tag{6.9}$$

*with*

$$\begin{aligned}
\widetilde{M}_0 &:= \frac{1}{k}\sum_{a<b}\Big(\widetilde{\Omega}_\alpha^{(a,b)} + \big(\widetilde{\Omega}_\alpha^{(a,b)}\big)^T + \widetilde{\Omega}_\beta^{(a,b)} + \big(\widetilde{\Omega}_\beta^{(a,b)}\big)^T\Big) + \widetilde{C}_0^{-1}\otimes\widetilde{C}_0,\\
\widetilde{\Omega}_\alpha^{(a,b)} &:= \big(\widetilde{C}_0^{-1}\otimes Id_p\big)\bigg(\frac{k}{N}\sum_{u=1}^{\frac{N}{k}-1}\big(\boldsymbol{X}_{u+1}^{(a)}(\boldsymbol{X}_{u+1}^{(b)})^T\big)\otimes\big(\boldsymbol{X}_u^{(a)}(\boldsymbol{X}_u^{(b)})^T\big)\bigg)\big(\widetilde{C}_0^{-1}\otimes Id_p\big),\\
\widetilde{\Omega}_\beta^{(a,b)} &:= \big(\widetilde{C}_0^{-1}\otimes Id_p\big)\bigg(\frac{k}{N}\sum_{u=1}^{\frac{N}{k}-2}\big(\boldsymbol{X}_{u+2}^{(a)}(\boldsymbol{X}_{u+1}^{(b)})^T\big)\otimes\big(\boldsymbol{X}_{u+1}^{(a)}(\boldsymbol{X}_u^{(b)})^T\big)\bigg)\big(\widetilde{C}_0^{-1}\otimes Id_p\big),\\
\widetilde{M}_1 &:= \sum_{\nu=1}^{k-1}\big(\widetilde{C}_0^{-1}\widetilde{C}_\nu^T\widetilde{C}_0^{-1}\big)\otimes\widetilde{C}_{k-\nu},\\
\widetilde{M}_2 &:= \sum_{\nu=1}^{k-1}\big(\widetilde{C}_0^{-1}\widetilde{C}_\nu^T\widetilde{C}_0^{-1}\big)\otimes\widetilde{C}_\nu^T + \sum_{\nu=0}^{k-1}\big(\widetilde{C}_0^{-1}\widetilde{C}_\nu\widetilde{C}_0^{-1}\big)\otimes\widetilde{C}_\nu.
\end{aligned}$$

*Hence, reject $H_0$ at significance level $\alpha\in(0;1)$ if and only if $Q > \chi^2_{p^2H,1-\alpha}$.*

In theory we are able to construct a hypothesis test without sample splitting. In this case we could calculate the overall cross-correlation matrices $\widetilde{C}_0^{-1}\widetilde{C}_h$ for all $h\in\{1,\dots,H\}$. However, under $(k-1)$-dependence the sample cross-covariance matrices $\widetilde{C}_h$ converge to some non-zero matrices in probability for all $h\in\{1,\dots,k-1\}$. We would need to know the latter (and to adjust $\Xi$) to construct some test statistics of the form

$$N(\widetilde{R}-\mu_R)^T\widetilde{\Xi}^{-1}(\widetilde{R}-\mu_R).$$

However, in practice we do not have any idea about $\mu_R$. Sample splitting makes us avoid this problem, because all subsample cross-covariances have expectation zero.

### 6.1.1 Proofs for the First Test

We will prove Lemmata 6.1, 6.2 and 6.3 (implied by 6.2 in the same way as [Brck, Theorem 7.3.1] implies [Brck, Theorem 7.3.2]) successively, which imply the claim of Theorem 6.4. The proof of Lemma 6.2 is the longest, because many tedious calculations concerning Kronecker products are involved.
The idea of this subsection is to express the subvectors of $\widetilde{R}$ by expressions of the form $\widetilde{C}_{hk}$, $h\in\{1,\dots,H\}$, to determine the joint asymptotical covariance matrix of the latter, to conclude their asymptotical distribution and finally to derive the hypothesis test.

**Proof of Lemma 6.1**

Take any $h \in \{1, \dots, H\}$. We plug in the definitions

$$\begin{aligned}
\widetilde{R}_h &\overset{(6.5)}{=} \frac{1}{k}\sum_{a=1}^{k} \widetilde{R}_h^{(a)} \overset{(6.4)}{=} \frac{1}{k}\widetilde{C}_0^{-1}\sum_{a=1}^{k}\widetilde{C}_h^{(a)} \overset{(6.3)}{=} \frac{1}{k}\widetilde{C}_0^{-1}\sum_{a=1}^{k}\frac{k}{N}\sum_{t=1}^{\frac{N}{k}-h}\mathbf{X}_{t+h}^{(a)}\big(\mathbf{X}_t^{(a)}\big)^T \\
&\overset{(6.2)}{=} \frac{1}{N}\widetilde{C}_0^{-1}\sum_{a=1}^{k}\sum_{t=1}^{\frac{N}{k}-h}\mathbf{X}_{(t+h-1)k+a}\mathbf{X}_{(t-1)k+a}^T \\
&= \frac{1}{N}\widetilde{C}_0^{-1}\sum_{a=1}^{k}\sum_{t=1}^{\frac{N}{k}-h}\mathbf{X}_{(t-1)k+a+hk}\mathbf{X}_{(t-1)k+a}^T.
\end{aligned}$$

We obtain two sums, where the indices $a \in \{1, \dots, k\}$ and $t \in \{1, \dots, \frac{N}{k} - h\}$ appear in the expression $(t-1)k+a$. This is the the notation of some number $u \in \{1, \dots, N-hk\}$ in the modulo representation with respect to $k$. Thus each $a$ and $t$ determine $u = (t-1)k+a$ uniquely. Therefore we are allowed to replace the two sums by one sum

$$\Big(\frac{1}{k}\sum_{a=1}^{k}\widetilde{C}_h^{(a)} = \Big)\frac{1}{N}\sum_{a=1}^{k}\sum_{t=1}^{\frac{N}{k}-h}\mathbf{X}_{(t-1)k+a+hk}\mathbf{X}_{(t-1)k+a}^T = \frac{1}{N}\sum_{u=1}^{N-hk}\mathbf{X}_{u+hk}\mathbf{X}_u^T \overset{(6.1)}{=} \widetilde{C}_{hk}. \quad (6.10)$$

Hence it turns out that $\widetilde{R}_h$ depends on the cross-covariance matrix $\widetilde{C}_{hk}$

$$\widetilde{R}_h = \widetilde{C}_0^{-1}\widetilde{C}_{hk}.$$

Since $h$ was chosen arbitrarily, the claim follows. □

**Proof of Lemma 6.2**

Choose any $h_1, h_2 \in \{1, \dots, H\}$.

$$\begin{aligned}
\lim_{N\to\infty} N\mathrm{cov}\big(\widetilde{C}_{h_1k}, \widetilde{C}_{h_2k}\big) &\overset{(6.10)}{=} \lim_{N\to\infty} N\mathrm{cov}\big(\frac{1}{k}\sum_{a=1}^{k}\widetilde{C}_{h_1}^{(a)}, \frac{1}{k}\sum_{b=1}^{k}\widetilde{C}_{h_2}^{(b)}\big) \\
&\overset{\text{Def. 6.5}}{=} \lim_{N\to\infty}\frac{N}{k^2}\sum_{a=1}^{k}\sum_{b=1}^{k}\mathbb{E}\Big[\mathrm{vec}\big(\widetilde{C}_{h_1}^{(a)}\big)\mathrm{vec}\big(\widetilde{C}_{h_2}^{(b)}\big)^T\Big] \\
&\overset{(6.3)}{=} \sum_{a=1}^{k}\sum_{b=1}^{k}\lim_{N\to\infty}\frac{N}{k^2}\Big(\frac{k}{N}\Big)^2\sum_{t=1}^{\frac{N_1}{k}-h_1}\sum_{u=1}^{\frac{N_2}{k}-h_2} \\
&\qquad \mathbb{E}\Big[\mathrm{vec}\big(\mathbf{X}_{t+h_1}^{(a)}(\mathbf{X}_t^{(a)})^T\big)\mathrm{vec}\big(\mathbf{X}_{u+h_2}^{(b)}(\mathbf{X}_u^{(b)})^T\big)^T\Big]
\end{aligned}$$

Fix any $a, b \in \{1, \dots, k\}$ and define $\widetilde{\Gamma}^{(a,b)}_{h_1,h_2}$ as the elements of the sum above over $a$ and $b$ before taking the limit

$$\begin{aligned}
\widetilde{\Gamma}^{(a,b)}_{h_1,h_2} &:= \frac{N}{k^2}\mathbb{E}\Big[\operatorname{vec}\big(\widetilde{C}^{(a)}_{h_1}\big)\operatorname{vec}\big(\widetilde{C}^{(b)}_{h_2}\big)^T\Big] \\
&= \frac{1}{N}\sum_{t=1}^{\frac{N}{k}-h_1}\sum_{u=1}^{\frac{N}{k}-h_2}\mathbb{E}\Big[\operatorname{vec}\big(\mathbf{X}^{(a)}_{t+h_1}(\mathbf{X}^{(a)}_t)^T\big)\operatorname{vec}\big(\mathbf{X}^{(b)}_{u+h_2}(\mathbf{X}^{(b)}_u)^T\big)^T\Big] \\
&\overset{(6.15)}{=} \frac{1}{N}\sum_{t=1}^{\frac{N}{k}-h_1}\sum_{u=1}^{\frac{N}{k}-h_2}\mathbb{E}\Big[\big(\mathbf{X}^{(a)}_{t+h_1}\otimes\mathbf{X}^{(a)}_t\big)\big(\mathbf{X}^{(b)}_{u+h_2}\otimes\mathbf{X}^{(b)}_u\big)^T\Big] \\
&\overset{\text{Lemma 6.11}}{=} \frac{1}{N}\sum_{t=1}^{\frac{N}{k}-h_1}\sum_{u=1}^{\frac{N}{k}-h_2}\mathbb{E}\Big[\big(\mathbf{X}^{(a)}_{t+h_1}\otimes\mathbf{X}^{(a)}_t\big)\big((\mathbf{X}^{(b)}_{u+h_2})^T\otimes(\mathbf{X}^{(b)}_u)^T\big)\Big] \\
&\overset{\text{Lemma 6.12}}{=} \frac{1}{N}\sum_{t=1}^{\frac{N}{k}-h_1}\sum_{u=1}^{\frac{N}{k}-h_2}\mathbb{E}\Big[\big(\mathbf{X}^{(a)}_{t+h_1}(\mathbf{X}^{(b)}_{u+h_2})^T\big)\otimes\big(\mathbf{X}^{(a)}_t(\mathbf{X}^{(b)}_u)^T\big)\Big].
\end{aligned}$$

Now the goal is to analyze the cases where $\widetilde{\Gamma}^{(a,b)}_{h_1,h_2}$ does not become 0. At first we consider $a < b$. Then there are three possible choices for $h_1$ and $h_2$ such that $\widetilde{\Gamma}^{(a,b)}_{h_1,h_2}$ does not equal zero.

- Consider $h_1+1 = h_2$ and denote $h = h_2$. Then there is only one possible combination for $t$ and $u$ such that $\widetilde{\Gamma}^{(a,b)}_{h,h+1}$ does not vanish:

| subsample | … | $a$ | … | $b$ | … |
|---|---|---|---|---|---|
| | | $\vdots$ | | $\vdots$ | |
| | … | $\vdots$ | … | $\mathbf{X}^{(b)}_u$ | … |
| | … | $\mathbf{X}^{(a)}_{u+1}$ | … | $\vdots$ | … |
| | | $\vdots$ | | $\vdots$ | |
| | … | $\mathbf{X}^{(a)}_{u+1+h}$ | … | $\mathbf{X}^{(b)}_{u+h+1}$ | … |
| | | $\vdots$ | | $\vdots$ | |

This implies $t = u + 1$. Hence the double sum in $\widetilde{\Gamma}^{(a,b)}_{h_1,h_2}$ can be reduced to one single sum. Furthermore we can use stationarity of $(\mathbf{X}_n)_{n\in\{1,\dots,N\}}$ and use the fact that

$\mathbf{X}^{(a)}_{2+h}(\mathbf{X}^{(b)}_{2+h})^T$ and $\mathbf{X}^{(a)}_{2}(\mathbf{X}^{(b)}_{1})^T$ are independent

$$\begin{aligned}
\widetilde{\Gamma}^{(a,b)}_{h,h+1} &= \frac{1}{N}\sum_{t=1}^{\frac{N}{k}-h}\sum_{u=1}^{\frac{N}{k}-(h+1)} \mathbb{E}\Big[\big(\mathbf{X}^{(a)}_{t+h}(\mathbf{X}^{(b)}_{u+h+1})^T\big)\otimes\big(\mathbf{X}^{(a)}_{t}(\mathbf{X}^{(b)}_{u})^T\big)\Big] \\
&= \frac{1}{N}\sum_{u=1}^{\frac{N}{k}-(h+1)} \mathbb{E}\Big[\big(\mathbf{X}^{(a)}_{u+1+h}(\mathbf{X}^{(b)}_{u+h+1})^T\big)\otimes\big(\mathbf{X}^{(a)}_{u+1}(\mathbf{X}^{(b)}_{u})^T\big)\Big] \\
&= \frac{1}{N}\sum_{u=1}^{\frac{N}{k}-(h+1)} \mathbb{E}\Big[\big(\mathbf{X}^{(a)}_{2+h}(\mathbf{X}^{(b)}_{2+h})^T\big)\otimes\big(\mathbf{X}^{(a)}_{2}(\mathbf{X}^{(b)}_{1})^T\big)\Big] \\
&= \frac{1}{N}\Big(\frac{N}{k}-(h+1)\Big)\mathbb{E}\Big[\big(\mathbf{X}^{(a)}_{2+h}(\mathbf{X}^{(b)}_{2+h})^T\big)\otimes\big(\mathbf{X}^{(a)}_{2}(\mathbf{X}^{(b)}_{1})^T\big)\Big] \\
&= \frac{1}{N}\Big(\frac{N}{k}-(h+1)\Big)\Big(\mathbb{E}\big[\mathbf{X}^{(a)}_{2+h}(\mathbf{X}^{(b)}_{2+h})^T\big]\Big)\otimes\Big(\mathbb{E}\big[\mathbf{X}^{(a)}_{2}(\mathbf{X}^{(b)}_{1})^T\big]\Big) \\
&\overset{(6.2)}{=} \frac{1}{N}\Big(\frac{N}{k}-(h+1)\Big)\Big(\mathbb{E}\big[\mathbf{X}_{(1+h)k+a}(\mathbf{X}_{(1+h)k+b})^T\big]\Big)\otimes\Big(\mathbb{E}\big[\mathbf{X}_{k+a}(\mathbf{X}_{b})^T\big]\Big) \\
&\overset{(6.1)}{=} \frac{1}{N}\Big(\frac{N}{k}-(h+1)\Big)C^T_{b-a}\otimes C_{k-(b-a)} \overset{N\to\infty}{\longrightarrow} \frac{1}{k}C^T_{b-a}\otimes C_{k-(b-a)}.
\end{aligned}$$

The next cases are argued in a very similar way.

