Technische Universität München
Fakultät für Mathematik
Lehrstuhl für Mathematische Statistik

Time series analysis in Hilbert spaces: Estimation of functional linear processes and prediction of traffic

Johannes Julian Klepsch

Vollständiger Abdruck der von der Fakultät für Mathematik der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktors der Naturwissenschaften (Dr. rer. nat.)

genehmigten Dissertation.

Vorsitzende(r):		Prof. Dr. Felix Krahmer
Prüfer der Dissertation:	1.	Prof. Dr. Claudia Klüppelberg
	2.	Prof. Dr. Klaus Mainzer
	3.	Prof. Dr. Alexander Aue (University of California Davis)

Die Dissertation wurde am 10.01.2017 bei der Technischen Universität München eingereicht und durch die Fakultät für Mathematik am 20.03.2017 angenommen.

Summary

When observations are curves over some natural time interval, the field of functional data analysis comes into play. The curves are considered to be elements of an infinite-dimensional Hilbert space, often the space of square integrable functions on the unit interval. In this thesis, the case where the observed curves are dependent in time is of interest. The temporal dependence between different curves is modelled with so called functional linear processes (FLP).

A special case of these FLPs is considered in the first part of the thesis, where the focus is on functional ARMA processes of order (p, q). Sufficient conditions for the existence of a unique stationary solution to the model equations are derived. It is then shown that a FLP can naturally be approximated by a vector ARMA(p, q) model by applying the concept of functional principal components. The finite-dimensional stationary vector model is used to predict the functional process, and a bound for the normed difference between vector and functional best linear predictor is derived. Finally, functional ARMA processes are applied for the modelling and prediction of highway traffic data.

The second part of the thesis is more technical. In this part the existence of the best linear predictor of a functional time series is investigated on a population level. The infinite dimensionality makes the problem of finding the best linear predictor of a FLP non-standard and difficult. By the construction of a sequence of increasing nested subspaces of the underlying Hilbert space, the well-known Innovations Algorithm from multivariate time series analysis is adapted to function spaces. Depending on the decay rate of the eigenvalues of the covariance and the spectral density operator, the resulting predictor converges with a certain rate to the theoretically best linear predictor. Several side results characterize subprocesses of functional linear models and special attention is paid to the interesting class of

functional MA models.

The third part of the thesis continues this line of research by proposing a first methodologically sound approach to estimate invertible functional time series by fitting functional moving average processes. In a finite sample setting, the concept of weak dependence of functional time series is used to consistently estimate the covariance operators of the FLP. Then conditions are derived such that the functional Innovations Algorithm, introduced in the second part of the thesis, provides estimators for the coefficient operators of a functional moving average model. The consistency of the estimators is derived in two different settings: first the sequence of increasing nested subspaces needed for the construction of the Innovations Algorithm is assumed to be known. Later on the proof is generalized to the case where the sequence has to be estimated. Different criteria for model selection are introduced and compared in a simulation study. In a real data example, highway traffic is investigated to compare the performance of our Innovations Algorithm estimator to known methods for the estimation of FMA(1) models.

Zusammenfassung

Der Bereich der funktionalen Datenanalyse kommt zur Anwendung wenn Beobachtungen als Kurven über ein natürliches Zeitintervall gesehen werden können. Die Kurven werden als Elemente eines unendlichdimensionalen Hilbert-Raumes interpretiert. Oftmals wird hierfür der Raum der quadratischen integrierbaren Funktionen auf dem Einheitsintervall verwendet. Diese Arbeit beschäftigt sich mit dem Fall, bei dem die beobachteten Kurven zeitlich voneinander abhängen. Die zeitliche Dynamik der Kurven wird mit funktionalen linearen Prozessen (FLP) modelliert.

Ein spezieller Fall dieser FLP wird im ersten Teil der Arbeit betrachtet, wo der Fokus auf funktionalen ARMA-Prozessen der Ordnung (p, q) liegt. Es werden Bedingungen für die Existenz einer eindeutigen stationären Lösung der ARMA-Modellgleichungen hergeleitet. Weiterhin wird gezeigt, dass ein FLP durch ein Vektor-ARMA(p, q) Modell in natürlicher Weise angenähert werden kann. Dabei wird das Konzept der funktionalen Hauptkomponentenanalyse angewendet. Das stationäre Vektormodell wird anschließend verwendet, um eine Prognose für den funktionalen Prozess aufzustellen. Es wird gezeigt, dass die normierte Differenz zwischen dem linearen Prädiktor, basierend auf dem Vektormodell, und dem besten funktionalen linearen Prädiktor eine obere Schranke besitzt. Schließlich werden funktionale ARMA-Prozesse zur Modellierung und Vorhersage von Autobahnverkehrsdaten genutzt.

Im zweiten Teil der Arbeit wird die Existenz des besten linearen Prädiktors einer funktionalen Zeitreihe in Hilberträumen untersucht. Die Suche nach dem besten linearen Prädiktor eines FLP ist schwierig, da der zugrundeliegende Funktionenraum unendlichdimensional ist. Durch den Aufbau einer Folge von monoton wachsenden Unterräumen des Hilbertraums wird der Innovationsalgorithmus, welcher aus der multivariaten Zeitreihenanalyse bekannt ist, an Funktionenräume angepasst. Die

Anwendung Innovationsalgorithmus führt zu einem linearen Prädiktor für FLPs. Unter Bedingungen an die Eigenwerte des Kovarianz- und des Spektraldichteoperators konvergiert dieser mit einer explizit gegebenen Rate gegen den theoretisch besten linearen Prädiktor . Weitere Ergebnisse charakterisieren endlichdimensionale Projektionen von funktionalen linearen Prozessen. Insbesondere wird die interessante Klasse der funktionalen Moving-Average-Modelle untersucht.

Der dritte Teil der Arbeit setzt diese Forschungsrichtung fort und schlägt einen ersten fundierten Ansatz zur Schätzung von invertierbaren FLPs vor. Das Konzept der schwachen Abhängigkeit von funktionalen Zeitreihen wird verwendet, um konsistente Schätzer der Kovarianzoperatoren eines FLP zu erhalten. Anschließend werden Bedingungen hergeleitet, sodass der im zweiten Teil eingeführte funktionale Innovationsalgorithmus Schätzer für die Koeffizientenoperatoren eines funktionalen Moving-Average-Modells liefert. Die Konsistenz der Schätzer wird in zwei Fällen bewiesen: Zunächst wird die Reihe der monoton wachsenden Unterräume, die für die Konstruktion des Algorithmus benötigt wird, als bekannt vorausgesetzt. Im Anschluss wird der Beweis verallgemeinert. Unterschiedliche Kriterien für die Modellauswahl werden eingeführt und in einer Simulationsstudie und mit Autobahnverkehrsdaten verglichen, und die Schätzer auf ihre Genauigkeit getestet.

Acknowledgments

I first of all want to thank my supervisor Claudia Klüppelberg. She always seemed to know where everything would eventually lead, both scientifically and personally, which was a big comfort. She provided me with the opportunity to work in a modern, quickly developing field and encouraged me to discuss my research with renowned experts. I here want to thank Siegfried Hörmann and Victor Panaretos for fruitful discussions when they visited the TUM.

Next, I can't thank Alexander Aue enough for hosting me at UC Davis. Working with him was both productive and diverting. His experience in functional time series and in soccer proved to be very inspiring.

Furthermore I thank the Autobahndirektion Südbayern and especially J. Grötsch for their support and for providing the traffic data.

The Munich Center for Technology and Society (MCTS) provided financial support and gave me the chance to work in an interdisciplinary team on the very interesting topics of highly automated driving and mobility of the future. I am extremely grateful for this opportunity.

Last I thank my family, friends and my colleagues at the Chair of Mathematical Statistics for making the last three years very enjoyable.

Introduction

Time series analysis

Whenever measurements are recorded repeatedly over time, the resulting dataset is referred to as a time series. Early examples of such recordings go back to measurements of the height of the Nile river, which has been reported annually since 622. Nowadays time series can be found in any domain of applied science that involves temporal measurements. Common examples include stock price returns, temperature or rainfall data, vehicle traffic flow and particle concentration or pollution data, just to name a few.

Theoretical developments in modern time series analysis started in the beginning of the last century. Since then stochastic movements are no longer merely regarded as residuals without significance for the future structure of the process. The modern approach rather assumes that stochastic impacts influence all components of a time series. An observed time series is since then seen as a realization of a data generating process (Kirchgässner et al. [33], Chapter 1). First attempts in this direction were made in the 1920s and 1930s by Yule and Slutzky. Wold systematised their work in his thesis [56], introducing the autoregressive moving average (ARMA) model to describe stationary time series. It took until the 1970s before G.E.P. Box and G.M. Jenkins published the classic book Box and Jenkins [12], which contained the first full modelling procedure for univariate time series. This led to a widespread application of modern time series analysis, and is still used and known as the Box-Jenkins method. From thereon the field of time series analysis developed quickly with generalizations from univariate linear stationary time series to more and more complex models. Important reference books include Brockwell and Davis [13] for a systematic account of linear time series models, Hamilton [22] for a theoretical introduction to traditional time series analysis combined with a review of more recent research, Hannan [23], Lütkepohl [39] and Tsay [53] for an introduction to multivariate time series, Shumway and Stoffer [49] for an overview on the use of recursive computation in the state space model, and Priestley [46] for an introduction to the spectral analysis of time series.

Since the beginning of this century, as noted for example in a survey in Tsay [52], an important driving force of research are advances in high-volume data acquisition. In fact, the advent of complex data challenges traditional time series techniques and requires the development of modern statistical technology (for example Jacod and Protter [31]). One option is the use of functional data analysis.

Functional data analysis

In recent years functional data analysis has established itself as an important and dynamic area of statistics. Functional data come in many forms, but it always consists of functions, often smooth curves. In some cases, the original observations are interpolated from longitudinal data, in other cases data are curves observed on a surface or in space. Quite regularly functional data are collected sequentially over time, and the different curves of functional observations are obtained by separating a continuous time record into disjoint natural time intervals, for example hours, days or years. One often anticipates that the recorded curves show similar shapes. These similarities may then be useful to the statistician in terms of complexity reduction.

One then assumes that the curve, say $X(t), t \in [0,1]$, is the representation of a random variable taking values in a function space. The parametrization $t \in [0,1]$ is standard in functional data analysis but can easily be generalized. Useful function spaces are the Hilbert space $L^2[0,1]$, the space of square integrable functions on $[0,1]$, and $C[0,1]$, the space of continuous functions on $[0,1]$. A key fact in functional data analysis is that one is dealing with infinite-dimensional objects. Therefore, most of the techniques known from multivariate data analysis cannot be directly applied.

Literature on functional data analysis is growing quickly. Ramsay and Silverman [47] and [48] offer an introduction to and applications of functional data analysis, Hsing and Eubank [29] furnishes theoretical foundations and an introduction to functional linear regression models, and an overview of recent developments is given in Ferraty and Romain [17] and Ferraty and Vieu [18].

Functional time series

All the above references have in common that they treat the functional observations as independent. However, in the case of sequentially recorded observations, this assumption may not hold. For instance, a return curve of some financial asset on day i is very likely to depend on the curves on days $i-j$ for $j<i$, where i and j are natural numbers.

As an example, Figure 1 shows how discretely observed data-points can be seen as functional observations. The plot shows highway traffic speed data observed at a fixed point on a highway. The recordings are measured every minute over six months, but instead of treating them as individual data points, the dataset is split up in days, and each day is considered as one datapoint consisting of a curve.

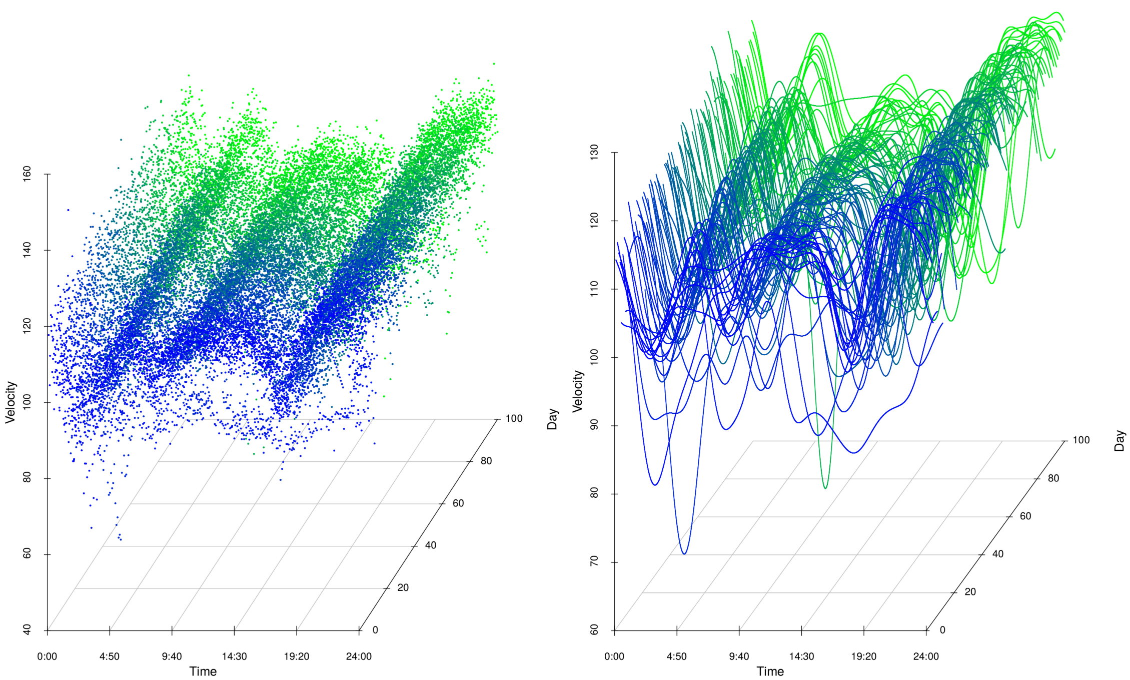

Figure 1: Raw and functional highway traffic speed data on 100 consecutive workings days in 2014

The topic of dependent functional data is quickly developing. The pioneering work of Bosq summarized in Bosq [8] is the most important reference to this thesis. It introduces the functional autoregressive process, develops estimation of its second order structure and derives conditions for consistency of estimators of the model parameters.

Frequency domain approaches have seen a quick rise in popularity in functional time series analysis since Panaretos and Tavakoli [44] introduced functional Fourier analysis based on functional cumulant conditions. As in multivariate time series analysis, functional frequency domain techniques may facilitate the understanding of temporal dependence in infinite-dimensional function spaces, as can be seen by the definition of dynamic functional principal components in Hörmann et al. [25] and Panaretos and Tavakoli [45]. Furthermore the frequency domain allows for the development of tests for stationarity (Aue and Delft [2]) or periodicity (Hörmann et al. [26]).

In the time domain analysis of functional time series, key references include Hörmann and Kokoszka [24] who developed a theoretical framework to describe a weak dependence concept of stationary functional time series, allowing to prove fundamental limit results in a general setting. Furthermore Aue et al. [5] developed a prediction technique for functional time series. This is a key publication as it builds a bridge between known tools from multivariate time series analysis and techniques from functional data analysis.

Time domain approaches often rely on functional principal component analysis (PCA) (see Horvàth and Kokoszka [27], Chapter 3) as the key tool in functional data analysis. Analogously to its multivariate counterpart, functional PCA relies on a decomposition of the variability of functional observations into principal directions. The principal directions are orthogonal functions, each explaining a proportion of the total variability of the data. It can be shown that, for independent and identically distributed data, a PCA yields the optimal finite-dimensional representation of a functional object in the L^2-norm.

The most intuitive link between multivariate time series analysis and functional data is therefore to use PCA to project the infinite-dimensional curves on a d-dimensional subspace and to then use the isometric isomorphy between d-dimensional Hilbert spaces and $\mathbb{R}^d$. As soon as data is represented as vectors in

$\mathbb{R}^d$, the entire literature on multivariate time series can be used (see Aue et al. [5]). However, there is a drawback. The subspace generated by the first d principal directions is a good choice in terms of representing the variability of the process, but it may not capture the dependence of the process. Principal directions not explaining much variability might still be relevant in terms of capturing the dependence of the model (see Kargin and Onatski [32]). Furthermore, even if the dynamics of the functional process are known, determining the dynamics of a finite-dimensional projection of the process is a highly nontrivial task. Projecting for example a functional autoregressive process of order p on a d-dimensional subspace does not generally yield a d-dimensional vector autoregressive process of order p.

The aim of this thesis is to contribute to the discussion about the consequences of projecting dependent data on the subspace generated by functional principal directions. Assuming that the dependence in a given functional dataset can be modelled with a functional linear process, what are the consequences of projecting the data on some finite-dimensional subspace? How can estimation and prediction be carried out? What is the error induced by dimension reduction? Can consistency be achieved not only in the finite-dimensional space but also in the functional setting?

Main results of this thesis

The thesis is structured in three main parts. The following paragraphs summarize the results of these chapters individually.

In Chapter 1 functional autoregressive moving average (FARMA) processes, a particular class of functional linear processes, are investigated. First sufficient conditions for the existence of a unique stationary solution to the model equations are derived. The sufficient conditions are natural extensions of the conditions developed in Bosq [8] for functional autoregressive (AR) models. The focus of the chapter is on prediction of FARMA(p, q) processes. As indicated, techniques known from multivariate analysis can not be directly applied as the objects under investigation live in function spaces. With the approach of Aue et al. [5], who used the methodology in the context of functional (AR) models, the functional process is projected on a finite-dimensional subspace using functional PCA. However the resulting multivariate process does not a priori follow a vector ARMA model. It is shown that the vector process can still be naturally approximated by a vector ARMA(p, q) model. Conditions for the approximation to be exact are investigated. The stationary vector model is then used to predict the functional process. The main result (Theorem 1.4.11) of the chapter quantifies the normed difference between vector and functional best linear predictor. The obtained bound naturally depends on two terms: one originating from the stochastic error and another from the error induced by reducing the dimension. Finally the methodology is applied to real data. The goal is a realistic time series model for traffic speed, which captures the day to day dependence. The analysis can support short term traffic regulation realised in real-time by electronic devices during the day, which may benefit from a more precise and parsimonious day-to-day prediction. An important factor in the application is the choice of the dimension of the underlying finite-dimensional space and the choice of the model order. Using cross validation, a criterion based on the functional prediction error is developed that, when minimized, yields optimal dimension and model order. The appeal of the methodology is its ease of application. Well-known `R` software packages (`fda` and `mts`) make the implementation straightforward. Furthermore, the generality of dependence induced by ARMA models gives rise to a wide range of application of functional time series.

In Chapter 2, the true dependence structure of a finite-dimensional projection of a functional linear process is investigated. Projecting a process of the general class of functional linear processes on a d-dimensional space results in a functional subprocess, which is isomorph to a d-dimensional vector process. The Wold decomposition is used on the subprocess to parametrize the true second order dependence structure. A special case is the class of functional moving average (MA) processes: it is shown in Theorem 2.4.7 that every subprocess of a functional MA process of order q is isomorph to a vector MA process of order q^*, with $q^* \le q$. A useful and interesting side result is that every subprocess of an invertible functional linear process is invertible (Proposition 2.4.3). The main contribution of the chapter is the development of a functional Innovations Algorithm. The multivariate Innovations Algorithm (Brockwell and Davis [13], Chapter 11) is extended to function spaces. The key problem in the generalization of the algorithm is, as often in functional data analysis, the non-invertibility of covariance operators. To resolve this problem, a similar tool as in Bosq [8], for the estimation of functional autoregressive processes of order 1, is proposed. The number of principal directions, generating the space on which covariance operators are inverted, is increased together with the sample size. However, since the Innovations Algorithm is based on a recursion, the recursion start always fails if the number of principal directions is chosen too large. The solution lies in an iterative increase of the number of principal directions included in the Innovations Algorithm. This results in the construction of a sequence of increasing nested subspaces of the function space, which is the real novelty of this approach. The algorithm is used to construct linear predictors of functional linear processes. Even though the existence of the best linear predictor in function spaces has been shown in Bosq [10], its practical implementation is highly nontrivial and appears to be not well understood in the literature. The functional Innovations Algorithm is a first attempt in this direction. Under conditions on the decay rate of the eigenvalues of the covariance and the spectral density operator, the Innovations Algorithm constructs a computationally tractable functional linear predictor for stationary invertible functional linear processes. As the sample size increases, the predictor is equivalent to the best linear predictor introduced in [10] (Theorem 2.5.3). Explicit rates of convergence can be derived. They are given by a combination of two tail sums, one involving operators of the inverse representation of the process, and the other the eigenvalues of the covariance operator.

In Chapter 2 it is assumed that quantities such as covariance operators determining the second order structure of the functional linear process are known. Chapter 3, however, deals with the finite sample case. Chapter 3 proposes a first methodologically sound approach to estimate invertible functional linear processes by fitting functional MA models. Making use of the property shown in Chapter 2 that subprocesses of functional MA processes are isometrically isomorph to vector MA of smaller or equal order, the idea is to estimate the coefficient operators in a functional linear filter. To this end the functional Innovations Algorithm of Chapter 2 is utilized as a starting point to estimate the corresponding moving average operators via suitable projections into principal directions. The main result is the proof of consistency of the proposed estimators (Theorem 3.3.5). The difficulty is that in order to apply the functional Innovations Algorithm, one has to estimate both the covariance operator of the functional process and the sequence of increasing nested subspaces of the function space. To ensure appropriate large-sample properties of the proposed estimators, the growth-rate of the sequence of nested subspaces has to depend on the decay rate of the infimum of the eigenvalues of the spectral density operator. For practical purposes, several strategies to select the number of principal directions in the estimation procedure as well as the choice of order of the functional moving average process are discussed. An independence test is introduced to select the dimension of the principal projection subspace, which can be used as a starting point for the suggested order selection procedures based on AICC and Ljung-Box criteria. Additionally, an fFPE criterion is established that jointly selects dimension d and order q. Their empirical performance is evaluated through Monte-Carlo studies and an application to vehicle traffic data.

Final remarks

The above does not qualify as a full introduction to the individual subsequent chapters. Neither a detailed literature review nor notational conventions or theoretical background is given. The individual chapters are self-contained in the sense that each of them introduces the notation, methodology and literature needed to be comprehensible. Notations and abbreviations might differ from chapter to chapter since different notations seem reasonable in different settings.

All chapters are based on publications or are submitted for publication.

- Chaper 1 is based on the paper [35] that is published as: J. Klepsch, C. Küppelberg, and T. Wei. "Prediction of functional ARMA processes with an application to traffic data". *Econometrics and Statistics*, 1:128-149, 2017.

- Chapter 2 is based on the paper [34] that is published as: J. Klepsch and C. Klüppelberg. "An Innovations Algorithm for the prediction of functional linear processes". *Journal of Multivariate Analysis*, 155:252-271, 2017.

- Chapter 3 is based on the paper [3] that is submitted for publication as: A. Aue and J. Klepsch. "Estimating functional time series by moving average model fitting, *preprint at arXiv:1701.00770[ME]*, 2017

Contents

List of Figures

Chapter 1:
Prediction of functional ARMA processes with an application to traffic data

1.1 Introduction

A *macroscopic highway traffic model* involves velocity, flow (number of vehicles passing a reference point per unit of time), and density (number of vehicles on a given road segment). The relation among these three variables is depicted in diagrams of "velocity-flow relation" and "flow-density relation". The diagram of "flow-density relation" is also called *fundamental diagram of traffic flow* and can be used to determine the capacity of a road system and give guidance for inflow regulations or speed limits. Figures 1.1 and 1.2 depict these quantities for traffic data provided by the Autobahndirektion Südbayern. At a critical traffic density (65 veh/km) the state of flow on the highway will change from stable to unstable.

In this chapter we develop a *statistical highway traffic model* and apply it to the above data. As can be seen from Figures 1.4 and 1.5 the data show a certain pattern over the day, which we want to capture utilising tools from functional data analysis. Functional data analysis is applied to represent the very high-dimensional traffic velocity data over the day by a random function $X(\cdot)$. This is a standard procedure, and we refer to Ramsay and Silverman [47] for details.

Given the functional data, we want to assess temporal dependence between different days; i.e., our goal is a realistic time series model for functional data, which captures the day-to-day dependence. Our analysis can support short term traffic regulation realised in real-time by electronic devices during the day, which may benefit

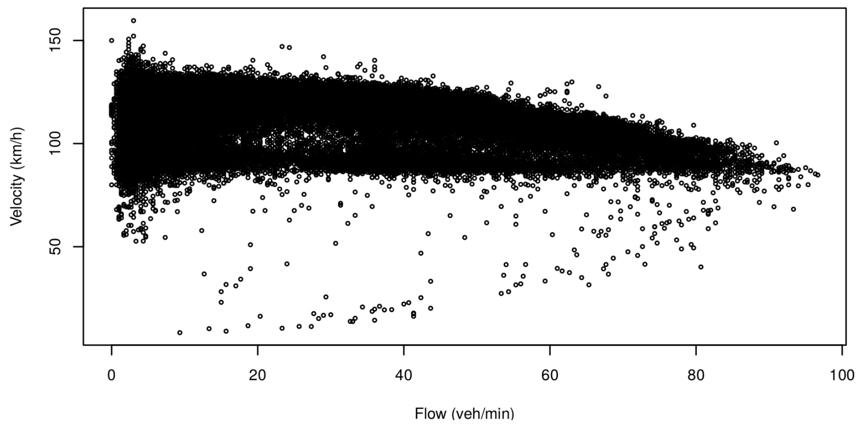

Figure 1.1: Velocity-flow relation on highway A92 in Southern Bavaria. Depicted are average velocities per 3 min versus number of vehicles within these 3 min during the period 01/01/2014 0:00 to 30/06/2014 23:59.

from a more precise and parsimonious day-to-day prediction.

From a statistical point of view we are interested in the prediction of a functional ARMA(p,q) process for arbitrary orders p and q. In scalar and multivariate time series analysis there exist several prediction methods, which can be easily implemented like the Durbin-Levinson and the Innovations Algorithm (see e.g Brockwell and Davis [13]). For functional time series, Bosq [8] has proposed the *functional best linear predictor* for a general linear process. However, implementation of the predictor is in general not feasible, because explicit formulas of the predictor can not be derived. The class of functional AR(p) processes is an exception, where explicit prediction formulas have been given (e.g. Bosq [8], Chapter 3, and Kargin and Onatski [32]). The functional AR(1) model has also been applied to the prediction of traffic data in Besse and Cardot [7].

In Aue et al. [5] a prediction algorithm is proposed, which combines the idea of functional principal component analysis (FPCA) and functional time series analysis. The basic idea is to reduce the infinite-dimensional functional data by FPCA to vector data. Thus, the task of predicting a functional time series is transformed to the prediction of a multivariate time series. In Aue et al. [5] this algorithm is used to predict the functional AR(p) process.

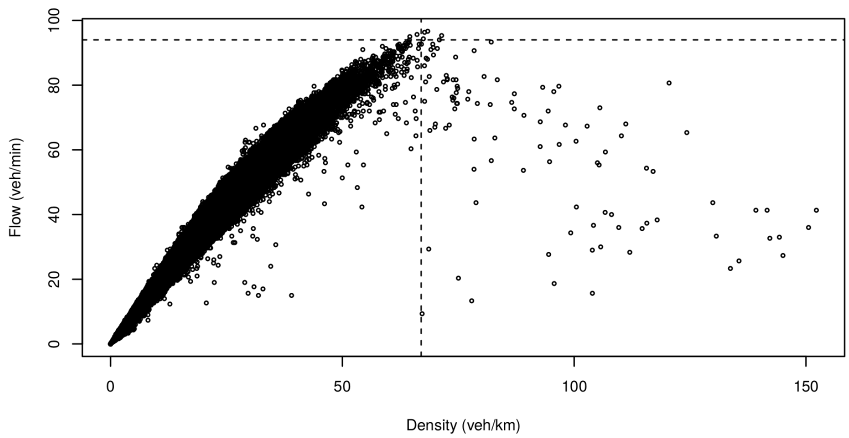

Figure 1.2: Flow-density relation for the data from Figure 1.1 with critical traffic density of 65 veh/km.

In this chapter we focus on functional ARMA(p, q) processes. We start by providing sufficient conditions for the existence of a stationary solution to functional ARMA(p, q) models. Then we obtain a vector process by projecting the functional process on the linear span of the d most important eigenfunctions of the covariance operator of the process. We derive conditions such that the projected process follows a vector ARMA(p, q). If these conditions do not hold, we show that the projected process can at least be approximated by a vector ARMA(p, q) process, and we assess the quality of the approximation. We present conditions such that the vector model equation has a unique stationary solution. This leads to prediction methods for functional ARMA(p, q) processes. An extension of the prediction algorithm of Aue et al. [5] can be applied, and makes sense under stationarity of both the functional and the vector ARMA(p, q) process. We derive bounds for the difference between vector and functional best linear predictor.

An extended simulation study can be found in Wei [55], Chapter 5, and confirms that approximating the projection of a functional ARMA process by a vector ARMA process of the same order works reasonably well.

This chapter is organised as follows. In Section 1.2 we introduce the necessary Hilbert space theory and notation, that we use throughout. We present the

Karhunen-Loève Theorem and describe the FPCA based on the functional covariance operator. In Section 1.3 we turn to functional time series models with special emphasis on functional ARMA(p, q) processes. Section 1.3.1 is devoted to stationarity conditions for the functional ARMA(p, q) model. In Section 1.3.2 we study the vector process obtained by projection of the functional process onto the linear span of the d most important eigenfunctions of the covariance operator. We investigate its stationarity and prove that a vector ARMA process approximates the functional ARMA process in a natural way. Section 1.4 investigates the prediction algorithm for functional ARMA(p, q) processes invoking the vector process, and compares it to the functional best linear predictor. Finally, in Section 1.5 we apply our results to traffic data of velocity measurements.

1.2 Methodology

We summarize some concepts which we shall use throughout the chapter. For details and more background we refer to the monographs Bosq [8], Horvàth and Kokoszka [27] and Hsing and Eubank [29]. Let $H = L^2([0,1])$ be the real separable Hilbert space of square integrable functions $x : [0,1] \to \mathbb{R}$ with norm $\|x\| = (\int_0^1 x^2(s)ds)^{1/2}$ generated by the inner product

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

We shall often use Parseval's equality, which ensures that for an orthonormal basis (ONB) $(e_i)_{i\in\mathbb{N}}$

$$\langle x, y \rangle = \sum_{i=1}^{\infty} \langle x, e_i \rangle \langle e_i, y \rangle, \quad x, y \in H. \tag{1.2.1}$$

We denote by $\mathcal{L}$ the space of bounded linear operators acting on H. If not stated differently, we take the standard operator norm defined for a bounded operator $\Psi \in \mathcal{L}$ by $\|\Psi\|_{\mathcal{L}} := \sup_{\|x\|\le 1} \|\Psi x\|$.

A bounded linear operator Ψ is a *Hilbert-Schmidt* operator if it is compact and for every ONB $(e_i)_{i\in\mathbb{N}}$ of H

$$\sum_{i=1}^{\infty} \|\Psi e_i\|^2 < \infty.$$

We denote by $\mathcal{S}$ the space of Hilbert-Schmidt operators acting on H, which is again a separable Hilbert space equipped with the following inner product and corresponding *Hilbert-Schmidt norm*:

$$\langle \Psi_1, \Psi_2 \rangle_{\mathcal{S}} := \sum_{i=1}^{\infty} \langle \Psi_1 e_i, \Psi_2 e_i \rangle \quad \text{and} \quad \|\Psi\|_{\mathcal{S}} := \sqrt{\langle \Psi, \Psi \rangle_{\mathcal{S}}} = \sqrt{\sum_{i=1}^{\infty} \|\Psi e_i\|^2} < \infty.$$

If Ψ is a Hilbert-Schmidt operator, then

$$\|\Psi\|_{\mathcal{L}} \leq \|\Psi\|_{\mathcal{S}}.$$

Let $\mathcal{B}_H$ be the Borel σ-algebra of subsets of H. All random functions are defined on some probability space $(\Omega, \mathcal{A}, P)$ and are $\mathcal{A} - \mathcal{B}_H$-measurable. Then the space of square integrable random functions $L_H^2 := L_H^2(\Omega, \mathcal{A}, P)$ is a Hilbert space with inner product $E\langle X, Y \rangle = E\int_0^1 X(s)Y(s)ds$ for $X, Y \in L_H^2$. We call such X an H-*valued random function*. For $X \in L_H^2$ there is a unique function $\mu \in H$, the *functional mean* of X, such that $E\langle y, X \rangle = \langle y, \mu \rangle$ for $y \in H$, satisfying

$$\mu(t) = E[X(t)], \quad t \in [0,1].$$

We assume throughout that $\mu = 0$, since under weak assumptions on X the functional mean can be estimated consistently from the data (see Remark 1.3.10).

Definition 1.2.1. The *covariance operator* C_X of X acts on H and is defined as

$$C_X : x \mapsto E\left[\langle X, x \rangle X\right], \quad x \in H. \tag{1.2.2}$$

More precisely,

$$(C_X x)(t) = E\left[\int_0^1 X(s)x(s)\mathrm{d}s\, X(t)\right] = \int_0^1 E\left[X(t)X(s)\right] x(s)ds,$$

where the change of integration order is allowed by Fubini. □

C_X is a symmetric, non-negative definite Hilbert-Schmidt operator with spectral representation

$$C_X x = \sum_{j=1}^{\infty} \lambda_j \langle x, \nu_j \rangle \nu_j, \quad x \in H,$$

for eigenpairs $(\lambda_j, \nu_j)_{j\in\mathbb{N}}$, where $(\nu_j)_{j\in\mathbb{N}}$ is an ONB of H and $(\lambda_j)_{j\in\mathbb{N}}$ is a sequence of positive real numbers such that $\sum_{j=1}^{\infty} \lambda_j < \infty$. When considering spectral representations we assume that the λ_j are decreasingly ordered and that no ties in the eigenvalues are allowed; i.e., $\lambda_i > \lambda_k$ for $i < k$. Every $X \in L_H^2$ can be represented as a linear combination of the eigenfunctions $(\nu_i)_{i\in\mathbb{N}}$. This is known as the *Karhunen-Loève representation.*

Theorem 1.2.2 (Karhunen-Loève Theorem). *For $X \in L_H^2$ with $EX = 0$*

$$X = \sum_{i=1}^{\infty} \langle X, \nu_i \rangle \, \nu_i, \tag{1.2.3}$$

where $(\nu_i)_{i\in\mathbb{N}}$ are the eigenfunctions of the covariance operator C_X. The scalar products $\langle X, \nu_i \rangle$ have mean-zero, variance λ_i and are uncorrelated; i.e., for all $i, j \in \mathbb{N}$, $i \neq j$,

$$E \langle X, \nu_i \rangle = 0, \quad E[\langle X, \nu_i \rangle \langle X, \nu_j \rangle] = 0, \quad \text{and} \quad E \langle X, \nu_i \rangle^2 = \lambda_i, \tag{1.2.4}$$

where $(\lambda_i)_{i\in\mathbb{N}}$ are the eigenvalues of C_X.

The scalar products $(\langle X, \nu_i \rangle)_{i\in\mathbb{N}}$ defined in (1.2.3) are called the *scores* of X. By the last equation in (1.2.4), we have

$$\sum_{j=1}^{\infty} \lambda_j = \sum_{j=1}^{\infty} E \langle X, \nu_j \rangle^2 = E\|X\|^2 < \infty, \quad X \in L_H^2. \tag{1.2.5}$$

Combining (1.2.4) and (1.2.5), every λ_j represents some proportion of the total variability of X.

Remark 1.2.3. [The CVP method] For $d \in \mathbb{N}$ consider the d largest eigenvalues $\lambda_1, \dots, \lambda_d$ of C_X. The *cumulative percentage of total variance* $\text{CPV}(d)$ is defined as

$$\text{CPV}(d) := \sum_{j=1}^{d} \lambda_j \Big/ \sum_{j=1}^{\infty} \lambda_j.$$

If we choose $d \in \mathbb{N}$ such that the $\text{CPV}(d)$ exceeds a predetermined high percentage value, then $\lambda_1, \dots, \lambda_d$ explain most of the variability of X. In this context $\nu_1, \dots, \nu_d$ are called the *functional principal components* (FPCs). □

1.3 Functional ARMA processes

In this section we introduce the functional ARMA(p,q) equations and derive sufficient conditions for the equations to have a stationary and causal solution, which we present explicitly as a functional linear process. We then project the functional linear process on a finite dimensional subspace of H. We approximate this finite dimensional process by a suitable vector ARMA process, and give conditions for the stationarity of this vector process. We also give conditions on the functional ARMA model such that the projection of the functional process onto a finite dimensional space follows an exact vector ARMA structure.

We start by defining functional white noise.

Definition 1.3.1. [Bosq [8], Definition 3.1]
Let $(\varepsilon_n)_{n\in\mathbb{Z}}$ be a sequence of H-valued random functions.
(i) $(\varepsilon_n)_{n\in\mathbb{Z}}$ is *H-white noise* (WN) if for all $n\in\mathbb{Z}$, $E[\varepsilon_n]=0$, $0<E\|\varepsilon_n\|^2=\sigma_\varepsilon^2<\infty$, $C_{\varepsilon_n}=C_\varepsilon$, and if $C_{\varepsilon_n,\varepsilon_m}(\cdot):=E[\langle\varepsilon_m,\cdot\rangle\,\varepsilon_n]=0$ for all $n\neq m$.
(ii) $(\varepsilon_n)_{n\in\mathbb{Z}}$ is *H-strong white noise* (SWN), if for all $n\in\mathbb{Z}$, $E[\varepsilon_n]=0$, $0<E\|\varepsilon_n\|^2=\sigma_\varepsilon^2<\infty$ and $(\varepsilon_n)_{n\in\mathbb{Z}}$ is i.i.d. □

We assume throughout that $(\varepsilon_n)_{n\in\mathbb{Z}}$ is WN with zero mean and $E\|\varepsilon_n\|^2=\sigma_\varepsilon^2<\infty$. When SWN is required, this will be specified.

1.3.1 Stationary functional ARMA processes

Formally we can define a functional ARMA process of arbitrary order.

Definition 1.3.2. Let $(\varepsilon_n)_{n\in\mathbb{Z}}$ be WN as in Definition 1.3.1(i). Let furthermore $\phi_1,\dots,\phi_p,\ \theta_1,\dots,\theta_q\in\mathcal{L}$. Then a solution of

$$X_n=\sum_{i=1}^{p}\phi_i X_{n-i}+\sum_{j=1}^{q}\theta_j\varepsilon_{n-j}+\varepsilon_n,\quad n\in\mathbb{Z},\tag{1.3.1}$$

is called a *functional* ARMA(p,q) *process*. □

We derive conditions such that (1.3.1) has a stationary solution. We begin with the functional ARMA$(1,q)$ process and need the following assumption.

Assumption 1.3.3. *There exists some* $j_0 \in \mathbb{N}$ *such that* $\|\phi_1^{j_0}\|_{\mathcal{L}} < 1$.

Theorem 1.3.4. *Let* $(X_n)_{n\in\mathbb{Z}}$ *be as in Definition 1.3.2 with* $p = 1$ *and set* $\phi_1 =: \phi$. *If Assumption 1.3.3 holds, there exists a unique stationary and causal solution to (1.3.1) given by*

$$\begin{aligned} X_n &= \varepsilon_n + (\phi + \theta_1)\varepsilon_{n-1} + (\phi^2 + \phi\theta_1 + \theta_2)\varepsilon_{n-2} \\ &\quad + \cdots + (\phi^{q-1} + \phi^{q-2}\theta_1 + \cdots + \theta_{q-1})\varepsilon_{n-(q-1)} \\ &\quad + \sum_{j=q}^{\infty} \phi^{j-q}(\phi^q + \phi^{q-1}\theta_1 + \cdots + \theta_q)\varepsilon_{n-j} \\ &= \sum_{j=0}^{q-1}\Big(\sum_{k=0}^{j} \phi^{j-k}\theta_k\Big)\varepsilon_{n-j} + \sum_{j=q}^{\infty}\phi^{j-q}\Big(\sum_{k=0}^{q}\phi^{q-k}\theta_k\Big)\varepsilon_{n-j}, \end{aligned} \tag{1.3.2}$$

where $\phi^0 = I$ *denotes the identity operator in* H. *Furthermore, the series in (1.3.2) converges in* L_H^2 *and with probability one.*

For the proof we need the following lemma.

Lemma 1.3.5 (Bosq [8], Lemma 3.1)**.** *For every* $\phi \in \mathcal{L}$ *the following are equivalent:*
(i) There exists some $j_0 \in \mathbb{N}$ *such that* $\|\phi^{j_0}\|_{\mathcal{L}} < 1$.
(ii) There exist $a > 0$ *and* $0 < b < 1$ *such that* $\|\phi^j\|_{\mathcal{L}} < ab^j$ *for every* $j \in \mathbb{N}$.

Proof of Theorem 1.3.4. We follow the lines of the proof of Proposition 3.1.1 of Brockwell and Davis [13] and Theorem 3.1 in Bosq [8]. First we prove L_H^2-convergence of the series (1.3.2). Take $m \geq q$ and consider the truncated series

$$\begin{aligned} X_n^{(m)} &:= \varepsilon_n + (\phi + \theta_1)\varepsilon_{n-1} + (\phi^2 + \phi\theta_1 + \theta_2)\varepsilon_{n-2} \\ &\quad + \cdots + (\phi^{q-1} + \phi^{q-2}\theta_1 + \cdots + \theta_{q-1})\varepsilon_{n-(q-1)} \\ &\quad + \sum_{j=q}^{m} \phi^{j-q}(\phi^q + \phi^{q-1}\theta_1 + \cdots + \theta_q)\varepsilon_{n-j}. \end{aligned} \tag{1.3.3}$$

Define

$$\beta(\phi, \theta) := \phi^q + \phi^{q-1}\theta_1 + \cdots + \phi\theta_{q-1} + \theta_q \in \mathcal{L}.$$

Since $(\varepsilon_n)_{n\in\mathbb{Z}}$ is WN, for all $m' > m \geq q$,

$$\begin{aligned} E\|X_n^{(m')} - X_n^{(m)}\|^2 &= E\Big\| \sum_{j=m}^{m'} \phi^{j-q}\beta(\phi,\theta)\varepsilon_{n-j}\Big\|^2 \\ &= \sum_{j=m}^{m'} E\left\|\phi^{j-q}\beta(\phi,\theta)\varepsilon_{n-j}\right\|^2 \\ &\leq \sigma_\varepsilon^2 \sum_{j=m}^{m'} \left\|\phi^{j-q}\right\|_{\mathcal{L}}^2 \|\beta(\phi,\theta)\|_{\mathcal{L}}^2. \end{aligned}$$

Lemma 1.3.5 applies, giving

$$\sum_{j=0}^{\infty} \|\phi^j\|_{\mathcal{L}}^2 < \sum_{j=0}^{\infty} a^2 b^{2j} = \frac{a^2}{1-b^2} < \infty. \tag{1.3.4}$$

Thus,

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

By the Cauchy criterion the series in (1.3.2) converges in L_H^2.

To prove convergence with probability one we investigate the following second moment, using that $(\varepsilon_n)_{n\in\mathbb{Z}}$ is WN:

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

Finiteness follows, since by (1.3.4),

$$\Big(\sum_{j=1}^{\infty} \|\phi^{j-q}\|_{\mathcal{L}}^2\Big)^2 < \Big(\sum_{j=1}^{\infty} a b^{j-q}\Big)^2 = \frac{a^2}{(1-b)^2} < \infty.$$

Thus, the series (1.3.2) converges with probability one.

Note that the solution (1.3.2) is stationary, since its second order structure only depends on $(\varepsilon_n)_{n\in\mathbb{Z}}$, which is shift-invariant as WN.

In order to prove that (1.3.2) is a solution of (1.3.1) with $p = 1$, we plug (1.3.2) into

(1.3.1), and obtain for $n \in \mathbb{Z}$,

$$\begin{aligned} X_n - \phi X_{n-1} = &\sum_{j=0}^{q-1} \Big(\sum_{k=0}^{j} \phi^{j-k}\theta_k\Big)\varepsilon_{n-j} + \sum_{j=q}^{\infty} \phi^{j-q}\Big(\sum_{k=0}^{q} \phi^{q-k}\theta_k\Big)\varepsilon_{n-j} \\ &- \phi\Big(\sum_{j=0}^{q-1} \Big(\sum_{k=0}^{j} \phi^{j-k}\theta_k\Big)\varepsilon_{n-1-j} + \sum_{j=q}^{\infty} \phi^{j-q}\Big(\sum_{k=0}^{q} \phi^{q-k}\theta_k\Big)\varepsilon_{n-1-j}\Big). \end{aligned} \tag{1.3.5}$$

The third term of the right-hand side can be written as

$$\begin{aligned} &\sum_{j=0}^{q-1} \Big(\sum_{k=0}^{j} \phi^{j+1-k}\theta_k\Big)\varepsilon_{n-1-j} + \sum_{j=q}^{\infty} \phi^{j+1-q}\Big(\sum_{k=0}^{q} \phi^{q-k}\theta_k\Big)\varepsilon_{n-1-j} \\ = &\sum_{j'=1}^{q} \Big(\sum_{k=0}^{j'-1} \phi^{j'-k}\theta_k\Big)\varepsilon_{n-j'} + \sum_{j'=q+1}^{\infty} \phi^{j'-q}\Big(\sum_{k=0}^{q} \phi^{q-k}\theta_k\Big)\varepsilon_{n-j'} \\ = &\sum_{j'=1}^{q} \Big(\sum_{k=0}^{j'} \phi^{j'-k}\theta_k - \phi^{j'-j'}\theta_{j'}\Big)\varepsilon_{n-j'} + \sum_{j'=q+1}^{\infty} \phi^{j'-q}\Big(\sum_{k=0}^{q} \phi^{q-k}\theta_k\Big)\varepsilon_{n-j'} \\ = &\sum_{j'=1}^{q} \Big(\sum_{k=0}^{j'} \phi^{j'-k}\theta_k\Big)\varepsilon_{n-j'} + \sum_{j'=q+1}^{\infty} \phi^{j'-q}\Big(\sum_{k=0}^{q} \phi^{q-k}\theta_k\Big)\varepsilon_{n-j'} - \sum_{j'=1}^{q} \theta_{j'}\varepsilon_{n-j'}. \end{aligned}$$

Comparing the sums in (1.3.5), the only remaining terms are

$$\begin{aligned} X_n - \phi X_{n-1} &= \varepsilon_n - \sum_{k=0}^{q} \phi^{q-k}\theta_k\varepsilon_{n-q} + \sum_{j'=1}^{q} \theta_{j'}\varepsilon_{n-j'} + \sum_{k=0}^{q} \phi^{q-k}\theta_k\varepsilon_{n-q} \\ &= \varepsilon_n + \sum_{j'=1}^{q} \theta_{j'}\varepsilon_{n-j'}, \quad n \in \mathbb{Z}, \end{aligned}$$

which shows that (1.3.2) is a solution of equation (1.3.1) with $p = 1$.

