Functional Time Series

Article *in* Handbook of Statistics · December 2012

DOI: 10.1007/978-1-4614-3655-3_16

CITATIONS
61

READS
3,201

2 authors, including:

Piotr Kokoszka
Colorado State University
173 PUBLICATIONS **5,876** CITATIONS

SEE PROFILE

Functional time series

Siegfried Hörmann and Piotr Kokoszka

June 3, 2011

Contents

1 Introduction **2**
1.1 Examples of functional time series 2

2 The Hilbert space model for functional data **4**
2.1 Operators 5
2.2 The space L^2 6
2.3 Functional mean and the covariance operator 6
2.4 Empirical functional principal components 7
2.5 Population functional principal components 8

3 Functional autoregressive model **10**
3.1 Existence 10
3.2 Estimation 11
3.3 Prediction 15

4 Weakly dependent functional time series **19**
4.1 Approximable functional sequences 19
4.2 Estimation of the mean function and the FPC's 21
4.3 Estimation of the long run variance 23

5 Further reading **28**

References **29**

1 Introduction

Functional data often arise from measurements obtained by separating an almost continuous time record into natural consecutive intervals, for example days. Examples include daily curves of financial transaction data and daily patterns of geophysical and environmental data. The functions thus obtained form a time series $\{X_k,\, k \in \mathbb{Z}\}$ where each X_k is a (random) function $X_k(t),\, t \in [a,b]$. We refer to such data structures as *functional time series*; examples are given in Section 1.1. A central issue in the analysis of such data is to take into account the temporal dependence of the observations, i.e. the dependence between events determined by $\{X_k,\, k \leq m\}$ and $\{X_k,\, k \geq m+h\}$. While the literature on scalar and vector time series is huge, there are relatively few contributions dealing with functional time series. The focus of functional data analysis has been mostly on iid functional observations. It is therefore hoped that the present survey will provide an informative account of a useful approach that merges the ideas of time series analysis and functional data analysis.

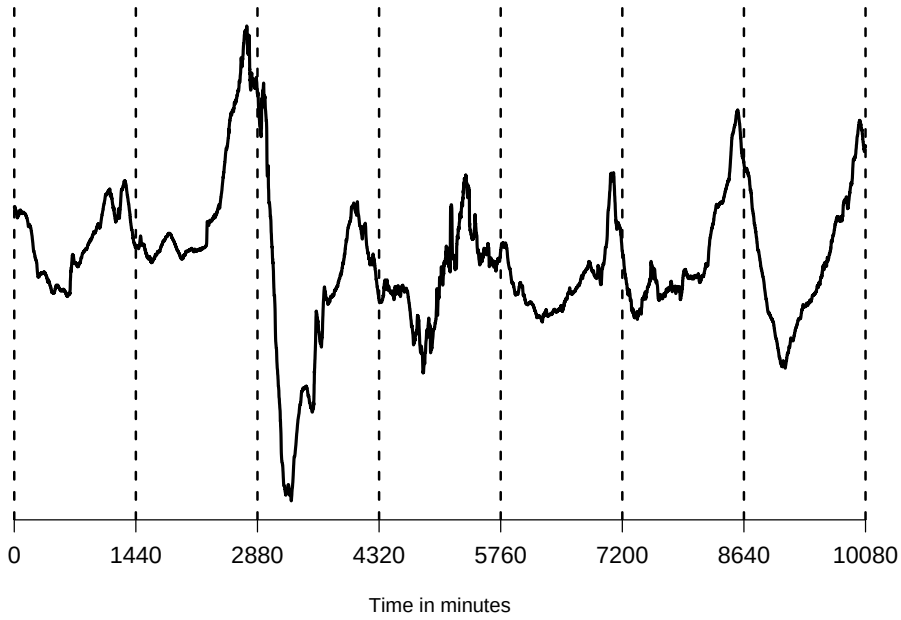
The monograph of Bosq (2000) studies the theory of linear functional time series, both in Hilbert and Banach spaces, focusing on the functional autoregressive model. For many functional time series it is however not clear what specific model they follow, and for many statistical procedures it is not necessary to assume a specific model. In such cases, it is important to know what the effect of the dependence on a given procedure is. Is it robust to temporal dependence, or does this type of dependence introduce a serious bias? To answer questions of this type, it is essential to quantify the notion of temporal dependence. Again, for scalar and vector time series, this question has been approached from many angles, but, except for the linear model of Bosq (2000), for functional time series, no general framework has been available. We present a moment based notion of weak dependence proposed in Hörmann and Kokoszka (2010).

To make this account as much self-contained as possible, we set in Section 2 the mathematical framework required for this contribution, and also report some results for iid data, in order to allow for a comparison between results for serially dependent and independent functional data. Next, in Section 3, we introduce the autoregressive model of Bosq (2000) and discuss its applications. In Section 4, we outline the notion of dependence proposed in Hörmann and Kokoszka (2010), and show how it can be applied to the analysis of functional time series. References to related topics are briefly discussed in Section 5.

1.1 Examples of functional time series

The data that motivated the research presented here are of the form $X_k(t),\ t \in [a,b]$. The interval $[a,b]$ is typically normalized to be a unit interval $[0,1]$. The treatment of the endpoints depends on the way the data are collected. For intradaily financial transactions data, a is the opening time and b is the closing time of an exchange, for example the NYSE, so both endpoints are included. Geophysical data are typically of the form $X(u)$ where u is measured at a very fine time grid. After normalizing to the unit interval, the curves are defined as $X_k(t) = X(k+t),\ 0 \leq t < 1$. In both cases, an observation is thus

FIGURE 1.1 The horizontal component of the magnetic field measured in one minute resolution at Honolulu magnetic observatory from 1/1/2001 00:00 UT to 1/7/2001 24:00 UT.

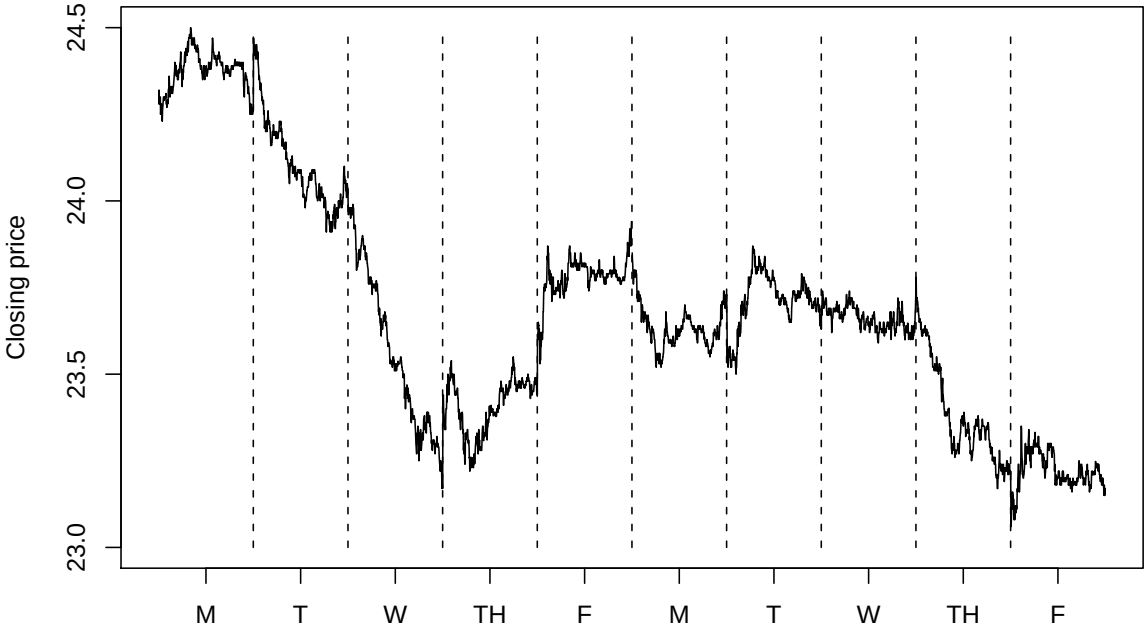


a curve.

Figure 1.1 shows a reading of a magnetometer over a period of one week. A magnetometer is an instrument that measures the three components of the magnetic field at a location where it is placed. There are over 100 magnetic observatories located on the surface of the Earth, and most of them have digital magnetometers. These magnetometers record the strength and direction of the field every five seconds, but the magnetic field exists at any moment of time, so it is natural to think of a magnetogram as an approximation to a continuous record. The raw magnetometer data are cleaned and reported as averages over one minute intervals. Such averages were used to produce Figure 1.1. Thus $7 \times 24 \times 60 = 10,080$ values (of one component of the field) were used to draw Figure 1.1. The vertical lines separate days in Universal Time (UT). It is natural to view a curve defined over one UT day as a single observation because one of the main sources influencing the shape of the record is the daily rotation of the Earth. When an observatory faces the Sun, it records the magnetic field generated by wind currents flowing in the ionosphere which are driven mostly by solar heating. Figure 1.1 thus shows seven consecutive observations of a functional time series.

Examples of data that can be naturally treated as functional also come from financial records. Figure 1.2 shows two consecutive weeks of Microsoft stock prices in one minute resolution. A great deal of financial research has been done using the closing daily price, i.e. the price in the last transaction of a trading day. However many assets are traded so frequently that one can practically think of a price curve that is defined at any moment

FIGURE 1.2 Microsoft stock prices in one-minute resolution, May 1-5, 8-12, 2006



of time. The Microsoft stock is traded several hundred times per minute. The values used to draw the graph in Figure 1.2 are the closing prices in one-minute intervals. It is natural to choose one trading day as the underlying time interval. If we do so, Figure 1.2 shows 10 consecutive functional observations. In contrast to magnetometer data, the price in the last minute of day k does not have to be close to the price in the first minute of day $k+1$.

2 The Hilbert space model for functional data

It is typically assumed that the observations X_k are elements of a separable Hilbert space H (i.e. a Hilbert space with a countable a basis $\{e_k,\, k \in \mathbb{Z}\}$) with inner product $\langle \cdot, \cdot \rangle$ which generates norm $\| \cdot \|$. This is what we assume in the following. An important example is the Hilbert space $L^2 = L^2([0,1])$ introduced in Section 2.2. Although we formally allow for a general Hilbert spaces, we call our H–valued data *functional observations*. All random functions are defined on some common probability space $(\Omega, \mathcal{A}, P)$. We say that X is integrable if $E\|X\| < \infty$, and we say it is square integrable if $E\|X\|^2 < \infty$. If $E\|X\|^p < \infty$, $p > 0$, we write $X \in L^p_H = L^p_H(\Omega, \mathcal{A}, P)$. Convergence of $\{X_n\}$ to X in L^p_H

means $E\|X_n - X\|^p \to 0$, whereas $\|X_n - X\| \to 0$ almost surely (a.s.) is referred to as almost sure convergence.

In this section, we follow closely the exposition in Bosq (2000). Good references on Hilbert spaces are Riesz and Sz.-Nagy (1990), Akhiezier and Glazman (1993) and Debnath and Mikusinski (2005). An in–depth theory of operators in a Hilbert space is developed in Gohberg *et al.* (1990).

2.1 Operators

We let $\langle \cdot, \cdot \rangle$ be the inner product in H which generates the norm $\| \cdot \|$, and denote by $\mathcal{L}$ the space of bounded (continuous) linear operators on H with the norm

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

An operator $\Psi \in \mathcal{L}$ is said to be *compact* if there exist two orthonormal bases $\{v_j\}$ and $\{f_j\}$, and a real sequence $\{\lambda_j\}$ converging to zero, such that

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

The λ_j are assumed positive because one can replace f_j by $-f_j$, if needed. Representation (2.1) is called the *singular value decomposition*. Compact operators are also called *completely continuous* operators.

