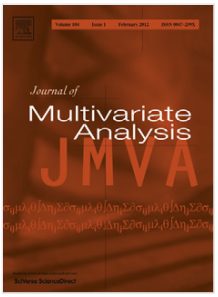
Accepted Manuscript



Conditional estimation for dependent functional data

Heather Battey, Alessio Sancetta

PII: S0047-259X(13)00063-8
DOI: http://dx.doi.org/10.1016/j.jmva.2013.04.009
Reference: YJMVA 3537

To appear in: *Journal of Multivariate Analysis*

Received date: 2 August 2012

Please cite this article as: H. Battey, A. Sancetta, Conditional estimation for dependent functional data, *Journal of Multivariate Analysis* (2013), http://dx.doi.org/10.1016/j.jmva.2013.04.009

Conditional estimation for dependent functional data

Heather Battey[a], Alessio Sancetta[b]

[a] *School of Mathematics, University of Bristol, University Walk, Clifton, BS8 1TW. Email:* h.s.battey@bristol.ac.uk

[b] *Department of Economics, Royal Holloway, University of London, Egham, TW20 0EX. Email:* asancetta@gmail.com

Abstract

Suppose we observe a Markov chain taking values in a functional space. We are interested in exploiting the time series dependence in these infinite dimensional data in order to make non-trivial predictions about the future. Making use of the Karhunen Loève (KL) representation of functional random variables in terms of the eigenfunctions of the covariance operator, we present a deliberately over-simplified nonparametric model, which allows us to achieve dimensionality reduction by considering one dimensional nearest neighbour (NN) estimators for the transition distribution of the random coefficients of the KL expansion. Under regularity conditions, we show that the NN estimator is consistent even when the coefficients of the KL expansion are estimated from the observations. This also allows us to deduce consistency of conditional regression function estimators for functional data. We show via simulations and two empirical examples that the proposed NN estimator outperforms the state of the art when data are generated both by the functional autoregressive (FAR) model of Bosq (2000) and by more general data generating mechanisms.

Keywords: functional data analysis, Karhunen-Loève expansion, dimension reduction, nearest neighbour estimator.

AMS (2000) subject classifications: primary: 62G20, 62M05; secondary: 62H25

1. Introduction

The statistical analysis of functional data has attracted substantial attention over the last fifteen years or so. Aside from the dramatic improvements in data collection technologies, which have allowed data to be collected over a denser collection of points, the increased popularity of functional data analysis has stemmed from its ability to exploit some assumed smoothness in the sample paths of the random process of interest. Much of the early work on functional data analysis (FDA) focussed on i.i.d. functional random variables, but recently there has been heightened interest in dependent functional data. The need to take account of dependence is particularly evident in cases where functional data arise from segmenting a long time series into natural consecutive intervals (e.g. days, weeks, etc.) of equal length, as discussed by Hörmann and Kokoszka (2010) and Bathia et. al. (2010). Electricity load curves, pollutant concentration curves and traffic volumes across the day are just a few examples of time series functional data studied in the literature (e.g. Cavallini et al., 1994, Besse and Cardot, 1996, Besse et al., 2000, Damon and Guillas, 2002). More generally, we are interested in dependent random functions whose domain is a higher dimensional set like $\mathcal{V} \subset \mathbb{R}^k$

($k \geq 1$), which is a case particularly relevant in fields such as brain imaging and geophysics (see e.g. Cohen and Jones, 1969). In these examples and others, it is not appropriate to assume functional data are generated independently of one another.

Much of FDA is based on variants of functional principal components analysis (FPCA) (see e.g. Ramsay and Silverman, 2005), a technique that permits dimensionality reduction by restricting attention to random functions taking values in a separable Hilbert space and decomposing in terms of orthogonal basis functions. More precisely, this basis consists of the eigenfunctions of the covariance operator. The random coefficients of this basis function expansion are termed *factor loadings* or *component scores*, and the principal factor loadings or component scores are those corresponding to the first few selected basis functions of the expansion; more details are provided in subsequent chapters.

Although many studies deal with estimation of principal factors using FPCA (see e.g. Silverman, 1996; Ramsay and Silverman, 2002; Lin and Carroll, 2000; Yao et al, 2005a; 2005b; Hall et al., 2006; Li and Hsing, 2010), there is relatively little research devoted to using the estimated factor loadings for prediction purposes. In this paper, we use the estimated factor loadings to propose an estimator of the transition distribution of functional observations, consistent in some suitable topology. In particular, predictions are based on the estimated factor loadings using nearest neighbour estimators for their transition distributions. Many other nonparametric approaches to functional data analysis are proposed in Ferraty and Vieu (2006).

Mean prediction of functional data has been studied before for the $\mathcal{V} \subset \mathbb{R}$ case; Bosq (2000) provides a comprehensive account of the linear functional autoregressive process of order 1, Bosq and Blanke (2007) provides a general definition of the the functional autoregressive process (see page 243), whilst Kargin and Onatski (2008) develop the predictive factors procedure for one-step prediction under the same assumption. Although the theoretical properties obtained by Kargin and Onatski (2008) can be used to justify their procedure, the recent study of Didericksen, Kokoszka and Zhang (2012) shows that in finite samples, the predictive factors procedure never outperforms the approach of Bosq (2000), and in some cases performs poorly, even when the data are generated artificially as a functional autoregressive process of order 1, which we henceforth refer to as FAR(1). Even more concerning, the same authors found that predictions based on the functional autoregression are often no better than those based on the mean function. These observations motivate a procedure that is able to outperform these procedures and handle more complex dependence structures. In this paper, we restrict attention to stationary ergodic Markov chains (MC), not necessarily linear; see Hörmann and Kokoszka (2010) for general weak dependence conditions on functional data. We propose a deliberately over-simplified nonparametric model based on the transition distributions of the unobservable factor loadings and present an estimator for these transition distributions based on the estimated factor loadings. We show that the consistency properties of this estimator are unaffected by the use of the estimated factor loadings rather than the true unobservable ones. In characterising the dependence structure in terms of the transition distributions, we avoid the need to parameterise the time series dynamics of the functional data. Although the deliberate over-simplification will induce unquantifiable bias if it fails to accurately describe the data generating mechanism, it allows us to significantly reduce the dimensionality of the problem, thereby reducing the estimation error that would otherwise be incurred.

This paper provides several new contributions. From a methodological perspective, we present a new estimator for the prediction of functional time series and functionals thereof. We establish the consistency of the proposed estimator and provide guidance on the rates at which tuning parameters should be allowed to grow or shrink with the sample

size. Our simulation study and real data examples illustrate the comparative gains of our procedure.

This paper proceeds as follows. Section 2 provides background material and established results that are used to motivate the methodology developed in subsequent sections. In Section 3, methodology for the estimation of several objects is elaborated, which ultimately allows us to propose a consistent estimator for the conditional expectation of Lipschitz functionals of our functional random variables. These methods are presented, in the first instance, using a general consistent estimator for the eigenfunctions that form our orthonormal projection subspace. In Section 4 we discuss estimation of the eigenfunctions further, elucidating estimators that satisfy the consistency and orthogonality requirements in the case of $\mathcal{V} \subset \mathbb{R}$ and discussing the conditions under which certain estimators may be preferred to others. The main contributions of this paper appear in the theorems of Section 3. Section 5 discusses in detail the results derived and the conditions imposed in previous sections. Sections 6 and 7 provide, respectively, simulation evidence for the performance of our proposed procedure and two empirical examples using geophysical data and electricity demand data. Finally, Section 8 gives details of the proofs.

2. Background

Suppose that $(Y_i)_{i\in\mathbb{N}}$ is a weakly stationary sequence of random functions, each taking values in real separable Hilbert space $\mathcal{H}$. We suppose that Y_i is mean zero and has jointly continuous covariance function $C_0(u,v) := \mathbb{E}(Y_i(u)Y_i(v))$ for $u, v \in \mathcal{V}$ where $\mathcal{V}$ is a compact subset of $\mathbb{R}^k$ ($k \geq 1$), and by stationarity the covariance function does not depend on i. The mean zero condition is simply equivalent to assuming that the mean function is known. In practice, the mean function can be estimated by e.g. the empirical mean and subtracted off; for independent or weakly dependent functional observations

$$\mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^{n} Y_i - \mathbb{E}Y_i\right\|^2 = O(n^{-1}).$$

Mercer's Theorem (Adler and Taylor, 2007) provides the following convergent series expansion,

$$C_0(u,v) = \sum_{s\in\mathbb{N}} \lambda_s \varphi_s(u)\varphi_s(v), \tag{2.1}$$

where convergence is uniform in u and v by the above continuity. Above, $(\varphi_s(v))_{s\in\mathbb{N}}$ is a collection of orthonormal real valued eigenfunctions of the integral operator with kernel $C_0(u,v)$ such that $\lambda_s \geq \lambda_{s+1} \ldots$ are the corresponding real ordered eigenvalues. By the L_2-separability of $\mathcal{H}$, the process admits the following Karhunen-Loève expansion with equality in L_2

$$Y_i(v) = \sum_{s\in\mathbb{N}} Z_{i,s}\varphi_s(v) \tag{2.2}$$

where, for each i, $(Z_{i,s})_{s\in\mathbb{N}}$ is a sequence of uncorrelated random variables such that $\mathbb{E}Z_{i,s} = 0$ with variance λ_s. Equation (2.2) is readily obtained from Mercer's Theorem with $Z_{i,s} := \int_{\mathcal{V}} Y_i(v)\varphi_s(v)dv = \langle Y_i, \varphi_s\rangle$ (see e.g. Adler and Taylor, 2007).

The goal is to estimate the transition distribution of Y_i, under the condition that $(Y_i)_{i\in\mathbb{N}}$ is a positive recurrent MC. A prototypical example is the following formal generalization of a multivariate stochastic difference equation to the

functional case (see Babillot *et. al.* (1997) for regularity conditions for positive recurrence in the multivariate case),

$$Y_i(v) = \int_{\mathcal{V}} A_i(v,u)\, Y_{i-1}(u)\, du + B_i(v), \tag{2.3}$$

where, using the same notation for an operator and its kernel function, $(A_i)_{i\in\mathbb{N}}$ is a sequence of i.i.d. linear random operators and $(B_i)_{i\in\mathbb{N}}$ are i.i.d random variables with values in $\mathcal{H}$. Bosq (2000) provides details on this class of linear models when $A_i = A$ is a nonrandom linear operator. By allowing A_i to be a random linear operator, we incorporate a great deal more flexibility. As an example, consider the functional generalisation of the ARCH(1) model as defined in Hörmann et al (2012) as

$$\begin{aligned} Y_i &= \epsilon_i \sigma_i \\ \sigma_i^2 &= \delta + \beta(Y_{i-1}^2) \end{aligned}$$

where $\{\epsilon_i\}$ is a sequence of independent and identically distributed random functions in $\mathcal{H}$, $\beta : \mathcal{H}^+ \to \mathcal{H}^+$ is a non-negative operator and $\delta \in \mathcal{H}^+$; $\mathcal{H}^+$ denotes the set of non-negative functions in $\mathcal{H}$. The functional ARCH(1) model is of the form (2.3) with $A_i = \epsilon_i^2 \beta(Y_{i-1}^2)$ and $B_i = \epsilon_i^2 \delta$.

We shall assume that for each $i \in \mathbb{N}$, $(Z_{i,s})_{s\in\mathbb{N}}$ are not only uncorrelated, but also independent. For example, in (2.3), independence of the principal component scores implies

$$A_i(v,u) = \sum_{s\in\mathbb{N}} W_{i,s}\varphi_s(v)\varphi_s(u),$$

where $\left\{(W_{i,s})_{i\in\mathbb{N}}; s \in \mathbb{N}\right\}$ are independent collections of i.i.d. random variables. A thorough discussion of this independence assumption is provided in Section 5.

