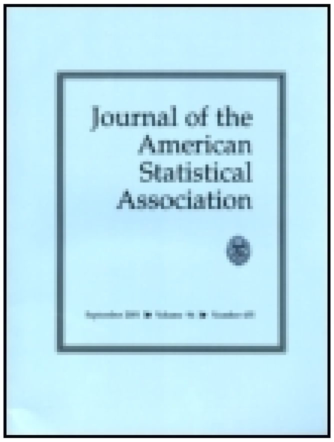

Journal of the American Statistical Association

Publication details, including instructions for authors and subscription information:
http://amstat.tandfonline.com/loi/uasa20

On the prediction of stationary functional time series

Alexander Aue, Diogo Dubart Norinho & Siegfried Hörmann

Accepted author version posted online: 18 Apr 2014.

On the prediction of stationary functional time series

Alexander Aue* Diogo Dubart Norinho Siegfried Hörmann

Abstract

This paper addresses the prediction of stationary functional time series. Existing contributions to this problem have largely focused on the special case of first-order functional autoregressive processes because of their technical tractability and the current lack of advanced functional time series methodology. It is shown here how standard multivariate prediction techniques can be utilized in this context. The connection between functional and multivariate predictions is made precise for the important case of vector and functional autoregressions. The proposed method is easy to implement, making use of existing statistical software packages, and may therefore be attractive to a broader, possibly non-academic, audience. Its practical applicability is enhanced through the introduction of a novel functional final prediction error model selection criterion that allows for an automatic determination of the lag structure and the dimensionality of the model. The usefulness of the proposed methodology is demonstrated in a simulation study and an application to environmental data, namely the prediction of daily pollution curves describing the concentration of particulate matter in ambient air. It is found that the proposed prediction method often significantly outperforms existing methods.

Keywords: Dimension reduction; Final prediction error, Forecasting, Functional autoregressions; Functional principal components, Functional time series; Particulate matter, Vector autoregressions

MSC 2010: Primary 62M10, 62M20; Secondary 62P12, 60G25

*Corresponding author

ACCEPTED MANUSCRIPT

1 Introduction

Functional data are often collected in sequential form. The common situation is a continuous-time record that can be separated into natural consecutive time intervals, such as days, for which a reasonably similar behavior is expected. Typical examples include the daily price and return curves of financial transactions data and the daily patterns of geophysical, meteorological and environmental data. The resulting functions may be described by a time series $(Y_k\colon k \in \mathbb{Z})$, each term in the sequence being a (random) function $Y_k(t)$ defined for t taking values in some interval $[a, b]$. Here, $\mathbb{Z}$ denotes the set of integers. The object $(Y_k\colon k \in \mathbb{Z})$ will be referred to as a functional time series (see Hörmann & Kokoszka (2012) for a recent survey on time series aspects, and Ferraty & Vieu (2010) and Ramsay & Silverman (2005) for general introductions to functional data analysis). Interest for this paper is in the functional modeling of concentration of particulate matter with an aerodynamic diameter of less than $10\mu m$ in ambient air, measured half-hourly in Graz, Austria. It is widely accepted that exposure to high concentrations can cause respiratory and related health problems. Local policy makers therefore monitor these pollutants closely. The prediction of concentration levels is then a particularly important tool for judging whether measures, such as partial traffic regulation, have to be implemented in order to meet standards set by the European Union.

Providing reliable predictions for future realizations is in fact one of the most important goals of any time series analysis. In the univariate and multivariate framework, this is often achieved by setting up general prediction equations that can be solved recursively by methods such as the Durbin-Levinson and innovations algorithms (see, for example, Brockwell & Davis (1991), Shumway & Stoffer (2011)). Prediction equations may be derived explicitly also for general stationary functional time series (see Section 1.6 of the monograph Bosq (2000)) but they seem difficult to solve and implement. As a consequence, much of the research in the area has focused on the first-order functional autoregressive model, shortly FAR(1). Bosq (2000) has derived one-step ahead predictors that are based on a functional form of the Yule-Walker equations. Besse et al. (2000) have proposed nonparametric kernel predictors and illustrated their methodology by forecasting climatological cycles caused by the El Niño phenomenon. While this paper, and also Besse & Cardot (1996), have adapted classical spline smoothing techniques, Antoniadis & Sapatinas (2003), see also Antoniadis et al. (2006, 2009), have studied FAR(1) curve prediction based on linear wavelet methods. Kargin & Onatski (2008) have introduced the predictive factor method, which seeks to replace functional principal components with directions most relevant for predictions. Didericksen et al. (2012) have evaluated several competing prediction models in a comparative simulation study, finding Bosq's (2000) method to have the best overall performance. Other contributions to the area are Aneiros-Pérez et al. (2010), and Aneiros-Pérez & Vieu (2008).

In spite of its statistical relevance and its mathematical appeal, functional time series modeling has still some unpleasant limitations for the practitioner. First, to date there are not many "ready to use" statistical software packages that can be utilized directly for estimation and prediction purposes. The only available packages that the authors are aware of are the `far` package of Damon & Guillas (2010) and the `ftsa` package of Hyndman & Shang (2012), both implemented for the statistical software `R`. The lack of tailor-made procedures often requires manual implementation. This may be challenging and therefore restrict use of the methodology to an academic audience. Second, the methodology developed for the FAR(1) case is difficult to generalize. If an FAR(1) approach is infeasible, one can use the multiple testing procedure of Kokoszka & Reimherr (2013) to determine an appropriate order p for a more general FAR(p) process. In addition, exogenous predictors can be incorporated using the work of Damon & Guillas (2002). These authors include exogenous covariates of FAR(1) type into a first-order autoregressive framework for functional ozone predictions. For more general cases functional theory and estimation have not yet been developed.

The goal of this paper is then to fill in this gap by promoting a simple alternative prediction algorithm which consists of three basic steps, all of which are easy to implement by means of existing software. First, use functional principal components analysis, FPCA, to transform the functional time series observations $Y_1, \ldots, Y_n$ into a vector time series of FPCA scores $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n$ of dimension d, where d is small compared to n. Second, fit a vector time series to the FPCA scores and obtain the predictor $\hat{\boldsymbol{Y}}_{n+1}$ for $\boldsymbol{Y}_{n+1}$. Third, utilize the Karhunen-Loève expansion to re-transform $\hat{\boldsymbol{Y}}_{n+1}$ into a curve predictor $\hat{Y}_{n+1}$. The first and the third step are simple and can be performed, for example, with the `fda` package in `R`. The second step may be tackled with standard multivariate time series methodology. Details are developed in Section 2. While the proposed approach is conceptually quite easy, several non-trivial questions need to be raised:

1. How does the resulting method differ from existing ones?
2. Why is this method justified from a theoretical standpoint?
3. In order to minimize the prediction error, how can the number of principal components in the dimension reduction for Step 1 be determined and how should model selection be performed in Step 2? Preferably, both choices should be made simultaneously.

These issues will be addressed in Section 3. In particular, a comparison to Bosq's (2000) classical benchmark FAR(p) prediction is made. A theoretical bound for the prediction error of the proposed methodology is established, which will imply asymptotic consistency. In Section 3.4 a novel functional final prediction error criterion is developed that jointly selects the order p and the dimensionality d of the FPC score vectors, thereby allowing for an automatic prediction process.

Functional principal components have been employed in other approaches to functional prediction, for example in Bosq's (2000) FAR(1) prediction method and in Aguilera et al. (1999). Roughly speaking, these and many other existing approaches have in common that Y_k is regressed onto the lagged observation Y_{k-1} by minimizing $E[\int [Y_k(t) - \Psi(Y_{k-1})(t)]^2 dt]$ with respect to a linear operator Ψ. The solution of this problem involves an infinite series representation of Ψ along FPCs. (More details will be given in Section 3.1.) In contrast, the proposed approach first uses dimension reduction via FPCA and then fits a model to the reduced data. No a priori knowledge of the functional model is needed and instead of a single estimator, a variety of existing tools for vector processes can be entertained. Further lags or exogenous covariates are also easily included into the prediction algorithm (see Section 4).

Hyndman & Ullah (2007) and Hyndman & Shang (2009) have suggested a curve prediction approach based on modeling FPC scores by scalar time series. They argue that scores are uncorrelated and that hence individual time series can be fit. Depending on the structure of the data, this can be quick and efficient in some cases but less accurate in other cases. The fact that FPC score vectors have no instantaneous correlation, does not imply that autocovariances at lags greater than zero remain diagonal. Hence univariate modeling may invoke a loss of valuable information hidden in the dependence of the data. This will be demonstrated in Section 6 as part of a simulation study. This issue can be avoided if one makes use of so-called *dynamic functional principal components* recently introduced in Hörmann et al. (2013) and Panaretos & Tavakoli (2013). These authors propose a methodology which produces score vectors with diagonal autocovariances via time invariant functional linear filters. Since the involved filters are two-sided (they require past and future observations), it is not clear how this methodology could be used for prediction.

It should be noted that, in this article, the data Y_k are assumed to be given in functional form, since the focus is on working out functional prediction methodology without getting into aspects of data preprocessing, which appears to be rather specific to the particular data at hand and therefore not conducive to a unified treatment. In practice, however, one observes only vectors $Y_k(t_1), \ldots, Y_k(t_L)$, with spacings, $t_\ell - t_{\ell-1}$, and number of intraday sampling points, L, potentially varying from day to day. The problem of transforming the vector observations into (smooth) functions has been treated in many articles and will not be detailed here. As an excellent starting point for reading in this direction the reader is referred to Chapters 3–7 of Ramsay & Silverman (2005). It is expected that the comparative results established in this paper as part of simulations and the application will hold also if the functions are not sampled equidistantly, with the rate of improvement of the proposed method over its competitors being of similar magnitude.

The remainder of the paper contains some possible extensions of the new prediction methodology in Section 5, a supporting simulation study in Section 6 and an application to the prediction of intraday patterns of particulate matter concentrations in Section 7. Section 8 concludes and technical proofs are given in Appendix A.

2 Methodology

In what follows, let $(Y_k \colon k \in \mathbb{Z})$ be an arbitrary stationary functional time series. It is assumed that the observations Y_k are elements of the Hilbert space $H = L^2([0,1])$ equipped with the inner product $\langle x, y\rangle = \int_0^1 x(t)y(t)dt$. Each Y_k is therefore a square integrable function satisfying $\|Y_k\|^2 = \int_0^1 Y_k^2(t)dt < \infty$. All random functions are defined on some common probability space $(\Omega, \mathcal{A}, P)$. The notation $Y \in L_H^p = L_H^p(\Omega, \mathcal{A}, P)$ is used to indicate that, for some $p > 0$, $E[\|Y\|^p] < \infty$. Any $Y \in L_H^1$ possesses then a mean curve $\mu = (E[Y(t)] \colon t \in [0,1])$, and any $Y \in L_H^2$ a covariance operator C, defined by $C(x) = E[\langle Y - \mu, x\rangle(Y - \mu)]$. The operator C is a kernel operator given by

$$C(x)(t) = \int_0^1 c(t,s)x(s)ds, \qquad c(t,s) = \mathrm{Cov}(Y(t), Y(s)).$$

As in the multivariate case, C admits the spectral decomposition

$$C(x) = \sum_{\ell=1}^{\infty} \lambda_\ell \langle v_\ell, x\rangle v_\ell,$$

where $(\lambda_\ell \colon \ell \in \mathbb{N})$ are the eigenvalues (in strictly descending order) and $(v_\ell \colon \ell \in \mathbb{N})$ the corresponding normalized eigenfunctions, so that $C(v_\ell) = \lambda_\ell v_\ell$ and $\|v_\ell\| = 1$. Here, $\mathbb{N}$ is the set of positive integers. The $(v_\ell \colon \ell \in \mathbb{N})$ form an orthonormal basis of $L^2([0,1])$. Hence Y allows for the Karhunen-Loéve representation $Y = \sum_{\ell=1}^{\infty} \langle Y, v_\ell\rangle v_\ell$. The coefficients $\langle Y, v_\ell\rangle$ in this expansion are called the FPC scores of Y.

Suppose now that we have observed $Y_1, \ldots, Y_n$. In practice μ as well as C and its spectral decomposition will be unknown and need to be estimated from the sample. We estimate μ by

$$\hat{\mu}_n(t) = \frac{1}{n}\sum_{k=1}^{n} Y_k(t), \qquad t \in [0,1],$$

and the covariance operator by

$$\hat{C}_n(x) = \frac{1}{n}\sum_{k=1}^{n}\langle Y_k - \hat{\mu}_n, x\rangle(Y_k - \hat{\mu}_n).$$

Under rather general weak dependence assumptions these estimators are $\sqrt{n}$-consistent. One may, for example, use the concept of L^p-m-approximability introduced in Hörmann & Kokoszka (2010) to prove that $E[\|\hat{\mu}_n - \mu\|^2] = O(n^{-1})$ and $E[\|\hat{C}_n - C\|_{\mathcal{L}}^2] = O(n^{-1})$, where the operator norm $\|\cdot\|_{\mathcal{L}}$ is, for any operator A, defined by

$$\|A\|_{\mathcal{L}} = \sup_{\|x\|\le 1}\|A(x)\|.$$

It is shown in Lemma A.1 of the Appendix that the general results (see Theorems 5 and 6 of Hörmann & Kokoszka (2012)) apply to the functional autoregressive processes studied in this paper. From $\hat{C}_n$, estimated eigenvalues $\hat{\lambda}_{1,n}, \ldots, \hat{\lambda}_{d,n}$ and estimated eigenfunctions $\hat{v}_{1,n}, \ldots, \hat{v}_{d,n}$ can be computed for an arbitrary fixed, but typically small, $d < n$. These estimators inherit $\sqrt{n}$-consistency from $\hat{C}_n$. See Theorem 7 in Hörmann & Kokoszka (2012). For notational convenience, $\hat{\lambda}_\ell$ and $\hat{v}_\ell$ will be used in place of $\hat{\lambda}_{\ell,n}$ and $\hat{v}_{\ell,n}$.