A compact operator admitting a representation (2.1) is said to be a *Hilbert–Schmidt* operator if $\sum_{j=1}^{\infty} \lambda_j^2 < \infty$. The space $\mathcal{S}$ of Hilbert–Schmidt operators is a separable Hilbert space with the scalar product

$$\langle \Psi_1, \Psi_2 \rangle_{\mathcal{S}} = \sum_{i=1}^{\infty} \langle \Psi_1(e_i), \Psi_2(e_i) \rangle, \tag{2.2}$$

where $\{e_i\}$ is an arbitrary orthonormal basis, the value of (2.2) does not depend on it. One can show that $\|\Psi\|_{\mathcal{S}}^2 = \sum_{j \geq 1} \lambda_j^2$ and

$$\|\Psi\|_{\mathcal{L}} \leq \|\Psi\|_{\mathcal{S}}. \tag{2.3}$$

An operator $\Psi \in \mathcal{L}$ is said to be *symmetric* if

$$\langle \Psi(x), y \rangle = \langle x, \Psi(y) \rangle, \quad x, y \in H,$$

and positive–definite if

$$\langle \Psi(x), x \rangle \geq 0, \quad x \in H.$$

(An operator with the last property is sometimes called positive semidefinite, and the term positive–definite is used when $\langle \Psi(x), x \rangle > 0$ for $x \neq 0$.)

A symmetric positive–definite Hilbert–Schmidt operator Ψ admits the decomposition

$$\Psi(x) = \sum_{j=1}^{\infty} \lambda_j \langle x, v_j \rangle v_j, \quad x \in H, \tag{2.4}$$

with orthonormal v_j which are the eigenfunctions of Ψ, i.e. $\Psi(v_j) = \lambda_j v_j$. The v_j can be extended to a basis by adding a complete orthonormal system in the orthogonal complement of the subspace spanned by the original v_j. The v_j in (2.4) can thus be assumed to form a basis, but some λ_j may be zero.

2.2 The space L^2

The space L^2 is the set of measurable real–valued functions x defined on $[0,1]$ satisfying $\int_0^1 x^2(t)dt < \infty$. It is a separable Hilbert space with the inner product

$$\langle x, y \rangle = \int x(t)y(t)dt.$$

An integral sign without the limits of integration is meant to denote the integral over the whole interval $[0,1]$. If $x, y \in L^2$, the equality $x = y$ always means $\int [x(t) - y(t)]^2 dt = 0$.

An important class of operators in L^2 are the integral operators defined by

$$\Psi(x)(t) = \int \psi(t,s)x(s)ds, \quad x \in L^2, \tag{2.5}$$

with the real kernel $\psi(\cdot,\cdot)$. Such operators are Hilbert–Schmidt if and only if $\iint \psi^2(t,s)dtds < \infty$, in which case

$$\|\Psi\|_S^2 = \iint \psi^2(t,s)dtds. \tag{2.6}$$

If $\psi(s,t) = \psi(t,s)$ and $\iint \psi(t,s)x(t)x(s)dtds \geq 0$, the integral operator Ψ is symmetric and positive–definite, and it follows from (2.4) that

$$\psi(t,s) = \sum_{j=1}^{\infty} \lambda_j v_j(t) v_j(s) \quad \text{in } L^2([0,1]\times[0,1]). \tag{2.7}$$

If ψ is continuous, the above expansions holds for all $s,t \in [0,1]$, and the series converges uniformly. This result is known as Mercer's theorem, see e.g. Riesz and Sz.-Nagy (1990).

2.3 Functional mean and the covariance operator

Let $X, X_1, X_2, \ldots$ be H–valued random functions. We call X weakly integrable if there is a $\mu \in H$ such that $E\langle X, y\rangle = \langle \mu, y\rangle$ for all $y \in H$. In this case μ is called the expectation of X, short EX. Some elementary results are: (a) EX is unique, (b) integrability implies

weak integrability and (c) $\|EX\| \le E\|X\|$. In the special case where $H = L^2$ one can show that $\{(EX)(t),\, t \in [0,1]\} = \{E(X(t)),\, t \in [0,1]\}$, i.e. one can obtain the mean function by point–wise evaluation. The expectation commutes with bounded operators, i.e. if $\Psi \in \mathcal{L}$ and X is integrable, then $E\Psi(X) = \Psi(EX)$.

For $X \in L^2_H$ the covariance operator of X is defined by

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

The covariance operator C is symmetric and positive–definite, with eigenvalues λ_i satisfying

$$\sum_{i=1}^{\infty} \lambda_i = E\|X - EX\|^2 < \infty. \tag{2.8}$$

Hence C is a symmetric positive–definite Hilbert–Schmidt operator admitting representation (2.4).

The sample mean and the sample covariance operator of $X_1, \ldots, X_N$ are defined as follows:

$$\hat{\mu}_N = \frac{1}{N}\sum_{k=1}^{N} X_k \quad \text{and} \quad \hat{C}_N(y) = \frac{1}{N}\sum_{k=1}^{N} \langle X_k - \hat{\mu}_N, y\rangle\,(X_k - \hat{\mu}_N), \quad y \in H.$$

The following result implies the consistency of the just defined estimators for iid samples.

THEOREM 2.1 *Let $\{X_k\}$ be an H-valued iid sequence with $EX = \mu$.*
(a) If $X_1 \in L^2_H$ then $E\|\hat{\mu}_N - \mu\|^2 = O(N^{-1})$.
(b) If $X_1 \in L^4_H$ then $E\|\hat{C}\|^2_{\mathcal{S}} < \infty$ and $E\|C - \hat{C}\|^2_{\mathcal{S}} = O(N^{-1})$.

In Section 4 we will prove Theorem 2.1 in a more general framework, namely for a stationary, weakly dependent sequence.

It is easy to see that for $H = L^2$,

$$C(y)(t) = \int c(t,s)y(s)ds, \quad \text{where} \quad c(t,s) = \mathrm{Cov}(X(t), X(s)).$$

The covariance kernel $c(t,s)$ is estimated by

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

2.4 Empirical functional principal components

Suppose we observe functions $x_1, x_2, \ldots, x_N$. In this section it is not necessary to view these functions as random, but we can think of them as the observed realizations of

random functions in some separable Hilbert space H. We assume that the data have been centered, i.e. $\sum_{i=1}^{N} x_i = 0$. Fix an integer $p < N$. We think of p as being much smaller than N, typically a single digit number. We want to find an orthonormal basis $u_1, u_2, \ldots, u_p$ such that

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

is minimized. Once such a basis is found, $\sum_{k=1}^{p} \langle x_i, u_k \rangle u_k$ is an approximation to x_i. For the p we have chosen, this approximation is uniformly optimal, in the sense of minimizing $\hat{S}^2$. This means that instead of working with infinitely dimensional curves x_i, we can work with p–dimensional vectors

$$\mathbf{x}_i = [\langle x_i, u_1 \rangle, \langle x_i, u_2 \rangle, \ldots, \langle x_i, u_p \rangle]^T.$$

This is a central idea of functional data analysis, as to perform any practical calculations we must reduce the dimension from infinity to a finite number. The functions u_j are called collectively the *optimal empirical orthonormal basis* or *natural orthonormal components*, the words "empirical" and "natural" emphasizing that they are computed directly from the functional data.

The functions $u_1, u_2, \ldots, u_p$ minimizing $\hat{S}^2$ are equal (up to a sign) to the normalized eigenfunctions, $\hat{v}_1, \hat{v}_2, \ldots, \hat{v}_p$, of the sample covariance operator, i.e. $\hat{C}(u_i) = \hat{\lambda}_i u_i$, where $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \cdots \geq \hat{\lambda}_p$. The eigenfunctions $\hat{v}_i$ are called the empirical functional principal components (EFPC's) of the data $x_1, x_2, \ldots, x_N$. The $\hat{v}_i$ are thus the natural orthonormal components and form the optimal empirical orthonormal basis.

2.5 Population functional principal components

Suppose $X_1, X_2, \ldots, X_N$ are zero mean functional observations in H having the same distribution as X. Parallel to Section 2.4 we can ask which orthonormal elements $v_1, \ldots, v_p$ in H minimize

$$E \left\| X - \sum_{i=1}^{p} \langle X, v_i \rangle v_i \right\|^2,$$

and in view of Section 2.5 the answer is not surprising. The eigenfunctions v_i of the covariance operator C allow for the "optimal" representation of X. The functional principal components (FPC's) are defined as the eigenfunctions of the covariance operator C of X. The representation

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

is called the *Karhunen-Loève* expansion.

The inner product $\langle X_i, v_j \rangle = \int X_i(t) v_j(t) dt$ is called the jth score of X_i. It can be interpreted as the weight of the contribution of the FPC v_j to the curve X_i.

We often estimate the eigenvalues and eigenfunctions of C, but the interpretation of these quantities as parameters, and their estimation, must be approached with care. The eigenvalues must be identifiable, so we must assume that $\lambda_1 > \lambda_2 > \ldots$. In practice, we can estimate only the p largest eigenvalues, and assume that $\lambda_1 > \lambda_2 > \ldots > \lambda_p > \lambda_{p+1}$, which implies that the first p eigenvalues are nonzero. The eigenfunctions v_j are defined by $C(v_j) = \lambda_j v_j$, so if v_j is an eigenfunction, then so is av_j, for any nonzero scalar a (by definition, eigenfunctions are nonzero). The v_j are typically normalized, so that $||v_j|| = 1$, but this does not determine the sign of v_j. Thus if $\hat{v}_j$ is an estimate computed from the data, we can only hope that $\hat{c}_j \hat{v}_j$ is close to v_j, where

$$\hat{c}_j = \text{sign}(\langle \hat{v}_j, v_j \rangle).$$

Note that $\hat{c}_j$ cannot be computed form the data, so it must be ensured that the statistics we want to work with do not depend on the $\hat{c}_j$.

With these preliminaries in mind, we define the estimated eigenelements by

$$\hat{C}_N(\hat{v}_j) = \hat{\lambda}_j \hat{v}_j, \qquad j = 1, 2, \ldots N. \tag{2.9}$$

The following result, established in Dauxois *et al.* (1982) and Bosq (2000), is used very often to develop asymptotic arguments.

THEOREM 2.2 *Assume that the observations $X_1, X_2, \ldots, X_N$ are iid in H and have the same distribution as X, which is assumed to be in L_H^4 with $EX = 0$. Suppose that*

$$\lambda_1 > \lambda_2 > \ldots > \lambda_d > \lambda_{d+1}. \tag{2.10}$$

Then, for each $1 \leq j \leq d$,

$$E\left[||\hat{c}_j \hat{v}_j - v_j||^2\right] = O(N^{-1}), \qquad E\left[|\lambda_j - \hat{\lambda}_j|^2\right] = O(N^{-1}). \tag{2.11}$$

Theorem 2.2 implies that, under regularity conditions, the population eigenfunctions can be consistently estimated by the empirical eigenfunctions. If the assumptions do not hold, the direction of the $\hat{v}_k$ may not be close to the v_k. Examples of this type, with many references, are discussed in Johnstone and Lu (2009). These examples show that if the iid curves are noisy, then (2.11) fails. Another setting in which (2.11) may fail is when the curves are sufficiently regular, but the dependence between them is too strong. Such examples are discussed in **?**.