Finally, we prove uniqueness of the solution. Assume that there is another stationary solution X_n' of (1.3.1). Iteration gives (cf. Spangenberg [51], eq. (4)) for all $r > q$,

$$\begin{aligned} X_n' = &\sum_{j=0}^{q-1} \Big(\sum_{k=0}^{j} \phi^{j-k}\theta_k\Big)\varepsilon_{n-j} + \sum_{j=q}^{r-1} \phi^{j-q}\Big(\sum_{k=0}^{q} \phi^{q-k}\theta_k\Big)\varepsilon_{n-j} \\ &+ \sum_{j=0}^{q-1} \phi^{r+j-q}\Big(\sum_{k=j+1}^{q} \phi^{q-k}\theta_k\Big)\varepsilon_{n-(r+j)} + \phi^r X_{n-r}'. \end{aligned}$$

Therefore, with $X^{(r)}$ as in (1.3.3), for $r > q$,

$$E\left\|X'_n - X_n^{(r)}\right\|^2 = E\left\|\sum_{j=0}^{q-1}\phi^{r+j-q}\Big(\sum_{k=j+1}^{q}\phi^{q-k}\theta_k\Big)\varepsilon_{n-(r+j)} + \phi^r X'_{n-r}\right\|^2$$
$$\leq 2E\left\|\sum_{j=0}^{q-1}\phi^{r+j-q}\Big(\sum_{k=j+1}^{q}\phi^{q-k}\theta_k\Big)\varepsilon_{n-(r+j)}\right\|^2 + 2\,E\left\|\phi^r X'_{n-r}\right\|^2$$
$$\leq 2\|\phi^{r-q}\|_{\mathcal{L}}^2 E\left\|\sum_{j=0}^{q-1}\phi^{j}\Big(\sum_{k=j+1}^{q}\phi^{q-k}\theta_k\Big)\varepsilon_{n-(r+j)}\right\|^2 + 2\|\phi^r\|_{\mathcal{L}}^2 E\left\|X'_{n-r}\right\|^2.$$

Since both $(\varepsilon_n)_{n\in\mathbb{Z}}$ and $(X'_n)_{n\in\mathbb{Z}}$ are stationary, Lemma 1.3.5 yields

$$E\left\|X'_n - X_n^{(r)}\right\|^2 \to 0, \quad r \to \infty.$$

Thus X'_n is in L^2_H equal to the limit X_n of $X_n^{(r)}$, which proves uniqueness. □

Remark 1.3.6. In Spangenberg [51] a strictly stationary, not necessarily causal solution of a functional ARMA(p,q) equation in Banach spaces is derived under minimal conditions. This extends known results considerably. □

For a functional ARMA(p,q) process we use the state space representation

$$\underbrace{\begin{pmatrix} X_n \\ X_{n-1} \\ \vdots \\ X_{n-p+1} \end{pmatrix}}_{Y_n} = \underbrace{\begin{pmatrix} \phi_1 & \cdots & \phi_{p-1} & \phi_p \\ I & & & 0 \\ & \ddots & & \vdots \\ & & I & 0 \end{pmatrix}}_{\widetilde{\phi}} \underbrace{\begin{pmatrix} X_{n-1} \\ X_{n-2} \\ \vdots \\ X_{n-p} \end{pmatrix}}_{Y_{n-1}} + \sum_{j=0}^{q} \underbrace{\begin{pmatrix} \theta_j & 0 & \cdots & 0 \\ 0 & 0 & & \vdots \\ \vdots & & \ddots & \\ 0 & & & 0 \end{pmatrix}}_{\widetilde{\theta}_j} \underbrace{\begin{pmatrix} \varepsilon_{n-j} \\ 0 \\ \vdots \\ 0 \end{pmatrix}}_{\delta_{n-j}}, \tag{1.3.6}$$

where $\theta_0 = I$, and I and 0 denote the identity and zero operators on H, respectively. We summarize this as

$$Y_n = \widetilde{\phi} Y_{n-1} + \sum_{j=0}^{q} \widetilde{\theta}_j \delta_{n-j}, \quad n \in \mathbb{Z}. \tag{1.3.7}$$

Since X_n and ε_n take values in H, Y_n and δ_n take values in the product Hilbert space $H^p := (L^2([0,1]))^p$ with inner product and norm given by

$$\langle x, y\rangle_p := \sum_{j=1}^{p} \langle x_j, y_j\rangle \quad \text{and} \quad \|x\|_p := \sqrt{\langle x, x\rangle_p}. \tag{1.3.8}$$

We denote by $\mathcal{L}(H^p)$ the space of bounded linear operators acting on H^p. The operator norm of $\widetilde{\phi} \in \mathcal{L}(H^p)$ is defined as usual by $\|\widetilde{\phi}\|_{\mathcal{L}} := \sup_{\|x\|_p \le 1} \|\widetilde{\phi} x\|_p$. The random vector $(\delta_n)_{n\in\mathbb{Z}}$ is WN in H^p.

Assumption 1.3.7. *There exists some $j_0 \in \mathbb{N}$ such that $\widetilde{\phi}$ as in* (1.3.6) *satisfies $\|\widetilde{\phi}^{j_0}\|_{\mathcal{L}} < 1$.*

Since the proof of Theorem 1.3.4 holds also in H^p, using the state space representation of a functional ARMA(p,q) in H as a functional ARMA$(1,q)$ in H^p, we get the following theorem as a consequence of Theorem 1.3.4. Let P_1 be the projection of H^p onto the first component; i.e.,

$$P_1 x = x_1, \quad x = (x_1, \dots, x_n)^\top \in H^p.$$

Theorem 1.3.8. *Under Assumption 1.3.7 there exists a unique stationary and causal solution to the functional* ARMA(p,q) *equations (1.3.1). The solution can be written as $X_n = P_1 Y_n$, where Y_n is the solution to the state space equation* (1.3.7), *given by*

$$\begin{aligned} Y_n &= \delta_n + (\widetilde{\phi} + \widetilde{\theta}_1)\delta_{n-1} + (\widetilde{\phi}^2 + \widetilde{\phi}\,\widetilde{\theta}_1 + \widetilde{\theta}_2)\delta_{n-2} \\ &\quad + \cdots + (\widetilde{\phi}^{q-1} + \widetilde{\phi}^{q-2}\,\widetilde{\theta}_1 + \cdots + \widetilde{\theta}_{q-1})\delta_{n-(q-1)} \\ &\quad + \sum_{j=q}^{\infty} \widetilde{\phi}^{j-q}(\widetilde{\phi}^q + \widetilde{\phi}^{q-1}\,\widetilde{\theta}_1 + \cdots + \widetilde{\theta}_q)\delta_{n-j}, \\ &= \sum_{j=0}^{q-1}\Big(\sum_{k=0}^{j} \widetilde{\phi}^{j-k}\widetilde{\theta}_k\Big)\delta_{n-j} + \sum_{j=q}^{\infty} \widetilde{\phi}^{j-q}\Big(\sum_{k=0}^{q} \widetilde{\phi}^{q-k}\,\widetilde{\theta}_k\Big)\delta_{n-j}, \end{aligned}$$

where $\widetilde{\phi}^0$ denotes the identity operator in H^p and Y_n, δ_n, $\widetilde{\phi}$ and $\widetilde{\theta}_1,\dots,\widetilde{\theta}_q$ are defined in (1.3.6)*. Furthermore, the series converges in L^2_H and with probability one.*

1.3.2 The vector ARMA(p,q) process

We project the stationary functional ARMA(p,q) process $(X_n)_{n\in\mathbb{Z}}$ on a finite-dimensional subspace of H. We fix $d \in \mathbb{N}$ and consider the projection of $(X_n)_{n\in\mathbb{Z}}$ onto the subspace $\overline{\text{sp}}\{\nu_1, \dots, \nu_d\}$ spanned by the d most important eigenfunctions of C_X giving

$$X_{n,d} = P_{\overline{\text{sp}}\{\nu_1,\dots,\nu_d\}} X_n = \sum_{i=1}^{d} \langle X_n, \nu_i \rangle \nu_i. \tag{1.3.9}$$

Remark 1.3.9. The dimension reduction based on the principal components is optimal for uncorrelated data in terms of its L^2-accuracy (cf. Horvàth and Kokoszka [27], Section 3.2). We consider time series data, where dimensions corresponding to eigenfunctions ν_l for $l > d$ can have an impact on subsequent elements of the time series, even if the corresponding eigenvalue λ_l is small. Hence FPCA might not be optimal for functional time series.

In Hörmann et al. [25] and Panaretos and Tavakoli [45] an optimal dimension reduction for dependent data is introduced. They propose a filtering technique based on a frequency domain approach, which reduces the dimension in such a way that the score vectors form a multivariate time series with diagonal lagged covariance matrices. However, as pointed out in Aue et al. [5], it is unclear how the technique can be utilized for prediction, since both future and past observations are required.

In order not to miss information valuable for prediction when reducing the dimension, we include cross validation on the prediction errors to choose the number of FPCs used to represent the data (see Section 5). This also allows us to derive explicit bounds for the prediction error in terms of the eigenvalues of C_X (see Section 4). □

In what follows we are interested in

$$\mathbf{X}_n := (\langle X_n, \nu_1 \rangle, \dots, \langle X_n, \nu_d \rangle)^\top . \tag{1.3.10}$$

$\mathbf{X}_n$ is d-dimensional and isometrically isomorph to $X_{n,d}$ (e.g. Hsing and Eubank [29], Theorem 2.4.17).

Remark 1.3.10. For theoretical considerations of the prediction problem we assume that C_X and its eigenfunctions are known. In a statistical data analysis the eigenfunctions have to be replaced by their empirical counterparts. In order to ensure consistency of the estimators we need slightly stronger assumptions on the innovation process $(\varepsilon_n)_{n\in\mathbb{Z}}$ and on the model parameters, similarly as for estimation and prediction in classical time series models (see Brockwell and Davis [13]).

In Hörmann and Kokoszka [24] it is shown that, under $L^4 - m$ approximability (a weak dependence concept for functional processes), empirical estimators of mean and covariance of the functional process are $\sqrt{n}$-consistent. Estimated eigenfunctions and eigenvalues inherit $\sqrt{n}$-consistency results from the estimated covariance operator

(Theorem 3.2 in Hörmann and Kokoszka [24]). Proposition 2.1 of Hörmann and Kokoszka [24] states conditions on the parameters of a linear process to ensure that the time series is $L^4 - m$ approximable, which are satisfied for stationary functional ARMA processes, where the WN has a finite 4-th moment. □

Our next result, which follows from the linearity of the projection operator, concerns the projection of the WN $(\varepsilon_n)_{n\in\mathbb{Z}}$ on $\overline{\text{sp}}\{\nu_1,\dots,\nu_d\}$.

Lemma 1.3.11. *Let $(e_i)_{i\in\mathbb{N}}$ be an arbitrary ONB of H. For $d\in\mathbb{N}$ we define the d-dimensional vector process*

$$\mathbf{Z}_n := (\langle\varepsilon_n, e_1\rangle, \dots, \langle\varepsilon_n, e_d\rangle)^\top, \quad n\in\mathbb{Z}.$$

(i) If $(\varepsilon_n)_{n\in\mathbb{Z}}$ is WN *as in Definition 1.3.1(i), then $(\mathbf{Z}_n)_{n\in\mathbb{Z}}$ is* WN *in $\mathbb{R}^d$.*
(ii) If $(\varepsilon_n)_{n\in\mathbb{Z}}$ is SWN *as in Definition 1.3.1(ii), then $(\mathbf{Z}_n)_{n\in\mathbb{Z}}$ is* SWN *in $\mathbb{R}^d$.*

As in Section 1.3.1 we start with the functional ARMA$(1,q)$ process for $q\in\mathbb{N}$ and are interested in the dynamics of $(X_{n,d})_{n\in\mathbb{Z}}$ of (1.3.9) for fixed $d\in\mathbb{N}$. Using the model equation (1.3.1) with $p=1$ and $\phi_1=\phi$, we get

$$\langle X_n, \nu_l\rangle = \langle\phi X_{n-1}, \nu_l\rangle + \sum_{j=0}^{q}\langle\theta_j\varepsilon_{n-j}, \nu_l\rangle, \quad l\in\mathbb{Z}. \tag{1.3.11}$$

For every l we expand $\langle\phi X_{n-1},\nu_l\rangle$, using that $(\nu_l)_{l\in\mathbb{N}}$ is an ONB of H as

$$\langle\phi X_{n-1},\nu_l\rangle = \Big\langle\phi\Big(\sum_{l'=1}^{\infty}\langle X_{n-1},\nu_{l'}\rangle\nu_{l'}\Big),\nu_l\Big\rangle = \sum_{l'=1}^{\infty}\langle\phi\nu_{l'},\nu_l\rangle\langle X_{n-1},\nu_{l'}\rangle,$$

and $\langle\theta_j\varepsilon_{n-j},\nu_l\rangle$ for $j=1,\dots,q$ as

$$\langle\theta_j\varepsilon_{n-j},\nu_l\rangle = \Big\langle\theta_j\Big(\sum_{l'=1}^{\infty}\langle\varepsilon_{n-j},\nu_{l'}\rangle\nu_{l'}\Big),\nu_l\Big\rangle = \sum_{l'=1}^{\infty}\langle\theta_j\nu_{l'},\nu_l\rangle\langle\varepsilon_{n-j},\nu_{l'}\rangle.$$

In order to study the d-dimensional vector process $(\mathbf{X}_n)_{n\in\mathbb{Z}}$, for notational ease, we restrict a precise presentation to the ARMA$(1,1)$ model. The presentation of the ARMA$(1,q)$ model is an obvious extension.

For a matrix representation of $\mathbf{X}_n$ given in (1.3.10) consider the notation:

$$\left(\begin{array}{c|c} \mathbf{\Phi} & \mathbf{\Phi}^\infty \\ \hline \vdots & \vdots \end{array}\right) = \left(\begin{array}{ccc|cc} \langle \phi\nu_1, \nu_1\rangle & \dots & \langle \phi\nu_d, \nu_1\rangle & \langle \phi\nu_{d+1}, \nu_1\rangle & \dots \\ \vdots & \ddots & \vdots & \vdots & \ddots \\ \langle \phi\nu_1, \nu_d\rangle & \dots & \langle \phi\nu_d, \nu_d\rangle & \langle \phi\nu_{d+1}, \nu_d\rangle & \dots \\ \hline \langle \phi\nu_1, \nu_{d+1}\rangle & \dots & \langle \nu_d, \nu_{d+1}\rangle & \langle \phi\nu_{d+1}, \nu_{d+1}\rangle & \dots \\ \vdots & \ddots & \vdots & \vdots & \ddots \end{array}\right).$$

The matrices $\mathbf{\Theta}$ and $\mathbf{\Theta}^\infty$ are defined analogously. For $q = 1$, with $\theta_0 = I$ and $\theta_1 = \theta$, (1.3.11) is given in matrix form by

$$\left(\begin{array}{c} \mathbf{X}_n \\ \hline \mathbf{X}_n^\infty \end{array}\right) = \left[\begin{array}{c|c} \mathbf{\Phi} & \mathbf{\Phi}^\infty \\ \hline \vdots & \vdots \end{array}\right] \left(\begin{array}{c} \mathbf{X}_{n-1} \\ \hline \mathbf{X}_{n-1}^\infty \end{array}\right) + \left(\begin{array}{c} \mathbf{E}_n \\ \hline \mathbf{E}_n^\infty \end{array}\right) + \left[\begin{array}{c|c} \mathbf{\Theta} & \mathbf{\Theta}^\infty \\ \hline \vdots & \vdots \end{array}\right] \left(\begin{array}{c} \mathbf{E}_{n-1} \\ \hline \mathbf{E}_{n-1}^\infty \end{array}\right), \tag{1.3.12}$$

where

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

The operators $\mathbf{\Phi}$ and $\mathbf{\Theta}$ in (1.3.12) are $d \times d$ matrices with entries $\langle \phi\nu_{l'}, \nu_l\rangle$ and $\langle \theta\nu_{l'}, \nu_l\rangle$ in the l-th row and l'-th column, respectively. Furthermore, $\mathbf{\Phi}^\infty$ and $\mathbf{\Theta}^\infty$ are $d \times \infty$ matrices with ll'-th entries $\langle \phi\nu_{l'+d}, \nu_l\rangle$ and $\langle \theta\nu_{l'+d}, \nu_l\rangle$, respectively.

By (1.3.12), $(\mathbf{X}_n)_{n\in\mathbb{Z}}$ satisfies the d-dimensional vector equation

$$\mathbf{X}_n = \mathbf{\Phi}\mathbf{X}_{n-1} + \mathbf{E}_n + \mathbf{\Theta}\mathbf{E}_{n-1} + \mathbf{\Delta}_{n-1}, \quad n \in \mathbb{Z}, \tag{1.3.13}$$

where

$$\mathbf{\Delta}_{n-1} := \mathbf{\Phi}^\infty \mathbf{X}_{n-1}^\infty + \mathbf{\Theta}^\infty \mathbf{E}_{n-1}^\infty. \tag{1.3.14}$$

By Lemma 1.3.11, $(\mathbf{E}_n)_{n\in\mathbb{Z}}$ is d-dimensional WN. Note that $\mathbf{\Delta}_{n-1}$ in (1.3.14) is a d-dimensional vector with l-th component

$$(\mathbf{\Delta}_{n-1})_l = \sum_{l'=d+1}^{\infty} \langle \phi\nu_{l'}, \nu_l\rangle \langle X_{n-1}, \nu_{l'}\rangle + \sum_{l'=d+1}^{\infty} \langle \theta\nu_{l'}, \nu_l\rangle \langle \varepsilon_{n-1}, \nu_{l'}\rangle. \tag{1.3.15}$$

Thus, the "error term" $\mathbf{\Delta}_{n-1}$ depends on X_{n-1}, and the vector process $(\mathbf{X}_n)_{n\in\mathbb{Z}}$ in (1.3.13) is in general not a vector ARMA$(1,1)$ process with innovations $(\mathbf{E}_n)_{n\in\mathbb{Z}}$. However, we can use a vector ARMA model as an approximation to $(\mathbf{X}_n)_{n\in\mathbb{Z}}$, where we can make $\mathbf{\Delta}_{n-1}$ arbitrarily small by increasing the dimension d.

Lemma 1.3.12. *Let $\|\cdot\|_2$ denote the Euclidean norm in $\mathbb{R}^d$, and let the d-dimensional vector $\mathbf{\Delta}_{n-1}$ be defined as in (1.3.14). Then $E\|\mathbf{\Delta}_{n-1}\|_2^2$ is bounded and tends to 0 as $d \to \infty$.*

Proof. From (1.3.14) we obtain

$$E\|\mathbf{\Delta}_{n-1}\|_2^2 \le 2\left(E\|\mathbf{\Phi}^\infty \mathbf{X}_{n-1}^\infty\|_2^2 + E\|\mathbf{\Theta}^\infty \mathbf{E}_{n-1}^\infty\|_2^2\right). \tag{1.3.16}$$

We estimate the two parts $E\|\mathbf{\Phi}^\infty \mathbf{X}_{n-1}^\infty\|_2^2$ and $E\|\mathbf{\Theta}^\infty \mathbf{E}_{n-1}^\infty\|_2^2$ separately. By (1.3.15) we obtain (applying Parseval's equality (1.2.1) in the third line),

$$\begin{aligned} E\|\mathbf{\Phi}^\infty \mathbf{X}_{n-1}^\infty\|_2^2 &= E\Big[\sum_{l=1}^{d}\Big(\sum_{l'=d+1}^{\infty}\langle\langle X_{n-1},\nu_{l'}\rangle\phi\nu_{l'},\nu_l\rangle\Big)^2\Big] \\ &\le E\Big[\sum_{l=1}^{\infty}\Big\langle\sum_{l'=d+1}^{\infty}\langle X_{n-1},\nu_{l'}\rangle\phi\nu_{l'},\nu_l\Big\rangle^2\Big] \\ &= E\Big\|\sum_{l'=d+1}^{\infty}\langle X_{n-1},\nu_{l'}\rangle\phi\nu_{l'}\Big\|^2. \end{aligned}$$

Since the scores $(\langle X_{n-1,l},\nu_l\rangle)_{l\in\mathbb{N}}$ are uncorrelated (cf. the Karhunen-Loève Theorem 1.2.2), and then using monotone convergence, we find

$$E\|\mathbf{\Phi}^\infty \mathbf{X}_{n-1}^\infty\|_2^2 \le E\sum_{l'=d+1}^{\infty}\langle X_{n-1},\nu_{l'}\rangle^2\|\phi\nu_{l'}\|^2 = \sum_{l'=d+1}^{\infty}E\left(\langle X_{n-1},\nu_{l'}\rangle\right)^2\|\phi\nu_{l'}\|^2.$$

Since by (1.2.4) $E\langle X_{n-1},\nu_{l'}\rangle^2 = \lambda_{l'}$, we get

$$\sum_{l'=d+1}^{\infty}E\left(\langle X_{n-1},\nu_{l'}\rangle\right)^2\|\phi\nu_{l'}\|^2 = \sum_{l'=d+1}^{\infty}\lambda_{l'}\|\phi\|_{\mathcal{L}}^2\|\nu_{l'}\|^2 \le \|\phi\|_{\mathcal{L}}^2\sum_{l'=d+1}^{\infty}\lambda_{l'}. \tag{1.3.17}$$

The bound for $E\|\mathbf{\Theta}^\infty \mathbf{E}_{n-1}^\infty\|_2^2$ can be obtained in exactly the same way, and we calculate

$$\begin{aligned} E\|\mathbf{\Theta}^\infty \mathbf{E}_{n-1}^\infty\|_2^2 &\le \sum_{l'=d+1}^{\infty}E\langle\varepsilon_{n-1},\nu_{l'}\rangle^2\|\theta\nu_{l'}\|^2 \\ &\le \|\theta\|_{\mathcal{L}}^2\sum_{l'=d+1}^{\infty}E\langle\langle\varepsilon_{n-1},\nu_{l'}\rangle\varepsilon_{n-1},\nu_{l'}\rangle \\ &= \|\theta\|_{\mathcal{L}}^2\sum_{l'=d+1}^{\infty}\langle C_\varepsilon\nu_{l'},\nu_{l'}\rangle, \end{aligned} \tag{1.3.18}$$

where C_ε is the covariance operator of the WN. As a covariance operator it has finite nuclear operator norm $\|C_\varepsilon\|_{\mathcal{N}} := \sum_{l'=1}^{\infty}\langle C_\varepsilon(\nu_{l'}), \nu_{l'}\rangle < \infty$. Hence, $\sum_{l'=d+1}^{\infty}\langle C_\varepsilon\nu_{l'}, \nu_{l'}\rangle \to 0$ for $d \to \infty$. Combining (1.3.16), (1.3.17) and (1.3.18) we find that $E\|\mathbf{\Delta}_{n-1}\|_2^2$ is bounded and tends to 0 as $d \to \infty$. □

For the vector ARMA$(1,q)$ model the proof of boundedness of $E\|\mathbf{\Delta}_{n-1}\|_2^2$ is analogous. We now summarize our findings for a functional ARMA$(1,q)$ process.

Theorem 1.3.13. *Consider a functional* ARMA$(1,q)$ *process for* $q \in \mathbb{N}$ *such that Assumption 1.3.3 holds. For* $d \in \mathbb{N}$*, the vector process of* (1.3.10) *has the representation*

$$\mathbf{X}_n = \mathbf{\Phi}\mathbf{X}_{n-1} + \mathbf{E}_n + \sum_{j=1}^{q}\mathbf{\Theta}_q\mathbf{E}_{n-j} + \mathbf{\Delta}_{n-1}, \quad n \in \mathbb{Z},$$

where

$$\mathbf{\Delta}_{n-1} := \mathbf{\Phi}^\infty\mathbf{X}_{n-1}^\infty + \sum_{j=1}^{q}\mathbf{\Theta}_j^\infty\mathbf{E}_{n-j},$$

and all quantities are defined analogously to (1.3.10), (1.3.13), *and* (1.3.14). *Define*

$$\check{\mathbf{X}}_n = \mathbf{\Phi}\check{\mathbf{X}}_{n-1} + \mathbf{E}_n + \sum_{j=1}^{q}\mathbf{\Theta}_j\mathbf{E}_{n-j}, \quad n \in \mathbb{Z}. \tag{1.3.19}$$

Then both the functional ARMA$(1,q)$ *process* $(X_n)_{n\in\mathbb{Z}}$ *in* (1.3.1) *and the* d*-dimensional vector process* $(\check{\mathbf{X}}_n)_{n\in\mathbb{Z}}$ *in* (1.3.19) *have a unique stationary and causal solution. Moreover,* $E\|\mathbf{\Delta}_{n-1}\|_2^2$ *is bounded and tends to 0 as* $d \to \infty$.

Proof. Recall from (1.3.12) the $d\times d$ matrix $\mathbf{\Phi}$ of the vector process (1.3.19). In order to show that (1.3.19) has a stationary solution, by Theorem 11.3.1 of Brockwell and Davis [13], it suffices to prove that every eigenvalue λ_k of $\mathbf{\Phi}$ with corresponding eigenvector $\mathbf{a}_k = (\mathbf{a}_{k,1}, \dots, \mathbf{a}_{k,d})$ satisfies $|\lambda_k| < 1$ for $k = 1, \dots, d$. Note that $|\lambda_k| < 1$ is equivalent to $|\lambda_k^{j_0}| < 1$ for all $j_0 \in \mathbb{N}$. Define $a_k := \mathbf{a}_{k,1}\nu_1 + \dots + \mathbf{a}_{k,d}\nu_d \in H$, then by Parseval's equality (1.2.1), $\|a_k\|^2 = \sum_{l=1}^{d}|\langle a_k, \nu_l\rangle|^2 = \sum_{l=1}^{d}\mathbf{a}_{k,l}^2 = \|\mathbf{a}_k\|_2^2 = 1$ for $k = 1, \dots, d$. With the orthogonality of $\nu_1, \dots, \nu_d$ we find $\|\mathbf{\Phi}\mathbf{a}_k\|_2^2 = \sum_{l=1}^{d}\big(\sum_{l'=1}^{d}\langle\phi\nu_{l'}, \nu_l\rangle\mathbf{a}_{k,l}\big)^2$. Defining $A_d = \overline{\mathrm{sp}}\{\nu_1, \dots, \nu_d\}$, we calculate

$$\|P_{A_d}\phi P_{A_d}a_k\|^2 = \sum_{l=1}^{d}\Big\langle\phi\Big(\sum_{l'=1}^{d}\mathbf{a}_{k,l'}\nu_{l'}\Big), \nu_l\Big\rangle^2\|\nu_l\|^2 = \sum_{l=1}^{d}\Big(\sum_{l'=1}^{d}\mathbf{a}_{k,l'}\langle\phi\nu_{l'}, \nu_l\rangle\Big)^2 = \|\Phi\mathbf{a}_k\|_2^2.$$

Hence, for j_0 as in Assumption 1.3.3,

$$|\lambda_k^{j_0}| = \|\lambda_k^{j_0}\mathbf{a}_k\|_2 = \|\mathbf{\Phi}^{j_0}\mathbf{a}_k\|_2 = \left\|\left(P_{A_d}\phi P_{A_d}\right)^{j_0} a_k\right\| \\ \leq \left\|\left(P_{A_d}\phi P_{A_d}\right)^{j_0}\right\|_{\mathcal{L}}\|a_k\| \leq \|\phi^{j_0}\|_{\mathcal{L}} < 1,$$

which finishes the proof. □

In order to extend approximation (1.3.19) of a functional ARMA$(1,q)$ process to a functional ARMA(p,q) process we use again the state space representation (1.3.7) given by

$$Y_n = \widetilde{\phi} Y_{n-1} + \sum_{j=0}^{q} \widetilde{\theta}_j \delta_{n-j}, \quad n \in \mathbb{Z},$$

where Y_n, $\widetilde{\theta}_0 = I$, $\widetilde{\phi}$, $\widetilde{\theta}_1, \dots, \widetilde{\theta}_q$ and δ_n are defined as in Theorem 1.3.8 and take values in $H_p = (L^2([0,1]))^p$; cf. (1.3.8).

Theorem 1.3.14. *Consider the functional* ARMA(p,q) *process as defined in* (1.3.1) *such that Assumption 1.3.7 holds. Then for $d \in \mathbb{N}$ the vector process of* (1.3.10) *has the representation*

$$\mathbf{X}_n = \sum_{i=1}^{p} \mathbf{\Phi}_i \mathbf{X}_{n-i} + \mathbf{E}_n + \sum_{j=1}^{q} \mathbf{\Theta}_j \mathbf{E}_{n-j} + \mathbf{\Delta}_{n-1}, \quad n \in \mathbb{Z}, \tag{1.3.20}$$

where

$$\mathbf{\Delta}_{n-1} := \sum_{i=1}^{p} \mathbf{\Phi}_i^{\infty} \mathbf{X}_{n-i}^{\infty} + \sum_{j=1}^{q} \mathbf{\Theta}_j^{\infty} \mathbf{E}_{n-j},$$

and all quantities are defined analogously to (1.3.10), (1.3.13), *and* (1.3.14). *Define*

$$\check{\mathbf{X}}_n = \sum_{i=1}^{p} \mathbf{\Phi}_i \check{\mathbf{X}}_{n-i} + \mathbf{E}_n + \sum_{j=1}^{q} \mathbf{\Theta}_j \mathbf{E}_{n-j}, \quad n \in \mathbb{Z}. \tag{1.3.21}$$

Then both the functional ARMA(p,q) *process* $(X_n)_{n\in\mathbb{Z}}$ *in* (1.3.1) *and the d-dimensional vector process* $(\check{\mathbf{X}}_n)_{n\in\mathbb{Z}}$ *in* (1.3.21) *have a unique stationary and causal solution. Moreover, $E\|\mathbf{\Delta}_{n-1}\|_2^2$ is bounded and tends to 0 as $d \to \infty$.*

We are now interested in conditions for $(\mathbf{X}_n)_{n\in\mathbb{Z}}$ to exactly follow a vector ARMA(p,q) model. A trivial condition is that the projections of ϕ_i and θ_j onto $A_d^{\perp}$, the orthogonal complement of $A_d = \overline{\mathrm{sp}}\{\nu_1, \dots, \nu_d\}$, satisfy

$$P_{A_d^{\perp}}\phi_i P_{A_d^{\perp}} = P_{A_d^{\perp}}\theta_j P_{A_d^{\perp}} = 0$$

for all $i = 1, \dots, p$ and $j = 1, \dots, q$. In that case $\check{\mathbf{X}}_n = \mathbf{X}_n$ for all $n \in \mathbb{Z}$.

However, as we show next, the assumptions on the moving average parameters $\theta_1, \dots, \theta_q$ are actually not required. We start with a well-known result that characterizes vector MA processes.

Lemma 1.3.15 (Brockwell and Davis [13], Proposition 3.2.1)**.** *If $(\mathbf{X}_n)_{n\in\mathbb{Z}}$ is a stationary vector process with autocovariance matrix $\mathbf{C}_{\mathbf{X}_h,\mathbf{X}_0} = E[\mathbf{X}_h \mathbf{X}_0^\top]$ with $\mathbf{C}_{\mathbf{X}_q,\mathbf{X}_0} \neq 0$ and $\mathbf{C}_{\mathbf{X}_h,\mathbf{X}_0} = 0$ for $|h| > q$, then $(\mathbf{X}_n)_{n\in\mathbb{Z}}$ is a vector* MA(q).

Proposition 1.3.16. *Let $A_d = \overline{\text{sp}}\{\nu_1, \dots, \nu_d\}$ and $A_d^\perp$ its orthogonal complement. If $P_{A_d^\perp}\phi_i P_{A_d^\perp} = 0$ for all $i = 1, \dots, p$, then the d-dimensional process $(\mathbf{X}_n)_{n\in\mathbb{Z}}$ as in (1.3.20) is a vector* ARMA(p, q) *process.*

Proof. Since ϕ_i for $i = 1, \dots, p$ only acts on A_d, from (1.3.20) we get

$$\begin{aligned} \mathbf{X}_n &= \sum_{i=1}^{p} \mathbf{\Phi}_i \mathbf{X}_{n-i} + \mathbf{E}_n + \sum_{j=1}^{q} \mathbf{\Theta}_j \mathbf{E}_{n-j} + \mathbf{\Delta}_{n-1} \\ &= \sum_{i=1}^{p} \mathbf{\Phi}_i \mathbf{X}_{n-i} + \mathbf{E}_n + \sum_{j=1}^{q} \mathbf{\Theta}_j \mathbf{E}_{n-j} + \sum_{j=1}^{q} \mathbf{\Theta}_j^\infty \mathbf{E}_{n-j}^\infty, \quad n \in \mathbb{Z}. \end{aligned}$$

To ensure that $(\mathbf{X}_n)_{n\in\mathbb{Z}}$ follows a vector ARMA(p, q) process, we have to show that

$$\mathbf{R}_n := \mathbf{E}_n + \sum_{j=1}^{q} \mathbf{\Theta}_j \mathbf{E}_{n-j} + \sum_{j=1}^{q} \mathbf{\Theta}_j^\infty \mathbf{E}_{n-j}^\infty, \quad n \in \mathbb{Z},$$

follows a vector MA(q) model. According to Lemma 1.3.15 it is sufficient to verify that $(\mathbf{R}_n)_{n\in\mathbb{Z}}$ is stationary and has an appropriate autocovariance structure.

Defining (with $\theta_0 = I$)

$$R_n := \sum_{j=0}^{q} \theta_j \varepsilon_{n-j}, \quad n \in \mathbb{Z},$$

where $\theta_1, \dots, \theta_q$ are as in (1.3.1), observe that $\mathbf{R}_n = (\langle R_n, \nu_1 \rangle, \dots, \langle R_n, \nu_d \rangle)$ is isometrically isomorph to $P_{A_d} R_n = \sum_{j=1}^{d} \langle R_n, \nu_j \rangle \nu_j$ for all $n \in \mathbb{Z}$. Hence, stationarity of $(\mathbf{R}_n)_{n\in\mathbb{Z}}$ immediately follows from the stationarity of $(R_n)_{n\in\mathbb{Z}}$. Furthermore,

$$E[\langle P_{A_d} R_0, \cdot \rangle P_{A_d} R_h] = P_{A_d} E[\langle R_0, \cdot \rangle R_h] P_{A_d} = P_{A_d} C_{R_h, R_0} P_{A_d}.$$

But since $(R_n)_{n\in\mathbb{Z}}$ is a functional MA(q) process, $C_{R_h,R_0} = 0$ for $|h| > q$. By the relation between $P_{A_d} R_n$ and $\mathbf{R}_n$ we also have $\mathbf{C}_{\mathbf{R}_h,\mathbf{R}_0} = 0$ for $|h| > q$ and, hence, $(\mathbf{R}_n)_{n\in\mathbb{Z}}$ is a vector MA(q). □

1.4 Prediction of functional ARMA processes

For $h \in \mathbb{N}$ we derive the best h-step linear predictor of a functional ARMA(p,q) process $(X_n)_{n\in\mathbb{Z}}$ based on $\mathbf{X}_1,\dots,\mathbf{X}_n$ as defined in (1.3.20). We then compare the vector best linear predictor to the functional best linear predictor based on $X_1,\dots,X_n$ and show that, under regularity conditions, the difference is bounded and tends to 0 as d tends to infinity.

1.4.1 Prediction based on the vector process

In finite dimensions the concept of a best linear predictor is well-studied. For a d-dimensional stationary time series $(\mathbf{X}_n)_{n\in\mathbb{Z}}$ we denote the *matrix linear span* of $\mathbf{X}_1,\dots,\mathbf{X}_n$ by

$$\mathbf{M}_1' := \Big\{ \sum_{i=1}^{n} \mathbf{A}_{ni}\mathbf{X}_i : \ \mathbf{A}_{ni} \text{ are real } d\times d \text{ matrices}, i=1,\dots,n \Big\}.$$

Then for $h \in \mathbb{N}$ the *h-step vector best linear predictor* $\widehat{\mathbf{X}}_{n+h}$ of $\mathbf{X}_{n+h}$ based on $\mathbf{X}_1,\dots,\mathbf{X}_n$ is defined as the projection of $\mathbf{X}_{n+h}$ onto the closure $\mathbf{M}_1$ of $\mathbf{M}_1'$ in $L^2_{\mathbb{R}^d}$; i.e.,

$$\widehat{\mathbf{X}}_{n+h} := P_{\mathbf{M}_1}\mathbf{X}_{n+h}. \tag{1.4.1}$$

Its properties are given by the projection theorem (e.g. Theorem 2.3.1 of Brockwell and Davis [13]) and can be summarized as follows.

Remark 1.4.1. Recall that $\|\cdot\|_2$ denotes the Euclidean norm in $\mathbb{R}^d$ and $\langle\, , \rangle_{\mathbb{R}^d}$ the corresponding scalar product.
(i) $E\langle \mathbf{X}_{n+h} - \widehat{\mathbf{X}}_{n+h}, \mathbf{Y}\rangle_{\mathbb{R}^d} = \mathbf{0}$ for all $\mathbf{Y} \in \mathbf{M}_1$.
(ii) $\widehat{\mathbf{X}}_{n+h}$ is the unique element in $\mathbf{M}_1$ such that

$$E\|\mathbf{X}_{n+h} - \widehat{\mathbf{X}}_{n+h}\|_2^2 = \inf_{\mathbf{Y}\in\mathbf{M}_1} E\|\mathbf{X}_{n+h} - \mathbf{Y}\|_2^2.$$

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

In analogy to the prediction algorithm suggested in Aue et al. [5], a method for finding the best linear predictor of X_{n+h} based on $\mathbf{X}_1,\dots,\mathbf{X}_n$ is the following:

Algorithm 1 [1]

(1) Fix $d \in \mathbb{N}$. Compute the FPC scores $\langle X_k, \nu_l \rangle$ for $l = 1, \dots, d$ and $k = 1, \dots, n$ by projecting each X_k on $\nu_1, \dots, \nu_d$. Summarize the scores in the vector

$$\mathbf{X}_k := (\langle X_k, \nu_1 \rangle, \dots, \langle X_k, \nu_d \rangle)^\top, \quad k = 1, \dots n.$$

(2) Consider the d-dimensional vectors $\mathbf{X}_1, \dots, \mathbf{X}_n$. For $h \in \mathbb{N}$ compute the vector best linear predictor of $\mathbf{X}_{n+h}$ by means of (1.4.1):

$$\widehat{\mathbf{X}}_{n+h} = (\langle \widehat{X_{n+h}, \nu_1} \rangle, \dots, \langle \widehat{X_{n+h}, \nu_d} \rangle)^\top.$$

(3) Re-transform the vector best linear predictor $\widehat{\mathbf{X}}_{n+h}$ into a functional form $\widehat{X}_{n+h}$ via the truncated Karhunen-Loève representation:

$$\widehat{X}_{n+h} := \langle \widehat{X_{n+h}, \nu_1} \rangle \nu_1 + \dots + \langle \widehat{X_{n+h}, \nu_d} \rangle \nu_d. \tag{1.4.2}$$

For functional AR(p) processes, Aue et al. [5] compare the resulting predictor (1.4.2) to the functional best linear predictor. Our goal is to extend these results to functional ARMA(p, q) processes. However, when moving away from AR models, the best linear predictor is no longer directly given by the process. We start by recalling the notion of best linear predictors in Hilbert spaces.

1.4.2 Functional best linear predictor

For $h \in \mathbb{N}$ we introduce the *h-step functional best linear predictor* $\widehat{X}_{n+h}$ of X_{n+h}, based on $X_1, \dots, X_n$, as proposed in Bosq [10]. It is the projection of X_{n+h} onto a large enough subspace of L^2_H containing $X_1, \dots, X_n$. More formally, we use the concept of *$\mathcal{L}$-closed subspaces* as in Definition 1.1 of Bosq [8].

Definition 1.4.2. Recall that $\mathcal{L}$ denotes the space of bounded linear operators acting on H. We call G an *$\mathcal{L}$-closed subspace* (LCS) of L^2_H, if
(1) G is a Hilbertian subspace of L^2_H.
(2) If $X \in G$ and $g \in \mathcal{L}$, then $g X \in G$. □

[1] Steps (1) and (3) are implemented in the `R` package `fda`, and (2) in the `R` package `mts`

We define

$$X^{(n)} := (X_n, \dots, X_1).$$

By Theorem 1.8 of Bosq [8] the LCS $G := G_{X^{(n)}}$ generated by $X^{(n)}$ is the closure in $L^2_{H^n}$ of $G'_{X^{(n)}}$, where

$$G'_{X^{(n)}} := \left\{ g_n X^{(n)} : \; g_n \in \mathcal{L}(H^n, H) \right\}.$$

For $h \in \mathbb{N}$ the *h-step functional best linear predictor* $\widehat{X}^G_{n+h}$ of X_{n+h} is defined as the projection of X_{n+h} onto G, which we write as

$$\widehat{X}^G_{n+h} := P_G X_{n+h} \in G. \tag{1.4.3}$$

Its properties are given by the projection theorem (e.g. Section 1.6 in Bosq [8]) and are summarized as follows.

Remark 1.4.3. (i) $E\langle X_{n+h} - \widehat{X}^G_{n+h}, Y\rangle = 0$ for all $Y \in G$.
(ii) $\widehat{X}^G_{n+h}$ is the unique element in G such that

$$E\|X_{n+h} - \widehat{X}^G_{n+h}\|^2 = \inf_{Y \in G} E\|X_{n+h} - Y\|^2.$$

(iii) The mean squared error of the functional best linear predictor $\widehat{X}^G_{n+h}$ is denoted by

$$\sigma^2_{n,h} := E\|X_{n+h} - \widehat{X}^G_{n+h}\|^2. \tag{1.4.4}$$

□

Since in general $G'_{X(n)}$ is not closed (cf. Bosq [10], Proposition 2.1), $\widehat{X}^G_{n+h}$ is not necessarily of the form $\widehat{X}^G_{n+h} = g_n^{(h)} X^{(n)}$ for some $g_n^{(h)} \in \mathcal{L}(H^n, H)$. However, the following result gives necessary and sufficient conditions for $\widehat{X}^G_{n+h}$ to be represented in terms of bounded linear operators.

Proposition 1.4.4 (Proposition 2.2, Bosq [10])**.** *For $h \in \mathbb{N}$ the following are equivalent:*
(i) There exists some $g \in \mathcal{L}(H^n, H)$ such that $C_{X^{(n)}, X_{n+h}} = g\, C_{X^{(n)}}$.
(ii) $P_G X_{n+h} = g\, X^{(n)}$ for some $g \in \mathcal{L}(H^n, H)$.

This result allows us to derive conditions, such that the difference between the predictors (1.4.1) and (1.4.3) can be computed. Weaker conditions are needed, if $\widehat{X}^G_{n+h}$ admits a representation $\widehat{X}^G_{n+h} = s_n^{(h)} X^{(n)}$ for some Hilbert-Schmidt operator $s_n^{(h)}$ from H^n to H ($s_n^{(h)} \in \mathcal{S}(H^n, H)$).

Proposition 1.4.5. *For $h \in \mathbb{N}$ the following are equivalent:*
(i) There exists some $s \in \mathcal{S}(H^n, H)$ such that $C_{X^{(n)},X_{n+h}} = s\, C_{X^{(n)}}$.
(ii) $P_G X_{n+h} = s\, X^{(n)}$ for some $s \in \mathcal{S}(H^n, H)$.

Proof. The proof is similar to the proof of Proposition 1.4.4. Assume that (i) holds. Then, since $C_{X^{(n)},s\,X^{(n)}} = E[\langle X^{(n)}, \cdot\rangle\, s\, X^{(n)}] = s\, C_{X^{(n)}}$, we have

$$C_{X^{(n)},X_{n+h}-s\,X^{(n)}} = 0.$$

Therefore, $X_{n+h} - s\, X^{(n)} \perp X^{(n)}$ and, hence, $X_{n+h} - s\, X^{(n)} \perp G$ which gives (ii).
For the reverse, note that (ii) implies

$$C_{X^{(n)},X_{n+h}-s\,X^{(n)}} = C_{X^{(n)},X_{n+h}-P_G X_{n+h}} = 0.$$

Thus, $C_{X^{(n)},X_{n+h}} = C_{X^{(n)},s\,X^{(n)}} = s\, C_{X^{(n)}}$, which finishes the proof. □

We proceed with examples of processes where Proposition 1.4.4 or Proposition 1.4.5 apply.

Example 1.4.6. *Let $(X_n)_{n\in\mathbb{Z}}$ be a stationary functional* AR(p) *process with representation*

$$X_n = \varepsilon_n + \sum_{j=1}^{p} \phi_j X_{n-j}, \quad n \in \mathbb{Z},$$

where $(\varepsilon_n)_{n\in\mathbb{Z}}$ is WN *and $\phi_j \in \mathcal{S}$ are Hilbert-Schmidt operators. Then for $n \geq p$, Proposition 1.4.5 applies for $h = 1$, giving the 1-step predictor $P_G X_{n+1} = s_n^{(1)} X^{(n)}$ for some $s_n^{(1)} \in \mathcal{S}$.*

Proof. We calculate

$$C_{X^{(n)},X_{n+1}}(\cdot) = E\Big[\langle X^{(n)}, \cdot\rangle (\phi_1, \dots, \phi_p, 0, \dots, 0) X^{(n)}\Big] = \phi C_{X^{(n)}}(\cdot),$$

where $\phi = (\phi_1, \dots, \phi_p, 0, \dots, 0) \in \mathcal{L}(H^n, H)$. Now let $(e_i)_{i\in\mathbb{N}}$ be an ONB of H. Then $(f_j)_{j\in\mathbb{N}}$ with $f_1 = (e_1, 0, \dots, 0)^\top$, $f_2 = (0, e_1, 0, \dots, 0)^\top$, … , $f_n = (0, \dots, 0, e_1)^\top$,

$f_{n+1} = (e_2, 0, \ldots, 0)^\top$, $f_{n+2} = (0, e_2, 0, \ldots, 0)^\top$, $\ldots$, $f_{2n} = (0, \ldots, 0, e_2)^\top$, $f_{2n+1} = (e_3, 0, \ldots, 0)^\top, \ldots$ is an ONB of H^n and, by orthogonality of $(e_i)_{i\in\mathbb{N}}$, we get

$$\|\phi\|_{\mathcal{S}}^2 = \sum_{j\in\mathbb{N}} \|\phi f_j\|^2 = \sum_{i\in\mathbb{N}} \sum_{j=1}^{p} \|\phi_j e_i\|^2 = \sum_{j=1}^{p} \sum_{i\in\mathbb{N}} \|\phi_j e_i\|^2 = \sum_{j=1}^{p} \|\phi_j\|_{\mathcal{L}}^2 < \infty,$$

since $\phi_j \in \mathcal{S}$ for every $j = 1, \ldots, p$, which implies that $\phi \in \mathcal{S}(H^n, H)$. □

Example 1.4.7. *Let* $(X_n)_{n\in\mathbb{Z}}$ *be a stationary functional* MA(1) *process*

$$X_n = \varepsilon_n + \theta\varepsilon_{n-1}, \quad n \in \mathbb{Z},$$

where $(\varepsilon_n)_{n\in\mathbb{Z}}$ *is* WN, $\|\theta\|_{\mathcal{L}} < 1$, $\theta \in \mathcal{S}$ *and* θ *nilpotent, such that* $\|\theta^j\|_{\mathcal{L}} = 0$ *for* $j > j_0$ *for some* $j_0 \in \mathbb{N}$. *Then for* $n > j_0$, *Proposition 1.4.5 applies.*

Proof. Since $\|\theta\|_{\mathcal{L}} < 1$, $(X_n)_{n\in\mathbb{Z}}$ is invertible, and since θ is nilpotent, $(X_n)_{n\in\mathbb{Z}}$ can be represented as an AR process of finite order, where the operators in the inverse representation are still Hilbert-Schmidt operators. Then the statement follows from the arguments of the proof of Example 1.4.6. □

Example 1.4.8. *Let* $(X_n)_{n\in\mathbb{Z}}$ *be a stationary functional* MA(1) *process*

$$X_n = \varepsilon_n + \theta\varepsilon_{n-1}, \quad n \in \mathbb{Z},$$

where $(\varepsilon_n)_{n\in\mathbb{Z}}$ *is* WN, *and denote by* C_ε *the covariance operator of the WN. Assume that* $\|\theta\|_{\mathcal{L}} < 1$. *If* θ *and* C_ε *commute, Proposition 1.4.5 applies.*

Proof. Stationarity of $(X_n)_{n\in\mathbb{Z}}$ ensures that $C_{X_n,X_{n+1}} = C_{X_0,X_1}$. Let θ^* denote the adjoint operator of θ. Since $\theta C_\varepsilon = C_\varepsilon\theta$, we have that $C_{X_1,X_0} = C_{X_0,X_1}$ which implies $\theta C_\varepsilon = C_\varepsilon\theta^* = C_\varepsilon\theta$. Hence, $C_\varepsilon = C_{X_0} - \theta C_\varepsilon\theta^* = C_{X_0} - \theta^2 C_\varepsilon$. Since $\|\theta\|_{\mathcal{L}} < 1$, the operator $I + \theta^2$ is invertible. Therefore, $C_\varepsilon = (I + \theta^2)^{-1} C_{X_0}$, and we get

$$C_{X_1,X_0} = \theta C_\varepsilon = (I + \theta^2)^{-1}\theta C_{X_0}.$$

Furthermore, since the space $\mathcal{S}$ of Hilbert-Schmidt operators forms an ideal in the space of bounded linear operators (e.g. Dunford and Schwartz [16], Theorem VI.5.4.) and $\theta \in \mathcal{S}$, also $(I + \theta^2)^{-1}\theta \in \mathcal{S}$. □

1.4.3 Bounds for the error of the vector predictor

We are now ready to derive bounds for the prediction error caused by the dimension reduction. More precisely, for $h \in \mathbb{N}$ we compare the vector best linear predictor $\widehat{X}_{n+h} = \sum_{j=1}^{d} \langle \widehat{X_{n+h}, \nu_j} \rangle \nu_j$ as defined in (1.4.2) with the functional best linear predictor $\widehat{X}_{n+h}^G = P_G X_{n+h}$ of (1.4.3). We first compare them on $\overline{\mathrm{sp}}\{\nu_1, \dots, \nu_d\}$, where the vector representations are given by

$$\begin{aligned}\widehat{\mathbf{X}}_{n+h} &= (\langle \widehat{X_{n+h}, \nu_1} \rangle, \dots, \langle \widehat{X_{n+h}, \nu_d} \rangle)^\top, \text{ and}\\ \widehat{\mathbf{X}}_{n+h}^G &:= \left(\left\langle \widehat{X}_{n+h}^G, \nu_1 \right\rangle, \dots, \left\langle \widehat{X}_{n+h}^G, \nu_d \right\rangle \right)^\top. \end{aligned} \tag{1.4.5}$$

We formulate assumptions such that for $d \to \infty$ the mean squared distance between the vector best linear predictor $\widehat{\mathbf{X}}_{n+h}$ and the vector $\widehat{\mathbf{X}}_{n+h}^G$ becomes arbitrarily small.