Functional linear prediction equations for general stationary processes have been derived in Section 1.6 of the monograph Bosq (2000). They appear to be impractical for actual data analysis as there do not seem to be either articles discussing applications to real life examples or contributions concerned with further foundational elaboration. As pointed out in the introduction, the notable exception is the FAR(1) process defined by the stochastic recursion

$$Y_k - \mu = \Psi(Y_{k-1} - \mu) + \varepsilon_k, \qquad k \in \mathbb{Z}, \tag{2.1}$$

where $(\varepsilon_k\colon k \in \mathbb{Z})$ are centered, independent and identically distributed innovations in L_H^2 and $\Psi : H \to H$ a bounded linear operator satisfying $\|\Psi^{k_0}\|_{\mathcal{L}} < 1$ for some $k_0 \geq 1$. The latter condition ensures that the recurrence equations (2.1) have a strictly stationary and causal solution in L_H^2. Bosq (2000) has in the FAR(1) case used the prediction equations to devise what is now often referred to as the common predictor. This one-step ahead prediction is based on an estimator $\tilde{\Psi}_n$ of Ψ and then given by $\tilde{Y}_{n+1} = \tilde{\Psi}_n Y_n$. Details of this method are given in Section 3, where it will be used as a benchmark to compare with the novel methodology to be introduced in the following. The new prediction technique avoids estimating operators directly and instead utilizes existing multivariate prediction methods.

The proposed prediction algorithm proceeds in three steps. First, select d, the number of principal components to be included in the analysis, for example by ensuring that a certain fraction of the data variation is explained. With the sample eigenfunctions, empirical FPC scores $y^e_{k,\ell} = \langle Y_k, \hat{v}_\ell \rangle$ can now be computed for each combination of observations Y_k, $k = 1, \ldots, n$, and sample eigenfunctions $\hat{v}_\ell$, $\ell = 1, \ldots, d$. The superscript e emphasizes that empirical versions are considered. Create from the FPC scores the vectors

$$\boldsymbol{Y}^e_k = (y^e_{k,1}, \ldots, y^e_{k,d})',$$

where $'$ signifies transposition. By nature of FPCA, the vector $\boldsymbol{Y}^e_k$ contains most of the information on the curve Y_k. In the second step, fix the prediction lag h. Then, use multivariate prediction techniques to produce the h-step ahead prediction

$$\hat{\boldsymbol{Y}}^e_{n+h} = (\hat{y}^e_{n+h,1}, \ldots, \hat{y}^e_{n+h,d})'$$

given the vectors $\boldsymbol{Y}^e_1, \ldots, \boldsymbol{Y}^e_n$. Standard methods such as the Durbin-Levinson and innovations algorithm can be readily applied, but other options such as exponential smoothing and nonparametric prediction algorithms are available as well. In the third and last step, the multivariate predictions are re-transformed to functional objects. This conversion is achieved by defining the truncated Karhunen-Loéve representation

$$\hat{Y}_{n+h} = \hat{y}^e_{n+h,1}\, \hat{v}_1 + \cdots + \hat{y}^e_{n+h,d} \hat{v}_d \tag{2.2}$$

based on the predicted FPC scores $\hat{y}^e_{k,\ell}$ and the estimated eigenfunctions $\hat{v}_\ell$. The resulting $\hat{Y}_{n+h}$ is then used as the h-step ahead functional prediction of Y_{n+h}. The three prediction steps are summarized in Algorithm 1.

ACCEPTED MANUSCRIPT

Algorithm 1 Functional Prediction

1. Fix d. For $k = 1, \ldots, n$, use the data $Y_1, \ldots, Y_n$ to compute the vectors

   $\boldsymbol{Y}_k^e = (y_{k,1}^e, \ldots, y_{k,d}^e)'$,

   containing the first d empirical FPC scores $y_{k,\ell}^e = \langle Y_k, \hat{v}_\ell \rangle$.
2. Fix h. Use $\boldsymbol{Y}_1^e, \ldots, \boldsymbol{Y}_n^e$ to determine the h-step ahead prediction

   $\hat{\boldsymbol{Y}}_{n+h}^e = (\hat{y}_{n+h,1}^e, \ldots, \hat{y}_{n+h,d}^e)'$

   for $\boldsymbol{Y}_{n+h}^e$ with an appropriate multivariate algorithm.
3. Use the functional object

   $\hat{Y}_{n+h} = \hat{y}_{n+h,1}^e \, \hat{v}_1 + \ldots + \hat{y}_{n+h,d}^e \hat{v}_d$

   as h-step ahead prediction for Y_{n+h}.

Several remarks are in order. The proposed algorithm is conceptually simple and allows for a number of immediate extensions and improvements as it is not bound by an assumed FAR structure or, in fact, any other particular functional time series specification. This is important because there is no well developed theory for functional versions of the the well-known linear ARMA time series models ubiquitous in univariate and multivariate settings. Moreover, if an FAR(p) structure is indeed imposed on $(Y_k \colon k \in \mathbb{Z})$, then it appears plausible that $\boldsymbol{Y}_1^e, \ldots, \boldsymbol{Y}_n^e$ should approximately follow a VAR(p) model. This statement will be made precise in Appendix A.

The FAR(1) model should in practice be employed only if it provides a reasonable approximation to the unknown underlying dynamics. To allow for more flexible predictions, higher-order FAR processes could be studied. The proposed methodology offers an automatic way to select the appropriate order p along with the dimensionality d (see Section 3.4). It can, in fact, be applied to any stationary functional time series. For example, by utilizing the multivariate innovations algorithm (see Section 11.4 in Brockwell & Davis (1991)) in the second step of Algorithm 1. How this is done in the present prediction setting is briefly outlined in Section 5 below.

ACCEPTED MANUSCRIPT

It should be emphasized that the numerical implementation of the new prediction methodology is convenient in R. For the first step, FPC score matrix $(\boldsymbol{Y}_1^e\colon\ \ldots : \boldsymbol{Y}_n^e)$ and corresponding empirical eigenfunctions can be readily obtained with the `fda` package. For the second step, forecasting for the FPC scores can be done in another routine step using the `vars` package in case VAR models are employed. The obtained quantities can be easily combined for obtaining (2.2).

3 Predicting functional autoregressions

The FAR(1) model (2.1) is the most often applied functional time series model. It will be used here as a benchmark to compare the proposed methodology to. Without loss of generality it is assumed that $\mu = E[Y_k] = 0$. More generally, the higher-order FAR(p) model

$$Y_k = \Psi_1(Y_{k-1}) + \cdots + \Psi_p(Y_{k-p}) + \varepsilon_k, \qquad k \in \mathbb{Z}, \tag{3.1}$$

is considered, assuming throughout that (i) $(\varepsilon_k\colon k \in \mathbb{Z})$ is an i.i.d. sequence in L_H^2 with $E[\varepsilon_k] = 0$, and (ii) the operators Ψ_j are such that equation (3.1) possesses a unique stationary and causal solution. All the above conditions are summarized as Assumption `FAR`.

3.1 The standard first-order predictor

In order to obtain Bosq's (2000) predictor, estimation of the autoregressive operator Ψ is briefly discussed. The approach is based on a functional version of the Yule-Walker equations. Let then $(Y_k\colon k \in \mathbb{Z})$ be the solution of (2.1). Applying $E[\langle\cdot, x\rangle Y_{k-1}]$ to (2.1) for any $x \in H$, leads to

$$E[\langle Y_k, x\rangle Y_{k-1}] = E[\langle\Psi(Y_{k-1}), x\rangle Y_{k-1}] + E[\langle\varepsilon_k, x\rangle Y_{k-1}] = E[\langle\Psi(Y_{k-1}), x\rangle Y_{k-1}].$$

Let again $C(x) = E[\langle Y_1, x\rangle Y_1]$ be the covariance operator of Y_1 and also let $D(x) = E[\langle Y_1, x\rangle Y_0]$ be the cross-covariance operator of Y_0 and Y_1. If Ψ' denotes the adjoint operator of Ψ, given by the requirement $\langle\Psi(x), y\rangle = \langle x, \Psi'(y)\rangle$, the operator equation $D(x) = C(\Psi'(x))$ is obtained. This formally gives $\Psi(x) = D'C^{-1}(x)$, where $D'(x) = E[\langle Y_0, x\rangle Y_1]$. The operator D' can be estimated by $\hat{D}'(x) = (n-1)^{-1}\sum_{k=2}^{n}\langle Y_{k-1}, x\rangle Y_k$. A more complicated object is the unbounded operator C^{-1}. Using the spectral decomposition of $\hat{C}_n$, it can be estimated by $\hat{C}_n^{-1}(x) = \sum_{\ell=1}^{d}\hat{\lambda}_\ell^{-1}\langle\hat{v}_\ell, x\rangle\hat{v}_\ell$ for an appropriately chosen d. Combining these results with an additional smoothing step, using the approximation $Y_k \approx \sum_{\ell=1}^{d}\langle Y_k, \hat{v}_\ell\rangle\hat{v}_\ell$, gives the estimator

$$\tilde{\Psi}_n(x) = \frac{1}{n-1}\sum_{k=2}^{n}\sum_{\ell=1}^{d}\sum_{\ell'=1}^{d}\hat{\lambda}_\ell^{-1}\langle x, \hat{v}_\ell\rangle\langle Y_{k-1}, \hat{v}_\ell\rangle\langle Y_k, \hat{v}_{\ell'}\rangle\hat{v}_{\ell'}. \tag{3.2}$$

for $\Psi(x)$. This is the estimator of Bosq (2000). It gives rise to the functional predictor

$$\tilde{Y}_{n+1} = \tilde{\Psi}_n(Y_n) \tag{3.3}$$

for Y_{n+1}. Theorem 8.7 of Bosq (2000) provides the strong consistency of $\tilde{\Psi}$ under certain technical assumptions. A recent result of Hörmann & Kidziński (2012) (see their Corollary 2.1) shows that consistent predictions (meaning that $\|\Psi(Y_n) - \tilde{\Psi}(Y_n)\| \xrightarrow{P} 0$) can be obtained in the present setting if the innovations $(\varepsilon_k : k \in \mathbb{Z})$ are elements of L_H^4. For these results to hold, it is naturally required that $d = d_n \to \infty$. The choice of d_n crucially depends on the decay rate of the eigenvalues of C as well as on the spectral gaps (distances between eigenvalues). As these parameters are unknown, a practical guideline for the dimension reduction is needed. An approach to this problem in the context of this paper will be provided in Section 3.4.

3.2 Fitting vector autoregressions to FPC scores

The goal of this section is to show that the one-step predictors $\hat{Y}_{n+1}$ in (2.2), based on fitting VAR(1) models in Step 2 of Algorithm 1, and $\tilde{Y}_{n+1}$ in (3.3) are asymptotically equivalent for FAR(1) processes. This statement is justified in the next theorem.

Theorem 3.1. *Suppose model* (2.1) *and let Assumption* FAR *hold. Assume that a VAR(1) model is fit to* $\boldsymbol{Y}_1^e, \ldots, \boldsymbol{Y}_n^e$ *by means of ordinary least squares. The resulting predictor* (2.2) *is asymptotically* *equivalent to* (3.3)*. More specifically, if for both estimators the same dimension d is chosen, then*

$$\|\hat{Y}_{n+1} - \tilde{Y}_{n+1}\| = O_P\left(\frac{1}{n}\right) \qquad (n \to \infty).$$

The proof of Theorem 3.1 is given in Section A.2, where the exact difference between the two predictors is detailed. These computations are based on a more detailed analysis given in Section A.1 which reveals that the FPC score vectors $\boldsymbol{Y}_1^e, \ldots, \boldsymbol{Y}_n^e$ follow indeed a VAR(1) model, albeit the non-standard one

$$\boldsymbol{Y}_k^e = B_d^e \boldsymbol{Y}_{k-1}^e + \boldsymbol{\delta}_k, \qquad k = 2, \ldots, n,$$

where the matrix B_d^e is random and the errors $\boldsymbol{\delta}_k$ depend on the lag $\boldsymbol{Y}_{k-1}^e$ (with precise definitions being given in Section A.1). Given this structure, one might suspect that the use of generalized least squares, GLS, could be advantageous. This is, however, not the case. Simulations not reported in this paper indicate that the gains in efficiency for GLS are negligible in the settings considered. This is arguably due to the fact that possible improvements may be significant only for small sample sizes for which, in turn, estimation errors more than make up the presumed advantage.