The proof of Theorem 2.2 is immediate from part (b) of Theorem 2.1, and Lemmas 2.1 and 2.2, which we will also use in Section 4. These two Lemmas appear, in a slightly more specialized form, as Lemmas 4.2 and 4.3 of Bosq (2000). Lemma 2.1 is proven in Section VI.1 of Gohberg *et al.* (1990), see their Corollary 1.6 on p. 99., while Lemma 2.2 is established in Horváth and Kokoszka (2011). To formulate Lemmas 2.1 and 2.2, we consider two compact operators $C, K \in \mathcal{L}$ with singular value decompositions

$$C(x) = \sum_{j=1}^{\infty} \lambda_j \langle x, v_j \rangle f_j, \quad K(x) = \sum_{j=1}^{\infty} \gamma_j \langle x, u_j \rangle g_j. \tag{2.12}$$

LEMMA 2.1 *Suppose $C, K \in \mathcal{L}$ are two compact operators with singular value decompositions* (2.12). *Then, for each $j \geq 1$, $|\gamma_j - \lambda_j| \leq ||K - C||_{\mathcal{L}}$.*

We now define

$$v_j' = c_j v_j, \quad c_j = \text{sign}(\langle u_j, v_j \rangle).$$

LEMMA 2.2 *Suppose $C, K \in \mathcal{L}$ are two compact operators with singular value decompositions* (2.12). *If C is symmetric, $f_j = v_j$ in* (2.12), *and its eigenvalues satisfy* (2.10), *then*

$$||u_j - v_j'|| \leq \frac{2\sqrt{2}}{\alpha_j} ||K - C||_{\mathcal{L}}, \quad 1 \leq j \leq d,$$

where $\alpha_1 = \lambda_1 - \lambda_2$ and $\alpha_j = \min(\lambda_{j-1} - \lambda_j, \lambda_j - \lambda_{j+1})$, $2 \leq j \leq d$.

We note that if C is a covariance operator, then it satisfies the conditions imposed on C in Lemma 2.2. The v_j are then the eigenfunctions of C. Since these eigenfunctions are determined only up to a sign, it is necessary to introduce the functions v_j'.

This section has merely set out the fundamental definitions and properties. Interpretation and estimation of the functional principal components has been a subject of extensive research, in which concepts of smoothing and regularization play a major role, see Chapters 8, 9, 10 of Ramsay and Silverman (2005).

3 Functional autoregressive model

The theory of autoregressive and more general linear processes in Hilbert and Banach spaces is developed in the monograph of Bosq (2000), on which Sections 3.1 and 3.2 are based, and which we also refer to for the proofs. We present only a few selected results which provide an introduction to the central ideas. Section 3.3 is devoted to prediction by means of the functional autoregressive (FAR) process. To lighten the notation, we set in this chapter, $\|\cdot\|_{\mathcal{L}} = \|\cdot\|$.

3.1 Existence

We say that a sequence $\{X_n, \ -\infty < n < \infty\}$ of mean zero functions in H follows a functional AR(1) model if

$$X_n = \Psi(X_{n-1}) + \varepsilon_n, \tag{3.1}$$

where $\Psi \in \mathcal{L}$ and $\{\varepsilon_n, \ -\infty < n < \infty\}$ is a sequence of iid mean zero errors in H satisfying $E\|\varepsilon_n\|^2 < \infty$.

The above definition defines a somewhat narrower class of processes than that considered by Bosq (2000) who does not assume that the $\{\varepsilon_n\}$ are iid, but rather that they are uncorrelated in an appropriate Hilbert space sense, see his Definitions 3.1 and 3.2. The theory of estimation for the process (3.1) is however developed only under the assumption that the errors are iid.

Scalar AR(1) equations, $X_n = \psi X_{n-1} + \varepsilon_n$, admit the unique causal solution $X_n = \sum_{j=0}^{\infty} \psi^j \varepsilon_{n-j}$, if $|\psi| < 1$. Our goal in this section is to state a condition analogous to $|\psi| < 1$ for functional AR(1) equations (3.1). We begin with the following lemma:

LEMMA 3.1 *For any $\Psi \in \mathcal{L}$, the following two conditions are equivalent:*

C0: The exists an integer j_0 such that $\|\Psi^{j_0}\| < 1$.

C1: There exist $a > 0$ and $0 < b < 1$ such that for every $j \geq 0$, $\|\Psi^j\| \leq ab^j$.

Note that condition C0 is weaker than the condition $\|\Psi\| < 1$; in the scalar case these two conditions are clearly equivalent. Nevertheless, C1 is a sufficiently strong condition to ensure the convergence of the series $\sum_j \Psi^j(\varepsilon_{n-j})$, and the existence of a stationary causal solution to functional AR(1) equations, as stated in Theorem 3.1.

Note that (3.1) can be viewed as an iterated random function system, see Diaconis and Freeman (1999) and Wu and Shao (2004). Condition C1 then refers to a geometric contraction property needed to obtain stationary solutions for such processes. Since iterated random function systems have been studied on general metric spaces, we could use this methodology to investigate extensions of the functional AR process to non-linear functional Markov processes of the form $X_t = \Psi_{Z_t}(X_{t-1})$.

THEOREM 3.1 *If condition C0 holds, then there is a unique strictly stationary causal solution to* (3.1). *This solution is given by*

$$X_n = \sum_{j=0}^{\infty} \Psi^j(\varepsilon_{n-j}). \tag{3.2}$$

The series converges almost surely and in L^2_H.

EXAMPLE 3.1 Consider an integral Hilbert–Schmidt operator on L^2 defined by (2.5), which satisfies

$$\iint \psi^2(t,s)dtds < 1. \tag{3.3}$$

Recall from Section 2.2 that the left–hand side of (3.3) is equal to $\|\Psi\|^2_{\mathcal{S}}$. Since $\|\Psi\| \leq \|\Psi\|_{\mathcal{S}}$, we see that (3.3) implies condition C0 of Lemma 3.1 with $j_0 = 1$.

3.2 Estimation

This section is devoted to the estimation of the autoregressive operator Ψ, but first we state a theorem on the convergence of the EFPC's and the corresponding eigenvalues, which follows from Example 4.1, Theorem 4.3 and Lemma 3.1. In essence, Theorem 3.2 states that bounds (2.11) also hold if the X_n follow an FAR(1) model.

THEOREM 3.2 *Suppose the operator* Ψ *in* (3.1) *satisfies condition C0 of Lemma 3.1, and the solution* $\{X_n\}$ *satisfies* $E\|X_0\|^4 < \infty$. *If* (2.10) *holds, then, for each* $1 \le j \le d$, *relations* (2.11) *hold.*

We now turn to the estimation of the autoregressive operator Ψ. It is instructive to focus first on the univariate case $X_n = \psi X_{n-1} + \varepsilon_n$, in which all quantities are scalars. We assume that $|\psi| < 1$, so that there is a stationary solution such that ε_n is independent of X_{n-1}. Then, multiplying the AR(1) equation by X_{n-1} and taking the expectation, we obtain $\gamma_1 = \psi\gamma_0$, where $\gamma_k = E[X_n X_{n+k}] = \mathrm{Cov}(X_n, X_{n+k})$. The autocovariances γ_k are estimated by the sample autocovariances $\hat{\gamma}_k$, so the usual estimator of ψ is $\hat{\psi} = \hat{\gamma}_1/\hat{\gamma}_0$. This estimator is optimal in many ways, see Chapter 8 of Brockwell and Davis (1991), and the approach outlined above, known as the Yule-Walker estimation, works for higher order and multivariate autoregressive processes. To apply this technique to the functional model, note that by (3.1), under condition C0 of Lemma 3.1,

$$E\left[\langle X_n, x\rangle\, X_{n-1}\right] = E\left[\langle \Psi(X_{n-1}), x\rangle\, X_{n-1}\right], \qquad x \in H.$$

Define the lag–1 autocovariance operator by

$$C_1(x) = E[\langle X_n, x\rangle\, X_{n+1}]$$

and denote with superscript $\cdot^T$ the adjoint operator. Then, $C_1^T = C\Psi^T$ because, by a direct verification, $C_1^T = E\left[\langle X_n, x\rangle\, X_{n-1}\right]$, i.e.

$$C_1 = \Psi C. \tag{3.4}$$

The above identity is analogous to the scalar case, so we would like to obtain an estimate of Ψ by using a finite sample version of the relation $\Psi = C_1 C^{-1}$. The operator C does not however have a bounded inverse on the whole of H. To see it, recall that C admits representation (2.4), which implies that $C^{-1}(C(x)) = x$, where

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

The operator C^{-1} is defined if all λ_j are positive. Since $\|C^{-1}(v_n)\| = \lambda_n^{-1} \to \infty$, as $n \to \infty$, it is unbounded. This makes it difficult to estimate the bounded operator Ψ using the relation $\Psi = C_1 C^{-1}$. A practical solution is to use only the first p most important EFPC's $\hat{v}_j$, and to define

$$\widehat{IC}_p(x) = \sum_{j=1}^{p} \hat{\lambda}_j^{-1}\, \langle x, \hat{v}_j\rangle\, \hat{v}_j.$$

The operator $\widehat{IC}_p$ is defined on the whole of L^2, and it is bounded if $\hat{\lambda}_j > 0$ for $j \le p$. By judiciously choosing p we find a balance between retaining the relevant information in the sample, and the danger of working with the reciprocals of small eigenvalues $\hat{\lambda}_j$. To derive

a computable estimator of Ψ, we use an empirical version of (3.4). Since C_1 is estimated by

$$\widehat{C}_1(x) = \frac{1}{N-1}\sum_{k=1}^{N-1} \langle X_k, x\rangle\, X_{k+1},$$

we obtain, for any $x \in H$,

$$\begin{aligned}\widehat{C}_1\widehat{IC}_p(x) &= \widehat{C}_1\left(\sum_{j=1}^{p} \hat{\lambda}_j^{-1}\langle x, \hat{v}_j\rangle\, \hat{v}_j\right)\\ &= \frac{1}{N-1}\sum_{k=1}^{N-1}\left\langle X_k, \sum_{j=1}^{p}\hat{\lambda}_j^{-1}\langle x,\hat{v}_j\rangle\,\hat{v}_j\right\rangle X_{k+1}\\ &= \frac{1}{N-1}\sum_{k=1}^{N-1}\sum_{j=1}^{p}\hat{\lambda}_j^{-1}\langle x,\hat{v}_j\rangle\langle X_k,\hat{v}_j\rangle\, X_{k+1}.\end{aligned}$$

The estimator $\widehat{C}_1\widehat{IC}_p$ can be used in principle, but typically an additional smoothing step is introduced by using the approximation $X_{k+1} \approx \sum_{i=1}^{p}\langle X_{k+1},\hat{v}_i\rangle\,\hat{v}_i$. This leads to the estimator

$$\widehat{\Psi}_p(x) = \frac{1}{N-1}\sum_{k=1}^{N-1}\sum_{j=1}^{p}\sum_{i=1}^{p}\hat{\lambda}_j^{-1}\langle x,\hat{v}_j\rangle\langle X_k,\hat{v}_j\rangle\langle X_{k+1},\hat{v}_i\rangle\,\hat{v}_i. \tag{3.5}$$

To establish the consistency of this estimator, it must be assumed that $p = p_N$ is a function of the sample size N. Theorem 8.7 of Bosq (2000) then establishes sufficient conditions for $\|\widehat{\Psi}_p - \Psi\|$ to tend to zero. They are technical, but, intuitively, they mean that the λ_j and the distances between them cannot tend to zero too fast.