For $l = 1, \dots, d$, the l-th component of $\widehat{\mathbf{X}}_{n+h}^G$ is given by

$$\begin{aligned}\left\langle \widehat{X}_{n+h}^G, \nu_l \right\rangle = \left\langle \sum_{i=1}^{n} g_{ni}^{(h)} X_i, \nu_l \right\rangle &= \left\langle \sum_{i=1}^{n} \sum_{l'=1}^{\infty} \langle X_i, \nu_{l'} \rangle \, g_{ni}^{(h)} \nu_{l'}, \nu_l \right\rangle \\ &= \sum_{i=1}^{n} \sum_{l'=1}^{\infty} \langle X_i, \nu_{l'} \rangle \left\langle g_{ni}^{(h)} \nu_{l'}, \nu_l \right\rangle . \end{aligned} \tag{1.4.6}$$

Using the vector representation (1.4.5), we write

$$\begin{aligned}\widehat{\mathbf{X}}_{n+h}^G &= \sum_{i=1}^{n} \left(\begin{array}{ccc|cc} \left\langle g_{ni}^{(h)} \nu_1, \nu_1 \right\rangle & \dots & \left\langle g_{ni}^{(h)} \nu_d, \nu_1 \right\rangle & \left\langle g_{ni}^{(h)} \nu_{d+1}, \nu_1 \right\rangle & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \left\langle g_{ni}^{(h)} \nu_1, \nu_d \right\rangle & \dots & \left\langle g_{ni}^{(h)} \nu_d, \nu_d \right\rangle & \left\langle g_{ni}^{(h)} \nu_{d+1}, \nu_d \right\rangle & \dots \end{array} \right) \left(\begin{array}{c} \langle X_i, \nu_1 \rangle \\ \vdots \\ \langle X_i, \nu_d \rangle \\ \hline \langle X_i, \nu_{d+1} \rangle \\ \vdots \end{array} \right) \\ &=: \sum_{i=1}^{n} \mathbf{G}_{ni}^{(h)} \mathbf{X}_i + \sum_{i=1}^{n} \mathbf{G}_{ni}^{(h)\infty} \mathbf{X}_i^\infty , \end{aligned} \tag{1.4.7}$$

where $\mathbf{G}_{ni}^{(h)}$ are $d \times d$ matrices with ll'-th component $\langle g_{ni}^{(h)} \nu_{l'}, \nu_l \rangle$ and $\mathbf{G}_{ni}^\infty$ are $d \times \infty$ matrices with ll'-th component $\langle g_{ni}^{(h)} \nu_{d+l'}, \nu_l \rangle$.

Moreover, for all $Y \in G$ there exist $n \in \mathbb{N}$ and (possibly unbounded) linear operators $t_{n1}, \dots, t_{nn}$ such that

$$Y = \sum_{i=1}^{n} t_{ni} X_i. \tag{1.4.8}$$

Similarly as in (1.4.6), we project $Y \in G$ on $\overline{\text{sp}}\{\nu_1, \dots, \nu_d\}$, which results in

$$\begin{aligned}\mathbf{Y} &:= \left(\langle Y, \nu_1\rangle, \dots, \langle Y, \nu_d\rangle\right)^\top \\ &= \Big(\Big\langle \sum_{i=1}^n t_{ni} X_i, \nu_1\Big\rangle, \dots, \Big\langle \sum_{i=1}^n t_{ni} X_i, \nu_d\Big\rangle\Big)^\top \\ &=: \sum_{i=1}^n \mathbf{T}_{ni}\mathbf{X}_i + \sum_{i=1}^n \mathbf{T}_{ni}^\infty \mathbf{X}_i^\infty. \end{aligned} \tag{1.4.9}$$

The $d \times d$ matrices $\mathbf{T}_{ni}$ and the $d \times \infty$ matrices $\mathbf{T}_{ni}^\infty$ in (1.4.8) are defined in the same way as $\mathbf{G}_{ni}^{(h)}$ and $\mathbf{G}_{ni}^{(h)\infty}$ in (1.4.7). We denote by $\mathbf{M}$ the space of all $\mathbf{Y}$:

$$\mathbf{M} := \left\{\mathbf{Y} = \left(\langle Y, \nu_1\rangle, \dots, \langle Y, \nu_d\rangle\right)^\top : Y \in G\right\}.$$

Recall $\mathbf{M}_1$ as defined in (1.4.1). Observing that for all $\mathbf{Y}_1 \in \mathbf{M}_1$ there exist $d \times d$ matrices $\mathbf{A}_{n1}, \dots, \mathbf{A}_{nn}$ such that $\mathbf{Y}_1 = \sum_{i=1}^n \mathbf{A}_{ni}\mathbf{X}_i$, there also exist operators t_{ni} such that $\mathbf{T}_{ni} = \mathbf{A}_{ni}$, and $\mathbf{T}_{ni}^\infty = \mathbf{0}$, which then gives $\mathbf{Y}_1 \in \mathbf{M}$. Hence $\mathbf{M}_1 \subseteq \mathbf{M}$.

Now that we have introduced the notation and the setting, we are ready to compute the mean squared distance $E\|\widehat{\mathbf{X}}_{n+h} - \widehat{\mathbf{X}}_{n+h}^G\|_2^2$.

Theorem 1.4.9. *Suppose* $(X_n)_{n\in\mathbb{Z}}$ *is a functional* ARMA(p,q) *process such that Assumption 1.3.7 holds. For* $h \in \mathbb{N}$ *let* $\widehat{X}_{n+h}^G$ *be the functional best linear predictor of* X_{n+h} *as defined in* (1.4.3) *and* $\widehat{\mathbf{X}}_{n+h}^G$ *as in (1.4.5). Let furthermore* $\widehat{\mathbf{X}}_{n+h}$ *be the vector best linear predictor of* $\mathbf{X}_{n+h}$ *based on* $\mathbf{X}_1, \dots, \mathbf{X}_n$ *as in* (1.4.1).
(i) In the framework of Proposition 1.4.4, and if $\sum_{l=1}^\infty \sqrt{\lambda_l} < \infty$*, for all* $d \in \mathbb{N}$*,*

$$E\left\|\widehat{\mathbf{X}}_{n+h} - \widehat{\mathbf{X}}_{n+h}^G\right\|_2^2 \le 4\Big(\sum_{i=1}^n \|g_{ni}^{(h)}\|_{\mathcal{L}}\Big)^2 \Big(\sum_{l=d+1}^\infty \sqrt{\lambda_l}\Big)^2 < \infty.$$

(ii) In the framework of Proposition 1.4.5, for all $d \in \mathbb{N}$*,*

$$E\left\|\widehat{\mathbf{X}}_{n+h} - \widehat{\mathbf{X}}_{n+h}^G\right\|_2^2 \le 4\Big(\sum_{i=1}^n \Big(\sum_{l=d+1}^\infty \|g_{ni}^{(h)}\nu_l\|^2\Big)^{\frac{1}{2}}\Big)^2 \sum_{l=d+1}^\infty \lambda_l < \infty.$$

In both cases, $E\|\widehat{\mathbf{X}}_{n+h} - \widehat{\mathbf{X}}_{n+h}^G\|_2^2$ *tends to 0 as* $d \to \infty$*.*

We start with a technical lemma, which we need for the proof of the above Theorem.

Lemma 1.4.10. *Suppose* $(X_n)_{n\in\mathbb{Z}}$ *is a stationary and causal functional* ARMA(p,q) *process and* $(\nu_l)_{l\in\mathbb{N}}$ *are the eigenfunctions of its covariance operator* C_X. *Then for all* $j,l\in\mathbb{N}$,

$$E\left[\left\langle X_{n+h}-\widehat{X}^G_{n+h},\nu_l\right\rangle\langle Y,\nu_j\rangle\right]=0,\quad Y\in G.$$

Proof. For all $j,l\in\mathbb{N}$ we set $s_{l,j}(\cdot):=\langle\cdot,\nu_l\rangle\nu_j$. First note that for all $x\in H$ with $\|x\|\le 1$,

$$\|s_{l,j}x\|=\|\langle x,\nu_l\rangle\,\nu_j\|\le\|x\|\le 1,$$

hence, $s_{l,j}\in\mathcal{L}$. Since G is an $\mathcal{L}$-closed subspace, $Y\in G$ implies $s_{l,j}(Y)\in G$ and we get with Remark 1.4.3(i) for all $j,l\in\mathbb{N}$,

$$E\left\langle X_{n+h}-\widehat{X}^G_{n+h},s_{l,j}Y\right\rangle=E\left[\left\langle X_{n+h}-\widehat{X}^G_{n+h},\nu_l\right\rangle\langle Y,\nu_j\rangle\right]=0.$$

□

Proof of Theorem 1.4.9. First note that under both conditions (i) and (ii), there exist $g^{(h)}_{ni}\in\mathcal{L}$ such that $\widehat{X}^G_{n+h}=\sum_{i=1}^n g^{(h)}_{ni}X_{n+h-i}$ and that $\mathcal{S}\subset\mathcal{L}$. With the matrix representation of $\widehat{\mathbf{X}}^G_{n+h}$ in (1.4.7) and Lemma 1.4.10 we obtain

$$\begin{aligned}\sum_{j=1}^d E\Big[\langle Y,\nu_j\rangle\Big\langle X_{n+h}-\widehat{X}^G_{n+h},\nu_j\Big\rangle\Big]&=E\Big\langle\mathbf{Y},\mathbf{X}_{n+h}-\widehat{\mathbf{X}}^G_{n+h}\Big\rangle_{\mathbb{R}^d}\\&=E\Big\langle\mathbf{Y},\mathbf{X}_{n+h}-\sum_{i=1}^n\mathbf{G}^{(h)}_{ni}\mathbf{X}_i-\sum_{i=1}^n\mathbf{G}^{(h)\infty}_{ni}\mathbf{X}^\infty_i\Big\rangle_{\mathbb{R}^d}=0,\quad Y\in G,\end{aligned}\tag{1.4.10}$$

where $\mathbf{Y}$ is defined as in (1.4.9). Since (1.4.10) holds for all $\mathbf{Y}\in\mathbf{M}$ and $\mathbf{M}_1\subseteq\mathbf{M}$, it especially holds for all $\mathbf{Y}_1\in\mathbf{M}_1$; i.e.,

$$E\Big\langle\mathbf{Y}_1,\mathbf{X}_{n+h}-\sum_{i=1}^n\mathbf{G}^{(h)}_{ni}\mathbf{X}_i-\sum_{i=1}^n\mathbf{G}^{(h)\infty}_{ni}\mathbf{X}^\infty_i\Big\rangle_{\mathbb{R}^d}=0,\quad\mathbf{Y}_1\in\mathbf{M}_1.\tag{1.4.11}$$

Combining (1.4.11) and Remark 1.4.3(i), we get

$$E\Big\langle\mathbf{Y}_1,\widehat{\mathbf{X}}_{n+h}-\sum_{i=1}^n\mathbf{G}^{(h)}_{ni}\mathbf{X}_i\Big\rangle_{\mathbb{R}^d}=E\Big\langle\mathbf{Y}_1,\sum_{i=1}^n\mathbf{G}^{(h)\infty}_{ni}\mathbf{X}^\infty_i\Big\rangle_{\mathbb{R}^d},\quad\mathbf{Y}_1\in\mathbf{M}_1.\tag{1.4.12}$$

Since both $\widehat{\mathbf{X}}_{n+h}$ and $\sum\limits_{i=1}^n\mathbf{G}^{(h)}_{ni}\mathbf{X}_i$ are in $\mathbf{M}_1$, (1.4.12) especially holds, when

$$\mathbf{Y}_1=\widehat{\mathbf{X}}_{n+h}-\sum_{i=1}^n\mathbf{G}^{(h)}_{ni}\mathbf{X}_i\in\mathbf{M}.\tag{1.4.13}$$

We plug $\mathbf{Y}_1$ as defined in (1.4.13) into (1.4.12) and obtain

$$E\Big\langle \widehat{\mathbf{X}}_{n+h} - \sum_{i=1}^{n} \mathbf{G}_{ni}^{(h)} \mathbf{X}_i, \widehat{\mathbf{X}}_{n+h} - \sum_{i=1}^{n} \mathbf{G}_{ni}^{(h)} \mathbf{X}_i \Big\rangle_{\mathbb{R}^d}$$

$$= E\Big\langle \widehat{\mathbf{X}}_{n+h} - \sum_{i=1}^{n} \mathbf{G}_{ni}^{(h)} \mathbf{X}_i, \sum_{i=1}^{n} \mathbf{G}_{ni}^{(h)\infty} \mathbf{X}_i^{\infty} \Big\rangle_{\mathbb{R}^d}. \tag{1.4.14}$$

From the left hand side of (1.4.14) we read off

$$E\Big\langle \widehat{\mathbf{X}}_{n+h} - \sum_{i=1}^{n} \mathbf{G}_{ni}^{(h)} \mathbf{X}_i, \widehat{\mathbf{X}}_{n+h} - \sum_{i=1}^{n} \mathbf{G}_{ni}^{(h)} \mathbf{X}_i \Big\rangle_{\mathbb{R}^d} = E\Big\| \widehat{\mathbf{X}}_{n+h} - \sum_{i=1}^{n} \mathbf{G}_{ni}^{(h)} \mathbf{X}_i \Big\|_2^2, \tag{1.4.15}$$

and for the right-hand side of (1.4.14), applying the Cauchy-Schwarz inequality twice, we get

$$\begin{aligned} E\Big\langle \widehat{\mathbf{X}}_{n+h} - \sum_{i=1}^{n} \mathbf{G}_{ni}^{(h)} \mathbf{X}_i, \sum_{i=1}^{n} \mathbf{G}_{ni}^{(h)\infty} \mathbf{X}_i^{\infty} \Big\rangle_{\mathbb{R}^d} & \\ \le E\Big[\Big\| \widehat{\mathbf{X}}_{n+h} - \sum_{i=1}^{n} \mathbf{G}_{ni}^{(h)} \mathbf{X}_i \Big\|_2 \Big\| \sum_{i=1}^{n} \mathbf{G}_{ni}^{(h)\infty} \mathbf{X}_i^{\infty} \Big\|_2 \Big] & \\ \le \Big(E\Big\| \widehat{\mathbf{X}}_{n+h} - \sum_{i=1}^{n} \mathbf{G}_{ni}^{(h)} \mathbf{X}_i \Big\|_2^2 \Big)^{\frac{1}{2}} \Big(E\Big\| \sum_{i=1}^{n} \mathbf{G}_{ni}^{(h)\infty} \mathbf{X}_i^{\infty} \Big\|_2^2 \Big)^{\frac{1}{2}}. & \end{aligned} \tag{1.4.16}$$

Dividing the right-hand side of (1.4.15) by the first square root on the right-hand side of (1.4.16) we find

$$E\Big\| \widehat{\mathbf{X}}_{n+h} - \sum_{i=1}^{n} \mathbf{G}_{ni}^{(h)} \mathbf{X}_i \Big\|_2^2 \le E\Big\| \sum_{i=1}^{n} \mathbf{G}_{ni}^{(h)\infty} \mathbf{X}_i^{\infty} \Big\|_2^2.$$

Hence, for the mean squared distance we obtain

$$\begin{aligned} E\Big\| \widehat{\mathbf{X}}_{n+h} - \widehat{\mathbf{X}}_{n+h}^{G} \Big\|_2^2 &= E\Big\| \widehat{\mathbf{X}}_{n+h} - \sum_{i=1}^{n} \mathbf{G}_{ni}^{(h)} \mathbf{X}_i - \sum_{i=1}^{n} \mathbf{G}_{ni}^{(h)\infty} \mathbf{X}_i^{\infty} \Big\|_2^2 \\ &\le 2E\Big\| \widehat{\mathbf{X}}_{n+h} - \sum_{i=1}^{n} \mathbf{G}_{ni}^{(h)} \mathbf{X}_i \Big\|_2^2 + 2E\Big\| \sum_{i=1}^{n} \mathbf{G}_{ni}^{(h)\infty} \mathbf{X}_i^{\infty} \Big\|_2^2 \\ &\le 4E\Big\| \sum_{i=1}^{n} \mathbf{G}_{ni}^{(h)\infty} \mathbf{X}_i^{\infty} \Big\|_2^2. \end{aligned}$$

What remains to do is to bound $\sum\limits_{i=1}^{n} \mathbf{G}_{ni}^{(h)\infty} \mathbf{X}_i^{\infty}$, which, by (1.4.7), is a d-dimensional vector with l-th component $\sum_{i=1}^{n} \sum\limits_{l'=d+1}^{\infty} \langle X_i, \nu_{l'} \rangle \langle g_{ni}^{(h)} \nu_l', \nu_l \rangle$.

(i) First we consider the framework of Proposition 1.4.4.
We abbreviate $x_{i,l'} := \langle X_i, \nu'_l \rangle$ and calculate

$$\begin{aligned} E\Big\| \sum_{i=1}^{n} \mathbf{G}_{ni}^{(h)\infty} \mathbf{X}_i^{\infty} \Big\|_2^2 &= E\Big\| \sum_{l=1}^{d} \Big(\sum_{i=1}^{n} \sum_{l'=d+1}^{\infty} x_{i,l'} \langle g_{ni}^{(h)} \nu'_l, \nu_l \rangle \Big) \nu_l \Big\|^2 \\ &\leq E\Big\| \sum_{l=1}^{\infty} \Big(\sum_{i=1}^{n} \sum_{l'=d+1}^{\infty} x_{i,l'} \langle g_{ni}^{(h)} \nu'_l, \nu_l \rangle \Big) \nu_l \Big\|^2 \\ &= E\Big\| \sum_{i=1}^{n} \sum_{l'=d+1}^{\infty} x_{i,l'} g_{ni}^{(h)} \nu'_l \Big\|^2 \end{aligned} \tag{1.4.17}$$

by Parseval's equality (1.2.1). Then we proceed using the orthogonality of ν_l and the Cauchy-Schwarz inequality,

$$\begin{aligned} &= E\Big[\Big\langle \sum_{i=1}^{n} \sum_{l=d+1}^{\infty} x_{i,l} g_{ni}^{(h)} \nu_l, \sum_{j=1}^{n} \sum_{l'=d+1}^{\infty} x_{j,l'} g_{nj}^{(h)} \nu'_l \Big\rangle \Big] \\ &= \sum_{i,j=1}^{n} \sum_{l,l'=d+1}^{\infty} E(x_{i,l} x_{j,l'}) \langle g_{ni}^{(h)} \nu_l, g_{nj}^{(h)} \nu_{l'} \rangle \\ &\leq \Big(\sum_{i=1}^{n} \sum_{l=d+1}^{\infty} \sqrt{E(x_{i,l})^2} \| g_{ni}^{(h)} \nu_l \| \Big)^2 \\ &= \Big(\sum_{i=1}^{n} \sum_{l=d+1}^{\infty} \sqrt{\lambda_l} \| g_{ni}^{(h)} \nu_l \| \Big)^2, \end{aligned} \tag{1.4.18}$$

since $E\langle X_i, \nu_l \rangle^2 = \lambda_l$ by (1.2.4). The right-hand side of (1.4.18) is bounded above by

$$\Big(\sum_{i=1}^{n} \sum_{l=d+1}^{\infty} \sqrt{\lambda_l} \| g_{ni}^{(h)} \|_{\mathcal{L}} \| \nu_l \| \Big)^2 = \Big(\sum_{i=1}^{n} \| g_{ni}^{(h)} \|_{\mathcal{L}} \Big)^2 \Big(\sum_{l=d+1}^{\infty} \sqrt{\lambda_l} \Big)^2,$$

since $\|\nu_l\| = 1$ for all $l \in \mathbb{N}$. Since $g_{ni}^{(h)} \in \mathcal{L}$, we have $\sum_{i=1}^{n} \|g_{ni}^{(h)}\|_{\mathcal{L}} < \infty$ for all $n \in \mathbb{N}$ and with $\sum_{l=1}^{\infty} \sqrt{\lambda_l} < \infty$, the right-hand side tends to 0 as $d \to \infty$.

(ii) In the framework of Proposition 1.4.5 there exist $g_{ni}^{(h)} \in \mathcal{S}$ such that $\widehat{X}_{n+h}^{G} = \sum_{i=1}^{n} g_{ni}^{(h)} X_{n+h-i}$. By the Cauchy-Schwarz inequality we estimate

$$\begin{aligned} E\Big\| \sum_{i=1}^{n} \mathbf{G}_{ni}^{(h)\infty} \mathbf{X}_i^{\infty} \Big\|_2^2 &\leq \Big(\sum_{i=1}^{n} \sum_{l=d+1}^{\infty} \sqrt{\lambda_l} \| g_{ni}^{(h)} \nu_l \| \Big)^2 \\ &\leq \Big(\sum_{i=1}^{n} \Big(\sum_{l=d+1}^{\infty} \lambda_l \Big)^{\frac{1}{2}} \Big(\sum_{l=d+1}^{\infty} \| g_{ni}^{(h)} \nu_l \|^2 \Big)^{\frac{1}{2}} \Big)^2 \\ &= \Big(\sum_{i=1}^{n} \Big(\sum_{l=d+1}^{\infty} \| g_{ni}^{(h)} \nu_l \|^2 \Big)^{\frac{1}{2}} \Big)^2 \sum_{l=d+1}^{\infty} \lambda_l. \end{aligned} \tag{1.4.19}$$

Now note that $\sum_{l=d+1}^{\infty} \|g_{ni}^{(h)} \nu_l\|^2 \le \|g_{ni}^{(h)}\|_{\mathcal{S}} < \infty$. Thus, (1.4.19) is bounded by

$$\Big(\sum_{i=1}^{n} \Big(\sum_{l=d+1}^{\infty} \|g_{ni}^{(h)} \nu_l\|^2\Big)^{1/2}\Big)^2 \sum_{l=d+1}^{\infty} \lambda_l \le \Big(\sum_{i=1}^{n} \|g_{ni}^{(h)}\|_{\mathcal{S}}^{1/2}\Big)^2 \sum_{l=d+1}^{\infty} \lambda_l < \infty,$$

such that (1.4.19) tends to 0 as $d \to \infty$. □

We are now ready to derive bounds of the mean squared prediction error $E\|X_{n+h} - \widehat{X}_{n+h}\|^2$.

Theorem 1.4.11. *Consider a stationary and causal functional* ARMA(p,q) *process as in* (1.3.1)*. Then, for $h \in \mathbb{N}$, $\widehat{X}_{n+h}$ as defined in* (1.4.2)*, and $\sigma_{n,h}^2$ as defined in* (1.4.4)*, we obtain*

$$E\left\|X_{n+h} - \widehat{X}_{n+h}\right\|^2 \le \sigma_{n,h}^2 + \gamma_{d;n;h},$$

where $\gamma_{d;n;h}$ can be specified as follows.

(i) In the framework of Proposition 1.4.4, and if $\sum_{l=1}^{\infty} \sqrt{\lambda_l} < \infty$, for all $d \in \mathbb{N}$,

$$\gamma_{d;n;h} = 4\Big(\sum_{i=1}^{n} \|g_{ni}^{(h)}\|_{\mathcal{L}}\Big)^2 \Big(\sum_{l=d+1}^{\infty} \sqrt{\lambda_l}\Big)^2 + \sum_{l=d+1}^{\infty} \lambda_l.$$

(ii) In the framework of Proposition 1.4.5, for all $d \in \mathbb{N}$,

$$\gamma_{d;n;h} = \sum_{l=d+1}^{\infty} \lambda_l \left(4\, g_{n;d;h} + 1\right) \quad \text{with} \quad g_{n;d;h} = \sum_{i=1}^{n} \Big(\sum_{l=d+1}^{\infty} \left\|g_{ni}^{(h)} \nu_l\right\|^2\Big)^{1/2} \le \sum_{i=1}^{n} \|g_{ni}^{(h)}\|_{\mathcal{S}}^2.$$

In both cases, $E\left\|X_{n+h} - \widehat{X}_{n+h}\right\|^2$ tends to $\sigma_{n,h}^2$ as $d \to \infty$.

Proof. With (1.2.4) and since $(\nu_l)_{l \in \mathbb{N}}$ is an ONB, we get

$$\begin{aligned} E\left\|X_{n+h} - \widehat{X}_{n+h}\right\|^2 &= E\Big\|\sum_{l=1}^{d} \langle X_{n+h} - \widehat{X}_{n+h}, \nu_l\rangle \nu_l + \sum_{l=d+1}^{\infty} \langle X_{n+h}, \nu_l\rangle \nu_l\Big\|^2 \\ &= \sum_{l=1}^{d} E\left\|\langle X_{n+h} - \widehat{X}_{n+h}, \nu_l\rangle \nu_l\right\|^2 + \sum_{l=d+1}^{\infty} E\left\|\langle X_{n+h}, \nu_l\rangle \nu_l\right\|^2 \\ &= \sum_{l=1}^{d} E\langle X_{n+h} - \widehat{X}_{n+h}, \nu_l\rangle^2 + \sum_{l=d+1}^{\infty} \lambda_l. \end{aligned} \tag{1.4.20}$$

Now note that by definition of the Euclidean norm,

$$\sum_{l=1}^{d} E\langle X_{n+h} - \widehat{X}_{n+h}, \nu_l\rangle^2 = E\|\mathbf{X}_{n+h} - \widehat{\mathbf{X}}_{n+h}\|_2^2.$$

Furthermore, by Definition 1.4.2 of $\mathcal{L}$-closed subspaces and Remark 1.4.3(i), $E\langle X_{n+h} - \widehat{X}^G_{n+h}, Y\rangle = 0$ for all $Y \in G$. Observing that $\widehat{X}^G_{n+h} - \widehat{X}_{n+h} \in G$, we conclude that

$$E\langle X_{n+h} - \widehat{X}^G_{n+h}, \widehat{X}^G_{n+h} - \widehat{X}_{n+h}\rangle = 0,$$

and, by Lemma 1.4.10,

$$E\langle X_{n+h} - \widehat{X}^G_{n+h}, \nu_l\rangle\langle \widehat{X}^G_{n+h} - \widehat{X}_{n+h}, \nu_{l'}\rangle = 0, \quad l, l' \in \mathbb{N}.$$

Hence,

$$E\|\mathbf{X}_{n+h} - \widehat{\mathbf{X}}_{n+h}\|_2^2 = E\|\mathbf{X}_{n+h} - \widehat{\mathbf{X}}^G_{n+h}\|_2^2 + E\|\widehat{\mathbf{X}}^G_{n+h} - \widehat{\mathbf{X}}_{n+h}\|_2^2, \tag{1.4.21}$$

where for the first term of the right-hand side,

$$\begin{aligned} E\|\mathbf{X}_{n+h} - \widehat{\mathbf{X}}^G_{n+h}\|_2^2 &= E\sum_{l=1}^{d}\langle X_{n+h} - \widehat{X}^G_{n+h}, \nu_l\rangle^2 \le \sum_{l=1}^{\infty}\langle X_{n+h} - \widehat{X}^G_{n+h}, \nu_l\rangle^2 \\ &= E\|X_{n+h} - \widehat{X}^G_{n+h}\|^2 = \sigma^2_{n,h}, \end{aligned} \tag{1.4.22}$$

and the last equality holds by Remark 1.4.3(iii). For the second term of the right-hand side of (1.4.21) we use Theorem 1.4.9. We finish the proof of both (i) and (ii) by plugging (1.4.21) and (1.4.22) into (1.4.20). □

Since the prediction error decreases with d, Theorem 1.4.11 can not be applied as a criterion for the choice of d. In a data analysis, when quantities such as covariance operators and its eigenvalues have to be estimated, the variance of the estimators increases with d. Small errors in the estimation of small empirical eigenvalues may have severe consequences on the prediction error (see Bernard [6]). To avoid this problem a conservative choice of d is suggested. Theorem 1.4.11 allows for an interpretation of the prediction error for fixed d. This is similar as for Theorem 3.2 in Aue et al. [5], here for a more general model class of ARMA models.

1.5 Traffic data analysis

We apply the functional time series prediction theory of Section 4 to highway traffic data provided by the Autobahndirektion Südbayern, thus extending previous work

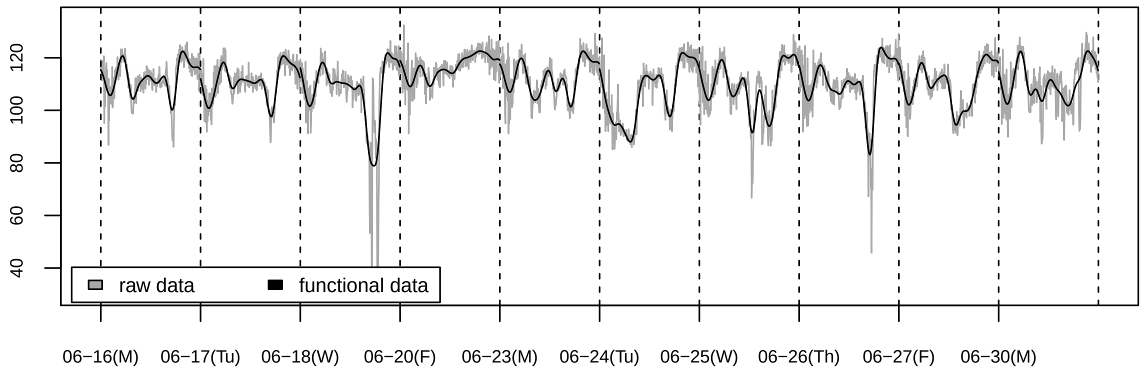

Figure 1.3: Functional velocity data (black) and raw data (grey) both in km/h on the last ten working days in June 2014 (June 19th 2014 was a catholic holiday).

in Besse and Cardot [7]. Our dataset consists of measurements at a fixed point on a highway (A92) in Southern Bavaria, Germany. Recorded is the average velocity per minute from 1/1/2014 00:00 to 30/06/2014 23:59 on three lanes. After taking care of missing values and data outliers, we average the velocity per minute over the three lanes, weighted by the number of vehicles per lane. Then we transform the cleaned daily high-dimensional data to functional data using the first 30 Fourier basis functions. The two standard bases of function spaces used in FDA are Fourier and B-spline basis functions (see Ramsay and Silverman [47], Section 3.3). We choose Fourier basis functions as they allow for a more parsimonious representation of the variability: a Fourier representation needs only 4 FPCs to explain 80% of the variability in the data, whereas a B-spline representation requires 6 (see Wei [55], Section 6.1). In Figure 1.3 we depict the resulting curves for the working days of two weeks in June 2014. More information on the transformation from discrete time observation to functional data and details on the implementation in `R` are provided in Wei [55], Chapter 6.

As can be seen in Figure 1.4, different weekdays have different mean velocity functions. To account for the difference between weekdays we subtract the empirical individual daily mean from all daily data (Monday mean from Monday data, etc.). The effect is clearly visible in Figure 1.5. However, even after deduction of the daily

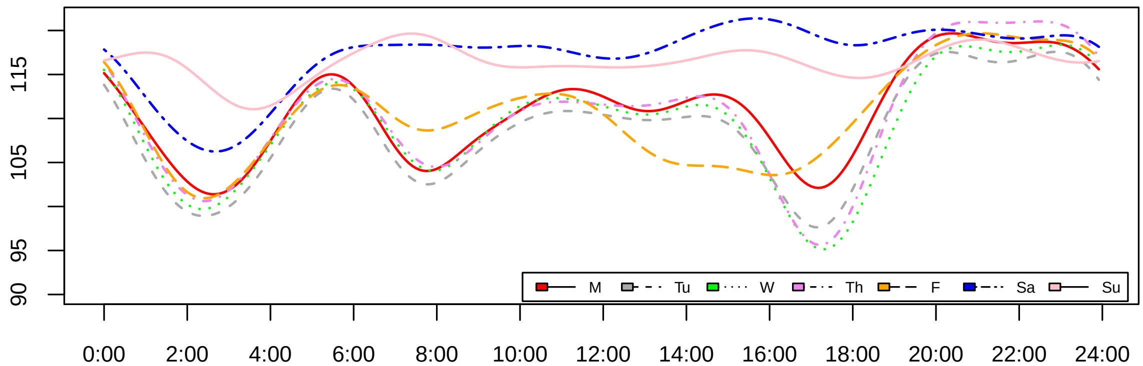

Figure 1.4: Empirical functional mean velocity (in km/h) on the 7 days of the week, over the day

mean, the functional stationarity test of Horvàth et al. [28] based on projection rejects stationarity of the time series. This is due to traffic flow on weekends: Saturday and Sunday traffic show different patterns than weekdays, even after mean correction. Consequently, we restrict our investigation to working days (Monday-Friday), resulting in a functional time series X_n for $n = 1, \ldots, N = 119$, for which the stationarity test suggested in Horvàth et al. [28] does not reject the stationarity assumption.

A portmanteau test of Gabrys and Kokoszka [20] applied to X_n for $n = 1, \ldots, N = 119$ working days rejects (with a p-value as small as 10^{-6}) that the daily functional data are uncorrelated. The assumption of temporal dependence in the data is in line with the results in Chrobok et al. [15] who use linear models to predict inner city traffic flow, and with results in Besse and Cardot [7] who use the temporal dependence for the prediction of traffic volume with a functional AR(1) model.

Next we show the prediction method at work for our data. More precisely, we estimate the 1-step predictors for the last 10 working days of our dataset and present the final result in Figure 1.9, where we compare the functional velocity data with their 1-step predictor. We explain the procedure in detail.

We start by estimating the covariance operator (recall Remark 1.3.10). Figure 1.6 shows the empirical covariance kernel of the functional traffic velocity data based on 119 working days (the empirical version of $E[(X(t) - \mu(t))(X(s) - \mu(s))]$ for

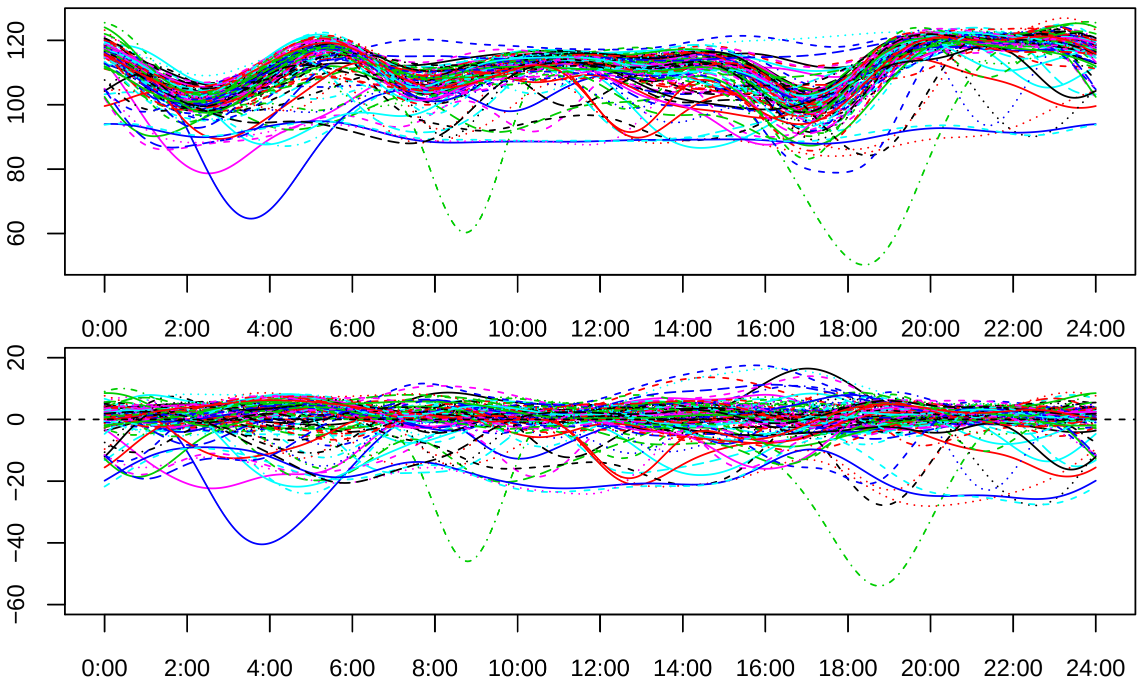

Figure 1.5: Functional velocity data (in km/h) over the day for 30 working days smoothed by a Fourier basis before and after subtracting the weekday mean

$0 \leq t, s \leq 1$).

As indicated by the arrows, the point $(t, s) = (0, 0)$ is at the bottom right corner and estimates the variance at midnight. The empirical variance over the day is represented along the diagonal from the bottom right to the top left corner. The valleys and peaks along the diagonal represent phases of low and high traffic density: for instance, the first peak represents the variance at around 05:00 a.m., where traffic becomes denser, since commuting to work starts. Peaks away from the diagonal represent high dependencies between different time points during the day. For instance, high traffic density in the early morning correlates with high traffic density in the late afternoon, again due to commuting.

Next we compute the empirical eigenpairs (λ_j^e, ν_j^e) for $j = 1, \ldots, N$ of the empirical covariance operator. The first four eigenfunctions are depicted in Figure 1.7.

Now we apply the CPV method from Remark 1.2.3 to the functional highway

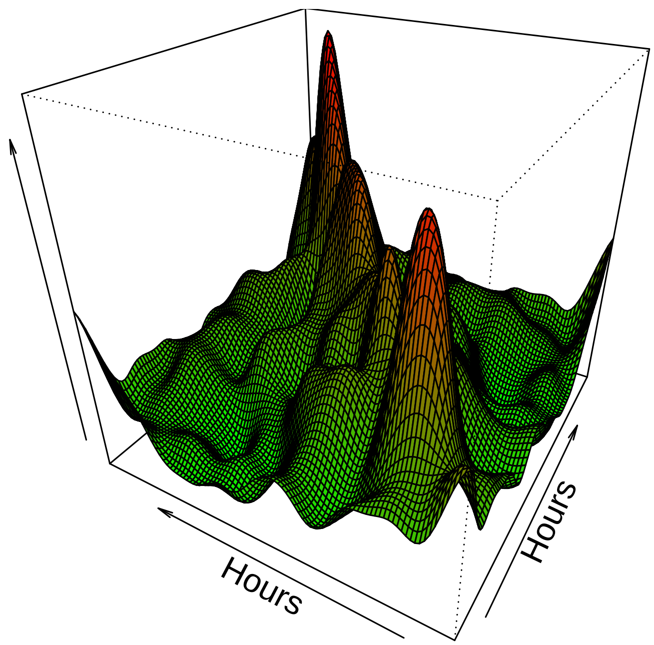

Figure 1.6: Empirical covariance kernel of functional velocity data on 119 working days.

velocity data. From a "CPV(d) vs. d" plot we read off that $d = 4$ FPCs explain 80% of the variability of the data.

Obviously, the choice of d is critical. Choosing d too small induces a loss of information as seen in Theorem 4.13. Choosing d too large makes the estimation of the vector model difficult and may result in imprecise predictors: the prediction error may explode (see Bernard [6]). As a remedy we perform cross validation on the prediction error based on a different number d of relevant scores. This furthermore ensures that the dependence structure of the data is not ignored when it is relevant for prediction.

Since the prediction is not only based on the number of scores, but also on the chosen ARMA model, we perform cross validation on the number of scores in combination with cross validation on the orders of the ARMA models.

Thus, we apply Algorithm 1 of Section 4.1 to the functional velocity data and implement the following steps for $d = 2, \dots, 6$ and $N = 119$.

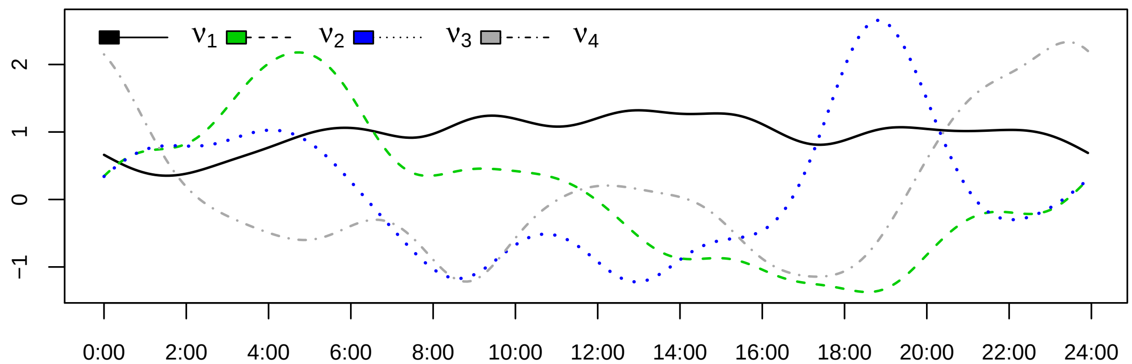

Figure 1.7: Four empirical eigenfunctions of the $N = 119$ working days functional velocity data. The criterion is 80%; i.e., $\nu_1, \nu_2, \nu_3, \nu_4$ explain together 80% of the total data variability.

(1) For each day $n \in \{1, \dots, N\}$, truncate the Karhunen-Loève representation (Theorem 1.2.2) of the daily functional velocity curve X_n at d. This yields

$$X_{n,d} := \sum_{j=1}^{d} \left\langle X_n, \nu_j^e \right\rangle \nu_j^e, \quad n = 1, \dots, N.$$

(Figure 1.8 depicts the (centered) functional velocity data and the corresponding truncated data for $d = 4$.) Store the d scores in the vector $\mathbf{X}_n$,

$$\mathbf{X}_n = (\langle X_n, \nu_1^e \rangle, \dots, \langle X_n, \nu_d^e \rangle)^\top, \quad n = 1, \dots, N.$$

(2) Fit different vector ARMA(p, q) models to the d-dimensional score vector. Compute the best linear predictor $\widehat{\mathbf{X}}_{n+1}$ based on the vector model iteratively by the Durbin-Levinson or the Innovations Algorithm (see e.g Brockwell and Davis [13]).

(3) Re-transform the vector best linear predictor $\widehat{\mathbf{X}}_{n+1}$ into its functional form $\widehat{X}_{n+1}$. Compare the goodness of fit of the models by their functional prediction errors $\|X_{n+1} - \widehat{X}_{n+1}\|^2$. (In Table 1.1 root mean squared errors (RMSE) and mean absolute errors (MAE) for the different models are summarized.) Fix the optimal d and the optimal ARMA(p, q) model.

As a result we find minimal 1-step prediction errors for $d = 4$, which confirms the choice proposed by the CPV method, and for the VAR(2) and the vector MA(1) model. Both models yield the same RMSE, and the MAE of the vector MA(1)

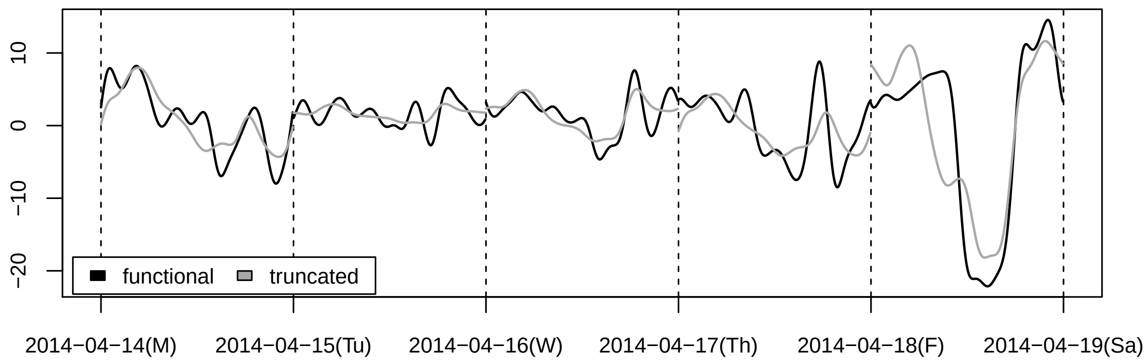

Figure 1.8: Functional velocity raw data on 5 consecutive working days (black) versus the truncated data by the Karhunen-Loève representation (grey). The criterion is 80% and the number d of FPCs is 4.

model is slightly smaller than that of the vector AR(2) model. Since we opt for a parsimonious model, we choose the vector MA(1) model, for which the predictor is depicted in Figure 1.9.

Finally, we compare the performance of the 1-step prediction based on the functional MA(1) model with standard non-parametric prediction methods. Our approach definitely outperforms prediction methods like exponential smoothing, naive prediction with the last observation, or using the mean of the time series as a predictor. Details are given in Wei [55], Section 6.3.

1.6 Conclusions

We have investigated functional ARMA(p, q) models and a corresponding approximating vector model, which lives on the closed linear span of the first d eigenfunctions of the covariance operator. We have presented conditions for the existence of a unique stationary and causal solution to both functional ARMA(p, q) and approximating vector model. Furthermore, we have derived conditions such that the approximating vector model is exact. Interestingly, and in contrast to AR or ARMA models, for a functional MA process of finite order the approximate vector process is automatically again a MA process of equal or smaller order.

	(p,q)	$(1,0)$	$(2,0)$	$(0,1)$	$(0,2)$	$(1,1)$	$(2,1)$	$(1,2)$
d=2	RMSE	5.15	5.09	5.02	5.15	5.13	4.96	5.09
	MAE	3.82	3.77	3.73	3.83	3.80	3.66	3.76
d=3	RMSE	4.97	4.87	4.86	5.30	4.94	4.89	5.08
	MAE	3.70	3.62	3.61	3.87	3.68	3.63	3.69
d=4	RMSE	4.98	**4.83**	**4.83**	5.55	4.92	4.90	5.23
	MAE	3.67	3.55	**3.54**	4.13	3.62	3.61	3.83
d=5	RMSE	5.06	5.15	4.91	5.80	5.04	5.20	5.46
	MAE	3.76	3.77	3.63	4.38	3.76	3.80	4.02
d=6	RMSE	5.12	5.28	5.09	6.47	5.12	5.34	5.97
	MAE	3.78	3.88	3.82	4.87	3.81	3.91	4.50

Table 1.1: Average 1-step prediction errors of the predictors for the last 10 observations for all working days for different ARMA models and number of principal components.