- Consider $h_1 = h_2+1$ and denote $h = h_1$. Again there is only one possible combination for $t$ and $u$ such that $\widetilde{\Gamma}^{(a,b)}_{h+1,h}$ does not vanish:

| subsample | ... | $a$ | ... | $b$ | ... |
|---|---|---|---|---|---|
| | | $\vdots$ | | $\vdots$ | |
| | ... | $\mathbf{X}^{(a)}_{u}$ | ... | $\mathbf{X}^{(b)}_{u}$ | ... |
| | | $\vdots$ | | $\vdots$ | |
| | ... | $\vdots$ | ... | $\mathbf{X}^{(b)}_{u+h}$ | ... |
| | ... | $\mathbf{X}^{(a)}_{u+h+1}$ | ... | $\vdots$ | ... |
| | | $\vdots$ | | $\vdots$ | |

This implies $t = u$. Since the calculations become similar to the previous case, some steps are skipped.

$$\begin{aligned}
\widetilde{\Gamma}^{(a,b)}_{h+1,h} &= \frac{1}{N}\sum_{u=1}^{\frac{N}{k}-(h+1)} \mathbb{E}\Big[\big(\mathbf{X}^{(a)}_{u+1+h}(\mathbf{X}^{(b)}_{u+h})^T\big)\otimes\big(\mathbf{X}^{(a)}_{u}(\mathbf{X}^{(b)}_{u})^T\big)\Big] \\
&= \frac{1}{N}\Big(\frac{N}{k}-(h+1)\Big)\Big(\mathbb{E}\big[\mathbf{X}^{(a)}_{2+h}(\mathbf{X}^{(b)}_{1+h})^T\big]\Big)\otimes\Big(\mathbb{E}\big[\mathbf{X}^{(a)}_{1}(\mathbf{X}^{(b)}_{1})^T\big]\Big) \\
&\overset{(6.2)}{=} \frac{1}{N}\Big(\frac{N}{k}-(h+1)\Big)\Big(\mathbb{E}\big[\mathbf{X}_{(1+h)k+a}(\mathbf{X}_{hk+b})^T\big]\Big)\otimes\Big(\mathbb{E}\big[\mathbf{X}_{a}(\mathbf{X}_{b})^T\big]\Big) \\
&\overset{(6.1)}{=} \frac{1}{N}\Big(\frac{N}{k}-(h+1)\Big)C_{k-(b-a)}\otimes C^T_{b-a} \overset{N\to\infty}{\longrightarrow} \frac{1}{k}C_{k-(b-a)}\otimes C^T_{b-a}.
\end{aligned}$$

- Consider $h_1 = h_2 =: h$. This is the most complicated case, because there are two possible combinations for $t$ and $u$ such that $\widetilde{\Gamma}_{h,h}^{(a,b)}$ does not vanish:

| subsample | $\dots$ | $a$ | $\dots$ | $b$ | $\dots$ |
|---|---|---|---|---|---|
| | | $\vdots$ | | $\vdots$ | |
| | $\dots$ | $\mathbf{X}_u^{(a)}$ | $\dots$ | $\mathbf{X}_u^{(b)}$ | $\dots$ |
| | | $\vdots$ | | $\vdots$ | |
| | $\dots$ | $\mathbf{X}_{u+h}^{(a)}$ | $\dots$ | $\mathbf{X}_{u+h}^{(b)}$ | $\dots$ |
| | | $\vdots$ | | $\vdots$ | |

and

| subsample | $\dots$ | $a$ | $\dots$ | $b$ | $\dots$ |
|---|---|---|---|---|---|
| | | $\vdots$ | | $\vdots$ | |
| | $\dots$ | $\vdots$ | $\dots$ | $\mathbf{X}_u^{(b)}$ | $\dots$ |
| | $\dots$ | $\mathbf{X}_{u+1}^{(a)}$ | $\dots$ | $\vdots$ | $\dots$ |
| | | $\vdots$ | | $\vdots$ | |
| | $\dots$ | $\vdots$ | $\dots$ | $\mathbf{X}_{u+h}^{(b)}$ | $\dots$ |
| | $\dots$ | $\mathbf{X}_{u+h+1}^{(a)}$ | $\dots$ | $\vdots$ | $\dots$ |
| | | $\vdots$ | | $\vdots$ | |

Therefore merging the double sum results in two separate sums

$$\begin{aligned}
\widetilde{\Gamma}_{h,h}^{(a,b)} &= \frac{1}{N}\Bigg( \sum_{u=1}^{\frac{N}{k}-h} \mathbb{E}\Big[\big(\mathbf{X}_{u+h}^{(a)}(\mathbf{X}_{u+h}^{(b)})^T\big) \otimes \big(\mathbf{X}_u^{(a)}(\mathbf{X}_u^{(b)})^T\big)\Big] \\
&\quad + \sum_{u=1}^{\frac{N}{k}-(h+1)} \mathbb{E}\Big[\big(\mathbf{X}_{u+1+h}^{(a)}(\mathbf{X}_{u+h}^{(b)})^T\big) \otimes \big(\mathbf{X}_{u+1}^{(a)}(\mathbf{X}_u^{(b)})^T\big)\Big]\Bigg) \\
&= \frac{1}{N}\Bigg( \Big(\frac{N}{k} - h\Big)\mathbb{E}\Big[\big(\mathbf{X}_{1+h}^{(a)}(\mathbf{X}_{1+h}^{(b)})^T\big) \otimes \big(\mathbf{X}_1^{(a)}(\mathbf{X}_1^{(b)})^T\big)\Big] \\
&\quad + \Big(\frac{N}{k} - (h+1)\Big)\mathbb{E}\Big[\big(\mathbf{X}_{2+h}^{(a)}(\mathbf{X}_{1+h}^{(b)})^T\big) \otimes \big(\mathbf{X}_2^{(a)}(\mathbf{X}_1^{(b)})^T\big)\Big]\Bigg).
\end{aligned}$$

In contrast to the previous cases the factors of the Kronecker products are independent only if $h > 1$. For $h = 1$ we define

$$\Delta_<^{(a,b)} := \mathbb{E}\Big[\big(\mathbf{X}_2^{(a)}(\mathbf{X}_2^{(b)})^T\big) \otimes \big(\mathbf{X}_1^{(a)}(\mathbf{X}_1^{(b)})^T\big)\Big] + \mathbb{E}\Big[\big(\mathbf{X}_3^{(a)}(\mathbf{X}_2^{(b)})^T\big) \otimes \big(\mathbf{X}_2^{(a)}(\mathbf{X}_1^{(b)})^T\big)\Big]. \tag{6.11}$$

Consequently

$$\widetilde{\Gamma}_{1,1}^{(a,b)} \overset{N\to\infty}{\longrightarrow} \frac{1}{k}\Delta_<^{(a,b)}.$$

For $h > 1$ we obtain

$$\begin{aligned}\widetilde{\Gamma}_{h,h}^{(a,b)} &= \frac{1}{N}\Big(\Big(\frac{N}{k}-h\Big)\Big(\mathbb{E}\big[\mathbf{X}_{1+h}^{(a)}(\mathbf{X}_{1+h}^{(b)})^T\big]\otimes\mathbb{E}\big[\mathbf{X}_1^{(a)}(\mathbf{X}_1^{(b)})^T\big]\Big)\\ &\quad+\Big(\frac{N}{k}-(h+1)\Big)\Big(\mathbb{E}\big[\mathbf{X}_{2+h}^{(a)}(\mathbf{X}_{1+h}^{(b)})^T\big]\otimes\mathbb{E}\big[\mathbf{X}_2^{(a)}(\mathbf{X}_1^{(b)})^T\big]\Big)\Big)\\ &= \frac{1}{N}\Big(\Big(\frac{N}{k}-h\Big)\Big(C_{b-a}^T\otimes C_{b-a}^T\Big)+\Big(\frac{N}{k}-(h+1)\Big)\Big(C_{k-(b-a)}\otimes C_{k-(b-a)}\Big)\Big)\\ &\overset{N\to\infty}{\longrightarrow} \frac{1}{k}C_{b-a}^T\otimes C_{b-a}^T+\frac{1}{k}C_{k-(b-a)}\otimes C_{k-(b-a)}.\end{aligned}$$

The same procedure has to be done for $a > b$. We will skip the arguments and show only the results. As before, there are three cases where $\widetilde{\Gamma}_{h_1,h_2}^{(a,b)}$ does not vanish:

- For $h_1 + 1 = h_2 =: h$ we obtain

$$\widetilde{\Gamma}_{h,h+1}^{(a,b)} = \frac{1}{N}\Big(\frac{N}{k}-(h+1)\Big)C_{k-(a-b)}^T\otimes C_{a-b},$$

  which corresponds to $\big(\widetilde{\Gamma}_{h+1,h}^{(a,b)}\big)^T$ for $b > a$.

- For $h := h_1 = h_2 + 1$ we obtain

$$\widetilde{\Gamma}_{h+1,h}^{(a,b)} = \frac{1}{N}\Big(\frac{N}{k}-(h+1)\Big)C_{a-b}\otimes C_{k-(a-b)}^T,$$

  which corresponds to $\big(\widetilde{\Gamma}_{h,h+1}^{(a,b)}\big)^T$ for $b > a$.

- For $h_1 = h_2 =: h$ we again obtain two subcases, i.e. $h = 1$ and $h > 1$. It turns out

$$\begin{aligned}\widetilde{\Gamma}_{1,1}^{(a,b)} &= \frac{1}{N}\Big(\Big(\frac{N}{k}-h\Big)\mathbb{E}\Big[\big(\mathbf{X}_2^{(a)}(\mathbf{X}_2^{(b)})^T\big)\otimes\big(\mathbf{X}_1^{(a)}(\mathbf{X}_1^{(b)})^T\big)\Big]\\ &\quad+\Big(\frac{N}{k}-(h+1)\Big)\mathbb{E}\Big[\big(\mathbf{X}_2^{(a)}(\mathbf{X}_3^{(b)})^T\big)\otimes\big(\mathbf{X}_1^{(a)}(\mathbf{X}_2^{(b)})^T\big)\Big]\Big)\end{aligned}$$

  and

$$\widetilde{\Gamma}_{h,h}^{(a,b)} = \frac{1}{N}\Big(\Big(\frac{N}{k}-h\Big)\Big(C_{a-b}\otimes C_{a-b}\Big)+\Big(\frac{N}{k}-(h+1)\Big)\Big(C_{k-(a-b)}^T\otimes C_{k-(a-b)}^T\Big)\Big)$$

  for $h > 1$. The former satisfies

$$\begin{aligned}\widetilde{\Gamma}_{1,1}^{(a,b)} &\overset{N\to\infty}{\longrightarrow} \frac{1}{k}\Big(\mathbb{E}\Big[\big(\mathbf{X}_2^{(a)}(\mathbf{X}_2^{(b)})^T\big)\otimes\big(\mathbf{X}_1^{(a)}(\mathbf{X}_1^{(b)})^T\big)\Big]\\ &\quad+\mathbb{E}\Big[\big(\mathbf{X}_2^{(a)}(\mathbf{X}_3^{(b)})^T\big)\otimes\big(\mathbf{X}_1^{(a)}(\mathbf{X}_2^{(b)})^T\big)\Big]\Big) =: \frac{1}{k}\Delta_{>}^{(a,b)} \qquad (6.12)\end{aligned}$$

  The latter coresponds to $\big(\widetilde{\Gamma}_{h,h}^{(a,b)}\big)^T$ for $b > a$.

By contrast $a = b$ is much simpler, because we consider observations of the same subsample which was designed to consist of independent elements. The only possibility for $h_1$ and $h_2$ such that $\widetilde{\Gamma}^{(a,a)}_{h_1,h_2}$ does not equal zero is $h_1 = h_2 =: h$, which implies

$$\begin{aligned}
\widetilde{\Gamma}^{(a,a)}_{h,h} &= \frac{1}{N}\sum_{u=1}^{\frac{N}{k}-h} \mathbb{E}\Big[\big(\mathbf{X}^{(a)}_{u+h}(\mathbf{X}^{(a)}_{u+h})^T\big) \otimes \big(\mathbf{X}^{(a)}_{t}(\mathbf{X}^{(a)}_{u})^T\big)\Big] \\
&= \frac{1}{N}\Big(\frac{N}{k} - h\Big)\Big(\mathbb{E}\big[\mathbf{X}^{(a)}_{1+h}(\mathbf{X}^{(a)}_{1+h})^T\big] \otimes \mathbb{E}\big[\mathbf{X}^{(a)}_{1}(\mathbf{X}^{(a)}_{1})^T\big]\Big) \\
&= \frac{1}{N}\Big(\frac{N}{k} - h\Big)C_0 \otimes C_0 \overset{N\to\infty}{\longrightarrow} \frac{1}{k}C_0 \otimes C_0
\end{aligned}$$

Back to the asymptotical covariance of $\widetilde{C}_{h_1 k}$ and $\widetilde{C}_{h_2 k}$: By using $\widetilde{\Gamma}^{(a,b)}_{h_1,h_2}$ we obtain five cases all in all:

- For $h_2 = h_1 + 1$ we get with $h := h_1$

$$\begin{aligned}
\Psi_{h,h+1} &= \sum_{a<b}\lim_{N\to\infty}\widetilde{\Gamma}^{(a,b)}_{h,h+1} + \sum_{a>b}\lim_{N\to\infty}\widetilde{\Gamma}^{(a,b)}_{h,h+1} \\
&= \sum_{\nu=1}^{k-1}\frac{k-\nu}{k}C_\nu^T \otimes C_{k-\nu} + \sum_{\nu=1}^{k-1}\frac{k-\nu}{k}C_{k-\nu}^T \otimes C_\nu \\
&\overset{\tau:=k-\nu}{=} \sum_{\nu=1}^{k-1}\Big(1-\frac{\nu}{k}\Big)C_\nu^T \otimes C_{k-\nu} + \sum_{\tau=1}^{k-1}\frac{\tau}{k}C_\tau^T \otimes C_{k-\tau} \\
&= \sum_{\nu=1}^{k-1} C_\nu^T \otimes C_{k-\nu}.
\end{aligned}$$

  The transformation from the double sum $\sum_{b>a}(\cdot)$ to $\sum_{\nu=1}^{k-1}(k-\nu)(\cdot)$ results from the fact that the indices $a$ and $b$ actually occur only in the difference $b-a$. Instead of counting the possible values of $a$ and $b$ we can count the possible values of the whole difference, which makes the double sum merge to one single sum. This justifies the second step.

- For $h_1 = h_2 + 1$ we get with $h := h_2$

$$\begin{aligned}
\Psi_{h+1,h} &= \sum_{a<b}\lim_{N\to\infty}\widetilde{\Gamma}^{(a,b)}_{h+1,h} + \sum_{a>b}\lim_{N\to\infty}\widetilde{\Gamma}^{(a,b)}_{h+1,h} \\
&= \sum_{\nu=1}^{k-1}\frac{k-\nu}{k}C_{k-\nu} \otimes C_\nu^T + \sum_{\nu=1}^{k-1}\frac{k-\nu}{k}C_\nu \otimes C_{k-\nu}^T \\
&\overset{\tau:=k-\nu}{=} \sum_{\nu=1}^{k-1}\Big(1-\frac{\nu}{k}\Big)C_{k-\nu} \otimes C_\nu^T + \sum_{\tau=1}^{k-1}\frac{\tau}{k}C_{k-\tau} \otimes C_\tau^T \\
&= \sum_{\nu=1}^{k-1} C_{k-\nu} \otimes C_\nu^T = \sum_{\nu=1}^{k-1} C_\nu \otimes C_{k-\nu}^T.
\end{aligned}$$

- For $h > 1$ we get

$$\begin{aligned}
\Psi_{h,h} &= \sum_{a<b} \lim_{N\to\infty} \widetilde{\Gamma}_{h,h}^{(a,b)} + \sum_{a>b} \lim_{N\to\infty} \widetilde{\Gamma}_{h,h}^{(a,b)} + \sum_{a=1}^{k} \lim_{N\to\infty} \widetilde{\Gamma}_{h,h}^{(a,a)} \\
&= \sum_{\nu=1}^{k-1} \frac{k-\nu}{k} \big(C_\nu^T \otimes C_\nu^T + C_{k-\nu} \otimes C_{k-\nu}\big) \\
&\quad + \sum_{\nu=1}^{k-1} \frac{k-\nu}{k} \big(C_\nu \otimes C_\nu + C_{k-\nu}^T \otimes C_{k-\nu}^T\big) + \frac{k}{k} C_0 \otimes C_0 \\
&\overset{\tau:=k-\nu}{=} \sum_{\nu=1}^{k-1} \Big(1 - \frac{\nu}{k}\Big) \big(C_\nu^T \otimes C_\nu^T + C_{k-\nu} \otimes C_{k-\nu}\big) \\
&\quad + \sum_{\tau=1}^{k-1} \frac{\tau}{k} \big(C_{k-\tau} \otimes C_{k-\tau} + C_\tau^T \otimes C_\tau^T\big) + C_0 \otimes C_0 \\
&= \sum_{\nu=1}^{k-1} \big(C_\nu^T \otimes C_\nu^T + C_{k-\nu} \otimes C_{k-\nu}\big) + C_0 \otimes C_0 \\
&= \sum_{\nu=1}^{k-1} C_\nu^T \otimes C_\nu^T + \sum_{\nu=0}^{k-1} C_\nu \otimes C_\nu.
\end{aligned}$$

- For $h = 1$ we get

$$\begin{aligned}
\Psi_{1,1} &= \sum_{a<b} \lim_{N\to\infty} \widetilde{\Gamma}_{1,1}^{(a,b)} + \sum_{a>b} \lim_{N\to\infty} \widetilde{\Gamma}_{1,1}^{(a,b)} + \sum_{a=1}^{k} \lim_{N\to\infty} \widetilde{\Gamma}_{1,1}^{(a,a)} \\
&= \sum_{a<b} \Delta_<^{(a,b)} + \sum_{a>b} \Delta_>^{(a,b)} + C_0 \otimes C_0 \\
&\overset{(6.11),(6.12)}{=} \frac{1}{k} \sum_{a<b} \Big( \mathbb{E}\Big[ \big(\mathbf{X}_2^{(a)} (\mathbf{X}_2^{(b)})^T\big) \otimes \big(\mathbf{X}_1^{(a)} (\mathbf{X}_1^{(b)})^T\big) \Big] \\
&\quad + \mathbb{E}\Big[ \big(\mathbf{X}_3^{(a)} (\mathbf{X}_2^{(b)})^T\big) \otimes \big(\mathbf{X}_2^{(a)} (\mathbf{X}_1^{(b)})^T\big) \Big] \\
&\quad + \mathbb{E}\Big[ \big(\mathbf{X}_2^{(b)} (\mathbf{X}_2^{(a)})^T\big) \otimes \big(\mathbf{X}_1^{(b)} (\mathbf{X}_1^{(a)})^T\big) \Big] \\
&\quad + \mathbb{E}\Big[ \big(\mathbf{X}_2^{(b)} (\mathbf{X}_3^{(a)})^T\big) \otimes \big(\mathbf{X}_1^{(b)} (\mathbf{X}_2^{(a)})^T\big) \Big] \Big) + C_0 \otimes C_0 \\
&= \frac{1}{k} \sum_{a<b} \Big( \Delta_\alpha^{(a,b)} + \big(\Delta_\alpha^{(a,b)}\big)^T + \Delta_\beta^{(a,b)} + \big(\Delta_\beta^{(a,b)}\big)^T \Big) + C_0 \otimes C_0
\end{aligned}$$

  with $\Delta_\alpha^{(a,b)}$ and $\Delta_\beta^{(a,b)}$ as defined in (6.6) and (6.7).

- In all other cases we get $\lim_{N\to\infty} N\mathrm{cov}\big(\widetilde{C}_{h_1 k}, \widetilde{C}_{h_2 k}\big) = 0$.