If $H = L^2$, the estimator (3.5) is a kernel operator with the kernel

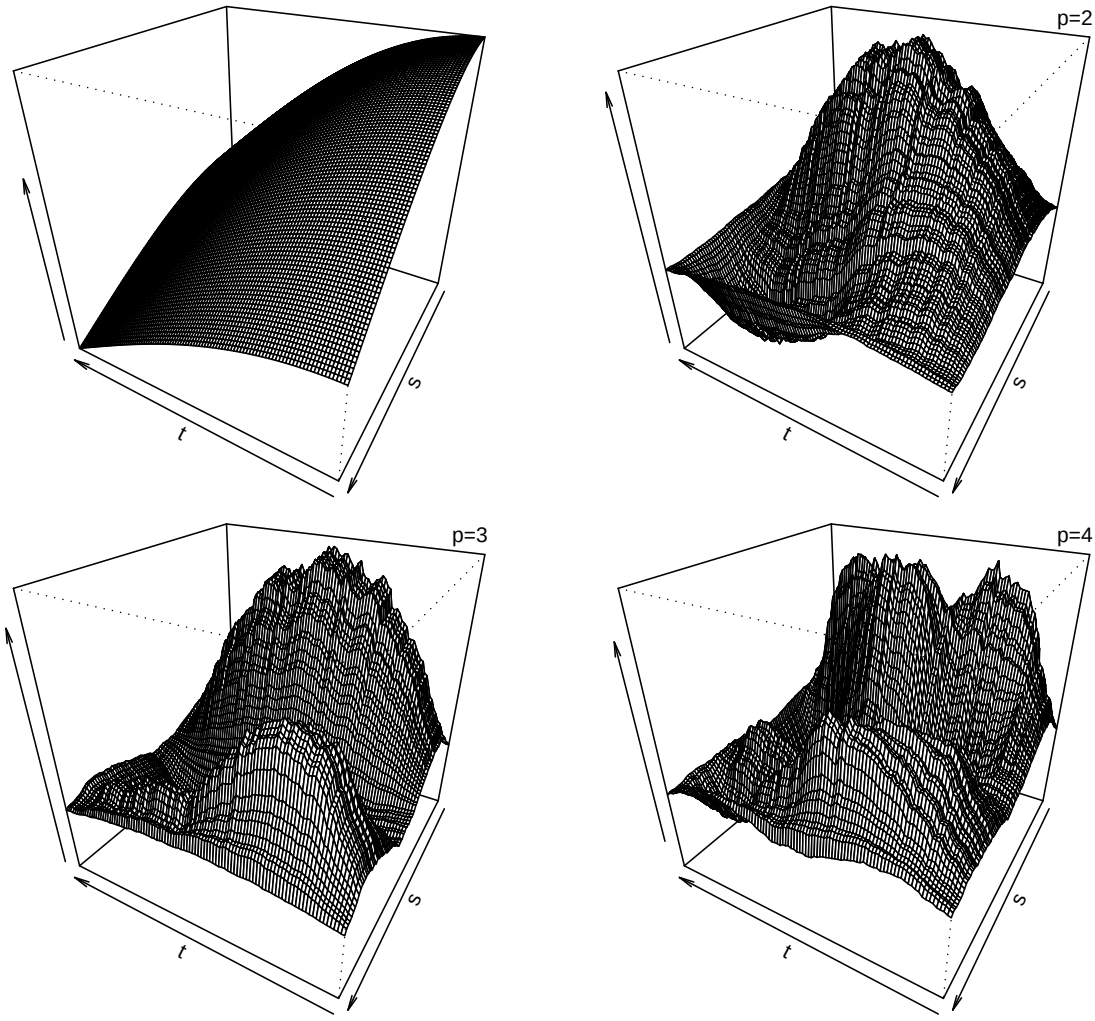
$$\hat{\psi}_p(t,s) = \frac{1}{N-1}\sum_{k=1}^{N-1}\sum_{j=1}^{p}\sum_{i=1}^{p}\hat{\lambda}_j^{-1}\langle X_k,\hat{v}_j\rangle\langle X_{k+1},\hat{v}_i\rangle\,\hat{v}_j(s)\hat{v}_i(t). \tag{3.6}$$

This is verified by noting that

$$\widehat{\Psi}_p(x)(t) = \int \hat{\psi}_p(t,s)x(s)ds.$$

All quantities at the right–hand side of (3.6) are available as output of the `R` function `pca.fd`, so this estimator is very easy to compute. Kokoszka and Zhang (2010) conducted a number of numerical experiments to determine how close the estimated surface $\hat{\psi}_p(t,s)$ is to the surface $\psi(t,s)$ used to simulate an FAR(1) process. Broadly speaking, for $N \le 100$, the discrepancies are very large, both in magnitude and in shape. This is illustrated in Figure 3.1, which shows the Gaussian kernel $\psi(t,s) = \alpha\exp\{-(t^2+s^2)/2\}$, with α chosen so that the Hilbert–Schmidt norm of ψ is $1/2$, and three estimates which use $p = 2, 3, 4$.

FIGURE 3.1 The kernel surface $\psi(t,s)$ (top left) and its estimates $\hat{\psi}_p(t,s)$ for $p = 2, 3, 4$.



The innovations ε_n were generated as Brownian bridges. Such discrepancies are observed for other kernels and other innovation processes as well. Moreover, by any reasonable measure of a distance between two surfaces, the distance between ψ and $\hat{\psi}_p$ increases as p increases. This is counterintuitive because by using more EFPC's $\hat{v}_j$, we would expect the approximation (3.6) to improve. For the FAR(1) used to produce Figure 3.1, the sums $\sum_{j=1}^{p} \hat{\lambda}_j$ explain, respectively, 74, 83 and 87 percent of the variance for $p = 2, 3$ and 4, but (for the series length $N = 100$), the absolute deviation distances between ψ and $\hat{\psi}_p$ are $0.40, 0.44$ and 0.55. The same pattern is observed for the RMSE distance $\|\hat{\psi} - \psi\|_{\mathcal{S}}$ and the relative absolute distance. As N increases, these distances decrease, but their tendency to increase with p remains. This problem is partially due to the fact that for many FAR(1) models, the estimated eigenvalues $\hat{\lambda}_j$ are very small, except $\hat{\lambda}_1$ and $\hat{\lambda}_2$, and so a small error in their estimation translates to a large error in the reciprocals $\hat{\lambda}_j^{-1}$ appearing in (3.6). Kokoszka and Zhang (2010) show that this problem can be alleviated to some extent by adding a positive baseline to the $\hat{\lambda}_j$. However, as we will see in Section 3.3, precise estimation of the kernel ψ is not necessary to obtain satisfactory predictions.

3.3 Prediction

In this section, we discuss some properties of forecasts with the FAR(1) model. Besse *et al.* (2000) apply several prediction methods, including traditional (nonfunctional) methods, to functional time series derived from real geophysical data. Their conclusion is that the method which we call below *Estimated Kernel* performs best. A different approach to prediction of functional data was proposed by Antoniadis *et al.* (2006). In this section, we mostly report the findings of **?**, whose simulation study includes a new method proposed by Kargin and Onatski (2008), which we call below *Predictive Factors*, and which seeks to replace the FPC's by directions which are most relevant for predictions.

We begin by describing the prediction methods we compare. This is followed by the discussion of their finite sample properties.

Estimated Kernel (EK). This method uses estimator (3.6). The predictions are calculated as

$$\hat{X}_{n+1}(t) = \int \hat{\psi}_p(t,s) X_n(s) ds = \sum_{i=1}^{p} \left(\sum_{j=1}^{p} \hat{\psi}_{ij} \langle X_n, \hat{v}_j \rangle \right) \hat{v}_i(t), \tag{3.7}$$

where

$$\hat{\psi}_{ij} = \hat{\lambda}_j^{-1} (N-1)^{-1} \sum_{n=1}^{N-1} \langle X_n, \hat{v}_j \rangle \langle X_{n+1}, \hat{v}_i \rangle. \tag{3.8}$$

There are several variants of this method which depend on where and what kind of smoothing is applied. In our implementation, all curves are converted to functional objects in `R` using 99 Fourier basis functions. The same minimal smoothing is used for the Predictive Factors method.

Predictive Factors (PF). Estimator (3.6) is not directly justified by the problem of prediction, it is based on FPC's, which may focus on the features of the data that are not relevant to prediction. An approach known as predictive factors may (potentially) be better suited for forecasting. It finds directions most relevant to prediction, rather than explaining the variability, as the FPC's do. Roughly speaking, it focuses on the optimal expansion of $\Psi(X_n)$, which is, theoretically, the best predictor of X_{n+1}, rather than the optimal expansion of X_n. Since Ψ is unknown, Kargin and Onatski (2008) developed a way of approximating such an expansion in finite samples. We describe only the general idea, as theoretical arguments developed by Kargin and Onatski (2008) are quite complex. As we will see, the PF method does not offer an advantage in finite samples, so we do not need to describe here all the details.

Denote by $\mathcal{R}_k$ the set of all rank k operators i.e. those operators which map L^2 into a subspace of dimension k. The goal is to find $A \in \mathcal{R}_k$ which minimizes $E||X_{n+1} - A(X_n)||^2$. To find a computable approximation to the operator A, a parameter $\alpha > 0$ must be introduced. Following the recommendation of Kargin and Onatski (2008), we used $\alpha = 0.75$. The prediction is computed as

$$\hat{X}_{n+1} = \sum_{i=1}^{k} \left\langle X_n, \hat{b}_{\alpha,i} \right\rangle \hat{C}_1(\hat{b}_{\alpha,i}),$$

where

$$\hat{b}_{\alpha,i} = \sum_{j=1}^{p} \hat{\lambda}_j^{-1/2} \langle \hat{x}_{\alpha,i}, \hat{v}_j \rangle \, \hat{v}_j + \alpha \hat{x}_{\alpha,i}.$$

The vectors $\hat{x}_{\alpha,i}$ are linear combinations of the EFPC $\hat{v}_i$, $1 \leq i \leq k$, and are approximations to the eigenfunctions of the operator Φ defined by the polar decomposition $\Psi C^{1/2} = U\Phi^{1/2}$, where C is the covariance operator of X_1 and U is a unitary operator. The operator $\hat{C}_1$ is the lag–1 autocovariance operator defined by

$$\hat{C}_1(x) = \frac{1}{N-1} \sum_{i=1}^{N-1} \langle X_i, x \rangle \, X_{i+1}, \quad x \in L^2.$$

The method depends on a selection of p and k. We selected p by the cumulative variance method and set $k = p$.

We selected five prediction methods for comparison, two of which do not use the autoregressive structure. To obtain further insights, we also included the errors obtained by assuming perfect knowledge of the operator Ψ. For ease of reference, we now describe these methods, and introduce some convenient notation.

MP (Mean Prediction) We set $\hat{X}_{n+1}(t) = 0$. Since the simulated curves have mean zero at every t, this corresponds to using the mean function as a predictor. This predictor is optimal if the data are uncorrelated.

NP (Naive Prediction) We set $\hat{X}_{n+1} = X_n$. This method does not attempt to model temporal dependence. It is included to see how much can be gained by utilizing the autoregressive structure of the data.

EX (Exact) We set $\hat{X}_{n+1} = \Psi(X_n)$. This is not really a prediction method because the autoregressive operator Ψ is unknown. It is included to see if poor predictions might be due to poor estimation of Ψ (cf. Section 3.2).

EK (Estimated Kernel) This method is described above.

EKI (Estimated Kernel Improved) This is method EK, but the $\hat{\lambda}_i$ in (3.8) are replaced by $\hat{\lambda}_i + \hat{b}$, as described in Section 3.2.

PF (Predictive Factors) This method is described above.