For arbitrary $h \in \mathbb{N}$ we have investigated the h-step functional best linear predictor of Bosq [10] and gave conditions for a representation in terms of operators in $\mathcal{L}$. We have compared the best linear predictor of the approximating vector model with the functional best linear predictor, and showed that the difference between the two predictors tends to 0 if the dimension of the vector model $d \to \infty$. The theory gives rise to a prediction methodology for stationary functional ARMA(p,q) processes similar to the one introduced in Aue et al. [5].

We have applied the new prediction theory to traffic velocity data. For finding an appropriate dimension d of the vector model, we applied the FPC criterion and cross validation on the prediction error. For our traffic data the cross validation leads to the same choice of $d = 4$ as the FPC criterion for CPV$(d) \geq 80\%$. The model selection is also performed via cross validation on the 1-step prediction error for different ARMA models resulting in an MA(1) model.

The appeal of the methodology is its ease of application. Well-known `R` software packages (`fda` and `mts`) make the implementation straightforward. Furthermore, the generality of dependence induced by ARMA models extends the range of application of functional time series, which was so far restricted to autoregressive dependence structures.

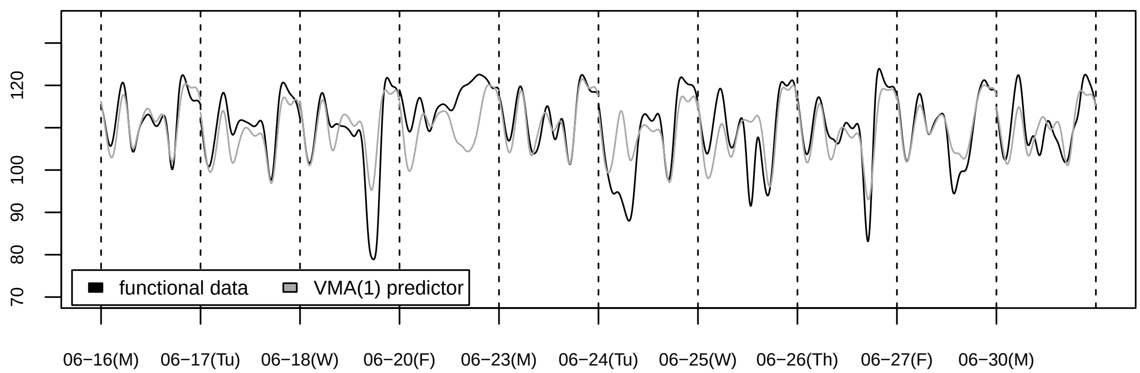

Figure 1.9: Functional velocity data in black and 1-step functional predictor based on VMA(1) in grey (in km/h) for the last 10 working days in June 2014

Chapter 2:

An Innovations Algorithm for the prediction of functional linear processes

2.1 Introduction

We consider observations which are consecutive curves over a fixed time interval within the field of functional data analysis (FDA). In this chapter curves are representations of a functional linear process. The data generating process is a time series $X = (X_n)_{n\in\mathbb{Z}}$ where each X_n is a random element $X_n(t)$, $t \in [0,1]$, of a Hilbert space, often the space of square integrable functions on $[0,1]$.

Several books contain a mathematical or statistical treatment of dependent functional data as Bosq [8], Horvàth and Kokoszka [27], and Bosq and Blanke [11]. The main source of this chapter is the book Bosq [8] on linear processes in function spaces, which gives the most general mathematical treatment of linear dependence in functional data, developing estimation, limit theorems and prediction for functional autoregressive processes. In Hörmann and Kokoszka [24] the authors develop limit theorems for the larger class of weakly dependent functional processes. More recently, Hörmann et al. [25] and Panaretos and Tavakoli [44] contribute to frequency domain methods of functional time series.

Solving the prediction equations in function spaces is problematic and research to-date has mainly considered first order autoregressive models. Contributions to functional prediction go hand in hand with an estimation method for the autoregressive parameter operator. Bosq [8] suggests a Yule-Walker type moment estimator, spline approximation is applied in Besse and Cardot [7], and Kargin and Onatski [32] proposes a predictive factor method where the principal components are replaced by directions which may be more relevant for prediction.

When moving away from the autoregressive process, results on prediction of functional time series become sparse. An interesting theory for the prediction of general functional linear processes is developed in Bosq [10]. Necessary and sufficient conditions are derived for the best linear predictor to take the form $\phi_n(X_1, \dots, X_n)$ with ϕ_n linear and bounded. However, due to the infinite-dimensionality of function spaces, boundedness of ϕ_n cannot be guaranteed. Consequently, most results, though interesting from a theoretical point of view, are not suitable for application.

More practical results are given for example in Antoniadis et al. [1], where prediction is performed non-parametrically with a functional kernel regression technique, or Chapter 1, Aue et al. [5] and Hyndman and Shang [30], where the dimensionality of the prediction problem is reduced via functional principal component analysis. In a multivariate setting, the Innovations Algorithm proposed in Brockwell and Davis [13] gives a established prediction method for linear processes. However, as often in functional data analysis, the non-invertibility of covariance operators prevents an ad-hoc generalization of the Innovations Algorithm to functional linear processes.

We suggest a computationally feasible linear prediction method extending the Innovations Algorithm to the functional setting. For a functional linear process $(X_n)_{n\in\mathbb{Z}}$ with values in a Hilbert space H and with innovation process $(\varepsilon_n)_{n\in\mathbb{Z}}$ our goal is the construction of a linear predictor $\widehat{X}_{n+1}$ based on $X_1, \dots, X_n$ such that $\widehat{X}_{n+1}$ is both computationally tractable and *consistent*. In other words, we want to find a bounded linear mapping ϕ_n with $\widehat{X}_{n+1} = \phi_n(X_1, \dots, X_n)$ such that the statistical prediction error converges to 0 for increasing sample size; i.e.,

$$\lim_{n\to\infty} \mathbb{E}\|X_{n+1} - \widehat{X}_{n+1}\|^2 = \mathbb{E}\|\varepsilon_0\|^2. \tag{2.1.1}$$

To achieve convergence in (2.1.1) we work with finite-dimensional projections of the functional process, similarly as in Aue et al. [5] and Chapter 1. We start with a representation of the functional linear model in terms of an arbitrary orthonormal basis of the Hilbert space. We then focus on a representation of the model based on only finitely many basis functions. An intuitive choice for the orthonormal basis consists of the eigenfunctions of the covariance operator of the process. Taking the eigenfunctions corresponding to the D largest eigenvalues results in a truncated

Karhunen-Loéve representation, and guarantees to capture most of the variance of the process (see Aue et al. [5]). Other applications may call for a different choice.

Though the idea of finite-dimensional projections is not new, our approach differs significantly from existing ones. Previous approaches consider the innovations of the projected process as the projection of the innovation of the original functional process. Though this may be sufficient in practice, it is in general not theoretically accurate.

The Wold decomposition enables us to work with the exact dynamics of the projected process, which then allows us to derive precise asymptotic results. The task set for this chapter is of a purely predictive nature: we assume knowing the dependence structure and do not perform model selection or covariance estimation of the functional process. This will be the topic of the subsequent chapter.

The truncated process $(X_{D,n})_{n\in\mathbb{Z}}$ based on D basis functions is called subprocess. We show that every subprocess of a stationary (and invertible) functional process is again stationary (and invertible). We then use an isometric isomorphy to a D-dimensional vector process to compute the best linear predictor of $(X_{D,n})_{n\in\mathbb{Z}}$ with the Multivariate Innovations Algorithm (see Brockwell and Davis [13]).

As a special example we investigate the functional moving average process of finite order. We prove that every subprocess is again a functional moving average process of same order or less. Moreover, for this process the Innovations Algorithm simplifies. Invertibility is a natural assumption in the context of prediction (see Brockwell and Davis [13], Section 5.5, and Nsiri and Roy [43]), and we require it when proving limit results. The theoretical results on the structure of $(X_{D,n})_{n\in\mathbb{Z}}$ enable us to quantify the prediction error in (2.1.1). As expected, it can be decomposed in two terms, one due to the dimension reduction, and the other due to the statistical prediction error of the D-dimensional model. However, the goal of consistency as in (2.1.1) is not satisfied, as the error due to dimension reduction does not depend on the sample size.

Finally, in order to satisfy (2.1.1), we propose a modified version of the Innovations Algorithm. The idea is to increase D together with the sample size. Hence the iterations of our modified Innovations Algorithm are based on increasing subspaces. Here we focus on the eigenfunctions of the covariance operator of X as orthonormal basis of the function space.

The main result of this chapter states that the prediction error is a combination of two tail sums, one involving operators of the inverse representation of the process, and the other the eigenvalues of the covariance operator. We obtain a computationally tractable functional linear predictor for stationary invertible functional linear processes. As the sample size tends to infinity the predictor satisfies (2.1.1) with a rate depending on the eigenvalues of the covariance operator and of the spectral density operator.

This chapter is organized as follows. After summarizing prerequisites of functional time series in Section 2.2, we recall in Section 2.3 the framework of prediction in infinite-dimensional Hilbert spaces, mostly based on the work of Bosq (see [8, 9, 10]). Here we also clarify the difficulties of linear prediction in infinite-dimensional function spaces. In Section 2.4 we propose an Innovations Algorithm based on a finite-dimensional subprocess of X. The predictor proposed in Section 2.4, though quite general, does not satisfy (2.1.1). Hence, in Section 2.5 we project the process on a finite-dimensional subspace spanned by the eigenfunctions of the covariance operator of X, and formulate the prediction problem in such a way that the dimension of the subprocess increases with the sample size. A modification of the Innovations Algorithm then yields a predictor which satisfies (2.1.1) and remains computationally tractable. The proof of this result requires some work and is deferred to Section 2.6 along with some auxiliary results.

2.2 Methodology

Let $H = L^2([0,1])$ be the real Hilbert space of square integrable functions with norm $\|x\| = (\int_0^1 x^2(s)ds)^{1/2}$ generated by the inner product $\langle x, y\rangle = \int_0^1 x(s)y(s)ds$ for $x, y \in H$. We denote by $\mathcal{L}$ the space of bounded linear operators acting on H. If not stated differently, for $A \in \mathcal{L}$ we take the standard operator norm $\|A\|_{\mathcal{L}} = \sup_{\|x\|\le 1} \|Ax\|$. Its adjoint A^* is defined by $\langle Ax, y\rangle = \langle x, A^*y\rangle$ for $x, y \in H$. The operator $A \in \mathcal{L}$ is called nuclear operator (denoted by $\mathcal{N}$), if it admits a representation $A = \sum_{j=1}^{\infty} \lambda_j \langle e_j, \cdot\rangle f_j$ with $\sum_{j=1}^{\infty} |\lambda_j| < \infty$ for two orthonormal bases (ONB) $(e_j)_{j\in\mathbb{N}}$ and $(f_j)_{j\in\mathbb{N}}$ of H. In that case $\|A\|_{\mathcal{N}} = \sum_{j=1}^{\infty} |\lambda_j| < \infty$. We shall also use the estimate $\|AB\|_{\mathcal{N}} \le \|A\|_{\mathcal{L}}\|B\|_{\mathcal{N}}$ for $A \in \mathcal{L}$ and $B \in \mathcal{N}$. For an introduction and more insight into Hilbert spaces we refer to Chapters 3.2 and 3.6 in Simon [50].

Let $\mathcal{B}_H$ be the Borel σ-algebra of subsets of H. All random functions are defined on a probability space $(\Omega, \mathcal{A}, \mathcal{P})$ and are $\mathcal{A} - \mathcal{B}_H$-measurable. The space of square integrable random functions $L^2_H := L^2(\Omega, \mathcal{A}, \mathcal{P})$ is a Hilbert space with inner product $\mathbb{E}\langle X, Y\rangle = \mathbb{E}\int_0^1 X(s)Y(s)ds$ for $X, Y \in L^2_H$. Furthermore, we say that X is integrable if $\mathbb{E}\|X\| = \mathbb{E}[(\int_0^1 X^2(t)dt)^{1/2}] < \infty$.

From Lemma 1.2 of Bosq [8] we know that X is a random function with values in H if and only if $\langle \mu, X\rangle$ is a real random variable for every $\mu \in H$. Hence, the following definitions are possible.

Definition 2.2.1. (i) If $X \in L^2_H$ is integrable, then there exists a unique $\mu \in H$ such that $\mathbb{E}\langle y, X\rangle = \langle y, \mu\rangle$ for $y \in H$. It follows that $\mathbb{E}X(t) = \mu(t)$ for almost all $t \in [0,1]$, and $\mathbb{E}X \in H$ is called the *expectation* of X.
(ii) If $X \in L^2_H$ and $\mathbb{E}X = 0 \in H$, the *covariance operator* of X is defined as

$$C_X(y) = \mathbb{E}[\langle X, y\rangle X], \quad y \in H.$$

(iii) If $X, Y \in L^2_H$ and $\mathbb{E}X = \mathbb{E}Y = 0$, the *cross covariance operator* of X and Y is defined as

$$C_{X,Y}(y) = C^*_{Y,X}(y) = \mathbb{E}[\langle X, y\rangle Y], \quad y \in H.$$

□

The operators C_X and $C_{Y,X}$ are in $\mathcal{N}$ (see Bosq [8], Section 1.5). Furthermore, C_X is a self-adjoint ($C_X = C^*_X$) and non-negative definite operator with spectral representation

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

for eigenpairs $(\lambda_j, \nu_j)_{j\in\mathbb{N}}$, where $(\nu_j)_{j\in\mathbb{N}}$ is an ONB of H and $(\lambda_j)_{j\in\mathbb{N}}$ is a sequence of positive real numbers such that $\sum_{j=1}^{\infty} \lambda_j < \infty$. When considering spectral representations, we assume that the λ_j are ordered decreasingly; i.e., $\lambda_i \geq \lambda_k$ for $i < k$.

For ease of notation we introduce the operator

$$x \otimes y(\cdot) = \langle x, \cdot\rangle y,$$

which allows us to write $C_X = \mathbb{E}[X \otimes X]$ and $C_{X,Y} = \mathbb{E}[X \otimes Y]$. Using that $\|x \otimes y\|_{\mathcal{N}} = \|x\| \|y\|$, we get (see Bosq [8], Eq. (1.59))

$$\mathbb{E}\|X\|^2 = \mathbb{E}\|X \otimes X\|_{\mathcal{N}} = \|C_X\|_{\mathcal{N}}. \tag{2.2.1}$$

Additionally, the following equalities are useful: for $A \in \mathcal{L}$ and $x_i, y_i \in H$ for $i = 1, 2$ we have

$$\begin{aligned} A(x_1 \otimes y_1) &= A(\langle x_1, \cdot \rangle y_1) = \langle x_1, \cdot \rangle A y_1 = x_1 \otimes A y_1, \\ (x_1 + x_2) \otimes (y_1 + y_2) &= x_1 \otimes y_1 + x_1 \otimes y_2 + x_2 \otimes y_1 + x_2 \otimes y_2. \end{aligned} \tag{2.2.2}$$

We define now functional linear processes and state some of their properties, taken from Bosq [8], Section 1.5 and Section 3.1. We first define the driving noise sequence.

Definition 2.2.2. $(\varepsilon_n)_{n \in \mathbb{Z}}$ is *white noise (WN) in* L^2_H if $\mathbb{E}\,\varepsilon_n = 0$, $0 < \mathbb{E}\|\varepsilon_n\|^2 = \sigma^2 < \infty$, $C_{\varepsilon_n} = C_\varepsilon$ is independent of n, and if $C_{\varepsilon_n, \varepsilon_m} = 0$ for all $n, m \in \mathbb{Z}$, $n \neq m$. □

Definition 2.2.3. Let $(\varepsilon_n)_{n \in \mathbb{Z}}$ be WN and $(\psi_j)_{j \in \mathbb{N}}$ a sequence in $\mathcal{L}$. Define $\psi_0 = I_H$, the identity operator on H, and let $\mu \in H$. We call $(X_n)_{n \in \mathbb{Z}}$ satisfying

$$X_n = \mu + \sum_{j=0}^{\infty} \psi_j \varepsilon_{n-j}, \quad n \in \mathbb{Z}, \tag{2.2.3}$$

a *functional linear process* in L^2_H with mean μ. The series in (2.2.3) converges in probability. □

Note that by definition a functional linear process is causal. We now state assumptions to ensure stronger convergence of the above series.

Lemma 2.2.4 (Bosq [8], Lemma 7.1(2))**.** *Let* $(\varepsilon_n)_{n \in \mathbb{Z}}$ *be WN and* $\sum_{j=0}^{\infty} \|\psi_j\|^2_{\mathcal{L}} < \infty$*. Then the series in* (2.2.3) *converges in* L^2_H *and a.s., and* $(X_n)_{n \in \mathbb{Z}}$ *is (weakly) stationary.*

Strict stationarity of a functional linear process can be enforced by assuming that $(\varepsilon_n)_{n \in \mathbb{Z}}$ is additionally independent. In our setting weak stationarity will suffice. From here on, without loss of generality we set $\mu = 0$. For a stationary process $(X_n)_{n \in \mathbb{Z}}$, the covariance operator with lag h is denoted by

$$C_{X;h} = \mathbb{E}[X_0 \otimes X_h], \quad h \in \mathbb{Z}. \tag{2.2.4}$$

We now define the concept of invertibility of a functional linear process, which is a natural assumption in the context of prediction; see Brockwell and Davis [13], Chapter 5.5 and Nsiri and Roy [43].

Definition 2.2.5. A functional linear process $(X_n)_{n\in\mathbb{Z}}$ is said to be *invertible* if it admits the representation

$$X_n = \varepsilon_n + \sum_{j=1}^{\infty} \pi_j X_{n-j}, \quad n \in \mathbb{Z}, \tag{2.2.5}$$

for $\pi_j \in \mathcal{L}$ and $\sum_{j=1}^{\infty} \|\pi_j\|_{\mathcal{L}} < \infty$. □

In Theorem 7.2.1 of Bosq [8] a sufficient condition for the invertibility of functional linear processes is given. Note that every stationary causal functional autoregressive moving average (FARMA) process is a functional linear process (see Spangenberg [51], Theorem 2.3). Special cases include functional autoregressive processes of order $p \in \mathbb{N}$ (FAR(p)), which have been thoroughly investigated. Our focus is on functional linear models, with the functional moving average process of order $q \in \mathbb{N}$ (FMA(q)) as an illustrating example, which we investigate in Section 2.4.2.

Definition 2.2.6. For $q \in \mathbb{N}$ a FMA(q) is a functional linear process $(X_n)_{n\in\mathbb{Z}}$ in L^2_H such that for WN $(\varepsilon_n)_{n\in\mathbb{Z}}$ and $\psi_j \in \mathcal{L}$ for $j = 1, \dots, q$,

$$X_n = \varepsilon_n + \sum_{j=1}^{q} \psi_j \varepsilon_{n-j}, \quad n \in \mathbb{Z}. \tag{2.2.6}$$

□

A FMA(q) process can be characterized as follows.

Proposition 2.2.7 (Bosq and Blanke [11], Proposition 10.2)**.** *A stationary functional linear process $(X_n)_{n\in\mathbb{Z}}$ in L^2_H is an* FMA(q) *for some $q \in \mathbb{N}$ if and only if $C_{X;q} \neq 0$ and $C_{X;h} = 0$ for $|h| > q$.*

2.3 Prediction in Hilbert spaces

In a finite-dimensional setting when the random elements take values in $\mathbb{R}^d$ equipped with the Euclidean norm, the concept of linear prediction of a random vector is well known (see Brockwell and Davis [13], Section 11.4). The best linear approximation of a random vector $\mathbf{X}$ based on vectors $\mathbf{X}_1, \dots, \mathbf{X}_n$ is the orthogonal projection of each component of $\mathbf{X}$ on the smallest closed linear subspace of $L^2_{\mathbb{R}}(\Omega, \mathcal{A}, \mathbb{P})$ generated by the components of $\mathbf{X}_i$. This results in

$$\widehat{\mathbf{X}} := \sum_{i=1}^{n} \mathbf{\Phi}_{n,i} \mathbf{X}_i$$

for $\mathbf{\Phi}_{n,i} \in \mathbb{R}^{d \times d}$. In infinite-dimensional Hilbert spaces one proceeds similarly but needs a rich enough subspace on which to project. The concept of linear prediction in infinite-dimensional Hilbert spaces was introduced by Bosq; see Section 1.6 in Bosq [8]. We start by recalling the notion of $\mathcal{L}$-closed subspaces (LCS), introduced in Fortet [19].

Definition 2.3.1. $\mathcal{G}$ is said to be an *$\mathcal{L}$-closed subspace* (LCS) of L^2_H if $\mathcal{G}$ is a Hilbertian subspace of L^2_H, and if $X \in \mathcal{G}$ and $\ell \in \mathcal{L}$ imply $\ell X \in \mathcal{G}$. □

We now give a characterization of an LCS generated by a subset of L^2_H.

Proposition 2.3.2 (Bosq [8], Theorem 1.8)**.** *Let $F \subseteq L^2_H$. Then the* LCS *generated by F, denoted by* LCS(F)*, is the closure with respect to $\|\cdot\|$ of*

$$F' = \Big\{ \sum_{i=1}^{k} \ell_i X_i : \ \ell_i \in \mathcal{L}, \ X_i \in F, \ k \geq 1 \Big\}.$$

We are now ready to define the best linear predictor in an infinite-dimensional Hilbert space analogous to the finite-dimensional setting.

Definition 2.3.3. Let $X_1, \dots, X_n$ be zero mean random elements in L^2_H. Define

$$F_n = \{X_1, \dots, X_n\} \quad \text{and} \quad \widehat{X}_{n+1} = P_{\mathrm{LCS}(F_n)}(X_{n+1}), \tag{2.3.1}$$

i.e., $\widehat{X}_{n+1}$ is the orthogonal projection of X_{n+1} on LCS(F_n). $\widehat{X}_{n+1}$ is called *best linear functional predictor* of X_{n+1} based on LCS(F_n). □

Note however that, since F' is not closed, $\widehat{X}_{n+1}$ as in (2.3.1) has in general not the form $\widehat{X}_{n+1} = \sum_{i=1}^{n} \ell_i X_i$ for $\ell_i \in \mathcal{L}$ (see Bosq [10], Proposition 2.2). Therefore, the practical relevance of (2.3.1) is limited. In the following we develop an alternative approach for the computation of the best linear predictor based on finite-dimensional projections of the functional process.

2.4 Prediction based on a finite-dimensional projection

For a stationary functional linear process $(X_n)_{n\in\mathbb{Z}}$ the infinite-dimensional setting makes the computation of $\widehat{X}_{n+1}$ as in (2.3.1) generally impossible. A natural solution lies in finite-dimensional projections of the functional process $(X_n)_{n\in\mathbb{Z}}$. For fixed $D\in\mathbb{N}$ we define

$$A_D = \overline{\mathrm{sp}}\{\nu_1,\dots,\nu_D\}, \tag{2.4.1}$$

where $(\nu_i)_{i\in\mathbb{N}}$ is some ONB of H, and consider the projection of a functional random element on A_D. In Aue et al. [5] the authors consider the projection of an FAR process $(X_n)_{n\in\mathbb{Z}}$ on A_D, where $\nu_1,\dots,\nu_D$ are the eigenfunctions corresponding to the largest eigenvalues of C_X. In Chapter 1 we proceed similarly with ARMA(p,q) models. However, instead of considering the true dynamics of the subprocess, they work with an approximation which lies in the same model class as the original functional process; e.g. projections of functional AR(p) models are approximated by multivariate AR(p) models. The following example clarifies this concept.

Example 2.4.1. Consider an FAR(1) process $(X_n)_{n\in\mathbb{Z}}$ as defined in Section 3.2 of Bosq [8] by

$$X_n = \Phi X_{n-1} + \varepsilon_n, \qquad n\in\mathbb{Z}, \tag{2.4.2}$$

for some $\Phi\in\mathcal{L}$ and WN $(\varepsilon_n)_{n\in\mathbb{Z}}$. Let furthermore $(\nu_i)_{i\in\mathbb{N}}$ be an arbitrary ONB of H. Using Parseval's identity, (2.4.2) can be rewritten in terms of $(\nu_i)_{i\in\mathbb{N}}$ as

$$\left(\begin{array}{c}\langle X_n,\nu_1\rangle\\ \vdots\\ \langle X_n,\nu_D\rangle\\ \hline \langle X_n,\nu_{D+1}\rangle\\ \vdots\end{array}\right) = \left(\begin{array}{ccc|cc}\langle\phi\nu_1,\nu_1\rangle & \dots & \langle\phi\nu_D,\nu_1\rangle & \langle\phi\nu_{D+1},\nu_1\rangle & \dots\\ \vdots & \ddots & \vdots & \vdots & \ddots\\ \langle\phi\nu_1,\nu_D\rangle & \dots & \langle\phi\nu_D,\nu_D\rangle & \langle\phi\nu_{D+1},\nu_D\rangle & \dots\\ \hline \langle\phi\nu_1,\nu_{D+1}\rangle & \dots & \langle\phi\nu_D,\nu_{D+1}\rangle & \langle\phi\nu_{D+1},\nu_{D+1}\rangle & \dots\\ \vdots & \ddots & \vdots & \vdots & \ddots\end{array}\right)\left(\begin{array}{c}\langle X_{n-1},\nu_1\rangle\\ \vdots\\ \langle X_{n-1},\nu_D\rangle\\ \hline \langle X_{n-1},\nu_{D+1}\rangle\\ \vdots\end{array}\right) + \left(\begin{array}{c}\langle\varepsilon_n,\nu_1\rangle\\ \vdots\\ \langle\varepsilon_n,\nu_D\rangle\\ \hline \langle\varepsilon_n,\nu_{D+1}\rangle\\ \vdots\end{array}\right),$$

which we abbreviate as

$$\begin{pmatrix} \mathbf{X}_{D,n} \\ \hline \mathbf{X}_n^\infty \end{pmatrix} = \left[\begin{array}{c|c} \mathbf{\Phi}_D & \mathbf{\Phi}_D^\infty \\ \hline \vdots & \vdots \end{array}\right] \begin{pmatrix} \mathbf{X}_{D,n-1} \\ \hline \mathbf{X}_{n-1}^\infty \end{pmatrix} + \begin{pmatrix} \mathbf{E}_{D,n} \\ \hline \mathbf{E}_n^\infty \end{pmatrix}. \tag{2.4.3}$$

We are interested in the dynamics of the D-dimensional subprocess $(\mathbf{X}_{D,n})_{n\in\mathbb{Z}}$. From (2.4.3) we find that $(\mathbf{X}_{D,n})_{n\in\mathbb{Z}}$ satisfies

$$\mathbf{X}_{D,n} = \mathbf{\Phi}_D \mathbf{X}_{D,n-1} + \mathbf{\Phi}_D^\infty \mathbf{X}_{n-1}^\infty + \mathbf{E}_{D,n}, \quad n \in \mathbb{Z}, \tag{2.4.4}$$

which does in general not define an FAR(1) process. This can be seen from the following example, similar to Example 3.7 in Bosq [8]. For some $a \in \mathbb{R}$ with $0 < a < 1/\sqrt{2}$ let

$$\Phi(x) = a \sum_{j=1}^{\infty} \langle x, \nu_j \rangle \nu_1 + a \sum_{i=1}^{\infty} \langle x, \nu_i \rangle \nu_{i+1}, \qquad x \in H.$$

Furthermore, assume that $\mathbb{E}\langle \varepsilon_n, \nu_1 \rangle^2 > 0$ but $\mathbb{E}\langle \varepsilon_n, \nu_j \rangle^2 = 0$ for all $j > 1$. Since $\|\Phi\|_{\mathcal{L}} = \sqrt{2}\,a < 1$, $(X_n)_{n\in\mathbb{Z}}$ defined by (2.4.2) with Φ as above is a stationary FAR(1) process. However, with (2.4.4) for $D = 1$,

$$\begin{aligned} \mathbf{X}_{1,n} &= \langle X_n, \nu_1 \rangle = a \sum_{j=1}^{\infty} \langle X_{n-1}, \nu_j \rangle + \langle \varepsilon_n, \nu_1 \rangle \\ &= a \langle X_{n-1}, \nu_1 \rangle + a \sum_{j=2}^{\infty} \Big\langle \Big(a \sum_{j'=1}^{\infty} \langle X_{n-2}, \nu_{j'} \rangle \nu_1 + a \sum_{i=1}^{\infty} \langle X_{n-2}, \nu_i \rangle \nu_{i+1} + \varepsilon_{n-1} \Big), \nu_j \Big\rangle + \langle \varepsilon_n, \nu_1 \rangle \\ &= a \langle X_{n-1}, \nu_1 \rangle + a^2 \langle X_{n-2}, \nu_1 \rangle + a^2 \sum_{j=2}^{\infty} \langle X_{n-2}, \nu_j \rangle + \langle \varepsilon_n, \nu_1 \rangle \\ &= \sum_{j=1}^{\infty} a^j \mathbf{X}_{1,n-j} + \mathbf{E}_{n,1}. \end{aligned}$$

Hence, $(\mathbf{X}_{1,n})_{n\in\mathbb{Z}}$ follows an AR(∞) model and $(\mathbf{X}_{1,n}\nu_1)_{n\in\mathbb{Z}}$ a FAR(∞) model.

In Chapter 1 and in Aue et al. [5] $(\mathbf{X}_{D,n})_{n\in\mathbb{Z}}$ is approximated by $(\tilde{\mathbf{X}}_{D,n})_{n\in\mathbb{Z}}$ satisfying

$$\tilde{\mathbf{X}}_{D,n} = \mathbf{\Phi}_D \tilde{\mathbf{X}}_{D,n-1} + \mathbf{E}_{D,n}, \quad n \in \mathbb{Z},$$

such that $(\tilde{\mathbf{X}}_{D,n})_{n\in\mathbb{Z}}$ follows a vector AR(1) process. □

We pursue the idea of Example 2.4.1 for functional linear processes and work with the true dynamics of a finite-dimensional subprocess.

2.4.1 Prediction of functional linear processes

For a functional linear process $(X_n)_{n\in\mathbb{Z}}$ we focus on the orthogonal projection

$$X_{D,n} = P_{A_D}(X_n) = \sum_{j=1}^{D} \langle X_n, \nu_j \rangle \nu_j, \quad n \in \mathbb{Z}, \tag{2.4.5}$$

for $(\nu_i)_{i\in\mathbb{N}}$ some ONB of H and A_D as in (2.4.1). We define for fixed $D \in \mathbb{N}$

$$F_{D,n} = \{X_{D,1}, \dots, X_{D,n}\}.$$

We will often use the following isometric isomorphism between two Hilbert spaces of the same dimension.

Lemma 2.4.2. *Define A_D as in* (2.4.1)*. The map $T : A_D \to \mathbb{R}^D$ defined by $Tx = (\langle x, \nu_i \rangle)_{i=1,\dots,D}$ is a bijective linear mapping with $\langle Tx, Ty \rangle_{\mathbb{R}^D} = \langle x, y \rangle$ for all $x, y \in A_D$. Hence,* LCS$(F_{D,n})$ *is isometrically isomorphic to $\overline{\text{sp}}\{\mathbf{X}_{D,1}, \dots, \mathbf{X}_{D,n}\}$. Moreover, $(X_{D,n})_{n\in\mathbb{Z}}$ as defined in* (2.4.5) *is isometrically isomorphic to the D-dimensional vector process*

$$\mathbf{X}_{D,n} := (\langle X_n, \nu_1 \rangle, \dots, \langle X_n, \nu_D \rangle)^\top, \quad n \in \mathbb{Z}. \tag{2.4.6}$$

When choosing $(\nu_i)_{i\in\mathbb{N}}$ as the eigenfunctions of the covariance operator C_X of $(X_n)_{n\in\mathbb{Z}}$, the representation (2.4.5) is a truncated version of the Karhunen-Loéve decomposition (see Bosq [8], Theorem 1.5).

As known from Example 2.4.1, the structure of $(X_n)_{n\in\mathbb{Z}}$ does in general not immediately reveal the dynamics of $(X_{D,n})_{n\in\mathbb{Z}}$. Starting with the representation of $(X_{D,n})_{n\in\mathbb{Z}}$ as in (2.2.3) with $\psi_0 = I_H$ and using similar notation as in (2.4.4), the D-dimensional vector process $(\mathbf{X}_{D,n})_{n\in\mathbb{Z}}$ can be written as

$$\mathbf{X}_{D,n} = \mathbf{E}_{D,n} + \sum_{j=1}^{\infty} \left(\mathbf{\Psi}_{D,j} \mathbf{E}_{D,n-j} + \mathbf{\Psi}_{D,j}^{\infty} \mathbf{E}_{n-j}^{\infty} \right), \quad n \in \mathbb{Z}, \tag{2.4.7}$$

where the blocks $\mathbf{\Psi}_{D,j}$, $\mathbf{\Psi}_{D,j}^{\infty}$, $\mathbf{E}_{D,n} = (\langle \varepsilon_n, \nu_1 \rangle, \dots, \langle \varepsilon_n, \nu_D \rangle)^\top$, and $\mathbf{E}_n^{\infty} = (\langle \varepsilon_n, \nu_{D+1} \rangle, \langle \varepsilon_n, \nu_{D+2} \rangle, \dots)^\top$ are defined analogously to the blocks in (2.4.3). Note that (2.4.7) is in general not a vector MA(∞) representation of a process with innovation $(\mathbf{E}_{D,n})_{n\in\mathbb{Z}}$.

The following proposition summarizes general results on the structure of $(X_{D,n})_{n\in\mathbb{Z}}$. Its proof is given in Section 2.6.

Proposition 2.4.3. *Let $(X_n)_{n\in\mathbb{Z}}$ be a stationary (and invertible) functional linear process with WN $(\varepsilon_n)_{n\in\mathbb{Z}}$, such that all eigenvalues of the covariance operator C_ε of $(\varepsilon_n)_{n\in\mathbb{Z}}$ are positive. Then $(X_{D,n})_{n\in\mathbb{Z}}$ is also a stationary (and invertible) functional linear process with some WN $(\tilde{\varepsilon}_{D,n})_{n\in\mathbb{Z}}$. $(\tilde{\varepsilon}_{D,n})_{n\in\mathbb{Z}}$ is isometrically isomorphic to the D-dimensional vector process $(\tilde{\mathbf{E}}_{D,n})_{n\in\mathbb{Z}}$, defined by $\tilde{\mathbf{E}}_{D,n} := (\langle\tilde{\varepsilon}_{D,n},\nu_1\rangle,\dots,\langle\tilde{\varepsilon}_{D,n},\nu_D\rangle)^\top$. Furthermore define $\mathcal{M}_{D,n} = \overline{\mathrm{sp}}\{\mathbf{X}_{D,t}, -\infty < t \le n\}$. Then*

$$\tilde{\mathbf{E}}_{D,n} = \mathbf{E}_{D,n} + \mathbf{\Psi}^\infty_{D,1}(\mathbf{E}^\infty_{n-1} - P_{\mathcal{M}_{D,n-1}}(\mathbf{E}^\infty_{n-1})) =: \mathbf{E}_{D,n} + \mathbf{\Delta}_{D,n-1}, \quad n \in \mathbb{Z}. \tag{2.4.8}$$

The lagged covariance operator $C_{X_D;h}$ of $(X_{D,n})_{n\in\mathbb{Z}}$ is given by

$$C_{X_D;h} = \mathbb{E}[P_{A_D}X_0 \otimes P_{A_D}X_h] = P_{A_D}\mathbb{E}[X_0\otimes X_h]P_{A_D} = P_{A_D}C_{X;h}P_{A_D}, \quad h\in\mathbb{Z}. \tag{2.4.9}$$

By Lemma 2.4.2, $(X_{D,n})_{n\in\mathbb{Z}}$ is isomorphic to the D-dimensional vector process $(\mathbf{X}_{D,n})_{n\in\mathbb{Z}}$ as defined in (2.4.6). The prediction problem can therefore be solved by methods from multivariate time series analysis. More precisely, we define for fixed $D\in\mathbb{N}$

$$\widehat{X}_{D,n+1} = P_{\mathrm{LCS}(F_{D,n})}(X_{n+1}),$$

i.e., $\widehat{X}_{D,n+1}$ is the best linear functional predictor based on $F_{D,n}$ for $n\in\mathbb{N}$. We formulate the Multivariate Innovations Algorithm for this setting.

Proposition 2.4.4 (Brockwell and Davis [13], Proposition 11.4.2)**.** *Let $(X_n)_{n\in\mathbb{Z}}$ be a stationary functional linear process and $(X_{D,n})_{n\in\mathbb{Z}} = (P_{A_D}X_n)_{n\in\mathbb{Z}}$ as in (2.4.5). If C_{X_D} is invertible on A_D then the best linear functional predictor $\widehat{X}_{D,n+1}$ of X_{n+1} based on* LCS$(F_{D,n})$ *can be computed by the following set of recursions:*

$$\widehat{X}_{D,1} = 0 \quad \text{and} \quad V_{D,0} = C_{X_D;0},$$

$$\widehat{X}_{D,n+1} = \sum_{i=1}^{n}\theta_{D,n,i}(X_{D,n+1-i} - \widehat{X}_{D,n+1-i}), \tag{2.4.10}$$

$$\theta_{D,n,n-i} = \Big(C_{X_D;n-i} - \sum_{j=0}^{i-1}\theta_{D,n,n-j}\ V_{D,j}\ \theta^*_{D,i,i-j}\Big)V^{-1}_{D,i}, \quad i = 1,\dots,n-1, \tag{2.4.11}$$

$$V_{D,n} = C_{X_{D,n+1}-\widehat{X}_{D,n+1}} = C_{X_D;0} - \sum_{j=0}^{n-1}\theta_{D,n,n-j}V_{D,j}\theta^*_{D,n,n-j}. \tag{2.4.12}$$

The recursions can be solved explicitly in the following order: $V_{D,0}$, $\theta_{D,1,1}$, $V_{D,1}$, $\theta_{D,2,2}$, $\theta_{D,2,1}\dots$. Thus we found a predictor which is easy to compute in contrast to $\widehat{X}_{n+1}$ from (2.3.1). However, since we are not using all available information, we lose predictive power. To evaluate this loss we bound the prediction error. We show that the error bound can be decomposed into two terms. One is due to the dimension reduction and the other to the statistical prediction error of the finite-dimensional model.

Theorem 2.4.5. *Let $(X_n)_{n\in\mathbb{Z}}$ be a stationary functional linear process with WN $(\varepsilon_n)_{n\in\mathbb{Z}}$ such that all eigenvalues of C_ε are positive. Assume furthermore that C_X is invertible on A_D. Recall the best linear functional predictor $\widehat{X}_{n+1}$ from Definition 2.3.3.*
(i) Then for all $n\in\mathbb{N}$ the prediction error is bounded:

$$E\|X_{n+1}-\widehat{X}_{n+1}\|^2 \le E\|X_{n+1}-\widehat{X}_{D,n+1}\|^2 = \sum_{i>D}\langle C_X\nu_i,\nu_i\rangle + \|V_{D,n}\|_{\mathcal{N}}^2. \tag{2.4.13}$$

(ii) If additionally $(X_n)_{n\in\mathbb{Z}}$ is invertible, then

$$\lim_{n\to\infty} E\|X_{n+1}-\widehat{X}_{D,n+1}\|^2 = \sum_{i>D}\langle C_X\nu_i,\nu_i\rangle + \|C_{\tilde{\varepsilon}_D}\|_{\mathcal{N}}^2.$$

Proof. (i) Since $\widehat{X}_{D,n+1} = P_{\mathrm{LCS}(F_{D,n})}(X_{n+1})$ and $\widehat{X}_{n+1} = P_{\mathrm{LCS}(F_n)}(X_{n+1})$, and since $\mathrm{LCS}(F_{D,n})\subseteq\mathrm{LCS}(F_n)$, the first inequality follows immediately from the projection theorem. Furthermore, since $X_{n+1}-X_{D,n+1}\in A_D^\perp$ (the orthogonal complement of A_D) and $X_{D,n+1}-\widehat{X}_{D,n+1}\in A_D$, we have $\langle X_{n+1}-X_{D,n+1}, X_{D,n+1}-\widehat{X}_{D,n+1}\rangle = 0$. Therefore,

$$\begin{aligned}\mathbb{E}\|X_{n+1}-\widehat{X}_{D,n+1}\|^2 &= \mathbb{E}\|X_{n+1}-X_{D,n+1}+X_{D,n+1}-\widehat{X}_{D,n+1}\|^2\\ &= \mathbb{E}\|X_{n+1}-X_{D,n+1}\|^2 + \mathbb{E}\|X_{D,n+1}-\widehat{X}_{D,n+1}\|^2.\end{aligned}$$

By (2.2.1), $\mathbb{E}\|X_{D,n+1}-\widehat{X}_{D,n+1}\|^2 = \|\mathbb{E}[(X_{D,n+1}-\widehat{X}_{D,n+1})\otimes(X_{D,n+1}-\widehat{X}_{D,n+1})]\|_{\mathcal{N}}$, which is equal to $\|V_{D,n}\|_{\mathcal{N}}$ by (2.4.12). Furthermore,

$$\begin{aligned}\mathbb{E}\|X_{n+1}-X_{D,n+1}\|^2 &= \mathbb{E}\Big\langle \sum_{i>D}\langle X_{n+1},\nu_i\rangle\nu_i, \sum_{j>D}\langle X_{n+1},\nu_j\rangle\nu_j\Big\rangle\\ &= \sum_{i,j>D}\mathbb{E}\Big\langle X_{n+1}\langle X_{n+1},\nu_i\rangle,\nu_j\Big\rangle\langle\nu_i,\nu_j\rangle\\ &= \sum_{i>D}\langle C_X\nu_i,\nu_i\rangle.\end{aligned}$$

(ii) By (i), what is left to show is that $\|V_{D,n}\|_{\mathcal{N}}^2 \to \|C_{\tilde{\varepsilon}_D}\|_{\mathcal{N}}^2$ for $n \to \infty$. However, this is an immediate consequence of the Multivariate Innovations Algorithm under the assumption that $(X_{D,n})_{n\in\mathbb{Z}}$ is invertible (see Remark 4 in Chapter 11 of Brockwell and Davis [13]). Invertibility of $(X_{D,n})_{n\in\mathbb{Z}}$ is given by Proposition 2.4.3, which finishes the proof. □

The above theorem states that for a stationary, invertible functional linear process, for increasing sample size the predictor restricted to the D-dimensional space performs arbitrarily well in the sense that in the limit only the statistical prediction error remains. However, our goal in (2.1.1) is not satisfied. The dimension reduction induces the additional error term $\sum_{i>D}\langle C_X(\nu_i), \nu_i\rangle$ independently of the sample size. If A_D is spanned by eigenfunctions of the covariance operator C_X with eigenvalues λ_i, the prediction error due to dimension reduction is $\sum_{i>D}\lambda_i$.

We now investigate the special case of functional moving average processes with finite order.

2.4.2 Prediction of FMA(q)

FMA(q) processes for $q \in \mathbb{N}$ as in Definition 2.2.6 are an interesting and not very well studied class of functional linear processes. We start with the FMA(1) process as an example.

Example 2.4.6. Consider an FMA(1) process $(X_n)_{n\in\mathbb{Z}}$ defined by

$$X_n = \psi\varepsilon_{n-1} + \varepsilon_n, \quad n \in \mathbb{Z},$$

for some $\psi \in \mathcal{L}$ and WN $(\varepsilon_n)_{n\in\mathbb{Z}}$. The representation of (2.4.7) reduces to

$$\mathbf{X}_{D,n} = \mathbf{\Psi}_D\mathbf{E}_{D,n-1} + \mathbf{\Psi}_D^{\infty}\mathbf{E}_{n-1}^{\infty} + \mathbf{E}_{D,n}, \quad n \in \mathbb{Z}.$$

As $\mathbf{X}_{D,n}$ depends on $\mathbf{E}_{D,n-1}$, $\mathbf{E}_{n-1}^{\infty}$ and $\mathbf{E}_{D,n}$, it is in general not a vector MA(1) process with WN $(\mathbf{E}_{D,n})_{n\in\mathbb{Z}}$. □ □

However, we can state the dynamics of a finite-dimensional subprocess of an FMA(q) process.

Theorem 2.4.7. *Let* $(X_n)_{n\in\mathbb{Z}}$ *be a stationary* FMA(q) *process for* $q \in \mathbb{N}$ *and* A_D *be as in* (2.4.1). *Then* $(X_{D,n})_{n\in\mathbb{Z}} = (P_{A_D} X_n)_{n\in\mathbb{Z}}$ *as defined in* (2.4.5) *is a stationary* FMA(q^*) *process for* $q^* \leq q$ *satisfying*

$$X_{D,n} = \sum_{j=1}^{q^*} \tilde{\psi}_{D,j} \tilde{\varepsilon}_{D,n-j} + \tilde{\varepsilon}_{D,n}, \qquad n \in \mathbb{Z}, \tag{2.4.14}$$

where $\tilde{\psi}_{D,j} \in \mathcal{L}$ *for* $j = 1, \ldots, q^*$ *and* $(\tilde{\varepsilon}_{D,n})_{n\in\mathbb{Z}}$ *is WN. Moreover,* $(\tilde{\varepsilon}_{D,n})_{n\in\mathbb{Z}}$ *is isometrically isomorphic to* $(\tilde{\mathbf{E}}_{D,n})_{n\in\mathbb{Z}}$ *as defined in* (2.4.8). *If* $q^* = 0$, *then* $(X_{D,n})_{n\in\mathbb{Z}}$ *is WN.*

Proof. By Proposition 2.4.3 $(X_{D,n})_{n\in\mathbb{Z}}$ is stationary. Furthermore, by (2.4.9) and Proposition 2.2.7 $C_{X_D;h} = P_{A_D} C_{X;h} P_{A_D} = 0$ for $h > q$, since $C_{X;h} = 0$ for $h > q$. Hence, again by Proposition 2.2.7 $(X_{D,n})_{n\in\mathbb{Z}}$ is a FMA(q^*) process, where q^* is the largest lag $j \leq q$ such that $C_{X_D;j} = P_{A_D} C_{X;j} P_{A_D} \neq 0$. Thus, (2.4.14) holds for some linear operators $\tilde{\psi}_{D,j} \in \mathcal{L}$ and WN $(\tilde{\varepsilon}_{D,n})_{n\in\mathbb{Z}}$. The fact that $(\tilde{\varepsilon}_{D,n})_{n\in\mathbb{Z}}$ is isometrically isomorphic to $(\tilde{\mathbf{E}}_{D,n})_{n\in\mathbb{Z}}$ as in (2.4.8) is again a consequence of the Wold decomposition of $(X_{D,n})_{n\in\mathbb{Z}}$ and follows from the proof of Proposition 2.4.3. □

The fact that every subprocess of a FMA(q) is a FMA(q^*) with $q^* \leq q$ simplifies the algorithm of Proposition 2.4.4, since $C_{X_D;h} = 0$ for $|h| > q$ modifies (2.4.10)-(2.4.12) as follows: for $n > q^*$,

$$\begin{aligned}
\widehat{X}_{D,n+1} &= \sum_{i=1}^{q^*} \theta_{D,n,i} (X_{D,n+1-i} - \widehat{X}_{D,n+1-i}) \\
\theta_{D,n,k} &= \Big(C_{X_D;k} - \sum_{j=0}^{n-k-1} \theta_{D,n,n-j} \; V_{D,j} \; \theta^*_{D,n-k,n-k-j} \Big) V^{-1}_{D,n-k}, \quad k = 1, \ldots, q^*, \\
V_{D,n} &= C_{X_{D,n+1} - \widehat{X}_{D,n+1}} = C_{X_d;0} - \sum_{j=1}^{q^*} \theta_{D,n,j} V_{D,n-j} \theta^*_{D,n,j}.
\end{aligned}$$

We now investigate the prediction error $E\|X_{n+1} - \widehat{X}_{D,n+1}\|^2$ of Theorem 2.4.5 for a functional linear process. For $D \to \infty$, $\sum_{i>D} \langle C_{X_0}(\nu_i), \nu_i \rangle \to 0$. However, the second term $\|V_{D,n}\|_{\mathcal{N}}$ on the right-hand-side of (2.4.13) is not defined in the limit, since the inverse of $V_{D,j}$ in (2.4.11) is no longer bounded when $D \to \infty$. To see this, take $V_{D,0}$. By (2.4.12), since $\widehat{X}_{D,1} = 0$ and since $(X_{D,n})_{n\in\mathbb{Z}}$ is stationary,

$$V_{D,0} = C_{X_{D,1} - \widehat{X}_{D,1}} = C_{X_{D,1}} = C_{X_D}.$$

By (2.4.9) for $h = 0$ we find $C_{X_D} = P_{A_D} C_X P_{A_D}$, hence for all $x \in H$, $\|(C_X - C_{X_D})(x)\| \to 0$ for $D \to \infty$. But, since C_X is not invertible on the entire H, neither is $\lim_{D\to\infty} C_{X_D}$. Therefore, $\lim_{D\to\infty} \widehat{X}_{D,n+1}$ is not defined.