Throughout the paper we shall use the following notation. For a (non-random) operator A from $\mathcal{H}$ into itself, we shall also use A to denote its kernel, so that we may switch between operator notation and integrals using the kernel, e.g. $Af = \int_{\mathcal{V}} A(v,u) f(u)\, du$. We use the Frobenius (Hilbert-Schmidt) norm as the operator's norm, $|A|_F = (trace(A^*A))^{1/2}$ where A^* is the adjoint of A. The operators considered in this paper are all Hermitian (symmetric) hence, $trace(A^*A) = trace(AA) = \int\int |A(u,v)|^2\, dudv$. Moreover, $\mathcal{H}$ is equipped with the inner product $\langle y, x\rangle = \int y(u)\, x(u)\, du$, $x, y \in \mathcal{H}$, and the distance between elements is given by $|y - x| = (\langle y - x, y - x\rangle)^{1/2}$. It should be clear by the context when $|\bullet|$ is the absolute value or the norm for $\mathcal{H}$. $\mathbb{1}\{\cdot\}$ is the indicator function of a set. Finally, $\lesssim$ is inequality up to a finite absolute constant (i.e. the left hand side is big O of the right hand side).

3. Conditional estimation

Suppose initially that we observe n realisations of the random functions $(Y_i)_{i=1}^n$ over the whole of the domain $\mathcal{V}$. In light of equation (2.2), the objects of interest are the eigenfunctions, $(\varphi_s(v))_{s\in\mathbb{N}}$, of the integral operator with kernel $C_0(u,v)$, from which one can construct the random variables, $(Z_{i,s})_{i=1}^n$, i.e. the factor loadings.

Condition 1. *(i) For each* $s \in \mathbb{N}$ $(Z_{i,s})_{i=1}^{n}$ *is a strictly stationary Markov chain with strong mixing coefficients*

$$\alpha_s(i) := \sup\{|\Pr(A \cap B) - \Pr(A)\Pr(B)| : A \in \sigma(Z_{0,s}), B \in \sigma(Z_{i,s})\}$$

bounded by $\alpha(i) \lesssim i^{-a}$, $a > 1$. *(ii) For each* $i = 1, \ldots, n$, $(Z_{i,s})_{s\in\mathbb{N}}$ *is a sequence of independent mean zero random variables.*

Referring to Chapter 1.8 of van der Vaart and Wellner (2000) the Markov process, $(Y_i)_{i=1}^{n}$, with stochastic representation in $(Z_{i,s})_{s\in\mathbb{N}}$, independent across s for every i (Condition 1), exists in a Hilbert space under Condition 2.

Condition 2. *For some* $\nu \geq 2$, $\mathbb{E}|Y_i|^{\nu} < \infty$.

Condition 3. $C_0(u,v)$ *has* $d \geq 2$ *mixed partial derivatives which are Lebesgue square integrable, and the eigenvalues of the covariance operator* C_0 *are distinct.*

Suppose we are interested in estimating the transition distribution of $(Y_i(v))_{i=1}^{n}$, say P_Y, for Lebesgue almost all v. By virtue of equation (2.2),

$$\begin{aligned} P_Y(y(v)|x) := \Pr(Y_i(v) \leq y(v)\,|Y_{i-1} = x) &= \Pr\left(Y_i(v) \leq y(v)\,|\int_{\mathcal{V}} |Y_{i-1}(u) - x(u)|^2\,du = 0\right) \\ &\quad \text{[equality in } L_2\text{]} \\ &= \Pr(Y_i(v) \leq y(v)\,|\,\langle\varphi_s, Y_{i-1}\rangle = x_s,\ s \in \mathbb{N}), \end{aligned}$$

which follows by function representation in $\mathcal{H}$ and orthonormality of the eigenfunctions, defining $x_s := \langle\varphi_s, x\rangle$. Let $P_s(z_s|x_s) := \Pr(\langle\varphi_s, Y_i\rangle \leq z_s|\,\langle\varphi_s, Y_{i-1}\rangle = x_s)$, i.e. the transition distribution of $(\langle\varphi_s, Y_i\rangle)_{i\in\{1,\ldots,n\}}$, $s \in \mathbb{N}$. Using independence of the $(Z_{i,s})_{s\in\mathbb{N}}$ for any i,

$$P_Y(y(v)|x) = \lim_{S\to\infty} \int_{\mathbb{R}^S} \mathbb{1}\left\{\sum_{s=1}^{S} \varphi_s(v)\, z_s \leq y(v)\right\} \prod_{s=1}^{S} dP_s(z_s|x_s),$$

hence

$$\mathbb{E}[g(Y_i)\,|Y_{i-1} = x] = \int_{\mathbb{R}} g(y(v))dP(y(v)|x) = \lim_{S\to\infty} \int_{\mathbb{R}^S} g\left(\sum_{s=1}^{S} \varphi_s(v)\, z_s\right) \prod_{s=1}^{S} dP_s(z_s|x_s),$$

The symbolic representation above emphasises the important quantities in the estimation procedure to be discussed. By independence, we only need S one dimensional estimators for P_s, rather than a multivariate one. We also require an estimator for the orthogonal eigenfunctions $\{\varphi_s(\cdot), s = 1, \ldots, S\}$ of the covariance operator, C_0. The next condition places very general requirements on the eigenfunction estimator. In practice, we will want to know the details of how to estimate $\{\varphi_s, s = 1 \ldots, S\}$, and this issue is addressed in Section 4.

Condition 4. $\{\widehat{\varphi}_s(\cdot), s = 1, \ldots, S\}$ *are orthonormal estimators of* $\{\varphi_s(\cdot), s = 1, \ldots, S\}$ *such that* $\sup_v |\widehat{\varphi}_s(v)| < \infty$ *for any* $s \geq 1$, *and which fulfill*

$$|\widehat{\varphi}_s - \varphi_s| = O_p(\Lambda_s^{-1} n^{-\tau})$$

$\tau > 0$, *where the* $\{\Lambda_s : s = 1, \ldots, S\}$ *depend on spacings between the eigenvalues. More specifically* $\Lambda_s \asymp |\lambda_s - \lambda_{s-1}| + |\lambda_s - \lambda_{s+1}|$, $\Lambda_1 \asymp |\lambda_s - \lambda_{s+1}|$.

Let $\Psi_{\delta,s} := \{\psi : |\psi - \varphi_s| \leq \delta, |\psi| = 1\}$. Throughout the rest of this paper, we shall work with $Z_i(\psi) := \langle \psi, Y_i \rangle$ and just drop the subscript s in $\Psi_{\delta,s}$ unless needed. Note that $(Z_i(\psi))_{i\in\mathbb{N}}$ is also a Markov chain, as $Z_i(\psi)$ is just a linear functional of $(Y_i(v))_{v\in\mathcal{V}}$. Condition 5 places more structure on the time series dependence of the factor loadings than that provided by Condition 1 and is discussed further in Section 5.

Condition 5. *There is a* $\beta \in (0, 1]$ *and a* $\delta > 0$, *such that for any* r *small enough, all* $z \in \mathbb{R}$ *and almost all* z',

$$\max_{s>0} \sup_{\psi\in\Psi_{\delta,s}} |\Pr(Z_1(\psi) \leq z | Z_0(\psi) = z' - r) - \Pr(Z_1(\psi) \leq z | Z_0(\psi) = z' + r)| \lesssim r^{\beta}.$$

Conditioning on $Z_0(\psi) = z'$, *for almost all* z', $Z_1(\psi)$ $(\psi \in \Psi_{\delta,s})$ *has a tight measure absolutely continuous with respect to the Lebesgue measure.*

Let $P_s(z|x_s) = \Pr(Z_{i,s} \leq z | Z_{i-1,s} = x_s)$. Let $B(x_s, r_s)$ be an interval of Lebesgue measure r_s centered at $x_s \in \mathbb{R}$. We define our nearest neighbour estimator of $P_s(z_s|x_s)$ as

$$P_{sn}(z_s | B_s(\langle x, \widehat{\varphi}_s \rangle, r_s)) := \frac{\sum_{i=2}^{n} \mathbb{1}\{\langle \widehat{\varphi}_s, Y_{i-1} \rangle \in B(\langle x, \widehat{\varphi}_s \rangle, r_s), \langle \widehat{\varphi}_s, Y_i \rangle \leq z_s\}}{\sum_{i=2}^{n} \mathbb{1}\{\langle \widehat{\varphi}_s, Y_{i-1} \rangle \in B(\langle x, \widehat{\varphi}_s \rangle, r_s)\}}. \tag{3.1}$$

Note that $x_s = \langle x, \varphi_s \rangle$ is unknown if the eigenfunction φ_s is unknown, hence we must also account for this in the estimator. It is tacitly assumed that r_s is taken large enough to ensure that the denominator in (3.1) does not vanish. We shall show that, in some suitable topology,

$$P_{sn}(z_s | B(\langle x, \widehat{\varphi}_s \rangle, r_s)) = P_s(z_s | \langle x, \varphi_s \rangle) + o_p(1),$$

$s \leq S$, for large n under suitable conditions on $S = S(n)$ and $r_s = r_s(n)$, increasing and decreasing sequences respectively. This result is enough to consistently estimate $\mathbb{E}_{i-1} g(Y_i)$ under regularity conditions on g, where $\mathbb{E}_{i-1}$ is expectation conditional on the previous observation Y_{i-1} (e.g. Sancetta, 2009; Linton and Sancetta, 2009).

3.1. Summary of the estimation procedure

1. Estimate the first S orthonormal eigenfunctions $\{\varphi_s(\cdot), s = 1, \ldots, S\}$ of the covariance operator; estimators satisfying Condition 4 are presented and discussed in Section 4.
2. Use the estimated eigenfunctions $\{\widehat{\varphi}_s : s = 1, \ldots, S\}$ to obtain $\left\{(Z_{i,sn})_{i\in\{1,\ldots,n\}} ; s = 1, ..., S\right\}$, where $Z_{i,sn} := Z_{i,s,n} = \langle Y_i, \widehat{\varphi}_s \rangle$;
3. Use $\left\{(Z_{i,sn})_{i\in\{1,\ldots,n\}} ; s = 1, ..., S\right\}$ to derive the formal estimator

$$P_n(y(v)|x) := \int_{\mathbb{R}} \mathbb{1}\left\{\sum_{s=1}^{S} \widehat{\varphi}_s(v) z_s \leq y(v)\right\} \prod_{s=1}^{S} dP_{sn}(z_s | B_s(\langle x, \widehat{\varphi}_s \rangle, r_s)) \tag{3.2}$$

of $P_Y(y|x)$. In equation (3.2) we use the usual linear functional notation in which integration over the empirical distribution is just summation over a finite set of z_s values corresponding to the points at which we observe the data. With this notation, our estimator of the conditional expectation of any Lipschitz function $g : \mathbb{R} \to \mathbb{R}$ can be written as

$$\int g(y(v))P_n(y(v)|x) = \sum_{i=2}^{n}\left[\prod_{s=1}^{S}\frac{\mathbb{1}\{Z_{i-1,sn} \in B_s(x_{s,n}, r_s)\}}{\sum_{i=2}^{n}\mathbb{1}\{Z_{i-1,sn} \in B_s(x_{s,n}, r_s)\}}\right] g\left(\sum_{s=1}^{S} Z_{i,sn}\widehat{\varphi}_s(v)\right), \tag{3.3}$$

where $x_{s,n} := \langle x, \widehat{\varphi}_s\rangle$.

Remark 1. When $g(y) = y$, the estimator is simply

$$\int y(v)dP_n(y(v)|x) = \sum_{s=1}^{S}\overline{Z}_{sn}\widehat{\varphi}_s(v), \tag{3.4}$$

where

$$\overline{Z}_{sn} = \sum_{i=2}^{n} Z_{i,sn}\frac{\mathbb{1}\{Z_{i-1,sn} \in B_s(x_{s,n}, r_s)\}}{\sum_{i=2}^{n}\mathbb{1}\{Z_{i-1,sn} \in B_s(x_{s,n}, r_s)\}}.$$

The following results show that the procedure is expected to provide results as good as if the $\{P_s(z_s|x_s) : s = 1, \ldots, S\}$ were known.