Turning to the case of FAR(p) processes, notice first that Theorem 3.1 can be established for the more general autoregessive Hilbertian model (ARH(1)). In this case, the space $L^2([0,1])$ is replaced by a general separable Hilbert space. The proof remains literally unchanged. Using this fact, a version of Theorem 3.1 for higher-order functional autoregressions can be derived by a change of Hilbert space. Following the approach in Section 5.1 of Bosq (2000), write the FAR(p) process (3.1) in state space form

$$\begin{pmatrix} Y_k \\ Y_{k-1} \\ \vdots \\ Y_{k-p+1} \end{pmatrix} = \begin{pmatrix} \Psi_1 & \cdots & \Psi_{p-1} & \Psi_p \\ \mathrm{Id} & & & 0 \\ & \ddots & & \vdots \\ & & \mathrm{Id} & 0 \end{pmatrix} \begin{pmatrix} Y_{k-1} \\ Y_{k-2} \\ \vdots \\ Y_{k-p} \end{pmatrix} + \begin{pmatrix} \varepsilon_k \\ 0 \\ \vdots \\ 0 \end{pmatrix}. \qquad (3.4)$$

The left-hand side of (3.4) is a p-vector of functions. It takes values in the space $H_p = (L^2[0,1])^p$. The matrix on the right-hand side of (3.4) is a matrix of operators which will be denoted by

Ψ^*. The components Id and 0 stand for the identity and the zero operator on H, respectively. Equipped with the inner product $\langle x, y\rangle_p = \sum_{j=1}^p \langle x_j, y_j\rangle$ the space H_p defines a Hilbert space. Setting $X_k = (Y_k, \ldots, Y_{k-p+1})'$ and $\delta_k = (\varepsilon_k, 0, \ldots, 0)'$, equation (3.4) can be written as $X_k = \Psi^*(X_{k-1}) + \delta_k$, with $\delta_k \in L^2_{H_p}$. Now, in analogy to (2.2) and (3.3), one can derive the vector-functional predictors $\hat{X}_k = (\hat{X}_k^{(1)}, \ldots, \hat{X}_k^{(p)})'$ and $\tilde{X}_k = (\tilde{X}_k^{(1)}, \ldots, \tilde{X}_k^{(p)})'$ and obtain that $\|\hat{X}_k - \tilde{X}_k\|_p = O_P(1/n)$, where $\|x\|_p = \sqrt{\langle x, x\rangle_p}$. Then, the following corollary is immediate.

Corollary 3.1. *Consider the FAR(p) model* (3.1) *and let Assumption* FAR *hold. Further suppose that* $\|(\Psi^*)^{k_0}\|_{\mathcal{L}} < 1$ *for some* $k_0 \geq 1$. *Then setting* $\hat{Y}_k = \hat{X}_k^{(1)}$ *and* $\tilde{Y}_k = \tilde{X}_k^{(1)}$ *one obtains* $\|\hat{Y}_{n+1} - \tilde{Y}_{n+1}\| = O_P(1/n)$, *as* $n \to \infty$.

3.3 Assessing the error caused by dimension reduction

Assume the underlying functional time series to be the causal FAR(p) process. In the population setting, meaning the model is fully known, the best linear one-step ahead prediction (in the sense of mean-squared loss) is $Y^*_{n+1} = \Psi_1(Y_n) + \cdots \Psi_p(Y_{n-p+1})$, provided $n \geq p$. In this case, the smallest attainable mean-squared prediction error is $\sigma^2 := E[\|\varepsilon_{n+1}\|^2]$. Both estimation methods described in Sections 3.1 and 3.2, however, give predictions that live on a d-dimensional subspace of the original function space. This dimension reduction step clearly introduces a bias, whose magnitude is bounded in this section. It turns out that the bias becomes negligible as $d \to \infty$, thereby providing a theoretical justification for the proposed methodology described in the next subsection.

Unlike in the previous section, it will be avoided to build the proposed procedure on the state space representation (3.4). Rather a VAR(p) model is directly fit by means of ordinary least squares to the d-dimensional score sequence. Continuing to work on the population level, the theoretical predictor

$$\hat{Y}_{n+1} = \hat{y}_{n+1,1} v_1 + \ldots + \hat{y}_{n+1,d} v_d,$$

is analyzed, where $y_{k,\ell} = \langle Y_k, v_\ell\rangle$ and $\hat{y}_{k,\ell}$ its one-step ahead linear prediction. Recall that a bounded linear operator A is called Hilbert-Schmidt if, for some orthonormal basis $(e_\ell \colon \ell \in \mathbb{N})$, $\|A\|_{\mathcal{S}}^2 = \sum_{\ell=1}^{\infty} \|A(e_\ell)\|^2 < \infty$. Note that $\|\cdot\|_{\mathcal{S}}$ defines a norm on the space of compact operators which can be shown to be independent of the choice of basis $(e_\ell \colon \ell \in \mathbb{N})$.

Theorem 3.2. *Consider the FAR(p) model* (3.1) *and suppose that Assumption* `FAR` *holds. Suppose further that* $\Psi_1, \ldots, \Psi_p$ *are Hilbert-Schmidt operators. Then*

$$E[\|Y_{n+1} - \hat{Y}_{n+1}\|^2] \leq \sigma^2 + \gamma_d, \tag{3.5}$$

where

$$\gamma_d = \left(1 + \left[\sum_{j=1}^{p} \psi_{j;d}\right]^2\right) \sum_{\ell=d+1}^{\infty} \lambda_\ell \qquad \textit{and} \qquad \psi_{j;d}^2 = \sum_{\ell=d+1}^{\infty} \|\Psi_j(v_\ell)\|^2.$$

The proof of Theorem 3.2 is given in Appendix A.3.

The constant γ_d bounds the additional prediction error due to dimension reduction. It decomposes into two terms. The first is given by the fraction of variance explained by the principal components $(v_\ell \colon \ell > d)$. The second term gives the contribution these principal components make to the Hilbert-Schmidt norm of the Ψ_j. Note that $\psi_{j;d} \leq \|\Psi_j\|_{\mathcal{S}}$ and that $\sum_{\ell=1}^{\infty} \lambda_\ell = \sigma^2$. As a simple consequence, the error in (3.5) tends indeed to σ^2 for $d \to \infty$.

This useful result, however, does not provide a practical guideline for choosing d in the proposed algorithm because the bound in (3.5) becomes smaller with increasing d. Rather γ_d has to be viewed as the asymptotic error due to dimension reduction, when d is fixed and $n \to \infty$. In practice one does not have full information on the model for the observations $Y_1, \ldots, Y_n$ and consequently several quantities, such as the autocovariance structure of the score vectors, have to be estimated. Then, with larger d, the variance of these estimators increases. In the next section, a novel criterion is provided that allows to simultaneously choose the dimension d and the order p in dependence of the sample size n. This is achieved with the objective of minimizing the mean-squared prediction error MSE.

3.4 Model and dimension selection

Given that the objective of this paper is prediction, it makes sense to choose the model to be fitted to the data as well as the dimension d of the proposed approach such that the MSE is minimized. Population principal components are still considered (recalling that estimators are $\sqrt{n}$-consistent), but in contrast to the previous section estimated processes are studied. The resulting additional estimation error will now be taken into account.

Let (Y_k) be a centered functional time series in L^2_H. Motivated by Corollary 3.1 VAR(p) models are fitted to the score vectors. The target is to propose a fully automatic criterion for choosing d and p. By orthogonality of the eigenfunctions $(v_\ell\colon \ell \in \mathbb{N})$ and the fact that the FPC scores $(y_{n,\ell}\colon \ell \in \mathbb{N})$ are uncorrelated, the MSE can be decomposed as

$$E[\|Y_{n+1} - \hat{Y}_{n+1}\|^2] = E\left[\left\| \sum_{\ell=1}^{\infty} y_{n+1,\ell} v_\ell - \sum_{\ell=1}^{d} \hat{y}_{n+1,\ell} v_\ell \right\|^2\right] = E[\|\boldsymbol{Y}_{n+1} - \hat{\boldsymbol{Y}}_{n+1}\|^2] + \sum_{\ell=d+1}^{\infty} \lambda_\ell,$$

where $\|\cdot\|$ is also used to denote the Euclidean norm of vectors. The process $(\boldsymbol{Y}_n)$ is again stationary. Assuming that it follows a d-variate VAR(p) model, that is,

$$\boldsymbol{Y}_{n+1} = \Phi_1 \boldsymbol{Y}_n + \cdots + \Phi_p \boldsymbol{Y}_{n-p+1} + \boldsymbol{Z}_{n+1},$$

with some appropriate white noise $(\boldsymbol{Z}_n)$, it can be shown (see, for example, Lütkepohl (2006)) that

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \stackrel{d}{\to} \mathcal{N}_{pd^2}(\boldsymbol{0}, \Sigma_{\boldsymbol{Z}} \otimes \Gamma_p^{-1}), \tag{3.6}$$

where $\boldsymbol{\beta} = \mathrm{vec}([\Phi_1, \ldots, \Phi_p]')$ and $\hat{\boldsymbol{\beta}} = \mathrm{vec}([\hat{\Phi}_1, \ldots, \hat{\Phi}_p]')$ is its least squares estimator, and where $\Gamma_p = \mathrm{Var}(\mathrm{vec}[\boldsymbol{Y}_p, \ldots, \boldsymbol{Y}_1])$ and $\Sigma_{\boldsymbol{Z}} = E[\boldsymbol{Z}_1 \boldsymbol{Z}_1']$. Suppose now that the estimator $\hat{\boldsymbol{\beta}}$ has been obtained from some independent training sample $(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n) \stackrel{d}{=} (\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n)$. Such an assumption is common in the literature. See, for example, the discussion on page 95 of Lütkepohl (2006). It follows then that

$$
\begin{aligned}
E[||\boldsymbol{Y}_{n+1} - \hat{\boldsymbol{Y}}_{n+1}||^2] &= E[||\boldsymbol{Y}_{n+1} - (\hat{\Phi}_1\boldsymbol{Y}_n + \cdots + \hat{\Phi}_p\boldsymbol{Y}_{n-p+1})||^2] \\
&= E[||\boldsymbol{Z}_{n+1}||^2] + E[||(\Phi_1 - \hat{\Phi}_1)\boldsymbol{Y}_n + \cdots + (\Phi_p - \hat{\Phi}_p)\boldsymbol{Y}_{n-p+1})||^2] \\
&= \mathrm{tr}(\Sigma_{\boldsymbol{Z}}) + E[||[I_p \otimes (\boldsymbol{Y}_n', \ldots, \boldsymbol{Y}_{n-p+1}')](\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})||^2].
\end{aligned}
$$

The independence of $\hat{\boldsymbol{\beta}}$ and $(\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n)$ yields that

$$
\begin{aligned}
E[||[I_p \otimes (\boldsymbol{Y}_n', \ldots, \boldsymbol{Y}_{n-p+1}')](\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})||^2] &= E[\mathrm{tr}\{(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})'[I_p \otimes \Gamma_p](\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})\}] \\
&= \mathrm{tr}\{[I_p \otimes \Gamma_p]E[(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})']\}.
\end{aligned}
$$

Using (3.6), it follows that the last term is

$$
\frac{1}{n}\left(\mathrm{tr}\left[\Sigma_{\boldsymbol{Z}} \otimes I_{pd}\right] + o(1)\right) \sim \frac{pd}{n}\,\mathrm{tr}(\Sigma_{\boldsymbol{Z}}).
$$

(Here $a_n \sim b_n$ means $a_n/b_n \to 1$.) Combining the previous estimates and replacing $\mathrm{tr}(\Sigma_{\boldsymbol{Z}})$ by $n(n-pd)^{-1}\mathrm{tr}(\hat{\Sigma}_{\boldsymbol{Z}})$, leads to

$$
E[||Y_{n+1} - \hat{Y}_{n+1}||^2] \approx \frac{n+pd}{n-pd}\,\mathrm{tr}(\hat{\Sigma}_{\boldsymbol{Z}}) + \sum_{\ell>d}\lambda_\ell.
$$

It is therefore proposed to jointly select the order p and the dimension d as the minimizers of the functional final prediction error-type criterion

$$
\mathrm{fFPE}(p,d) = \frac{n+pd}{n-pd}\,\mathrm{tr}(\hat{\Sigma}_{\boldsymbol{Z}}) + \sum_{\ell>d}\hat{\lambda}_\ell. \tag{3.7}
$$

With the use of the functional FPE criterion, the proposed prediction methodology becomes fully data driven and does not need the additional subjective specification of tuning parameters. It is in particular noteworthy that the selection of d is now made in dependence of the sample size n. The excellent practical performance of this method is demonstrated in Sections 6 and 7.

It should finally be noted that in a multivariate context Akaike (1969) originally suggested the use of the log-determinant in place of the trace in (3.7) so as to make his FPE criterion equivalent to the AIC criterion (see Lütkepohl (2006)). Here, however, the use of the trace is recommended, since this puts the two terms in (3.7) on the same scale.