To resolve this problem, we propose a tool used before in functional data analysis, for instance in Bosq [8] for the estimation of FAR(1). We increase the dimension D together with the sample size n by choosing $d_n := D(n)$ and $d_n \to \infty$ with $n \to \infty$. However, since the Innovations Algorithm is based on a recursion, it will always start with $V_{d_n,0} = C_{X_{d_n}}$, which again is not invertible on H for $d_n \to \infty$. For the Innovations Algorithm we increase D iteratively such that $V_{d_1,0}$ is inverted on A_1, $V_{d_2,1}$ is inverted on $A_2, \dots$ and so on. To quantify a convergence rate in Theorem 2.5.3 below we restrict our analysis to projections on eigenspaces of the covariance operator C_X of the underlying process.

2.5 Prediction based on projections on increasing subspaces of H

In this section we propose a functional version of the Innovations Algorithm. Starting with the same idea as in Section 2.4, we project the functional data on a finite-dimensional space. However, we now let the dimension of the space on which we project depend on the sample size. More precisely, let $(X_n)_{n\in\mathbb{Z}}$ be a stationary functional linear process with covariance operator C_X. For some positive, increasing sequence $(d_n)_{n\in\mathbb{N}}$ in $\mathbb{N}$ such that $d_n \to \infty$ with $n \to \infty$, we define

$$A_{d_n} = \overline{\text{sp}}\{\nu_1, \dots, \nu_{d_n}\}, \quad n \in \mathbb{N}, \tag{2.5.1}$$

where $(\nu_i)_{i\in\mathbb{N}}$ are now chosen as the eigenfunctions of C_X, and $(A_{d_n})_{n\in\mathbb{N}}$ is an increasing sequence of subspaces of H. Instead of applying the Innovations Algorithm to $(P_{A_D} X_1, \dots, P_{A_D} X_n)$ as in Proposition 2.4.4, we apply it now to $(P_{A_{d_1}} X_1, \dots, P_{A_{d_n}} X_n)$.

Proposition 2.5.1. *Let $(X_n)_{n\in\mathbb{Z}}$ be a stationary functional linear process with covariance operator C_X with eigenpairs $(\lambda_j, \nu_j)_{j\in\mathbb{N}}$, where $\lambda_j > 0$ for all $j \in \mathbb{N}$. Let $(d_n)_{n\in\mathbb{N}}$ be a positive sequence in $\mathbb{N}$ such that $d_n \uparrow \infty$ as $n \to \infty$. Define furthermore* *the best linear predictor of* X_{n+1} *based on* $\mathrm{LCS}(F'_{d_n,n})$ *for* $n \in \mathbb{N}$ *as*

$$F'_{d_n,n} = \{X_{d_1,1}, \dots, X_{d_n,n}\} \quad \textit{and} \quad \widehat{X}_{d_{n+1},n+1} = P_{\mathrm{LCS}(F'_{d_n,n})}(X_{n+1}). \tag{2.5.2}$$

Then $\widehat{X}_{d_{n+1},n+1}$ *is given by the following set of recursions:*

$$\widehat{X}_{d_1,1} = 0 \quad \textit{and} \quad V_{d_1,0} = C_{X_{d_1}},$$

$$\widehat{X}_{d_{n+1},n+1} = \sum_{i=1}^{n} \theta_{d_{n-i+1},n,i}(X_{d_{n+1-i},n+1-i} - \widehat{X}_{d_{n+1-i},n+1-i}), \tag{2.5.3}$$

$$\theta_{d_{i+1},n,n-i} = \Big(P_{A_{d_{n+1}}} C_{X;n-i} P_{A_{d_{i+1}}} - \sum_{j=0}^{i-1} \theta_{d_{j+1},n,n-j} \; V_{d_{j+1},j} \; \theta^*_{d_{j+1},i,i-j}\Big) V^{-1}_{d_{i+1},i}, \quad i = 1, \dots, n-1, \tag{2.5.4}$$

$$V_{d_{n+1},n} = C_{X_{d_{n+1},n+1} - \widehat{X}_{d_{n+1},n+1}} = C_{X_{d_{n+1}}} - \sum_{j=0}^{n-1} \theta_{d_{j+1},n,n-j} V_{d_{j+1},j} \theta^*_{d_{j+1},n,n-j}. \tag{2.5.5}$$

Proof. The proof is based on the proof of Proposition 11.4.2 in Brockwell and Davis [13]. First notice that the representation

$$\widehat{X}_{d_{n+1},n+1} = \sum_{i=1}^{n} \theta_{d_{n-i+1},n,i}(X_{d_{n+1-i},n+1-i} - \widehat{X}_{d_{n+1-i},n+1-i}), \quad n \in \mathbb{N},$$

results from the definition of $\widehat{X}_{d_{n+1},n+1} = P_{\mathrm{LCS}(F'_{d_n,n})}(X_{n+1})$. Multiplying both sides of (2.5.3) with $\langle X_{d_{k+1},k+1} - \widehat{X}_{d_{k+1},k+1}, \cdot \rangle$ for $0 \le k \le n$ and taking expectations, we get

$$\begin{aligned}
\mathbb{E}\big[(X_{d_{k+1},k+1} - \widehat{X}_{d_{k+1},k+1}) \otimes \widehat{X}_{d_{n+1},n+1}\big]
&= \sum_{i=1}^{n} \theta_{d_{n-i+1},n,i} \mathbb{E}\big[(X_{d_{k+1},k+1} - \widehat{X}_{d_{k+1},k+1}) \otimes (X_{d_{n+1-i},n+1-i} - \widehat{X}_{d_{n+1-i},n+1-i})\big] \\
&= \theta_{d_{k+1},n,n-k} \mathbb{E}\big[(X_{d_{k+1},k+1} - \widehat{X}_{d_{k+1},k+1}) \otimes (X_{d_{k+1},k+1} - \widehat{X}_{d_{k+1},k+1})\big],
\end{aligned}$$

where we used that $\mathbb{E}\langle X_{d_{n+1},n+1} - \widehat{X}_{d_{n+1},n+1}, X_{d_{k+1},k+1} - \widehat{X}_{d_{k+1},k+1}\rangle = 0$ for $k \neq n$ (see Brockwell and Davis [13], Eq. (11.4.24)). Now with the definition of $V_{d_{k+1},k}$ in (2.5.5),

$$\mathbb{E}\big[(X_{d_{k+1},k+1} - \widehat{X}_{d_{k+1},k+1}) \otimes X_{d_{n+1},n+1}\big] = \theta_{d_{k+1},n,n-k} V_{d_{k+1},k}. \tag{2.5.6}$$

By representation (2.5.3) for $n = k$ and the fact that $V_{d_{k+1},k}$ is finite-dimensional and therefore invertible, since all eigenvalues of C_X are positive,

$$\theta_{d_{k+1},n,n-k} = \Big(P_{A_{d_{n+1}}} C_{X;n-k} P_{A_{d_{k+1}}} - \sum_{i=1}^{k} \mathbb{E}\big[(X_{d_i,i} - \widehat{X}_{d_i,i}) \otimes X_{d_{n+1},n+1}\big] \theta^*_{d_i,k,k-i-1}\Big) V^{-1}_{d_{k+1},k}.$$

However, with (2.5.6) the expectation on the right-hand-side can be replaced by $\theta_{d_i,n,n+1-i}V_{d_i,i-1}$, for $i = 1,\dots,k$, which leads to

$$\theta_{d_{k+1},n,n-k} = \Big(P_{A_{d_{n+1}}} C_{X;n-k} P_{A_{d_{k+1}}} - \sum_{i=1}^{k} \theta_{d_i,n,n+1-i} V_{d_i,i-1} \theta^*_{d_i,k,k-i-1}\Big) V^{-1}_{d_{k+1},k}.$$

Finally, the projection theorem gives

$$\begin{aligned} V_{d_{n+1},n} &= C_{X_{d_{n+1},n+1}-\widehat{X}_{d_{n+1},n+1}} = C_{X_{d_{n+1}}} - C_{\widehat{X}_{d_{n+1},n+1}} \\ &= C_{X_{d_{n+1}}} - \sum_{j=0}^{n-1} \theta_{d_{j+1},n,n-j} V_{d_{j+1},j} \theta^*_{d_{j+1},n,n-j}. \end{aligned}$$

□

Remark 2.5.2. Notice that $X_{d_1,1}, X_{d_2,2},\dots,X_{d_n,n}$ is not necessarily stationary. However, the recursions above can still be applied, since stationarity is not required for the application of the Innovations Algorithm in finite dimensions, see Proposition 11.4.2 in Brockwell and Davis [13]. □

If $(X_n)_{n\in\mathbb{Z}}$ is invertible, we can derive asymptotics for $\widehat{X}_{d_{n+1},n+1}$ as $d_n \to \infty$ and $n \to \infty$.

Theorem 2.5.3. *Let $(X_n)_{n\in\mathbb{Z}}$ be a stationary, invertible functional linear process with WN $(\varepsilon_n)_{n\in\mathbb{Z}}$ such that all eigenvalues of C_ε are positive. Assume furthermore that all eigenvalues λ_j, $j \in \mathbb{N}$, of C_X are positive.*
(i) Let $m_n \to \infty$, $m_n < n$ and $m_n/n \to 0$ for $n \to \infty$ and $d_n \to \infty$ for $n \to \infty$ be two positive increasing sequences in $\mathbb{N}$. Then

$$\mathbb{E}\|X_{n+1} - \widehat{X}_{d_{n+1},n+1} - \varepsilon_{n+1}\|^2 = O\Big(\sum_{j>m_n} \|\pi_j\|_{\mathcal{L}} + \sum_{j>d_{n-m_n}} \lambda_j\Big) \to 0, \quad n\to\infty. \tag{2.5.7}$$

(ii) Denote by $C_{\mathbf{X}_{d_n};h}$ the covariance matrix of the subprocess $(\mathbf{X}_{d_n})_{n\in\mathbb{Z}}$ as defined in Lemma 2.4.2. Then all eigenvalues of the spectral density matrix $f_{\mathbf{X}_{d_n}}[\omega] := \frac{1}{2\pi}\sum_{h\in\mathbb{Z}} e^{-ih\omega} C_{\mathbf{X}_{d_n};h}$ for $-\pi < \omega \le \pi$ are positive. Denote by $\alpha_{d_n} > 0$ the infimum of these eigenvalues. If

$$\frac{1}{\alpha_{d_n}}\Big(\sum_{j>m_n} \|\pi_j\|_{\mathcal{L}} + \sum_{j>d_{n-m_n}} \lambda_j\Big) \to 0, \quad n\to\infty. \tag{2.5.8}$$

then for $i = 1,\dots,n$ and for all $x \in H$,

$$\|(\theta_{d_n,n,i} - \gamma_i)(x)\| \to 0, \quad n\to\infty.$$

The proof of this Theorem is given in Section 2.6.

Remark 2.5.4. (a) Part (i) of Theorem 2.5.3 requires only that $d_n \to \infty$ as $n \to \infty$. No rate is required, and we do not impose any coupling condition of d_n with m_n. The theory would suggest to let d_n increase as fast as possible. In practice, when quantities such as the lagged covariance operators of the underlying process have to be estimated, the variance of the estimators of $P_{d_n} C_{X;h} P_{d_n}$ increases with d_n. In fact, for instance, for the estimation of $\theta_{d_1,1,1}$ the statistician is faced with the inversion of $P_{d_1} C_X P_{d_1}$. Small errors in the estimation of small empirical eigenvalues of $P_{d_1} C_X P_{d_1}$ may have severe consequences for the estimation of $\theta_{d_1,1,1}$. This suggests a conservative choice for d_n. The problem is similar to the choice of k_n in Chapter 9.2 of Bosq [8] concerned with the estimation of the autoregressive parameter operator in a FAR(1). The authors propose to choose k_n based on validation of the empirical prediction error. In Aue et al. [5] the authors suggest a functional FPE type criterion, which also aims at minimizing the prediction error over different choices of d_n.

(b) The choice of m_n in (2.5.7) allows us to calibrate two error terms: under the restriction that $m_n/n \to 0$, choosing a larger m_n increases $\sum_{j>d_{n-m_n}} \lambda_j$, the error caused by dimension reduction. Choosing a smaller m_n will on the other hand increase $\sum_{j>m_n} \|\pi_j\|$. □

2.6 Proofs

Before presenting a proof of Theorem 2.5.3 we give some notation and auxiliary results. Recall that throughout I_H denotes the identity operator on H. We also recall the notation and results provided in Section 2.2, which we shall use below without specific referencing.

Let $(X_n)_{n\in\mathbb{Z}}$ be a stationary functional linear process. Then for $n \in \mathbb{N}$ define the covariance operator of the vector $(X_n, \dots, X_1)$ by

$$\Gamma_n := \begin{pmatrix} \mathbb{E}[X_n \otimes X_n] & \mathbb{E}[X_n \otimes X_{n-1}] & \dots & \mathbb{E}[X_n \otimes X_1] \\ \mathbb{E}[X_{n-1} \otimes X_1] & \mathbb{E}[X_{n-1} \otimes X_{n-1}] & \dots & \vdots \\ \vdots & & \ddots & \\ \mathbb{E}[X_1 \otimes X_n] & \dots & & \mathbb{E}[X_1 \otimes X_1] \end{pmatrix}$$

$$= \begin{pmatrix} C_X & C_{X;1} & \dots & C_{X;n-1} \\ C_{X;-1} & C_X & \dots & \vdots \\ \vdots & & \ddots & \\ C_{X;-(n-1)} & \dots & & C_X \end{pmatrix}, \tag{2.6.1}$$

i.e., Γ_n is an operator acting on H^n, where H^n is the Cartesian product of n copies of H. Recall that H^n is again a Hilbert space, when equipped with the scalar product

$$\langle x, y\rangle_n = \sum_{i=1}^{n} \langle x_i, y_i\rangle$$

(see Bosq [8], Section 5 for details). As covariance operator of $(X_n, X_{n-1}, \dots, X_1)$, Γ_n is self-adjoint, nuclear, and has the spectral representation (see Theorem 5.1 in Gohberg et al. [21])

$$\Gamma_n = \sum_{j=1}^{\infty} \lambda_j^{(n)} \nu_j^{(n)} \otimes \nu_j^{(n)}, \quad n \in \mathbb{N},$$

with eigenpairs $(\lambda_j^{(n)}, \nu_j^{(n)})_{j\in\mathbb{N}}$.

Furthermore, define the operators $P_{(d_n)}$ and P_D acting on H^n by

$$P_{(d_n)} = \operatorname{diag}\big(P_{A_{d_n}}, \dots, P_{A_{d_1}}\big) \quad \text{and} \quad P_D = \operatorname{diag}\big(P_{A_D}, \dots, P_{A_D}\big). \tag{2.6.2}$$

Additionally, define the operators $\Gamma_{(d_n),n}$ and $\Gamma_{D,n}$ by

$$\Gamma_{(d_n),n} := P_{(d_n)} \Gamma_n P_{(d_n)} \quad \text{and} \quad \Gamma_{D,n} := P_D \Gamma_n P_D.$$

Note that $\Gamma_{(d_n),n}$ is in fact the covariance operator of $(X_{d_n,n}, \dots, X_{d_1,1})$ and has rank $k_n := \sum_{i=1}^{n} d_i$, whereas $\Gamma_{D,n}$ is the covariance operator of $(X_{D,n}, \dots, X_{D,1})$ and has rank $D\cdot n$. The operators $\Gamma_{(d_n),n}$ and $\Gamma_{D,n}$ are therefore self-adjoint nuclear operators with spectral representations

$$\Gamma_{(d_n),n} = \sum_{j=1}^{k_n} \lambda_{(d_n),j}^{(n)} e_{(d_n),j}^{(n)} \otimes e_{(d_n),j}^{(n)} \quad \text{and} \quad \Gamma_{D,n} = \sum_{j=1}^{D\cdot n} \lambda_{D,j}^{(n)} e_{D,j}^{(n)} \otimes e_{D,j}^{(n)}. \tag{2.6.3}$$

We need the following auxiliary results.

Lemma 2.6.1 (Theorem 1.2 in Mitchell [41])**.** *Let $(\mathbf{X}_{D,n})_{n\in\mathbb{Z}}$ be a $D-$variate stationary, invertible linear process satisfying*

$$\mathbf{X}_{D,n} = \sum_{i=1}^{\infty} \mathbf{\Psi}_i \mathbf{E}_{n-i} + \mathbf{E}_n, \quad n \in \mathbb{Z},$$

with $\sum_{i=1}^{\infty} \|\mathbf{\Psi}_i\|_2 < \infty$ ($\|\cdot\|_2$ denotes the Euclidean matrix norm) and WN $(\mathbf{E}_{D,n})_{n\in\mathbb{Z}}$ in $L^2_{\mathbb{R}^D}$ with non-singular covariance matrix $C_{\mathbf{E}_D}$. Let $C_{\mathbf{X}_D}$ be the covariance matrix of $(\mathbf{X}_{D,n})_{n\in\mathbb{Z}}$. Then the spectral density matrix $f_{\mathbf{X}_D}[\omega] := \frac{1}{2\pi}\sum_{h\in\mathbb{Z}} e^{-ih\omega} C_{\mathbf{X}_{D;h}}$ for $-\pi < \omega \le \pi$ has only positive eigenvalues. Let α_D be their infimum. Then the eigenvalues $(\lambda_{D,i}^{(n)})_{i=1,\dots,D\cdot n}$ of $\Gamma_{D,n}$ as in (2.6.3) are bounded below as follows:

$$0 < 2\pi\alpha_D \le \lambda_{D\cdot n}^{(n)} \le \ldots \le \lambda_1^{(n)}.$$

The following is a consequence of the Cauchy-Schwarz inequality.

Lemma 2.6.2. *For $j, \ell \in \mathbb{N}$ let (λ_j, ν_j) and (λ_ℓ, ν_ℓ) be eigenpairs of C_X. Then for $h \in \mathbb{Z}$,*

$$\langle C_{X;h}\nu_j, \nu_\ell\rangle \le \lambda_j^{1/2}\lambda_\ell^{1/2}. \tag{2.6.4}$$

Proof. With the definition of the lagged covariance operators in (2.2.4) and then the Cauchy-Schwarz inequality, we get by stationarity of $(X_n)_{n\in\mathbb{Z}}$

$$\begin{aligned}\langle C_{X;h}\nu_j, \nu_\ell\rangle &= \langle \mathbb{E}[\langle X_0, \nu_j\rangle X_h], \nu_\ell\rangle = \mathbb{E}[\langle X_0, \nu_j\rangle\langle X_h, \nu_\ell\rangle] \\ &\le \left(\mathbb{E}\langle X_0, \nu_j\rangle^2\right)^{1/2}\left(\mathbb{E}\langle X_h, \nu_\ell\rangle^2\right)^{1/2}.\end{aligned}$$

We find $\mathbb{E}\langle X_0, \nu_j\rangle^2 = \mathbb{E}\langle\langle X_0, \nu_j\rangle X_0, \nu_j\rangle = \langle C_X\nu_j, \nu_j\rangle = \lambda_j$, which implies (2.6.4). □

So far we only considered the real Hilbert space $H = L^2([0,1])$. There is a natural extension to the complex Hilbert space by defining the scalar product $\langle x, y\rangle = \int_0^1 x(t)\bar{y}(t)dt$ for complex valued functions $x, y : [0,1] \to \mathbb{C}$. As in Section 7.2 of Bosq [8], for a sequence $(\psi_j)_{j\in\mathbb{N}}$ of operators in $\mathcal{L}$, we define the *complex* operators

$$A[z] := \sum_{j=0}^{\infty} z^j \psi_j, \quad z \in \mathbb{C}, \tag{2.6.5}$$

such that the series converges in the operator norm. We need some methodology on frequency analysis of functional time series, recently studied in Panaretos and

Tavakoli [44]. The functional discrete Fourier transform of $(X_1, \dots, X_n)$ is defined by

$$S_n(\omega) = \sum_{j=1}^{n} X_j e^{-ij\omega}, \quad \omega \in (-\pi, \pi].$$

By Theorem 4 of Cerovecki and Hörmann [14], for all $\omega \in (-\pi, \pi]$, if $(X_n)_{n\in\mathbb{Z}}$ is a linear process with $\sum_{i=1}^{\infty} \|\psi_j\|_{\mathcal{L}} < \infty$, then $\frac{1}{\sqrt{n}} S_n(\omega)$ converges in distribution as $n \to \infty$ to a complex Gaussian random element with covariance operator

$$2\pi \mathcal{F}_X[\omega] := \sum_{h\in\mathbb{Z}} C_{X;h} e^{-ih\omega}.$$

The *spectral density operator* $\mathcal{F}_X[\omega]$ of $(X_n)_{n\in\mathbb{Z}}$ is non-negative, self-adjoint and nuclear (see Proposition 2.1 in Panaretos and Tavakoli [44]).

Theorem 1 and 4 of Cerovecki and Hörmann [14] infer the following duality between $C_{X;h}$ and $\mathcal{F}_X[\omega]$, with $A[z]$ as in (2.6.5) and adjoint $A[z]^*$:

$$\begin{aligned} C_{X;h} &= \int_{-\pi}^{\pi} \mathcal{F}_X[\omega] e^{ih\omega} d\omega, \quad h \in \mathbb{Z} \quad \text{and} \\ \mathcal{F}_X[\omega] &= \frac{1}{2\pi} A[e^{-i\omega}] C_\varepsilon A[e^{-i\omega}]^*, \quad \omega \in (-\pi, \pi]. \end{aligned} \tag{2.6.6}$$

The following Lemma is needed for the subsequent proofs, but may also be of interest by itself.

Lemma 2.6.3. *Let $(X_n)_{n\in\mathbb{Z}}$ be a stationary, invertible functional linear process with WN $(\varepsilon_n)_{n\in\mathbb{Z}}$, such that all eigenvalues of C_ε are positive. Then for all $\omega \in (-\pi, \pi]$ the spectral density operator $\mathcal{F}_X[\omega]$ has only positive eigenvalues.*

Proof. The proof is an extension of the proof of Theorem 3.1 in Nsiri and Roy [43] to the infinite-dimensional setting. Define for $A[z]$ as in (2.6.5) and $(\pi_i)_{i\in\mathbb{N}}$ as in (2.2.5)

$$P[z] := \sum_{j=0}^{\infty} z^j \pi_j \quad \text{and} \quad D[z] := P[z]A[z], \quad z \in \mathbb{C}.$$

Since $A[z]$ and $P[z]$ are power-series, $D[z]$ can also be represented as a power-series by

$$D[z] = \sum_{j=0}^{\infty} z^j \delta_j, \quad z \in \mathbb{C},$$

for $\delta_j \in \mathcal{L}$ for all $j \in \mathbb{N}$. Let B be the backshift operator. Then $X_n = A[B]\varepsilon_n$ and $\varepsilon_n = P[B]X_n$; in particular,

$$\varepsilon_n = P[B]X_n = P[B]A[B]\varepsilon_n = D[B]\varepsilon_n, \qquad n \in \mathbb{Z}. \tag{2.6.7}$$

Since all eigenvalues of C_ε are positive, by equating the coefficients in (2.6.7), $D[z] = I_H$ for all $z \in \mathbb{C}$.

Assume that there exists some non-zero $v \in H$ such that $\mathcal{F}_X[\omega](v) = 0$. Then by (2.6.6),

$$\frac{1}{2\pi}A[e^{-i\omega}]C_\varepsilon A[e^{-i\omega}]^*(v) = 0.$$

But since all eigenvalues of C_ε are positive, there exists some non-zero $u \in H$ such that $A[e^{i\omega}](u) = 0$. However, since $D[z] = P[z]A[z] = I_H$ for all $z \in \mathbb{C}$, this is a contradiction, and $\mathcal{F}_X[\omega]$ can only have positive eigenvalues for all $\omega \in (-\pi, \pi]$. □

2.6.1 Proof of Proposition 2.4.3

Stationarity of $(X_{D,n})_{n\in\mathbb{Z}}$ follows immediately from stationarity of $(X_n)_{n\in\mathbb{Z}}$, since P_{A_D} is a linear shift-invariant transformation. The functional Wold decomposition (see Definition 3.1 in Bosq [9]) gives a representation of $(X_{D,n})_{n\in\mathbb{Z}}$ as a linear process with WN, say $(\tilde{\varepsilon}_n)_{n\in\mathbb{Z}}$ in L^2_H. By Lemma 2.4.2, $(X_{D,n})_{n\in\mathbb{Z}}$ is isometrically isomorphic to the vector process $(\mathbf{X}_{D,n})_{n\in\mathbb{Z}}$ as in (2.4.6). Analogously, $(\tilde{\mathbf{E}}_{D,n})_{n\in\mathbb{Z}}$ defined by $\tilde{\mathbf{E}}_{D,n} := (\langle \tilde{\varepsilon}_{D,n}, \nu_1\rangle, \ldots, \langle \tilde{\varepsilon}_{D,n}, \nu_D\rangle)^\top$ is isometrically isomorphic to $(\tilde{\varepsilon}_{D,n})_{n\in\mathbb{Z}}$. We give a representation of $(\tilde{\mathbf{E}}_{D,n})_{n\in\mathbb{Z}}$.

Define $\mathcal{M}_{D,n} = \overline{\text{sp}}\{\mathbf{X}_{D,t}, -\infty < t \le n\}$. Then, from the multivariate Wold decomposition, the WN of $(\mathbf{X}_{D,n})_{n\in\mathbb{Z}}$ in $L^2_{\mathbb{R}^D}$ is defined by

$$\tilde{\mathbf{E}}_{D,n} = \mathbf{X}_{D,n} - P_{\mathcal{M}_{D,n-1}}(\mathbf{X}_{D,n}), \quad n \in \mathbb{Z}. \tag{2.6.8}$$

Now recall (2.4.7) in the following form

$$\mathbf{X}_{D,n} = \mathbf{E}_{D,n} + \sum_{j=1}^{\infty} \mathbf{\Psi}_{D,j}\mathbf{E}_{D,n-j} + \mathbf{\Psi}^{\infty}_{D,1}\mathbf{E}^{\infty}_{n-1} + \sum_{j=2}^{\infty} \mathbf{\Psi}^{\infty}_{D,j}\mathbf{E}^{\infty}_{n-j}, \quad n \in \mathbb{Z}.$$

We apply the projection operator onto $\mathcal{M}_{D,n-1}$ to all terms. $P_{\mathcal{M}_{D,n-1}}(\mathbf{E}_{D,n}) = 0$, and $\mathbf{E}_{D,n-j}$ and $\mathbf{E}^{\infty}_{n-j-1}$ belong to $\mathcal{M}_{D,n-1}$ for all $j \ge 1$. Hence,

$$P_{\mathcal{M}_{D,n-1}}(\mathbf{X}_{D,n}) = \sum_{j=1}^{\infty} \mathbf{\Psi}_{D,j}\mathbf{E}_{D,n-j} + \sum_{j=2}^{\infty} \mathbf{\Psi}^{\infty}_{D,j}\mathbf{E}^{\infty}_{n-j} + \mathbf{\Psi}^{\infty}_{D,1}P_{\mathcal{M}_{D,n-1}}(\mathbf{E}^{\infty}_{n-1}), \quad n \in \mathbb{Z},$$

which together with (2.6.8) implies (2.4.8).

We now show that $(\mathbf{X}_{D,n})_{n\in\mathbb{Z}}$ is invertible. The Wold decomposition gives the following representation

$$\mathbf{X}_{D,n} = \sum_{i=1}^{\infty} \tilde{\mathbf{\Psi}}_{D,i}(\tilde{\mathbf{E}}_{D,n-i}) + \tilde{\mathbf{E}}_{D,n}, \quad n \in \mathbb{Z}, \tag{2.6.9}$$

for appropriate $\tilde{\mathbf{\Psi}}_{D,i}$ and innovation process as in (2.6.8). Theorem 1 of Nsiri and Roy [43] gives conditions for the invertibility of the stationary D-variate linear process $(\mathbf{X}_{D,n})_{n\in\mathbb{Z}}$ satisfying (2.6.9).

We verify these conditions one by one.

(1) We start by showing that for all $\omega \in (-\pi, \pi]$ the matrix $\mathcal{F}_{\mathbf{X}_D}[\omega]$ is invertible, equivalently, $\langle \mathcal{F}_{\mathbf{X}_D}[\omega]\mathbf{x}, \mathbf{x}\rangle_{\mathbb{R}^D} > 0$ for all non-zero $\mathbf{x} \in \mathbb{R}^D$. By the isometric isomorphy between $\mathbb{R}^D$ and A_D from Lemma 2.4.2 we have

$$\left\langle \mathcal{F}_{\mathbf{X}_D}[\omega]\mathbf{x}, \mathbf{x}\right\rangle_{\mathbb{R}^D} = \left\langle \mathcal{F}_{X_D}[\omega]x, x\right\rangle.$$

By (2.4.9) the spectral density operator $\mathcal{F}_{X_D}[\omega]$ of $(X_{D,n})_{n\in\mathbb{Z}}$ satisfies

$$\begin{aligned}\mathcal{F}_{X_D}[\omega] &= \frac{1}{2\pi}\sum_{h\in\mathbb{Z}} C_{X_D;h}e^{-ih\omega} = \frac{1}{2\pi}\sum_{h\in\mathbb{Z}} P_{A_D}C_{X;h}P_{A_D}e^{-ih\omega} \\ &= P_{A_D}\Big(\frac{1}{2\pi}\sum_{h\in\mathbb{Z}} C_{X;h}e^{-ih\omega}\Big)P_{A_D} = P_{A_D}\mathcal{F}_X[\omega]P_{A_D}.\end{aligned} \tag{2.6.10}$$

However, since $(X_n)_{n\in\mathbb{Z}}$ is invertible and all eigenvalues of C_ε are positive, by Lemma 2.6.3 all eigenvalues of $\mathcal{F}_X[\omega]$ are positive for all $\omega \in (-\pi, \pi]$. Using first (2.6.10), then $x \in A_D$, and finally that all eigenvalues of $\mathcal{F}_X[\omega]$ are positive, together with $P_{A_D} = P_{A_D}^*$, we get

$$\left\langle \mathcal{F}_{X_D}[\omega]x, x\right\rangle = \left\langle P_{A_D}\mathcal{F}_X[\omega]P_{A_D}x, x\right\rangle = \left\langle \mathcal{F}_X[\omega]x, x\right\rangle > 0.$$

Hence, $\langle \mathcal{F}_{\mathbf{X}_D}[\omega]\mathbf{x}, \mathbf{x}\rangle_{\mathbb{R}^D} > 0$ and thus $\mathcal{F}_{X_D}[\omega]$ is invertible.

(2) We next show that the covariance matrix $C_{\tilde{\mathbf{E}}_D}$ of $(\tilde{\mathbf{E}}_{D,n})_{n\in\mathbb{Z}}$ as in (2.4.8) is invertible. Since $\mathbf{E}_{D,n}$ and $\mathbf{\Delta}_{D,n-1}$ from (2.4.8) are uncorrelated, $C_{\tilde{\mathbf{E}}_D} = C_{\mathbf{E}_D} + C_{\mathbf{\Delta}_D}$. All eigenvalues of C_ε are positive by assumption. For all $x \in A_D$ we get $\langle x, C_\varepsilon x\rangle = \langle \mathbf{x}, C_{\mathbf{E}_D}\mathbf{x}\rangle_{\mathbb{R}^d}$ where x and $\mathbf{x}$ are related by the isometric isomorphism T of Lemma 2.4.2. With the characterization of the eigenvalues of a self-adjoint

operator via the Rayleigh quotient as in Theorem 4.2.7 in Hsing and Eubank [29], all eigenvalues of $C_{\mathbf{E}_D}$ are positive. Therefore, all eigenvalues of $C_{\tilde{\mathbf{E}}_D} = C_{\mathbf{E}_D} + C_{\mathbf{\Delta}_D}$ are positive, and $C_{\tilde{\mathbf{E}}_D}$ is invertible.
(3) Finally, summability in Euclidean matrix norm of the matrices $\tilde{\mathbf{\Psi}}_{D,i}$ over $i \in \mathbb{N}$ follows from the properties of the Wold decomposition (see Theorem 5.7.1 in Brockwell and Davis [13]) and from the summability of $\|\psi_i\|_{\mathcal{L}}$ over $i \in \mathbb{N}$.

Therefore, all conditions of Theorem 1 of Nsiri and Roy [43] are satisfied and $(\mathbf{X}_{D,n})_{n\in\mathbb{Z}}$ is invertible. □

2.6.2 Proof of Theorem 2.5.3 (i)

First note that by the projection theorem (see Theorem 2.3.1 in Brockwell and Davis [13]),

$$\mathbb{E}\|X_{n+1} - \widehat{X}_{d_{n+1},n+1}\|^2 \le \mathbb{E}\|X_{n+1} - \sum_{i=1}^{n} \eta_i X_{d_{n+1-i},n+1-i}\|^2, \quad n \in \mathbb{N}, \tag{2.6.11}$$

for all $\eta_i \in \mathcal{L}$, $i = 1, \dots, n$. Hence, (2.6.11) holds in particular for $\eta_i = \pi_i$ for $i = 1, \dots, n$, where π_i are the operators in the invertible representation of $(X_n)_{n\in\mathbb{Z}}$ of (2.2.5). Furthermore, by the orthogonality of ε_{n+1} and X_k for $k < n+1$ and $n \in \mathbb{N}$, and since $\mathbb{E}\langle X_{n+1}, \varepsilon_{n+1}\rangle = \mathbb{E}\|\varepsilon_{n+1}\|^2$, we get

$$\mathbb{E}\|X_{n+1} - \widehat{X}_{d_{n+1},n+1}\|^2 = \mathbb{E}\|X_{n+1} - \widehat{X}_{d_{n+1},n+1} - \varepsilon_{n+1}\|^2 + \mathbb{E}\|\varepsilon_{n+1}\|^2$$

Now (2.6.11) with $\eta_i = \pi_i$ and then the invertibility of $(X_n)_{n\in\mathbb{Z}}$ yield

$$\begin{aligned}
\mathbb{E}\|X_{n+1} - \widehat{X}_{d_{n+1},n+1} - \varepsilon_{n+1}\|^2 &\le \mathbb{E}\|X_{n+1} - \sum_{i=1}^{n} \pi_i X_{d_{n+1-i},n+1-i}\|^2 - \mathbb{E}\|\varepsilon_{n+1}\|^2 \\
&= \mathbb{E}\|\sum_{i=1}^{\infty} \pi_i X_{n+1-i} + \varepsilon_{n+1} - \sum_{i=1}^{n} \pi_i X_{d_{n+1-i},n+1-i}\|^2 - \mathbb{E}\|\varepsilon_{n+1}\|^2 \\
&= \mathbb{E}\|\sum_{i=1}^{n} \pi_i (X_{n+1-i} - X_{d_{n+1-i},n+1-i}) + \varepsilon_{n+1} + \sum_{i>n} \pi_i X_{n+1-i}\|^2 \\
&\quad - \mathbb{E}\|\varepsilon_{n+1}\|^2.
\end{aligned}$$

Again by the orthogonality of ε_{n+1} and X_k, for $k < n+1$, since $X_{d_n,n} = P_{A_{d_n}} X_n$, and then using that for $X, Y \in L^2_H$, $\mathbb{E}\|X+Y\|^2 \leq 2\mathbb{E}\|X\|^2 + 2\mathbb{E}\|Y\|^2$, we get

$$\begin{aligned}\mathbb{E}\|X_{n+1} - \widehat{X}_{d_{n+1},n+1} - \varepsilon_{n+1}\|^2 &\leq \mathbb{E}\|\sum_{i=1}^{n} \pi_i (I_H - P_{A_{d_{n+1-i}}}) X_{n+1-i} + \sum_{i>n} \pi_i X_{n+1-i}\|^2 \\ &\leq 2\mathbb{E}\|\sum_{i=1}^{n} \pi_i (I_H - P_{A_{d_{n+1-i}}}) X_{n+1-i}\|^2 + 2\mathbb{E}\|\sum_{i>n} \pi_i X_{n+1-i}\|^2 \qquad (2.6.12) \\ &=: 2J_1 + 2J_2.\end{aligned}$$

We consider the two terms in (2.6.12) separately. From (2.2.1) we get for the first term in (2.6.12)

$$J_1 = \Big\|\mathbb{E}\Big[\sum_{i=1}^{n} \pi_i (I_H - P_{A_{d_{n+1-i}}}) X_{n+1-i} \otimes \sum_{i=1}^{n} \pi_i (I_H - P_{A_{d_{n+1-i}}}) X_{n+1-i}\Big]\Big\|_{\mathcal{N}}.$$

Using the triangle inequality together with properties of the nuclear operator norm given in Section 2.2, and then the definition of $C_{X;h}$ in (2.2.4), we calculate

$$\begin{aligned}J_1 &\leq \sum_{i,j=1}^{n} \|\pi_i\|_{\mathcal{L}} \|\pi_j\|_{\mathcal{L}} \Big\|\mathbb{E}\big[(I_H - P_{A_{d_{n+1-i}}}) X_{n+1-i} \otimes (I_H - P_{A_{d_{n+1-j}}}) X_{n+1-j}\big]\Big\|_{\mathcal{N}} \\ &= \sum_{i,j=1}^{n} \|\pi_i\|_{\mathcal{L}} \|\pi_j\|_{\mathcal{L}} \Big\|(I_H - P_{A_{d_{n+1-i}}}) C_{X;i-j} (I_H - P_{A_{d_{n+1-j}}})\Big\|_{\mathcal{N}} \\ &=: \sum_{i,j=1}^{n} \|\pi_i\|_{\mathcal{L}} \|\pi_j\|_{\mathcal{L}} K(i,j). \qquad (2.6.13)\end{aligned}$$

By the definition of A_d in (2.5.1) and, since by (2.4.5) we have $(I_H - P_{A_{d_i}}) = \sum_{\ell > d_i} \nu_\ell \otimes \nu_\ell$,

$$\begin{aligned}K(i,j) &= \Big\|\Big(\sum_{\ell' > d_{n+1-i}} \nu_{\ell'} \otimes \nu_{\ell'}\Big) C_{X;i-j} \Big(\sum_{\ell > d_{n+1-j}} \nu_\ell \otimes \nu_\ell\Big)\Big\|_{\mathcal{N}} \\ &= \Big\|\sum_{\ell' > d_{n+1-i}} \sum_{\ell > d_{n+1-j}} \langle C_{X;i-j}(\nu_\ell), \nu_{\ell'}\rangle \nu_\ell \otimes \nu_{\ell'}\Big\|_{\mathcal{N}}.\end{aligned}$$

With Lemma 2.6.2, the definition of the nuclear norm given in Section 2.2 and the orthogonality of the $(\nu_i)_{i\in\mathbb{N}}$, we get

$$\begin{aligned} K(i,j) &\le \Big\| \sum_{\ell'>d_{n+1-i}} \sum_{\ell>d_{n+1-j}} \lambda_\ell^{1/2}\lambda_{\ell'}^{1/2} \nu_\ell \otimes \nu_{\ell'} \Big\|_{\mathcal{N}} \\ &= \sum_{k=1}^{\infty} \Big\langle \sum_{\ell'>d_{n+1-i}} \sum_{\ell>d_{n+1-j}} \lambda_\ell^{1/2}\lambda_{\ell'}^{1/2} \nu_\ell \otimes \nu_{\ell'}(\nu_k), \nu_k \Big\rangle \\ &= \sum_{k>\max(d_{n+1-j},d_{n+1-i})} \lambda_k \ \le \sum_{k>d_{n+1-j}} \lambda_k. \end{aligned} \tag{2.6.14}$$

Plugging (2.6.14) into (2.6.13), and recalling that $\sum_{i=1}^{\infty} \|\pi_i\|_{\mathcal{L}} =: M_1 < \infty$, we conclude

$$J_1 \le M_1 \sum_{j=1}^{n} \|\pi_j\|_{\mathcal{L}} \sum_{\ell>d_{n+1-j}} \lambda_\ell. \tag{2.6.15}$$

Now for some $m_n < n$,

$$\sum_{j=1}^{n} \|\pi_j\|_{\mathcal{L}} \sum_{\ell>d_{n+1-j}} \lambda_\ell = \sum_{j=1}^{m_n} \|\pi_j\|_{\mathcal{L}} \sum_{\ell>d_{n+1-j}} \lambda_\ell \ + \sum_{j=m_n}^{n} \|\pi_j\|_{\mathcal{L}} \sum_{\ell>d_{n+1-j}} \lambda_\ell. \tag{2.6.16}$$

Since $\sum_{j=1}^{m_n} \|\pi_j\|_{\mathcal{L}} \le \sum_{j=1}^{\infty} \|\pi_j\|_{\mathcal{L}} = M_1 < \infty$, the first term on the right-hand side of (2.6.16) can be bounded by

$$\sum_{j=1}^{m_n} \|\pi_j\|_{\mathcal{L}} \sum_{\ell>d_{n+1-j}} \lambda_\ell \le M_1 \sum_{\ell>d_{n+1-m_n}} \lambda_\ell. \tag{2.6.17}$$

Furthermore, since $\sum_{\ell>d_{n+1-j}} \lambda_\ell \le \sum_{\ell=1}^{\infty} \lambda_\ell = \|C_X\|_{\mathcal{N}} < \infty$, the second term of the right-hand-side in (2.6.16) can be bounded by

$$\sum_{j=m_n}^{n} \|\pi_j\|_{\mathcal{L}} \sum_{\ell>d_{n+1-j}} \lambda_\ell \le \|C_X\|_{\mathcal{N}} \sum_{j=m_n}^{n} \|\pi_j\|_{\mathcal{L}}. \tag{2.6.18}$$

Hence, from (2.6.15) together with (2.6.16), (2.6.17) and (2.6.18) we obtain

$$J_1 = O\Big(\sum_{j=m_n}^{n} \|\pi_j\|_{\mathcal{L}} + \sum_{\ell>d_{n+1-m_n}} \lambda_\ell \Big). \tag{2.6.19}$$

Concerning J_2, the second term of (2.6.12) with (2.2.2), and then the definition of $C_{X;h}$ in (2.2.4) yield

$$\begin{aligned} J_2 = \mathbb{E}\Big\| \sum_{i>n} \pi_i X_{n+1-i} \Big\|^2 &= \Big\| \mathbb{E}\Big[\sum_{i>n} \pi_i X_{n+1-i} \otimes \sum_{j>n} \pi_j X_{n+1-j} \Big] \Big\|_{\mathcal{N}} \\ &= \Big\| \sum_{i,j>n} \pi_i C_{X;i-j} \pi_j^* \Big\|_{\mathcal{N}} \le \sum_{i,j>n} \|\pi_i\|_{\mathcal{L}} \|\pi_j\|_{\mathcal{L}} \|C_{X;i-j}\|_{\mathcal{N}}. \end{aligned}$$

Since $C_{X;i-j} \in \mathcal{N}$ for all $i, j \in \mathbb{N}$, there exists a constant $M_2 < \infty$ such that $\|C_{X;i-j}\|_{\mathcal{N}} < M_2$ for all $i, j \in \mathbb{N}$ and for some $m_n < n$,

$$J_2 \le M_2\Big(\sum_{i>n} \|\pi_i\|_{\mathcal{L}}\Big)^2 = O\Big(\sum_{i>m_n} \|\pi_i\|_{\mathcal{L}}\Big). \tag{2.6.20}$$

Finally the combination of (2.6.12), (2.6.19) and (2.6.20) yields assertion (i). □

2.6.3 Proof of Theorem 2.5.3 (ii)

Note first that by the projection theorem there is an equivalent representation of $\widehat{X}_{d_{n+1},n+1}$ to (2.5.3) given by

$$\widehat{X}_{d_{n+1},n+1} = P_{\mathrm{LCS}(F'_{d_n,n})}(X_{n+1}) = \sum_{i=1}^{n} \beta_{d_{n+1-i},n,i} X_{d_{n+1-i},n+1-i} \tag{2.6.21}$$

for $F'_{d_n,n}$ as in (2.5.2) and $\beta_{d_{n+1-i},n,i} \in \mathcal{L}$ for $i = 1, \dots, n$. Furthermore, for $k = 1, \dots, n$, we define the best linear predictor of X_{n+1} based on $F'_{d_n,n}(k) = \{X_{d_{n+1-k},n+1-k}, X_{d_{n-k+2},n+2-k}, \dots, X_{d_n,n}\}$ by

$$\widehat{X}_{d_{n+1},n+1}(k) = P_{\mathrm{LCS}(F'_{d_n,n}(k))}(X_{n+1}) = \sum_{i=1}^{k} \beta_{d_{n+1-i},k,i} X_{d_{n+1-i},n+1-i}. \tag{2.6.22}$$

We start with the following Proposition, which is an infinite-dimensional extension to Proposition 2.2 in Mitchell and Brockwell [42].