Theorem 1. *Suppose $r_s \to 0$ such that $r_s\Lambda_s n^{\tau} \to \infty$, where τ and Λ_s are those of Condition 4. Under Conditions 1 - 5, for almost all $x \in \mathcal{H}$, for any $S < \infty$, and all $s \leq S$,*

$$\sup_{z\in\mathbb{R}}|P_{sn}(z|B(\langle x, \widehat{\varphi}_s\rangle, r_s)) - P_s(z|x_s)| = O_p\left(n^{-1/2}r_s^{-1} + r_s^{\beta}\right) = o_p(1).$$

If r_s is chosen optimally as $r_s \asymp n^{-1/(2\beta+2)}$, the above display is $O_p\left(n^{-\beta/(2\beta+2)}\right)$.

Using Theorem 1, we also prove the following result for the conditional expectation of Lipschitz functionals of $Y_i(v)$.

Theorem 2. *Let $r_s \asymp n^{-1/(2\beta+2)}$ and let $g : \mathbb{R} \to \mathbb{R}$ be bounded Lipschitz with Lipschitz constant one. If $S\left(n^{-\beta/(2\beta+2)} + n^{-2\tau}\min_{s\leq S}\Lambda_s^{-2}\right) \to 0$, under Conditions 1 - 5, for Lebesgue almost all v,*

$$\int_{\mathbb{R}} g(y(v))\, dP_n(y(v)|x) = \int_{\mathbb{R}} g(y(v))\, dP(y(v)|x) + o_p(1).$$

Remark 2. Despite offering some guidelines on the choice of S, Theorem 2 does not provide the practitioner with a concrete rule. In Section 6 we simply choose S to be fixed at some small value, which is seen to be reasonable from an empirical point of view. As an alternative, S may be chosen by cross validation.

4. Eigenfunction estimation

We present a class of estimators satisfying Condition 4 under Condition 3 and the additional Condition 6, which is presented below. This result is formally stated in Lemma 1. In this section, we restrict attention to the closed interval $\mathcal{V} = [a, b]$ and without loss of generality we work with $\mathcal{V} = [0, 1]$. Our proposed procedure involves expanding the eigenfunctions in the Fourier basis and estimating the coefficients of that expansion. Write the covariance function in terms of the bivariate Fourier basis, as

$$C_0(u, v) = \sum_{r,p\in\mathbb{Z}} c(r, p) \exp\{i(ru + pv)\} \tag{4.1}$$

where $c(r, p)$ is the $(r, p)^{th}$ Fourier coefficient of $C_0(u, v)$. A natural estimator of the coefficients of the Fourier basis expansion of the eigenfunctions is thus obtained by truncating the above Fourier expansion between $-S$ and S; estimating the Fourier coefficients of an estimator, $C_n(u, v)$, of $C_0(u, v)$; and deriving an estimator for the Fourier coefficients of interest based on Mercer's Theorem. In particular, our proposed estimator is

$$\varphi^S_{s,n}(v) = \sum_{r=-S}^{S} a_{s,n}(r) \exp\{irv\} \tag{4.2}$$

where $a_{s,n}(r)$ is the r^{th} entry of the s^{th} eigenvector of C^S_n where C^S_n is the $S \times S$ matrix of Fourier coefficients $c^S_n(r, p)$ of the truncated bivariate Fourier expansion of the estimated covariance function, i.e.

$$C^S_n(u, v) = \sum_{r=-S}^{S} \sum_{p=-S}^{S} c^S_n(r, p) \exp\{i(ru + pv)\}, \tag{4.3}$$

which is the truncated Fourier representation of $C_n(u, v)$.

Remark 3. Any other atomic representation $C_0 = \sum_{r,p\geq 1} c\,(r, p)\, e_r e_p$ where are orthonormal functions can be used in place of the complex exponential. For the direct application of the results of Lemma 1 it is only necessary that the coefficients decay at the same rate as the Fourier coefficients of smooth functions as in Condition 3. Hence, for numerical calculations, if $\mathcal{V} = [0, 1]$, one may use the trigonometric basis

$$e_1 = 1;\ e_{2r}\,(v) = 2^{1/2} \cos\,(2\pi rv)\,;\ e_{2r+1}\,(v) = 2^{1/2} \sin\,(2\pi rv)\,;\ r \geq 1.$$

Remark 4. When C_n is the empirical covariance function (e.g. Bosq, 2000, p.9), C^S_n is just the covariance matrix of the first S frequencies of the Fourier transformed data. This follows by the orthonormality of Fourier basis functions and corresponds to steps 1 and 3 in Silverman (1996) p. 6.

Condition 6. *The sequence* $(C_n)_{n\geq 1}$ *satisfies* $|C_n - C_0|_F = O_p\,(n^{-\tau})$, $\tau > 0$.

We denote by $\{\varphi^S_{sn} : s \in \mathbb{N}\}$ the set of eigenfunctions of C^S_n. $\{\varphi^S_{s,n} : s = 1, \ldots, S\}$ are clearly orthogonal, hence satisfy Condition 4 for an appropriate choice of S; furthermore $\varphi^S_{sn} := 0$, for $s > S$. The error in the truncation of the Fourier expansion is well known and its implications in the present context is summarised in Lemma 1.

Lemma 1. *Under Conditions 3 and 6, for any $s \geq 1$, for $\mathcal{V} = [0,1]$, and for $S \gtrsim n^{-\tau/d}$, $\left|\varphi_{s,n}^{S} - \varphi_s\right| = O_p\left(\Lambda_s^{-1} n^{-\tau}\right)$, and $\sup_v \left|\varphi_{s,n}^{S}(v)\right| < \infty$, for any n, where τ is that of Condition 6 and $\Lambda_s \asymp |\lambda_s - \lambda_{s-1}| + |\lambda_s - \lambda_{s+1}|$, $\Lambda_1 \asymp |\lambda_s - \lambda_{s+1}|$.*

5. Remarks on conditions and results

Condition 1

The polynomial rate of decay for the strong mixing coefficients of the Markov chain is weak (Bradley, 2005, for a review of mixing coefficients). Many stochastic processes satisfy this condition and actually geometric decay. For example GARCH models can be embedded into a multivariate version of the stochastic recurrence equation

$$Z_i = A_i Z_{i-1} + B_i,$$

which is the matrix version of (2.3). Under regularity conditions, the above is geometrically ergodic, hence strongly mixing with geometric decay (Basrak et al., 2002, Theorem 2.8). ARMA processes of any finite order are also strong mixing with geometric decay when the error distribution is absolutely continuous w.r.t. the Lebesgue measure (Mokkadem, 1988, Theorem 1). This allows us to easily control the time series dependence. Regarding Condition 1 (ii), note that the factor loadings are uncorrelated by construction and for several stochastic dependence concepts, zero correlation implies independence (Joe, 1997, Esary et al., 1967).

Condition 2

The moment condition on $|Y_i|$ implies moment conditions on the factor loadings

$$\mathbb{E}\left|Y_i\right|^{\nu} = \mathbb{E}\left[\int \left(\sum_{s=1}^{\infty} Z_{i,s}\varphi_s(v)\right)^2 dv\right]^{\nu/2} = \mathbb{E}\left(\sum_{s=1}^{\infty} Z_{i,s}^2\right)^{\nu/2} < \infty.$$

Since all quantities in the summation are positive,

$$\infty > \mathbb{E}\left(\sum_{s=1}^{\infty} Z_{i,s}^2\right)^{\nu/2} \geq \mathbb{E}\sum_{s=1}^{\infty}\left|Z_{1,s}\right|^{\nu},$$

i.e. absolute summability of $\mathbb{E}Z_{i,s}^{\nu} = \mathbb{E}\langle Y_i, \varphi_s\rangle^{\nu}$ for any i, by stationarity. This ensures $\sum_{s\in\mathbb{N}}\langle Y_i, \varphi_s\rangle^2$ converges almost surely (van der Vaart and Wellner, 2000, Lemma 1.8.1). Since $\mathcal{V} = [0,1]$ and C_0 is linear and continuous by Condition 3, then $\sup_{u,v}|C_0(u,v)| < \infty$. By Mercer's Theorem and existence of second moments, this implies that the eigenfunctions are also uniformly bounded: $\sup_{s\in\mathbb{N}}\sup_{v\in\mathcal{V}}|\varphi_k(v)| < \infty$. Bounded eigenfunctions ensure that, for $k \in \mathbb{N}$, $\sum_{s\in\mathbb{N}}\langle Y_i, \varphi_s\rangle\varphi_k$ converges whenever $\sum_{s\in\mathbb{N}}\langle Y_i, \varphi_s\rangle^2$ converges, giving completeness. Some form of measurability of $\langle Z_{i,s}, \varphi_s\rangle$ $(\forall i, s)$ is required for the existence of the separable process (2.2) (see e.g. van der Vaart and Wellner, 2000, Lemma 1.8.2). Throughout this paper we assume everything is measurable.

Condition 3

Condition 3 controls the rate of approximation using S principal factors. In particular, Condition 3 ensures that the process $(Y_i(v))_{v\in\mathcal{V}}$ is asymptotically finite dimensional in the sense of van der Vaart and Wellner (2000), hence tight by Condition 2 (van der Vaart and Wellner, 2000, Lemma 1.8.1). Eigenvalues of C_0 are distinct for most regular cases. For example, one can conjecture that the s^{th} eigenvalue of the parametric covariance function $\exp\left\{-\theta^2|u-v|^2\right\}$ is $\lambda_s \asymp \lambda^s$ where $\lambda \in (0,1)$ increases with $\theta(>0)$ (e.g. Rasmussen and Williams, 2006 eq. 4.39, for details, where integrals are with respect to the Gaussian measure rather than the Lebesgue measure over a compact interval).

Condition 4

Although the orthogonality condition rules out the use of the estimator of Silverman (1996), there are several possibilities for the choice of eigenfunctions satisfying Condition 4. For $\mathcal{V}$ a closed interval of the real line, the rate of convergence of orthogonal eigenfunction estimators of the form (4.2) is established in Lemma 1 under Condition 6 on the rate of convergence of the covariance function estimator. For $\mathcal{V}$ a compact subset of $\mathbb{R}^k$ $(k>1)$, it follows immediately by Lemmata 4.2 and 4.3 of Bosq (2000) that the eigenfunctions of the estimated covariance kernel converge to those of the true ones at rate $\Lambda_s^{-1}n^{-\tau}$ in inner product norm as long as Condition 6 holds. In general however, τ in the convergence rate of the covariance function estimator will depend on k. The exception is if the true covariance function is in some parametric class whose dimension does not depend on k. Moreover, in practice it is infeasible to obtain the eigenfunctions of the estimated covariance function directly and we need to propose an estimator for the eigenfunctions such as that of equation (4.2). These estimators are in general affected by the dimensionality of $\mathcal{V}$, for instance if an estimator analogous to that is equation (4.2) was to be employed, we would require kd derivatives in order for the Fourier approximation error to be of the same order. We therefore do not consider generalisations of Lemma 1 to the case of $k>1$.