4 Prediction with covariates

In many practical problems, such as in the particulate matter example presented in Section 7, predictions could not only contain lagged values of the functional time series of interest, but also other exogenous covariates. These covariates might be scalar, vector-valued and functional. Formally the goal is then to obtain a predictor $\hat{Y}_{n+h}$ given observations of the curves $Y_1, \ldots, Y_n$ and a number of covariates $X_n^{(1)}, \ldots, X_n^{(r)}$. The exogenous variables need not be defined on the same space. For example, $X_n^{(1)}$ could be scalar, $X_n^{(2)}$ a function and $X_n^{(3)}$ could contain lagged values of $X_n^{(2)}$. The following adaptation of the methodology given in Algorithm 1 is derived under the assumption that $(Y_k \colon k \in \mathbb{Z})$ as well as the covariates $(X_n^{(i)} \colon n \in \mathbb{N})$ are stationary processes in their respective spaces. The modified procedure is summarized in Algorithm 2.

The first step of Algorithm 2 is expanded compared to Algorithm 1. Step 1(a) performs FPCA on the response time series curves $Y_1, \ldots, Y_n$. In Step 1(b), all functional covariates are first transformed via FPCA into empirical FPC score vectors. For each functional covariate, a different number of principal components can be selected. Vector-valued and scalar covariates can be used directly. All exogenous covariates are finally combined into one vector $\boldsymbol{R}_n^e$ in Step 1(c).

Details for Step 2 and the one-step ahead prediction case $h = 1$ could be as follows. Since stationarity is assumed for all involved processes, the resulting FPC scores form stationary time series. Define hence

$$\Gamma_{\boldsymbol{YY}}(i) = \mathrm{Cov}(\boldsymbol{Y}_k^e, \boldsymbol{Y}_{k-i}^e), \qquad \Gamma_{\boldsymbol{YR}}(i) = \mathrm{Cov}(\boldsymbol{Y}_k^e, \boldsymbol{R}_{k-i}^e), \qquad \Gamma_{\boldsymbol{RR}} = \mathrm{Cov}(\boldsymbol{R}_k^e, \boldsymbol{R}_k^e)$$

and notice that these matrices are independent of k. Fix $m \in \{1, \ldots, n\}$. The best linear predictor $\hat{\boldsymbol{Y}}_{n+1}^e$ of $\boldsymbol{Y}_{n+1}^e$ given the vector variables $\boldsymbol{Y}_n^e, \ldots, \boldsymbol{Y}_{n-m+1}^e, \boldsymbol{R}_n^e$ can be obtained by projecting each component $y_{n+1,\ell}^e$ of $\boldsymbol{Y}_{n+1}^e$ onto $\overline{\mathrm{sp}}\{y_{k,i}^e, R_{nj}^e | 1 \leq i \leq d,\ 1 \leq j \leq r,\ n-m+1 \leq k \leq n\}$. Then there exist

Algorithm 2 Functional Prediction with Exogenous Covariates

1. (a) Fix d. For $k = 1, \ldots, n$, use the data $Y_1, \ldots, Y_n$ to compute the vectors

   $$\boldsymbol{Y}_k^e = (y_{k,1}^e, \ldots, y_{k,d}^e)',$$

   containing the first d empirical FPC scores $y_{k,\ell}^e = \langle Y_k, \hat{v}_\ell \rangle$.

   (b) For a functional covariate, fix d'. For $k = 1, \ldots, n$, use the data $X_1, \ldots, X_n$ to compute the vectors

   $$\boldsymbol{X}_k^e = (x_{k,1}^e, \ldots, x_{k,d'}^e)',$$

   containing the first d' empirical FPC scores $x_{k,\ell}^e = \langle X_k, \hat{w}_\ell \rangle$. Repeat this step for each functional covariate.

   (c) Combine all covariate vectors into one vector $\boldsymbol{R}_n^e = (R_{n1}^e, \ldots, R_{nr}^e)'$.

2. Fix h. Use $\boldsymbol{Y}_1^e, \ldots, \boldsymbol{Y}_n^e$ and $\boldsymbol{R}_n^e$ to determine the h-step ahead prediction

   $$\hat{\boldsymbol{Y}}_{n+h}^e = (\hat{y}_{n+h,1}^e, \ldots, \hat{y}_{n+h,d}^e)'$$

   for $\boldsymbol{Y}_{n+h}^e$ with an appropriate multivariate algorithm.

3. Use the functional object

   $$\hat{Y}_{n+h} = \hat{y}_{n+h,1}^e \, \hat{v}_1 + \cdots + \hat{y}_{n+h,d}^e \hat{v}_d$$

   as h-step ahead prediction for Y_{n+h}.

$d \times d$ matrices Φ_i and a $d \times r$ matrix Θ, such that

$$\hat{\boldsymbol{Y}}_{n+1}^e = \Phi_1 \boldsymbol{Y}_n^e + \Phi_2 \boldsymbol{Y}_{n-1}^e + \cdots + \Phi_m \boldsymbol{Y}_{n-m+1}^e + \Theta \boldsymbol{R}_n^e.$$

Using the projection theorem, it can be easily shown that the matrices $\Phi_1, \ldots, \Phi_m$ and Θ are characterized by the equations

$$\Gamma_{\boldsymbol{YY}}(i+1) = \Phi_1 \Gamma_{\boldsymbol{YY}}(i) + \cdots + \Phi_m \Gamma_{\boldsymbol{YY}}(i+1-m) + \Theta \Gamma_{\boldsymbol{RY}}(i), \quad i = 0, \ldots, m-1;$$

$$\Gamma_{\boldsymbol{YR}}(1) = \Phi_1 \Gamma_{\boldsymbol{YR}}(0) + \cdots + \Phi_m \Gamma_{\boldsymbol{YR}}(1-m) + \Theta \Gamma_{\boldsymbol{RR}}.$$

Let

$$\Gamma = \begin{pmatrix} \Gamma_{\boldsymbol{YY}}(0) & \Gamma_{\boldsymbol{YY}}(1) & \cdots & \Gamma_{\boldsymbol{YY}}(m-1) & \Gamma_{\boldsymbol{YR}}(0) \\ \Gamma_{\boldsymbol{YY}}(-1) & \Gamma_{\boldsymbol{YY}}(0) & \cdots & \Gamma_{\boldsymbol{YY}}(m-2) & \Gamma_{\boldsymbol{YR}}(-1) \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \Gamma_{\boldsymbol{YY}}(1-m) & \Gamma_{\boldsymbol{YY}}(2-m) & \cdots & \Gamma_{\boldsymbol{YY}}(0) & \Gamma_{\boldsymbol{YR}}(1-m) \\ \Gamma_{\boldsymbol{RY}}(0) & \Gamma_{\boldsymbol{RY}}(1) & \cdots & \Gamma_{\boldsymbol{RY}}(m-1) & \Gamma_{\boldsymbol{RR}}(0) \end{pmatrix}.$$

Assuming that Γ has full rank, it follows that

$$(\Phi_1, \Phi_2, \ldots, \Phi_m, \Theta) = (\Gamma_{\boldsymbol{YY}}(1), \ldots, \Gamma_{\boldsymbol{YY}}(m), \Gamma_{\boldsymbol{YR}}(1))\Gamma^{-1}.$$

The matrices $\Gamma_{\boldsymbol{YY}}(i)$, $\Gamma_{\boldsymbol{YR}}(i)$ and $\Gamma_{\boldsymbol{RR}}$ have to be replaced in practice by the corresponding sample versions. This explains why predictions should not be made conditional on all data $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n$. It would involve the matrices $\Gamma_{\boldsymbol{YY}}(n), \Gamma_{\boldsymbol{YY}}(n-1), \ldots$ which cannot be reasonably estimated from the sample. In the application of Section 7 a VARX(p) model of dimension d is fitted. The dimension d and the order p are selected by the adjusted fFPE criterion (7.1)

5 Additional options

5.1 Using the innovations algorithm

The proposed methodology has been developed with a focus on functional autoregressive processes. For this case a fully automatic prediction procedure has been constructed in Section 3.4. It should be noted, however, that other options are generally available to the practitioner as well if one seeks to go beyond the FAR framework. One way to do this would be to view the fitted FAR process as a best approximation to the underlying stationary functional time series in the sense of the functional FPE-type criterion in 3.4.

In certain cases a more parsimonious modeling could be achieved if one instead used the innovations algorithm in Step 2 of Algorithm 1. The advantage of the innovations algorithm is that it can be updated quickly when new observations arrive. It should be particularly useful if one has to predict functional moving average processes that have an infinite functional autoregressive representation with coefficient operators whose norms only slowly decay with the lag. The application of Algorithm 3 requires the estimation of covariances $\Gamma(k)$ for increasing lag k. Such estimates are less reliable the smaller n and the larger k. Therefore including too many lag values has a negative effect on the estimation accuracy. If estimated eigenfunctions and the covariance matrices $\hat{\Gamma}(k)$ are replaced by population analogues, then this algorithm gives the best linear prediction (in mean square sense) of the population FPC scores based on the last m observations.

5.2 Prediction bands

To assess the forecast accuracy, a method for computing uniform prediction bands is provided in this section. The target is to find parameters $\underline{\xi}_\alpha, \overline{\xi}_\alpha \geq 0$, such that, for a given $\alpha \in (0,1)$ and $\gamma\colon [0,1] \to [0,\infty)$,

$$P\Big(\hat{Y}_{n+1}(t) - \underline{\xi}_\alpha \gamma(t) \leq Y_{n+1}(t) \leq \hat{Y}_{n+1}(t) + \overline{\xi}_\alpha \gamma(t) \quad \text{for all } t \in [0,1]\Big) = \alpha.$$

Algorithm 3 The Innovations Algorithm for Step 2 in Algorithm 1

1. Fix $m \in \{1, \ldots, n\}$. The last m observations will be used to compute the predictor.
2. For $k = 0, 1, \ldots, m$, compute

$$\hat{\Gamma}(k) = \frac{1}{k} \sum_{j=1}^{k} (\hat{\boldsymbol{Y}}_j^e - \overline{\boldsymbol{Y}}^e)(\hat{\boldsymbol{Y}}_j^e - \overline{\boldsymbol{Y}}^e)',$$

where $\overline{\boldsymbol{Y}}^e = \frac{1}{n} \sum_{k=1}^{n} \hat{\boldsymbol{Y}}_k^e$.

3. Set

$$\hat{\boldsymbol{Y}}_{n+1}^e = \sum_{j=1}^{m} \Theta_{mj} (\boldsymbol{Y}_{n+1-j}^e - \hat{\boldsymbol{Y}}_{n+1-j}^e),$$

where

$$\begin{aligned}
\Theta_{00} &= \hat{\Gamma}(0), \\
\Theta_{m,m-k} &= \left(\hat{\Gamma}(n-k) - \sum_{j=0}^{k-1} \Theta_{m,m-j} \Theta_{j0} \Theta_{k,k-j}' \right) \Theta_{k0}^{-1}, \qquad k = 0, \ldots, m-1, \\
\Theta_{m0} &= \hat{\Gamma}(0) - \sum_{j=0}^{m-1} \Theta_{m,m-j} \Theta_{j0} \Theta_{m,m-j}'.
\end{aligned}$$

The recursion is solved in the order $\Theta_{00}; \Theta_{11}, \Theta_{10}; \Theta_{22}, \Theta_{21}, \Theta_{20}; \ldots$

Algorithm 4 Algorithm for determining prediction bands

1. Compute the d-variate score vectors $\boldsymbol{Y}_1^e, \ldots, \boldsymbol{Y}_n^e$ and the sample FPCs $\hat{v}_1, \ldots, \hat{v}_d$.
2. For $L > 0$ fix $k \in \{L+1, \ldots, n-1\}$ and compute
$$\hat{Y}_{k+1} = \hat{y}^e_{k+1,1}\,\hat{v}_1 + \cdots + \hat{y}^e_{k+1,d}\hat{v}_d,$$
where $\hat{y}^e_{k+1,1}, \ldots, \hat{y}^e_{k+1,d}$, are the components of the one-step ahead prediction obtained from $\boldsymbol{Y}_1^e, \ldots, \boldsymbol{Y}_k^e$ by means of a multivariate algorithm.
3. Let $M = n - L$. For $k \in \{1, \ldots, M\}$, define the residuals $\hat{\epsilon}_k = Y_{k+L} - \hat{Y}_{k+L}$.
4. For $t \in [0,1]$, define $\gamma(t) = \mathrm{sd}(\hat{\epsilon}_k(t) \colon k = 1 \ldots, M)$.
5. Determine $\underline{\xi}_\alpha, \overline{\xi}_\alpha$ such that $\alpha \times 100\%$ of the residuals satisfy
$$-\underline{\xi}_\alpha \gamma(t) \le \hat{\epsilon}_i(t) \le \overline{\xi}_\alpha \gamma(t) \quad \text{for all } t \in [0,1].$$

There is no a priori restriction on the function γ, but clearly it should account for the structure and variation of the data. Although this problem is very interesting from a theoretical standpoint, only a practical approach for the determination of $\underline{\xi}_\alpha, \overline{\xi}_\alpha$ and γ is proposed here. It is outlined in Algorithm 4.