Proposition 2.6.4. *Under the assumptions of Theorem 2.5.3 the following assertions hold:*
(i) The operators $\beta_{d_{n+1-i},n,i}$ from (2.6.21) and $\theta_{d_{n+1-i},n,i}$ from (2.5.3) are for $n \in \mathbb{N}$ related by

$$\theta_{d_{n+1-i},n,i} = \sum_{j=1}^{i} \beta_{d_{n+1-j},n,j} \theta_{d_{n+1-i},n-j,i-j}, \quad i = 1, \dots, n. \tag{2.6.23}$$

Furthermore, for every $i, j \in \mathbb{N}$ and $x \in H$, as $n \to \infty$,
(ii) $\big\|(\beta_{d_{n+1-i},n,i} - \pi_i)(x)\big\| \to 0,$
(iii) $\big\|(\beta_{d_{n+1-i},n,i} - \beta_{d_{n+1-i-j},n-j,i})(x)\big\| \to 0,$
(iv) $\big\|(\theta_{d_{n+1-i},n,i} - \theta_{d_{n+1-i-j},n-j,i})(x)\big\| \to 0.$

Proof. **(i)** Set $\theta_{d_{n+1},n,0} := I_H$. By adding the term $\theta_{d_{n+1},n,0}(X_{d_{n+1},n+1} - \widehat{X}_{d_{n+1},n+1})$ to both sides of (2.5.3), we get

$$X_{d_{n+1},n+1} = \sum_{j=0}^{n} \theta_{d_{n+1-j},n,j}(X_{d_{n+1-j},n+1-j} - \widehat{X}_{d_{n+1-j},n+1-j}), \quad n \in \mathbb{N}.$$

Plugging this representation of $X_{d_{n+1-i},n+1-i}$ into (2.6.21) for $i = 1, \dots, n$ yields

$$\widehat{X}_{d_{n+1},n+1} = \sum_{i=1}^{n} \beta_{d_{n+1-i},n,i}\Big(\sum_{j=0}^{n-i} \theta_{d_{n+1-i-j},n-i,j}(X_{d_{n+1-i-j},n+1-i-j} - \widehat{X}_{d_{n+1-i-j},n+1-i-j})\Big).$$

Equating the coefficients of the innovations $(X_{d_{n+1-i},n+1-i} - \widehat{X}_{d_{n+1-i},n+1-i})$ with the innovation representation (2.5.3) leads by linearity of the operators to (2.6.23).
(ii) Let

$$B_{(d_n),n} = (\beta_{d_n,n,1}, \dots, \beta_{d_1,n,n}) \quad \text{and} \quad \Pi_n = (\pi_1, \dots, \pi_n), \tag{2.6.24}$$

which are both operators from H^n to H defined as follows: let $x = (x_1, \dots, x_n) \in H^n$ with $x_i \in H$ for $i = 1, \dots, n$. Then $B_{(d_n),n}\ x = \sum_{i=1}^n \beta_{d_{n+1-i},n,i} x_i \ \in H$. By definition of the norm in H^n we have for all $x \in H$

$$\|(B_{(d_n),n} - \Pi_n)(x)\| = \sum_{i=1}^{n} \|(\beta_{d_{n+1-i},n,i} - \pi_i)(x_i)\|.$$

We show that this tends to 0 as $n \to \infty$, which immediately gives $\|(\beta_{d_{n+1-i},n,i} - \pi_i)(x_i)\| \to 0$ for all $i \in \mathbb{N}$. First notice that for $x \in H^n$ and with $P_{(d_n)}$ defined in (2.6.2), the triangular inequality yields

$$\begin{aligned}\|(B_{(d_n),n} - \Pi_n)(x)\| &\leq \|(B_{(d_n),n} - \Pi_n P_{(d_n)})(x)\| + \|\Pi_n(I_{H^n} - P_{(d_n)})(x)\| \\ &=: J_1(d_n, n)(x) + J_2(d_n, n)(x),\end{aligned}$$

with identity operator I_{H^n} on H^n. We find bounds for $J_1(d_n, n)(x)$ and $J_2(d_n, n)(x)$. Since uniform convergence implies pointwise convergence, we consider the operator norm of $J_1(d_n, n)$

$$J_1(d_n, n) = \|B_{(d_n),n} - \Pi_n P_{(d_n)}\|_{\mathcal{L}}$$

and show that $J_1(d_n, n) \to 0$ as $n \to \infty$. From Theorem 2.1.8 in Simon [50] we find

$$\|B_{(d_n),n} - \Pi_n P_{(d_n)}\|_{\mathcal{L}}^2 = \|(B_{(d_n),n} - \Pi_n P_{(d_n)})(B_{(d_n),n} - \Pi_n P_{(d_n)})^*\|_{\mathcal{L}}. \tag{2.6.25}$$

Recall the spectral representation of $\Gamma_{(d_n),n}$ as in (2.6.3). By the definition of $B_{(d_n),n}$ and $\Pi_n P_{(d_n)}$, note that $(B_{(d_n),n} - \Pi_n P_{(d_n)})P_{(d_n)} = B_{(d_n),n} - \Pi_n P_{(d_n)}$. Extracting the smallest positive eigenvalue $\lambda^{(n)}_{(d_n),k_n}$ of $\Gamma_{(d_n),n}$, we get

$$\begin{aligned}
&\Big\|(B_{(d_n),n} - \Pi_n P_{(d_n)})\Gamma_{(d_n),n}(B_{(d_n),n} - \Pi_n P_{(d_n)})^*\Big\|_{\mathcal{L}} \\
&\qquad = \Big\|(B_{(d_n),n} - \Pi_n P_{(d_n)})\sum_{j=1}^{k_n}\lambda^{(n)}_{(d_n),j}(e^{(n)}_{(d_n),j}\otimes e^{(n)}_{(d_n),j})(B_{(d_n),n} - \Pi_n P_{(d_n)})^*\Big\|_{\mathcal{L}} \\
&\qquad \geq \lambda^{(n)}_{(d_n),k_n}\Big\|(B_{(d_n),n} - \Pi_n P_{(d_n)})(B_{(d_n),n} - \Pi_n P_{(d_n)})^*\Big\|_{\mathcal{L}}. \qquad (2.6.26)
\end{aligned}$$

Since $A_{d_i} \subseteq A_{d_n}$ for all $i \leq n$ we obtain $\mathbf{A}_{(d_n)} := (A_{d_n}, A_{d_{n-1}}, \dots, A_{d_1}) \subseteq \mathbf{A}_{d_n} := (A_{d_n}, A_{d_n}, \dots, A_{d_n})$ and, therefore, $P_{d_n}P_{(d_n)} = P_{(d_n)}$. Together with the definition of $\Gamma_{(d_n),n}$ this implies

$$\Gamma_{(d_n),n} = P_{(d_n)}\Gamma_n P_{(d_n)} = P_{(d_n)}P_{d_n}\Gamma_n P_{d_n}P_{(d_n)} = P_{(d_n)}\Gamma_{d_n,n}P_{(d_n)}.$$

Since $\langle x, \Gamma_{(d_n),n}x\rangle = \langle x, \Gamma_{d_n,n}x\rangle$ for all $x \in \mathbf{A}_{d_n}$, and $\mathbf{A}_{(d_n)} \subseteq \mathbf{A}_{d_n}$, we get

$$\lambda^{(n)}_{(d_n),k_n} = \min_{x\in\mathbf{A}_{(d_n)}}\frac{\langle x, \Gamma_{(d_n),n}x\rangle}{\|x\|^2} = \min_{x\in\mathbf{A}_{(d_n)}}\frac{\langle x, \Gamma_{d_n,n}x\rangle}{\|x\|^2} \geq \min_{x\in\mathbf{A}_{d_n}}\frac{\langle x, \Gamma_{d_n,n}x\rangle}{\|x\|^2} = \lambda^{(n)}_{d_n,d_n\cdot n},$$

where the first and last equality hold by application of Theorem 4.2.7 in Hsing and Eubank [29]. Furthermore, by Lemma 2.6.1, $\lambda^{(n)}_{d_n,d_n\cdot n} \geq 2\pi\alpha_{d_n}$. Therefore,

$$\lambda^{(n)}_{d_n,k_n} \geq \lambda^{(n)}_{d_n,d_n\cdot n} \geq 2\pi\alpha_{d_n}. \qquad (2.6.27)$$

With (2.6.26) and (2.6.27), we get

$$\begin{aligned}
\|B_{(d_n),n} - \Pi_n P_{(d_n)}\|^2 &\leq \frac{1}{2\pi\alpha_{d_n}}\Big\|(B_{(d_n),n} - \Pi_n P_{(d_n)})\Gamma_{(d_n),n}(B_{(d_n),n} - \Pi_n P_{(d_n)})^*\Big\|_{\mathcal{L}} \\
&=: J_1'(d_n, n). \qquad (2.6.28)
\end{aligned}$$

Furthermore, since $\langle Ax, y\rangle = \langle x, A^*y\rangle$ for $A \in \mathcal{L}$ and $x, y \in H$, and by (2.6.24) and the structure of $\Gamma_{(d_n),n}$,

$$\begin{aligned}
&\Big\|(B_{(d_n),n} - \Pi_n P_{(d_n)})\Gamma_{(d_n),n}(B_{(d_n),n} - \Pi_n P_{(d_n)})^*\Big\|_{\mathcal{L}} \\
&\leq \Big\|\mathbb{E}\Big[\sum_{i=1}^{n}(\beta_{d_{n+1-i},n,i} - \pi_i P_{A_{d_{n+1-i}}})X_{d_{n+1-i},n+1-i} \\
&\quad \otimes \sum_{j=1}^{n}(\beta_{d_{n+1-j},n,j} - \pi_j P_{A_{d_{n+1-j}}})X_{d_{n-j+1},n-j+1}\Big]\Big\|_{\mathcal{L}}.
\end{aligned}$$

Now with (2.2.5) and (2.6.21) we get

$$\begin{aligned}
&\Big\|\mathbb{E}\Big[\sum_{i=1}^{n}(\beta_{d_{n+1-i},n,i}-\pi_i P_{A_{d_{n+1-i}}})X_{d_{n+1-i},n+1-i}\\
&\qquad\otimes\sum_{j=1}^{n}(\beta_{d_{n+1-j},n,j}-\pi_j P_{A_{d_{n+1-j}}})X_{d_{n-j+1},n-j+1}\Big]\Big\|_{\mathcal{L}}\\
&=\Big\|\mathbb{E}\Big[\Big(\widehat{X}_{d_{n+1},n+1}-X_{n+1}+\varepsilon_{n+1}+\sum_{i>n}\pi_i X_{n+1-i}+\sum_{i=1}^{n}\pi_i(I-P_{A_{d_{n+1-i}}})X_{n+1-i}\Big)\\
&\qquad\otimes\Big(\widehat{X}_{d_{n+1},n+1}-X_{n+1}+\varepsilon_{n+1}+\sum_{j>n}\pi_j X_{n+1-j}+\sum_{j=1}^{n}\pi_j(I-P_{A_{d_{n+1-j}}})X_{n+1-j}\Big)\Big]\Big\|_{\mathcal{L}}
\end{aligned}$$

With the triangular inequality, the above equation decomposes in the following four terms giving with (2.6.28):

$$\begin{aligned}
&2\pi\alpha_{d_n}\|B_{(d_n),n}-\Pi_n P_{(d_n)}\|^2\\
&\le\Big\|\mathbb{E}\Big[\big(\widehat{X}_{d_{n+1},n+1}-X_{n+1}+\varepsilon_{n+1}\big)\otimes\big(\widehat{X}_{d_{n+1},n+1}-X_{n+1}+\varepsilon_{n+1}\big)\Big]\Big\|_{\mathcal{L}}\\
&\quad+\Big\|\mathbb{E}\Big[\Big(\sum_{i>n}\pi_i X_{n+1-i}+\sum_{i=1}^{n}\pi_i(I-P_{A_{d_{n+1-i}}})X_{n+1-i}\Big)\\
&\qquad\otimes\Big(\sum_{j>n}\pi_j X_{n+1-j}+\sum_{j=1}^{n}\pi_j(I-P_{A_{d_{n+1-j}}})X_{n+1-j}\Big)\Big]\Big\|_{\mathcal{L}}\\
&\quad+\Big\|\mathbb{E}\Big[\Big(\widehat{X}_{d_{n+1},n+1}-X_{n+1}+\varepsilon_{n+1}\Big)\otimes\Big(\sum_{j>n}\pi_j X_{n+1-j}+\sum_{j=1}^{n}\pi_j(I-P_{A_{d_{n+1-j}}})X_{n+1-j}\Big)\Big]\Big\|_{\mathcal{L}}\\
&\quad+\Big\|\mathbb{E}\Big[\Big(\sum_{i>n}\pi_i X_{n+1-i}+\sum_{i=1}^{n}\pi_i(I-P_{A_{d_{n+1-i}}})X_{n+1-i}\Big)\otimes\Big(\widehat{X}_{d_{n+1},n+1}-X_{n+1}+\varepsilon_{n+1}\Big)\Big]\Big\|_{\mathcal{L}}.
\end{aligned}\tag{2.6.29}$$

Define

$$f(n,d_n,m_n):=\Big(\sum_{j>m_n}\|\pi_j\|_{\mathcal{L}}+\sum_{j>d_{n-m_n}}\lambda_j\Big).$$

By Theorem 2.5.3 the first term of (2.6.29) is of the order $f(n,d_n,m_n)$. The second term of (2.6.29) is of the same order by the calculations following (2.6.12). Concerning the remaining two terms, using first that $\|C_{X,Y}\|_{\mathcal{L}}\le\mathbb{E}\|X\|\|Y\|$, and then applying the Cauchy-Schwarz inequality gives

$$\begin{aligned}
&\Big\|\mathbb{E}\Big[\Big(\widehat{X}_{d_{n+1},n+1} - X_{n+1} + \varepsilon_{n+1}\Big) \otimes \Big(\sum_{j>n} \pi_j X_{n+1-j} + \sum_{j=1}^{n} \pi_j (I - P_{A_{d_{n+1-j}}}) X_{n+1-j}\Big)\Big]\Big\|_{\mathcal{L}}^2 \\
&\quad \le \Big(\mathbb{E}\Big\|\widehat{X}_{d_{n+1},n+1} - X_{n+1} + \varepsilon_{n+1}\Big\|_{\mathcal{L}} \\
&\qquad \Big\|\sum_{j>n} \pi_j X_{n+1-j} + \sum_{j=1}^{n} \pi_j (I - P_{A_{d_{n+1-j}}}) X_{n+1-j}\Big\|\Big)^2 \qquad (2.6.30)\\
&\quad \le \mathbb{E}\Big\|\widehat{X}_{d_{n+1},n+1} - X_{n+1} + \varepsilon_{n+1}\Big\|_{\mathcal{L}}^2 \mathbb{E}\Big\|\sum_{j>n} \pi_j X_{n+1-j} + \sum_{j=1}^{n} \pi_j (I - P_{A_{d_{n+1-j}}}) X_{n+1-j}\Big\|_{\mathcal{L}}^2.
\end{aligned}$$

Both terms are of the order $f(n, d_n, m_n)$ by Theorem 2.5.3(i). Hence, $\|B_{(d_n),n} - \Pi_n P_{(d_n)}\|^2$ is of the order $f(n, d_n, m_n)/\alpha_{d_n}$, and with the assumption (2.5.8),

$$J_1(d_n, n)^2 \to 0, \quad n \to \infty. \qquad (2.6.31)$$

We now estimate $J_2(d_n, n)(x)$, which we have to consider pointwise. For every $x = (x_1, \dots, x_n) \in H^n$ with $x_i \in H$ for $1 \le i \le n$ and $\|x\| \le 1$,

$$\begin{aligned}
J_2(d_n, n) &= \|\Pi_n (I - P_{(d_n)})(x)\| \\
&= \Big\|\Big(\pi_1 (I_H - P_{A_{d_n}}), \pi_2 (I_H - P_{A_{d_{n-1}}}), \dots, \pi_n (I_H - P_{A_{d_1}})\Big)(x)\Big\| \\
&= \sum_{i=1}^{n} \|\pi_i (I_H - P_{A_{d_{n+1-i}}})(x_i)\|.
\end{aligned}$$

(iii) Similarly to the proof of (ii), we start by defining for every $n \in \mathbb{N}$,

$$\widetilde{B}_{(d_n),n-j} := (\beta_{d_n,n-j,1}, \beta_{d_{n-1},n-j,2}, \dots, \beta_{d_{j+1},n-j,n-j}, 0_H, \dots, 0_H), \quad j = 1, \dots, n,$$

where the last j entries are 0_H, the null operator on H. Then $\widetilde{B}_{(d_n),n-j}$ is a bounded linear operator from H^n to H. Analogously to the beginning of the proof of (ii), we show that $\|\widetilde{B}_{(d_n),n} - \widetilde{B}_{(d_n),n-j}\|_{\mathcal{L}} \to 0$ for $n \to \infty$. With the same calculation as deriving (2.6.28) from (2.6.25), we obtain

$$\begin{aligned}
\|\widetilde{B}_{(d_n),n} - \widetilde{B}_{(d_n),n-j}\|_{\mathcal{L}}^2 &\le \frac{1}{2\pi\alpha_{d_n}} \|(\widetilde{B}_{(d_n),n} - \widetilde{B}_{(d_n),n-j})\Gamma_{(d_n),n}(\widetilde{B}_{(d_n),n} - \widetilde{B}_{(d_n),n-j})^*\|_{\mathcal{L}} \\
&=: \frac{1}{2\pi\alpha_{d_n}} \widetilde{J}_1'(d_n, n).
\end{aligned}$$

Applying the same steps as when bounding $J_1(d_n, n)$ in the proof of (ii), and setting $\beta_{d_{n+j},n,m} = 0$ for $m > n$, we obtain

$$\begin{aligned}\widetilde{J}_1'(d_n, n) &= \Big\| \mathbb{E}\Big[\Big(\sum_{i=1}^{n} (\beta_{d_{n-i+1},n,i} - \beta_{d_{n-i+1},n-j,i}) X_{d_{n+1-i},n+1-i} \Big) \\ &\qquad \otimes \Big(\sum_{\ell=1}^{n} (\beta_{d_{n-\ell+1},n,\ell} - \beta_{d_{n-\ell+1},n-j,\ell}) X_{d_{n+1-\ell},n+1-\ell} \Big) \Big] \Big\|_{\mathcal{L}} \\ &= \Big\| \mathbb{E}\Big[\Big(\widehat{X}_{d_{n+1},n+1} - \widehat{X}_{d_{n+1},n+1}(n-j) \Big) \otimes \Big(\widehat{X}_{d_{n+1},n+1} - \widehat{X}_{d_{n+1},n+1}(n-j) \Big) \Big] \Big\|_{\mathcal{L}},\end{aligned}$$

where $\widehat{X}_{d_{n+1},n+1}(k) = \sum_{\ell=1}^{k} \beta_{d_{n-\ell+1},k,\ell} X_{d_{n+1-\ell},n+1-\ell}$ is defined as in (2.6.22). By adding and subtracting $X_{d_{n+1},n+1} + \varepsilon_{n+1}$ and then using the linearity of the scalar product we get

$$\begin{aligned}&\widetilde{J}_1'(K, n) \\ &= \Big\| \mathbb{E}\Big[\Big((\widehat{X}_{d_{n+1},n+1} - X_{d_{n+1},n+1} - \varepsilon_{n+1}) - (\widehat{X}_{d_{n+1},n+1}(n-j) - X_{d_{n+1},n+1} - \varepsilon_{n+1}) \Big) \\ &\quad \otimes \Big((\widehat{X}_{d_{n+1},n+1} - X_{d_{n+1},n+1} - \varepsilon_{n+1}) - (\widehat{X}_{d_{n+1},n+1}(n-j) - X_{d_{n+1},n+1} - \varepsilon_{n+1}) \Big) \Big] \Big\|_{\mathcal{L}} \\ &\le \Big\| \mathbb{E}\Big[\Big(\widehat{X}_{d_{n+1},n+1} - X_{d_{n+1},n+1} - \varepsilon_{n+1} \Big) \otimes \Big(\widehat{X}_{d_{n+1},n+1} - X_{d_{n+1},n+1} - \varepsilon_{n+1} \Big) \Big] \Big\|_{\mathcal{L}} \\ &\quad + \Big\| \mathbb{E}\Big[\Big(\widehat{X}_{d_{n+1},n+1}(n-j) - X_{d_{n+1},n+1} - \varepsilon_{n+1} \Big) \otimes \Big(\widehat{X}_{d_{n+1},n+1}(n-j) - X_{d_{n+1},n+1} - \varepsilon_{n+1} \Big) \Big] \Big\|_{\mathcal{L}} \\ &\quad + \Big\| \mathbb{E}\Big[\Big(\widehat{X}_{d_{n+1},n+1} - X_{d_{n+1},n+1} - \varepsilon_{n+1} \Big) \otimes \Big(\widehat{X}_{d_{n+1},n+1}(n-j) - X_{d_{n+1},n+1} - \varepsilon_{n+1} \Big) \Big] \Big\|_{\mathcal{L}} \\ &\quad + \Big\| \mathbb{E}\Big[\Big(\widehat{X}_{d_{n+1},n+1}(n-j) - X_{d_{n+1},n+1} - \varepsilon_{n+1} \Big) \otimes \Big(\widehat{X}_{d_{n+1},n+1} - X_{d_{n+1},n+1} - \varepsilon_{n+1} \Big) \Big] \Big\|_{\mathcal{L}}.\end{aligned}$$

For $n \to \infty$ the first term converges to 0 by Theorem 2.5.3 (i). For every fixed $j \in \{1, \ldots, n\}$ the second term converges to 0 by the exact same arguments. Similar arguments as in the proof of (ii) show that the third and fourth terms also converge to 0. Indeed, applying the Cauchy-Schwarz inequality, we find as in (2.6.30),

$$\begin{aligned}&\Big\| \mathbb{E}\Big[\Big(\widehat{X}_{d_{n+1},n+1} - X_{d_{n+1},n+1} - \varepsilon_{n+1} \Big) \otimes \Big(\widehat{X}_{d_{n+1},n+1}(n-j) - X_{d_{n+1},n+1} - \varepsilon_{n+1} \Big) \Big] \Big\|_{\mathcal{L}}^2 \\ &\qquad \le \mathbb{E} \Big\| \widehat{X}_{d_{n+1},n+1} - X_{d_{n+1},n+1} - \varepsilon_{n+1} \Big\|_{\mathcal{L}}^2 \, \mathbb{E} \Big\| \widehat{X}_{d_{n+1},n+1}(n-j) - X_{d_{n+1},n+1} - \varepsilon_{n+1} \Big\|_{\mathcal{L}}^2.\end{aligned}$$

Since both these terms tend to 0 for $n \to \infty$, $\widetilde{J}_1'(d_n, n) \to 0$ for $n \to \infty$, which finishes the proof of (iii).

(iv) By (2.6.23)

$$\theta_{d_{n+1-k},n,k} = \sum_{\ell=1}^{k} \beta_{d_{n+1-\ell},n,\ell}\theta_{d_{n+1-k},n-\ell,k-\ell}, \quad k = 1,\dots,n,$$

and we get $\theta_{d_n,n,1} = \beta_{d_n,n,1}$. Hence, for $n \to \infty$ and fixed $j \in \mathbb{N}$,

$$\|(\theta_{d_n,n,1} - \theta_{d_n,n-j,1})(x)\| = \|(\beta_{d_n,n,1} - \beta_{d_n,n-j,1})(x)\| \to 0. \tag{2.6.32}$$

For some fixed $j \in \mathbb{N}$ by a shift of (2.6.23), we obtain

$$\theta_{d_{n+1-k},n-j,k} = \sum_{\ell=1}^{k} \beta_{d_{n+1-\ell},n-j,\ell}\theta_{d_{n+1-\ell},n-j-\ell,k-\ell}. \tag{2.6.33}$$

With (2.6.33) and then the triangular equality after adding and subtracting $\beta_{d_{n+1-\ell},n,\ell}\theta_{d_{n+1-\ell},n-j-\ell,k-\ell}(x)$ for $\ell = 1,\dots,k$,

$$\begin{aligned}
&\Big\|(\theta_{d_{n+1-k},n,k} - \theta_{d_{n+1-k},n-j,k})(x)\Big\| \\
&= \Big\|\Big(\sum_{\ell=1}^{k} \beta_{d_{n+1-\ell},n,\ell}\theta_{d_{n+1-k},n-\ell,k-\ell} - \beta_{d_{n+1-\ell},n-j,\ell}\theta_{d_{n+1-\ell},n-j-\ell,k-\ell}\Big)(x)\Big\| \\
&\le \Big\|\sum_{\ell=1}^{k} \beta_{d_{n+1-\ell},n,\ell}(\theta_{d_{n+1-k},n-\ell,k-\ell} - \theta_{d_{n+1-\ell},n-j-\ell,k-\ell})(x)\Big\| \\
&\quad + \Big\|(\beta_{d_{n+1-\ell},n,\ell} - \beta_{d_{n+1-\ell},n-j,\ell})\theta_{d_{n+1-\ell},n-j-\ell,k-\ell}(x)\Big\|
\end{aligned}$$

By (iii) $\|(\beta_{d_{n+1-\ell},n,\ell} - \beta_{d_{n+1-\ell},n-j,\ell})(x)\| \to 0$ as $n \to \infty$. Furthermore, if for all $\ell = 1,\dots,i-1$, $\|(\theta_{d_{n+1-\ell},n,\ell} - \theta_{d_{n+1-\ell},n-j,\ell}(x)\| \to 0$, then $\|(\theta_{d_{n+1-i},n,i} - \theta_{d_{n+1-i},n-j,i}(x)\| \to 0$. The proof then follows by induction with the initial step given in (2.6.32). □

We are now ready to prove Theorem 2.5.3(ii).

Proof of Theorem 2.5.3(ii). Set $\pi_0 := -I_H$. By (2.2.5) and the definition of a linear process (2.2.3)

$$-\varepsilon_n = \sum_{i=0}^{\infty} \pi_i(X_{n-i}) = \sum_{i=0}^{\infty} \pi_i\Big(\sum_{j=0}^{\infty} \psi_j \varepsilon_{n-i-j}\Big), \quad n \in \mathbb{Z}.$$

Setting $k = i + j$, this can be rewritten as

$$-\varepsilon_n = \sum_{i=0}^{\infty} \pi_i\Big(\sum_{j=0}^{\infty} \psi_j \varepsilon_{n-i-j}\Big) = \sum_{k=0}^{\infty}\Big(\sum_{i+j=k} \pi_j\psi_i\Big)\varepsilon_{n-k} = \sum_{k=0}^{\infty}\sum_{j=0}^{k} \pi_j\psi_{k-j}\varepsilon_{n-k}.$$

Equating the coefficients we get $\sum_{j=0}^{k}\pi_j\psi_{k-j} = 0$ for $k > 0$. Since $-\pi_0 = I_H$, extracting the first term of the series, $\sum_{j=1}^{k}\pi_j\psi_{k-j} - I_H\psi_k = 0$, hence,

$$\sum_{j=1}^{k}\pi_j\psi_{k-j} = \psi_k.$$

Furthermore, by (2.6.23) we get for all $x \in H$,

$$\begin{aligned}
\Big\|\big(\theta_{d_{n+1-i},n,i} - \psi_i\big)(x)\Big\| &= \Big\|\Big(\sum_{j=1}^{i}\beta_{d_{n+1-j},n,j}\theta_{d_{n+1-i},n-j,i-j} - \sum_{j=1}^{i}\pi_j\psi_{i-j}\Big)(x)\Big\| \\
&= \Big\|\sum_{j=1}^{i}(\beta_{d_{n+1-j},n,j} - \pi_j)\theta_{d_{n+1-i},n-j,i-j}(x) \\
&\quad - \sum_{j=1}^{i}\pi_j(\psi_{i-j} - \theta_{d_{n+1-i},n-j,i-j})(x)\Big\| \\
&\le \Big\|\sum_{j=1}^{i}(\beta_{d_{n+1-j},n,j} - \pi_j)\theta_{d_{n+1-i},n-j,i-j}(x)\Big\| \\
&\quad + \Big\|\sum_{j=1}^{i}\pi_j(\psi_{i-j} - \theta_{d_{n+1-i},n,i-j})(x)\Big\| \\
&\quad + \Big\|\Big(\sum_{j=1}^{i}\pi_j(\theta_{d_{n+1-i},n,i-j} - \theta_{d_{n+1-i},n-j,i-j})\Big)(x)\Big\|,
\end{aligned}$$

where we have added and subtracted $\theta_{d_{n+1-i},n,i-j}$ and applied the triangular inequality for the last equality. Now, for $n \to \infty$, the last term tends to 0 by Proposition 2.6.4 (iv). The first term tends to 0 by Proposition 2.6.4 (ii). The second term tends to 0 by induction, where the initial step is clear, since $\psi_1 = -\pi_1$ and $\theta_{d_n,n,1} = \beta_{d_n,n,1}$. □

Chapter 3:

Estimating functional time series by moving average model fitting

3.1 Introduction

With the advent of complex data came the need for methods to address novel statistical challenges. Among the new methodologies, functional data analysis provides a particular set of tools for tackling questions related to observations conveniently viewed as entire curves rather than individual data points. The current state of the field may be reviewed in one of the comprehensive monographs written by Bosq [8], Ramsay and Silverman [47], Horváth and Kokoszka [27], and Hsing and Eubank [29]. Many of the applications discussed there point to an intrinsic time series nature of the underlying curves. This has led to an upsurge in contributions to the functional time series literature. The many recent works in this area include papers on time-domain methods such as Hörmann and Kokoszka [24], who introduced a framework to describe weakly stationary functional time series, and the work in Chapter 1 and 2, Aue et al. [5] and Hyndman and Shang [30], where functional prediction methodology is developed; as well as frequency domain methods such as Panaretos and Tavakoli [44], who utilized functional cumulants to justify their functional Fourier analysis, Hörmann et al. [25], who defined the concept of dynamic functional principal components, and Aue and van Delft [2], who designed stationarity tests based on functional periodogram properties.

This chapter is concerned with functional moving average (FMA) processes as a building block to estimate potentially more complicated functional time series.

Together with the functional autoregressive (FAR) processes, the FMA processes comprise one of the basic functional time series model classes. They are used, for example, as a building block in the L^p-m-approximability concept of Hörmann and Kokoszka [24], which is based on the idea that a sufficiently close approximation with truncated linear processes may adequately capture more complex dynamics, based on a causal infinite MA representation. It should be noted that, while there is a significant number of papers on the use of both FMA and FAR processes, the same is not the case for the more flexible functional autoregressive moving average (FARMA) processes. This is due to the technical difficulties that arise from transitioning from the multivariate to the functional level. One advantage that FMA processes enjoy over other members of the FARMA class is that their projections remain multivariate MA processes (of potentially lower order). This is one of the reasons that makes them attractive for further study.

Here interest is in estimating the dynamics of an invertible functional linear process through fitting FMA models. The operators in the FMA representation, a functional linear filter, are estimated using a functional Innovations Algorithm. This counterpart of the well-known univariate and multivariate Innovations Algorithms is introduced in Chapter 2 where its properties are analyzed on a population level. These results are here extended to the sample case and used as a first step in the estimation. The proposed procedure uses projections to a number of principal directions, estimated through functional principal components analysis (see, for example, Ramsay and Silverman [47]). To ensure appropriate large-sample properties of the proposed estimators, the dimensionality of the principle directions space is allowed to grow slowly with the sample size. In this framework, the consistency of the estimators of the functional linear filter is the main theoretical contribution. It is presented in Section 3.3.

The theoretical results are accompanied by selection procedures to guide the selection of the order of the approximating FMA process and the dimension of the subspace of principal directions. To choose the dimension of the subspace a sequential test procedure is proposed. Order selection based on AICC, Box–Ljung and FPE type criteria are suggested. Details of the proposed model selection procedures are given in Section 3.4. Their practical performance is highlighted in Section 3.5, where results of a simulation study are reported, and Section 3.6, where an application to real-world data on vehicle traffic data is discussed.

To summarize, this chapter is organized as follows. Section 3.2 briefly reviews basic notions of Hilbert-space valued random variables before introducing the setting and the main assumptions. The proposed estimation methodology for functional time series is detailed in Section 3.3. Section 3.4 discusses in some depth the practical selection of the dimension of the projection space and the order of the approximating FMA process. These suggestions are tested in a Monte Carlo simulation study and an application to traffic data in Sections 3.5 and 3.6, respectively. Section 3.7 concludes and proofs of the main results can be found in Section 3.8.

3.2 Setting

Functional data is often conducted in $H = L^2[0,1]$, the Hilbert-space of square-integrable functions, with canonical norm $\|x\| = \langle x,x\rangle^{1/2}$ induced by the inner product $\langle x,y\rangle = \int_0^1 x(s)y(s)ds$ for $x,y \in H$. For an introduction to Hilbert spaces from a functional analytic perspective, the reader is referred to Chapters 3.2 and 3.6 in Simon [50]. All random functions considered in this chapter are defined on a probability space $(\Omega,\mathcal{A},\mathcal{P})$ and are assumed to be $\mathcal{A}$-$\mathcal{B}_H$-measurable, where $\mathcal{B}_H$ denotes the Borel σ-algebra of subsets of H. Note that the space of square integrable random functions $L^2_H = L^2(\Omega,\mathcal{A},\mathcal{P})$ is a Hilbert space with inner product $\mathbb{E}[\langle X,Y\rangle] = \mathbb{E}[\int_0^1 X(s)Y(s)ds]$ for $X,Y \in L^2_H$. Similary, denote by $L^p_H = L^p(\Omega,\mathcal{A},\mathcal{P})$ the space of H-valued functions such that $\nu_p(X) = (\mathbb{E}[\|X\|^p])^{1/p} < \infty$. Let $\mathbb{Z}$, $\mathbb{N}$ and $\mathbb{N}_0$ denote the set of integers, positive integers and non-negative integers, respectively.

Interest in this chapter is in fitting techniques for functional time series $(X_j\colon j \in \mathbb{Z})$ taking values in L^2_H. To describe a wide variety of temporal dynamics, the framework is established for functional linear processes $(X_j\colon j \in \mathbb{Z})$ defined through the series expansion

$$X_j = \sum_{\ell=0}^{\infty} \psi_\ell \varepsilon_{j-\ell}, \qquad j \in \mathbb{Z}, \tag{3.2.1}$$

where $(\psi_\ell\colon \ell \in \mathbb{N}_0)$ is a sequence in $\mathcal{L}$, the space of bounded linear operators acting on H, equipped with the standard norm $\|A\|_{\mathcal{L}} = \sup_{\|x\|\leq 1}\|Ax\|$, and $(\varepsilon_j\colon j \in \mathbb{Z})$ is assumed to be an independent and identically distributed sequence in L_H^2. Additional summability conditions are imposed on the sequence of coefficient operators $(\psi_\ell : \ell \in \mathbb{N}_0)$ if it is necessary to control the rate of decay of the temporal dependence. Whenever the terminology "functional linear process" is used in this chapter it is understood to be in the sense of (3.2.1). Note that, as for univariate and multivariate time series models, every stationary causal functional autoregressive moving average (FARMA) process is a functional linear process (see Spangenberg [51], Theorem 2.3). Special cases include functional autoregressive processes of order p, FAR(p), which have been thoroughly investigated in the literature, and the *functional moving average process of order* q, FMA(q), which is given by the equation

$$X_j = \sum_{\ell=1}^{q} \theta_\ell \varepsilon_{j-\ell} + \varepsilon_j, \qquad j \in \mathbb{Z}, \tag{3.2.2}$$

with $\theta_1, \ldots, \theta_q \in \mathcal{L}$.

While the functional linear process in (3.2.1) is the prototypical causal time series, in the context of prediction, the concept of invertibility naturally enters; see Chapter 5.5 of Brockwell and Davis [13], and Nsiri and Roy [43]. For a functional time series $(X_j : j \in \mathbb{Z})$ to be *invertible*, it is required that

$$X_j = \sum_{\ell=1}^{\infty} \pi_\ell X_{j-\ell} + \varepsilon_j, \qquad j \in \mathbb{Z}, \tag{3.2.3}$$

for $(\pi_\ell : \ell \in \mathbb{N})$ in $\mathcal{L}$ such that $\sum_{\ell=1}^{\infty} \|\pi_\ell\|_{\mathcal{L}} < \infty$; see Merlevède [40]. A sufficient condition for invertibility of a functional linear process, which is assumed throughout, is given in Theorem 7.2 of Bosq [8].

The definition of a functional linear process in (3.2.1) provides a convenient framework for the formulation of large-sample results and their verification. In order to analyze time series characteristics in practice, however, most statistical methods require a more in-depth understanding of the underlying dependence structure. This is typically achieved through the use of autocovariances which determine the second-order structure. Observe first that any random variable in L_H^p with $p \geq 1$ possesses a unique *mean function* in H, which allows for a pointwise definition; see Bosq [8]. For what follows, it is assumed without loss of generality that $\mu = 0$, the zero function. If $X \in L_H^p$ with $p \geq 2$ such that $\mathbb{E}[X] = 0$, then the *covariance operator* of X exists and is given by

$$C_X(y) = \mathbb{E}[\langle X, y\rangle X], \qquad y \in H.$$

If $X, Y \in L^p_H$ with $p \geq 2$ such that $\mathbb{E}[X] = \mathbb{E}[Y] = 0$, then the *cross covariance operator* of X and Y exists and is given by

$$C_{X,Y}(y) = C^*_{Y,X}(y) = \mathbb{E}[\langle X, y\rangle Y], \qquad y \in H.$$

where $C^*_{Y,X}$ denotes the adjoint of $C_{Y,X}$, noting that the adjoint A^* of an operator A is defined by the equality $\langle Ax, y\rangle = \langle x, A^* y\rangle$ for $x, y \in H$. The operators C_X and $C_{Y,X}$ belong to $\mathcal{N}$, the class of *nuclear operators*, whose elements A have a representation $A = \sum_{j=1}^{\infty} \lambda_j \langle e_j, \cdot\rangle f_j$ with $\sum_{j=1}^{\infty} |\lambda_j| < \infty$ for two orthonormal bases (ONB) $(e_j : j \in \mathbb{N})$ and $(f_j : j \in \mathbb{N})$ of H. In that case $\|A\|_{\mathcal{N}} = \sum_{j=1}^{\infty} |\lambda_j| < \infty$; see Section 1.5 of Bosq [8]. Furthermore, C_X is self-adjoint ($C_X = C_X^*$) and non-negative definite with spectral representation

$$C_X(y) = \sum_{i=1}^{\infty} \lambda_i \langle y, \nu_i\rangle \nu_i, \qquad y \in H,$$

where $(\nu_i : i \in \mathbb{N})$ is an ONB of H and $(\lambda_i : i \in \mathbb{N})$ is a sequence of positive real numbers such that $\sum_{i=1}^{\infty} \lambda_i < \infty$. When considering spectral representations, it is standard to assume that the $(\lambda_i : i \in \mathbb{N})$ are ordered decreasingly and that there are no ties between consecutive λ_i.

For ease of notation, introduce the operator $x \otimes y(\cdot) = \langle x, \cdot\rangle y$ for $x, y \in H$. Then, $C_X = \mathbb{E}[X \otimes X]$ and $C_{X,Y} = \mathbb{E}[X \otimes Y]$. Moreover, for a stationary process $(X_j : j \in \mathbb{Z})$, the *lag-h covariance operator* can be written as

$$C_{X;h} = \mathbb{E}[X_0 \otimes X_h], \qquad h \in \mathbb{Z}. \tag{3.2.4}$$

The quantities in (3.2.4) are the basic building block in the functional Innovations Algorithm and the associated estimation strategy to be discussed in the next section.

3.3 Estimation methodology

3.3.1 Linear prediction in function spaces

Briefly recall the concept of linear prediction in Hilbert spaces as defined in Section 1.6 of Bosq [8]. Let $(X_j : j \in \mathbb{Z})$ be an invertible, functional linear process. Let $\bar{L}_{n,k}$

be the $\mathcal{L}$-closed subspace (LCS) generated by the stretch of functions $X_{n-k}, \dots, X_n$. LCS here is to be understood in the sense of Fortet [19], that is, $\bar{L}_{n,k}$ is the smallest subspace of H containing $X_{n-k}, \dots, X_n$, closed with respect to operators in $\mathcal{L}$. Then, the best linear predictor of X_{n+1} given $\{X_n, X_{n-1}, \dots, X_{n-k}\}$ at the population level is given by

$$\tilde{X}^f_{n+1,k} = P_{\bar{L}_{n,k}}(X_{n+1}), \tag{3.3.1}$$

where the superscript f in the predictor notation indicates the fully functional nature of the predictor and $P_{\bar{L}_{n,k}}$ denotes projection onto $\bar{L}_{n,k}$. Note that there are major differences to the multivariate prediction case. Due to the infinite dimensionality of function spaces, $\tilde{X}^f_{n+1,k}$ in (3.3.1) is not guaranteed to have a representation in terms of its past values and operators in $\mathcal{L}$, see for instance Proposition 2.2 in Bosq [10] and the discussion in Chapter 2.3. A typical remedy in FDA is to resort to projections into principal directions and then to let the dimension d of the projection subspace grow to infinity. At the subspace-level, multivariate methods may be applied to compute the predictors; for example the multivariate Innovations Algorithm; see Lewis and Reinsel [38] and Mitchell and Brockwell [42]. This, however, has to be done with care, especially if sample versions of the predictors in (3.3.1) are considered. Even at the population level, the rate at which d tends to infinity has to be calibrated scrupulously to ensure that the inversions of matrices occurring, for example, in the multivariate Innovations Algorithm are meaningful and well defined (see Theorem 2.5.3).

Therefore, the following alternative to the functional best linear predictor defined in (3.3.1) is proposed. Recall that $(\nu_j : j \in \mathbb{N})$ are the eigenfunctions of the covariance operator C_X. Let $\mathcal{V}_d = \overline{\mathrm{sp}}\{\nu_1, \dots, \nu_d\}$ be the subspace generated by the first d principal directions and let $P_{\mathcal{V}_d}$ be the projection operator projecting from H onto $\mathcal{V}_d$. Let furthermore $(d_i : i \in \mathbb{N})$ be an increasing sequence of positive integers and define

$$X_{d_i,j} = P_{\mathcal{V}_{d_i}} X_j, \qquad j \in \mathbb{Z},\ i \in \mathbb{N}. \tag{3.3.2}$$

Note that (3.3.2) allows for the added flexibility of projecting different X_j into different subspaces $\mathcal{V}_{d_i}$. Then, X_{n+1} can be projected into the LCS generated by

$X_{d_k,n}, X_{d_{k-1},n-1}, \ldots, X_{d_1,n-k}$, which is denoted by $\bar{F}_{n,k}$. Consequently, write

$$\tilde{X}_{n+1,k} = P_{\bar{F}_{n,k}}(X_{n+1}) \tag{3.3.3}$$

for the best linear predictor of X_{n+1} given $\bar{F}_{n,k}$. This predictor could be computed by regressing X_{n+1} onto $X_{d_k,n}, X_{d_{k-1},n-1}, \ldots, X_{d_1,n-k}$, but interest is here in the equivalent representation of $\tilde{X}_{n+1,k}$ in terms of one-step ahead prediction residuals given by

$$\tilde{X}_{n+1,k} = \sum_{i=1}^{k} \theta_{k,i}(X_{d_{k+1-i},n+1-i} - \tilde{X}_{n+1-i,k-i}), \tag{3.3.4}$$

where $\tilde{X}_{n-k,0} = 0$. On a population level, it is shown in Proposition 2.5.1 that the coefficients $\theta_{k,i}$ with $k, i \in \mathbb{N}$ can be computed with the following algorithm.

Algorithm 3.3.1 (Functional Innovations Algorithm) Let $(X_j : j \in \mathbb{Z})$ be a stationary functional linear process with covariance operator C_X possessing eigenpairs $(\lambda_i, \nu_i : i \in \mathbb{N})$ with $\lambda_i > 0$ for all $i \in \mathbb{N}$. The best linear predictor $\tilde{X}_{n+1,k}$ of X_{n+1} based on $\bar{F}_{n,k}$ defined in (3.3.4) can be computed by the recursions

$$\tilde{X}_{n-k,0} = 0 \qquad \text{and} \qquad V_1 = P_{\mathcal{V}_{d_1}} C_X P_{\mathcal{V}_{d_1}},$$

$$\tilde{X}_{n+1,k} = \sum_{i=1}^{k} \theta_{k,i}(X_{d_{k+1-i},n+1-i} - \tilde{X}_{n+1-i,k-i}),$$

$$\theta_{k,k-i} = \left(P_{\mathcal{V}_{d_{k+1}}} C_{X;k-i} P_{\mathcal{V}_{d_{i+1}}} - \sum_{j=0}^{i-1} \theta_{k,k-j} \; V_j \; \theta^*_{i,i-j} \right) V_i^{-1}, \qquad i = 1, \ldots, n-1, \tag{3.3.5}$$

$$V_k = C_{X_{d_{k+1}} - \tilde{X}_{n+1,k}} = C_{X_{d_{k+1}}} - \sum_{i=0}^{k-1} \theta_{k,k-i} V_i \theta^*_{k,k-i}. \tag{3.3.6}$$

Note that $\theta_{k,k-i}$ and V_i are operators in $\mathcal{L}$ for all $i = 1, \ldots, k$.

The first main goal is now to show how a finite sample version of this algorithm can be used to estimate the operators in (3.2.2), as these FMA processes will be used to approximate the more complex processes appearing in Definition 3.8.1. Note that Hörmann and Kokoszka [24] give assumptions under which $\sqrt{n}$-consistent estimators can be obtained for the lag-h autocovariance operator $C_{X;h}$, for $h \in \mathbb{Z}$. However, in (3.3.5), estimators are required for the more complicated quantities

$P_{\mathcal{V}_{d_{k+1}}} C_{X;k-i} P_{\mathcal{V}_{d_{i+1}}}$, for $k, i \in \mathbb{N}$. If, for $i \in \mathbb{N}$, the projection subspace $\mathcal{V}_{d_i}$ is known, consistent estimators of $P_{\mathcal{V}_{d_{k+1}}} C_{X;k-i} P_{\mathcal{V}_{d_{i+1}}}$ can be obtained by estimating $C_{X;k-i}$ and projecting the operator on the desired subspace. This case will be dealt with in Section 3.3.2. In practice, however, the subspaces $\mathcal{V}_{d_i}$, $i \in \mathbb{N}$, need to be estimated. This is a further difficulty that will be addressed separately in an additional step as part of Section 3.3.3.

Now, introduce additional notation. For $k \in \mathbb{N}$, denote by $(X_j(k) \colon j \in \mathbb{Z})$ the functional process taking values in H^k such that

$$X_j(k) = (X_j, X_{j-1}, \dots, X_{j-k+1})^\top,$$

where $^\top$ signifies transposition. Let

$$\Gamma_k = C_{X_n(k)} \qquad \text{and} \qquad \Gamma_{1,k} = C_{X_{n+1}, X_n(k)} = \mathbb{E}\big[X_{n+1} \otimes X_n(k)\big].$$

Based on a realization $X_1, \dots, X_n$ of $(X_j \colon j \in \mathbb{Z})$, estimators of the above operators are given by

$$\hat{\Gamma}_k = \frac{1}{n-k} \sum_{j=k}^{n-1} X_j(k) \otimes X_j(k) \qquad \text{and} \qquad \hat{\Gamma}_{1,k} = \frac{1}{n-k} \sum_{j=k}^{n-1} X_{j+1} \otimes X_j(k). \tag{3.3.7}$$

The following theorem establishes the $\sqrt{n}$-consistency of the estimator $\hat{\Gamma}_k$ of Γ_k defined in (3.3.7).

Theorem 3.3.1. *If $(X_j \colon j \in \mathbb{Z})$ is a functional linear process defined in* (3.2.1) *such that the coefficient operators $(\psi_\ell \colon \ell \in \mathbb{N}_0)$ satisfy the summability condition $\sum_{m=1}^{\infty} \sum_{\ell=m}^{\infty} \|\psi_\ell\|_{\mathcal{L}} < \infty$ and with independent, identically distributed innovations $(\varepsilon_j \colon j \in \mathbb{Z})$ such that $\mathbb{E}[\|\varepsilon_0\|^4] < \infty$, then*

$$(n-k)\, \mathbb{E}\big[\|\hat{\Gamma}_k - \Gamma_k\|_{\mathcal{N}}^2\big] \le k\, U_X,$$

where U_X is a constant that does not depend on n.

The proof of Theorem 3.3.1 is given in Section 3.8. There, an explicit expression for the constant U_X is derived that depends on moments of the underlying functional linear process and on the rate of decay of the temporal dependence implied by the summability condition on the coefficient operators $(\psi_\ell \colon \ell \in \mathbb{N}_0)$.