? studied the errors E_n and R_n, $N - 50 < n < N$, defined by

$$E_n = \sqrt{\int_0^1 \left(X_n(t) - \hat{X}_n(t)\right)^2 dt} \quad \text{and} \quad R_n = \int_0^1 \left|X_n(t) - \hat{X}_n(t)\right| dt.$$

for $N = 50, 100, 200$, and $||\Psi||_{\mathcal{S}} = 0.5, 0.8$. They considered several kernels and innovation processes, including smooth errors obtained as sum of two trigonometric function, irregular errors generated as Brownian bridges, and intermediate errors obtained by adding small multiples of Brownian bridges to smooth innovations. Examples of boxplots are shown in Figures 3.2 and 3.3. In addition to boxplots, **?** reported the averages of the E_n and R_n, $N - 50 < n < N$, and the standard errors of these average, which allow to assess if the differences in the performance of the predictors are statistically significant. Their conclusions can be summarized as follows:

1. Taking the autoregressive structure into account reduces prediction errors, but, in some settings, this reduction is not statistically significant relative to method MP, especially if $\|\Psi\| = 0.5$. Generally if $\|\Psi\| = 0.8$, using the autoregressive structure significantly and visibly improves the predictions.

2. None of the Methods EX, EK, EKI uniformly dominates the other. In most cases method EK is the best, or at least as good at the others.

3. In some cases, method PF performs visibly worse than the other methods, but always better than NP.

4. Using the improved estimation described in Section 3.2 does not generally reduce prediction errors.

? also applied all prediction methods to mean corrected precipitation data studied in Besse *et al.* (2000). For this data set, the averages of the E_n and the R_n are not significantly different between the first five methods, method PF performs significantly

FIGURE 3.2 Boxplots of the prediction errors E_n (left) and R_n (right); Brownian bridge innovations, $\psi(t,s) = Ct$, $N = 100$, $p = 3$, $||\Psi|| = 0.5$.

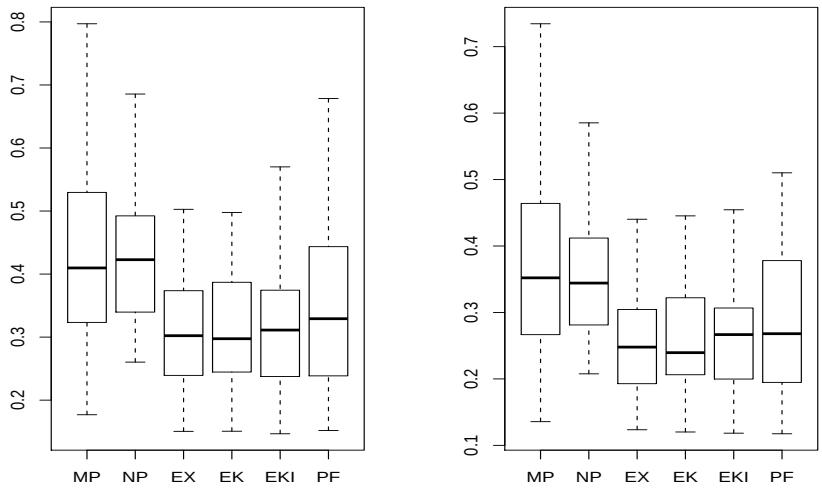
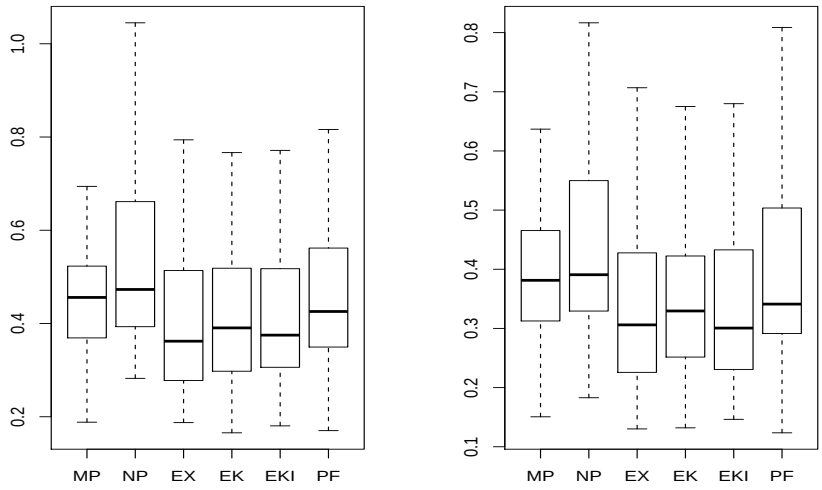


FIGURE 3.3 Boxplots of the prediction errors E_n (left) and R_n (right); Brownian bridge innovations, $\psi(t,s) = Ct$, $N = 100$, $p = 3$, $||\Psi|| = 0.8$.



worse than the others. We should point out that method PF depends on the choice of the parameters α and k. It is possible that its performance can be improved by better tuning these parameters. On the other hand, our simulations show that method EK essentially reaches the limit of what is possible, it is comparable to the theoretically perfect method EX. While taking into account the autoregressive structure of the observations does reduce prediction errors, many prediction errors are comparable to those of the trivial MP method. To analyze this observation further, we present in Figure 3.4 six consecutive trajectories of a FAR(1) process with $||\Psi|| = 0.5$, and Brownian bridge innovations, together with EK predictions. Predictions obtained with other nontrivial methods look similar. We see that the predictions look much smoother than the observations, and their range is much smaller. If the innovations ε_n are smooth, the observations are also smooth, but the predicted curves have a visibly smaller range than the observations. The smoothness of the predicted curves follows from representation (3.7), which shows that each predictor is a linear combination of a few EFPC's, which are smooth curves themselves. The smaller range of the the predictors is not peculiar to functional data, but is enhanced in the functional setting. For a mean zero scalar AR(1) process $X_n = \psi X_{n-1} + \varepsilon_n$, we have $\mathrm{Var}(X_n) = \psi^2 \mathrm{Var}(X_{n-1}) + \mathrm{Var}(\varepsilon_n)$, so the variance of the predictor $\hat{\psi} X_{n-1}$ is about ψ^{-2} times smaller than the variance of X_n. In the functional setting, the variance of $\hat{X}_n(t)$ is close to $\mathrm{Var}[\int \psi(t,s) X_n(s) ds]$. If the kernel ψ admits the decomposition $\psi(t,s) = \psi_1(t)\psi_2(s)$, as all the kernels we use do, then

$$\mathrm{Var}\left[\hat{X}_n(t)\right] \approx \psi_1^2(t) \mathrm{Var}\left[\int_0^1 \psi_2(s) X_{n-1}(s) ds\right].$$

If the function ψ_1 is small for some values of $t \in [0,1]$, it will automatically drive down the predictions. If ψ_2 is small for some $s \in [0,1]$, it will reduce the integral $\int_0^1 \psi_2(s) X_{n-1}(s) ds$. The estimated kernels do not admit a factorization of this type, but are always weighted sums of products of orthonormal functions (the EFPC's $\hat{v}_k$). A conclusion of this discussion is that the predicted curves will in general look smoother and "smaller" than the data. This somewhat disappointing performance is however not due to poor prediction methods, but to a natural limit of predictive power of the FAR(1) model; the curves $\Psi(X_n)$ share the general properties of the curves $\hat{\Psi}(X_n)$, no matter how Ψ is estimated.

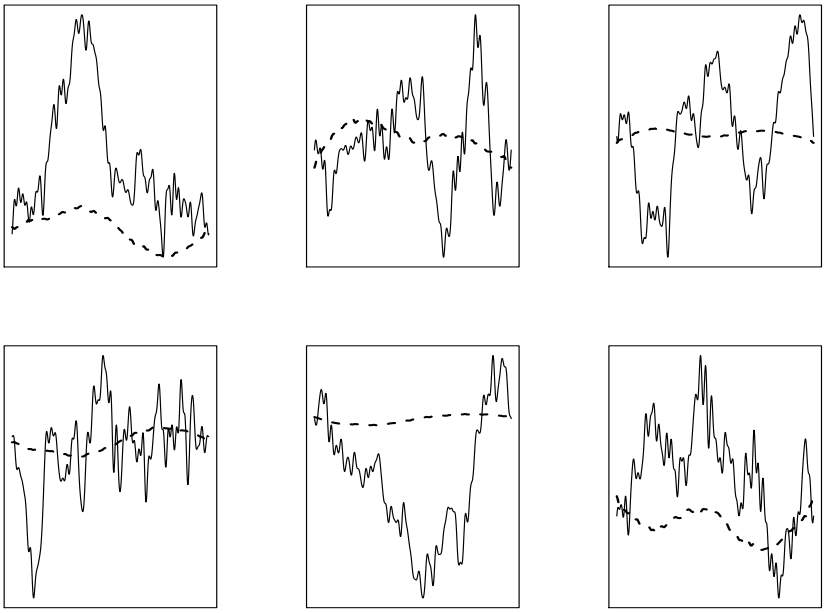
4 Weakly dependent functional time series

What distinguishes time series analysis from other fields of statistics is attention to temporal dependence of the data. In this section, we describe a general framework that accommodates the temporal dependence of functional time series, and illustrate it with several examples.

4.1 Approximable functional sequences

The notion of weak dependence has been formalized in many ways. Perhaps the most popular are various mixing conditions, see Doukhan (1994), Bradley (2007), but in re-

FIGURE 3.4 Six consecutive trajectories of the FAR(1) process with Gaussian kernel, $||\Psi|| = 0.5$, and Brownian bridge innovations. Dashed lines show EK predictions with $p = 3$.



cent years several other approaches have also been introduced, see Doukhan and Louhichi (1999) and Wu (2005, 2007), among others. In time series analysis, moment based measures of dependence, most notably autocorrelations and cumulants, have gained wide acceptance. The measure we consider below is a moment type quantity, but it is also related to the mixing conditions as it considers σ–algebras m time units apart, with m tending to infinity. A most direct relaxation of independence is the m–dependence. Suppose $\{X_n\}$ is a sequence of random elements taking values in a measurable space S. Denote by $\mathcal{F}_k^- = \sigma\{\ldots X_{k-2}, X_{k-1}, X_k\}$ and $\mathcal{F}_k^+ = \sigma\{X_k, X_{k+1}, X_{k+2}, \ldots\}$, the σ–algebras generated by the observations up to time k and after time k, respectively. Then the sequence $\{X_n\}$ is said to be m-dependent if for any k, the σ–algebras $\mathcal{F}_k^-$ and $\mathcal{F}_{k+m}^+$ are independent. Most time series models are not m–dependent. Rather, various measures of dependence decay sufficiently fast, as the distance m between the σ–algebras $\mathcal{F}_k^-$ and $\mathcal{F}_{k+m}^+$ increases. However, m–dependence can be used as a tool to study properties of many nonlinear sequences, see e.g. Hörmann (2008) and Berkes *et al.* (2010) for recent applications. The general idea is to approximate $\{X_n,\, n \in \mathbb{Z}\}$ by m–dependent processes $\{X_n^{(m)},\, n \in \mathbb{Z}\}$, $m \geq 1$. The goal is to establish that for every n the sequence $\{X_n^{(m)},\, m \geq 1\}$ converges in some sense to X_n, if we let $m \to \infty$. If the convergence is fast enough, then one can obtain the limiting behavior of the original process from corresponding results for m–dependent sequences. Definition 4.1 formalizes this idea and sets up the necessary framework for the construction of such m–dependent approximation sequences. The idea of approximating scalar sequences by m–dependent nonlinear moving averages appears

already in Section 21 of Billingsley (1968), and it was developed in several direction by Pötscher and Prucha (1997). A version of Definition 4.1 for vector valued processes was used in Aue *et al.* (2009).