The purpose of the parameter L is to ensure a reasonable sample size for the predictions in Step 2 of Algorithm 4. The residuals $\hat{\epsilon}_1, \ldots \hat{\epsilon}_M$ are then expected to be approximately stationary and, by a law of large numbers effect, to satisfy

$$\begin{aligned}\frac{1}{M}\sum_{k=1}^{M} I\Big(-\underline{\xi}_\alpha \gamma(t) \le \hat{\epsilon}_k(t) \le \overline{\xi}_\alpha \gamma(t) \quad \text{for all } t \in [0,1]\Big) \\ \approx P\Big(-\underline{\xi}_\alpha \gamma(t) \le Y_{n+1}(t) - \hat{Y}_{n+1}(t) \le \overline{\xi}_\alpha \gamma(t) \quad \text{for all } t \in [0,1]\Big).\end{aligned}$$

Note that, in Step 1, the principal components $\hat{v}_1, \ldots, \hat{v}_d$ have been obtained from the entire sample $Y_1, \ldots, Y_n$ and not just from the first k observations. The choice of γ in Step 4 clearly accounts for the variation of the data. For an intraday time exhibiting a higher volatility there should also be a broader prediction interval. Typically the constants $\underline{\xi}_\alpha$ and $\overline{\xi}_\alpha$ are chosen equal, but there may be situations when this is not desired.

One advantage of this method is that it does not require particular model assumptions. If two competing prediction methods exist, then the one which is performing better on the sample will lead to narrower prediction bands. Simulation results not reported in this paper indicate that Algorithm 4 performs well in finite samples even for moderate sample sizes.

6 Simulations

6.1 General setting

To analyze the finite sample properties of the new prediction method, a comparative simulation study was conducted. The proposed method was tested on a number of functional time series, namely first- and second-order FAR processes, first-order FMA processes and FARMA processes of order (1,2). In each simulation run, $n = 200$ (or 1000) observations were generated of which the first $m = 180$ (or 900) were used for parameter estimation as well as order and dimension selection with the fFPE(p, d) criterion (3.7). On the remaining 20 (or 100) observations one-step ahead predictions and the corresponding squared prediction errors were computed. From these mean (MSE), median (medSE) and standard deviation (SD) were calculated. If not otherwise mentioned, this procedure was repeated $N = 100$ times. More details and a summary of the results are given in Sections 6.2–6.4.

Since in simulations one can only work in finite dimensions, the setting consisted of D Fourier basis functions $v_1, \ldots, v_D$ on the unit interval $[0, 1]$, which together determine the (finite-dimensional) space $H = \mathrm{sp}\{v_1, \ldots, v_D\}$. Note that an arbitrary element $x \in H$ has the representation $x(t) = \sum_{\ell=1}^{D} c_\ell v_\ell(t)$ with coefficients $\mathbf{c} = (c_1, \ldots, c_D)'$. If $\Psi\colon H \to H$ is a linear operator, then

$$\Psi(x) = \sum_{\ell=1}^{D} c_\ell \Psi(v_\ell) = \sum_{\ell=1}^{D} \sum_{\ell'=1}^{D} c_\ell \langle \Psi(v_\ell), v_{\ell'} \rangle v_{\ell'} = (\boldsymbol{\Psi}\mathbf{c})'\mathbf{v},$$

where $\mathbf{\Psi}$ is the matrix whose ℓ-th column and ℓ'-th row is $\langle \Psi(v_\ell), v_{\ell'} \rangle$, and $\mathbf{v} = (v_1, \ldots, v_D)'$ is the vector of basis functions. The linear operators needed to simulate the functional time series of interest can thus be represented by a $D \times D$ matrix that acts on the coefficients in the basis function representation of the curves. The corresponding innovations were generated according to

$$\varepsilon_k(t) = \sum_{\ell=1}^{D} A_{k,\ell} v_\ell(t), \tag{6.1}$$

where $A_{k,\ell}$ are i.i.d. normal random variables with mean zero and standard deviations σ_ℓ that will be specified below.

6.2 Comparison with scalar prediction

As mentioned in the introduction, a special case of the proposed method was considered by Hyndman & Ullah (2007) and Hyndman & Shang (2009). Motivated by the fact that PCA score vectors have uncorrelated components, these authors have proposed to predict the scores individually as univariate time series. This will be referred to as the *scalar method*, in contrast to the *vector method* promoted in this paper. The scalar method is fast and works well as long as the cross-spectra related to the score vectors are close to zero. However, in general the score vectors have non-diagonal autocorrelations. Then, scalar models are not theoretically justified. To explore the effect of neglecting cross-sectional dependence, FAR(1) time series of length $n = 200$ were generated as described above. For the purpose of demonstration $D = 3$ and $\sigma_1 = \sigma_2 = \sigma_3 = 1$ were chosen. Two autocovariance operators $\Psi^{(1)}$ and $\Psi^{(2)}$ with corresponding matrices

$$\mathbf{\Psi}^{(1)} = \begin{pmatrix} -0.05 & -0.23 & 0.76 \\ 0.80 & -0.05 & 0.04 \\ 0.04 & 0.76 & 0.23 \end{pmatrix} \quad \text{and} \quad \mathbf{\Psi}^{(2)} = 0.8 \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix},$$

were tested. Both matrices are orthogonal with norm 0.8. In these simple settings it is easy to compute the population autocorrelation function (ACF) of the 3-dimensional FPCA score vectors.

The ACF related to the score sequences of the process generated by $\Psi^{(1)}$ is displayed in Figure A.1. It shows that two scores are uncorrelated at lag zero and that there is almost no temporal correlation in the individual score sequences. However, at lags greater than 1 there is considerable dependence in the cross-correlations between the first and the third score sequence. The analogous plot for $\Psi^{(2)}$ would reveal a contrary behavior: while the autocorrelations of the individual score sequences decay slowly, cross-correlations are zero at all lags.

Given these observations, it is expected that the scalar method will do very well in forecasting the scores when data are generated by operator $\Psi^{(2)}$, while it should be not competitive with the vector method if $\Psi^{(1)}$ is used. This conjecture is confirmed in Figure A.2 which shows histograms of the ratios

$$r_i = \frac{\text{MSE vector method}}{\text{MSE scalar method}}, \qquad i = 1, \ldots, 1000, \tag{6.2}$$

obtained from 1000 simulation runs. The grey histogram refers to the time series generated by $\Psi^{(2)}$. It indicates that the scalar method is a bit favorable, as the ratios tend to be slightly larger than one. Contrary to this, a clear superiority of the vector method can be seen when data stem from the sequence generated by $\Psi^{(1)}$. In a majority of the cases, the MSE resulting from the vector methods is less than half as large as the corresponding MSE obtained by the scalar method. It should also be mentioned that p and d where estimated for the proposed method, while they were fixed at the true values $p = 1$ and $d = 3$ for the scalar predictions.

6.3 Comparison with standard functional prediction

In this section the proposed prediction is compared on FAR(2) processes $Y_k = \Psi_1 Y_{k-1} + \Psi_2 Y_{k-2} + \varepsilon_k$ to the standard predicton of Bosq (2000). For the latter, the multiple testing procedure of Kokoszka & Reimherr (2013) was utilized to determine the order p of the FAR model to be fitted. Following these authors, d was chosen as the smallest integer such that the first d principal components explain at least 80% of the variance of the data. To ensure that the multiple testing procedure keeps an overall asymptotic level of 10%, the levels in three subtests (so testing up to a maximal order $p = 3$) were chosen to be 5%, 3% and 2%, respectively. For ease of reference this method will be referred to as the BKR method. Owing to the results of Section 3, both methods are expected to yield similar results if the order p was known and if the same dimension d was chosen for the two predictors.

The operators were generated such that $\Psi_1 = \kappa_1 \Psi$ and $\Psi_2 = \kappa_2 \Psi$ with $|\kappa_1| + |\kappa_2| < 1$ to ensure stationarity. The case $\kappa_2 = 0$ yields the FAR(1) process. The operator Ψ was chosen at random. More precisely, choosing $D = 21$, a $D \times D$ matrix of independent, zero-mean normal random variables with corresponding standard deviations $\sigma_{\ell\ell'}$ was generated. This matrix was then scaled so that the resulting matrix $\boldsymbol{\Psi}$ has induced norm equal to 1. In every iteration of the simulation runs $\boldsymbol{\Psi}$ was newly generated. Two types of standard deviations for the innovations in (6.1) were chosen, namely

$$(\sigma 1) \quad \boldsymbol{\sigma}_1' = (\ell^{-1} \colon \ell = 1, \ldots, D) \qquad \text{and} \qquad (\sigma 2) \quad \boldsymbol{\sigma}_2' = (1.2^{-\ell} \colon \ell = 1, \ldots, D).$$

Note that if $\Psi \colon L^2 \to L^2$, then $\langle \Psi(v_\ell), v_{\ell'} \rangle \to 0$ if $\ell \to \infty$ or $\ell' \to \infty$ by the Riemann-Lebesgue lemma. This will be reflected in the corresponding matrices by choosing $\sigma_{\ell\ell'}$ as a decaying sequence in ℓ and ℓ'. In particular we have chosen $((\sigma_{\ell\ell'})) = \boldsymbol{\sigma}_1 \boldsymbol{\sigma}_1'$ for setting $(\sigma 1)$ and $((\sigma_{\ell\ell'})) = \boldsymbol{\sigma}_2 \boldsymbol{\sigma}_2'$ for setting $(\sigma 2)$.

Results for four pairs of values (κ_1, κ_2) are shown in Table A.3. The numbers are averages from 100 iterations of the simulation setting explained in Section 6.1. Recall that 10% of the data was used in each simulation run to compute out-of-sample predictions. This means that the MSE's are based on 2,000 forecasts when $n = 200$ and 10,000 forecasts when $n = 1,000$. The quantity MSE_a refers to the MSE produced by the proposed method and MSE_b to the MSE obtained from the BKR method. Similarly, PVE_a and PVE_b give the respective averages of the proportions of variance explained by d principal components, where d is the chosen dimension of the predictor. In summary, the following was found:

- The proposed approach had slight advantages over BKR in almost all considered settings. For $\kappa_1 = 0$ and $\kappa_2 = 0.8$, the BKR method almost always failed to choose the correct order p (see Table A.3). In this case MSE_b was about 30%–40% larger than MSE_a.

- With increasing sample size MSE_a decreases and approaches the value of the fFPE. The latter is an estimate for the minimal possible MSE. Contrary to the BKR method, the dimension parameter d chosen by fFPE grows with increasing sample size. This is visualized in Figure A.3.

- When both methods choose the correct order p, MSE_a still had a tendency to be smaller than MSE_b. This may arguably be due to the fact that a data driven criterion was applied to optimally select the dimension parameter d. It can also be seen that the mean squared prediction errors are relatively robust with respect to the choice of d but quite sensitive to the choice of p. In particular, underestimating p can lead to a non-negligible increase of MSE.

- We have also experimented with $D = 51$. The conclusions remain very similar.

6.4 Beyond functional autoregressions

To test the proposed procedure also for non-autoregressive functional time series, it was applied to the functional FMA(2) and FARMA(1,2) processes respectively given by the equations

$$Y_k = \varepsilon_k + \Theta\varepsilon_{k-2}, \tag{6.3}$$

$$Y_k = \Psi_1 Y_{k-1} + \varepsilon_k + \Theta_1\varepsilon_{k-1} + \Theta_2\varepsilon_{k-2}, \tag{6.4}$$

with operators $\Theta = .8\Psi$, $\Psi_1 = .1\Psi$, $\Theta_1 = .1\Psi$ and $\Theta_2 = .9\Psi$ randomly generated as above. Both the fFPE-based proposed procedure and the BKR method were applied to time series of length $n = 1000$. Since a fitting of long autoregressions is expected the maximal order was set to be 10. The rejection levels for the individual tests of the BKR method were set to achieve an overall level of approximately 10%. The simulation results are displayed in Table 6.3. The conclusions of the previous section still hold true. In particular, MSE reductions of 15%–25% are seen, with the reduction being slightly greater for the FARMA(1,2) process. The proposed method approximates the given time series structure generally with longer FAR processes with average orders (taken over 100 simulation runs) between $p = 4$ and $p = 5$ in all four cases. On the other hand, the BKR method largely fails to make adjustments and selects $p = 0$ more than 90% of the time.

7 Predicting particulate matter concentrations

In order to demonstrate its practical usefulness, the new methodology has been applied to environmental data on pollution concentrations. The observations are half-hourly measurements of the concentration (measured in μgm^{-3}) of particulate matter with an aerodynamic diameter of less than $10\mu m$, abbreviated PM10, in ambient air taken in Graz, Austria from October 1, 2010 until March 31, 2011. Since epidemiological and toxicological studies have pointed to negative health effects, European Union (EU) regulation sets pollution standards for the level of the concentration. Policy makers have to ensure compliance with these EU rules and need reliable statistical tools to determine, and justify to the public, appropriate measures such as partial traffic regulation (see Stadlober et al. (2008)). Accurate predictions are therefore paramount for well informed decision making.