Condition 5

Suppose for a moment that $\langle\psi, Y_i\rangle = Z_{i,s}$, (e.g. $\psi\in\Psi_{\delta,s}$, $\delta=0$), then, the condition is standard. It says that the transition distribution is smooth in the conditioning variable, though not necessarily differentiable and also that the transition density exists. The Lipschitz continuity condition is necessary to control the rate of convergence in the nearest neighbour estimator of the transition distribution of the $\{Z_{i,s}: s=1,\ldots,S\}$. It is however not standard that the condition needs to hold uniformly in $\psi\in\Psi_{\delta,s}$, as in this case we usually have $\langle\psi,Y_i\rangle \neq Z_{i,s}$. The plausibility of the condition follows from the fact that the Markovian structure is preserved by the inner product $\langle\psi,Y_i\rangle$. For definiteness, suppose that the simple stationary AR(1) structure

$$Z_{i,s} = \rho_s Z_{i-1,s} + B_{is}, \tag{5.1}$$

where $\sup_s|\rho_s|<1$ and $(B_{is})_{i\in\mathbb{Z}}$ is i.i.d. Gaussian with mean zero and variance $\left(1-\rho_s^2\right)\lambda_s$, using the fact that $\mathbb{E}Z_{i,s}^2=\lambda_s$ by construction. Recall that the $(Z_{i,s})_{s\in\mathbb{N}}$ are independent across s. Then, for any bounded ψ using the fact that also the eigenfunctions are bounded, let $\gamma_s=\langle\psi,\varphi_s\rangle<\infty$. Then, $\langle\psi,Y_i\rangle=\sum_{s=1}^{\infty}\gamma_s Z_{i,s}$ is Gaussian with

mean zero and variance $\sum_{s=1}^{\infty} \gamma_s \lambda_s < \infty$ (recall λ_s is the s^{th} eigenvalue). From (5.1), $\langle \psi, Y_i \rangle$ has conditional mean

$$\begin{aligned} \mathbb{E}_{i-1} \langle \psi, Y_i \rangle &= \sum_{s=1}^{\infty} \gamma_s \mathbb{E}_{i-1} Z_{i,s} \\ &= \sum_{s=1}^{\infty} \gamma_s \rho_s Z_{i-1,s} \end{aligned}$$

and conditional variance, using independence across s,

$$\begin{aligned} Var_{i-1} \left(\langle \psi, Y_i \rangle \right) &= \sum_{s=1}^{\infty} \gamma_s^2 \mathbb{E}_{i-1} B_{is}^2 \\ &= \sum_{s=1}^{\infty} \gamma_s^2 \left(1 - \rho_s^2 \right) \lambda_s. \end{aligned}$$

The transition distribution of $\langle \psi, Y_i \rangle$ is Gaussian, it has a density and it is Lipschitz of order $\beta = 1$ in the conditioning argument. Here, Gaussianity makes the calculations feasible. The example shows that the $\{(Z_{i,s})_{s \in \mathbb{N}} : i \in \mathbb{N}\}$ are not changed by the function ψ, and the dependence of $\langle \psi, Y_i \rangle$ on the past only comes from the $\{(Z_{i,s})_{s \in \mathbb{N}} : i \in \mathbb{N}\}$ which are independent across s.

Condition 6

For $\tau = 1/2$, this condition is satisfied by a correctly specified parametric covariance model as well as by the empirical covariance function estimator, (see e.g. Bosq, 2000, Theorem 4.1 using Condition 1 to control the time series dependence). The empirical covariance will only be suitable in practice if the functions are observed at a dense enough grid of points in $\mathcal{V}$. Alternatively, a smooth estimator via two dimensional nonparametric smoothing techniques can be considered (e.g. Yao et al., 2005; Hall et al., 2006). In this case, $\tau < 1/2$, with exact optimal rate depending on the smoother and the properties of C_0. Under regularity conditions, a first order kernel would lead to $\tau = 1/3$. Under suitable conditions, Hall et al. (2006) show improved convergence rates for the estimated eigenfunctions: compare Lemma 1 in Section 8 with Theorem 1-2 in Hall et al. (2006).

Theorem 1

For observable factor loadings, $r_s \asymp n^{-1/(2\beta+2)}$ would still be optimal with the same convergence rate stated in Theorem 1. Hence, estimated factor loadings and eigenfunctions do not affect the rate of convergence as long as $r_s n^{\tau} \Lambda_s \to \infty$. Substituting the optimal rate for r, and noting that $\beta \leq 1$, this is the case when $n^{-\tau} \Lambda_s^{-1} = o\left(n^{-1/4}\right)$. From the remark on Condition 6, $\tau > 1/3$ in most practical cases when $\mathcal{V}$ is a closed interval of the real line. As long as the convergence rate of the estimator of the eigenfunctions is faster than the convergence rate of the nearest neighbour estimator of the transition distribution of the factor loadings, we do not need to worry about not directly observing the factor loadings. Theorem 1 provides a bound for the Kolmogorov's metric of the joint distribution of the factor loadings. This implies convergence of estimated conditional expectation of any function of bounded variation of the factor loadings (here bounded variation is intended in the sense of Hardy; see e.g. Sancetta, 2009, for details).

The results are in terms of the $(\Lambda_s)_{s=1}^S$ of Condition 4. This implies that the estimator is less precise for higher order factors. Of course, a faster decay of the eigenvalues, requires a smaller number of principal components to obtain a good approximation. In practice, one usually restricts attention to a small number of principal factors.

Theorem 2

The result of part (ii) of the Theorem also applies to $\int_{\mathcal{V}} \int_{\mathbb{R}} g\left(y\left(v\right)\right) dP_n\left(y\left(v\right)|x\right) dv$ by dominated convergence if $g_{Pn}\left(v\right) := \left|\int_{\mathbb{R}} g\left(y\left(v\right)\right) dP_n\left(y\left(v\right)|x\right)\right|$ is bounded by some integrable function. The error rate is only increased by $|\mathcal{V}|$ which is finite since $|\mathcal{V}| < \infty$.

6. Simulation performance

Here we compare several versions of our method to the linear autoregressive estimator (LA) of Bosq (2000) and to the predictive factors estimator (PF) of Kargin and Onatski (2008) when the simulated data were generated by a model that satisfies the linear autoregressive assumption under which LA and PF were designed to be effective. For comparison, we also present the performance of the mean function and the last available curve in the sample. In the next sub-section we present results for a more general data generating mechanism.

6.1. Data generated by an FAR(1)

We follow the experimental design of Didericksen et al (2012), generating a functional autoregression (FAR) according to

$$Y_i(v) = \int_{\mathcal{V}} k(u,v) Y_{i-1}(u) du + \epsilon_i(v) \tag{6.1}$$

where $\mathcal{V} = [0,1]$ and where we choose $k(u,v)$ to be the Gaussian kernel $k(u,v) = C\exp\{-(u^2+v^2)/2\}$ where C is a normalising constant chosen such that $\|k\|_F = 0.8$. The error process $\epsilon(v)$ is taken to be the Brownian Bridge process

$$\epsilon(v) = W(v) - vW(1)$$

where $W(\cdot)$ is the standard Wiener process generated as

$$W\left(\frac{k}{K}\right) = \frac{1}{\sqrt{K}} \sum_{j=1}^{k} Z_j \quad j = 0, 1, \ldots, K$$

with $(Z_j)_{j=1}^K$ independent standard normal random variables and $Z_0 = 0$. As in Didericksen et al (2012), we use a burn-in sample of 50 functional observations.

In each simulation we compute the integrated square error as realisations of

$$\int_{\mathcal{V}} (Y_{n+1}(v) - \widehat{Y}_{n+1}(v))^2 dv \tag{6.2}$$

where $\widehat{Y}_{n+1}$ is the estimator of the $(n+1)^{th}$ functional data object given observations up to time n. In practice, we replace integration by summation over a finite grid of points.

In the first version of the NN estimator (E-FNN), we simply consider discretising the empirical covariance on a regular grid and taking the singular value decomposition of the resulting matrix in order to estimate the eigenfunctions at a finite set of points. The second version of the NN estimator (E-SNN) uses the empirical covariance function and estimates the eigenfunctions on a finite set of points as the eigenvectors of a matrix of evaluation points of the Fourier transformed empirical covariance function, which is equivalent to the empirical covariance of the Fourier transformed data (see e.g. Bosq, 2000); this equivalence is useful for functional data that are observed at a different set of points along each functional data unit. We use PF to denote the predictive factors estimator of Kargin and Onatski (2008), LA to denote the estimator of Bosq, Mean to denote the mean function, and Previous to denote the last available curve in the sample. For the estimator of Bosq (2000), the eigenfunctions and eigenvalues are estimated by the smoothed functional principal components approach of Silverman (1996) using the publicly available `fda` package of Ramsay, Wickham, Hooker and Graves (2009).

The NN estimation is performed by first estimating $\{\varphi_s : s = 1, \ldots, S\}$ by one of the procedures mentioned in the previous paragraph, and transforming the realisations of $(Y_i)_{i=1}^n$ to realisations of the scalar random variables in the dual space $\{(Z_{i,sn})_{i=1}^n : s = 1, \ldots, S\}$ by taking the inner product $\{\langle Y_i, \widehat{\varphi}_s \rangle, i = 1, \ldots, n; s = 1, \ldots, S\}$. In practice we replace the integral by a sum over the finite grid of points over which each of the $(Y_i)_{i=1}^n$ is observed. We then construct a ball of radius r_s, around $x_{sn} = \langle Y_n, \widehat{\varphi}_s \rangle$ for each $s = 1, \ldots, S$. In the experiments below, we take the radius of the set $B(x_{sn}, r_s)$ in the NN estimator to be $r_s = c_s \sigma_s n^{-1/4}$, and $S = 2$. For good performance of the NN estimator, S can be taken larger (as in Theorems 1 and 2), but this causes the performance of the Bosq (2000) and PF estimators to deteriorate. Here, c_s is a tuning constant and σ_s is the standard deviation of the s^{th} estimated component score from the sample of size n. We carried out experiments with different values of c_s and found that the choice $c_s = 1.5$ worked well in our case. We then construct the NN estimates of each of the S transition distributions over a large set of values for z_s as in equation (3.1). We estimate the conditional expectation of $Z_{n+1,s}$ by replacing the integral $\int z_s dP_{sn}(z_s | B(x_{sn}, r_s))$ by a sum over the finite grid of z_s values corresponding to the sample data (i.e. by implementing equation (3.4)), which subsequently allows us to estimate the conditional expectation of Y_{n+1} by replacing the $(Z_{n+1,s})_{s=1}^S$ in the truncated Karhunen-Loève representation by their expected values.

FIGURE 1 HERE

Figure 1: **Boxplots of square root ISE for E-SNN, E-FNN, LA, PF, Mean and Previous for $n \in \{50, 100, 500, 1000\}$. 4000 Monte Carlo replications with functional data drawn from the FAR model described by equation (6.1).**

6.2. General data generating mechanisms with dependence

Although the results presented in the previous subsection are favourable to our procedure, the FAR(1) model is very restrictive and we would like to consider more general nonlinear data generating processes. Since interest lies in the performance of our estimator over a range of different data generating mechanisms, we follow Friedman (2001) in that

for each of our simulated examples, we simulate a random curve time series of size $n+1$ from the random model

$$Y_i(v) = \sum_{k=1}^{K} R_{i,k} e_k(v), \tag{6.3}$$

$$e_1 = a_1;\ e_{2k}(v) = s^{-b} a_{2k} \cos(2\pi k v);\ e_{2k+1}(v) = k^{-b} a_{2k+1} \sin(2\pi k v),$$

where $b \in [1,4]$, and $a_k \in [0,1]$, $R_{i,k} = \phi R_{i-1,k} + \zeta_{i,k}$, $\phi \in [-0.99, 0.99]$ and the $\{\zeta_{i,k}\}$ are i.i.d. in i and k and have t-distribution with $\nu \in [3,30]$ degrees of freedom. For each of the simulations, the parameters $(a_1, a_2, ..., a_K, b, \nu, \rho)$ are generated from a uniform distribution with support as defined above. Moreover, for each sample, the model is only observed over a fixed grid, $\mathcal{V}'$ of 20 points in $\mathcal{V} = [0,1]$. Note that $(e_k)_{k>1}$ are proportional to the eigenfunctions, and $(R_{i,k})_{k>1}$ are proportional to the principal factor loadings.

We consider both the case of an equally spaced grid, and a randomly generated grid of 20 points, uniformly distributed on $\mathcal{V}$. Table 1. provides the step-ahead forecast performance of several versions of the nearest neighbour estimator for training samples of size $n = 50, 100, 500, 1000$, as measured by the mean integrated square error (MISE) and its standard deviation over 5000 simulations. We report the same summary statistics for the PF estimator, the LA estimator, the unconditional mean function, and the last available function in the training sample.

Estimation is performed as in the previous section, except now we also consider using the true covariance function rather than the estimated one (T-SNN), as this is directly available from equation (6.3). For space considerations, we do not present the PF estimator, whose performance was worse than the other methods.