At the end all the possible cases were analyzed and the claim follows. □

**Proof of Lemma 6.3**

The goal is to apply central limit theorem in order to gain asymptotical normality. One can show that for any $h \in \{1, \dots, N-1\}$ the estimators

$$\widetilde{C}_h^* := \frac{1}{N} \sum_{n=1}^{N} \mathbf{X}_{n+h} \mathbf{X}_n^T. \tag{6.13}$$

have the same asymptotic behavior as $\widetilde{C}_h$, because

$$\sqrt{N}\big(\widetilde{C}_h^* - \widetilde{C}_h\big) \xrightarrow{P} 0.$$

Thus we will show the result by using $\widetilde{C}_{hk}^*$ instead of $\widetilde{C}_{hk}$ for all $h \in \{1, \dots, H\}$. We therefore consider

$$\begin{pmatrix} \operatorname{vec}(\widetilde{C}_k^*) \\ \operatorname{vec}(\widetilde{C}_{2k}^*) \\ \vdots \\ \operatorname{vec}(\widetilde{C}_{Hk}^*) \end{pmatrix} \overset{\text{Lemma 6.7}}{=} \frac{1}{N} \left[ \begin{pmatrix} \operatorname{vec}\big(\mathbf{X}_{1+k}\mathbf{X}_1^T\big) \\ \operatorname{vec}\big(\mathbf{X}_{1+2k}\mathbf{X}_1^T\big) \\ \vdots \\ \operatorname{vec}\big(\mathbf{X}_{1+Hk}\mathbf{X}_1^T\big) \end{pmatrix} + \dots + \begin{pmatrix} \operatorname{vec}\big(\mathbf{X}_{n+k}\mathbf{X}_n^T\big) \\ \operatorname{vec}\big(\mathbf{X}_{n+2k}\mathbf{X}_n^T\big) \\ \vdots \\ \operatorname{vec}\big(\mathbf{X}_{n+Hk}\mathbf{X}_n^T\big) \end{pmatrix} \right]$$
$$=: \frac{1}{N} \sum_{n=1}^{N} \mathbf{Z}_n.$$

Thus consider the sequence $(\mathbf{Z}_n)_{n\in\mathbb{N}}$. It is strictly stationary because of strict stationarity of $(\mathbf{X}_n)_{n\in\mathbb{N}}$. Moreover since $C_h > 0$ for all $h \geq k$ holds, all $\mathbf{Z}_n$ have expectation zero. $(\mathbf{Z}_n)_{n\in\mathbb{N}}$ is $((H+1)k)$-dependent, because $(\mathbf{X}_n)_{n\in\mathbb{N}}$ is $(k-1)$-dependent and each $\mathbf{Z}_n$ contains $\mathbf{X}_{n+Hk}$ as the latest observation. Strict stationarity, expectation zero and $((H+1)k)$-dependence hold for the univariate process $(\lambda^T \mathbf{Z}_n)_{n\in\mathbb{N}}$ with any arbitrary $\lambda \in \mathbb{R}^{Hp^2}$ as well. Due to

$$\begin{aligned} \lim_{N\to\infty} \mathbb{V}\mathrm{ar}\Big(\frac{\sqrt{N}}{N} \sum_{n=1}^{N} \lambda^T \mathbf{Z}_n\Big) &= \lim_{N\to\infty} N\lambda^T \mathbb{V}\mathrm{ar}\Big(\frac{1}{N} \sum_{n=1}^{N} \mathbf{Z}_n\Big)\lambda \\ &= \lambda^T \Big( \lim_{N\to\infty} N \mathbb{V}\mathrm{ar}\big((\operatorname{vec}(\widetilde{C}_k^*)^T, \dots, \operatorname{vec}(\widetilde{C}_{Hk}^*)^T)^T\big)\Big)\lambda \\ &\overset{\text{Lemma 6.2}}{=} \lambda^T \Psi \lambda > 0 \end{aligned}$$

for all $\lambda \in \mathbb{R}^{Hp^2}$ such that $\lambda^T \Psi \lambda > 0$ Lemma 6.17 yields

$$\begin{aligned} 0 < \lim_{N\to\infty} N \mathbb{V}\mathrm{ar}\Big(\frac{1}{N} \sum_{n=1}^{N} \lambda^T \mathbf{Z}_n\Big) &\overset{\text{Lemma 6.17}}{=} \lim_{N\to\infty} \sum_{|h|<N} \big(1 - \frac{|h|}{N}\big)\gamma(h) \\ &\overset{((H+1)k)-\text{dep.}}{=} \lim_{N\to\infty} \sum_{|h|<(H+1)k} \big(1 - \frac{|h|}{N}\big)\gamma(h) = \sum_{|h|<(H+1)k} \gamma(h) \end{aligned}$$

with $\gamma(\cdot)$ as the autocovariance function for $(\lambda^T \mathbf{Z}_n)_{n\in\mathbb{N}}$. All in all we are allowed to apply the Central limit theorem for $((H+1)k)$-dependent data (Lemma 6.18). It states the

convergence

$$\frac{1}{\sqrt{N}}\sum_{n=1}^{N}\lambda^T\mathbf{Z}_n \xrightarrow{\mathscr{D}} \mathcal{N}(0,\lambda^T\Psi\lambda)$$

for all $\lambda\in\mathbb{R}^{Hp^2}$ such that $\lambda^T\Psi\lambda>0$. For $\lambda\in\mathbb{R}^{Hp^2}$ such that $\lambda^T\Psi\lambda=0$ we verify that

$$\frac{1}{\sqrt{N}}\sum_{n=1}^{N}\lambda^T\mathbf{Z}_n=0\xrightarrow{\mathscr{D}}0,$$

where 0 as a constant is normally distributed with mean 0 and variance 0. All in all due to arbitrariness of $\lambda\in\mathbb{R}^{Hp^2}$ Cramér-Wold device (Lemma 6.19) yields

$$\sqrt{N}(\mathrm{vec}(\widetilde{C}_k^*)^T,\ldots,\mathrm{vec}(\widetilde{C}_{Hk}^*)^T)^T\big)=\frac{1}{\sqrt{N}}\sum_{n=1}^{N}\mathbf{Z}_n\xrightarrow{\mathscr{D}}\mathcal{N}(0,\Psi).$$

Hence the result follows. □

**Proof of Theorem 6.4**

Note

$$\begin{aligned}\widetilde{R} &= \begin{pmatrix}\mathrm{vec}(\widetilde{R}_1)\\ \vdots\\ \mathrm{vec}(\widetilde{R}_H)\end{pmatrix}\overset{\text{Lemma 6.1}}{=}\begin{pmatrix}\mathrm{vec}(\widetilde{C}_0^{-1}\widetilde{C}_k)\\ \vdots\\ \mathrm{vec}(\widetilde{C}_0^{-1}\widetilde{C}_{Hk})\end{pmatrix}=\begin{pmatrix}\mathrm{vec}(\widetilde{C}_0^{-1}\widetilde{C}_k Id_p)\\ \vdots\\ \mathrm{vec}(\widetilde{C}_0^{-1}\widetilde{C}_{Hk} Id_p)\end{pmatrix}\\ &\overset{(6.14)}{=}\begin{pmatrix}\big(\widetilde{C}_0^{-1}\otimes Id_p\big)\mathrm{vec}(\widetilde{C}_k)\\ \vdots\\ \big(\widetilde{C}_0^{-1}\otimes Id_p\big)\mathrm{vec}(\widetilde{C}_{Hk})\end{pmatrix}=\begin{pmatrix}\widetilde{C}_0^{-1}\otimes Id_p & & 0\\ & \ddots & \\ 0 & & \widetilde{C}_0^{-1}\otimes Id_p\end{pmatrix}\begin{pmatrix}\mathrm{vec}(\widetilde{C}_k)\\ \vdots\\ \mathrm{vec}(\widetilde{C}_{Hk})\end{pmatrix}\end{aligned}$$

Now note that continuous mapping theorem (Lemma 6.15) holds for random matrices as well, because they can be vectorized. Thus

$$\begin{pmatrix}\widetilde{C}_0^{-1}\otimes Id_p & & 0\\ & \ddots & \\ 0 & & \widetilde{C}_0^{-1}\otimes Id_p\end{pmatrix}\xrightarrow{P}\begin{pmatrix}C_0^{-1}\otimes Id_p & & 0\\ & \ddots & \\ 0 & & C_0^{-1}\otimes Id_p\end{pmatrix}.$$

Furthermore Lemma 6.3 states

$$\sqrt{N}(\mathrm{vec}(\widetilde{C}_k)^T,\ldots,\mathrm{vec}(\widetilde{C}_{Hk})^T)^T\xrightarrow{\mathscr{D}}\mathcal{N}_{Hp^2}(0,\Psi).$$

Now remember the formula the formula $\mathbb{V}\mathrm{ar}(AW)=A\mathbb{V}\mathrm{ar}(W)A^T$ for any random vector $W$ and any real matrix $A$ with suitable dimensions. Altogether we gain by Slutsky's theorem (Lemma 6.16) and by Lemma 6.10

$$\sqrt{N}\widetilde{R}\xrightarrow{\mathscr{D}}\mathcal{N}_{Hp^2}(0,\Xi)$$

with the $Hp^2 \times Hp^2$ covariance matrix

$$\Xi := \begin{pmatrix} C_0^{-1} \otimes Id_p & & 0 \\ & \ddots & \\ 0 & & C_0^{-1} \otimes Id_p \end{pmatrix} \Psi \begin{pmatrix} C_0^{-1} \otimes Id_p & & 0 \\ & \ddots & \\ 0 & & C_0^{-1} \otimes Id_p \end{pmatrix}.$$

Analogously to what we did for $\Psi$ we analyze all blocks of $\Xi$, which are called

$$\Xi_{h_1,h_2} = \big(C_0^{-1} \otimes Id_p\big)\Psi_{h_1,h_2}\big(C_0^{-1} \otimes Id_p\big)$$

for $h_1, h_2 \in \{1, \ldots, H\}$.

- For $h_2 = h_1 + 1$ we get for $h := h_1$

$$\Xi_{h,h+1} \overset{\text{Lemma 6.12}}{=} \sum_{\nu=1}^{k-1} \big(C_0^{-1} C_\nu^T C_0^{-1}\big) \otimes C_{k-\nu}.$$

- For $h_1 = h_2 + 1$ we get for $h := h_2$

$$\Xi_{h+1,h} \overset{\text{Lemma 6.12}}{=} \sum_{\nu=1}^{k-1} \big(C_0^{-1} C_\nu C_0^{-1}\big) \otimes C_{k-\nu}^T = \Xi_{h,h+1}^T.$$

- For $h := h_1 = h_2 > 1$ we get

$$\Xi_{h,h} \overset{\text{Lemma 6.12}}{=} \sum_{\nu=1}^{k-1} \big(C_0^{-1} C_\nu^T C_0^{-1}\big) \otimes C_\nu^T + \sum_{\nu=0}^{k-1} \big(C_0^{-1} C_\nu C_0^{-1}\big) \otimes C_\nu.$$

- For $h_1 = h_2 = 1$ we get

$$\begin{aligned} \Xi_{1,1} \overset{\text{Lemma 6.12}}{=} & \frac{1}{k} \sum_{a<b} \Bigg( \mathbb{E}\Big[\big(C_0^{-1} \mathbf{X}_2^{(a)} (\mathbf{X}_2^{(b)})^T C_0^{-1}\big) \otimes \big(\mathbf{X}_1^{(a)} (\mathbf{X}_1^{(b)})^T\big)\Big] \\ & + \mathbb{E}\Big[\big(C_0^{-1} \mathbf{X}_2^{(b)} (\mathbf{X}_2^{(a)})^T C_0^{-1}\big) \otimes \big(\mathbf{X}_1^{(b)} (\mathbf{X}_1^{(a)})^T\big)\Big] \\ & + \mathbb{E}\Big[\big(C_0^{-1} \mathbf{X}_3^{(a)} (\mathbf{X}_2^{(b)})^T C_0^{-1}\big) \otimes \big(\mathbf{X}_2^{(a)} (\mathbf{X}_1^{(b)})^T\big)\Big] \\ & + \mathbb{E}\Big[\big(C_0^{-1} \mathbf{X}_2^{(b)} (\mathbf{X}_3^{(a)})^T C_0^{-1}\big) \otimes \big(\mathbf{X}_1^{(b)} (\mathbf{X}_2^{(a)})^T\big)\Big] \Bigg) + \big(C_0^{-1} C_0 C_0^{-1}\big) \otimes C_0 \\ = & \frac{1}{k} \sum_{a<b} \Big( \Omega_\alpha^{(a,b)} + \big(\Omega_\alpha^{(a,b)}\big)^T + \Omega_\beta^{(a,b)} + \big(\Omega_\beta^{(a,b)}\big)^T \Big) + C_0^{-1} \otimes C_0 \end{aligned}$$

  with

$$\begin{aligned} \Omega_\alpha^{(a,b)} &:= \mathbb{E}\Big[\big(C_0^{-1} \mathbf{X}_2^{(a)} (\mathbf{X}_2^{(b)})^T C_0^{-1}\big) \otimes \big(\mathbf{X}_1^{(a)} (\mathbf{X}_1^{(b)})^T\big)\Big], \\ \Omega_\beta^{(a,b)} &:= \mathbb{E}\Big[\big(C_0^{-1} \mathbf{X}_3^{(a)} (\mathbf{X}_2^{(b)})^T C_0^{-1}\big) \otimes \big(\mathbf{X}_2^{(a)} (\mathbf{X}_1^{(b)})^T\big)\Big]. \end{aligned}$$

- In all other cases we trivially obtain $\Xi_{h_1,h_2} = 0$, because $\Psi_{h_1,h_2} = 0$.

The estimator $\widetilde{\Xi}$ given in (6.9) consists of sample (cross-)covariance matrices $\widetilde{C}_h$, $h \in \{0,\dots,m\}$, $\widetilde{C}_0^{-1}$ and the empirical fourth moments. All of them are consistent. Thus they converge to their empirical counterparts $C_h$, $h \in \{0,\dots,m\}$ in probability. Moreover Lemma 6.14 also holds for random matrices, because they can be vectorized. By applying Lemma 6.14 and the continuous mapping theorem (Lemma 6.15) several times it turns out

$$\widetilde{\Xi} \xrightarrow{P} \Xi.$$

Since both $\widetilde{\Xi}$ and $\Xi$ are symmetric and positive definite, they possess real positive definite square roots $\widetilde{\Xi}^{\frac{1}{2}}$ and $\Xi^{\frac{1}{2}}$ by the spectral theorem with the property

$$\widetilde{\Xi}^{\frac{1}{2}} \xrightarrow{P} \Xi^{\frac{1}{2}}.$$

Consequently Slutsky's Theorem implies

$$\widetilde{Q}_N = \sqrt{N}\widetilde{\Xi}^{-\frac{1}{2}}\widetilde{R} \xrightarrow{\mathscr{D}} \mathcal{N}_{p^2H}(0, Id_{p^2H}) \overset{\mathscr{D}}{=}: Z,$$

Since the squared euclidean norm $g(x) = \|x\|_2^2$ is continuous, continuous mapping theorem (Lemma 6.15) implies

$$N\widetilde{R}^T\widetilde{\Xi}^{-1}\widetilde{R} = (\sqrt{N}\widetilde{\Xi}^{-\frac{1}{2}}\widetilde{R})^T\sqrt{N}\widetilde{\Xi}^{-\frac{1}{2}}\widetilde{R} = g\big(\sqrt{N}\widetilde{\Xi}^{-\frac{1}{2}}\widetilde{R}\big) \xrightarrow{\mathscr{D}} g(Z) \overset{\mathscr{D}}{=} \chi^2_{p^2H}.$$

The last part follows by definition of the $\chi^2$ distribution and by the fact that independence and uncorrelatedness are equivalent for normally distributed random variables. □

**Remark:** We consider two special cases:

- Suppose $p = 1$. Then $Q_N \xrightarrow{\mathscr{D}} \chi^2_H$ holds under $H_0$ with $\widetilde{R} \in \mathbb{R}^H$ and $\widetilde{\Xi} \in \mathbb{R}^{H\times H}$. This is exactly what Lemma 2.9 of [Moon] is about.

- Suppose we want to test independence, which means $k = 1$. Then one can show that $(\widetilde{C}_h)_{h\geq 0}$ are independent. Consequently so are $(Z_h)_{h\geq 0}$ and $(R_h)_{h\geq 0}$ (c.f. [GaKo]). Furthermore it turns out that the asymptotical covariance matrix of $R_h$ equals $C_0^{-1} \otimes C_0 \in \mathbb{R}^{p^2\times p^2}$ for all $h \in \{1,\dots,H\}$. Hence it suffices to consider any $h \in \{1,\dots,H\}$ and to take the sum over $h$ at the end, because here $\Xi$ looks like

$$\begin{pmatrix} C_0^{-1}\otimes C_0 & & 0 \\ & \ddots & \\ 0 & & C_0^{-1}\otimes C_0 \end{pmatrix}.$$

  Therefore use Lemma 6.10 and Lemma 6.13 to obtain

$$\begin{aligned} \mathrm{vec}(C_0^{-1}Z_h)^T(C_0\otimes C_0^{-1})\mathrm{vec}(C_0^{-1}Z_h) &= \mathrm{vec}(C_0^{-1}Z_h)^T\mathrm{vec}((C_0^{-1}Z_h^T)^T) \\ &= \mathrm{vec}(C_0^{-1}Z_h)^T\mathrm{vec}(Z_hC_0^{-1}). \end{aligned}$$

  This is exactly the limit of $\mathrm{vec}(\widetilde{C}_0^{-1}\widetilde{C}_h)^T\mathrm{vec}(\widetilde{C}_h\widetilde{C}_0^{-1})$. All in all we get

$$Q_N = N\sum_{h=1}^{H}\mathrm{vec}(\widetilde{C}_0^{-1}\widetilde{C}_h)^T\mathrm{vec}(\widetilde{C}_h\widetilde{C}_0^{-1}),$$

  which equals the test statistics in [GaKo].