Functional data were obtained as follows. In a first step, very few missing intra-day data points were replaced through linear interpolation. A square-root transformation was then applied to the data in order to stabilize the variance. A visual inspection of the data revealed several extreme outliers around New Years Eve known to be caused by firework activities. The corresponding week was removed from the sample. The data was then centered and adjusted for weekly seasonality by subtracting from each observation the corresponding weekday average. This is done because PM10 concentration levels are significantly different for the weekends when traffic volume is much lower. In the next step, 48 observations for a given day were combined into vectors and transformed

ACCEPTED MANUSCRIPT

into functional data using ten cubic B-spline basis functions and least squares fitting. The `fda` package available for the statistical software `R` was applied here. Eventually, 175 daily functional observations, say, $Y_1, \ldots, Y_{175}$, were obtained, roughly representing one winter season for which pollution levels are known to be high. They are displayed in the upper left panel of Figure A.4. Shown in this figure are also the effect of the first three FPCs on the mean curve. Following Ramsay & Silverman (2005), a multiple (using the factor .5) of the ℓth empirical eigenfunction $\hat{v}_\ell$ was added to and subtracted from the overall estimated mean curve $\hat{\mu}$ to study the effect of large (small) first, second or third FPC score. Notice that

$$Y_k \approx \hat{\mu} + y^e_{k1}\hat{v}_1 + y^e_{k2}\hat{v}_2 + y^e_{k3}\hat{v}_3, \qquad k = 1, \ldots, 175,$$

where $y^e_{k\ell} = \langle Y_k, \hat{v}_\ell \rangle$ are the empirical FPC scores. These combine to explain about 89% of variability in the data. The upper right panel of Figure A.4 indicates that if the first FPC score y^e_{k1}, which explains about 72% of the variation, is large (small), then a positive (negative) shift of the mean occurs. The second and third FPCs are contrasts, explaining respectively 10% and 7% of variation, with the second FPC describing an intraday trend and the third FPC indicating whether the diurnal peaks are more or less pronounced (see the lower panel of Figure A.4).

For the comparison of the quality of the competing prediction methods, the following was adopted. First, five blocks of consecutive functional observations $Y_{k+1}, \ldots, Y_{k+100}$ were chosen, with $k = 0, 15, 30, 45, 60$. Each block was then used to estimate parameters and fit a certain model. Then, out-of-sample predictions for the values of $Y_{k+100+\ell}$, $\ell = 1, \ldots, 15$, were made. Finally, the resulting squared prediction errors

$$\int_0^1 \left[Y_{k+100+\ell}(t) - \mathtt{Pr}(Y_{k+100+\ell})(t)\right]^2 dt, \qquad \ell = 1, \ldots, 15,$$

were computed, where `Pr` can stand for any of the prediction methods tested. From the 15 resulting numbers, median ($\mathrm{MED}_{\mathtt{Pr}}$) and mean ($\mathrm{MSE}_{\mathtt{Pr}}$) were computed. Results are reported in Table A.4. With the exception of the first period ($k = 0$), MSE and MED obtained from the new method are significantly smaller than the ones resulting from the BKR method. In fact, during the second and third period ($k = 15$ and $k = 30$) prediction errors are on average only about half as big as the ones obtained via BKR. This may arguably be due to an underestimation of the order by BKR method (as evidenced in the simulations).

PM10 concentrations are known to be high at locations suffering from severe temperature inversions such as the basin areas of the Alps. Following Stadlober et al. (2008), temperature difference between Graz (350m above sea level) and Kalkleiten (710m above sea level) can be utilized to model this phenomenon. Temperature inversion is often seen as a key factor influencing PM10 concentrations because temperatures increasing with sea level result in a sagging exchange of air, thereby yielding a higher pollutant load at the lower elevation.

To illustrate functional prediction with covariates, temperature difference curves of Graz and Kalkleiten have been included as a dependent variable. For the overall sample, the first two FPCs of the temperature difference curves describe about 92% of the variance. Hence, FPCA was used for covariate dimension reduction, leading to the inclusion of a two-dimensional exogenous regressor (which is almost equivalent to the true regressor curve) in the second step of Algorithm 2. Then a d-variate VARX(p) model was fit with d and p selected by the functional final prediction error-type criterion adjusted for the covariate:

$$\mathrm{fFPE}(p, d) = \frac{n + pd + r}{n - pd - r} \operatorname{tr}(\hat{\Sigma}_{\boldsymbol{Z}}) + \sum_{\ell > d} \hat{\lambda}_\ell. \tag{7.1}$$

Here r is the dimension of the regressor vector (in the present case, $r = 2$) and $\hat{\Sigma}_{\boldsymbol{Z}}$ is the covariance matrix of the residuals when a model of order p and dimension d is fit. The latter method is referred to as FPEX. The corresponding prediction results are summarized in Table A.4. A further significant improvement in the mean and median square (out-of-sample) prediction error can be observed.

8 Conclusions

This paper proposes a new prediction methodology for functional time series that appears to be widely and easily applicable. It is based on the idea that dimension reduction with functional principal components analysis should lead to a vector-valued time series of FPC scores that can be predicted with any existing multivariate methodology, parametric and nonparametric. The multivariate prediction is then transformed to a functional prediction using a truncated Karhunen-Loéve decomposition.

The proposed methodology seems to be advantageous for several reasons. Among them is its intuitive appeal, made rigorous for the predominant FAR(p) case, but also its ease of application as existing software packages can be readily used, even by non-experts. It is in particular straightforward to extend the procedure to include exogenous covariates into the prediction algorithm. Simulations and an application to pollution data suggest that the proposed method leads to predictions that are always competitive with and often superior to the benchmark predictions in the field.

It is hoped that the present article can spawn interest among researchers working in the active area of functional time series.

A Theoretical considerations

It is stated in Section 2 that empirical mean and covariance are $\sqrt{n}$-consistent estimators for their population counterparts for a large class of functional time series. The following lemma makes this statement precise for FAR(p) processes. The notation of Section 3.2 is adopted.

Lemma A.1. *Consider the FAR(p) model* (3.1) *and suppose that Assumption* FAR *holds. Further suppose that* $\|(\Psi^*)^{k_0}\|_{\mathcal{L}} < 1$ *for some* $k_0 \geq 1$. *Then (i)* $E[\|\hat{\mu}_n - \mu\|^2] = O(1/n)$. *(ii) If in addition* (ε_k) *in* L^4_H, *then* $E[\|\hat{C}_n - C\|^2] = O(1/n)$.

Proof. If follows from Proposition 2.1 in Hörmann & Kokoszka (2010) and Theorem 3.1 in Bosq (2000) that (X_k) is L^2-m-approximable under (i) and L^4-m-approximable under (ii). L^p-m-approximability is inherited by the projection $\pi(X_k) = X_k^{(1)} = Y_k$. Now the proof follows from Theorems 5 and 6 in Hörmann & Kokoszka (2012). □

A.1 The VAR structure

In case of a VAR(1), Step 2. of Algorithm 1 can be performed with least squares. To explicitly calculate $\hat{\boldsymbol{Y}}^e_{n+1}$, apply $\langle \cdot, \hat{v}_\ell \rangle$ to both sides of $Y_k = \Psi(Y_{k-1}) + \varepsilon_k$ to obtain

$$\begin{aligned}\langle Y_k, \hat{v}_\ell \rangle &= \langle \Psi(Y_{k-1}), \hat{v}_\ell \rangle + \langle \varepsilon_k, \hat{v}_\ell \rangle \\ &= \sum_{\ell'=1}^{\infty} \langle Y_{k-1}, \hat{v}_{\ell'} \rangle \langle \Psi(\hat{v}_{\ell'}), \hat{v}_\ell \rangle + \langle \varepsilon_k, \hat{v}_\ell \rangle \\ &= \sum_{\ell'=1}^{d} \langle Y_{k-1}, \hat{v}_{\ell'} \rangle \langle \Psi(\hat{v}_{\ell'}), \hat{v}_\ell \rangle + \delta_{k,\ell}, \end{aligned} \tag{A.1}$$

with remainder terms $\delta_{k,\ell} = d_{k,\ell} + \langle \varepsilon_k, \hat{v}_\ell \rangle$ where

$$d_{k,\ell} = \sum_{\ell'=d+1}^{\infty} \langle Y_{k-1}, \hat{v}_{\ell'} \rangle \langle \Psi(\hat{v}_{\ell'}), \hat{v}_\ell \rangle,$$

noting that $(\hat{v}_\ell)$ can always be extended to an orthonormal basis of L^2. Some notation is needed. Set $\boldsymbol{e}_k = (\langle \varepsilon_k, v_1 \rangle, \ldots, \langle \varepsilon_k, v_d \rangle)'$ and $\boldsymbol{u}_k = (u_{k,1}, \ldots, u_{k,d})'$, where $u_{k,\ell} = \sum_{\ell'>d} \langle Y_{k-1}, v_{\ell'} \rangle \langle \Psi(v_{\ell'}), v_\ell \rangle$, and let $B_d \in \mathbb{R}^{d \times d}$ be the matrix with entry $\langle \Psi(v_\ell), v_{\ell'} \rangle$ in the ℓth row and the ℓ'th column, $\ell, \ell' = 1, \ldots, d$. Let moreover $\boldsymbol{\beta} = \mathrm{vec}(B_d')$, $\boldsymbol{Z} = (\boldsymbol{Y}_2', \ldots, \boldsymbol{Y}_n')'$, $\boldsymbol{E} = (\boldsymbol{e}_2', \ldots, \boldsymbol{e}_n')'$, $\boldsymbol{U} = (\boldsymbol{u}_2', \ldots, \boldsymbol{u}_n')'$, $X_k = I_d \otimes \boldsymbol{Y}_k'$ and $X = (X_1' : \ldots : X_{n-1}')'$. Replacing the eigenfunctions v_ℓ by their sample counterparts $\hat{v}_\ell$, empirical versions of the above variables are denoted by $\boldsymbol{Y}_k^e$, $\boldsymbol{Z}^e$, X_k^e, X^e, B_d^e and $\boldsymbol{\beta}_d^e$. For a vector $\mathbf{x} \in \mathbb{R}^{d^2}$, the operation mat($\mathbf{x}$) creates a $d \times d$ matrix, whose ℓ-th column contains the elements $v_{(1-\ell)d+1}, \ldots, v_{\ell d}$. Define now $\boldsymbol{\delta}_k = (\delta_{k,1}, \ldots, \delta_{k,d})'$ to arrive at the equations

$$\boldsymbol{Y}_k^e = B_d^e \boldsymbol{Y}_{k-1}^e + \boldsymbol{\delta}_k, \qquad k = 2, \ldots, n. \tag{A.2}$$

The equations in (A.2) formally resemble VAR(1) equations. Notice, however, that it is a nonstandard formulation, since the errors $\boldsymbol{\delta}_k$ are generally not centered and dependent. Furthermore, $\boldsymbol{\delta}_k$ depends in a complex way on $\boldsymbol{Y}^e_{k-1}$, so that the errors are not uncorrelated with past observations. The coefficient matrix B^e_d is also random, but fixed for fixed sample size n. In the sequel these effects are ignored. Utilizing some matrix algebra, (A.2) can be written as the linear regression

$$\boldsymbol{Z}^e = X^e \boldsymbol{\beta}^e_d + \boldsymbol{\Delta}, \tag{A.3}$$

where $\boldsymbol{\Delta} = (\boldsymbol{\delta}'_2, \ldots, \boldsymbol{\delta}'_n)'$. The ordinary least squares estimator is then $\hat{\boldsymbol{\beta}}^e_d = (X^{e\prime} X^e)^{-1} X^{e\prime} \boldsymbol{Z}^e$, and the prediction equation

$$\hat{\boldsymbol{Y}}^e_{n+1} = \hat{B}^e_d \boldsymbol{Y}^e_n = (\hat{y}^e_{n+1,1}, \ldots, \hat{y}^e_{n+1,d})', \tag{A.4}$$

follows directly, defining $\hat{B}^e_d = \mathrm{mat}(\hat{\boldsymbol{\beta}}^e_d)'$.