For $X \in L_H^p$ we define

$$\nu_p(X) = \big(E\|X\|^p\big)^{1/p} < \infty. \tag{4.1}$$

DEFINITION 4.1 A sequence $\{X_n\} \in L_H^p$ is called *L^p–m–approximable* if each X_n admits the representation

$$X_n = f(\varepsilon_n, \varepsilon_{n-1}, \ldots), \tag{4.2}$$

where the ε_i are iid elements taking values in a measurable space S, and f is a measurable function $f : S^\infty \to H$. Moreover we assume that if $\{\varepsilon_i'\}$ is an independent copy of $\{\varepsilon_i\}$ defined on the same probability space, then letting

$$X_n^{(m)} = f(\varepsilon_n, \varepsilon_{n-1}, \ldots, \varepsilon_{n-m+1}, \varepsilon_{n-m}', \varepsilon_{n-m-1}', \ldots), \tag{4.3}$$

we have

$$\sum_{m=1}^{\infty} \nu_p\big(X_m - X_m^{(m)}\big) < \infty. \tag{4.4}$$

The applicability of Definition 4.1 was demonstrated in Hörmann and Kokoszka (2010) for several linear and non-linear functional time series. The variables ε_n are typically model errors. The general idea is that in the nonlinear moving average representation (4.2), the impact on X_n of the ε_{n-m} becomes so small as $m \to \infty$ that they can be replaced by different errors. We illustrate it for the FAR(1) model.

EXAMPLE 4.1 (Functional autoregressive process) Let $\{X_n, n \in \mathbb{Z}\}$ be a functional AR(1) model as given as in(3.1), with $\|\Psi\| < 1$. As we have obtained in Theorem 3.1 the AR(1) sequence admits the expansion $X_n = \sum_{j=0}^{\infty} \Psi^j(\varepsilon_{n-j})$, where Ψ^j is the j-th iterate of the operator Ψ. We thus set $X_n^{(m)} = \sum_{j=0}^{m-1} \Psi^j(\varepsilon_{n-j}) + \sum_{j=m}^{\infty} \Psi^j(\varepsilon_{n-j}')$. It is easy to verify that for every A in $\mathcal{L}$, $\nu_p(A(Y)) \le \|A\|\,\nu_p(Y)$. Since $X_m - X_m^{(m)} = \sum_{j=m}^{\infty} \big(\Psi^j(\varepsilon_{m-j}) - \Psi^j(\varepsilon_{m-j}')\big)$, it follows that $\nu_p(X_m - X_m^{(m)}) \le 2\sum_{j=m}^{\infty} \|\Psi\|^j\ \nu_p(\varepsilon_0) = O(1)\nu_p(\varepsilon_0)\|\Psi\|^m$. By assumption $\nu_2(\varepsilon_0) < \infty$ and therefore $\sum_{m=1}^{\infty} \nu_2(X_m - X_m^{(m)}) < \infty$, so condition (4.4) holds with $p \ge 2$, as long as $\nu_p(\varepsilon_0) < \infty$.

4.2 Estimation of the mean function and the FPC's

With the notion of weak-dependence just defined at hand, we can obtain analogs of Theorems 2.1 and 2.2 for time series. We include the proofs because they illustrate the ease with which the condition of m–approximability is applied. We let $\hat{\mu}_N$ and $\hat{C}_N$ be defined as in Section 2.3.

THEOREM 4.1 *Assume that $\{X_k\}$ is an H-valued L^2–m–approximable process with $EX = \mu$. Then $E\|\hat{\mu}_N - \mu\|^2 = O(N^{-1})$.*

PROOF: Observe that for any $h > 0$ we have

$$X_0 = f(\varepsilon_0, \varepsilon_{-1}, \ldots), \qquad X_h^{(h)} = f^{(h)}(\varepsilon_h, \varepsilon_{h-1}, \ldots, \varepsilon_1, \varepsilon_0^{(h)}, \varepsilon_{-1}^{(h)}, \ldots),$$

and thus the random variables X_0 and $X_h^{(h)}$ are independent. Stationarity of $\{X_k\}$, independence of X_0 and $X_h^{(h)}$ and the Cauchy-Schwarz inequality yield that

$$\begin{aligned} NE\|\hat{\mu}_N - \mu\|^2 &= \sum_{h=-(N-1)}^{N-1} \left(1 - \frac{|h|}{N}\right) E\langle X_0 - \mu, X_h - \mu\rangle \\ &\le \sum_{h\in\mathbb{Z}} |E\langle X_0 - \mu, X_h - \mu\rangle| \\ &\le E\|X_0 - \mu\|^2 + 2\sum_{h\ge 1} |E\langle X_0 - \mu, X_h - X_h^{(h)}\rangle| \\ &\le \nu_2(X_0 - \mu) \times \left(\nu_2(X_0 - \mu) + 2\sum_{h\ge 1} \nu_2\left(X_h - X_h^{(h)}\right)\right) < \infty. \end{aligned}$$

THEOREM 4.2 *Suppose $\{X_n\} \in L_H^4$ is an L^4–m–approximable sequence with covariance operator C. Then there is some constant $U_X < \infty$, which does not depend on N, such that*

$$E\|\hat{C} - C\|_{\mathcal{S}}^2 \le U_X\, N^{-1}. \tag{4.5}$$

Before we give the proof we state the following important result that follows immediately from Theorem 4.2 and from Lemmas 2.1 and 2.2.

THEOREM 4.3 *Suppose $\{X_n,\, n \in \mathbb{Z}\} \in L_H^4$ is an L^4–m–approximable sequence and assumption (2.10) holds. Then, for $1 \le j \le d$,*

$$E\left[|\lambda_j - \hat{\lambda}_j|^2\right] = O(N^{-1}) \quad \textit{and} \quad E\left[\|\hat{c}_j\hat{v}_j - v_j\|^2\right] = O(N^{-1}). \tag{4.6}$$

Theorems 4.1–4.3 show that the standard estimates for the functional mean and the FPC's employed for iid data are robust to a sufficiently weak violation of the independence assumption.

PROOF OF THEOREM 4.2: We assume for simplicity that $EX = 0$. For $k \in \mathbb{Z}$ define the operators $B_k(y) = \langle X_k, y\rangle\, X_k - C(y)$, $y \in H$. Then since B_k are stationary, we have

$$\begin{aligned} E\|\hat{C}_N - C\|_{\mathcal{S}}^2 &= E\left\|\frac{1}{N}\sum_{k=1}^{N} B_k\right\|_{\mathcal{S}}^2 \\ &= \frac{1}{N}\sum_{k=-(N-1)}^{N-1} \left(1 - \frac{|k|}{N}\right) E\langle B_0, B_k\rangle_{\mathcal{S}} \\ &\le \frac{1}{N}\left(E\|B_0\|_{\mathcal{S}}^2 + 2\sum_{k\ge 1} |E\langle B_0, B_k\rangle_{\mathcal{S}}|\right), \end{aligned}$$

and it remains to show that $|E\langle B_0, B_k\rangle_{\mathcal{S}}|$ decays sufficiently fast. We let $\lambda_1 \geq \lambda_2 \geq \cdots$ be the eigenvalues of the operator C and we let $\{e_i\}$ be the corresponding eigenfunctions. It can be readily verified that

$$E\left\langle B_0, B_k\right\rangle_{\mathcal{S}} = E\left\langle X_0, X_k\right\rangle^2 - \sum_{j\geq 1}\lambda_j^2, \quad k \geq 1.$$

Furthermore using the independence of X_0 and $X_k^{(k)}$ we have

$$E\left\langle X_0, X_k^{(k)}\right\rangle^2 = \sum_{j\geq 1}\lambda_j^2, \quad k \geq 1,$$

showing that

$$E\left\langle B_0, B_k\right\rangle_{\mathcal{S}} = E\left\langle X_0, X_k\right\rangle^2 - E\left\langle X_0, X_k^{(k)}\right\rangle^2. \tag{4.7}$$

For ease of notation we set $X_k' = X_k^{(k)}$. Then we have

$$\begin{aligned}\left\langle X_0, X_k - X_k'\right\rangle^2 &= \left\langle X_0, X_k\right\rangle^2 + \left\langle X_0, X_k'\right\rangle^2 - 2\left\langle X_0, X_k\right\rangle\left\langle X_0, X_k'\right\rangle \\ &= \left\langle X_0, X_k\right\rangle^2 - \left\langle X_0, X_k'\right\rangle^2 - 2\left\langle X_0, X_k - X_k'\right\rangle\left\langle X_0, X_k'\right\rangle.\end{aligned}$$

Thus

$$\left\langle X_0, X_k\right\rangle^2 - \left\langle X_0, X_k'\right\rangle^2 = \left\langle X_0, X_k - X_k'\right\rangle^2 + 2\left\langle X_0, X_k - X_k'\right\rangle\left\langle X_0, X_k'\right\rangle$$

and by repeated application of Cauchy-Schwarz it follows that

$$\left|E\left\langle X_0, X_k\right\rangle^2 - E\left\langle X_0, X_k'\right\rangle^2\right| \leq \nu_4^2(X_0)\nu_4^2\left(X_k - X_k'\right) + 2\nu_4^2(X_0)\nu_2(X_0)\nu_2\left(X_k - X_k'\right). \tag{4.8}$$

Combining (4.7) and (4.8) and using the Definition of L^4–m–approximability yields the proof of our theorem, with U_X equal to the sum over $k \geq 1$ of the right hand side of (4.8).

4.3 Estimation of the long run variance

The main results of this section are Corollary 4.1 and Proposition 4.1 which state that the long run variance matrix obtained by projecting the data on the functional principal components can be consistently estimated. We start with some preliminaries which lead to the main results. For illustration, we present the proof of Lemma 4.1. More details can be found in Hörmann and Kokoszka (2010).

Let $\{X_n\}$ be a scalar (weakly) stationary sequence. Its long run variance is defined as $\sigma^2 = \sum_{j\in\mathbb{Z}}\gamma_j$, where $\gamma_j = \mathrm{Cov}(X_0, X_j)$, provided this series is absolutely convergent. Our first lemma shows that this is the case for L^2–m–approximable sequences.

LEMMA 4.1 *Suppose $\{X_n\}$ is a scalar L^2–m–approximable sequence. Then its autocovariance function $\gamma_j = \mathrm{Cov}(X_0, X_j)$ is absolutely summable, i.e. $\sum_{j=-\infty}^{\infty}|\gamma_j| < \infty$.*

PROOF: As we have noted in the proof of Theorem 4.1, X_0 and $X_j^{(j)}$ are independent and thus $\mathrm{Cov}(X_0, X_j^{(j)}) = 0$. It follows that $|\gamma_j| \leq [EX_0^2]^{1/2}[E(X_j - X_j^{(j)})^2]^{1/2}$, which proves the lemma.

The summability of the autocovariances is the fundamental property of weak dependence because then $N\mathrm{Var}[\bar{X}_N] \to \sum_{j=-\infty}^{\infty} \gamma_j$, i.e. the variance of the sample mean converges to zero at the rate N^{-1}, the same as for iid observations. A popular approach to the estimation of the long-run variance is to use the kernel estimator

$$\hat{\sigma}^2 = \sum_{|j|\leq q} \omega_q(j)\hat{\gamma}_j, \qquad \hat{\gamma}_j = \frac{1}{N}\sum_{i=1}^{N-|j|}(X_i - \bar{X}_N)(X_{i+|j|} - \bar{X}_N).$$

Various weights $\omega_q(j)$ have been proposed and their optimality properties studied, see Andrews (1991) and Anderson (1994), among others. In theoretical work, it is typically assumed that the bandwith q is a deterministic function of the sample size such that $q = q(N) \to \infty$ and $q = o(N^r)$.