### 6.1.2 Appendix

Matrix and vector distributions, the Kronecker product with some properties and convergence of random variables are needed for the proofs.

**Definition 6.5 (Covariance matrix of random vectors)** *Let $X$ and $Y$ be two random vectors in $\mathbb{R}^p$ and $\mathbb{R}^q$ respectively. Then we define*

$$\operatorname{cov}(X,Y) := \mathbb{E}\big[(X-\mathbb{E}[X])(Y-\mathbb{E}[Y])^T\big] = \mathbb{E}[XY^T] - \mathbb{E}[X]\mathbb{E}[Y]^T \in \mathbb{R}^{p\times q}$$

*as the* covariance matrix *of them. The variance of a random vector $X$ in $\mathbb{R}^p$ is defined intuitively $\mathbb{V}\mathrm{ar}(X) := \operatorname{cov}(X,X) \in \mathbb{R}^{p\times p}$.*

In this chapter the row and column indices of matrices are written in brackets.

**Definition 6.6 (Rowwise vectorizer, c.f. [GaKo, App. B])** *For any matrix $A \in \mathbb{R}^{n_1\times n_2}$ the function* vec *applied on $A$ denotes vectorizing $A$ row by row*

$$\operatorname{vec}(A) = (A(1,1),\dots,A(1,n_2),A(2,1),\dots,A(n_1,1),\dots,A(n_1,n_2))^T \in \mathbb{R}^{n_1 n_2}$$

*as a column vector of length $n_1 \cdot n_2$. Thus* vec *counts the last index before the first one.*

**Lemma 6.7 (Linearity of** vec**)** *For any $\lambda \in \mathbb{R}$ and for any $A, B \in \mathbb{R}^{n_1\times n_2}$*

$$\operatorname{vec}(\lambda A) = \lambda \operatorname{vec}(A), \qquad \operatorname{vec}(A+B) = \operatorname{vec}(A) + \operatorname{vec}(B).$$

*Thus* vec *is linear.*

Proof:

$$\begin{aligned}\operatorname{vec}(\lambda A) &= (\lambda A(1,1),\dots,\lambda A(1,n_2),\lambda A(2,1),\dots,\lambda A(n_1,1),\dots,\lambda A(n_1,n_2))^T \\ &= \lambda(A(1,1),\dots,A(1,n_2),A(2,1),\dots,A(n_1,1),\dots,A(n_1,n_2))^T = \lambda\operatorname{vec}(A)\end{aligned}$$

and

$$\begin{aligned}\operatorname{vec}(A+B) &= (A(1,1)+B(1,1),\dots,A(1,n_2)+B(1,n_2),A(2,1)+B(2,1),\dots, \\ &\quad A(n_1,1)+B(n_1,1),\dots,A(n_1,n_2)+B(n_1,n_2))^T \\ &= (A(1,1),\dots,A(1,n_2),A(2,1),\dots,A(n_1,1),\dots,A(n_1,n_2))^T \\ &\quad +(B(1,1),\dots,B(1,n_2),B(2,1),\dots,B(n_1,1),\dots,B(n_1,n_2))^T \\ &= \operatorname{vec}(A)+\operatorname{vec}(B).\end{aligned}$$

Hence linearity is proven. □

**Definition 6.8 (Matrix normal distribution)** *An $n_1 \times n_2$ random matrix $X$ is said to be normally distributed if $\operatorname{vec}(X) \sim \mathcal{N}_{n_1\cdot n_2}(\eta,\Psi)$ holds with $\eta = \mathbb{E}[\operatorname{vec}(X)]$ and $\Psi = \mathbb{V}\mathrm{ar}(\operatorname{vec}(X))$. We then write*

$$X \sim \mathcal{N}_{n_1\times n_2}(M,\Psi)$$

*with $M = \mathbb{E}[X]$ and $\Psi$ as before.*

**Definition 6.9 (Kronecker product, c.f. [Steeb, Sec. 1.1])** *For two matrices* $A \in \mathbb{R}^{n_1 \times n_2}$ *and* $B \in \mathbb{R}^{n_3 \times n_4}$ *the symbol* $\otimes$ *denotes the* Kronecker product

$$A \otimes B := \begin{pmatrix} A(1,1)B & \dots & A(1,n_2)B \\ \vdots & \ddots & \vdots \\ A(n_1,1)B & \dots & A(n_1,n_2)B \end{pmatrix} \in \mathbb{R}^{(n_1 n_3)\times(n_2 n_4)}.$$

*Thus the Kronecker product can be indexed by double indices: For any* $i \in \{1,\dots,n_1\}$, $j \in \{1,\dots,n_2\}$, $w \in \{1,\dots,n_3\}$ *and* $l \in \{1,\dots,n_4\}$ *we write*

$$(A \otimes B)((i,w),(j,l)) := A(i,j)B(w,l),$$

*such that the indices of* $B$ *are counted before the indices of* $A$.

We will need some properties of the Kronecker product.

**Lemma 6.10 (Inverse of the Kronecker product, c.f. [Steeb, Sec. 1.3])** $(A \otimes B)$ *is invertible*

$$(A \otimes B)^{-1} = A^{-1} \otimes B^{-1}$$

*for all invertible squared matrices* $A, B \in \mathbb{R}^{v \times v}$.

Proof: See [Steeb, Sec. 1.3]. □

**Lemma 6.11 (Transpose of the Kronecker product, c.f. [Steeb, Sec. 1.2])** *Transposing a Kronecker product equals transposing its factors*

$$(A \otimes B)^T = A^T \otimes B^T$$

*for all matrices* $A \in \mathbb{R}^{n_1 \times n_2}$ *and* $B \in \mathbb{R}^{n_3 \times n_4}$.

Proof: See [Steeb, Sec. 1.2]. □

**Lemma 6.12 (Product of Kronecker products, c.f. [Steeb, Sec. 1.3])** *Let* $A \in \mathbb{R}^{n_1 \times n_2}$, $B \in \mathbb{R}^{n_3 \times n_4}$, $C \in \mathbb{R}^{n_5 \times n_6}$ *and* $D \in \mathbb{R}^{n_7 \times n_8}$ *be any given matrices. Then*

$$(A \otimes B)(C \otimes D) = (AB) \otimes (CD),$$

*which means that the matrix product of two Kronecker products equals the Kronecker product of the matrix products of the individual factors.*

Proof: See [Steeb, Sec. 1.3]. □

**Lemma 6.13 (Kronecker product and** vec**, c.f. [Steeb, Sec. 1.12])** *Let* $A \in \mathbb{R}^{n_1 \times n_2}$, $X \in \mathbb{R}^{n_2 \times n_3}$ *and* $B \in \mathbb{R}^{n_3 \times n_4}$. *Then*

$$AXB = M \iff (B^T \otimes A)\mathrm{vec}(X^T) = \mathrm{vec}(M^T). \tag{6.14}$$

*A consequence of this is*

$$\mathrm{vec}(W_1 W_2^T) = \mathrm{vec}\big((W_2 \cdot 1 \cdot W_1^T)^T\big) \overset{(6.14)}{=} (W_1 \otimes W_2)\mathrm{vec}(1) = (W_1 \otimes W_2) \tag{6.15}$$

*for any vectors* $W_1 \in \mathbb{R}^{n_1}$ *and* $W_2 \in \mathbb{R}^{n_2}$.

Proof: We will prove the claim elementwise. Thus let $i \in \{1, \dots, n_1\}$ and $j \in \{1, \dots, n_4\}$. Recall that transposing of any matrix is defined by interchanging its indices. Thus the calculation

$$\begin{aligned}(AXB)(i,j) &= \sum_{w=1}^{n_2}\sum_{l=1}^{n_3} A(i,w)X(w,l)B(l,j) = \sum_{w=1}^{n_2}\sum_{l=1}^{n_3}(B^T)(j,l)A(i,w)(X^T)(l,w)\\ &\overset{\text{Def. 6.9}}{=} \sum_{w=1}^{n_2}\sum_{l=1}^{n_3}(B^T \otimes A)((j,i),(l,w))\mathrm{vec}(X^T)(l,w)\\ &= \big((B^T \otimes A)\mathrm{vec}(X^T)\big)(j,i) = \Big(\big((B^T \otimes A)\mathrm{vec}(X^T)\big)^T\Big)(i,j)\end{aligned}$$

gives us the equivalence. □

**Lemma 6.14 (c.f. [Vaart, Thm. 2.7 (vi)])** *Let $(X_n)_{n\in\mathbb{N}}$ and $(Y_n)_{n\in\mathbb{N}}$ be sequences of real random column vectors in $\mathbb{R}^{n_1}$ and respectively $\mathbb{R}^{n_2}$ converging to random vectors $X$ and respectively $Y$ in probability. Then*

$$(X_n^T, Y_n^T)^T \xrightarrow{P} (X^T, Y^T)^T,$$

*for $n \to \infty$.*

Proof: See [Vaart, Thm. 2.7 (vi)]. □

**Lemma 6.15 (Continuous mapping theorem, c.f. [Vaart, Thm. 2.3])** *Let $n_1, n_2 \in \mathbb{N}$, $(X_n)_{n\in\mathbb{N}}$ be a sequence of random vectors in $\mathbb{R}^{n_1}$, $X$ be a random vector in $\mathbb{R}^{n_1}$ and $g : \mathbb{R}^{n_1} \to \mathbb{R}^{n_2}$ a function that is continuous at every point of a set $S \subset \mathbb{R}^{n_1}$ such that $\mathbb{P}(X \in S) = 1$ holds. Then*

*(i) if $X_n \xrightarrow{\mathscr{D}} X$, then $g(X_n) \xrightarrow{\mathscr{D}} g(X)$;*

*(ii) if $X_n \xrightarrow{P} X$, then $g(X_n) \xrightarrow{P} g(X)$ and*

*(iii) if $X_n \xrightarrow{\text{a.s.}} X$, then $g(X_n) \xrightarrow{\text{a.s.}} g(X)$.*

Proof: See [Vaart, Thm. 2.3]. □

**Lemma 6.16 (Slutsky-Theorem, c.f. [Vaart, Lemma 2.8])** *Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of random vectors in $\mathbb{R}^p$, $p \in \mathbb{N}$, converging to some random vector $X$ in $\mathbb{R}^p$ in distribution. Furthermore let $(\Sigma_n)_{n\in\mathbb{N}}$ be a sequence of invertible random matrices in $\mathbb{R}^{p\times p}$ converging to some invertible non-random matrix $\Sigma \in \mathbb{R}^{p\times p}$ in probability. Then $\Sigma_n^{-1}X_n \xrightarrow{\mathscr{D}} \Sigma^{-1}X$ holds for $n \to \infty$.*

Proof: See [Vaart, Thm. 2.3 & Thm. 2.7 (v)]. □

**Lemma 6.17 (c.f. [Brck, Theorem 7.1.1])** *Let $(Z_t)_{t\in\mathbb{Z}}$ be a strictly stationary sequence of scalar random variables with mean 0 and autocovariance function $\gamma$. Then*

$$n\mathbb{V}\mathrm{ar}(\overline{Z_n}) = \sum_{|h|<n}\big(1 - \frac{|h|}{n}\big)\gamma(h)$$

*with $\overline{Z_n} := \frac{1}{n}\sum_{i=1}^{n} Z_i$.*

Proof: We plug in the definition of $\overline{Z_n}$. Note that for univariate time series $\gamma(h) = \gamma(-h)$ holds for all $h \in \mathbb{Z}$. Then

$$\begin{aligned} n\mathbb{V}\mathrm{ar}\Big(\frac{1}{n}\sum_{i=1}^{n} Z_i\Big) &= \frac{1}{n}\sum_{i,j=1}^{n} \mathrm{cov}(Z_i, Z_j) = \frac{1}{n}\sum_{i,j=1}^{n} \gamma(i-j) \\ &\overset{h:=i-j}{=} \frac{1}{n}\sum_{|h|<n} (n-|h|)\gamma(h) = \sum_{|h|<n} \big(1 - \frac{|h|}{n}\big)\gamma(h) \end{aligned}$$

holds by counting the differences instead of the single indices $i$ and $j$. □

**Lemma 6.18 (Central limit theorem for $m$-dependent data, c.f. [Brck, Theorem 6.4.2])** *Let $(Z_t)_{t\in\mathbb{Z}}$ be a strictly stationary $m$-dependent sequence of scalar random variables with mean 0 and autocovariance function $\gamma$. If $v_m = \sum_{|h|<m} \gamma(h) \neq 0$, then*

*(i)* $\lim_{n\to\infty} n\mathbb{V}\mathrm{ar}(\overline{Z_n}) = v_m$ *and*

*(ii)* $\sqrt{n}\overline{Z_n} \xrightarrow{\mathscr{D}} \mathcal{N}(0, v_m)$ *for* $n \to \infty$

*hold.*

**Lemma 6.19 (Cramér-Wold device, c.f. [Brck, Theorem 6.3.1])** *Let $(Z_n)_{n\in\mathbb{N}}$ be a sequence of random vectors in $\mathbb{R}^v$. Then we have*

$$Z_n \xrightarrow{\mathscr{D}} Z \qquad \Leftrightarrow \qquad \lambda^T Z_n \xrightarrow{\mathscr{D}} \lambda^T Z \qquad \forall \lambda^T = (\lambda_1, \ldots, \lambda_v) \in \mathbb{R}^v$$

*for some random vector $Z$ in $\mathbb{R}^v$.*

## 6.2 Second Hypothesis Test

Theorem 6.4 in combination with Lemma 6.1 states that it suffices to consider all cross-covariance matrices with multiple lags of $k$. Indeed, if $Q$ is too large, then $H_0$ cannot be assumed. However, the power (i.e. one minus the type II error) of this test is too low, as in Subsection 6.3.2 shows for an $MAH(3)$ process $(Y_n)_{n\in\mathbb{Z}}$ of the form

$$Y_n = \varepsilon_n + l\varepsilon_{n-3}, \qquad \forall n \in \mathbb{Z}.$$

Its cross-covariance matrices with lags 1 and 2 equal zero, but not the one with lag 3. Imagine we wish to test 1-dependence. Therefore the test in Theorem 6.4 considers only the cross-covariance matrices with even lags, which in this case equal zero altogether. Hence the null hypothesis $H_0$ of 1-dependence cannot be rejected, although this process is 3-dependent actually, which means that the alternative hypothesis $H_1$ holds.

This example illustrates that for testing $m$-dependence it is worth taking all the remaining covariance matrices with lags greater than $m$ into account in addition to the ones with multiple lags of $m + 1 = k$. However, if we consider

$$(\mathrm{vec}(\widetilde{C}_k)^T, \mathrm{vec}(\widetilde{C}_{k+1})^T, \ldots)^T,$$

then the covariance matrix of this vector becomes very tedious to evaluate. The benefit of working with multiple lags of $k$ is the block band matrix structure of $\Xi$. Thus a compromise of augmenting the number of lags in order to increase the statistical power and of keeping the covariance matrix as simple as possible is derived in this section. The idea is to take all the random vectors

$$\begin{pmatrix} \mathrm{vec}(\widetilde{C}_k) \\ \mathrm{vec}(\widetilde{C}_{2k}) \\ \vdots \\ \mathrm{vec}(\widetilde{C}_{Hk}) \end{pmatrix}, \begin{pmatrix} \mathrm{vec}(\widetilde{C}_{k+1}) \\ \mathrm{vec}(\widetilde{C}_{2k+1}) \\ \vdots \\ \mathrm{vec}(\widetilde{C}_{Hk+1}) \end{pmatrix}, \ldots, \begin{pmatrix} \mathrm{vec}(\widetilde{C}_{2k-1}) \\ \mathrm{vec}(\widetilde{C}_{3k-1}) \\ \vdots \\ \mathrm{vec}(\widetilde{C}_{(H+1)k-1}) \end{pmatrix} \tag{6.16}$$

into account. All of them contain cross-covariance matrices $\widetilde{C}_h$ with $h$ such that $+k$ is added. Therefore covariance matrices of these vectors are block band matrices with block band width 1, such that the bands only consist of

$$\sum_{\nu=1}^{k-1} C_\nu^T \otimes C_{k-\nu} \qquad \text{and} \qquad \sum_{\nu=1}^{k-1} C_\nu \otimes C_{k-\nu}^T$$

respectively analogously to Lemma 6.2. These formulas can be derived in a similar way as in Subsection 6.1.1 – by imagining the possible combinations of matchings, but on a horizontal line and not between subsamples. The reason for the independence of $\widetilde{C}_h$ and $\widetilde{C}_{h+wk}$ with $w \in \{2, 3, \ldots\}$ for all $h \in \mathbb{Z}$ is that for an expression of the form

$$\mathbb{E}[\mathrm{vec}(\mathbf{X}_{i+h+wk}\mathbf{X}_i^T)\mathrm{vec}(\mathbf{X}_{j+h}\mathbf{X}_j^T)^T]$$

it is impossible for any fixed $i \in \{1, \ldots, N-h+wk\}$ to choose $j \in \{1, \ldots, N-h\}$ such that all of those four $\mathbf{X}_{(\cdot)}$ depend on another $\mathbf{X}_{(\cdot)}$, i.e. in each case there is a vector $\mathbf{X}_{(\cdot)}$ left which is independent from the others, so that the whole expectation becomes zero.
In light of 6.3 all the $k$ random vectors in (6.16) are multivariate normally distributed with mean zero and with block band matrices as covariance matrices. Now the idea is to construct $k$ test statistics of the same form as in (6.8). For this it remains to find out the diagonal blocks of their covariance matrices. It turns out that they do not change from one specific lag on.