A.2 Proof of Theorem 3.1

Recall the notations introduced above equation (A.2). In order to prove the asymptotic equivalence between $\tilde{Y}_{n+1}$ in (2.2) and $\hat{Y}_{n+1}$ in (3.3) for the case of FAR(1) functional time series, observe first that

$$\left(\frac{1}{n-1} X^{e\prime} X^e\right)^{-1} = I_d \otimes \hat{\Gamma}^{-1},$$

where $\hat{\Gamma}$ is the $d \times d$ matrix with entries $\hat{\Gamma}(\ell, \ell') = \frac{1}{n-1} \sum_{k=1}^{n-1} y^e_{k,\ell} y^e_{k,\ell'}$ determined by the FPC scores $y^e_{k,\ell} = \langle Y_k, \hat{v}_\ell \rangle$, and $\otimes$ signifies the Kronecker product. With the help of (A.4), the VAR(1) based predictor (2.2) can be written in the form

$$\hat{Y}_{n+1} = \frac{1}{n-1} \left\{ \left(\mathrm{mat}\left([I_d \otimes \hat{\Gamma}^{-1}] X^{e\prime} \boldsymbol{Z}^e \right) \right)' \boldsymbol{Y}^e_n \right\}' \hat{\boldsymbol{v}},$$

with $\hat{\boldsymbol{v}} = (\hat{v}_1, \ldots, \hat{v}_d)'$ being the vector of the first d empirical eigenfunctions. On the other hand, defining the $d \times d$ matrix $\tilde{\Gamma}$ by the entries $\tilde{\Gamma}(\ell, \ell') = \frac{1}{n} \sum_{k=1}^{n} y^e_{k,\ell} y^e_{k,\ell'} = \mathrm{diag}(\hat{\lambda}_1, \ldots, \hat{\lambda}_d)$, direct verification shows that (3.3) takes the form

$$\tilde{Y}_{n+1} = \frac{1}{n-1} \left\{ \left(\mathrm{mat}\left([I_d \otimes \tilde{\Gamma}^{-1}] X^{e\prime} \boldsymbol{Z}^e \right) \right)' \boldsymbol{Y}^e_n \right\}' \hat{\boldsymbol{v}}.$$

The only formal difference between the two predictors under consideration is therefore in the matrices $\hat{\Gamma}$ and $\tilde{\Gamma}$. Now, for any $\ell, \ell' = 1, \ldots, d$,

$$\begin{aligned}\hat{\Gamma}(\ell, \ell') &= \tilde{\Gamma}(\ell, \ell') + \frac{1}{n-1}\frac{1}{n}\sum_{k=1}^{n} y^e_{k,\ell} y^e_{k,\ell'} - \frac{1}{n-1} y^e_{n,\ell} y^e_{n,\ell'} \\ &= \tilde{\Gamma}(\ell, \ell') + \frac{1}{n-1}\left(\hat{\lambda}_\ell I\{\ell = \ell'\} - y^e_{n,\ell} y^e_{n,\ell'}\right),\end{aligned}$$

so that $Y_n \in L^2_H$ implies

$$\left|\hat{\Gamma}(\ell, \ell') - \tilde{\Gamma}(\ell, \ell')\right| \leq \frac{1}{n-1}\left(\frac{1}{n}\sum_{k=1}^{n} \|Y_k\|^2 + \|Y_n\|^2\right) = O_P\left(\frac{1}{n}\right).$$

In the following $\|\cdot\|$ will be used for the L^2 norm, the Euclidean norm in $\mathbb{R}^d$ and matrix norm $\|A\| = \sup_{\|\mathbf{x}\|=1} \|A\mathbf{x}\|$, for a square matrix $A \in \mathbb{R}^{d\times d}$. Let

$$\Delta = \mathrm{mat}\left([I_d \otimes (\hat{\Gamma}^{-1} - \tilde{\Gamma}^{-1})]\frac{1}{n-1} X^{e\prime} \boldsymbol{Z}^e\right).$$

The orthogonality of the $\hat{v}_\ell$ together with Pythagoras' theorem and Bessel's inequality imply that

$$\|\hat{Y}_{n+1} - \tilde{Y}_{n+1}\| = \left\|\Delta' \boldsymbol{Y}^e_n\right\| \leq \|\Delta\|\|\boldsymbol{Y}^e_n\| = \|\Delta\|\left(\sum_{\ell=1}^{d} (y^e_{n,\ell})^2\right)^{1/2} \leq \|\Delta\|\|Y_n\|.$$

Define $S = \mathrm{mat}(\frac{1}{n-1} X^{e\prime}\boldsymbol{Z}^e)$ and notice that $\Delta = (\hat{\Gamma}^{-1} - \tilde{\Gamma}^{-1})S$ and hence $\|\Delta\| \leq \left\|\hat{\Gamma}^{-1} - \tilde{\Gamma}^{-1}\right\|\|S\|$.

Let $\boldsymbol{w} = (w_1, \ldots, w_d)'$. Since $S(\ell, \ell') = \frac{1}{n-1}\sum_{k=1}^{n-1} y^e_{k,\ell} y^e_{k+1,\ell'}$, iterative applications of the Cauchy-Schwarz inequality yield

$$\begin{aligned}\|S\|^2 &= \sup_{\|\boldsymbol{w}\|=1} \sum_{\ell=1}^{d}\left(\sum_{\ell'=1}^{d} \frac{1}{n-1}\sum_{k=1}^{n-1} y^e_{k,\ell} y^e_{k+1,\ell'} w_{\ell'}\right)^2 \\ &\leq \sum_{\ell=1}^{d}\sum_{\ell'=1}^{d}\left(\frac{1}{n-1}\sum_{k=1}^{n-1} y^e_{k,\ell} y^e_{k+1,\ell'}\right)^2 \\ &\leq \sum_{\ell=1}^{d}\sum_{\ell'=1}^{d} \frac{1}{n-1}\sum_{k=1}^{n} (y^e_{k,\ell})^2 \frac{1}{n-1}\sum_{k=1}^{n} (y^e_{k,\ell'})^2 \\ &\leq \left(\frac{1}{n-1}\sum_{k=1}^{n} \|Y_k\|^2\right)^2 \\ &= O_P(1).\end{aligned}$$

It remains to estimate $\|\hat{\Gamma}^{-1} - \tilde{\Gamma}^{-1}\|$. The next step consists of using the fact that, for any $A, B \in \mathbb{R}^{d\times d}$, it holds that $(A+B)^{-1} = A^{-1} - A^{-1}(I + BA^{-1})^{-1}BA^{-1}$, provided all inverse matrices exist. Now choose $A = \tilde{\Gamma}$ and $B = \hat{\Gamma} - \tilde{\Gamma}$. Since in the given setting the time series (Y_n) is stationary and ergodic, it can be deduced that $\hat{\lambda}_d \to \lambda_d$ with probability one. Thus $\hat{\lambda}_d^{-1}\|\tilde{\Gamma} - \hat{\Gamma}\| < 1$ for large enough n, and consequently

$$\begin{aligned}
\left\|\hat{\Gamma}^{-1} - \tilde{\Gamma}^{-1}\right\| &= \left\|\tilde{\Gamma}^{-1}[I_d + (\hat{\Gamma} - \tilde{\Gamma})\tilde{\Gamma}^{-1}]^{-1}(\hat{\Gamma} - \tilde{\Gamma})\tilde{\Gamma}^{-1}\right\| \\
&\le \left\|\tilde{\Gamma}^{-1}\right\|^2 \left\|\hat{\Gamma} - \tilde{\Gamma}\right\| \left\|[I_d + (\hat{\Gamma} - \tilde{\Gamma})\tilde{\Gamma}^{-1}]^{-1}\right\| \\
&\le \frac{\left\|\tilde{\Gamma} - \hat{\Gamma}\right\|}{\hat{\lambda}_d^2} \sum_{\ell=0}^{\infty} \left(\frac{\|\tilde{\Gamma} - \hat{\Gamma}\|}{\hat{\lambda}_d}\right)^{\ell} \\
&= O_P\left(\frac{1}{n}\right).
\end{aligned}$$

It has been assumed here that $\lambda_d > 0$. If $\lambda_d = 0$, then the model has dimension $d' < d$. In this case both estimators will of course be based on at most d' principal components.

Putting together all results, the statement of Theorem 3.1 is established.

A.3 Proof of Theorem 3.2

Using the results and notations of Section 3.4, it follows that

$$E[\|Y_{n+1} - \hat{Y}_{n+1}\|^2] = E[\|\boldsymbol{Y}_{n+1} - \hat{\boldsymbol{Y}}_{n+1}\|^2] + \sum_{i>d} \lambda_i.$$

Some algebra shows that

$$\boldsymbol{Y}_{n+1} = \boldsymbol{\Psi}_1 \boldsymbol{Y}_n + \cdots + \boldsymbol{\Psi}_p \boldsymbol{Y}_{n-p+1} + \boldsymbol{E}_n,$$

where the $d\times d$ matrices $\boldsymbol{\Psi}_j$ have entry $\langle \Psi_j(v_{\ell'}), v_\ell\rangle$ in the ℓ'th column and ℓth row, and $\boldsymbol{E}_n = \boldsymbol{T}_n + \boldsymbol{S}_n$ with d-variate vectors $\boldsymbol{T}_n$ and $\boldsymbol{S}_n$ taking the respective values $\sum_{j=1}^{p} \sum_{\ell'=d+1}^{\infty} y_{n+1-j,\ell'} \langle \Psi_j(v_{\ell'}), v_\ell\rangle$ and $\langle \varepsilon_{n+1}, v_\ell\rangle$ in the ℓth coordinate.

The best linear predictor $\hat{\boldsymbol{Y}}_{n+1}$ of $\boldsymbol{Y}_{n+1}$ based on $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n$ satisfies

$$E[\|\boldsymbol{Y}_{n+1} - \hat{\boldsymbol{Y}}_{n+1}\|^2] \leq E[\|\boldsymbol{Y}_{n+1} - (\boldsymbol{\Psi}_1 \boldsymbol{Y}_n + \cdots + \boldsymbol{\Psi}_p \boldsymbol{Y}_{n-p+1})\|^2] = E[\|\boldsymbol{S}_n\|^2] + E[\|\boldsymbol{T}_n\|^2].$$

The last equality comes from the fact that, due to causality, the components in $\boldsymbol{S}_n$ and in $\boldsymbol{T}_n$ are uncorrelated. Observe next that, by Bessel's inequality, $E[\|\boldsymbol{S}_n\|^2] = \sum_{\ell=1}^d E[\langle \varepsilon_{n+1}, \nu_\ell \rangle^2] \leq \sigma^2$. It remains to bound $E[\|\boldsymbol{T}_n\|^2]$. For this term, it holds

$$\begin{aligned}
E[\|\boldsymbol{T}_n\|^2] &= E\Bigg[\sum_{\ell=1}^{d}\Bigg(\sum_{j=1}^{p}\sum_{\ell'=d+1}^{\infty} y_{n+1-j,\ell'}\langle \Psi_j(\nu_{\ell'}), \nu_\ell\rangle\Bigg)^2\Bigg] \\
&\leq E\Bigg[\sum_{d=1}^{\infty}\Bigg\langle\sum_{j=1}^{p}\sum_{\ell'=d+1}^{\infty} y_{n+1-j,\ell'}\Psi_j(\nu_{\ell'}), \nu_\ell\Bigg\rangle^2\Bigg] \\
&= E\Bigg[\Bigg\|\sum_{j=1}^{p}\sum_{\ell'=d+1}^{\infty} y_{n+1-j,\ell'}\Psi_j(\nu_{\ell'})\Bigg\|^2\Bigg],
\end{aligned}$$

where Parseval's identity was applied in the final step. Repeatedly using the Cauchy-Schwarz inequality, the last expectation can be estimated as

$$\begin{aligned}
&\sum_{j,j'=1}^{p}\sum_{\ell,\ell'=d+1}^{\infty} E[y_{n+1-j,\ell}\, y_{n+1-j',\ell'}]\langle \Psi_j(\nu_\ell), \Psi_{j'}(\nu_{\ell'})\rangle \\
&\quad\leq \sum_{j,j'=1}^{p}\Bigg(\sum_{\ell=d+1}^{\infty}\sqrt{\lambda_\ell}\|\Psi_j(\nu_\ell)\|\Bigg)\Bigg(\sum_{\ell'=d+1}^{\infty}\sqrt{\lambda_{\ell'}}\|\Psi_{j'}(\nu_{\ell'})\|\Bigg) \\
&\quad\leq \sum_{\ell''=d+1}^{\infty}\lambda_{\ell''}\sum_{j,j'=1}^{p}\Bigg(\sum_{\ell=d+1}^{\infty}\|\Psi_j(\nu_\ell)\|^2\Bigg)^{1/2}\Bigg(\sum_{\ell'=d+1}^{\infty}\|\Psi_{j'}(\nu_{\ell'})\|^2\Bigg)^{1/2} \\
&\quad= \sum_{\ell=d+1}^{\infty}\lambda_\ell\Bigg(\sum_{j=1}^{p}\Bigg[\sum_{\ell=d+1}^{\infty}\|\Psi_j(\nu_\ell)\|^2\Bigg]^{1/2}\Bigg)^2.
\end{aligned}$$

Collecting all estimates finishes the proof.

Acknowledgment

Alexander Aue is Associate Professor, Department of Statistics, University of California, Davis, CA 95616 (E-mail: `aaue@ucdavis.edu`). Diogo Dubart Norinho is graduate student, Department of Computer Science, University College London, London WC1E 6BT, UK (E-mail: ucabdub@ucl.ac.uk. Siegfried Hörmann is Chargé de cours, Department of Mathematics, Université libre de Bruxelles, B-1050 Brussels, Belgium (E-mail: shormann@ulb.ac.be). The authors are grateful to the editor, the associate editor and the reviewers for constructive comments and support. This research was partially supported by NSF grants DMS 0905400, DMS 1209226 and DMS 1305858, Communauté française de Belgique—Actions de Recherche Concertées (2010–2015) and the IAP research network grant nr. P7/06 of the Belgian government (Belgian Science Policy). Part of this work was completed during a Research in Pairs stay (Aue and Hörmann) at the Mathematical Research Institute Oberwolfach.

References

Aguilera, A. M., Ocaña, F. A. & Valderrama, M. J. (1999), 'Forecasting time series by functional PCA. Discussion of several weighted approaches', *Computational Statistics* **14**, 443–467.

Akaike, H. (1969), 'Fitting autoregressive models for prediction', *The Annals of the Institute of Statistical Mathematics* **21**, 243–247.

Aneiros-Pérez, G., Cao, R. & Vilar-Fernánez, J. M. (2010), 'Functional methods for time series prediction: A nonparametric approach', *Journal of Forecasting* **30**, 377–392.

Aneiros-Pérez, G. & Vieu, P. (2008), 'Nonparametric time series prediction: A semi-functional partial linear modeling', *Journal of Multivariate Analysis* **99**, 834–857.

Antoniadis, A., Paparoditis, E. & Sapatinas, T. (2006), 'A functional wavelet-kernel approach for time series prediction', *Journal of the Royal Statistical Society, Series B* **68**, 837–857.