We consider the vector case in which the data are of the form

$$\mathbf{X}_n = [X_{1n}, X_{2n}, \ldots, X_{dn}]^T, \qquad n = 1, 2, \ldots, N.$$

The estimation of the mean by the sample mean does not effect the limit of the kernel long–run variance estimators, so *we assume that* $EX_{in} = 0$ and define the autocovariances as

$$\gamma_r(i,j) = E[X_{i0}X_{jr}], \qquad 1 \leq i, j \leq d.$$

If $r \geq 0$, $\gamma_r(i,j)$ is estimated by $N^{-1}\sum_{n=1}^{N-r} X_{in}X_{j,n+r}$ but if $r < 0$ it is estimated by $N^{-1}\sum_{n=1}^{N-|r|} X_{i,n+|r|}X_{j,n}$. The autocovariance matrices are thus

$$\hat{\mathbf{\Gamma}}_r = \begin{cases} N^{-1}\displaystyle\sum_{n=1}^{N-r} \mathbf{X}_n\mathbf{X}_{n+r}^T & \text{if} \quad r \geq 0, \\ N^{-1}\displaystyle\sum_{n=1}^{N-|r|} \mathbf{X}_{n+|r|}\mathbf{X}_n^T & \text{if} \quad r < 0. \end{cases}$$

The variance $\mathrm{Var}[N^{-1}\bar{\mathbf{X}}_N]$ has (i,j)–entry

$$N^{-2}\sum_{m,n=1}^{N} E[X_{im}X_{jn}] = N^{-1}\sum_{|r|<N}\left(1 - \frac{|r|}{N}\right)\gamma_r(i,j),$$

so the long–run variance is

$$\mathbf{\Sigma} = \sum_{r=-\infty}^{\infty} \mathbf{\Gamma}_r, \qquad \mathbf{\Gamma}_r := [\gamma_r(i,j),\ 1 \leq i, j \leq d],$$

and its kernel estimator is

$$\hat{\mathbf{\Sigma}} = \sum_{|r|\leq q} \omega_q(r)\hat{\mathbf{\Gamma}}_r. \tag{4.9}$$

We consider the weights $\omega_q(j) = K(j/q)$, where K is a kernel satisfying the following assumption.

Assumption 4.1

(*i*) $K(0) = 1$

(*ii*) *K is a symmetric, Lipschitz function*

(*iii*) *K has a bounded support*

(*iv*) *$\hat{K}$, the Fourier transform of K, is also Lipschitz and integrable*

The following theorem is proven in Horváth and Kokoszka (2011).

Theorem 4.4 *Suppose $\{\mathbf{X}_n\}$ is an L^2–m–approximable sequence. If Assumption 4.1 holds and $q \to \infty$, $q/N \to 0$, then $\hat{\mathbf{\Sigma}}_N \xrightarrow{P} \mathbf{\Sigma}$.*

In contrast to many classical results, see e.g. Newey and West (1987), Theorem 4.4 does not impose fourth order conditions, and replaces mixing or linearity conditions by L^2–m–approximability. Assumption 4.1 is standard, except its condition (iv). The following example shows that it holds for the Bartlett kernel; the Fourier transforms of other commonly used kernels are smoother and decay faster.

Example 4.2 The Bartlett kernel is

$$K(s) = \begin{cases} 1 - |s|, & |s| \leq 1, \\ 0, & \text{otherwise} \end{cases}$$

This kernel clearly satisfies parts (i)–(iii) of Assumption 4.1. Its Fourier transform is

$$\hat{H}(u) = \left\{ \frac{1}{\pi u} \sin\left(\frac{u}{2}\right) \right\}^2.$$

Thus, to verify part (iv), we must check that the function

$$F(t) = \left\{ \frac{\sin(t)}{t} \right\}^2$$

is integrable and Lipschitz. The integrability follows because $|F(t)| \leq t^{-2}$ and $F(t) \to 1$, as $t \to 0$.

The derivative of F for $t \neq 0$ is

$$F'(t) = \frac{2\sin(t)}{t} \left\{ \frac{t\cos(t) - \sin(t)}{t^2} \right\}.$$

This function is clearly bounded outside any neighborhood of zero. Using the Taylor expansion of the sine and cosine functions, it is easy to verify that $F'(t) = o(t)$, as $t \to 0$. In a similar fashion, one can verify that $F(t) - F(0) = o(t^2)$, as $t \to 0$. Thus F is Lipschitz on the whole line.

We are now able to turn to functional data. Suppose $\{X_n\} \in L_H^2$ is a zero mean sequence and $e_1, e_2, \ldots, e_d$ is any set of orthonormal functions in H. Define $X_{in} = \langle X_n, e_i \rangle$, $\mathbf{X}_n = [X_{1n}, X_{2n}, \ldots, X_{dn}]^T$ and $\mathbf{\Gamma}_r = \mathrm{Cov}(\mathbf{X}_0, \mathbf{X}_r)$. A direct verification shows that if $\{X_n\}$ is L^p–m–approximable, then so is the vector sequence $\{\mathbf{X}_n\}$. We thus obtain the following corollary.

COROLLARY 4.1 *a) If $\{X_n\} \in L_H^2$ is an L^2–m–approximable sequence, then the series $\sum_{r=-\infty}^{\infty} \mathbf{\Gamma}_r$ converges absolutely. b) If, in addition, Assumption 4.1 holds and $q \to \infty$ with $q = o(N)$, then $\hat{\mathbf{\Sigma}} \xrightarrow{P} \mathbf{\Sigma}$.*

In Corollary 4.1, the functions $e_1, e_2, \ldots, e_d$ form an arbitrary orthonormal deterministic basis. In many applications, a random basis consisting of the EFPC's $\hat{v}_1, \hat{v}_2, \ldots, \hat{v}_d$ is used. The scores with respect to this basis are defined by

$$\hat{\eta}_{\ell i} = \left\langle X_i - \bar{X}_N, \hat{v}_\ell \right\rangle, \quad 1 \le \ell \le d.$$

To use the results established so far, it is convenient to decompose the stationary sequence $\{X_n\}$ into its mean and a zero mean process, i.e. we set $X_n = \mu + Y_n$, where $EY_n = 0$. We introduce the unobservable quantities

$$\beta_{\ell n} = \langle Y_n, v_\ell \rangle, \qquad \hat{\beta}_{\ell n} = \langle Y_n, \hat{v}_\ell \rangle, \quad 1 \le \ell \le d.$$

The following proposition is useful in the development of asymptotic arguments for many statistical procedures for functional time series. The boldface symbols refer to vectors with the coordinates just defined and $\hat{\mathbf{\Sigma}}(\boldsymbol{\delta})$ is the estimator (4.9) calculated from the observation vectors $\boldsymbol{\delta}_1, \ldots, \boldsymbol{\delta}_N$.

PROPOSITION 4.1 *Let $\hat{\mathbf{C}} = \mathrm{diag}(\hat{c}_1, \ldots, \hat{c}_d)$, with $\hat{c}_i = \mathrm{sign}(\langle v_i, \hat{v}_i \rangle)$. Suppose $\{X_n\} \in L_H^4$ is L^4–m–approximable and that (2.10) holds. Assume further that Assumption 4.1 holds with a stronger condition $q^4/N \to 0$. Then*

$$|\hat{\mathbf{\Sigma}}(\boldsymbol{\beta}) - \hat{\mathbf{\Sigma}}(\hat{\mathbf{C}}\hat{\boldsymbol{\beta}})| = o_P(1) \quad \textit{and} \quad |\hat{\mathbf{\Sigma}}(\hat{\boldsymbol{\eta}}) - \hat{\mathbf{\Sigma}}(\hat{\boldsymbol{\beta}})| = o_P(1). \tag{4.10}$$

The point of Proposition 4.1 is that if $\hat{\mathbf{\Sigma}}(\boldsymbol{\beta})$ is consistent under some conditions, e.g. those stated in Theorem 4.4, then so is $\hat{\mathbf{\Sigma}}(\hat{\boldsymbol{\eta}})$. Before presenting the proof of Proposition 4.1, we note that for functional data it is also often useful to consider the long-run covariance kernel which is defined in terms of the kernels $c_h(t, s) = \mathrm{Cov}(X_i(t), X_{i+h}(t))$. Kernel estimators can be defined analogously to (4.9), and their consistency can be established under L^2–m–approximability and additional technical conditions; we refer to Horváth *et al.* (2011).

PROOF OF PROPOSITION 4.1. We only prove the left relation in (4.10). We introduce the constant

$$\kappa := \sup_{q\geq 1} \frac{1}{q} \sum_{j=-q}^{q} w_q(j)$$

which by Assumption 4.1 is finite (and converges to $2\int_{-1}^{1} K(x)dx$). The element in the k-th row and ℓ-th column of $\hat{\mathbf{\Sigma}}(\boldsymbol{\beta}) - \hat{\mathbf{\Sigma}}(\hat{\mathbf{C}}\hat{\boldsymbol{\beta}})$ is given by

$$\begin{aligned} &\sum_{h=0}^{q} \frac{w_q(h)}{N} \sum_{1\leq n\leq N-h} \left(\beta_{kn}\beta_{\ell,n+h} - \hat{c}_k\hat{\beta}_{kn}\hat{c}_\ell\hat{\beta}_{\ell,n+h}\right) \\ &\quad + \sum_{h=1}^{q} \frac{w_q(h)}{N} \sum_{1\leq n\leq N-h} \left(\beta_{k,n+h}\beta_{\ell,n} - \hat{c}_k\hat{\beta}_{k,n+h}\hat{c}_\ell\hat{\beta}_{\ell,n}\right). \end{aligned} \tag{4.11}$$

For reasons of symmetry it suffices to study (4.11), which can be decomposed into

$$\begin{aligned} &\sum_{h=0}^{q} \frac{w_q(h)}{N} \sum_{1\leq n\leq N-h} \beta_{kn}\left(\beta_{\ell,n+h} - \hat{c}_\ell\hat{\beta}_{\ell,n+h}\right) \\ &\quad + \sum_{h=0}^{q} \frac{w_q(h)}{N} \sum_{1\leq n\leq N-h} \hat{c}_\ell\hat{\beta}_{\ell,n+h}\left(\beta_{kn} - \hat{c}_k\hat{\beta}_{kn}\right). \end{aligned} \tag{4.12}$$

As both summands above can be treated similarly, we will only treat (4.12). For any $\varepsilon > 0$ we have

$$\begin{aligned} &P\left(\left|\sum_{h=0}^{q} \frac{w_q(h)}{N} \sum_{1\leq n\leq N-h} \beta_{kn}\left(\beta_{\ell,n+h} - \hat{c}_\ell\hat{\beta}_{\ell,n+h}\right)\right| > \varepsilon\kappa\right) \\ &\quad \leq P\left(\left|\sum_{h=0}^{q} \frac{w_q(h)}{N} \sum_{1\leq n\leq N-h} \beta_{kn}\left(\beta_{\ell,n+h} - \hat{c}_\ell\hat{\beta}_{\ell,n+h}\right)\right| > \frac{\varepsilon}{q}\sum_{h=0}^{q} w_q(h)\right) \\ &\quad \leq \sum_{h=0}^{q} P\left(\frac{1}{N}\left|\sum_{1\leq n\leq N-h} \beta_{kn}\left(\beta_{\ell,n+h} - \hat{c}_\ell\hat{\beta}_{\ell,n+h}\right)\right| > \frac{\varepsilon}{q}\right). \end{aligned} \tag{4.13}$$

In order to show that (4.13) tends to 0 as $N \to \infty$, we introduce a slowly increasing sequence $\alpha_N \to \infty$ such that $q^4\alpha_N/N \to 0$ and we let C_0 such that $N \max_{1\leq\ell\leq d} E\|v_\ell -$