**Lemma 6.20** *Under strong stationarity and m-dependence*

$$\lim_{N\to\infty} N\mathbb{V}\mathrm{ar}(\widetilde{C}_h) = \sum_{\nu=-(k-1)}^{k-1} C_\nu \otimes C_\nu$$

*for all* $h \geq 2m+1 (= 2k-1)$ *analogously to Lemma 6.2.*

Sketch of proof: Choose any $h \geq 2m+1$ and consider the expectation on the right hand side of

$$N\mathbb{V}\mathrm{ar}(\widetilde{C}_h) = \frac{1}{N}\sum_{i=1}^{N-h}\sum_{j=1}^{N-h} \mathbb{E}[\mathrm{vec}(\mathbf{X}_{i+h}\mathbf{X}_i^T)\mathrm{vec}(\mathbf{X}_{j+h}\mathbf{X}_j^T)^T]. \tag{6.17}$$

Fix any $i \in \{1, \ldots, N-h\}$. Due to $m$-dependence and due to the choice of $h$ $\mathbf{X}_i$ and $\mathbf{X}_{i+h}$ are so far away that they cannot depend on any common $\mathbf{X}_{(\cdot)}$, which is illustrated in Figure 6.1. Consequently for all choices of $j \in \{1, \ldots, N-h\}$ the expression $\mathbb{E}[\mathrm{vec}(\mathbf{X}_{i+h}\mathbf{X}_i^T)\mathrm{vec}(\mathbf{X}_{j+h}\mathbf{X}_j^T)^T]$ becomes either zero or the Kronecker product $C_{i-j} \otimes C_{i-j}$. Thus the limit of the right hand side of (6.17) leads to the claim and does not depend on $h \geq 2m+1$. □

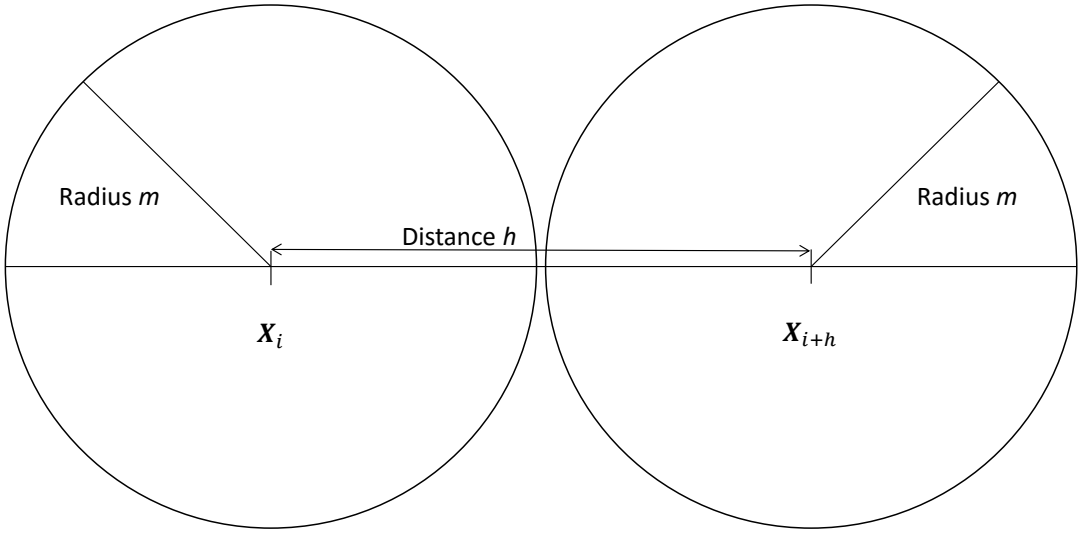


Figure 6.1: Univariate illustration of the time series with $\mathbf{X}_i$ and $\mathbf{X}_{i+h}$.

Now it remains to compute the asymptotical covariance matrices of each of the estimators $\widetilde{C}_k, \ldots, \widetilde{C}_{2k-2}$. In other words we are interested in

$$N\mathbb{V}\mathrm{ar}(\widetilde{C}_{k+q}) = \frac{1}{N}\sum_{i=1}^{N-k-q}\sum_{j=1}^{N-k-q}\mathbb{E}[\mathrm{vec}(\mathbf{X}_{i+k+q}\mathbf{X}_i^T)\mathrm{vec}(\mathbf{X}_{j+k+q}\mathbf{X}_j^T)^T]$$

for all $q \in \{0, \ldots, k-2\}$. Here Figure 6.2 demonstrates the existence of an index interval where the corresponding random vectors $\mathbf{X}_{(\cdot)}$ can depend both on $\mathbf{X}_i$ and $\mathbf{X}_{i+k+q}$.

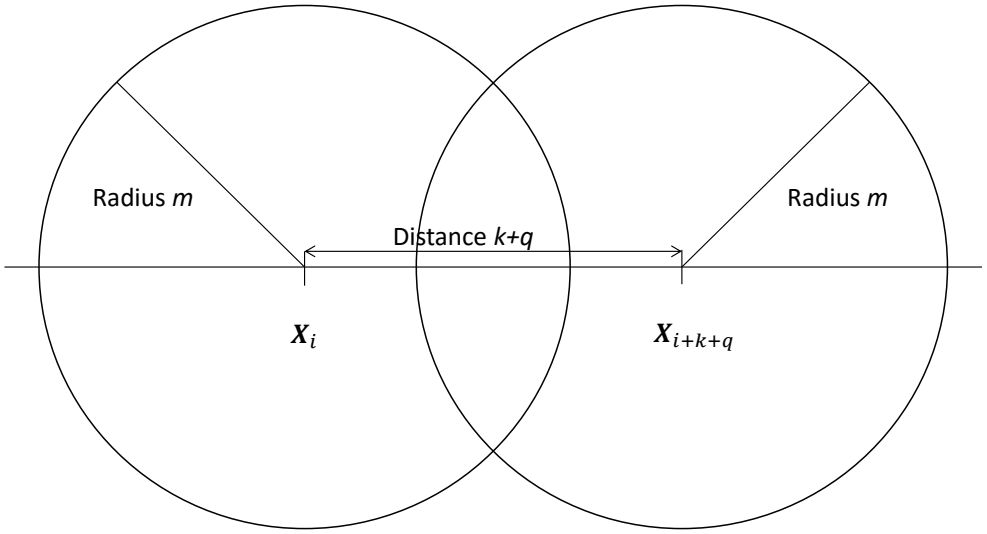


Figure 6.2: Univariate illustration of the time series with $\mathbf{X}_i$ and $\mathbf{X}_{i+k+q}$.

As in the sketch of proof of Lemma 6.20 we consider the expressions of the form

$$\mathbb{E}[\mathrm{vec}(\mathbf{X}_{i+k+q}\mathbf{X}_i^T)\mathrm{vec}(\mathbf{X}_{j+k+q}\mathbf{X}_j^T)^T],$$

fix any $i \in \{1,\ldots,N-k-q\}$ and check $j \in \{1,\ldots,N-k-q\}$. There are three cases, where this expectation does not equal zero: if $\mathbf{X}_{j+k+q}$ is in that intersection area (so that $\mathbf{X}_j$ is located within the left circle), if $\mathbf{X}_j$ is in that intersection area (so that $\mathbf{X}_{j+k+q}$ is located within the right circle) and if $\mathbf{X}_j$ and $\mathbf{X}_{j+k+q}$ are within the union of the two circles excluding their intersection. This results in three double sums

$$\begin{aligned} N\mathbb{V}\mathrm{ar}(\widetilde{C}_{k+q}) &= \frac{1}{N}\sum_{i=k}^{N-k-q}\mathbb{E}\Big[\sum_{j=i-k+1}^{i-q-1}(\mathbf{X}_{i+k+q}\mathbf{X}_{j+k+q}^T)\otimes(\mathbf{X}_i\mathbf{X}_j^T)^T\Big] \\ &\quad+\frac{1}{N}\sum_{i=1}^{N-2k-q+1}\mathbb{E}\Big[\sum_{j=i+q+1}^{i+k-1}(\mathbf{X}_{i+k+q}\mathbf{X}_{j+k+q}^T)\otimes(\mathbf{X}_i\mathbf{X}_j^T)^T\Big] \\ &\quad+\frac{1}{N}\sum_{i=1}^{N-k-q}\sum_{j=i-q}^{i+q}C_{i-j}\otimes C_{i-j}. \end{aligned}$$

Taking the limit on both sides makes the outer sum of the fourth moments vanish such that

$$\begin{aligned}\lim_{N\to\infty} N\mathbb{V}\mathrm{ar}(\widetilde{C}_{k+q}) &= \mathbb{E}\Big[\sum_{j=i_1-k+1}^{i_1-q-1}(\mathbf{X}_{i_1+k+q}\mathbf{X}_{j+k+q}^T)\otimes(\mathbf{X}_{i_1}\mathbf{X}_j^T)^T\Big] \\ &\quad +\mathbb{E}\Big[\sum_{j=i_2+q+1}^{i_2+k-1}(\mathbf{X}_{i_2+k+q}\mathbf{X}_{j+k+q}^T)\otimes(\mathbf{X}_{i_2}\mathbf{X}_j^T)^T\Big]+\sum_{\nu=-q}^{q}C_\nu\otimes C_\nu\end{aligned}$$

holds for all $i_1 \in \{k,\dots,N-k-q\}$ and for all $i_2 \in \{1,\dots,N-2k-q+1\}$. Note that this expression is a generalization of the special cases Lemma 6.20 for $q > k-2$ and $\Psi_{1,1}$ in Lemma 6.2, where that fourth moment expression with the sum over $a < b$ is equivalent to the upper expression. Note that $C_{-\nu} = C_\nu^T$ holds as well for their empirical counterparts $\widetilde{C}_{-\nu} = \widetilde{C}_\nu^T$. This asymptotical covariance matrix $\lim_{N\to\infty} N\mathbb{V}\mathrm{ar}(\widetilde{C}_{k+q})$ can be estimated by means of replacing the expectation operators by empirical averages

$$\begin{aligned}\widetilde{\Psi}_{1,1}^{(q+1)} &= \frac{1}{N}\sum_{i=k}^{N-k-q}\sum_{j=i-k+1}^{i-q-1}(\mathbf{X}_{i+k+q}\mathbf{X}_{j+k+q}^T)\otimes(\mathbf{X}_i\mathbf{X}_j^T)^T \\ &\quad +\frac{1}{N}\sum_{i=1}^{N-2k-q+1}\sum_{j=i+q+1}^{i+k-1}(\mathbf{X}_{i+k+q}\mathbf{X}_{j+k+q}^T)\otimes(\mathbf{X}_i\mathbf{X}_j^T)^T+\sum_{\nu=-q}^{q}\widetilde{C}_\nu\otimes\widetilde{C}_\nu. \quad (6.18)\end{aligned}$$

Eventually we are ready to establish an improved version of the hypothesis test in Theorem 6.4. The random vectors in (6.16) multiplied by the $Hp^2 \times Hp^2$-matrix

$$\begin{pmatrix}\widetilde{C}_0^{-1}\otimes Id & & 0\\ & \ddots & \\ 0 & & \widetilde{C}_0^{-1}\otimes Id\end{pmatrix}$$

imply $k$ test statistics $\widetilde{Q}_N^{(1)},\dots,\widetilde{Q}_N^{(k)}$ (in the same way as Theorem 6.4 describes), which altogether are asymptotically $\chi^2_{Hp^2}$ distributed under strong stationarity and $m$-dependence. Consequently if at least one of those test statistics is not asymptotically $\chi^2_{Hp^2}$ distributed, $m$-dependence cannot hold. Now consider the overall significance level. As usual we wish the type I error not to be larger than $\alpha \in (0;1)$. However, if we set the significance level of each subtest to be $\frac{\alpha}{k}$, the calculation

$$\begin{aligned}\text{type I error} &= \mathbb{P}(\text{Reject } H_0|H_0) = \mathbb{P}(\exists\tau\in\{1,\dots,k\}:\ \widetilde{Q}_N^{(\tau)} > \chi^2_{Hp^2,1-\frac{\alpha}{k}}|H_0) \\ &= \mathbb{P}(\bigcup_{\tau=1}^{k}\widetilde{Q}_N^{(\tau)} > \chi^2_{Hp^2,1-\frac{\alpha}{k}}|H_0) \le \sum_{\tau=1}^{k}\mathbb{P}(\widetilde{Q}_N^{(\tau)} > \chi^2_{Hp^2,1-\frac{\alpha}{k}}|H_0) \approx \sum_{\tau=1}^{k}\frac{\alpha}{k} = \alpha\end{aligned}$$

reveals that the overall significance level becomes $\alpha$. This consideration is called *Bonferroni criterion.*

**Theorem 6.21 (Improved $m$-Dependence Hypothesis Test)** *Suppose strong stationarity and*

$$H_0:\ \text{Data } m-\text{dependent}; \qquad H_1:\ \text{Data at least } (m+1)-\text{dependent}.$$

*Then under $H_0$ the test statistics $\widetilde{Q}_N^{(1)}, \ldots, \widetilde{Q}_N^{(k)}$, derived from (6.16) in the same manner as in Theorem 6.4, are asymptotically $\chi^2_{Hp^2}$ distributed. Therefore reject $H_0$ at significance level $\alpha \in (0;1)$ if and only if there exists a $\tau \in \{1, \ldots, k\}$ such that $\widetilde{Q}_N^{(\tau)} > \chi^2_{Hp^2, 1-\frac{\alpha}{k}}$.*

Sketch of proof: See above.

## 6.3 Simulation results

This section deals with examples of test applications. The principle is to simulate a functional ARMA process, discretized onto $n \in \{100, 1000, 10000\}$ knots and based on a Brownian bridge as a white noise, 1000 times and then to apply the hypothesis tests while varying $H \in \{3, 5\}$, $p \in \{3, 4, 5\}$ and $\alpha \in \{0.01, 0.05, 0.1\}$ according to [GaKo]. For all 1000 simulations we count how often the test statistic exceed the quantile $\chi^2_{Hp^2, 1-\alpha}$ or respectively $\chi^2_{Hp^2, 1-\frac{\alpha}{k}}$ and divide that by 1000 to get an average. Those numbers are displayed in tables showing the *empirical size* or the *empirical power*.

**Definition 6.22 (Size and Power)** *Suppose an hypothesis test wit null hypothesis $H_0$ and alternative hypothesis $H_1 = \neg H_0$. Then*

- *the* size *of the test is defined as* $\mathbb{P}(\text{Reject } H_0 | H_0)$ *(type I error) and*
- *the* power *of the test is defined as* $\mathbb{P}(\text{Reject } H_0 | H_1)$ *(1− type II error).*

*Both take the rejection of $H_0$ into account.*

At first we will analyze the results only of the test in Theorem 6.4. We will test 1-dependence for an $ARH(1)$ process and for a strong white noise as well as 3-dependence for an $MAH(m)$ process with $m \in \{1, \ldots, 5\}$. Afterwards we will compare this test with its improvement in Theorem 6.21.

### 6.3.1 Results for the test in Theorem 6.4

Before discussing the simulation results here is one remark about how to estimate the test statistic $\widetilde{Q}_N$: All the blocks and subvectors of $\widetilde{R}$ and $\widetilde{\Xi}$ can be estimated straightforwardly – except the fourth moments in the first block of $\widetilde{\Xi}$. There is no routine that calculates multivariate fourth moments in contrast to second moments, where the `R`-function `cov` returns sample (cross-)covariance matrices. In Theorem 6.4, one fourth moment estimator is suggested before left- and right-multiplication by $\widetilde{C}_0^{-1} \otimes Id$ . For ease of coding in `R`, we transform this formula to the representation

$$\begin{aligned}\widetilde{\Psi}_{1,1} &= \frac{1}{N}\sum_{i=k}^{N-k}\sum_{j=i-k+1}^{i-1}(\mathbf{X}_{i+k}\mathbf{X}_{j+k}^T)\otimes(\mathbf{X}_i\mathbf{X}_j^T)^T \\ &\quad + \frac{1}{N}\sum_{i=1}^{N-2k+1}\sum_{j=i+1}^{i+k-1}(\mathbf{X}_{i+k}\mathbf{X}_{j+k}^T)\otimes(\mathbf{X}_i\mathbf{X}_j^T)^T + \widetilde{C}_0\otimes\widetilde{C}_0\end{aligned}$$

from (6.18) with $q = 0$. Both formulas are equivalent, but the latter does not use sample-splitting for the implementation. (In fact, sample-splitting only gave an idea of how to derive the test.) Consequently, the top left block of $\Xi$ can be estimated by

$$\widetilde{\Xi}_{1,1} = \big(\widetilde{C}_0^{-1} \otimes Id_p\big)\widetilde{\Psi}_{1,1}\big(\widetilde{C}_0^{-1} \otimes Id_p\big).$$

Now we go into testing 1-dependence. We approximate the size of the test by a strong white noise (which is 0-dependent and thus 1-dependent) where each observation is a Brownian bridge and where the observations are independent. We use B-splines for interpolation here and in the remainder. For the strong white noise the values converge to $\alpha$ for $N \to \infty$, see Table 6.1. This makes sense because they are empirical estimates of the test size and by the strong law of large numbers (which holds because the replicates are i.i.d.) they are converge to their expectation, the theoretical test size. In other words let

$$\widetilde{Q}_{N,1}, \widetilde{Q}_{N,2}, \ldots, \widetilde{Q}_{N,S}$$

be the $S \in \mathbb{R}$ (here $S = 1000$) replicates of $\widetilde{Q}_N$ (and thus identically distributed). Since they arise from $S$ different simulations, they are independent. Hence the strong law of large numbers holds

$$\frac{1}{S}\sum_{s=1}^{S} \mathbb{1}_{\widetilde{Q}_{N,s} > \chi^2_{Hp^2,1-\alpha}} \overset{S\to\infty}{\longrightarrow} \mathbb{E}\Big[\mathbb{1}_{\widetilde{Q}_{N,1} > \chi^2_{Hp^2,1-\alpha}}\Big] = \mathbb{P}(\widetilde{Q}_N > \chi^2_{Hp^2,1-\alpha}) \overset{N\to\infty}{\longrightarrow} \alpha \qquad \text{a.e.}$$

The number of PCs $p$ does not seem to cause any tendency that the results get better or worse. The same holds for $H$. Since the test is based on the central limit theorem, the "results are likely to be worse for non-normal data" according to [GaKo, Section 3], which is not analyzed in this thesis either.

| $WNH$ | $p=3$ | | | $p=4$ | | | $p=5$ | | |
|---|---|---|---|---|---|---|---|---|---|
| $\alpha$ : | 0.1 | 0.05 | 0.01 | 0.1 | 0.05 | 0.01 | 0.1 | 0.05 | 0.01 |
| $N = 100$ | | | | | | | | | |
| $H = 3$ | 0.078 | 0.034 | 0.009 | 0.136 | 0.093 | 0.052 | 0.113 | 0.086 | 0.054 |
| $H = 5$ | 0.051 | 0.030 | 0.006 | 0.104 | 0.071 | 0.036 | 0.081 | 0.057 | 0.040 |
| $N = 1000$ | | | | | | | | | |
| $H = 3$ | 0.084 | 0.040 | 0.008 | 0.095 | 0.045 | 0.005 | 0.079 | 0.039 | 0.012 |
| $H = 5$ | 0.086 | 0.039 | 0.008 | 0.094 | 0.048 | 0.009 | 0.082 | 0.039 | 0.009 |
| $N = 10000$ | | | | | | | | | |
| $H = 3$ | 0.106 | 0.050 | 0.006 | 0.109 | 0.054 | 0.012 | 0.111 | 0.054 | 0.014 |
| $H = 5$ | 0.118 | 0.066 | 0.014 | 0.109 | 0.057 | 0.012 | 0.110 | 0.052 | 0.012 |

Table 6.1: Average rejections for $H_0$: "1-dependence" over 1000 replications of a strong white noise with the test in Theorem 6.4.