3.3.2 Known projection subspaces

In this section, conditions are established that ensure consistency of estimators of a functional linear process under the assumption that the projection subspaces $\mathcal{V}_{d_i}$ are known in advance. In this case as well as in the unknown subspace case, the following general strategy is pursued; see Mitchell and Brockwell [42]. Start by providing consistency results for the regression estimators of $\beta_{k,1},\dots,\beta_{k,k}$ in the linear model formulation

$$\tilde{X}_{n+1,k} = \beta_{k,1}X_{d_k,n} + \beta_{k,2}X_{d_{k-1},n-1} + \cdots + \beta_{k,k}X_{d_1,n-k+1}$$

of (3.3.3). To obtain the consistency of the estimators $\theta_{k,1},\dots,\theta_{k,k}$ exploit then that regression operators and Innovations Algorithm coefficient operators are, for $k \in \mathbb{N}$, linked through the recursions

$$\theta_{k,i} = \sum_{j=1}^{i} \beta_{k,j}\theta_{k-j,i-j}, \qquad i = 1,\dots,k. \tag{3.3.8}$$

Define furthermore $P_{(k)} = \operatorname{diag}(P_{\mathcal{V}_{d_k}},\dots,P_{\mathcal{V}_{d_1}})$, the operator from H^k to H^k whose ith diagonal entry is given by the projection operator onto $\mathcal{V}_{d_i}$. One verifies that $P_{(k)}X_n(k) = (X_{d_k,n}, X_{d_{k-1},n-1},\dots,X_{d_1,n-k})^\top$, $C_{P_{(k)}X(k)} = P_{(k)}\Gamma_k P_{(k)} = \Gamma_{k,d}$ and $C_{X,P_{(k)}X(k)} = P_{(k)}\Gamma_{1,k} = \Gamma_{1,k,d}$. With this notation, it can be shown that $B(k) = (\beta_{k,1},\dots,\beta_{k,k})$ satisfies the population Yule–Walker equations

$$B(k) = \Gamma_{1,k,d}\,\Gamma_{k,d}^{-1},$$

of which sample versions are needed. In the known subspace case, estimators of $\Gamma_{1,k,d}$ and $\Gamma_{k,d}$ are given by

$$\widehat{\Gamma}_{k,d} = P_{(k)}\widehat{\Gamma}_k P_{(k)} \qquad \text{and} \qquad \widehat{\Gamma}_{1,k,d} = \widehat{\Gamma}_{1,k}P_{(k)}, \tag{3.3.9}$$

where $\widehat{\Gamma}_k$ and $\widehat{\Gamma}_{1,k}$ are as in (3.3.7). With this notation, $B(k)$ is estimated by the sample Yule–Walker equations

$$\widehat{B}(k) = \widehat{\Gamma}_{1,k,d}\widehat{\Gamma}_{k,d}^{-1}. \tag{3.3.10}$$

Furthermore, the operators $\theta_{k,i}$ in (3.3.4) are estimated by $\widehat{\theta}_{k,i}$, resulting from Algorithm 3.3.1 applied to the estimated covariance operators with $\mathcal{V}_{d_i}$ known. In order to derive asymptotic properties of $\widehat{\beta}_{k,i}$ and $\widehat{\theta}_{k,i}$ as both k and n tend to infinity, the following assumptions are imposed. Let α_{d_k} denote the infimum of the eigenvalues of all spectral density operators of $(X_{d_k,j} \colon j \in \mathbb{Z})$.

Assumption 3.3.2. *As $n \to \infty$, let $k = k_n \to \infty$ and $d_k \to \infty$ such that*

(i) $(X_j \colon j \in \mathbb{Z})$ *is as in Theorem 3.3.1 and invertible.*

(ii) $k^{1/2}(n-k)^{-1/2}\alpha_{d_k}^{-1} \to 0$ *as* $n \to \infty$.

(iii) $k^{1/2}\alpha_{d_k}^{-1}\big(\sum_{\ell>k}\|\pi_\ell\|_{\mathcal{L}} + \sum_{\ell=1}^{k}\|\pi_\ell\|_{\mathcal{L}}\sum_{i>d_{k+1-\ell}}\lambda_i\big) \to 0$ *as* $n \to \infty$.

Invertibility imposed in part *(i)* of Assumption 3.3.2 is a standard requirement in the context of prediction and is also necessary for the univariate Innovations Algorithm to be consistent. Assumption *(ii)* describes the restrictions on the relationship between k, d_k and n. The corresponding multivariate assumption in Mitchell and Brockwell [42] is $k^3/n \to 0$ as $n \to \infty$. Assumption *(iii)* is already required in the population version of the functional Innovations Algorithm in the proof of Theorem 2.5.3. It ensures that the best linear predictor based on the last k observations converges to the conditional expectation for $k \to \infty$. The corresponding multivariate condition in Brockwell and Mitchell [42] is $k^{1/2}\sum_{\ell>k}\|\pi_\ell\| \to 0$ as $n \to \infty$, where $(\pi_\ell \colon \ell \in \mathbb{N})$ here denote the matrices in the invertible representation of a multivariate linear process.

The main result concerning the asymptotic behavior of the estimators $\widehat{\beta}_{k,i}$ and $\widehat{\theta}_{k,i}$ is given next.

Theorem 3.3.3. *Let $\mathcal{V}_{d_i}$ be known for all $i \in \mathbb{N}$ and let Assumption 3.3.2 be satisfied. Then, for all $x \in H$ and all $i \in \mathbb{N}$ as $n \to \infty$,*

(i) $\|(\widehat{\beta}_{k,i} - \pi_i)(x)\| \xrightarrow{p} 0$,

(ii) $\|(\widehat{\theta}_{k,i} - \psi_i)(x)\| \xrightarrow{p} 0$.

If the operators $(\psi_\ell \colon \ell \in \mathbb{N})$ and $(\pi_\ell \colon \ell \in \mathbb{N})$ in the respective causal and invertible representations are assumed Hilbert–Schmidt, then the convergence in (i) and (ii) is uniform.

The proof of Theorem 3.3.3 is given in Section 3.8. The theorem establishes the pointwise convergence of the estimators needed in order to get a sample proxy for the functional linear filter $(\pi_\ell : \ell \in \mathbb{N})$. This filter encodes the second-order dependence in the functional linear process and can therefore be used for estimating the underlying dynamics for the case of known projection subspaces.

3.3.3 Unknown projection subspaces

The goal of this section is to remove the assumption of known $\mathcal{V}_{d_i}$. Consequently, the standard estimators for the eigenfunctions $(\nu_i : i \in \mathbb{N})$ of the covariance operator C_X are used, obtained as the sample eigenfunctions $\widehat{\nu}_j$ of $\widehat{C}_X$. Therefore, for $i \in \mathbb{N}$, the estimators of $\mathcal{V}_{d_i}$ and $P_{\mathcal{V}_{d_i}}$ are

$$\widehat{\mathcal{V}}_{d_i} = \overline{\mathrm{sp}}\{\widehat{\nu}_1, \widehat{\nu}_2, \ldots, \widehat{\nu}_{d_i}\} \qquad \text{and} \qquad \widehat{P}_{\mathcal{V}_{d_i}} = P_{\widehat{\mathcal{V}}_{d_i}}. \tag{3.3.11}$$

For $i \in \mathbb{N}$, let $\widehat{\nu}_i' = c_i \widehat{\nu}_i$, where $c_i = \mathrm{sign}(\langle \widehat{\nu}_i, \nu_i \rangle)$. Then, Theorem 3.1 in Hörmann and Kokoszka [24] implies the consistency of $\widehat{\nu}_i'$ for $\widehat{\nu}_i$, with the quality of approximation depending on the spectral gaps of the eigenvalues $(\lambda_i : i \in \mathbb{N})$ of C_X. With this result in mind, define

$$\widehat{\widehat{\Gamma}}_{k,d} = \widehat{P}_{(k)} \widehat{\Gamma}_k \widehat{P}_{(k)} \qquad \text{and} \qquad \widehat{\widehat{\Gamma}}_{1,k,d} = \widehat{\Gamma}_{1,k} \widehat{P}_{(k)}. \tag{3.3.12}$$

Now, if the projection subspace $\mathcal{V}_{d_i}$ is not known, the operators appearing in (3.3.8) and can be estimated by solving the estimated Yule–Walker equations

$$\widehat{\widehat{B}}(k) = \widehat{\widehat{\Gamma}}_{1,k,d} \widehat{\widehat{\Gamma}}_{k,d}^{-1}. \tag{3.3.13}$$

The coefficient operators in Algorithm 3.3.1 obtained from estimated covariance operators and estimated projection space $\widehat{P}_{\mathcal{V}_{d_i}}$ are denoted by $\widehat{\widehat{\theta}}_{k,i}$. In order to derive results concerning their asymptotic behavior, an additional assumption concerning the decay of the spectral gaps of C_X is needed. Let $\delta_1 = \lambda_1 - \lambda_2$ and $\delta_j = \min\{\lambda_{j-1} - \lambda_j, \lambda_j - \lambda_{j+1}\}$ for $j \geq 2$.

Assumption 3.3.4. *As $n \to \infty$, $k = k_n \to \infty$ and $d_k \to \infty$ such that*

(iv) $k^{3/2} \alpha_{d_k}^{-2} n^{-1} (\sum_{\ell=1}^{d_k} \delta_\ell^{-2})^{1/2} \to 0.$

This type of assumption dealing with the spectral gaps is typically encountered when dealing with the estimation of eigenelements of functional linear processes (see, for example, Bosq [8], Theorem 8.7). We are now ready to derive the asymptotic result of the estimators in the general case that A_{d_i} is not known.

Theorem 3.3.5. *Let Assumptions 3.3.2 and 3.3.4 be satisfied. Then, for all $x \in H$ and $i \in \mathbb{N}$ as $n \to \infty$,*

(i) $\|(\widehat{\widehat{\beta}}_{k,i} - \pi_i)(x)\| \xrightarrow{p} 0,$

(ii) $\|(\widehat{\widehat{\theta}}_{k,i} - \psi_i)(x)\| \xrightarrow{p} 0.$

If the operators $(\psi_\ell : \ell \in \mathbb{N})$ and $(\pi_\ell : \ell \in \mathbb{N})$ are Hilbert–Schmidt, then the convergence is uniform.

The proof of Theorem 3.3.5 is given in Section 3.8. The theoretical results quantify the large-sample behavior of the estimates of the linear filter operators in the causal and invertible representations of the strictly stationary functional time series $(X_j : j \in \mathbb{Z})$. How to guide the application of the proposed method in finite samples is addressed in the next section.

3.4 Selection of principal directions and FMA order

Model selection is a difficult problem when working with functional time series. Contributions to the literature have been made in the context of functional autoregressive models by Kokoszka and Reimherr [36], who devised a sequential test to decide on the FAR order, and Aue et al. [5], who introduced an FPE-type criterion. To the best of our knowledge, there are no contributions in the context of model selection in functional moving average models. This section introduces several procedures. A method for the selection of the subspace dimension is introduced in Section 3.4.1, followed by a method for the FMA order selection in Section 3.4.2. A criterion for the simultaneous selection is in Section 3.4.3.

3.4.1 Selection of principal directions

The most well-known method for the selection of d in functional data analysis is based on total variance explained, TVE, where d is chosen such that the first d eigenfunctions of the covariance operator explain a predetermined amount P of the variability; see, for example, Horváth and Kokoszka [27]. In order to apply the TVE criterion in the functional time series context, one has to ensure that no essential parts of the dependence structure in the data are omitted after the projection into principal directions. This is achieved as follows. First choose an initial d^* with the TVE criterion such with a fraction P of variation in the data is explained. This should be done conservatively. Then apply the portmanteau test of Gabrys and Kokoszka [20] to check whether the non-projected part $(I_H - P_{\mathcal{V}_{d^*}})X_1, \dots, (I_H - P_{\mathcal{V}_{d^*}})X_n$ of the observed functions $X_1, \dots, X_n$ can be considered independent. Modifying their test to the current situation, yields the statistic

$$Q_n^{d^*} = n \sum_{h=1}^{\bar{h}} \sum_{\ell,\ell'=d^*+1}^{d^*+p} f_h(\ell,\ell') b_h(\ell,\ell'), \tag{3.4.1}$$

where $f_h(\ell,\ell')$ and $b_h(\ell,\ell')$ denote the (ℓ,ℓ')th entries of $C_{\mathbf{X}^*;0}^{-1} C_{\mathbf{X}^*;h}$ and $C_{\mathbf{X}^*;h} C_{\mathbf{X}^*;0}^{-1}$, respectively, and $(\mathbf{X}_j^* : j \in \mathbb{Z})$ is the p-dimensional vector process consisting of the $d+1$st to $d+p$th eigendirections of the covariance operator C_X. Following Gabrys and Kokoszka [20], it follows under the assumption of independence of the non-projected series that $Q_n^{d^*} \to \chi^2_{p^2\bar{h}}$ in distribution. If the assumption of independence is rejected, set $d^* = d^* + 1$. Repeat the test until the independence hypothesis cannot be rejected and choose $d = d^*$ to estimate the functional linear filters. This leads to the following algorithm.

Algorithm 3.4.1 (Test for independence) Perform the following steps.

(1) For given observed functional time series data $X_1, \dots, X_n$, estimate the eigenpairs $(\widehat{\lambda}_1, \widehat{\nu}_1), \dots, (\widehat{\lambda}_n, \widehat{\nu}_n)$ of the covariance operator C_X. Select d^* such that

$$\mathrm{TVE}(d^*) = \frac{\sum_{i=1}^{d^*} \widehat{\lambda}_i}{\sum_{i=1}^{n} \widehat{\lambda}_i} \geq P$$

for some prespecified $P \in (0,1)$.

(2) While $Q_n^{d^*} > q_{\chi^2_{p^2\bar{h}},\alpha}$, set $d^* = d^* + 1$.

(3) If $Q_n^{d^*} \leq q_{\chi^2_{p^2\bar{h}},\alpha}$ stop and apply Algorithm 3.3.1 with $d_i = d^*$, for all $i \leq k$.

Note that the Algorithm 3.4.1 does not specify the choices of P, p, H and α. Recommendations on their selection are given in Section 3.5. Multiple testing could potentially be an issue, but intensive simulation studies have shown that, since d^* is initialized with the TVE criterion, usually no more than one or two iterations and tests are required for practical purposes. Therefore the confidence level is not adjusted, even though it would be feasible to incorporate this additional step into the algorithm.

3.4.2 Selection of FMA order

For a fixed d, multivariate model selection procedures can be applied to choose q. In fact, it is shown in Theorem 2.4.7 that the projection of an FMA(q) process on a finite-dimensional space is a VMA(q^*) with $q^* \leq q$. Assuming that the finite-dimensional space is chosen such that no information on the dependence structure of the process is lost, $q = q^*$. Then, the FMA order q may be chosen by performing model selection on the d-dimensional vector model given by the first d principal directions of $(X_j : j \in \mathbb{Z})$. Methods for selecting the order of VMA models are described, for example, in Chapter 11.5 of Brockwell and Davis [13], and Chapter 3.2 of Tsai [53].

The latter book provides arguments for the identification of the VMA order via cross correlation matrices. This Ljung–Box (LB) method for testing the null hypothesis $H_0 : C_{\mathbf{X};\underline{h}} = C_{\mathbf{X};\underline{h}+1} = \cdots = C_{\mathbf{X};\overline{h}} = 0$ versus the alternative that $C_{\mathbf{X};h} \neq 0$ for a lag h between $\underline{h}$ and $\overline{h}$ is based on the statistic

$$Q_{\underline{h},\overline{h}} = n^2 \sum_{h=\underline{h}}^{\overline{h}} \frac{1}{n-h} \mathrm{tr}(\widehat{C}_{\mathbf{X};h}^{\top} \widehat{C}_{\mathbf{X};0}^{-1} \widehat{C}_{\mathbf{X};h} C_{\mathbf{X};0}^{-1}). \tag{3.4.2}$$

Under regularity conditions $Q_{\underline{h},\overline{h}}$ is asymptotically distributed as a $\chi^2_{d^2(\overline{h}-\underline{h}+1)}$ random variable if the multivariate procss $(\mathbf{X}_j : j \in \mathbb{Z})$ on the first d principal directions follows a VMA(q) model and $\underline{h} > q$. For practical implementation, one computes iteratively $Q_{1,\overline{h}}, Q_{2,\overline{h}}, \ldots$ and selects the order q as the largest $\underline{h}$ such that $Q_{\underline{h},\overline{h}}$ is significant, but $Q_{\underline{h}+h,\overline{h}}$ is insignificant for all $h > 0$.

Alternatively, the well-known AICC criterion could be utilized. Algorithm 3.3.1 allows for the computationally efficient maximization of the likelihood function through the use of its innovation form; see Chapter 11.5 of Brockwell and Davis [13]. The AICC criterion is then given by

$$\mathrm{AICC}(q) = -2 \ln L(\Theta_1, \ldots, \Theta_q, \Sigma) + \frac{2nd(qd^2+1)}{nd - qd^2 - 2}, \tag{3.4.3}$$

where $\Theta_1, \ldots, \Theta_q$ are the fitted VMA coefficient matrices and Σ its fitted covariance matrix. The minimizer of (3.4.3) is selected as order of the FMA process. Both methods are compared in Section 3.5.

3.4.3 Functional FPE criterion

In this section a criterion that allows to choose d and q simultaneously is introduced. A similar criterion was established in Aue et al. [5], based on a decomposition of the functional mean squared prediction error. Note that, due to the orthogonality of the eigenfunctions $(\nu_i : i \in \mathbb{N})$ and the fact that $\widehat{X}_{n+1,k}$ lives in $\mathcal{V}_d$,

$$\mathbb{E}\Big[\|X_{n+1} - \widehat{X}_{n+1,k}\|^2\Big] = \mathbb{E}\Big[\|P_{\mathcal{V}_d}(X_{n+1} - \widehat{X}_{n+1,k})\|^2\Big] + \mathbb{E}\Big[\|(I_H - P_{\mathcal{V}_d})X_{n+1}\|^2\Big]. \quad (3.4.4)$$

The second summand in (3.4.4) satisfies

$$\mathbb{E}[\|(I_H - P_{\mathcal{V}_d})X_{n+1}\|^2] = \mathbb{E}[\|\sum_{i>d}\langle X_{n+1}, \nu_i\rangle\nu_i\|^2] = \sum_{i>d}\lambda_i.$$

The first summand in (3.4.4) is, due to the isometric isomorphy between $\mathcal{V}_d$ and $\mathbb{R}^d$ equal to the mean squared prediction error of the vector model fit on the d dimensional principal subspace. It can be shown using the results of Lai and Lee [37] that it is of order $\mathrm{tr}(C_{\mathbf{Z}}) + qd\,\mathrm{tr}(C_{\mathbf{Z}})/n$, where $C_{\mathbf{Z}}$ denotes the covariance matrix of the innovations of the vector process. Using the matrix version $\mathbf{V}_n$ of the operator V_n given through Algorithm 3.3.1 as a consistent estimator for $C_{\mathbf{Z}}$, the functional FPE criterion

$$\mathrm{fFPE}(d,q) = \frac{n + q\,d}{n}\,\mathrm{tr}(\mathbf{V}_n) + \sum_{i>d}\hat{\lambda}_i \quad (3.4.5)$$

is obtained. It can be minimized over both d and q to select the dimension of the principal subspace and the order of the FMA process jointly. As is noted in Aue et al. [5], where a similar criterion is proposed for the selection of the order of an FAR(p) model, the fFPE method is fully data driven: no further selection of tuning parameters is required.

3.5 Simulation evidence

3.5.1 Simulation setting

In this section, results from Monte Carlo simulations are reported. The simulation setting was as follows. Using the first D Fourier basis functions $f_1, \ldots, f_D$, the D-dimensional subspace $G^D = \overline{\mathrm{sp}}\{f_1, \ldots, f_D\}$ of H was generated following the setup in Aue et al. [5], then the isometric isomorphy between $\mathbb{R}^D$ and G^D is utilized to represent elements in G^D by D-dimensional vectors and operators acting on G^D by $D \times D$ matrices. Therefore $n + q$ D-dimensional random vectors as innovations for an FMA(q) model and q $D \times D$ matrices as operators were generated. Two different settings were of interest: processes possessing covariance operators with slowly and quickly decaying eigenvalues. Those cases were represented by selecting two sets of standard deviations for the innovation process, namely

$$\sigma_{\text{slow}} = (i^{-1}\colon i = 1, \ldots, D) \qquad \text{and} \qquad \sigma_{\text{fast}} = (2^{-i}\colon i = 1, \ldots, D). \tag{3.5.1}$$

With this, innovations

$$\varepsilon_j = \sum_{i=1}^{D} c_{j,i} f_i, \qquad j = 1 - q, \ldots, n,$$

were simulated, where $c_{j,i}$ are independent normal random variables with mean 0 and standard deviation $\sigma_{\cdot,i}$, the $\cdot$ being replaced by either slow or fast, depending on the setting. The parameter operators $\tilde{\theta}_\ell$, for $\ell = 1, \ldots, q$, were chosen at random by generating $D \times D$ matrices, whose entries $\langle \tilde{\theta}_\ell f_i, f_{i'} \rangle$ were independent zero mean normal random variables with variance $\sigma_{\cdot,i}\sigma_{\cdot,i'}$. The matrices were then rescaled to have spectral norm 1. Combining the forgoing, the FMA(q) process

$$X_j = \sum_{\ell=1}^{q} \theta_\ell \varepsilon_{j-\ell} + \varepsilon_j, \qquad j = 1, \ldots, n \tag{3.5.2}$$

were simulated, where $\theta_\ell = \kappa_\ell \tilde{\theta}_\ell$ with κ_ℓ being chosen to ensure invertibility of the FMA process. In the following section, the performance of the proposed estimator is evaluated, and compared and contrasted to other methods available in the literature for the special case of FMA(1) processes, in a variety of situations.

3.5.2 Estimation of FMA(1) processes

In this section, the performance of the proposed method is compared to two approaches introduced in Turbillon et al. [54] for the special case of FMA(1) processes. These methods are based on the following idea. Denote by C_ε the covariance operator of $(\varepsilon_n : n \in \mathbb{Z})$. Observe that since $C_{X;1} = \theta_1 C_\varepsilon$ and $C_X = C_\varepsilon + \theta_1 C_\varepsilon \theta_1^*$, it follows that $\theta_1 C_X = \theta_1 C_\varepsilon + \theta_1^2 C_\varepsilon \theta_1^* = C_{X;1} + \theta_1^2 C_{X;1}^*$, and especially

$$\theta_1^2 C_{X;1}^* - \theta_1 C_X + C_{X;1} = 0. \tag{3.5.3}$$

The estimators in Turbillon et al. [54] are based on solving the quadratic equation in (3.5.3) for θ_1. The first of these only works under the restrictive assumption that θ_1 and C_ε commute. Then, solving (3.5.3) is equivalent to solving univariate equations generated by individually projecting (3.5.3) onto the eigenfunctions of C_X. The second approach is inspired by the Riesz–Nagy method. It relies on regarding (3.5.3) as a fixed-point equation and therefore establishing a fixed-point iteration. Since solutions may not exist in H, suitable projections have to be applied. Consistency of both estimators is established in Turbillon et al. [54].

To compare the performance of the methods, FMA(1) time series were simulated as described in Section 3.5.1. As measure of comparison the estimation error $\|\theta_1 - \widehat{\theta}_1\|_{\mathcal{L}}$ was used after computing $\widehat{\theta}_1$ with the three competing procedures. Rather than selecting the dimension of the subspace via Algorithm 3.4.1, the estimation error is computed for $d = 1, \ldots, 5$.

The results are summarized in Table 3.1, where estimation errors were averaged over 1000 repetitions for each specification, using sample sizes $n = 100, 500$ and 1,000. For all three sample sizes, the operator kernel estimated with the proposed algorithm is closest to the real kernel. As can be expected, the optimal dimension increases with the sample size, especially for the case where the eigenvalues decay slowly. The projection method does not perform well, which is also to be expected, because the condition of commuting θ_1 and C_ε is violated. One can see that the choice of d is crucial: especially for small sample sizes for the proposed method, the estimation error explodes for large d. In order to get an intuition for the shape of the estimators, the kernels of the estimators resulting from the different estimation methods, using $n = 500$ and $\kappa_1 = 0.8$, are plotted in Figure 3.1. It can again be

		$n = 100$			$n = 500$			$n = 1000$		
	d	Proj	Iter	Inn	Proj	Iter	Inn	Proj	Iter	Inn
σ_{fast}	1	0.539	0.530	0.514	0.527	0.521	0.513	0.518	0.513	0.508
	2	0.528	0.433	**0.355**	0.508	0.391	0.287	0.500	0.386	0.277
	3	0.533	0.534	0.448	0.512	0.467	**0.235**	0.503	0.460	**0.197**
	4	0.534	0.650	0.582	0.513	0.573	0.276	0.504	0.567	0.216
	5	0.534	0.736	0.646	0.513	0.673	0.311	0.504	0.662	0.239
σ_{slow}	1	0.610	0.602	0.588	0.579	0.574	0.566	0.575	0.573	0.569
	2	0.614	0.527	**0.513**	0.581	0.487	0.434	0.577	0.483	0.422
	3	0.618	0.552	0.610	0.583	0.504	**0.389**	0.578	0.500	0.362
	4	0.620	0.591	0.861	0.584	0.531	0.402	0.579	0.522	**0.344**
	5	0.620	0.630	1.277	0.584	0.556	0.448	0.579	0.548	0.358

Table 3.1: Estimation error $\|\theta_1 - \widehat{\theta}_1\|_{\mathcal{L}}$, with $\theta_1 = \kappa_1\tilde{\theta}_1$ and $\kappa_1 = 0.8$, with $\widehat{\theta}_1$ computed with the projection method (Proj) and the iterative method (Iter) of [54], and the proposed method based on the functional Innovations Algorithm (Inn). The smallest estimation error is highlighted in bold for each case.

seen that the projection method yields results that are significantly different from both the truth and the other two methods who produce estimated operator kernels, whose shapes look roughly similar to the truth.

3.5.3 Model selection

In this section, the performance of the different model selection methods introduced in Section 3.4 is demonstrated. To do so, FMA(1) processes with weights $\kappa_1 = 0.4$ and 0.8 were simulated as in the previous section. In addition, two different FMA(3) processes were simulated according to the setting described in Section 3.5.1, namely

- Model 1: $\kappa_1 = 0.8$, $\kappa_2 = 0.6$, and $\kappa_3 = 0.4$.
- Model 2: $\kappa_1 = 0$, $\kappa_2 = 0$, and $\kappa_3 = 0.8$.

For sample sizes $n = 100$, 500 and 1,000, 1,000 processes of both Model 1 and 2 were

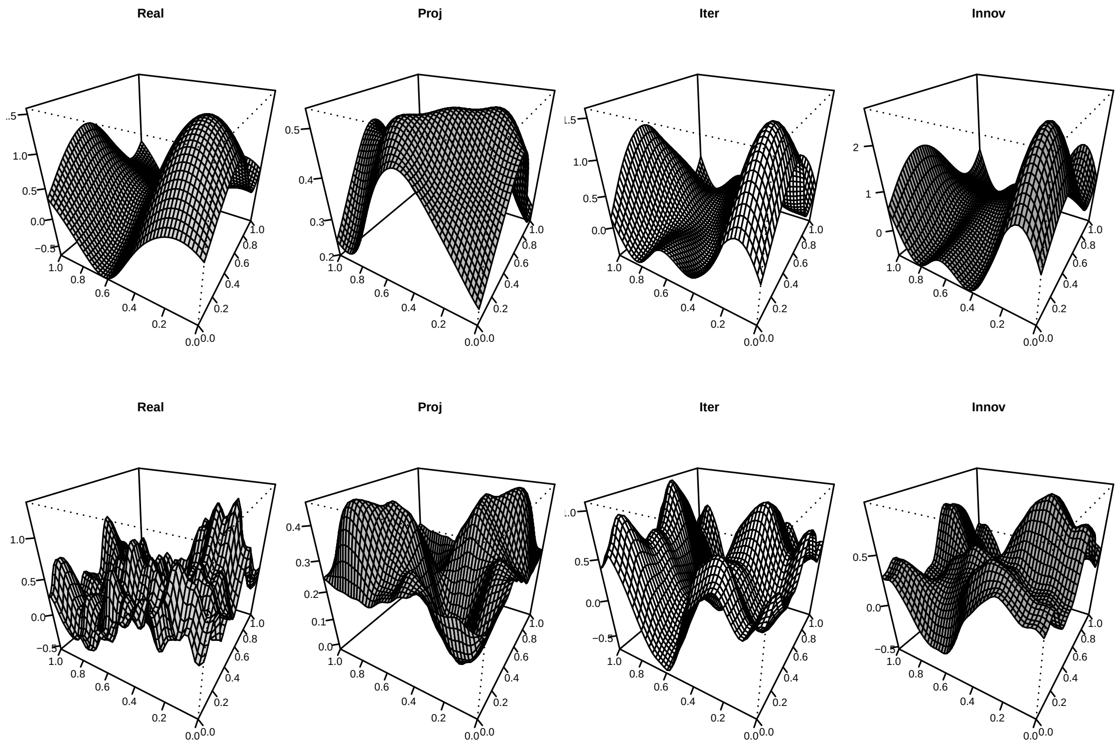

Figure 3.1: Estimated operator kernel of simulated FMA(1) process with $\kappa_1 = 0.8$, $d = 3$ and σ_{fast} (first row) and σ_{slow} (second row), using $n = 500$ sampled functions. Labeling of procedures is as in Table 3.1.

simulated using σ_{slow} and σ_{fast}. The estimation process was done as follows. First, the dimension d of the principal projection subspace was chosen using Algorithm 3.4.1 with TVE such that $P = 0.8$. With this selection of d, the LB and AICC criteria described in Section 3.4.2 were applied to choose q. Second, the fFPE criterion was used for a simultaneous selection of d and q. The results are summarized in Figures 3.2 and 3.3.

Figures 3.2 and 3.3 allow for a number of interesting observations. For both the FMA(1) and the FMA(3) example, the model order is estimated well. In all cases, especially for sample sizes larger than 100, all three selection methods (AIC, LB, FPEq) for the choice of q yield the correct model order (1 or 3). The Ljung–Box

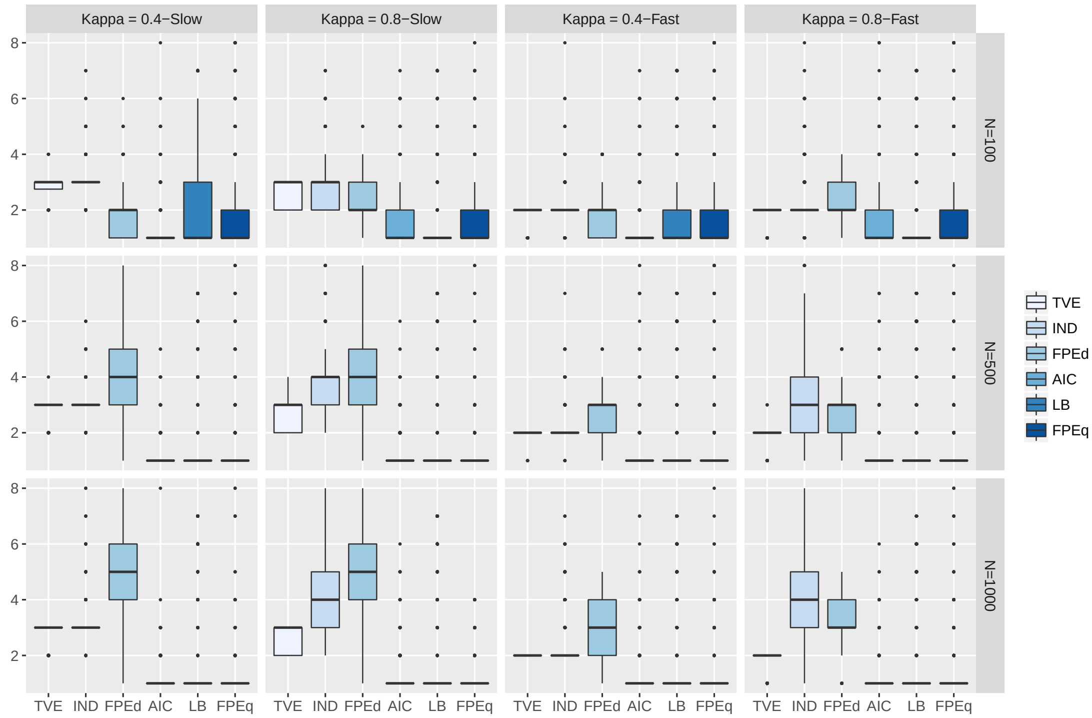

Figure 3.2: Model selection for different MA(1) processes. The left three plots in each small figure give the d chosen by total variation explained with $P = 0.8$ (TVE), Algorithm 3.4.1 (IND) and the functional FPE criterion (FPEd). The right three plots in each small figure give the selected order q by AICC, LB and fFPE.

(LB) method seems to have the most stable results. The methods for the choice of d are more heterogeneous. The TVE method yields the most stable results among different sample sizes. For σ_{fast}, it almost always selects $d = 2$ and for σ_{slow} the choice varies between $d = 2$ and $d = 3$. However, the TVE method seems to underestimate d. Often there appears to be dependence left in the data, as one can see from the selection of d by Algorithm 3.4.1. Especially in the FMA(3) case and Model 1, this algorithm yields some large choices for d of about 7 or 8. The choice of FPEd seems to increase with increasing sample size: this is to be expected as for increasing sample size the variance of the estimators decreases and the resulting predictors get more precise, even for high-dimensional models. This is valid especially for σ_{slow} where

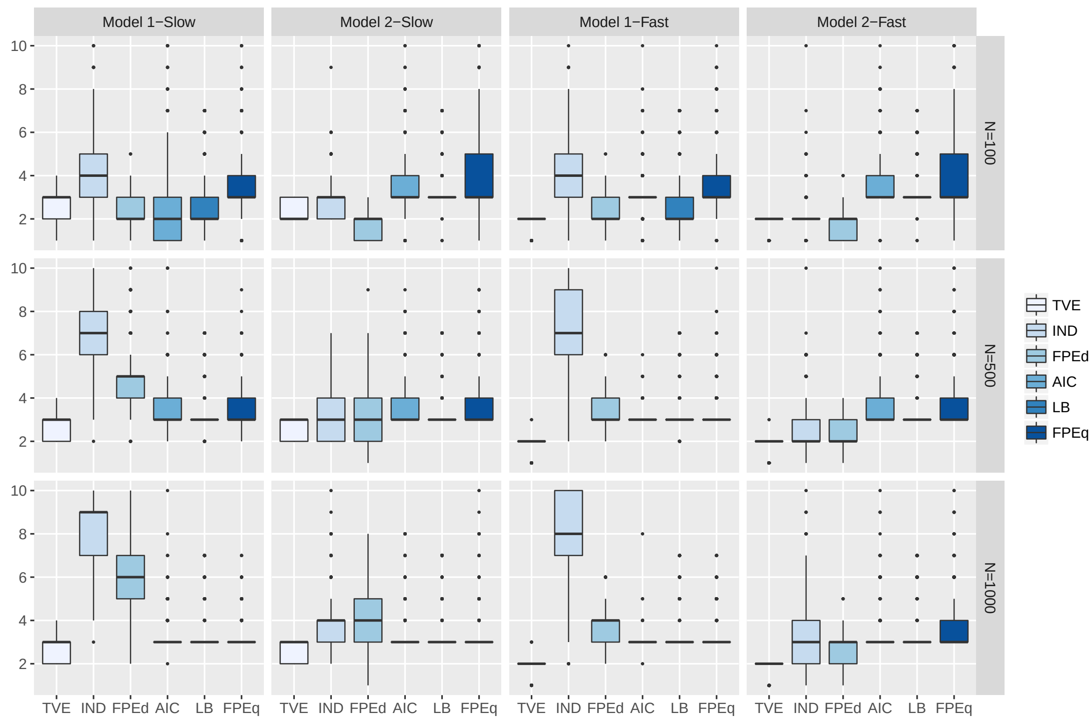

Figure 3.3: Model selection for different MA(3) processes. Labeling of procedures is as in Figure 3.2.

a larger d is needed to explain the dynamics of the functional process. A similar trade-off is occasionally observed for Algorithm 3.4.1.

3.6 Application to traffic data

In this section, the proposed estimation method is applied to vehicle traffic data provided by the Autobahndirektion Südbayern. The dataset consists of measurements at a fixed point on a highway (A92) in Southern Bavaria, Germany. Recorded is the average velocity per minute from 1/1/2014 00:00 to 30/06/2014 23:59 on three lanes. After taking care of missing values and outliers, the velocity per minute was averaged over the three lanes, weighted by the number of vehicles per lane. This leads to 1440 preprocessed and cleaned data points per day, which were transformed into functional data using the first 30 Fourier basis functions with the `R` package `fda`. The result is a functional time series $(X_j\colon j = 1,\dots,n = 119)$, which is deemed stationary and exhibits temporal dependence, as evidenced in Chapter 1.5.

The goal then is to approximate the temporal dynamics in this stationary functional time series with an FMA fit. Observe that the plots of the spectral norms $\|\widehat{C}_{\mathbf{X};h}\widehat{C}_{\mathbf{X};0}^{-1}\|_{\mathcal{L}}$ for $h = 0,\dots,5$ in Figure 3.4 display a pattern typical for MA models of low order. Here $\mathbf{X}$ stands for the multivariate auxiliary model of dimension d obtained from projection into the corresponding principal subspace. Consequently,

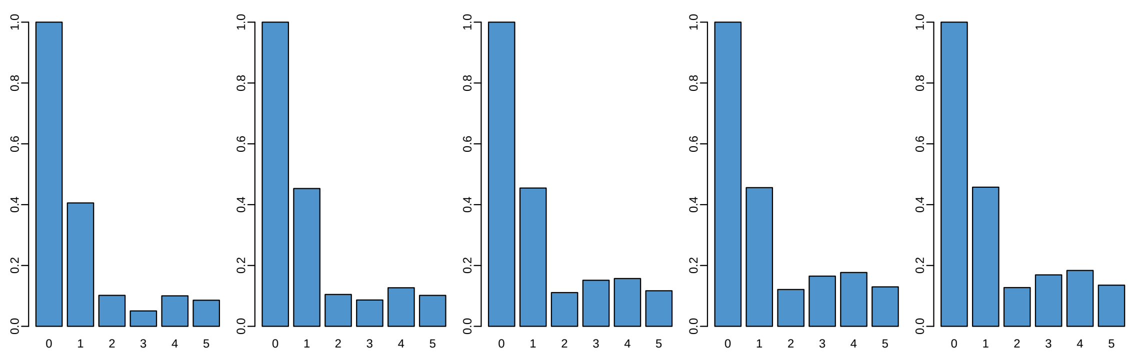

Figure 3.4: Spectral norm of estimated cross-correlation matrices for lags $h = 1,\dots,5$ of the vector model based on principal subspaces of dimension $d = 1$ to $d = 5$ (from left to right).

the methodology introduced in Section 3.3 and 3.4 was applied to the data. First, the covariance operator $C_{X;0}$ and its first 15 eigenelements $(\lambda_1,\nu_1),\dots,(\lambda_{15},\nu_{15})$ were estimated to construct the vector process $(\hat{\mathbf{X}}_j\colon j = 1,\dots,n)$, where $\hat{\mathbf{X}}_j = (\langle X_j,\hat{\nu}_1\rangle,\dots,\langle X_j,\hat{\nu}_{15}\rangle)^\top$. Then, the methods described in Sections 3.4 were applied to choose the appropriate dimension d and model order q.

The first four sample eigenfunctions explained 81% of the variability, hence the TVE criterion with $P = 0.8$ gave $d^* = 4$ to initialize Algorithm 3.4.1. The hypothesis of independence of the left-out score vector process $(\hat{\mathbf{X}}_j[4\colon 15]\colon j = 1,\dots,n)$ was rejected with p-value 0.03. Here $\mathbf{X}_j[i\colon i']$ is used as notation for the vector comprised of coordinates $i,\dots,i'$, with $i \le i'$, of the original 15-dimensional vector $\hat{\mathbf{X}}_j$. In the next step of Algorithm 3.4.1, d^* is increased to 5. A second independence test was run on $(\hat{\mathbf{X}}_j[5\!:\!15]\!:j=1,\dots,n)$ and did not result in a rejection; the corresponding p-value was 0.25.

This analysis led to using $d=5$ as dimension of the principal subspace to conduct model selection with the methods of Section 3.4.2. Since TVE indicated $d=4$, the selection procedures were applied also with this choice. In both cases, the AICC criterion in (3.4.3) and LB criterion in (3.4.2) opted for $q=1$, in accordance with the spectral norms observed in Figure 3.4. Simultaneously choosing d and q with the fFPE criterion of Section 3.4.3 yields $d=3$ and $q=1$.

After the model selection step, the operator of the chosen FMA(1) process was estimated using Algorithm 3.3.1. Similarly the methods introduced in Section 3.5.2 were applied. Figure 3.5 displays the kernels of the estimated integral operator for all methods, selecting for $d=3$ and $d=5$. The plots indicate that, on this particular data set, all three methods produce estimated operators that lead to kernels of roughly similar shape. The similarity is also reflected in the covariance of the estimated innovations. For $d=3$, the trace of the covariance matrix is 43.14, 45.4 and 44.41 for the Innovations Algorithm, iterative method and projective method, respectively. For $d=4$, the trace of the covariance of the estimated innovations is 48.19, 46.00 and 45.74 for the different methods in the same order.

3.7 Conclusions

This chapter introduces a complete methodology to estimate any stationary, causal and invertible functional time series. This is achieved by approximating the functional linear filters in the causal representation with functional moving average processes obtained from an application of the functional Innovations Algorithm. The consistency of the estimators is verified as the main theoretical contribution. The proof relies on the fact that d-dimensional projections of FMA(q) processes are isomorph to d dimensional VMA(q^*) models, with $q^* \leq q$. Introducing appropriate sequences of increasing subspaces of H, consistency can be established in the two cases of known and unknown principal projection subspaces. This line of reasoning follows multivariate techniques given in Lewis and Reinsel [38] and Mitchell and Brockwell [42].

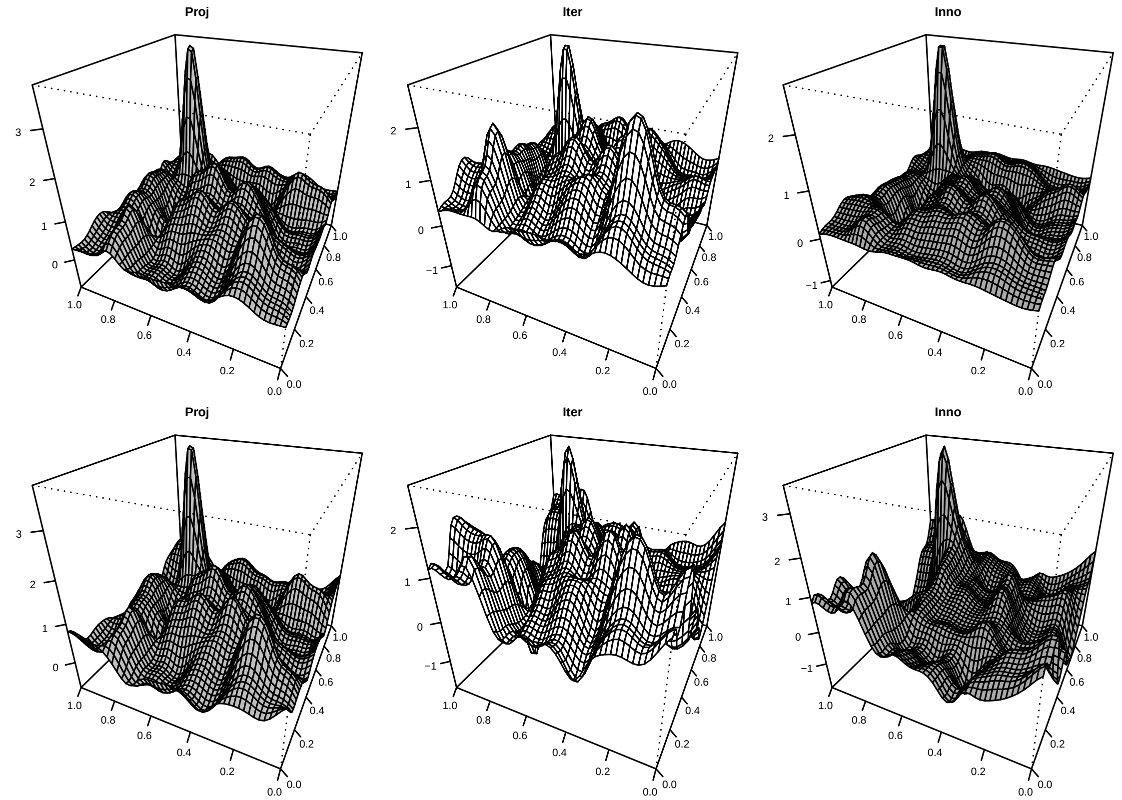

Figure 3.5: Estimated FMA(1) kernel with the three methods for $d = 3$ (first row) and $d = 4$ (second row)

The theoretical underpinnings are accompanied by model selection procedures facilitating the practical implementation of the proposed method. An independence test is introduced to select the dimension of the principal projection subspace, which can be used as a starting point for the suggested order selection procedures based on AICC and Ljung–Box criteria. Additionally, an fFPE criterion is established that jointly selects dimension d and order q. Illustrative results from a simulation study and the analysis of traffic velocity data show that the practical performance of the proposed method is satisfactory and at least competitive with other methods available in the literature for the case of FMA(1) processes.

Future research could focus on an extension of the methodology to FARMA processes in order to increase parsimony in the estimation. It should be noted, however, that this not a straightforward task as identifying the dynamics of the projection of an FARMA(p,q) to a finite-dimensional space is a non-resolved problem. In addition, the proposed methodology could be applied to offer an alternative route to estimate the spectral density operator, a principal object in the study of functional time series in the frequency domain; see Aue and van Delft [2], Hörmann et al. [25] and Panaretos and Tavakoli [44].

3.8 Proofs

The notion of L^p-m-approximability is utilized for the proofs. A version of this notion was used for multivariate time series in Aue et al. [4] and then translated to the functional domain by Hörmann and Kokoszka [24]. The definition is as follows.

Definition 3.8.1. Let $p \geq 1$. A sequence $(X_j : j \in \mathbb{Z})$ with values in L_H^p is called *L^p-m-approximable* if

$$X_j = f(\varepsilon_j, \varepsilon_{j-1}, \ldots), \qquad j \in \mathbb{Z},$$

can be represented as a functional Bernoulli shift with a sequence of independent, identically distributed random elements $(\varepsilon_j : j \in \mathbb{Z})$ taking values in the measurable space S, potentially different from H, and a measurable function $f: S^\infty \to H$ such that

$$\sum_{m=0}^{\infty} \left(\mathbb{E}[\|X_j - X_j^{(m)}\|^p]\right)^{1/p} < \infty,$$

where $X_j^{(m)} = f(\varepsilon_j, \ldots, \varepsilon_{j-m+1}, \varepsilon_{j-m}^{(j)}, \varepsilon_{j-m-1}^{(j)}, \ldots)$ with $(\varepsilon_j^{(i)} : j \in \mathbb{Z})$, $i \in \mathbb{N}_0$, being independent copies of $(\varepsilon_j : j \in \mathbb{Z})$.

Conditions can be established for most of the common linear and nonlinear functional time series models to be L^p-m-approximable. In particular, the functional linear processes $(X_j : j \in \mathbb{Z})$ defined in (3.2.1) are naturally included if the summability condition $\sum_{m=1}^{\infty} \sum_{\ell=m}^{\infty} \|\psi_\ell\|_{\mathcal{L}} < \infty$ is met (see Proposition 2.1 in Hörmann and Kokoszka [24]).