Antoniadis, A., Paparoditis, E. & Sapatinas, T. (2009), 'Bandwidth selection for functional time series prediction', *Statistics & Probability Letters* **79**, 733–740.

ACCEPTED MANUSCRIPT

Antoniadis, A. & Sapatinas, T. (2003), 'Wavelet methods for continuous time prediction using Hilbert-valued autoregressive processes', *Journal of Multivariate Analysis* **87**, 133–158.

Besse, P. & Cardot, H. (1996), 'Approximation spline de la prévision d'un processus fonctionnel autorégressif d'ordre 1', *Canadian Journal of Statistics* **24**, 467–487.

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

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

Brockwell, P. J. & Davis, R. A. (1991), *Time Series Analysis: Theory and Methods (2nd ed.)*, Springer-Verlag, New York.

Damon, J. & Guillas, S. (2002), 'The inclusion of exogenous variables in functional autoregressive ozone forecasting', *Environmetrics* **13**, 759–774.

Damon, J. & Guillas, S. (2010), 'The `far` package for `R`'.
URL: `http://cran.r-project.org/web/packages/far/far.pdf`

Didericksen, D., Kokoszka, P. & Zhang, X. (2012), 'Empirical properties of forecasts with the functional autoregressive model', *Computational Statistics* **27**, 285–298.

Ferraty, F. & Vieu, P. (2010), *Nonparametric Functional Data Analysis*, Springer-Verlag, New York.

Hörmann, S. & Kidziński, Ł. (2012), A note on estimation in the Hilbertian linear models. Technical report.
URL: `http://arxiv.org/pdf/1208.2895v2.pdf`

Hörmann, S., Kidziński, L. & Hallin, M. (2013), Dynamic functional principal components. Technical report.
URL: `http://arxiv.org/pdf/1210.7192v3.pdf`

ACCEPTED MANUSCRIPT

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

Hörmann, S. & Kokoszka, P. (2012), 'Functional time series'. In: *Handbook of Statistics: Time Series Analysis—Methods and Applications* **30**, 157–186.

Hyndman, R. J. & Shang, H. L. (2009), 'Forecasting functional time series (with discussion)', *Journal of the Korean Statistical Society* **38**, 199–221.

Hyndman, R. J. & Shang, H. L. (2012), 'The `ftsa` package for `R`'. **URL:** `http://cran.r-project.org/web/packages/ftsa/ftsa.pdf`

Hyndman, R. J. & Ullah, M. S. (2007), 'Robust forecasting of mortality and fertility rates: A functional data approach', *Computational Statistics and Data Analysis* **51**, 4942–4956.

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

Kokoszka, P. & Reimherr, M. (2013), 'Determining the order of the functional autoregressive model', *Journal of Time Series Analysis* **34**, 116–129.

Lütkepohl, H. (2006), *New Introduction to Multiple Time Series Analysis*, Springer-Verlag, New York.

Panaretos, V. & Tavakoli, S. (2013), 'Cramér-Karhunen-Loève representation and harmonic principal component analysis of functional time series', *Stochastic Processes and Their Applications* **123**, 2779–2807.

Ramsay, J. O. & Silverman, B. W. (2005), *Functional Data Analysis (2nd ed.)*, Springer-Verlag, New York.

Shumway, R. H. & Stoffer, D. S. (2011), *Time Series and Its Applications (3rd ed.)*, Springer-Verlag, New York.

Stadlober, E., Hörmann, S. & Pfeiler, B. (2008), 'Quality and performance of a PM10 daily forecasting model', *Atmospheric Environment* **42**, 1098–1109.

κ_1	κ_2	(σ1) fFPE	MSE$_a$	MSE$_b$	PVE$_a$	PVE$_b$	(σ2) fFPE	MSE$_a$	MSE$_b$	PVE$_a$	PVE$_b$
0.2	0.0	2.29	2.32	2.31	0.40	0.83	1.61	1.59	1.58	0.71	0.83
		2.29	2.31	2.31	0.72	0.84	1.61	1.59	1.59	0.81	0.82
0.8	0.0	2.37	2.37	2.47	0.90	0.83	1.64	1.71	1.81	0.89	0.85
		2.30	2.29	2.37	0.97	0.83	1.61	1.62	1.70	0.95	0.85
0.4	0.4	2.38	2.40	2.43	0.73	0.83	1.67	1.65	1.69	0.84	0.84
		2.31	2.33	2.36	0.92	0.83	1.63	1.64	1.71	0.90	0.84
0.0	0.8	2.42	2.48	2.94	0.83	0.83	1.66	1.72	2.28	0.87	0.85
		2.32	2.34	2.94	0.95	0.83	1.63	1.62	2.27	0.93	0.86

Table 6.1: Functional final prediction error (fFPE), mean squared prediction error based on the fFPE criterion (MSE$_a$), mean squared prediction error based on BKR (MSE$_b$), and the corresponding proportions of variance explained by the chosen number of FPCs (PVE$_a$, PVE$_b$). The first row in each setting (κ_1, κ_2) corresponds to $n = 200$, the second row to $n = 1000$.

κ_1	κ_2	$n = 200$: $p = 0$	$p = 1$	$p = 2$	$p = 3$	$n = 1000$: $p = 0$	$p = 1$	$p = 2$	$p = 3$
0.2	0.0	40	**48**	8	4	2	**94**	3	1
		48	**51**	1	0	0	**98**	2	0
0.8	0.0	0	**97**	3	0	0	**100**	0	0
		0	**95**	5	0	0	**81**	17	2
0.4	0.4	1	3	**90**	6	0	0	**99**	1
		3	3	**94**	0	0	0	**95**	5
0.0	0.8	0	0	**95**	5	0	0	**99**	1
		94	0	**5**	1	93	0	**7**	0

Table 6.2: Selected order for different choices of κ_1 and κ_2 from 100 iterations under setting (σ1). For each choice the top (bottom) row represents the order obtained via fFPE (BKR). The number of correctly selected orders is shown in bold.

	(σ1) fFPE	MSE$_a$	MSE$_b$	PVE$_a$	PVE$_b$	(σ2) fFPE	MSE$_a$	MSE$_b$	PVE$_a$	PVE$_b$
FMA(2)	2.37	2.39	2.80	0.92	0.83	1.65	1.64	2.12	0.90	0.85
FARMA(1,2)	2.38	2.42	2.96	0.82	0.83	1.65	1.67	2.24	0.90	0.85

Table 6.3: As in Table 6.1, but for the functional time series in (6.3) and (6.4) for $n = 1000$.

k	p_a	p_b	p_c	d_a	d_b	d_c	MSE_a	MSE_b	MSE_c	MED_a	MED_b	MED_c
0	1	1	2	3	3	3	1.33	**1.28**	1.32	1.28	1.23	**0.88**
15	3	1	3	3	3	3	2.69	5.23	**2.50**	2.38	5.34	**1.45**
30	4	1	3	3	2	3	2.05	4.05	**1.93**	1.33	2.56	**1.26**
45	3	1	3	3	2	3	2.25	2.44	**1.83**	1.34	1.67	**1.14**
60	2	1	1	3	2	5	1.22	1.82	**1.05**	1.12	1.60	**0.89**

Table 7.1: Comparison of the 3 prediction methods. Subscript a (b, c) corresponds to method FPE (BKR, FPEX). We report mean (MSE) and median (MED) of the 15 predictions from each block as well as the values of d and p chosen by the respective methods.

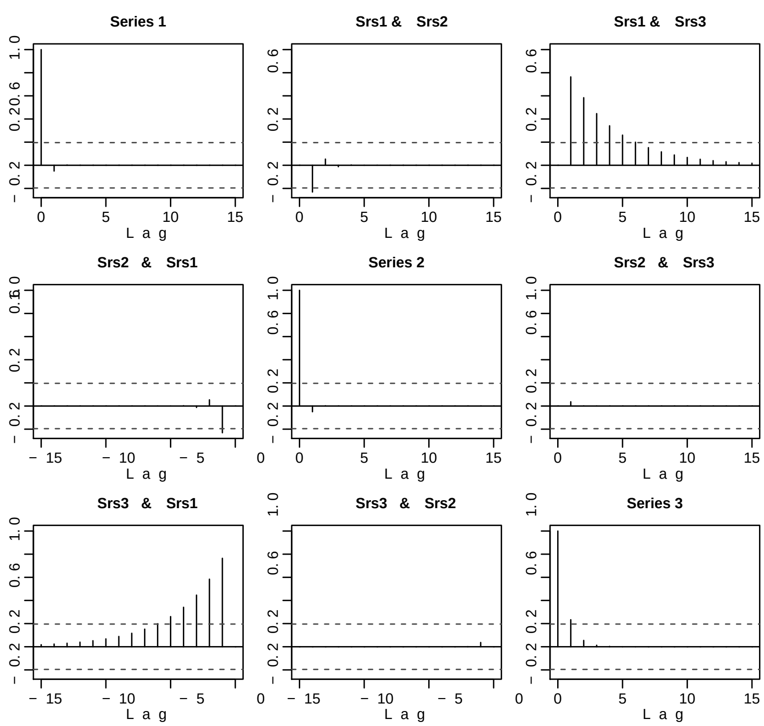

Figure 6.1: Autocorrelation function for the scores related to the sequence generated from operator $\Psi^{(1)}$.

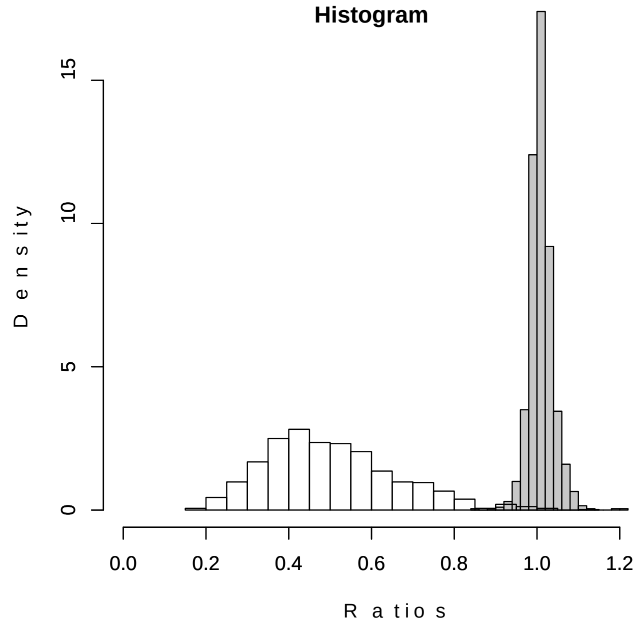

Figure 6.2: Histogram of the ratios r_i in (6.2) for the FAR(1) processes given by the operators $\Psi^{(1)}$ (white) and $\Psi^{(2)}$ (grey).

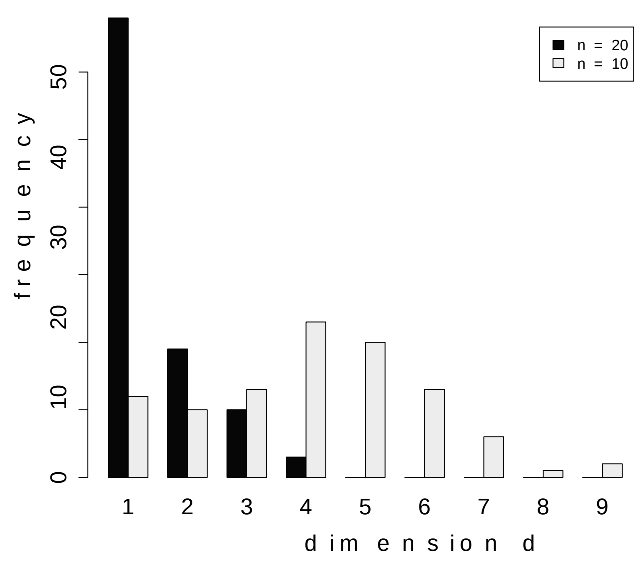

Figure 6.3: Frequencies of the dimensions d chosen by fFPE in 100 simulation runs under setting $(\sigma 1)$ and $(\kappa_1, \kappa_2) = (0.2, 0.0)$.

ACCEPTED MANUSCRIPT

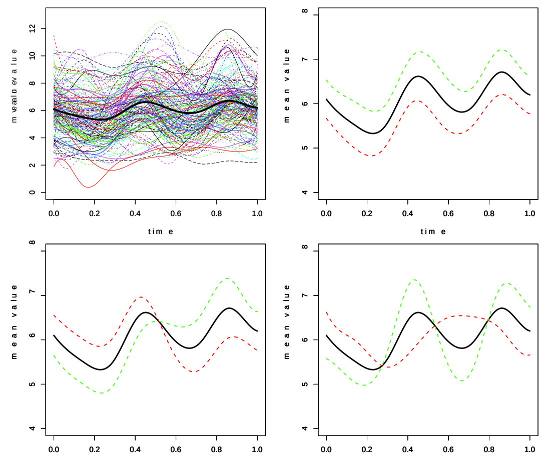

Figure 7.1: Square-root transformed `PM10` observations with fat overall mean curve (upper left panel), effect of the first FPC (upper right panel), effect of the second FPC (lower left panel), and effect of the third FPC (lower right panel).