The power of the test is approximated by the $ARH(1)$ process $(Y_n)_{n\in\mathbb{Z}}$ such that

$$Y_n = lY_{n-1} + \varepsilon_n, \ \forall n \in \mathbb{Z}.$$

Its integral operator $l \in \mathcal{L}$ is uniquely defined by the integral kernel

$$0.5 \exp\left(-\frac{s^2+t^2}{2}\right), \forall s,t \in [0,1]. \tag{6.19}$$

The results can be found in Table 6.2. In contrast to the strong white noise for the $ARH(1)$ process the empirical type II error becomes zero rapidly for large $N$. For increasing $H$ the empirical power decreases, because the quantiles, depending on $Hp^2$, grow faster than the test statistics according to [GaKo, Section 3] and because we take more estimates into account, which makes the test statistics more imprecise. We can see that this test is not very good for low $N$, which is reasonable too, because we estimate large matrices containing fourth moments, which requires a large number of observations.

| $ARH(1)$ | $p=3$ | | | $p=4$ | | | $p=5$ | | |
|---|---|---|---|---|---|---|---|---|---|
| $\alpha$ : | 0.1 | 0.05 | 0.01 | 0.1 | 0.05 | 0.01 | 0.1 | 0.05 | 0.01 |
| $N=100$ | | | | | | | | | |
| $H=3$ | 0.105 | 0.052 | 0.014 | 0.140 | 0.094 | 0.062 | 0.140 | 0.101 | 0.077 |
| $H=5$ | 0.070 | 0.037 | 0.007 | 0.102 | 0.074 | 0.043 | 0.108 | 0.088 | 0.056 |
| $N=1000$ | | | | | | | | | |
| $H=3$ | 0.784 | 0.659 | 0.382 | 0.625 | 0.491 | 0.241 | 0.657 | 0.521 | 0.244 |
| $H=5$ | 0.650 | 0.506 | 0.243 | 0.500 | 0.379 | 0.149 | 0.517 | 0.365 | 0.159 |
| $N=10000$ | | | | | | | | | |
| $H=3$ | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| $H=5$ | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |

Table 6.2: Average rejections for $H_0$: "1-dependence" over 1000 replications of an $ARH(1)$ process with the test in Theorem 6.4.

For 3-dependence we consider $MAH(m)$ processes $(Y_n)_{n\in\mathbb{Z}}$ such that

$$Y_n = \varepsilon_n + l\sum_{\nu=1}^{m} \varepsilon_{n-\nu}, \ \forall n \in \mathbb{Z},$$

with $m \in \{1,\dots,5\}$ and with $l$ as above. Consequently the test results for $m \leq 3$ in Table 6.3 up to Table 6.5 display the empirical size, whereas the test results for $m > 3$ in Table 6.6 and Table 6.7 display the empirical power. Here we can conclude the same as in the 1-dependence example, concerning behavior of the empirical sizes and the empirical power for $N$ increasing. The values for $N = 1000$ are even better than in the previous example, because the moving average lag structure is clearer for a lower number of observations than the autoregressive lag structure. As it can be seen in Table 6.2 as well, there are less rejections for $H = 5$ than for $H = 3$ in the empirical power tables and the power estimates for low $N$ are low.

| $MAH(1)$ | $p=3$ | | | $p=4$ | | | $p=5$ | | |
|---|---|---|---|---|---|---|---|---|---|
| $\alpha:$ | 0.1 | 0.05 | 0.01 | 0.1 | 0.05 | 0.01 | 0.1 | 0.05 | 0.01 |
| $N=100$ | | | | | | | | | |
| $H=3$ | 0.096 | 0.070 | 0.038 | 0.067 | 0.054 | 0.042 | 0.043 | 0.037 | 0.031 |
| $H=5$ | 0.056 | 0.039 | 0.024 | 0.049 | 0.042 | 0.028 | 0.031 | 0.028 | 0.024 |
| $N=1000$ | | | | | | | | | |
| $H=3$ | 0.095 | 0.043 | 0.005 | 0.086 | 0.039 | 0.004 | 0.067 | 0.029 | 0.006 |
| $H=5$ | 0.095 | 0.042 | 0.012 | 0.081 | 0.034 | 0.011 | 0.075 | 0.034 | 0.006 |
| $N=10000$ | | | | | | | | | |
| $H=3$ | 0.085 | 0.047 | 0.009 | 0.088 | 0.040 | 0.005 | 0.082 | 0.048 | 0.009 |
| $H=5$ | 0.095 | 0.047 | 0.007 | 0.081 | 0.048 | 0.008 | 0.083 | 0.047 | 0.010 |

Table 6.3: Average rejections for $H_0$: "3-dependence" over 1000 replications of an $MAH(1)$ process with the test in Theorem 6.4.

| $MAH(2)$ | $p=3$ | | | $p=4$ | | | $p=5$ | | |
|---|---|---|---|---|---|---|---|---|---|
| $\alpha:$ | 0.1 | 0.05 | 0.01 | 0.1 | 0.05 | 0.01 | 0.1 | 0.05 | 0.01 |
| $N=100$ | | | | | | | | | |
| $H=3$ | 0.091 | 0.068 | 0.052 | 0.069 | 0.057 | 0.043 | 0.044 | 0.034 | 0.026 |
| $H=5$ | 0.068 | 0.056 | 0.040 | 0.051 | 0.043 | 0.033 | 0.026 | 0.021 | 0.016 |
| $N=1000$ | | | | | | | | | |
| $H=3$ | 0.081 | 0.048 | 0.010 | 0.078 | 0.044 | 0.013 | 0.083 | 0.046 | 0.010 |
| $H=5$ | 0.088 | 0.047 | 0.009 | 0.081 | 0.042 | 0.009 | 0.071 | 0.030 | 0.007 |
| $N=10000$ | | | | | | | | | |
| $H=3$ | 0.104 | 0.056 | 0.012 | 0.106 | 0.047 | 0.009 | 0.096 | 0.046 | 0.007 |
| $H=5$ | 0.087 | 0.040 | 0.009 | 0.088 | 0.047 | 0.011 | 0.097 | 0.047 | 0.005 |

Table 6.4: Average rejections for $H_0$: "3-dependence" over 1000 replications of an $MAH(2)$ process with the test in Theorem 6.4.

| $MAH(3)$ | $p=3$ | | | $p=4$ | | | $p=5$ | | |
|---|---|---|---|---|---|---|---|---|---|
| $\alpha:$ | 0.1 | 0.05 | 0.01 | 0.1 | 0.05 | 0.01 | 0.1 | 0.05 | 0.01 |
| $N=100$ | | | | | | | | | |
| $H=3$ | 0.100 | 0.080 | 0.040 | 0.077 | 0.065 | 0.048 | 0.063 | 0.049 | 0.041 |
| $H=5$ | 0.075 | 0.056 | 0.030 | 0.059 | 0.048 | 0.039 | 0.057 | 0.052 | 0.037 |
| $N=1000$ | | | | | | | | | |
| $H=3$ | 0.106 | 0.053 | 0.013 | 0.101 | 0.055 | 0.008 | 0.098 | 0.045 | 0.013 |
| $H=5$ | 0.101 | 0.048 | 0.009 | 0.099 | 0.051 | 0.016 | 0.090 | 0.048 | 0.013 |
| $N=10000$ | | | | | | | | | |
| $H=3$ | 0.102 | 0.051 | 0.008 | 0.103 | 0.050 | 0.009 | 0.098 | 0.049 | 0.011 |
| $H=5$ | 0.100 | 0.048 | 0.009 | 0.091 | 0.038 | 0.008 | 0.088 | 0.045 | 0.011 |

Table 6.5: Average rejections for $H_0$: "3-dependence" over 1000 replications of an $MAH(3)$ process with the test in Theorem 6.4.

| $MAH(4)$ | $p=3$ | | | $p=4$ | | | $p=5$ | | |
|---|---|---|---|---|---|---|---|---|---|
| $\alpha:$ | 0.1 | 0.05 | 0.01 | 0.1 | 0.05 | 0.01 | 0.1 | 0.05 | 0.01 |
| $N=100$ | | | | | | | | | |
| $H=3$ | 0.133 | 0.102 | 0.059 | 0.082 | 0.064 | 0.048 | 0.043 | 0.031 | 0.022 |
| $H=5$ | 0.090 | 0.066 | 0.040 | 0.060 | 0.044 | 0.034 | 0.044 | 0.034 | 0.026 |
| $N=1000$ | | | | | | | | | |
| $H=3$ | 1.000 | 0.999 | 0.986 | 0.996 | 0.979 | 0.879 | 0.989 | 0.973 | 0.873 |
| $H=5$ | 0.995 | 0.977 | 0.875 | 0.949 | 0.874 | 0.621 | 0.925 | 0.849 | 0.586 |
| $N=10000$ | | | | | | | | | |
| $H=3$ | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| $H=5$ | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |

Table 6.6: Average rejections for $H_0$: "3-dependence" over 1000 replications of an $MAH(4)$ process with the test in Theorem 6.4.

| $MAH(5)$ | $p=3$ | | | $p=4$ | | | $p=5$ | | |
|---|---|---|---|---|---|---|---|---|---|
| $\alpha:$ | 0.1 | 0.05 | 0.01 | 0.1 | 0.05 | 0.01 | 0.1 | 0.05 | 0.01 |
| $N=100$ | | | | | | | | | |
| $H=3$ | 0.122 | 0.088 | 0.054 | 0.089 | 0.070 | 0.050 | 0.064 | 0.053 | 0.041 |
| $H=5$ | 0.089 | 0.065 | 0.043 | 0.072 | 0.056 | 0.047 | 0.055 | 0.048 | 0.039 |
| $N=1000$ | | | | | | | | | |
| $H=3$ | 1.000 | 0.998 | 0.986 | 0.993 | 0.974 | 0.857 | 0.972 | 0.926 | 0.704 |
| $H=5$ | 0.994 | 0.982 | 0.874 | 0.926 | 0.856 | 0.582 | 0.845 | 0.720 | 0.426 |
| $N=10000$ | | | | | | | | | |
| $H=3$ | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| $H=5$ | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |

Table 6.7: Average rejections for $H_0$: "3-dependence" over 1000 replications of an $MAH(5)$ process with the test in Theorem 6.4.

### 6.3.2 Comparison between Theorem 6.4 and Theorem 6.21

Given $l \in \mathcal{L}$ as in (6.19) we here consider an $MAH(3)$ process $(Y_n)_{n\in\mathbb{Z}}$ such that

$$Y_n = \varepsilon_n + l\varepsilon_{n-3}, \ \forall n \in \mathbb{Z}, \tag{6.20}$$

holds, as already mentioned in Section 6.2. When testing 1-dependence, we know that $H_1$ holds, because $(Y_n)_{n\in\mathbb{Z}}$ suffices $C_3 \neq 0$. Nevertheless according to Table 6.8 the test in Theorem 6.4 treats $(Y_n)_{n\in\mathbb{Z}}$ as if it was 1-dependent, because $C_h \equiv 0$ holds for all $h \in \mathbb{N}\backslash\{3\}$. Hence the empirical power converges to the significance level. By contrast, the test in Theorem 6.21 takes $C_3$ into account, so that $H_0$ is mostly rejected. The power of the latter test grows tremendously fast, which can be seen in Table 6.9.

| First Test | $p=3$ | | | $p=4$ | | | $p=5$ | | |
|---|---|---|---|---|---|---|---|---|---|
| $\alpha$ : | 0.1 | 0.05 | 0.01 | 0.1 | 0.05 | 0.01 | 0.1 | 0.05 | 0.01 |
| $N=100$ | | | | | | | | | |
| $H=3$ | 0.093 | 0.044 | 0.013 | 0.162 | 0.118 | 0.063 | 0.118 | 0.092 | 0.065 |
| $H=5$ | 0.069 | 0.031 | 0.012 | 0.121 | 0.081 | 0.045 | 0.097 | 0.081 | 0.056 |
| $N=1000$ | | | | | | | | | |
| $H=3$ | 0.113 | 0.058 | 0.011 | 0.098 | 0.051 | 0.009 | 0.095 | 0.048 | 0.006 |
| $H=5$ | 0.121 | 0.059 | 0.011 | 0.097 | 0.052 | 0.009 | 0.088 | 0.041 | 0.007 |
| $N=10000$ | | | | | | | | | |
| $H=3$ | 0.118 | 0.059 | 0.013 | 0.116 | 0.058 | 0.011 | 0.120 | 0.055 | 0.010 |
| $H=5$ | 0.121 | 0.065 | 0.016 | 0.118 | 0.063 | 0.013 | 0.123 | 0.068 | 0.014 |

Table 6.8: Average rejections for $H_0$: "1-dependence" over 1000 replications of the $MAH(3)$ process in (6.20) with the test in Theorem 6.4.

| Second Test | $p=3$ | | | $p=4$ | | | $p=5$ | | |
|---|---|---|---|---|---|---|---|---|---|
| $\alpha$ : | 0.1 | 0.05 | 0.01 | 0.1 | 0.05 | 0.01 | 0.1 | 0.05 | 0.01 |
| $N=100$ | | | | | | | | | |
| $H=3$ | 0.411 | 0.274 | 0.113 | 0.343 | 0.234 | 0.097 | 0.314 | 0.225 | 0.105 |
| $H=5$ | 0.263 | 0.174 | 0.065 | 0.201 | 0.133 | 0.060 | 0.185 | 0.112 | 0.061 |
| $N=1000$ | | | | | | | | | |
| $H=3$ | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| $H=5$ | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| $N=10000$ | | | | | | | | | |
| $H=3$ | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| $H=5$ | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |

Table 6.9: Average rejections for $H_0$: "1-dependence" over 1000 replications of the $MAH(3)$ process in (6.20) with the test in Theorem 6.21.

# Chapter 7

# Real Data Study

The theory and methods described in the previous chapters is going to be applied here. We will work with highway traffic data. Having explained the properties of the datasets, we will deal with clustering methods in order to gain some kind of homogeneity. Finally we will use the moving average hypothesis test and compute some forecasts.

## 7.1 Data Description

We are provided some traffic volume data and some traffic speed data by Autobahndirektion Südbayern. Both concern one specific location of Autobahn A92 and contain measurements per minute from January 1st, 2014 to June 30ths, 2014. The number of cars can be found in the volume dataset and the average speed in the speed dataset. Both datasets are preprocessed in the same way as in [Wei], i.e. merging from three lanes to one lane, constructing weighted speed data and aggregating the latter from one second to five seconds. We aggregate the volume data too by summing up for each five seconds interval. As well as in [Wei] we exclude the days January 20th, March 3rd and May 7th due to lack of information. Thus the number of observations equals $N = 178$. Afterwards we convert the two resulting datasets into functional data via the package `fda`. The observations were interpolated by linear combinations of some fourier basis functions mapping from $[0, 1]$ as recommended in [Wei]. Since the dates are known, the weekdays and holidays can be determined for each observation. We expect some sort of homogeneity for each weekday. For example some characteristics for all Mondays might be assumed which differentiate from the traffic behavior on Sundays. Therefore we classify the observations into their weekdays. However, we set Tuesday, Wednesday and Friday as one group and all non-weekend holidays as another one. Figure 7.1 shows the traffic volume and Figure 7.2 the traffic speed, both colored according to the weekdays. Note that the functional observations describe the behavior from 12 am to 12 am. The x-axis denotes the interval $[0, 1]$ because of the choice of the functional basis. In contrast to what we imagined, the groups by weekdays (and holidays) do not seem to be homogeneous. It is especially the speed dataset which has much variety in the weekdays.