		T-SNN		**E-SNN**		**E-FNN**		**LA**		**Mean**		**Previous**	
grid	n	MISE	s.e.	MISE	s.e	MISE	s.e	MISE	s.e	MISE	s.e	MISE	s.e
equi	50	1.1620	1.1857	1.1108	1.8057	2.0733	4.5433	2.8033	7.5049	2.8281	7.8394	5.9545	27.016
equi	100	1.1409	2.2164	1.0561	1.6986	2.0876	5.2641	2.9289	8.6935	2.9374	8.8427	6.4815	31.120
equi	500	1.0875	2.1620	1.0608	1.9875	2.1050	4.9924	2.9992	8.2723	3.0016	8.2854	5.9430	22.735
equi	1000	1.0607	1.9549	1.0449	1.8393	2.0298	4.7432	2.8200	8.3019	2.8213	8.3107	6.0760	30.145
rand	50	1.1847	2.6513	1.1392	2.7145	1.9830	5.2117	2.8774	8.0396	2.9033	8.3173	6.4219	29.094
rand	100	1.1325	2.4495	1.0887	2.3518	1.7895	4.5411	2.8342	9.6220	2.8489	9.9407	6.3363	38.260
rand	500	1.0209	2.0653	0.9955	1.6189	1.6845	3.5614	2.9669	7.7604	2.9728	7.7912	6.0124	21.580
rand	1000	1.0039	1.7040	1.0178	1.8293	1.6350	3.2332	2.9165	8.2991	2.9190	8.3182	6.2206	26.937

Table 1: **MISE and its standard error of T-SNN, E-SNN, E-FNN, LA, Mean and Previous for various values of n in 5000 simulations of the model described by equation (6.3). 'equi' denotes an equispaced grid of observation points and 'rand' denotes a randomly generated grid that is the same for all n observations.**

FIGURE 2 HERE

Figure 2: **Boxplots of square root ISE for T-SNN, E-SNN, E-FNN, LA, Mean and Previous for various values of n in 5000 simulations of the model described by equation (6.3) with $\mathcal{V}'$ equispaced. In both images, the groups correspond (from left to right) to $n = 50$, $n = 100$, $n = 500$ and $n = 1000$. These plots correspond to the same simulations as those described in the upper portion of Table 1.**

6.3. Summary of simulation performance

The results from this section can summarised as follows.

1. Predictive factors (PF) is competitive with the procedure of Bosq (2000) (LA) when the data are generated by a functional linear autoregression as described in Section 6.1, but the performance can be poor for more complicated data generating mechanisms such as those described in Section 6.2. It should be noted however, that the PF procedure involves some tuning parameters that we have not attempted to optimise, so it may be the case that the performance of PF can be improved. Nevertheless, our results are consistent with those of Didericksen et al (2012).
2. E-SNN uniformly outperforms all competing procedures when data are generated according to the FAR (1) model, with the performance improving with the sample size.
3. For the more general data generating mechanism, E-SNN can outperform even the estimator that uses the true covariance function, which can be computed explicitly with this particular data generating mechanism. The reason for this is that the singular value decomposition based on the empirical covariance function produces principal component scores that are uncorrelated in sample, whilst the PC scores based on the true covariance will only be uncorrelated in the limit as $n \to \infty$.
4. LA performs better than the mean function, but with the more general data generating mechanism, the improvement is only marginal and the estimator does not compete with the NN procedures.

7. Empirical performance

Empirical performance on geophysics data

The Southern Oscillation Index (SOI) is a measure of the difference in southern surface air pressure over time between Tahiti and Darwin, Australia. It is used as a measure of the strength of the Southern Oscillation, which is a major component of the el-Niño and la-Niña phenomena. El-Niño and la Niña can have a dramatic impact on weather conditions around the world, and the resulting floods, droughts and migration of fish populations can be particularly devastating for countries heavily dependent on fishing and agriculture. With this in mind, it would be highly beneficial to be able to predict future Southern Oscillation patterns based on a time series of annual SOI curves. This curve time series is very oscillatory, with peaks and troughs occurring at irregular intervals. We examine the performance of E-FNN and E-SNN, along with the mean function, the last available curve, the predictive factors approach of Kargin and Onatski (2008) and the linear autoregressive approach of Bosq (2000). The data for this example are publically available at http://www.bom.gov.au/climate/current/soi2.shtml.

Our sample is a time series of 135 annual Southern Oscillation curves. We remove the last 51 curves from the sample to form the initial training set, which we use to predict one step ahead. We sequentially increase the size of the training sample by one until we reach 134, each time predicting one step ahead so that we characterise the prediction performance of the competing estimators over a test set of 50 functional observations. We subtract the 7-year moving average from each curve before implementing each of the competing procedures. The MISE and their standard errors over the 50 test observations are presented in Table 3., where the error is scaled by the standard deviation function before computing the ISE and summary statistics. This ensures that the statistics we observe are unit-free and may be compared across different data sets.

Empirical performance on electricity demand data

We use a data set from the R package `fds` (Shang and Hyndman, 2011) which consists of observations on half-hourly electricity demand in Adelaide between 6 June 1997 and 31 March 2007. Because commodity demand is characterised by day-of-the-week effects, which would potentially make the curve time series non-stationary between curves, we decide to segment the curve time series into 7 different curve time series, one corresponding to each day of the week. We focus only on the Monday series here. We subtract the 52-week moving average from each curve before implementing each of the competing procedures; this captures both the seasonal component and the observed trend that electricity demand in Adelaide is increasing over time.

The same split into training and test data is done as in the SOI example, and we again consider the prediction performance of competing estimators over a test set of 50 functional observations. The MISE and their standard errors over the 50 test observations are presented in Table 3., where again, the error is scaled by the standard deviation function before computing the ISE and summary statistics. The Mean estimator in this case is the 52-week moving average at the end of the training sample.

		E-FNN		**E-SNN**		**LA**		**Mean**		**Previous**	
data	n	MISE	std. err.	MISE	std. err.	MISE	std. err.	MISE	std. err.	MISE	std. err.
SOI	135	1.3368	0.8338	1.3192	0.8306	1.3461	0.8347	1.3453	0.8343	2.3713	1.6690
Elec	508	1.2556	1.9258	1.1038	2.0268	1.2436	1.8136	1.3133	1.8201	1.5344	2.3825

Table 3: **MISE and its standard error of E-FNN, E-SNN, PF, Mean and Previous, over a test set of 30, for the SOI and electricity data sets.**

In Figure 3. below, we present the last observation in each data set along with its prediction by several competing methods. Although E-FNN does very well at predicting this particular realisation for the Electricity data, E-SNN performs better over the entire test set, so we present both predictions in the right hand panel.

FIGURE 3 HERE

Figure 3: **Left panel: SOI data: plot of the target (realisation at time $n+1$) and its prediction by several methods: the last available curve (dashed red line); E-SNN prediction (dot-dashed black line); Bosq LA prediction (dotted blue line). Right panel: Electricity demand data: plot of the target (realisation at time $n+1$) and its prediction by several methods: the last available curve (dashed red line); E-SNN prediction (dot-dashed black line); E-FNN prediction (solid marked black line); Bosq LA prediction (dotted line).**

8. Proofs

The proof of the results is based on several steps. For the sake of clarity, we shall follow the same order as in Section 3.1 to establish all the relevant steps in the proofs. We shall then prove Lemma 1 which shows our proposed estimator of Section 1 satisfies Condition 4 under Conditions 6 and 3.

Properties of the estimator of $\left\{(Z_{i,s})_{i\in\{1,...,n\}}; s=1,...,S\right\}$ *based on the* $(\varphi_{s,n})_{s=1}^{S}$.

We have the following preliminary Lemma

Lemma 2. *Under Conditions 2 and 4, for any* $S \geq 1$,

$$\int \left| \sum_{s=1}^{S} \left(\widehat{\varphi}_s(v) - \varphi_s(v)\right) Z_{i,s} \right|^2 dv = O_p\left(\left(\frac{S}{\underline{\Lambda}_S^2 n^{2\tau}} \right)^{\frac{\nu}{\nu+2}} \right),$$

where $\underline{\Lambda}_S := \min_{s\leq S} \Lambda_s$ *with* Λ_s *as in Condition 4.*

Proof. Note that, by orthogonality of the eigenfunctions (Condition 4),

$$\begin{aligned} & \left\{ \int \left| \sum_{s=1}^{S} \left(\widehat{\varphi}_s(v) - \varphi_s(v)\right) Z_{i,s} \right|^2 dv > \epsilon \right\} \\ = & \left\{ \sum_{s=1}^{S} Z_{i,s}^2 \int \left(\widehat{\varphi}_s(v) - \varphi_s(v)\right)^2 dv > \epsilon \right\} \\ \subseteq & \left\{ M \max_{s\leq S} \int \left|\widehat{\varphi}_s(v)(v) - \varphi_s(v)\right|^2 dv > \epsilon \right\} \cup \left\{ \sum_{s=1}^{S} Z_{i,s}^2 > M \right\}. \end{aligned}$$

Hence,

$$\Pr\left(\int \left| \sum_{s=1}^{S} \left(\widehat{\varphi}_s(v) - \varphi_s(v)\right) Z_{i,s} \right|^2 dv > \epsilon \right) \leq MS\underline{\Lambda}_S^{-2} n^{-2\tau} + M^{-\nu/2}$$

where the first term follows by Condition 4 and the second by Condition 2. Equating the two terms, the result follows. □

The following lemma ensures the $\{(Z_{i,s})_{i\in\{1,...,n\}} : s = 1,\ldots,S\}$ based on the known eigenfunctions and the estimated ones are close.

Lemma 3. *Let* $Z_{i,sn} := \langle Y_i, \widehat{\varphi}_s\rangle$ *and* $Z_{i,s} := \langle Y_i, \varphi_s\rangle$. *Then, under Conditions 2 and 4, for any* $s \geq 1$, *and* $i \in \{1,\ldots,n\}$, $|Z_{i,sn} - Z_{i,s}| = O_p\left(\Lambda_s^{-1} n^{-\tau}\right)$.

Proof. By definition, $|Z_{i,sn} - Z_{i,s}| = |\langle Y_i, (\widehat{\varphi}_s - \varphi_s)\rangle| \leq |Y_i|\,|\widehat{\varphi}_s - \varphi_s|$, and the bound follows by Condition 4 because $|Y_i|$ is bounded in probability by Condition 2 □

Properties of the estimator for the transition distribution

This section establishes convergence of the estimator in (3.1).

Convergence of the marginals under estimation error

Recall that $Z_i(\psi) := \langle Y_i, \psi \rangle$, $\forall \psi \in \Psi_\delta$, as defined before Condition 5. For ease of notation, $Z_i := Z_i(\psi)$. For any $x \in \mathcal{H}$, and $\psi \in \Psi_\delta$, define

$$P_n(z|B,\psi) := \frac{\sum_{i=2}^n \mathbb{1}\{\langle \psi, Y_{i-1}\rangle \in B\left(\langle \psi, x\rangle, r\right), \langle \psi, Y_i\rangle \leq z\}}{\sum_{i=2}^n \mathbb{1}\{\langle \psi, Y_{i-1}\rangle \in B\left(\langle \psi, x\rangle, r\right)\}},$$

$$P(z|B,\psi) := \Pr\left(\langle Y_1, \psi\rangle \leq z | \langle Y_0, \psi\rangle \in B\left(\langle \psi, x\rangle, r\right)\right)$$

and,

$$P(z|x,\psi) := \Pr\left(\langle Y_1, \psi\rangle \leq z | \langle Y_0, \psi\rangle = \langle \psi, x\rangle\right)$$

when $r = 0$. With this notation we now prove Theorem 1.

Proof. [Theorem 1] By Condition 4 for large enough n, we do have $\widehat{\varphi}_s \in \Psi_{\delta,s}$ in probability for some $\delta = O\left(\Lambda_s^{-1} n^{-\tau}\right)$. By the triangle inequality,

$$\begin{aligned} & \sup_{z\in\mathbb{R}} |P_n(z|B,\widehat{\varphi}_s) - P(z|x,\varphi_s)| \\ \leq & \sup_{z\in\mathbb{R},\psi\in\Psi_\delta} |P_n(z|B,\psi) - P(z|B,\psi)| \\ & + \sup_{z\in\mathbb{R},\psi\in\Psi_\delta} |P(z|B,\psi) - P(z|x,\varphi_s)| \\ = & \mathrm{I} + \mathrm{II}. \end{aligned}$$

We shall control each term separately.