Proof of Theorem 3.3.1. Using that $(X_j : j \in \mathbb{Z})$ is L^4-m-approximable, write

$$\begin{aligned} X_j(k) &= (f(\varepsilon_j, \varepsilon_{j-1}, \ldots), \ldots, f(\varepsilon_{j-k+1}, \varepsilon_{j-k}, \ldots))^\top \\ &= g(\varepsilon_j, \varepsilon_{j-1}, \ldots), \end{aligned}$$

where $g\colon H^\infty \to H^k$ is defined accordingly. For $k, m \in \mathbb{N}$ and $j \in \mathbb{Z}$, define

$$\begin{aligned} X_j^{(m)}(k) &= \big(f(\varepsilon_j, \ldots, \varepsilon_{j-m+1}, \varepsilon_{j-m}^{(j)}, \varepsilon_{j-m-1}^{(j)}, \ldots), \ldots, \\ &\qquad f(\varepsilon_{j-k+1}, \ldots, \varepsilon_{j-m+1}, \varepsilon_{j-m}^{(j)}, \varepsilon_{j-m-1}^{(j)}, \ldots)\big)^\top \\ &= g(\varepsilon_j, \varepsilon_{j-1}, \ldots, \varepsilon_{j-m+1}, \varepsilon_{j-m}^{(j)}, \varepsilon_{j-m-1}^{(j)}, \ldots). \end{aligned}$$

Now, by definition of the norm in H^k,

$$\begin{aligned} \sum_{m=k}^{\infty} \big(\mathbb{E}\big[\|X_m(k) - X_m^{(m)}(k)\|^4\big]\big)^{1/4} &= \sum_{m=k}^{\infty} \left(\sum_{i=0}^{k-1} \mathbb{E}\big[\|X_{m-i} - X_{m-i}^{(m-i)}\|^4\big]\right)^{1/4} \\ &\le \sum_{m=k}^{\infty} \left(\sum_{i=0}^{k-1} \mathbb{E}\big[\|X_{m-i} - X_{m-i}^{(m-k)}\|^4\big]\right)^{1/4} \\ &= \sum_{m=k}^{\infty} \big(k\mathbb{E}\big[\|X_{m-k} - X_{m-k}^{(m-k)}\|^4\big]\big)^{1/4} \\ &= k^{1/4} \sum_{m=0}^{\infty} \big(\mathbb{E}\big[\|X_m - X_m^{(m)}\|^4\big]\big)^{1/4}, \end{aligned} \tag{3.8.1}$$

where the first inequality is implied by Assumption 3.3.2, since $\mathbb{E}[\|X_j - X_j^{(m-i)}\|^2] \le \mathbb{E}[\|X_j - X_j^{(m)}\|^2]$ for all $i \ge 0$, and the last inequality, since $\mathbb{E}[\|X_1 - X_1^{(m-k)}\|^2] = \mathbb{E}[\|X_j - X_j^{(m-k)}\|^2]$ by stationarity. But the right-hand side of (3.8.1) is finite because $(X_j\colon j \in \mathbb{Z})$ is L^4-m-approximable by assumption. This shows that $(X_j(k)\colon j \in \mathbb{Z})$ is also L^4-m approximable.

To prove the consistency of the estimator $\widehat{C}_{X(k)}$, note that the foregoing implies, by Theorem 3.1 in Hörmann and Kokoszka [24], that the bound

$$n\,\mathbb{E}\big[\|\widehat{C}_{X(k)} - C_{X(k)}\|_{\mathcal{N}}^2\big] \le U_{X(k)},$$

holds, where

$$U_{X(k)} = \mathbb{E}[\|X_1(k)\|^4] + 4\sqrt{2}(\mathbb{E}[\|X_1(k)\|^4])^{3/4} \sum_{m=0}^{\infty} (\mathbb{E}[\|X_m(k) - X_m^{(m)}(k)\|^4])^{1/4}$$

is a constant that does not depend on n. Since $\mathbb{E}[\|X_1(k)\|^4] = k\mathbb{E}[\|X_1\|^4]$, (3.8.1) yields that $U_{X(k)} = kU_X$, which is the assertion. □

Corollary 3.8.1 The operators $\widehat{\beta}_{k,i}$ from (3.3.10) and $\widehat{\theta}_{k,i}$ from (3.3.4) related through

$$\widehat{\theta}_{k,i} = \sum_{j=1}^{i} \widehat{\beta}_{k,j} \widehat{\theta}_{k-j,i-j}, \qquad i = 1, \ldots, k,\ k \in \mathbb{N}. \tag{3.8.2}$$

Proof. The proof is based on the finite-sample versions of the regression formulation of (3.3.1) and the innovations formulation given in (3.3.4). Details are omitted to conserve space. □

Proof of Theorem 3.3.3. *(i)* It is first shown that, for all $x \in H^k$,

$$\|(\widehat{B}(k) - \Pi(k))(x)\| \xrightarrow{p} 0 \qquad (n \to \infty),$$

where $\Pi(k) = (\pi_1, \ldots, \pi_k)^\top$ is the vector of the first k operators in the invertibility representation of the functional time series $(X_j : j \in \mathbb{Z})$. Define the process $(e_{j,k} : j \in \mathbb{Z})$ by letting

$$e_{j,k} = X_j - \sum_{\ell=1}^{k} \pi_\ell X_{j-\ell} \tag{3.8.3}$$

and let I_{H^k} be the identity operator on H^k. Note that

$$\begin{aligned}
\widehat{B}(k) - \Pi(k) &= \widehat{\Gamma}_{1,k,d}\widehat{\Gamma}_{k,d}^{-1} - \Pi(k)\widehat{\Gamma}_{k,d}\widehat{\Gamma}_{k,d}^{-1} + \Pi(k)(I_{H^k} - P_{(k)}) \\
&= \left(\Gamma_{1,k,d} - \Pi(k)\widehat{\Gamma}_{k,d}\right)\widehat{\Gamma}_{k,d}^{-1} + \Pi(k)(I_{H^k} - P_{(k)}).
\end{aligned}$$

Plugging in the estimators defined in (3.3.9) and subsequently using (3.8.3), it follows that

$$\begin{aligned}
\widehat{B}(k) - \Pi(k) &= \left(\frac{1}{n-k}\sum_{j=k}^{n-1}\left((P_{(k)}X_{j,k} \otimes X_{j+1}) - (P_{(k)}X_{j,k} \otimes \Pi(k)X_{j,k})\right)\right)\widehat{\Gamma}_{k,d}^{-1} \\
&\qquad + \Pi(k)(I_{H^k} - P_{(k)}) \\
&= \left(\frac{1}{n-k}\sum_{j=k}^{n-1}\left(P_{(k)}X_{j,k} \otimes (X_{j+1} - \Pi(k)X_{j,k})\right)\right)\widehat{\Gamma}_{k,d}^{-1} + \Pi(k)(I_{H^k} - P_{(k)}) \\
&= \left(\frac{1}{n-k}\sum_{j=k}^{n-1}\left(P_{(k)}X_{j,k} \otimes e_{j+1,k}\right)\right)\widehat{\Gamma}_{k,d}^{-1} + \Pi(k)(I_{H^k} - P_{(k)}).
\end{aligned}$$

Two applications of the triangle inequality imply that, for all $x \in H^k$,

$$
\begin{aligned}
&\|(\widehat{B}(k) - \Pi(k))(x)\| \\
&\le \left\| \left(\frac{1}{n-k} \sum_{j=k}^{n-1} \big(P_{(k)} X_j(k) \otimes e_{j+1,k} \big) \right) \widehat{\Gamma}_{k,d}^{-1}(x) \right\| + \|\Pi(k)(I_{H^k} - P_{(k)})(x)\| \\
&\le \left\| \left(\frac{1}{n-k} \sum_{j=k}^{n-1} \big(P_{(k)} X_j(k) \otimes (e_{j+1,k} - \varepsilon_{j+1}) \big) \right) \widehat{\Gamma}_{k,d}^{-1} \right\|_{\mathcal{L}} \\
&\qquad + \left\| \left(\frac{1}{n-k} \sum_{j=k}^{n-1} \big(P_{(k)} X_j(k) \otimes \varepsilon_{j+1} \big) \right) \widehat{\Gamma}_{k,d}^{-1} \right\|_{\mathcal{L}} + \|\Pi(k)(I_{H^k} - P_{(k)})(x)\| \\
&\le \big(\|U_{1n}\|_{\mathcal{L}} + \|U_{2n}\|_{\mathcal{L}} \big) \|\widehat{\Gamma}_{k,d}^{-1}\|_{\mathcal{L}} + \|\Pi(k)(I_{H^k} - P_{(k)})(x)\|,
\end{aligned}
\tag{3.8.4}
$$

where U_{1n} and U_{2n} have the obvious definitions. Arguments similar to those used in Proposition 2.6.4 yield that the second term on the right-hand side of (3.8.4) can be made arbitrarily small by increasing k. To be more precise, for $\delta > 0$, there is $k_\delta \in \mathbb{N}$ such that

$$
\|\Pi(k)(I_{H^k} - P_{(k)})(x)\| < \delta \tag{3.8.5}
$$

for all $k \ge k_\delta$ and all $x \in H^k$.

To estimate the first term on the right-hand side of (3.8.4), focus first on $\|\widehat{\Gamma}_{k,d}^{-1}\|_{\mathcal{L}}$. Using the triangular inequality, $\|\widehat{\Gamma}_{k,d}^{-1}\|_{\mathcal{L}} \le \|\widehat{\Gamma}_{k,d}^{-1} - \Gamma_{k,d}^{-1}\|_{\mathcal{L}} + \|\Gamma_{k,d}^{-1}\|_{\mathcal{L}}$. Theorem 1.2 in Mitchell [41] and Lemma 2.6.1 give the bound

$$
\|\Gamma_{k,d}^{-1}\|_{\mathcal{L}} \le \alpha_{d_k}^{-1}, \tag{3.8.6}
$$

where α_{d_k} is the infimum of the eigenvalues of all spectral density operators of $(X_{d_k,j} \colon j \in \mathbb{Z})$. Furthermore, using the triangle inequality and then again Lemma 2.6.1, we get

$$
\begin{aligned}
\|\widehat{\Gamma}_{k,d}^{-1} - \Gamma_{k,d}^{-1}\|_{\mathcal{L}} &= \|\widehat{\Gamma}_{k,d}^{-1}(\widehat{\Gamma}_{d,k} - \Gamma_{d,k})\Gamma_{k,d}^{-1}\|_{\mathcal{L}} \\
&\le \big(\|\widehat{\Gamma}_{k,d}^{-1} - \Gamma_{k,d}^{-1}\|_{\mathcal{L}} + \|\Gamma_{k,d}^{-1}\|_{\mathcal{L}} \big) \|\widehat{\Gamma}_{d,k} - \Gamma_{d,k}\|_{\mathcal{L}} \alpha_{d_k}^{-1}.
\end{aligned}
\tag{3.8.7}
$$

Hence, following arguments in the proof of Theorem 1 in Lewis and Reinsel [38],

$$
0 \le \frac{\|\widehat{\Gamma}_{k,d}^{-1} - \Gamma_{k,d}^{-1}\|_{\mathcal{L}}}{\alpha_{d_k}^{-1}(\|\widehat{\Gamma}_{k,d}^{-1} - \Gamma_{k,d}^{-1}\|_{\mathcal{L}} + \alpha_{d_k}^{-1})} \le \|\widehat{\Gamma}_{d,k} - \Gamma_{d,k}\|_{\mathcal{L}},
$$

by (3.8.7). This yields

$$\|\widehat{\Gamma}_{d,k}^{-1} - \Gamma_{d,k}^{-1}\|_{\mathcal{L}} \leq \frac{\|\widehat{\Gamma}_{d,k} - \Gamma_{d,k}\|_{\mathcal{L}}\alpha_{d_k}^{-2}}{1 - \|\widehat{\Gamma}_{d,k} - \Gamma_{d,k}\|_{\mathcal{L}}\alpha_{d_k}^{-1}}. \tag{3.8.8}$$

Note that, since $P_{(k)}P_k = P_{(k)}$, $\|\Gamma_{k,d}\|_{\mathcal{L}} = \|P_{(k)}P_k\Gamma_k P_k P_{(k)}\|_{\mathcal{L}} \leq \|P_k\Gamma_k P_k\|_{\mathcal{L}}$. Also, by Theorem 3.3.1, for some positive finite constant M_1, $\mathbb{E}[\|P_k\widehat{\Gamma}_k P_k - P_k\Gamma_k P_k\|^2] \leq M_1 k/(n-k)$. Therefore,

$$\|\widehat{\Gamma}_{d,k} - \Gamma_{d,k}\| = O_p\left(\sqrt{\frac{k}{n-k}}\right). \tag{3.8.9}$$

Hence, the second part of Assumption 3.3.2 and (3.8.8) lead first to $\|\widehat{\Gamma}_{d,k}^{-1} - \Gamma_{d,k}^{-1}\|_{\mathcal{L}} \xrightarrow{p} 0$ and, consequently, combining the above arguments,

$$\|\widehat{\Gamma}_{k,d}^{-1}\|_{\mathcal{L}} = O_p(\alpha_{d_k}^{-1}). \tag{3.8.10}$$

Next consider U_{1n} in (3.8.4). With the triangular and Cauchy–Schwarz inequalities, calculate

$$\begin{aligned}
\mathbb{E}[\|U_{1n}\|] &= \mathbb{E}\left[\left\|\frac{1}{n-k}\sum_{j=k}^{n-1} P_{(k)}X_j(k) \otimes (e_{j+1,k} - \varepsilon_{j+1})\right\|_{\mathcal{L}}\right] \\
&\leq \frac{1}{n-k}\sum_{j=k}^{n-1} \mathbb{E}\left[\left\|P_{(k)}X_j(k) \otimes (e_{j+1,k} - \varepsilon_{j+1})\right\|_{\mathcal{L}}\right] \\
&\leq \frac{1}{n-k}\sum_{j=k}^{n-1} \big(\mathbb{E}[\|P_{(k)}X_j(k)\|^2]\big)^{1/2}\big(\mathbb{E}[\|e_{j+1,k} - \varepsilon_{j+1}\|^2]\big)^{1/2}.
\end{aligned}$$

The stationarity of $(X_j : j \in \mathbb{Z})$ and the fact that $X_j \in L_H^2$ imply that, for a positive finite constant M_2,

$$\begin{aligned}
\mathbb{E}[\|U_{1n}\|_{\mathcal{L}}] &\leq \big(\mathbb{E}[\|P_{(k)}X_j(k)\|^2]\big)^{1/2}\big(\mathbb{E}[\|e_{j+1,k} - \varepsilon_{j+1}\|^2]\big)^{1/2} \\
&\leq \sqrt{k}\big(\mathbb{E}[\|P_{\mathcal{V}_{d_k}}X_0\|^2]\big)^{1/2}\left(\mathbb{E}\left[\left\|\sum_{\ell>k}\pi_\ell X_{1-\ell} + \sum_{\ell=1}^{k}\pi_\ell(I_H - P_{\mathcal{V}_{d_{k+1-\ell}}})X_{1-\ell}\right\|^2\right]\right)^{1/2} \\
&\leq \sqrt{k}\left(2\mathbb{E}\left[\|\sum_{\ell>k}\pi_\ell X_{j+1-\ell}\|^2\right] + 2\mathbb{E}\left[\left\|\sum_{\ell=1}^{k}\pi_\ell(I_H - P_{\mathcal{V}_{d_{k+1-\ell}}})X_{1-\ell}\right\|^2\right]\right)^{1/2} \\
&= M_2\sqrt{k(J_1 + J_2)} \\
&\leq M_2\sqrt{k}(\sqrt{J_1} + \sqrt{J_2}),
\end{aligned} \tag{3.8.11}$$

where J_1 and J_2 have the obvious definition. Since for $X \in L^2_H$, $\mathbb{E}[\|X\|^2] = \|C_X\|_{\mathcal{N}}$, the term J_1 can be bounded as follows. Observe that

$$\begin{aligned} J_1 &= \left\| \mathbb{E}\left[\sum_{\ell>k} \pi_\ell X_{1-\ell} \otimes \sum_{\ell'>k} \pi_{\ell'} X_{1-\ell'} \right] \right\|_{\mathcal{N}} \\ &= \left\| \sum_{\ell,\ell'>k} \pi_\ell C_{X;\ell-\ell'} \pi^*_{\ell'} \right\|_{\mathcal{N}} \\ &\le \sum_{\ell,\ell'>k} \|\pi_\ell\|_{\mathcal{L}} \|\pi_{\ell'}\|_{\mathcal{L}} \|C_{X;\ell-\ell'}\|_{\mathcal{N}}. \end{aligned}$$

Now $C_{X;\ell-\ell'} \in \mathcal{N}$ for all $\ell, \ell' \in \mathbb{Z}$, hence $\|C_{X;\ell-\ell'}\|_{\mathcal{N}} \le M_3$ and $J_1 \le M_3(\sum_{\ell>k} \|\pi_\ell\|_{\mathcal{L}})^2$. Concerning J_2, note first that, since $\mathbb{E}[\|X\|^2] = \|C_X\|_{\mathcal{N}}$,

$$J_2 = \left\| \mathbb{E}\left[\sum_{\ell=1}^{k} \pi_\ell (I_H - P_{\mathcal{V}_{d_{k+1-\ell}}}) X_{1-\ell} \otimes \sum_{\ell'=1}^{n} \pi_{\ell'} (I_H - P_{\mathcal{V}_{d_{k+1-\ell'}}}) X_{1-\ell'} \right] \right\|_{\mathcal{N}}.$$

Using the triangle inequality together with properties of the nuclear operator norm and the definition of $C_{X;h}$ in display (3.2.4) leads to

$$\begin{aligned} J_2 &\le \sum_{\ell,\ell'=1}^{k} \|\pi_\ell\|_{\mathcal{L}} \|\pi'_\ell\|_{\mathcal{L}} \left\| \mathbb{E}\left[(I_H - P_{\mathcal{V}_{d_{k+1-\ell}}}) X_{1-\ell} \otimes (I_H - P_{\mathcal{V}_{d_{k+1-\ell'}}}) X_{1-\ell'} \right] \right\|_{\mathcal{N}} \\ &= \sum_{\ell,\ell'=1}^{k} \|\pi_\ell\|_{\mathcal{L}} \|\pi'_\ell\|_{\mathcal{L}} \left\| (I_H - P_{\mathcal{V}_{d_{k+1-\ell}}}) C_{X;\ell-\ell'} (I_H - P_{\mathcal{V}_{d_{k+1-\ell'}}}) \right\|_{\mathcal{N}} \\ &= \sum_{\ell,\ell'=1}^{k} \|\pi_\ell\|_{\mathcal{L}} \|\pi'_\ell\|_{\mathcal{L}} K(\ell,\ell'). \end{aligned} \tag{3.8.12}$$

By the definition of $\mathcal{V}_d$ in (3.3.2) and since $(I_H - P_{\mathcal{V}_{d_i}}) = \sum_{r>d_i} \nu_r \otimes \nu_r$, it follows that

$$\begin{aligned} K(\ell,\ell') &= \left\| \sum_{s>d_{k+1-\ell'}} \sum_{r>d_{k+1-\ell}} \langle C_{X;\ell-\ell'}(\nu_r), \nu_s \rangle \nu_r \otimes \nu_s \right\|_{\mathcal{N}} \\ &\le \left\| \sum_{s>d_{k+1-\ell'}} \sum_{r>d_{k+1-\ell}} \sqrt{\lambda_r \lambda_s} \nu_r \otimes \nu_s \right\|_{\mathcal{N}} \\ &= \sum_{i=1}^{\infty} \left\langle \sum_{s>d_{k+1-\ell'}} \sum_{r>d_{k+1-\ell}} \sqrt{\lambda_r \lambda_s} \nu_r \otimes \nu_s(\nu_i), \nu_i \right\rangle \\ &\le \sum_{i>d_{k+1-\ell}} \lambda_i, \end{aligned} \tag{3.8.13}$$

where Lemma 2.6.2 was applied to give $\langle C_{X;\ell-\ell'}\nu_r, \nu_s\rangle \le \sqrt{\lambda_r\lambda_s}$. Plugging (3.8.13) into (3.8.12), and recalling that $\sum_{\ell=1}^{\infty} \|\pi_\ell\|_{\mathcal{L}} = M_4 < \infty$, gives that

$$J_2 \le M_4 \sum_{\ell=1}^{k} \|\pi_\ell\|_{\mathcal{L}} \sum_{i>d_{k+1-\ell}} \lambda_i. \tag{3.8.14}$$

Inserting the bounds for J_1 and J_2 into (3.8.11), for some $M < \infty$,

$$\begin{aligned}\mathbb{E}[\|U_{1n}\|] &\le \sqrt{k} M_2 (M_3\sqrt{J_1} + \sqrt{J_2}) \\ &\le \sqrt{k} M_2 \Bigg(M_3 \sum_{\ell>k} \|\pi_\ell\|_{\mathcal{L}} + M_4 \sum_{\ell=1}^{k} \|\pi_\ell\|_{\mathcal{L}} \sum_{i>d_{k+1-\ell}} \lambda_i \Bigg) \\ &\le \sqrt{k} M \Bigg(\sum_{\ell>k} \|\pi_\ell\|_{\mathcal{L}} + \Bigg(\sum_{\ell=1}^{k} \|\pi_\ell\|_{\mathcal{L}} \sum_{i>d_{k+1-\ell}} \lambda_i \Bigg)\Bigg). \end{aligned} \tag{3.8.15}$$

Concerning U_{2n} in (3.8.4), use the linearity of the scalar product, the independence of the innovations $(\varepsilon_j : j \in \mathbb{Z})$ and the stationarity of the functional time series $(X_j : j \in \mathbb{Z})$ to calculate

$$\begin{aligned} E[\|U_{2n}\|^2] &\le \left(\frac{1}{n-k}\right)^2 \sum_{j=k}^{n-1} \mathbb{E}\big[\|P_{(k)}X_j(k)\|^2\big]\mathbb{E}\big[\|\varepsilon_{j+1}\|^2\big] \\ &\le \frac{1}{n-k}\mathbb{E}\big[\|P_{(k)}X_0(k)\|^2\big]\mathbb{E}\big[\|\varepsilon_0\|^2\big] \\ &\le \frac{k}{n-k}\mathbb{E}\big[\|X_0\|^2\big]\mathbb{E}\big[\|\varepsilon_0\|^2\big]. \end{aligned}$$

Since both $(X_j : j \in \mathbb{Z})$ and $(\varepsilon_j : j \in \mathbb{Z})$ are in L_H^2, (3.8.10) implies that

$$\|U_{2n}\|_{\mathcal{L}} \|\widehat{\Gamma}_{k,d}^{-1}\|_{\mathcal{L}} = O_p\left(\frac{1}{\alpha_{d_k}} \sqrt{\frac{k}{n-k}} \right).$$

Furthermore, (3.8.10) and (3.8.15) show that

$$\|U_{1n}\|_{\mathcal{L}} \|\widehat{\Gamma}_{k,d}^{-1}\|_{\mathcal{L}} = O_p\Bigg(\frac{\sqrt{k}}{\alpha_{d_k}} \Bigg(\sum_{\ell>k} \|\pi_\ell\|_{\mathcal{L}} + \sum_{\ell=1}^{k} \|\pi_\ell\|_{\mathcal{L}} \sum_{i>d_{k+1-\ell}} \lambda_i \Bigg)\Bigg).$$

Thus Assumption 3.3.2, (3.8.4) and (3.8.5) assert that, for all $x \in H^k$, $\|\widehat{B}_k - \Pi(k)(x)\| \xrightarrow{p} 0$, which proves the first statement of the theorem.

(ii) First note that, for all $x \in H^k$, $\|(\widehat{\beta}_{k,i} - \beta_{k,i})(x)\| \le \|(\widehat{\beta}_{k,i} - \pi_i)(x)\| + \|(\pi_i - \beta_{k,i})(x)\| \xrightarrow{p} 0$ as $n \to \infty$. Now $\theta_{k,1} = \beta_{k,1}$ and by Corollary 3.8.1 $\widehat{\theta}_{k,1} = \widehat{\beta}_{k,1}$. Since

furthermore $\sum_{j=1}^{k}\pi_j\psi_{k-j}=\psi_k$ (see, for instance, the proof of Theorem 2.5.3), $\psi_1=\pi_1$. Therefore,

$$\|(\widehat{\theta}_{k,1}-\psi_1)(x)\|=\|(\widehat{\beta}_{k,1}-\pi_1)(x)\|\overset{p}{\to}0$$

as $n\to\infty$. This proves the statement for $i=1$. Proceed by assuming the statement of the theorem is true for $i=1,\dots,N\in\mathbb{N}$, and then use induction on N. Indeed, for $i=N+1$, the triangle inequality yields, for all $x\in H$,

$$\begin{aligned}\|(\widehat{\theta}_{k,N+1}-\psi_{N+1})(x)\| &= \left\|\left(\sum_{j=1}^{N+1}\widehat{\beta}_{k,j}\widehat{\theta}_{k-j,N+1-j}-\pi_j\psi_{N+1-j}\right)(x)\right\| \\ &\le \sum_{j=1}^{N+1}\|(\widehat{\beta}_{k,j}-\pi_j)\widehat{\theta}_{k-j,N+1-j}(x)\|+\|\pi_j(\widehat{\theta}_{k-j,N+1-j}-\psi_{N+1-j})(x)\|.\end{aligned}$$

Now, for $n\to\infty$, the first summand converges in probability to 0 by part *(i)*, while the second summand converges to 0 in probability by induction. Therefore the statement is proven. □

Proof of Theorem 3.3.5. *(i)* The proof is based again on showing that, for all $x\in H^k$, $\|(\widehat{\widehat{B}}(k)-\Pi(k))(x)\|\overset{p}{\to}0$ as $n\to\infty$, where $\widehat{\widehat{B}}(k)=(\widehat{\widehat{\beta}}_{k,1},\dots,\widehat{\widehat{\beta}}_{k,k})$. To this end, first note that

$$\|(\widehat{\widehat{B}}(k)-\Pi(k))(x)\|\le\|(\widehat{\widehat{B}}(k)-\widehat{B}(k))(x)\|+\|(\widehat{B}(k)-\Pi(k))(x)\|. \tag{3.8.16}$$

Under Assumptions 3.3.2, the second term of the right-hand side converges to 0 in probability for all $x\in H^k$ by part *(i)* of Theorem 3.3.3. The first term of the right-hand side of (3.8.16) can be investigated uniformly over H^k. Using the plug-in estimators defined as in (3.3.13), we get for $k\in\mathbb{N}$

$$\begin{aligned}\|\widehat{\widehat{B}}(k)-\widehat{B}(k)\|_{\mathcal{L}} &= \|\widehat{\widehat{\Gamma}}_{1,k,d}\widehat{\widehat{\Gamma}}_{k,d}^{-1}-\widehat{\Gamma}_{1,k,d}\widehat{\Gamma}_{k,d}^{-1}\|_{\mathcal{L}} \\ &\le\|\big(\widehat{\widehat{\Gamma}}_{1,k,d}-\widehat{\Gamma}_{1,k,d}\big)\widehat{\widehat{\Gamma}}_{k,d}^{-1}\|_{\mathcal{L}}+\|\widehat{\Gamma}_{1,k,d}\big(\widehat{\Gamma}_{k,d}^{-1}-\widehat{\widehat{\Gamma}}_{k,d}^{-1}\big)\|_{\mathcal{L}}.\end{aligned} \tag{3.8.17}$$

Following the same intuition as in the proof of Theorem 3.3.3, start by investigating the term $\|(\widehat{\Gamma}_{k,d}-\widehat{\widehat{\Gamma}}_{k,d})\|_{\mathcal{L}}$. Applying triangle inequality, linearity of the inner product and the inequalities $\|P_{(k)}X_j(k)\|\le\|X_j(k)\|$ and $\|\widehat{P}_{(k)}X_j(k)\|\le\|X_j(k)\|$, it follows that

$$\|(\widehat{\Gamma}_{k,d} - \widehat{\widehat{\Gamma}}_{k,d})\|_{\mathcal{L}} = \left\| \frac{1}{n-k} \sum_{j=k}^{n-1} \big(P_{(k)} X_j(k) \otimes P_{(k)} X_j(k) - \widehat{P}_{(k)} X_j(k) \otimes \widehat{P}_{(k)} X_j(k) \big) \right\|_{\mathcal{L}}$$

$$\leq \frac{2}{n-k} \sum_{j=k}^{n-1} \|X_j(k)\| \|P_{(k)} X_j(k) - \widehat{P}_{(k)} X_j(k)\|. \tag{3.8.18}$$

Note that, from the definitions of $X_j(k)$, $P_{(k)}$ and $\widehat{P}_{(k)}$,

$$P_{(k)} X_j(k) = \left(\sum_{i=1}^{d_k} \langle X_j, \nu_i \rangle \nu_i, \dots, \sum_{i=1}^{d_1} \langle X_{j-k}, \nu_i \rangle \nu_i \right)^\top,$$

$$\widehat{P}_{(k)} X_j(k) = \left(\sum_{i=1}^{d_k} \langle X_j, \widehat{\nu}_i \rangle \widehat{\nu}_i, \dots, \sum_{i=1}^{d_1} \langle X_{j-k}, \widehat{\nu}_i \rangle \widehat{\nu}_i \right)^\top.$$

These relations show that

$$\begin{aligned}
&\|P_{(k)} X_j(k) - \widehat{P}_{(k)} X_j(k)\| \\
&\quad = \left\| \left(\sum_{i=1}^{d_k} \langle X_j, \widehat{\nu}_i \rangle \widehat{\nu}_i - \langle X_j, \nu_i \rangle \nu_i, \dots, \sum_{i=1}^{d_1} \langle X_{j-k}, \widehat{\nu}_i \rangle \widehat{\nu}_i - \langle X_{j-k}, \nu_i \rangle \nu_i \right)^\top \right\| \\
&\quad = \left\| \left(\sum_{i=1}^{d_k} \langle X_j, \widehat{\nu}_i - \nu_i \rangle \widehat{\nu}_i, \dots, \sum_{i=1}^{d_1} \langle X_{j-k}, \widehat{\nu}_i - \nu_i \rangle \widehat{\nu}_i \right)^\top \right\| \\
&\qquad + \left\| \left(\sum_{i=1}^{d_k} \langle X_j, \nu_i \rangle (\nu_i - \widehat{\nu}_i), \dots, \sum_{i=1}^{d_1} \langle X_{j-k}, \nu_i \rangle (\nu_i - \nu_i) \right)^\top \right\|.
\end{aligned}$$

Observe that, for $x = (x_1, \dots, x_k) \in H^k$, $\|x\| = (\sum_{i=1}^k \|x_i\|^2)^{1/2}$, Then, applications of the Cauchy–Schwarz inequality and the orthonormality of $(\nu_i : i \in \mathbb{N})$ and $(\widehat{\nu}_i : i \in \mathbb{N})$ lead to

$$\begin{aligned}
&\|P_{(k)} X_j(k) - \widehat{P}_{(k)} X_j(k)\| \\
&\quad \leq \left(\sum_{i=0}^{k-1} \left\| \sum_{i=1}^{d_i} \langle X_{j-i}, \widehat{\nu}_l - \nu_l \rangle \widehat{\nu}_l \right\|^2 \right)^{1/2} + \left(\sum_{i=0}^{k-1} \left\| \sum_{l=1}^{d_i} \langle X_{j-i}, \nu_l \rangle (\nu_l - \widehat{\nu}_l) \right\|^2 \right)^{1/2} \\
&\quad \leq \left(\sum_{i=0}^{k-1} \sum_{l=1}^{d_i} \|X_{j-i}\|^2 \|\widehat{\nu}_l - \nu_l\|^2 \right)^{1/2} + \left(\sum_{i=0}^{k-1} \sum_{l=1}^{d_i} \|X_{j-i}\|^2 \|\nu_l - \widehat{\nu}_l\|^2 \right)^{1/2} \\
&\quad \leq 2 \left(\sum_{i=0}^{k-1} \sum_{l=1}^{d_k} \|X_{j-i}\|^2 \|\widehat{\nu}_l - \nu_l\|^2 \right)^{1/2} \\
&\quad \leq 2 \|X_j(k)\| \left(\sum_{l=1}^{d_k} \|\widehat{\nu}_l - \nu_l\|^2 \right)^{1/2}.
\end{aligned}$$

Plugging this relation back into (3.8.18), it follows that

$$\|\widehat{\Gamma}_{k,d} - \widehat{\widehat{\Gamma}}_{k,d}\|_{\mathcal{L}} \leq 4\Bigg(\sum_{l=1}^{d_k} \|\widehat{\nu}_l - \nu_l\|^2\Bigg)^{1/2} \frac{2}{n-k} \sum_{j=k}^{n-1} \|X_j(k)\|^2.$$

Since $(X_j : j \in \mathbb{Z})$ is L^4-m approximable, Theorems 3.1 and 3.2 in Hörmann and Kokoszka [24] imply that, for some finite positive constant C_1, $n\mathbb{E}[\|\widehat{\nu}_l - \nu_l\|^2] \leq C_1/\delta_l$, where δ_l is the l-th spectral gap. Hence,

$$\sum_{l=1}^{d_k} \|\widehat{\nu}_l - \nu_l\|^2 \leq \frac{C_1}{n} \sum_{l=1}^{d_k} \frac{1}{\alpha_l^2}.$$

Furthermore, note that

$$\frac{2}{n-k} \sum_{j=k}^{n-1} \mathbb{E}\big[\|X_j(k)\|^2\big] \leq 2 \sum_{i=0}^{k-1} \mathbb{E}\big[\|X_{k-i}\|^2\big] = 2k\|C_X\|_{\mathcal{N}}.$$

Therefore, collecting the previous results yields the rate

$$\|\widehat{\Gamma}_{k,d} - \widehat{\widehat{\Gamma}}_{k,d}\|_{\mathcal{L}} = O_p\Bigg(\frac{k}{n}\Bigg(\sum_{l=1}^{d_k} \frac{1}{\alpha_l^2}\Bigg)^{1/2}\Bigg). \tag{3.8.19}$$

Next, investigate $\|\widehat{\widehat{\Gamma}}_{k,d}^{-1}\|_{\mathcal{L}}$. Similarly as in the corresponding part of the proof of Theorem 3.3.3, it follows that $\|\widehat{\widehat{\Gamma}}_{k,d}^{-1}\|_{\mathcal{L}} \leq \|\widehat{\widehat{\Gamma}}_{k,d}^{-1} - \widehat{\Gamma}_{k,d}^{-1}\|_{\mathcal{L}} + \|\widehat{\Gamma}_{k,d}^{-1}\|_{\mathcal{L}}$. By (3.8.10), $\|\widehat{\Gamma}_{k,d}^{-1}\|_{\mathcal{L}} = O_p(\alpha_{d_k}^{-1})$. Furthermore, the same arguments as in (3.8.7) and (3.8.8) imply that

$$\|\widehat{\widehat{\Gamma}}_{k,d}^{-1} - \widehat{\Gamma}_{k,d}^{-1}\|_{\mathcal{L}} \leq \frac{\|\widehat{\widehat{\Gamma}}_{d,k} - \widehat{\Gamma}_{d,k}\|_{\mathcal{L}} \|\widehat{\Gamma}_{k,d}^{-1}\|_{\mathcal{L}}^2}{1 - \|\widehat{\widehat{\Gamma}}_{d,k} - \widehat{\Gamma}_{d,k}\|_{\mathcal{L}} \|\widehat{\Gamma}_{k,d}^{-1}\|_{\mathcal{L}}}. \tag{3.8.20}$$

Hence, by (3.8.10) and (3.8.19),

$$\|\widehat{\widehat{\Gamma}}_{d,k} - \widehat{\Gamma}_{d,k}\|_{\mathcal{L}} \|\widehat{\Gamma}_{k,d}^{-1}\|_{\mathcal{L}}^2 = O_p\Bigg(\frac{k}{n\alpha_{d_k}^2}\Bigg(\sum_{l=1}^{d_k} \frac{1}{\alpha_l^2}\Bigg)^{1/2}\Bigg).$$

Therefore, by Assumption 3.3.4 as $n \to \infty$, $\|\widehat{\widehat{\Gamma}}_{k,d}^{-1} - \widehat{\Gamma}_{k,d}^{-1}\|_{\mathcal{L}} \xrightarrow{p} 0$. Taken the previous calculations together, this gives the rate

$$\|\widehat{\widehat{\Gamma}}_{k,d}^{-1}\|_{\mathcal{L}} = O_p\Bigg(\frac{1}{\alpha_{d_k}}\Bigg). \tag{3.8.21}$$

Going back to (3.8.17) and noticing that $\|\widehat{\widehat{\Gamma}}_{1,k,d} - \widehat{\Gamma}_{1,k,d}\|_{\mathcal{L}} \leq \|(I_H, 0, \ldots, 0)(\widehat{\widehat{\Gamma}}_{k,d} - \widehat{\Gamma}_{k,d})\|_{\mathcal{L}}$, the first summand in this display can be bounded by

$$\begin{aligned}
\|\big(\widehat{\widehat{\Gamma}}_{1,k,d} - \widehat{\Gamma}_{1,k,d}\big)\widehat{\widehat{\Gamma}}^{-1}_{k,d}\|_{\mathcal{L}} &\leq \|\widehat{\widehat{\Gamma}}_{1,k,d} - \widehat{\Gamma}_{1,k,d}\|_{\mathcal{L}}\|\widehat{\widehat{\Gamma}}^{-1}_{k,d}\|_{\mathcal{L}} \\
&\leq \|(I_H, 0, \ldots, 0)(\widehat{\widehat{\Gamma}}_{k,d} - \widehat{\Gamma}_{k,d})\|_{\mathcal{L}}\|\widehat{\widehat{\Gamma}}^{-1}_{k,d}\|_{\mathcal{L}} \\
&= O_p\left(\frac{k}{n\alpha_{d_k}}\left(\sum_{l=1}^{d_k}\frac{1}{\alpha_l^2}\right)^{1/2}\right),
\end{aligned} \tag{3.8.22}$$

where the rate in (3.8.19) was used in the last step. For the second summand in (3.8.17), use the plug-in estimator for $\widehat{\Gamma}_{1,k,d}$ to obtain, for all $k < n$,

$$\|\widehat{\Gamma}_{1,k,d}\big(\widehat{\Gamma}^{-1}_{k,d} - \widehat{\widehat{\Gamma}}^{-1}_{k,d}\big)\|_{\mathcal{L}} \leq \left\|\frac{1}{n-k}\sum_{j=k}^{n-1} P_{(k)}X_j(k) \otimes X_{j+1}\right\|_{\mathcal{L}} \|\widehat{\Gamma}^{-1}_{k,d} - \widehat{\widehat{\Gamma}}^{-1}_{k,d}\|_{\mathcal{L}}.$$

Since

$$\begin{aligned}
\mathbb{E}\left[\left\|\frac{1}{n-k}\sum_{j=k}^{n-1} P_{(k)}X_j(k) \otimes X_{j+1}\right\|_{\mathcal{L}}\right] &\leq \frac{1}{n-k}\sum_{j=k}^{n-1}\mathbb{E}\big[\|P_{(k)}X_j(k) \otimes X_{j+1}\|_{\mathcal{L}}\big] \\
&\leq \frac{1}{n-k}\sum_{j=k}^{n-1}\big(\mathbb{E}[\|P_{(k)}X_j(k)\|^2]\big)^{1/2}\big(\mathbb{E}[\|X_{j+1}\|^2]\big)^{1/2} \\
&= \left(\sum_{l=0}^{k-1}\mathbb{E}[\|X_{j-l}\|^2]\right)^{1/2}\|C_X\|_{\mathcal{N}}^{1/2} \\
&= \sqrt{k}\|C_X\|_{\mathcal{N}},
\end{aligned}$$

the result in (3.8.20) implies that

$$\left\|\widehat{\Gamma}_{1,k,d}\big(\widehat{\Gamma}^{-1}_{k,d} - \widehat{\widehat{\Gamma}}^{-1}_{k,d}\big)\right\|_{\mathcal{L}} = O_p\left(\frac{k^{3/2}}{n\alpha_{d_k}^2}\left(\sum_{l=1}^{d_k}\frac{1}{\alpha_l^2}\right)^{1/2}\right). \tag{3.8.23}$$

Applying Assumption 3.3.4 to this rate and collecting the results in (3.8.16), (3.8.17), (3.8.22) and (3.8.23), shows that, for all $x \in H^k$ as $n \to \infty$, $\|(\widehat{\widehat{B}}(k) - \Pi(k))(x)\| \xrightarrow{p} 0$. This is the claim.

(ii) Similar to the proof of part *(ii)* of Theorem 3.3.3. □

Bibliography

[1] A. Antoniadis, E. Paparoditis, and T. Sapatinas. A functional wavelet-kernel approach for time series prediction. *Journal of the Royal Statistical Society, Series B*, 68:837–857, 2006.

[2] A. Aue and A. Van Delft. Testing for stationarity of functional time series in the frequency domain. *preprint: arXiv:1701.01741[ME]*, 2017.

[3] A. Aue and J. Klepsch. Estimating functional time series by moving average model fitting. *preprint at arXiv:1701.00770[ME]*, 2017.

[4] A. Aue, S. Hörmann, L. Horváth, and M. Reimherr. Detecting changes in the covariance structure of multivariate time series. *The Annals of Statistics*, 37: 4046–4087, 2009.

[5] A. Aue, D. Dubart Norinho, and S. Hörmann. On the prediction of stationary functional time series. *Journal of the American Statistical Association*, 110: 378–392, 2015.

[6] P. Bernard. Analyse de signaux physiologiques. *Mémoire Université Catholique Angers*, 1997.

[7] P. Besse and H. Cardot. Approximation spline de la prévision d'un processus fonctionnel autorégressif d'ordre 1. *Canadian Journal of Statistics*, 24:467–487, 1996.

[8] D. Bosq. *Linear Processes in Function Spaces: Theory and Applications.* Springer, New York, 2000.

[9] D. Bosq. General linear processes in Hilbert spaces and prediction. *Journal of Statistical Planning and Inference*, 137:879–894, 2007.

[10] D. Bosq. Computing the best linear predictor in a Hilbert space. Applications to general ARMAH processes. *Journal of Multivariate Analysis*, 124:436–450, 2014.

[11] D. Bosq and D. Blanke. *Inference and Prediction in Large Dimensions*. Wiley-Dunod, Chichester, 2007.

[12] G. E. P. Box and G. M. Jenkins. *Time Series Analysis - Forecasting and Control*. San Francisco: Holden Day, 1970.

[13] P.J. Brockwell and R.A. Davis. *Time Series: Theory and Methods (2nd Ed.)*. Springer, New York, 1991.

[14] C. Cerovecki and S. Hörmann. On the CLT for discrete Fourier transforms of functional time series. *Journal of Multivariate Analysis*, 154:282–295, 2017.

[15] R. Chrobok, O. Kaumann, J. Wahle, and M. Schreckenberg. Different methods of traffic forecast based on real data. *European Journal of Operational Research*, 155:558–568, 2004.

[16] N. Dunford and J.T. Schwartz. *Linear Operators Part 1 General Theory*. Wiley, 1988.

[17] F. Ferraty and Y. Romain. *The Oxford Handbook of Functional Data Analysis*. Oxford Press, 2011.

[18] F. Ferraty and P. Vieu. *Nonparametric Functional Data Analysis: Theory and Practice*. Springer, New York, 2006.

[19] R. Fortet. *Vecteurs, fonctions et distributions alÃľatoires dans les espaces de Hilbert*. Hermes, Paris, 1995.

[20] R. Gabrys and P. Kokoszka. Portmanteau test of independence for functional observations. *Journal of the American Statistical Association*, 102:1338–1348, 2007.

[21] I. Gohberg, S. Goldberg, and M.A. Kaashoek. *Basic Classes of Linear Operators*. Birkhäuser, Basel, 2003.

[22] J. D. Hamilton. *Time Series Analysis*. Princeton University Press, Princeton, 1994.

[23] E.J. Hannan. *Multiple Time Series*. Wiley, New York, 1970.

[24] S. Hörmann and P. Kokoszka. Weakly dependent functional data. *Annals of Statistics*, 38:1845–1884, 2010.

[25] S. Hörmann, L. Kidziński, and M. Hallin. Dynamic functional principal components. *Journal of the Royal Statistical Society: Series B*, 77:319–348, 2015.

[26] S. Hörmann, P. Kokoszka, and G. Nisol. Detection of periodicity in functional time series. *eprint: arXiv:1607.02017v1*, 2016.

[27] L. Horvàth and P. Kokoszka. *Inference for Functional Data with Applications*. Springer, New York, 2012.

[28] L. Horvàth, P. Kokoszka, and G. Rice. Testing stationarity of functional time series. *Journal of Econometrics*, 179:66–82, 2013.

[29] T. Hsing and R. Eubank. *Theoretical Foundations of Functional Data Analysis, with an Introduction to Linear Operators*. Wiley, West Sussex, UK, 2015.

[30] R.J. Hyndman and H.L. Shang. Forecasting functional time series. *Journal of the Korean Statistical Society*, 38:199–221, 2009.

[31] J. Jacod and P.E. Protter. *Discretization of Processes*. Springer, Berlin Heidelberg, 2012.

[32] V. Kargin and A. Onatski. Curve forecasting by functional autoregression. *Journal of Multivariate Analysis*, 99:2508–2526, 2008.

[33] G. Kirchgässner, J. Wolters, and U. Hassler. *Introduction to Modern Time Series Analysis*. Springer, New York, 2000.

[34] J. Klepsch and C. Klüppelberg. An Innovations Algorithm for the prediction of functional linear processes. *Journal of Multivariate Analysis*, 155:252–271, 2017.

[35] J. Klepsch, C. Klüppelberg, and T. Wei. Prediction of functional ARMA processes with an application to traffic data. *Econometrics and Statistics*, 1:128–149, 2017.

[36] P. Kokoszka and M. Reimherr. Determining the order of the functional autoregressive model. *Journal of Time Series Analysis*, 34:116–129, 2013.

[37] T.L. Lai and C.P. Lee. Information and prediction criteria for model selection in stochastic regression and ARMA models. *Statistica Sinica*, 7:285–309, 1997.

[38] R. Lewis and G.C. Reinsel. Prediction of multivariate time series by autoregressive model fitting. *Journal of Multivariate Analysis*, 16:393–411, 1985.

[39] H. Lütkepohl. *New Introduction to Multiple Time Series Analysis*. Springer Berlin Heidelberg, 2005.

[40] F. Merlevède. Sur l'inversibilité des processus linéaires à valeurs dans un espace de Hilbert. *Comptes Rendus de l'Académie des Sciences Paris, Serie I*, 321: 477–480, 1995.

[41] H. Mitchell. *Topics in Multiple Time Series*. PhD thesis, Royal Melbourne Institute of Technology, 1996.

[42] H. Mitchell and P.J. Brockwell. Estimation of the coefficients of a multivariate linear filter using the Innovations Algorithm. *Journal of Time Series Analysis*, 18:157–179, 1997.

[43] S. Nsiri and R. Roy. On the invertibility of multivariate linear processes. *Journal of Time Series Analysis*, 14:305–316, 1993.

[44] V. M. Panaretos and S. Tavakoli. Fourier analysis of stationary time series in function spaces. *Annals of Statistics*, 41:568–603, 2013.

[45] V. M. Panaretos and S. Tavakoli. Cramér-Karhunen-Loève representation and harmonic component analysis of functional time series. *Stochastic Processes and Their Applications*, 123:2779–2807, 2013.

[46] M. B. Priestley. *Spectral analysis and time series*. Academic Press, 1982.

[47] J. O. Ramsay and B. W. Silverman. *Functional Data Analysis*. Springer Series in Statistics, 2005.

[48] J.O. Ramsay and B.W. Silverman. *Applied Functional Data Analysis*. Springer Series in Statistics, 2002.

[49] R. H. Shumway and D. S. Stoffer. *Time Series Analysis and Its Applications, 3rd Edition*. Springer, New York, 2011.

[50] B. Simon. *Operator Theory - A Comprehensive Course in Analysis, Part 4*. AMS, 2015.

[51] F. Spangenberg. Strictly stationary solutions of ARMA equations in Banach spaces. *Journal of Multivariate Analysis*, 121:127–138, 2013.

[52] R. S. Tsay. Time series and forecasting: Brief history and future research. *Journal of the American Statistical Association*, 95:638–643, 2000.

[53] R.S. Tsay. *Multivariate Time Series Analysis*. Wiley, Hoboken, 2014.

[54] C. Turbillon, D. Bosq, J.M. Marion, and B. Pumo. Parameter estimation of moving averages in Hilbert spaces. *Comptes rendus de l'Académie des Sciences, Série I*, 346:347–350, 2008.

[55] T. Wei. Time series in functional data analysis. Master's thesis, Technical University of Munich, http://mediatum.ub.tum.de/node?id=1291335, 12 2015.

[56] H. Wold. *A Study in the Analysis of Stationary Time Series*. PhD thesis, Stockholm, 1938.