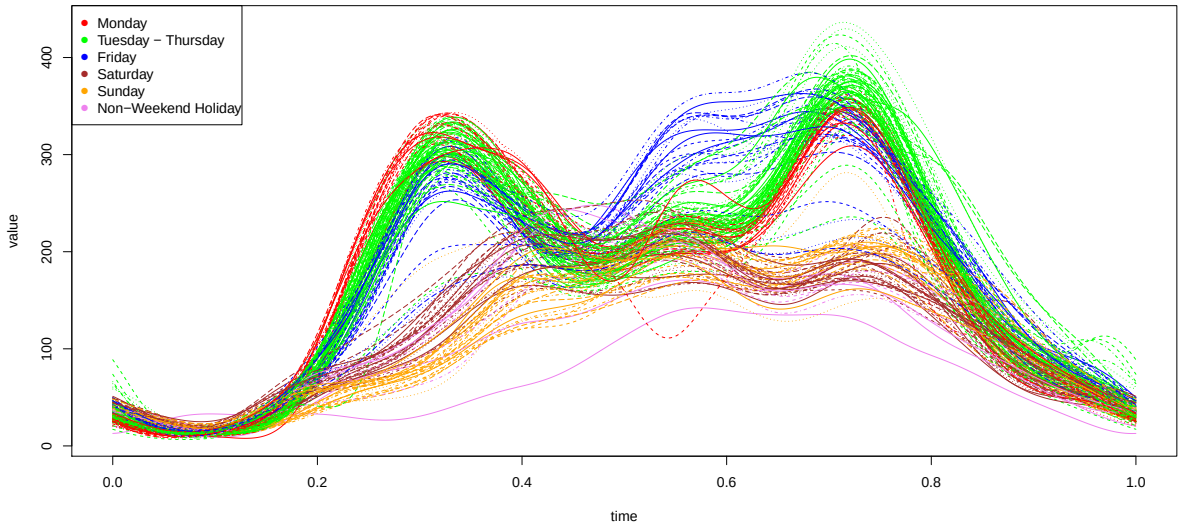


Figure 7.1: Functional volume data, colored by weekdays.

The curves make sense because of the rush hours (traffic jam, thus there are many cars and most of them have to drive slowly) in the morning and especially in the evening on working days, where people tend to go home earlier on Fridays. Furthermore less cars occur on non-working days, which drive faster than on working days.

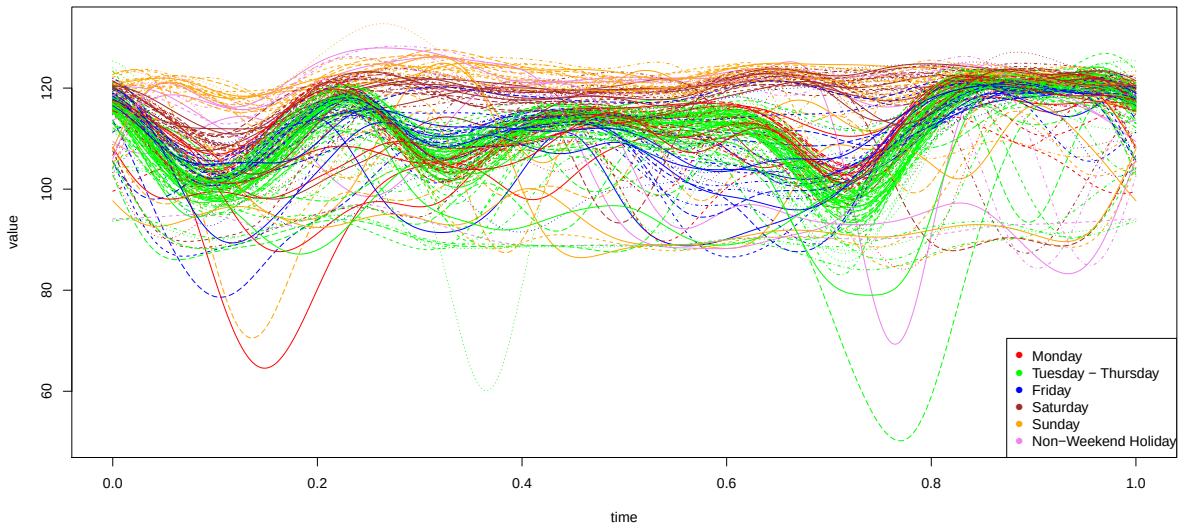


Figure 7.2: Functional speed data, colored by weekdays.

The plots also show that some kind of grouping the observations is recommended for predictions. The overall variety is too high to assume an $MAH(1)$ model. If the dates and consequently the weekdays had not been available, we would have to cluster the data. In this case the goal of clustering is to obtain an appropriate group of observations which can be described by an $MAH(1)$ process fairly and which hence yields good predictions. However, it turns out that this does not work here, because the clustering methods will look for similar curves, which might be from different weekdays. Consequently the day intervals from observation to observation within each cluster might differ tremendously. Moreover, since the curves within one cluster are too similar, they are highly correlated – even with respect to higher order lags, so that an $MAH(1)$ model is not recommended. Nevertheless we will discuss some clustering algorithms and analyze their results.

## 7.2 Data Clustering

Clustering means grouping the data into $k$ clusters, where $k \in \mathbb{N}$ is chosen arbitrarily before. We work only with the speed dataset. The same can be applied on the volume dataset as well. In this section we visualize the functional observations as points by projecting onto the first three functional principal components and by plotting the scores. Figure 7.3 shows the functional observations of the speed dataset as points colored by their weekdays. It demonstrates that the data can be clustered more appropriately.

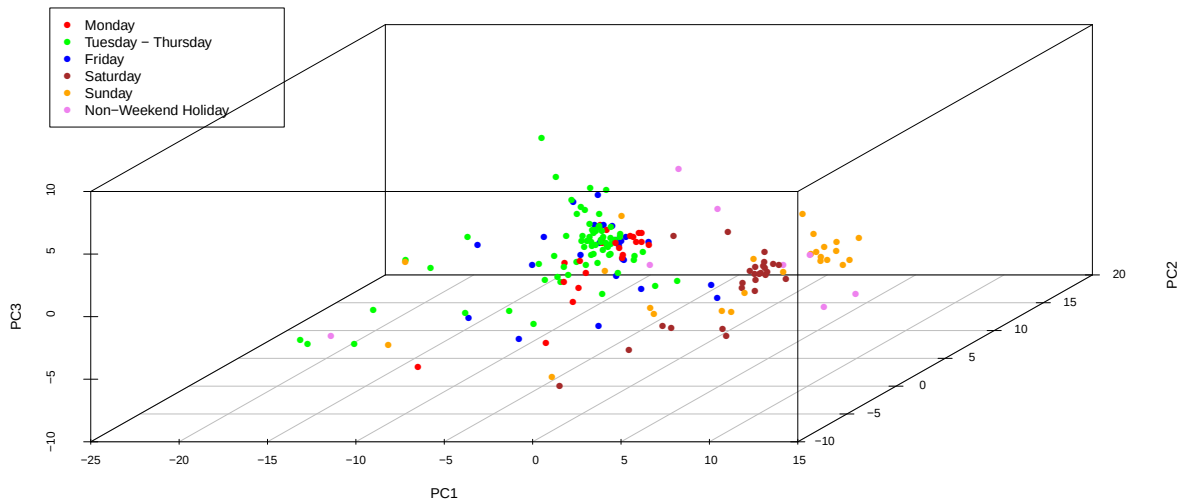


Figure 7.3: Speed data classified by weekdays.

There are many kinds of cluster algorithms. We will focus on $k$-means algorithm and complete linkage. Both methods require the number of clusters $k \in \mathbb{N}$ to be chosen before performing. We take $k = 6$. They are applied onto the PC scores of the speed data with $p = 7$ as the number of PCs. Moreover recall the number of observations $N = 178$.

### 7.2.1 $k$-means Algorithm

Let $\mu_1, \ldots, \mu_k \in \mathbb{R}^p$ the initial cluster centers. The idea of $k$-means is

(i) to put each observations $x_i$, $i \in \{1, \ldots, N\}$, into cluster $g \in \{1, \ldots, k\}$ according to the smallest distance to the given cluster centers

$$\min_{g \in \{1,\ldots,k\}} \|x_i - \mu_g\|$$

(classification) and

(ii) to update the cluster centers (with the number of observations per cluster $N_g$)

$$\mu_g = \frac{1}{N_g} \sum_{i:\ x_i \text{ in cluster } g} x_i$$

for all clusters $g \in \{1, \ldots, k\}$ (adjustment)

both iteratively until the observations are clustered in an appropriate way. `R` suggests the maximal number of iterations for the command `kmeans` to be ten. Here the `kmeans` routine terminates after four iterations and gives the result visualized in Figure 7.4 and in Figure 7.5.

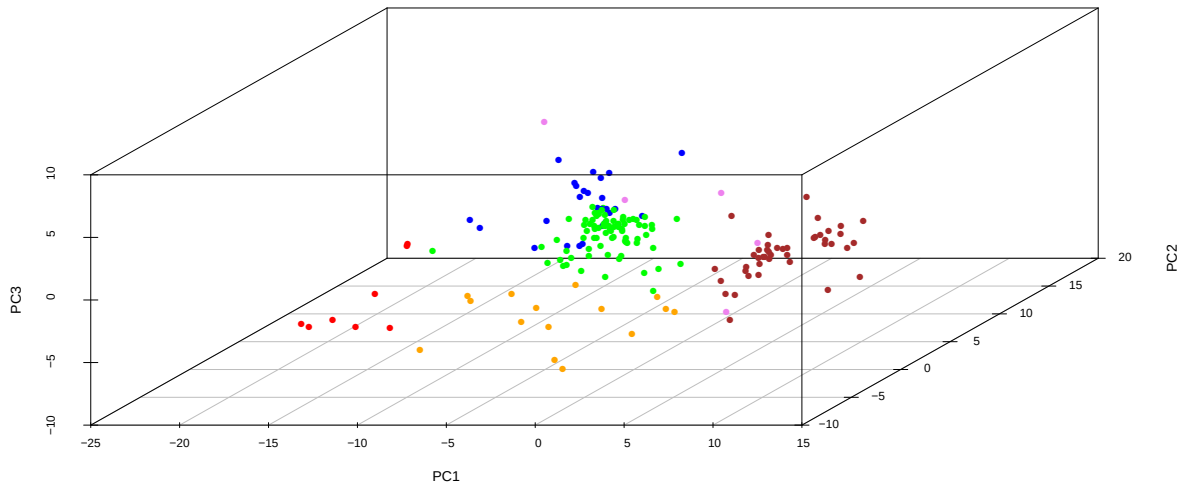


Figure 7.4: First three PC scores of speed data clustered by kmeans algorithm.

Both figures show that $k$-means separates the data better than the clustering by weekdays. Nevertheless the blue and green clusters overlap. Those observations can be separated more clearly.

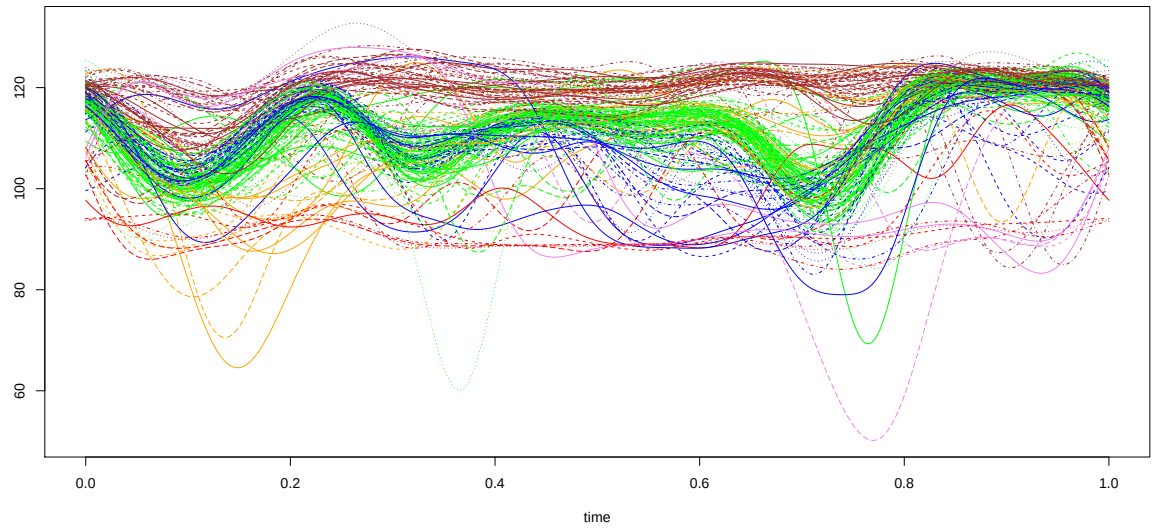


Figure 7.5: Speed data clustered by kmeans algorithm.

In general $k$-means algorithm is a good standard clustering procedure, but it does not perform well in some cases. The reason is that its concept (classifying the data and updating the centers iteratively) is quite limited for certain data, e.g. for a two-dimensional dataset consisting of one cluster as a circle and another one as a ring surrounding the circle. For those problems hierarchical clustering is often a better solution.

### 7.2.2 Complete Linkage

The basic idea of hierarchical clustering is either to consider each single observation as one cluster and to merge two clusters in each iteration step until we obtain $k$ clusters (agglomerative) or to consider the entire dataset as one cluster and to exclude one observation in each iteration step until we obtain $k$ clusters (divisive).
The *complete linkage* clustering is an agglomerative approach. Take any iteration step and any two clusters $A$ and $B$. Complete linkage measures the farthest distance between them, i.e. it compares all distances between one element of $A$ and one of $B$ and takes the farthest distance. Those cluster distances are computed for all clusters. The two clusters with the minimal cluster distance are merged to one cluster and the next iteration step starts. This procedure can be visualized in a schedule called dendrogram, see Figure 7.6. It is only the vertical and not the horizontal position of the observations that plays

a role. On the bottom all observations are one-element-clusters, whereas on the top all observations form one cluster.

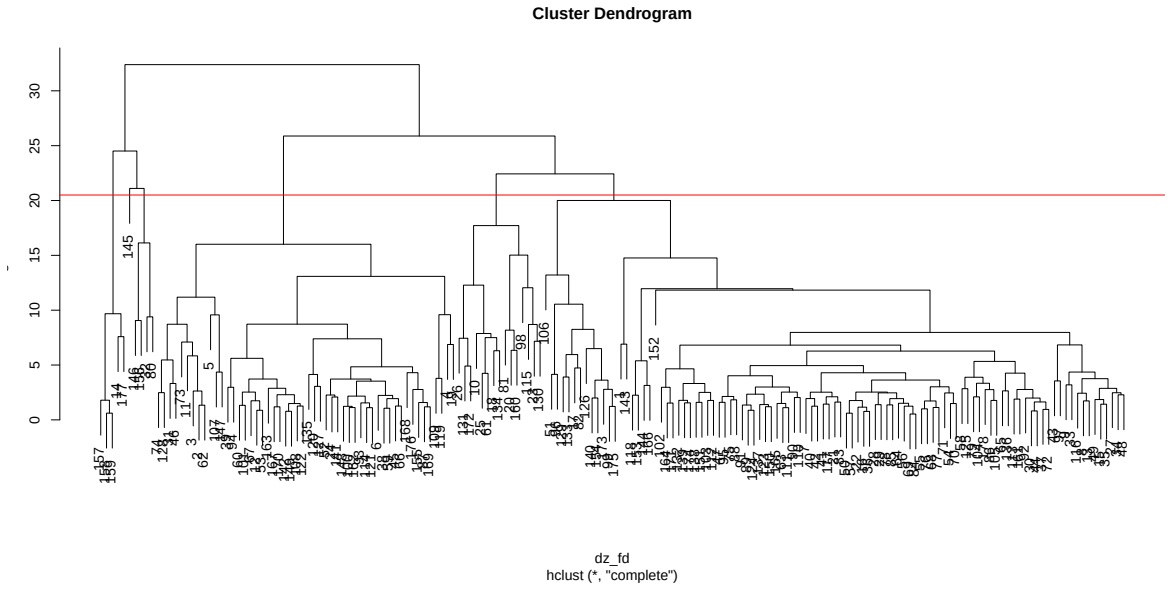


Figure 7.6: Dendogram of complete linkage on PC scores of speed data.

The R-user has to determine the clusters manually. Since we wish to get $k = 6$ groups, we choose the clusters along the red line. On the one hand the cluster sizes differ tremendously, but on the other hand the observations within each cluster are more homogeneous, which Figure 7.7 and Figure 7.8 reveal.

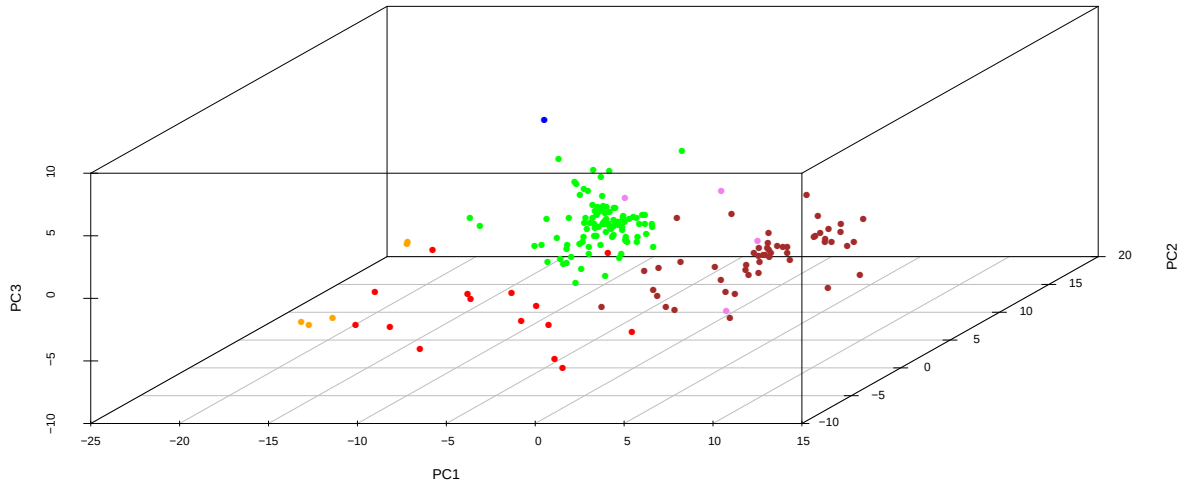


Figure 7.7: PC scores of speed data clustered by complete linkage.