Control over I. An application of Lemma 5 shows that $\mathrm{I} \lesssim K\left(\langle x, \varphi_s\rangle\right) n^{-1/2} r^{-1}$, as long as $\delta = o\left(r\right)$.

Control over II. By Lemma 3, $|\varphi_s - \psi| \leq \delta = O\left(\Lambda_s^{-1} n^{-\tau}\right)$ implies there are δ' and δ'' of order $O\left(\Lambda_s^{-1} n^{-\tau}\right)$, such that $\langle Y_i - x, \varphi_s - \psi\rangle \leq \delta'$, $\langle Y_i, \varphi_s - \psi\rangle \leq \delta''$ (in probability). Moreover,

$$\begin{aligned} & \{\langle Y_0, \psi\rangle \in B\left(\langle x, \psi\rangle, r\right)\} \\ = & \{\langle x, \psi\rangle + \langle Y_0, \varphi_s - \psi\rangle - r \leq \langle Y_0, \varphi_s\rangle \leq \langle x, \psi\rangle + \langle Y_0, \varphi_s - \psi\rangle + r\} \\ = & \{\langle x, \varphi_s\rangle + \langle Y_0 - x, \varphi_s - \psi\rangle - r \leq \langle Y_0, \varphi_s\rangle \leq \langle x, \varphi_s\rangle + \langle Y_0 - x, \varphi_s - \psi\rangle + r\} \\ = & \{\langle Y_0, \varphi_s\rangle \in B\left(\langle x, \varphi_s\rangle + \langle Y_0 - x, \varphi_s - \psi\rangle, r\right)\} \end{aligned}$$

and, by the previous remarks,

$$B\left(\langle x, \varphi_s\rangle + \langle Y_0 - x, \varphi_s - \psi\rangle, r\right) \subseteq B\left(\langle x, \varphi_s\rangle, r + \delta'\right),$$

in probability. Hence,

$$\begin{aligned}
\text{II} &= \left|\Pr\left(\langle Y_1,\psi\rangle \le z \,|\, \langle Y_0,\psi\rangle \in B\left(\langle x,\psi\rangle\right)\right) - \Pr\left(\langle Y_1,\varphi_s\rangle \le z \,|\, \langle Y_0,\varphi_s\rangle = \langle x,\varphi_s\rangle\right)\right| \\
&\le \sup_{z\in\mathbb{R},x'\in B(\langle x,\varphi_s\rangle,r+\delta')} \left|\Pr\left(\langle Y_1,\varphi_s\rangle \le z+\delta'' \,|\, \langle Y_0,\varphi_s\rangle = x'\right) - \Pr\left(\langle Y_1,\varphi_s\rangle \le z+\delta'' \,|\, \langle Y_0,\varphi_s\rangle = \langle x,\varphi_s\rangle\right)\right| \\
&\quad + \sup_{z\in\mathbb{R}} \left|\Pr\left(\langle Y_1,\varphi_s\rangle \le z+\delta'' \,|\, \langle Y_0,\varphi_s\rangle = \langle x,\varphi_s\rangle\right) - \Pr\left(\langle Y_1,\varphi_s\rangle \le z \,|\, \langle Y_0,\varphi_s\rangle = \langle x,\varphi_s\rangle\right)\right| \\
&= O\left(r^{\beta}+\delta''\right)
\end{aligned}$$

by Condition 5 using the fact that $\delta' = o(r)$ by the condition in the Theorem. Putting everything together, we have

$$\begin{aligned}
\text{I}+\text{II} &= O\left(K\left(\langle x,\varphi_s\rangle\right) n^{-1/2} r^{-1} + r^{\beta} + \Lambda_s^{-1} n^{-\tau}\right) \\
&= O\left(K\left(\langle x,\varphi_s\rangle\right) n^{-\beta/(2\beta+2)}\right)
\end{aligned}$$

for $r \asymp n^{-1/(2\beta+2)}$ if $\Lambda_s^{-1} n^{-\tau} = o\left(n^{-1/(2\beta+2)}\right)$, because $\beta \in (0,1]$.

□

We now prove Theorem 2 (ii); the proof of Theorem 2 (i) follows trivially from this by Remark 1.

Proof. [Theorem 2] Let $g : \mathbb{R} \to \mathbb{R}$ be Lipschitz of order one in the sense that

$$\mathbb{E}|g(Y(v)) - g(Y'(v))| \le \mathbb{E}|Y(v) - Y'(v)|.$$

This implies that g is also of bounded variation hence, with no loss of generality, we can take g to be increasing. For each integer S,

$$\mathbb{E}\int \left|\sum_{s>S} Z_{i,s}\varphi_s(v)\right|^2 dv \le \sum_{s>S} \mathbb{E}Z_{i,s}^2 = \sum_{s>S} \lambda_s.$$

By Condition 3, $\lambda_s = o\left(s^{-(d+1)}\right)$ (e.g. Reade, 1984), implying $\sum_{s>S} \lambda_s = o\left(S^{-d}\ln S\right)$, and consequently,

$$\left|g\left(\sum_{s=1}^{\infty} Z_{i,s}\varphi_s(v)\right) - g\left(\sum_{s=1}^{S} Z_{i,s}\varphi_s(v)\right)\right| = O_p\left(S^{-d/2}\ln^{1/2} S\right)$$

for Lebesgue almost all v, because g is Lipschitz.

Define $g_M := g\{g \le M\}$ and $g_{M^c} := g\{g > M\}$ and note that

$$\mathbb{E}\left|g\left(\sum_{s=1}^{S} Z_{i,s}\varphi_s(v)\right)\right|^2 \le \mathbb{E}\left|\sum_{s=1}^{\infty} Z_{i,s}\varphi_s(v)\right|^2 = C_0(v,v) < \infty,$$

by the Lipschitz condition, (2.2) and Condition 3. By Lemma 1, and the above display we also have $\mathbb{E}\left|g\left(\sum_{s=1}^{S} Z_{i,s}\varphi_{sn}^{S}(v)\right)\right|^2 < \infty$. Then, $\Pr(g > M) \lesssim M^{-2}$, by Markov's inequality, whether we use the true or estimated eigenfunctions. Hence, $\mathbb{E}|g - g_M| = \mathbb{E}g_{M^c} \lesssim M^{-1}$, by Holder's inequality on $g_{M^c} = g\{g > M\}$, and we can use a finite dimensional version of Y and g_M in place of g (again, whether we use the true or estimated eigenfunctions). Moreover, by Lemma 2,

$$\left|g_M\left(\sum_{s=1}^{S}\varphi_s(v)Z_{i,s}\right) - g_M\left(\sum_{s=1}^{S}\widehat{\varphi}_s(v)Z_{i,s}\right)\right| = O_p\left(\left(\frac{S}{\underline{\Lambda}_S^2 n}\right)^{\frac{\nu}{\nu+2}}\right),$$

for Lebesgue almost all v because g_M is Lipschitz.

We show the results for $g_M\left(\sum_{s=1}^{S}\widehat{\varphi}_s(v)z_s\right)$. To this end, define

$$g_M^{s,n}(z_s, v) := \int g_M\left(\sum_{t=1}^{S}\widehat{\varphi}_t(v)z_t\right)\prod_{t<s}dP_{tn}\left(z_t|B_t(\langle x,\widehat{\varphi}_t\rangle)\right)\prod_{t>s}dP_t\left(z_t|\langle x,\varphi_t\rangle\right),$$

where the empty product is one. Note that for each v, $g_M^{s,n}(z_s, v)$ is random, but bounded and monotonic. With this proviso, for Lebesgue almost all v, consider the telescoping sum

$$\begin{aligned}
&\left|\int g_M\left(\sum_{s=1}^{S}\widehat{\varphi}_s(v)z_s\right)\prod_{s=1}^{S}d\left[P_{sn}\left(z_s|B_s\left(\langle x,\widehat{\varphi}_s\rangle\right)\right) - P_s\left(z_s|\langle x,\varphi_s\rangle\right)\right]\right| \\
= &\left|\int\sum_{s=1}^{S}g_M^{s,n}\left(z_s,\widehat{\varphi}_s(v)\right)d\left[P_{sn}\left(z_s|B_s\left(\langle x,\widehat{\varphi}_s\rangle\right)\right) - P_s\left(z_s|\langle x,\varphi_s\rangle\right)\right]\right| \\
\leq &MS\max_{s\leq S}\sup_{z_s\in\mathbb{R}}\left|P_{sn}\left(z_s|B_s\left(\langle x,\widehat{\varphi}_s\rangle\right)\right) - P_s\left(z_s|\langle x,\varphi_s\rangle\right)\right| \\
&\text{[by the properties of the Kolmogorov metric, e.g. Sancetta (2007, 2009)]} \\
= &O_p\left(K\left(\langle x,\varphi_s\rangle\right)MSn^{-\beta/(2\beta+2)}\right)
\end{aligned}$$

by Theorem 1, for r as in the statement of the theorem. Putting everything together, we have

$$\begin{aligned}
&\int g_M\left(\sum_{s=1}^{S}\widehat{\varphi}_s(v)z_s\right)\prod_{s=1}^{S}dP_{sn}\left(z_s|B_s(\langle x,\widehat{\varphi}_s\rangle)\right) \\
= &\int g\left(\sum_{s=1}^{\infty}\varphi_s(v)z_s\right)\prod_{s=1}^{S}dP_s\left(z_s|\langle x,\varphi_s\rangle\right) \\
&+O_p\left(K\left(\langle x,\varphi_s\rangle\right)MSn^{-\beta/(2\beta+2)} + \left(\frac{S}{\underline{\Lambda}_S^2 n^{2\tau}}\right)^{\frac{v}{v+2}} + M^{-1} + S^{-d/2}\ln^{1/2}S\right).
\end{aligned}$$

Letting $M \to \infty$ slowly enough, the error term clearly goes to zero under the conditions of the Theorem. □

We next prove Lemma 1. The proof uses the following re-phrasing of Lemma 4.3 in Bosq (2000).

Lemma 4. *Suppose C_0, C_1 are covariance functions from a Hilbert space $\mathcal{H}$ into itself. Let $\{\psi_{s0}; s \in \mathbb{N}\}$ and $\{\psi_{s1}; s \in \mathbb{N}\}$ be the respective eigenfunctions (their range is assumed to be a subset of $\mathbb{R}$). Then, for $s \in \mathbb{N}$, $|\psi_{s0} - \psi_{s1}| \leq \Lambda_s^{-1} |C_0 - C_1|_F$, where $\Lambda_s \asymp |\lambda_s - \lambda_{s-1}| + |\lambda_s - \lambda_{s+1}|$, $\Lambda_1 \asymp |\lambda_s - \lambda_{s+1}|$.*

As we will see, in the proof of Lemma 1 we need to control for the approximation error due to the truncated number of Fourier coefficients as well as for the estimation error. Recall that

$$C_n^S(u,v) := \sum_{|r|,|p| \leq S} c_n(r,p) \exp\{i(ru+pv)\},$$

the truncated Fourier representation of C_n. As before, we denote by $\left\{\varphi_{s,n}^S : s \in \mathbb{N}\right\}$ the set of eigenfunctions of C_n^S.