$\hat{c}_\ell \hat{v}_\ell\|^2 \le C_0$. By Cauchy-Schwarz and Markov inequality we have

$$
\begin{aligned}
&P\left(\left|\sum_{1\le n\le N-h} \beta_{kn}\left(\beta_{\ell,n+h} - \hat{c}_\ell\hat{\beta}_{\ell,n+h}\right)\right| > \frac{\varepsilon\, N}{q}\right)\\
&\quad\le P\left(\sum_{n=1}^N \beta_{kn}^2 \sum_{n=1}^N \left(\beta_{\ell n} - \hat{c}_\ell\hat{\beta}_{\ell n}\right)^2 > \frac{\varepsilon^2\, N^2}{q^2}\right)\\
&\quad\le P\left(\frac{1}{N}\sum_{n=1}^N \beta_{kn}^2 > q\alpha_N\right) + P\left(\frac{1}{N}\sum_{n=1}^N \left(\beta_{\ell n} - \hat{c}_\ell\hat{\beta}_{\ell n}\right)^2 > \frac{\varepsilon^2}{q^3\alpha_N}\right)\\
&\quad\le \frac{E\beta_{k1}^2}{q\alpha_N} + P\left(\frac{1}{N}\sum_{n=1}^N \|Y_n\|^2\|v_\ell - \hat{c}_\ell\hat{v}_\ell\|^2 > \frac{\varepsilon^2}{q^3\alpha_N}\right)\\
&\quad\le \frac{E\|Y_1\|^2}{q\alpha_N} + P\left(\frac{1}{N}\sum_{n=1}^N \|Y_n\|^2 > 2E\|Y_1\|^2\right) + P\left(\|v_\ell - \hat{c}_\ell\hat{v}_\ell\|^2 > \frac{\varepsilon^2}{2E\|Y_1\|^2 q^3\alpha_N}\right)\\
&\quad\le \frac{E\|Y_1\|^2}{q\alpha_N} + \frac{\mathrm{Var}\left(\frac{1}{N}\sum_{n=1}^N \|Y_n\|^2\right)}{E^2\|Y_1\|^2} + \frac{2C_0\, E\|Y_1\|^2 q^3\alpha_N}{N\varepsilon^2}.
\end{aligned}
$$

It can be easily shown that for U, V in L_H^4

$$\nu_2\left(\|U\|^2 - \|V\|^2\right) \le \nu_4^2(U-V) + 2\left\{\nu_4(U) + \nu_4(V)\right\}\nu_4(U-V).$$

An immediate consequence is that L^4–m–approximability of $\{Y_n\}$ implies L^2–m–approximability of the scalar sequence $\{\|Y_n\|^2\}$. A basic result for stationary sequences gives

$$\mathrm{Var}\left(\frac{1}{N}\sum_{n=1}^N \|Y_n\|^2\right) \le \frac{1}{N}\sum_{h\in\mathbb{Z}} \left|\mathrm{Cov}\left(\|Y_0\|^2, \|Y_h\|^2\right)\right|,$$

where the by Lemma 4.1 the autocovariances are absolutely summable. Hence the summands in (4.13) are bounded by

$$C_1\left\{\frac{1}{q\alpha_N} + \frac{1}{N} + \frac{q^3\alpha_N}{N\varepsilon^2}\right\},$$

where the constant C_1 depends only on the law of $\{Y_n\}$. The proof of the proposition follows immediately from our assumptions on q and α_N.

5 Further reading

All topics discussed in this survey are presented in detail in Horváth and Kokoszka (2011). Bosq (2000) contains theoretical foundations for most results of Sections 2 and 3. Ramsay and Silverman (2005) provide an introduction to many fundamental concepts of FDA, while Ramsay *et al.* (2009) focus on implementation in `R` and `MATLAB`.

A topic of particular importance in time series analysis is *change point detection*. Most approaches to modeling time series assume that the data follow one model. If the stochastic structure of the data changes at some time point(s), both exploratory and inferential tools produce misleading results. Berkes *et al.* (2009) study the problem of testing for a change in the mean function assuming that the curves are collected over time, but are independent. Hörmann and Kokoszka (2010) extend their procedure to L^4–m–approximable functional time series. Asymptotic distributions of related change point estimators are studied in Aue *et al.* (2009). Horváth *et al.* (2010) develop a test for a change point in the autoregressive operator Ψ in the FAR(1) model. Gabrys *et al.* (2010b) use the framework of FDA to detect changes in the intraday volatility pattern.

A central topic in FDA is the *functional linear model* of the for $Y_n = \Psi(X_n) + \varepsilon_n$. In its most general form, the responses Y_n, the regressors X_n and the errors ε_n are functions, and Ψ is an integral kernel operator. Very extensive research is available under the assumption that the cases (X_n, Y_n) are independent and the errors ε_n are independent. Hörmann and Kokoszka (2010) showed that an estimator for Ψ developed under iid assumption remains consistent if the X_n form an L^4–m–approximable time series. Gabrys *et al.* (2010) developed procedures to test the assumption of iid ε_n against the alternative that the ε_n are correlated. Gabrys and Kokoszka (2007) developed a similar test, which is however applicable to directly observable curves, not to unobservable errors.

Dependence between curves plays a central role also for *spatial* functional data. In this context, we observe curves at many spatial locations, for example the precipitation over many decades at a number of measurement stations. In addition to the dependence between curves, spatial distribution of the locations must also be taken into account to develop informative statistical procedures. **?** develop asymptotic theory for the estimation of the mean function and the FPC's for such data. Gromenko and Kokoszka (2010) propose and compare several estimation procedures which improve on the standard simple mean and the EFPC's defined in Section 2. Other topics, including kriging for functional data, are discussed in Delicado *et al.* (2010).

Acknowledgements. This research was partially supported by the National Science Foundation, by the Banque nationale de Belgique, and by Communauté française de Belgique – Actions de Recherche Concertées (2010–2015).

References

Akhiezier, N. I. and Glazman, I. M. (1993). *Theory of Linear Operators in Hilbert Space*. Dover, New York.

Anderson, T. W. (1994). *The Statistical Analysis of Time Series*. Wiley and Sons.

Andrews, D. W. K. (1991). Heteroskedasticity and autocorrelation consistent covariance matrix estimation. *Econometrica*, **59,** 817–858.

Antoniadis, A., Paparoditis, E. and Sapatinas, T. (2006). A functional wavelet–kernel approach for time series prediction. *Journal of the Royal Statistical Society, Series* B, **68,** 837–857.

Aue, A., Gabrys, R., Horváth, L. and Kokoszka, P. (2009). Estimation of a change–point in the mean function of functional data. *Journal of Multivariate Analysis*, **100,** 2254–2269.

Aue, A., Hörmann, S., Horváth, L. and Reimherr, M. (2009). Break detection in the covariance structure of multivariate time series models. *The Annals of Statistics*, **37,** 4046–4087.

Berkes, I., Gabrys, R., Horváth, L. and Kokoszka, P. (2009). Detecting changes in the mean of functional observations. *Journal of the Royal Statistical Society (B)*, **71,** 927–946.

Berkes, I., Hörmann, S. and Schauer, J. (2010). Split invariance principles for stationary processes. *The Annals of Probability*, **00,** 0000–0000; Forthcoming.

Besse, P., Cardot, H. and Stephenson, D. (2000). Autoregressive forecasting of some functional climatic variations. *Scandinavian Journal of Statistics*, **27,** 673–687.

Billingsley, P. (1968). *Convergence of Probability Measures*. Wiley, New York.

Bosq, D. (2000). *Linear Processes in Function Spaces*. Springer, New York.

Bradley, R. C. (2007). *Introduction to Strong Mixing Conditions*, volume 1,2,3. Kendrick Press.

Brockwell, P. J. and Davis, R. A. (1991). *Time Series: Theory and Methods*. Springer, New York.

Dauxois, J., Pousse, A. and Romain, Y. (1982). Asymptotic theory for principal component analysis of a vector random function. *Journal of Multivariate Analysis*, **12,** 136–154.

Debnath, L. and Mikusinski, P. (2005). *Introduction to Hilbert Spaces with Applications*. Elsevier.

Delicado, P., Giraldo, R., Comas, C. and Mateu, J. (2010). Statistics for spatial functional data: some recent contributions. *Environmetrics*, **21,** 224–239.

Diaconis, P. and Freeman, D. (1999). Iterated random functions. *SIAM Review*, **41,** 45–76.

Doukhan, P. (1994). *Mixing: Properties and Examples*. Lecture Notes in Statistics. Springer.

Doukhan, P. and Louhichi, S. (1999). A new weak dependence and applications to moment inequalities. *Stochastic Processes and their Applications*, **84,** 313–343.

Gabrys, R., Hörmann, S. and Kokoszka, P. (2010b). Monitoring the intraday volatility pattern. Technical Report. Utah State University.

Gabrys, R., Horváth, L. and Kokoszka, P. (2010). Tests for error correlation in the functional linear model. *Journal of the American Statistical Association*, **105,** 1113–1125.

Gabrys, R. and Kokoszka, P. (2007). Portmanteau test of independence for functional observations. *Journal of the American Statistical Association*, **102,** 1338–1348.

Gohberg, I., Golberg, S. and Kaashoek, M. A. (1990). *Classes of Linear Operators*. Operator Theory: Advances and Applications, volume 49. Birkhaüser.

Gromenko, O. and Kokoszka, P. (2010). Estimation of the mean function and the principal components of spatially distributed curves. Technical Report. Utah State University.

Hörmann, S. (2008). Augmented GARCH sequences: Dependence structure and asymptotics. *Bernoulli*, **14,** 543–561.

Hörmann, S. and Kokoszka, P. (2010). Weakly dependent functional data. *The Annals of Statistics*, **38,** 1845–1884.

Horváth, L., Hušková, M. and Kokoszka, P. (2010). Testing the stability of the functional autoregressive process. *Journal of Multivariate Analysis*, **101,** 352–367.

Horváth, L. and Kokoszka, P. (2011). *Inference for Functional Data with Applications*. Springer Series in Statistics. Springer. Forthcoming.

Horváth, L., Kokoszka, P. and Reeder, R. (2011). Estimation of the mean of functional time series and a two sample problem. Technical Report. University of Utah.

Johnstone, I. M. and Lu, A. Y. (2009). On consistency and sparcity for principal components analysis in high dimensions. *Journal of the Americal Statistical Association*, **104,** 682–693.

Kargin, V. and Onatski, A. (2008). Curve forecasting by functional autoregression. *Journal of Multivariate Analysis*, **99,** 2508–2526.

Kokoszka, P. and Zhang, X. (2010). Improved estimation of the kernel of the functional autoregressive process. Technical Report. Utah State University.

Newey, W. K. and West, K. D. (1987). A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica*, **55,** 703–08.

Pötscher, B. and Prucha, I. (1997). *Dynamic Non–linear Econonometric Models. Asymptotic Theory*. Springer.

Ramsay, J., Hooker, G. and Graves, S. (2009). *Functional Data Analysis with R and MATLAB*. Springer.

Ramsay, J. O. and Silverman, B. W. (2005). *Functional Data Analysis*. Springer.

Riesz, F. and Sz.-Nagy, B. (1990). *Functional Analysis*. Dover.

Wu, W. (2005). *Nonlinear System Theory: Another Look at Dependence*. Proceedings of The National Academy of Sciences of the United States, volume 102. National Academy of Sciences.

Wu, W. (2007). Strong invariance principles for dependent random variables. *The Annals of Probability*, **35,** 2294–2320.

Wu, W.B. and Shao, X. (2004). Limit theorems for iterated random functions. *Journal of Applied Probability*, **41,** number 2.