The clusters resulting from complete linkage can be distinguished better than the ones resulting from $k$-means.

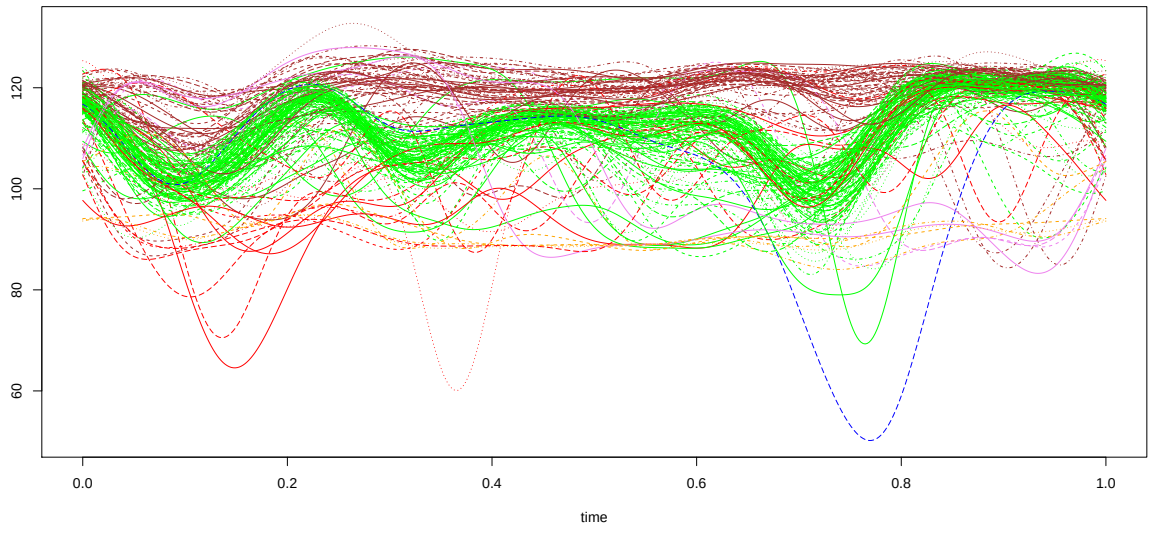


Figure 7.8: Speed data clustered by complete linkage.

Both plots show that the blue observation is an outlier. It is Wednesday, May 28th.

## 7.3 Data Prediction

We attempted to gain some suitable groups via clustering in order to apply the estimation approaches for $MAH(1)$ processes and to obtain some appropriate forecasts. Unfortunately it is only the classification by weekdays which allows us to assume the data to come from an $MAH(1)$ process, when we set all working days (119 days) as one group. Before estimating the covariance kernel function on the working days we need to center the data in order to get rather stationary mean zero data. Therefore we determine the means of the three clusters Monday, Tuesday-Thursday and Friday and subtract the working day observations by their corresponding means. For this centered dataset Figure 7.9 shows that the lags for this group suggest possessing an $MAH(1)$ structure, which does not hold for any other clustering groups.

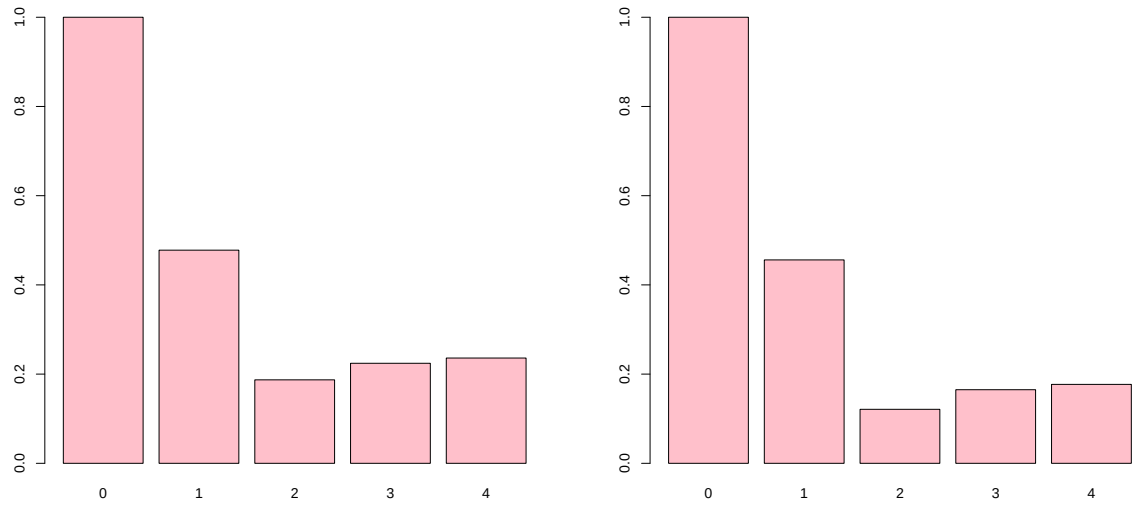


Figure 7.9: Lags of the centered speed data (left) and of the projection onto the first four PCs (right).

The $MAH(1)$ structure is confirmed by the hypothesis test in Theorem 6.4: We test 1-dependence, which means $k = 2$, by considering $H = 5$ lags. Moreover we eliminate the last observation out of the dataset, so that $N = 118$ divided by $k$ is a natural number. Furthermore it turns out that $p = 4$ is the optimal number of PC when using Fourier basis functions as stated in [Wei]. Thus we obtain

$$\widetilde{Q}_N = 86.3912 < 101.8795 = \chi^2_{5\cdot 4^2,0.95}.$$

Hence "$H_0$: data 1-dependent" cannot be rejected at significance level $\alpha = 0.05$. The same holds for the test in Theorem 6.21 (where by definition, $\widetilde{Q}_N^{(1)}$ equals $\widetilde{Q}_N$ from before)

$$\widetilde{Q}_N^{(1)} = 86.3912 < 106.6286 = \chi^2_{5\cdot 4^2,0.975},\ \widetilde{Q}_N^{(2)} = 89.6041 < 106.6286 = \chi^2_{5\cdot 4^2,0.975}.$$

(In fact the reason for neglecting the traffic volume dataset is its structure of lags, which rather suggests possessing an $ARH(1)$ structure, even if we work with clusters.)

After that we apply the iterative method and the innovation algorithm on this centered speed dataset of working days, but we compute the fitted values $\widehat{X}_{n-9}, \ldots, \widehat{X}_n$ for the last ten observations (with $\widehat{X}_i = \widehat{l}_{\text{iter/in}}\widehat{\varepsilon}_i^{\text{iter/in}}$). Therefore concerning estimation of the integral kernel we exclude one observation out of the dataset recursively until we removed ten observations. At the end, we take the mean

$$\frac{1}{10}\sum_{i=0}^{9} X_{n-i}$$

of the functional observations, of the observations projected onto the PCs (here with optimal number 4) and of the fitted values. The reason for predicting the last ten observations is to stabilize the results by taking the average. Figure 7.10 shows that the estimates are close to the mean of observations. We obtain 0.6333 as the relative forecast error for the iterative method and 0.5784 for the innovation algorithm. Hence it is again the latter which tends to perform better.

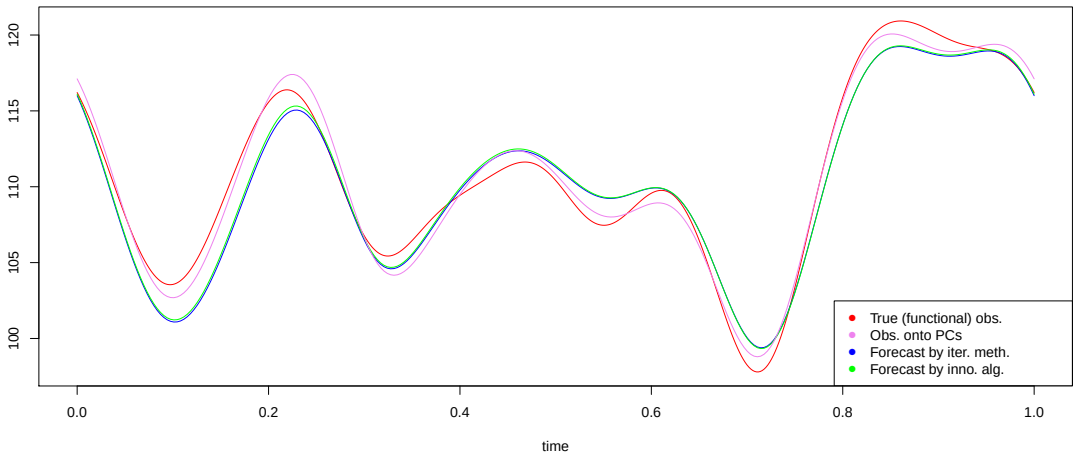


Figure 7.10: Time-averaged forecasts for speed data.

Figure 7.11 shows the means of the ten integral kernel estimates, calculated by the iterative method and by the innovation algorithm. It is hard to draw any conclusions for the traffic flow from the shape of the estimates. The integral kernels are slightly diagonal dominant.

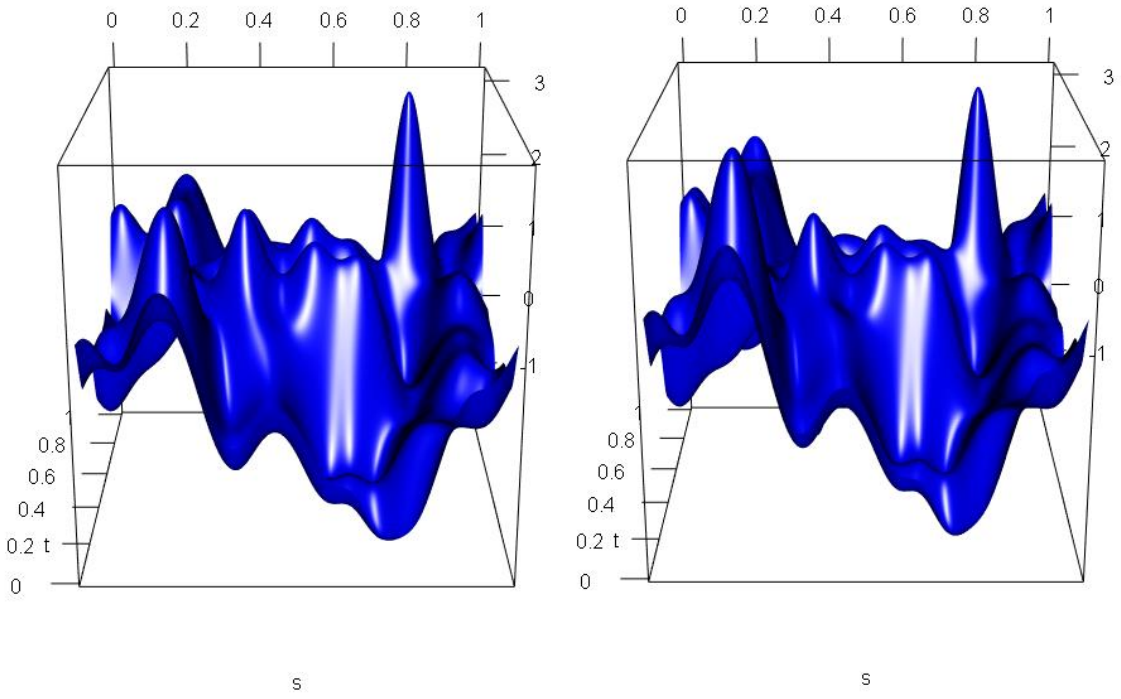


Figure 7.11: Average of ten estimated integral kernels for the speed dataset, calculated by the iterative method (left) and by the innovation algorithm (right).

Since we draw the plots of Figure 7.11 over $288 \times 288$ knots, the grids, which were used in all previous 3D-plots, would make the illustrations dark. Therefore we use the `R`-function `persp3d` from the package `rgl` instead of the `R`-function `persp` from the package `graphics` as before.

# Chapter 8

# Conclusions

This thesis consists of estimating $MAH(1)$ processes and of testing the dependence structure of a functional time series. The motivation of the tests is to choose a suitable functional time series model, provided that the dependence order is known. Having chosen 1-dependence, the estimation approaches set an appropriate model (fitted to the observations), under which we are able to forecast future observations.

Two estimation approaches for $l$ as in (3.3) have been discussed in Chapter 4. The projection method solves quadratic equations for the eigenvalues of the occurent operators. The iterative method is a fixed-point algorithm for those operators. As long as the number of observations is sufficiently large and all their conditions are satisfied, these approaches provide good estimates for $l$ as discussed in Chapter 5. However, when it comes to real data, the number of observations is mostly low and some of the theoretical assumptions do not hold. The simulation results show that the projection method performs better than the iterative method, but it requires more assumptions. When it comes to simulations, the choice for the white noise, which does not seem to be important in theory, plays a vital role for the results of the implementation. As an outlook, one could try to extend the estimation approaches to functional moving average processes of higher order. However, it might be very hard to solve the equations for the model operators.

The first hypothesis test from Theorem 6.4 was motivated by independence tests and sample-splitting. Its empirical size corresponds to the theory, but it has a low empirical power in some cases (see Subsection 6.3.2). Therefore, we improved the first test in Section 6.2 and derived a multiple test involving Bonferroni criterion. It is a compromise between considering all lags of greater order than the dependence order and summing them up to as few test statistics as possible. Concerning future work, one could compare the two hypothesis tests in more details. One could even establish a test that analyzes every lag successively and assess its goodness in comparison with the previous two tests.

For the highway traffic speed dataset, both the hypothesis test results and the averages of the predicted observations look reasonable, although the number of observations is low. By contrast, the results of the clustering routines do not correspond to the weekdays and holidays. However, this is a problem of the speed dataset and not of the clustering algorithms. One could go into details of traffic science and try to find out the reasons for heterogeneity within the weekdays. Based on this knowledge, one could develop some clustering methods detecting the weekdays and holidays from the traffic observations.

# Bibliography


[Aue] Alexander Aue, Diogo Dubart Norinho, Siegfried Hörmann, On the prediction of stationary functional time series, Journal of the American Statistical Association, 110:509, 378-392, 2015

[BB] Denis Bosq, Delphine Blanke, Inference and Prediction in Large Dimensions, Wiley, Chicheser, 2007

[Bosq] Denis Bosq, Linear Processes in Function Spaces, Theory and Applications, Springer, New York, 2000

[Brck] Peter J. Brockwell, Richard A. Davis, Time Series: Theory and Methods, Second Edition, Springer, New York, 1991

[GaKo] Robertas Gabrys, Piotr Kokoszka, Portmanteau Test of Independence for Functional Observations, Journal of the American Statistical Association, Vol. 102, pp. 1338-1348, 2007

[Hor] Lajos Horváth, Piotr Kokoszka, Inference for Functional Data with Applications, Springer, New York, 2012

[Hsing] Tailen Hsing, Randall Eubank, Theoretical Foundations of Functional Data Analysis, with an Introduction to Linear Operators, Wiley, Chichester, 2015

[Moon] Seongman Moon, Carlos Velasco, Tests for m-dependence based on sample splitting methods, Journal of Econometrics, Vol. 173, pp. 143-159, 2012

[Rams] James O. Ramsay, Bernard W. Silverman, Functional Data Analysis, Second Edition, Springer, New York, 2005

[Riesz-Nagy] Friedrich Riesz, Béla Sz.-Nagy, Vorlesungen über Funktionalanalysis, VEB Deutscher Verlag der Wissenschaften, Berlin, 1956

[Steeb] Willi-Hans Steeb, Kronecker Product of Matrices and Applications, BI-Wissenschaftsverlag, 1991

[Turb1] Céline Turbillon, Denis Bosq, Jean-Marie Marion, Besnik Pumo, Estimation du paramètre des Moyennes Mobiles hilbertiennes, ScienceDirect, Vol. 346, pp. 347-350, 2008

[Turb2] Céline Turbillon, Jean-Marie Marion, Besnik Pumo, Estimation of the moving-average operator in a Hilbert space, conference paper of "Recent Advances in Stochastic Modeling and Data Analysis", 2007

[Turb3] Denis Bosq, Céline Turbillon, Processus Moyennes Mobiles dans les espaces Hilbertiens, technical report

[TurbThese] Céline Turbillon, Estimation et prévision des processus Moyenne Mobile fonctionnels., Ph.D. Thesis, University of Paris 6, 2007

[Vaart] A.W. van der Vaart, Asymptotic Statistics, Cambridge University Press, 13.10.1998

[Vigier] Jean Pierre Vigier, Etude sur les suites infinies d'opérateurs hermitiens, Ph.D. Thesis, Geneva, 1946

[Wei] Taoran Wei, Time Series in Functional Data Analysis, Master's Thesis, Technical University of Munich, Department of Mathematics, 2015