Proof. [Lemma 1] By the triangle inequality

$$\left|\varphi_{s,n}^S - \varphi_s\right| \leq \left|\varphi_{s,n}^S - \varphi_s^S\right| + \left|\varphi_s^S - \varphi_s\right| =: \mathrm{I} + \mathrm{II},$$

where φ_s^S is the s^{th} eigenfunction of $C_0^S := \sum_{|r|,|p| \leq S} c(r,p) \exp\{i(ru+sv)\}$ with obvious notation using the Fourier representation in (4.3) for C_0. We shall apply Lemma 4 to the estimation error I and the approximation error II. To bound I, note that, using the Fourier basis functions as a common basis for both C_n and C_0,

$$\int_{\mathcal{V}} \int_{\mathcal{V}} |C_n(u,v) - C_0(u,v)|^2 \, dudv = \sum_{r,p} |c_n(r,p) - c(r,p)|^2 = O_p\left(n^{-2\tau}\right)$$

by Condition 6, where the first equality follows by orthonormality of the Fourier basis. By positivity of the elements in the sum

$$\sum_{r,p} |c_n(r,p) - c(r,p)|^2 \geq \sum_{|r|,|p| \leq S} |c_n(r,p) - c(r,p)|^2 = \int_{\mathcal{V}} \int_{\mathcal{V}} \left|C_n^S(u,v) - C_0^S(u,v)\right|^2 dudv$$

so that, by Lemma 4, $\mathrm{I} = O_p\left(\Lambda_s^{-1} n^{-1/2}\right)$. To bound II, we need a bound on the following:

$$\int_{\mathcal{V}} \int_{\mathcal{V}} \left|C_0^S(u,v) - C_0(u,v)\right|^2 dudv \quad = \quad \left(\sum_{|r|,|p|>S} + 2 \sum_{r \in \mathbb{Z}} \sum_{|p|>S}\right) |c(r,p)|^2$$

by orthogonality of trigonometric polynomials and symmetry of c. By Condition 3, for any non negative integers k and l with $k + l \leq d$ (with d as in Condition 3), the Fourier coefficients of C_0 satisfy

$$\sum_{|r|,|p| \leq S} \left|r^k p^l c(r,p)\right|^2 < \infty, \tag{8.1}$$

using the relationship between the Fourier coefficients of C_0 and the ones of its derivatives. Hence,

$$\left(\sum_{|r|,|p|>S} + 2 \sum_{r \in \mathbb{Z}} \sum_{|p|>S}\right) |c(r,p)|^2 \leq 3 \sum_{r \in \mathbb{Z}} \sum_{|p|>S} |c(r,p)|^2 \lesssim S^{-2d},$$

by (8.1) with $k = 0$ and $l = d$. Hence, applying Lemma 4, solve for $S^{-2d} = n^{-2\tau}$ to infer $S \gtrsim n^{\tau/d}$, if we want $\mathrm{II} = O\left(\Lambda_s^{-1} n^{-\tau}\right)$. Finally, we note that the estimated eigenfunctions are uniformly bounded because they are based on trigonometric polynomials (which are bounded) with square coefficients that need to sum up to one, as the eigenfunctions are orthonormal. □

8.1. Supplementary lemmata

Using the same notation as in the proof of Theorem 1 we have the following.

Lemma 5. *For $\delta = o(r)$ and $r > 0$, under Conditions 1 and 5,*

$$\sup_{z\in\mathbb{R},\psi\in\Psi_\delta} |P_n(z|B,\psi) - P(z|B,\psi)| \lesssim \frac{K(\langle x,\varphi\rangle)}{n^{1/2}r}$$

in probability, where $K(\langle x,\varphi\rangle)$ is as in (8.2) and for ease of notation we have dropped the subscript s, e.g. $\varphi = \varphi_s$.

Proof. To ease notation we may suppress the dependence on ψ when not needed, hence $B := B(\langle x,\psi\rangle)$ and $Z_i := Z_i(\psi)$ when explicit dependence on ψ and x is not used. As in Caires and Ferreira (2005), note that

$$\begin{aligned} P_n(z|B,\psi) - P(z|B,\psi) &= \frac{\sum_{i=1}^n (1-\mathbb{E})\mathbb{I}\{Z_{i-1}\in B, Z_i \le z\}}{\sum_{i=1}^n \mathbb{E}\mathbb{I}\{Z_{i-1}\in B\}} \\ &\quad - \left[\frac{\sum_{i=1}^n (1-\mathbb{E})\mathbb{I}\{Z_{i-1}\in B, Z_i \le z\}}{\sum_{i=1}^n \mathbb{E}\mathbb{I}\{Z_{i-1}\in B\}} + P(z|B,\psi)\right] \\ &\quad \times \left[\frac{\sum_{i=1}^n \mathbb{I}\{Z_{i-1}\in B\}}{\sum_{i=1}^n \mathbb{E}\mathbb{I}\{Z_{i-1}\in B\}} - 1\right] \frac{\sum_{i=1}^n \mathbb{E}\mathbb{I}\{Z_{i-1}\in B\}}{\sum_{i=1}^n \mathbb{I}\{Z_{i-1}\in B\}} \end{aligned}$$

To avoid trivialities in the notation, we used the summation $\sum_{i=1}^n$ rather than $\sum_{i=2}^n$. Define

$$\Delta_1 := \frac{\sum_{i=1}^n (1-\mathbb{E})\mathbb{I}\{Z_{i-1}\in B, Z_i \le z\}}{\sum_{i=1}^n \mathbb{E}\mathbb{I}\{Z_{i-1}\in B\}}$$

and

$$\Delta_2 := \frac{\sum_{i=1}^n \mathbb{I}\{Z_{i-1}\in B\}}{\sum_{i=1}^n \mathbb{E}\mathbb{I}\{Z_{i-1}\in B\}} - 1,$$

so that

$$|P_n(z|B,\psi) - P(z|B,\psi)| \le |\Delta_1| + |\Delta_1 + P(z|B,\psi)| \left|\frac{\Delta_2}{1+\Delta_2}\right| =: \mathrm{I} + \mathrm{II}.$$

The Lemma is proved if, in probability, we can find a uniform bound in $z\in\mathbb{R}$ and $\psi\in\Psi_\delta$ for the above display. By (8.2) we will replace $\mathbb{E}\{Z_0\in B\}$ with its minorand r/K.

Control over I. We show that, uniformly in $z\in\mathbb{R}$ and $\psi\in\Psi_\delta$, $\mathrm{I} = O_p\left(K(\langle x,\varphi\rangle)/\left(rn^{1/2}\right)\right)$. Hence,

$$\begin{aligned} &\Pr\left(\sup_{z\in\mathbb{R},\psi\in\Psi_\delta} \left|\frac{\sum_{i=1}^n (1-\mathbb{E})\mathbb{I}\{Z_{i-1}(\psi)\in B(\langle x,\psi\rangle), Z_i(\psi)\le z\}}{\sum_{i=1}^n \mathbb{E}\mathbb{I}\{Z_{i-1}(\psi)\in B(\langle x,\psi\rangle)\}}\right| > \frac{\epsilon K(\langle x,\varphi\rangle)}{rn^{1/2}}\right) \\ &\le \Pr\left(\sup_{z\in\mathbb{R},\psi\in\Psi_\delta} \left|\frac{\sum_{i=1}^n (1-\mathbb{E})\mathbb{I}\{Z_{i-1}(\psi)\in B(\langle x,\psi\rangle), Z_i(\psi)\le z\}}{n^{1/2}}\right| > \epsilon\right) \end{aligned}$$

by Lemma 6 and we shall control the above probability. For reasons of technical nature, we are going to complete the proof for the set $\{Z_{i-1}\in B, Z_i > z\}$ rather than$\{Z_{i-1}\in B, Z_i \le z\}$. Note that

$$\{Z_{i-1}\in B(\langle x,\psi\rangle), Z_i > z\} = \{Z_{i-1} > (\langle x,\psi\rangle - r), Z_i > z\} - \{Z_{i-1} > (\langle x,\psi\rangle + r), Z_i > z\}$$

so that it is enough to focus on $\{Z_{i-1} > (\langle x, \psi\rangle - r), Z_i > z\}$. Hence, let P_ψ be the marginal distribution of $Z_i := Z_i(\psi)$ ($\forall i$ by stationarity) and define $U_i := P_\psi(Z_i(\psi))$. Then,

$$\{Z_{i-1} > (\langle x, \psi\rangle - r), Z_i > z\} = \{U_{i-1} > P_\psi(\langle x, \psi\rangle - r), U_i > P_\psi(z)\},$$

so that

$$\begin{aligned} & \sup_{z\in\mathbb{R},\psi\in\Psi_\delta} \left| \frac{\sum_{i=1}^n (1-\mathbb{E})\mathbb{I}\{Z_{i-1}(\psi) \in B(\langle x,\psi\rangle), Z_i(\psi) > z\}}{n^{1/2}} \right| \\ \lesssim & \sup_{u,v\in[0,1]} \left| \frac{\sum_{i=1}^n (1-\mathbb{E})\mathbb{I}\{U_{i-1} > u, U_i > v\}}{n^{1/2}} \right|. \end{aligned}$$

Note that by construction U_i is $[0,1]$ uniform because P_ψ is continuous for every $\psi \in \Psi_\delta$. The r.h.s. is bounded if we show equicontinuity in probability. This follows from Proposition 7.3 in Rio (2000). In fact, by the previous remarks on Sklar's Theorem and the remarks in Section 5, the $(U_i)_{i\in\{1,\dots,n\}}$ are α-mixing with decaying rate as prescribed by Proposition 7.3 of Rio (2000).

Control over II. Define $\epsilon' := \epsilon K(\langle x, \varphi\rangle) / (rn^{1/2})$ and note that, making explicit the dependence of Δ_1 and Δ_2 on ψ,

$$\begin{aligned} & \left\{ \sup_{\psi\in\Psi_\delta} |\Delta_1(\psi) + P(z|B,\psi)| \left| \frac{\Delta_2(\psi)}{1+\Delta_2(\psi)} \right| > \epsilon' \right\} \\ \subseteq & \left\{ 2 \sup_{\psi\in\Psi_\delta} \left| \frac{\Delta_2(\psi)}{1+\Delta_2(\psi)} \right| > \epsilon' \right\} \\ & [\text{because } |\Delta_1(\psi) + P(z|B,\psi)| < 2] \\ \subseteq & \left[\left\{ 2 \sup_{\psi\in\Psi_\delta} \left| \frac{\Delta_2(\psi)}{1+\Delta_2(\psi)} \right| > \epsilon', \sup_{\psi\in\Psi_\delta} |\Delta_2(\psi)| < \frac{1}{2} \right\} \cup \left\{ 2 \sup_{\psi\in\Psi_\delta} \left| \frac{\Delta_2}{1+\Delta_2} \right| > \epsilon', \sup_{\psi\in\Psi_\delta} |\Delta_2(\psi)| > \frac{1}{2} \right\} \right] \\ \subseteq & \left[\left\{ \sup_{\psi\in\Psi_\delta} |\Delta_2(\psi)| > \frac{\epsilon'}{4} \right\} \cup \left\{ \sup_{\psi\in\Psi_\delta} |\Delta_2(\psi)| > \frac{1}{2} \right\} \right]. \end{aligned}$$

Hence, for $\epsilon' := \epsilon K(\langle x, \varphi\rangle) / (rn^{1/2}) \leq 2$,

$$\begin{aligned} \Pr\left(\sup_{\psi\in\Psi_\delta} \left| \frac{\Delta_2(\psi)}{1+\Delta_2(\psi)} \right| > \epsilon' \right) & \leq 2\Pr\left(\sup_{\psi\in\Psi_\delta} |\Delta_2(\psi)| > \frac{\epsilon'}{4} \right). \\ & \leq 2\Pr\left(\sup_{\psi\in\Psi_\delta} \left| \frac{\sum_{i=1}^n (1-\mathbb{E})\mathbb{I}\{Z_{i-1}(\psi) \in B(\langle x,\psi\rangle)\}}{rn/K(\langle x,\psi\rangle)} \right| > \frac{\epsilon'}{4n^{1/2}} \right) \\ & \quad [\text{by (8.2)}] \\ & \lesssim 2\Pr\left(\sup_{\psi\in\Psi_\delta} \left| \frac{\sum_{i=1}^n (1-\mathbb{E})\mathbb{I}\{Z_{i-1}(\psi) \in B(\langle x,\psi\rangle)\}}{n^{1/2}} \right| > \frac{\epsilon}{4} \right) \end{aligned}$$

by definition of ϵ' and Lemma 6. Using the same arguments as in the Control over I, we deduce that $\text{II} = O_p(K(\langle x, \varphi\rangle) / (rn^{1/2}))$. □

Lemma 6. *Let $\delta \to 0$ such that $\delta = o(r)$. Then, eventually, for $r \to 0$*

$$\inf_{\psi \in \Psi_\delta} \sum_{i=1}^{n} \mathbb{E}\mathbb{I}\{Z_{i-1}(\psi) \in B(\langle x, \psi \rangle)\} \asymp nr/K(\langle x, \varphi \rangle),$$

where $K := K(x_s)$ is a finite constant depending on x_s only, such that, for any s,

$$|B(x_s, r)| \leq K(x_s)\, \mathbb{E}\{Z_{0s} \in B(x_s, r)\}. \tag{8.2}$$

Proof. Under the conditions of the lemma $(\delta/r)|Y_0| = o_p(1)$ because $|Y_0|$ has a tight measure, by integrability. Hence, $r - \delta|Y_1| \asymp r$ in probability, implying

$$\begin{aligned} \inf_{\psi \in \Psi_\delta} \Pr(\langle Y_0, \psi \rangle \in [\langle \psi, x \rangle - r, \langle \psi, x \rangle + r]) &\asymp \Pr(\langle Y_0, \varphi \rangle \in [\langle \varphi, x \rangle - r, \langle \varphi, x \rangle + r]) \\ &\geq \frac{|r|}{K(\langle \varphi, x \rangle)} \end{aligned}$$

by (8.2). $K(x_s)$ can be taken as an upper bound for the Radon-Nykodym derivative of the Lebesgue measure with respect to the invariant measure of $(Z_{i,s})_{i \in \{1,\ldots,n\}}$ (see the proof of Lemma 2.2 in Devroye, 1981). This is clearly finite almost surely. □

Acknowledgement: We thank Rudolf Beran for pointing out the reference of Cohen and Jones (1969). We also thank Peter Green and the anonymous referees for comments and questions that greatly enhanced the quality and clarity of this work. HB gratefully acknowledges financial support from the ESRC and the EPSRC.

[1] Adler, R. J. and Taylor, J. E. (2007) *Random Fields and Geometry*. Springer Monographs in Mathematics. Springer, New York.

[2] Babillot, M., Bougerol, P., and Elie, L. (1997) The random difference equation $X_n = A_n X_{n-1} + B_n$ in the critical case. *Ann. Probab.*, 25(1):478–493.

[3] Basrak, B., Davis, R.A., and Mikosch, T. (2002) Regular Variation of GARCH Processes, Stochastic Processes and their Application 99, 95-115.

[4] Bathia, N., Yao, Q., and Ziegelmann, F. (2010) Identifying the finite dimensionality of curve time series. *Ann. Statist.*, 38:3352–3386.

[5] Besse, P. C. and H. Cardot (1996). Approximation Spline de la Prévision d'un Processus Fonctionnel Autoregressif d'Ordre 1. *Canad. J. Statist.* 24 467-487.

[6] Besse, P. C., Cardot, H. and D. B. Stephenson (2000). Autoregressive Forecasting of Some Functional Climatic Variations. *Scand. J. Statist.* 27 673-687.

[7] Bradley, R.C. (2005) Basic Properties of Strong Mixing Conditions. A survey and some open questions. *Probability Surveys* 2, 107–144.

[8] Bosq, D. *Linear processes on function spaces: theory and applications*. Springer-Verlag, New York, 2000.

[9] Bosq, D. and D. Blanke *Inference and prediction in large dimensions.* John Wiley & Sons Ltd., Chichester, 2007.

[10] Caires, S. and Ferreira, J. A. (2005) On the non-parametric prediction of conditionally stationary sequences. *Stat. Inference Stoch. Process.*, 8(2):151–184.

[11] Cavallini, A., Montanari, G. C., Loggini, M., Lessi, O. and M. Cacciari (1994) Nonparametric Prediction of Harmonic Levels in Electrical Networks. *Proceedings of IEEE ICHPS VI.* Bologna. 165-171.

[12] Cohen, A., Jones, R. H.(1969) Regression on a random field *J. Amer. Statist. Assoc.* 1172-1182.

[13] Damon, J. and S. Guillas (2002) The Inclusion of Exogenous Variables in Functional Autoregressive Ozone Forecasting. *Environmetrics*, 13 759-774.

[14] Devroye, L. (1981) On the almost everywhere convergence of nonparametric regression function estimates. *Ann. Statist.*, 9(6):1310–1319.

[15] Didericksen, D., Kokoszka, P., Zhang, X. (2012) Empirical properties of forecasts with the functional autoregressive model. *Comput. Statist.*

[16] Esary, J.D., F. Proschan, and D. W. Walkup (1967) Association of Random Variables, with Applications. *Annals of Mathematical Statistics* 38, 1466-1474.

[17] Ferraty, F. and P. Vieu *Nonparametric functional data analysis.* Springer, New York, 2006

[18] Friedman, Jerome H. (2001) Greedy function approximation: a gradient boosting machine. *Ann. Statist.*, 29(5): 1189–1232

[19] Hall, P., Müller, H.-G. and Wang, J.-J. (2006) Properties of principal component methods for functional and longitudinal data analysis. *Ann. Statist.* 34 (3) 1493–1517.

[20] Hormann, S. and Kokoszka, P. (2010) Weakly dependent functional data. *Ann. Statist.*, 38:1845–1884.

[21] Horvarth, L., Kokoszka, P. and Reeder, R. (2012) Estimation of the mean of functional time series and a two sample problem. *J. R. Statist. Soc. B*, 74(5).

[22] Joe, H. (1997) *Multivariate Models and Dependence Concepts.* London, Chapman and Hall.

[23] Kargin, V. and Onatski, A. (2008) Curve forecasting by functional autoregression. *J. Multivariate Anal.*, 99(10):2508–2526.

[24] Li, Y. and Hsing, T. (2010) Uniform convergence rates for nonparametric regression and principal component analysis in functional/longitudinal data. *Ann. Statist.*, 38, 3321–3351.

[25] Lin, N. X. and Carroll, L. R. J . (2000) Nonparametric function estimation for clustered data when the predictor is measured without/with error. *J. Amer. Statist. Assoc.* 95, 520–534.

[26] Linton, O. and Sancetta, A. (2009) Consistent Estimation of a General Nonparametric Regression Function in Time Series. *J. Econometrics* 152, 70–78.

[27] Mokkadem, A. (1988) Mixing Properties of ARMA Processes. Stochastic Processes and their Applications 29, 309-315.

[28] Ramsay, J. O. and Silverman, B. W. (2002) *Applied Functional Data Analysis*. Springer Series in Statistics. Springer-Verlag, New York. Methods and case studies.

[29] Ramsay, J. O. and Silverman, B. W. (2002) *Functional Data Analysis*. Springer Series in Statistics. Springer-Verlag, New York.

[30] Rasmussen C.E. and C.K.I. Williams (2006) *Gaussian Processes for Machine Learning*. Cambridge: MIT Press.

[31] Reade, J.B. (1984) Eigenvalues of Positive Definite Kernels II. *SIAM J. Math. Analysis*, 15, 137-142.

[32] Rio, E. (2000) *Théorie Asymptotique des Processus Aléatoires Faiblement Dépendants*. Springer, Paris.

[33] Sancetta, A. (2007) Weak Convergence of Laws on $\mathbb{R}^K$ with Common Marginals. *Journal of Theoretical Probability*, 20, 371–380.

[34] Sancetta, A. (2009) Nearest Neighbor Conditional Estimation for Harris Recurrent Markov Chains. *J. Multivar. Analysis*, 100, 2224–2236.

[35] Silverman, B.W. (1996) Smoothed Functional Principal Components Analysis by Choice of Norm. *Ann. Statist.*, 24, 1-24.

[36] Van der Vaart, A. W. and Wellner, J. A. (1996) *Weak Convergence and Empirical Processes with Applications to Statistics*. Springer Series in Statistics. Springer-Verlag, New York.

[37] Yao, F. , Muller, H.-G. and Wang, J.-L. (2005a) Functional data analysis for sparse longitudinal data. *J. Amer. Statist. Assoc.*, 100, 577–590.

[38] Yao, F. , Muller, H.-G. and Wang, J.-L. (2005b) Functional linear regression analysis for longitudinal data. *Ann. Statist.*, 33, 2873–2903.

Figure

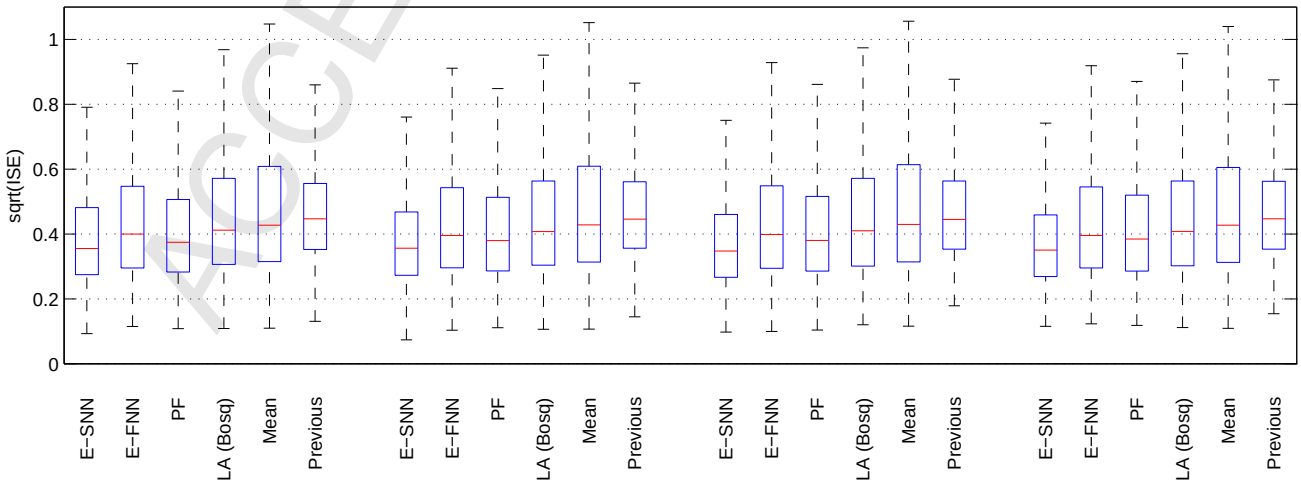


Figure2

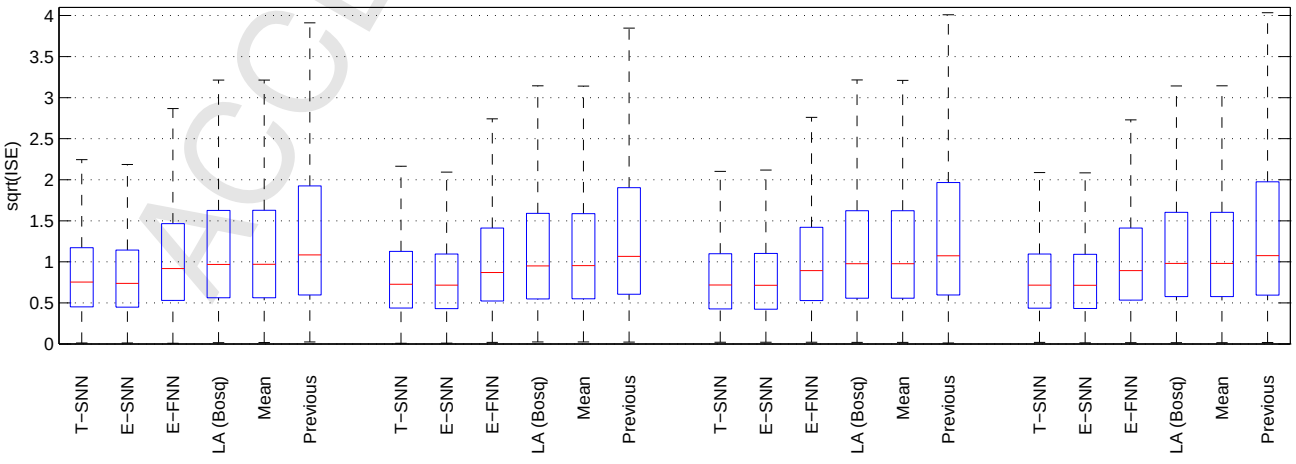


Figure3