Comput Stat (2012) 27:285–298
DOI 10.1007/s00180-011-0256-2

ORIGINAL PAPER

Empirical properties of forecasts with the functional autoregressive model

Devin Didericksen · Piotr Kokoszka · Xi Zhang

Received: 10 December 2010 / Accepted: 20 April 2011 / Published online: 3 May 2011

Abstract We study the finite sample performance of predictors in the functional (Hilbertian) autoregressive model $X_{n+1} = \Psi(X_n) + \varepsilon_n$. Our extensive empirical study based on simulated and real data reveals that predictors of the form $\hat{\Psi}(X_n)$ are practically optimal in a sense that their prediction errors are comparable with those of the infeasible perfect predictor $\Psi(X_n)$. The predictions $\hat{\Psi}(X_n)$ cannot be improved by an improved estimation of Ψ, nor by a more refined prediction approach which uses predictive factors rather than the functional principal components. We also discuss the practical limits of predictions that are feasible using the functional autoregressive model. These findings have not been established by theoretical work currently available, and may serve as a practical reference to the properties of predictors of functional data.

Keywords Autoregressive process · Functional data · Prediction

1 Introduction

Over the last two decades, functional data analysis (FDA) has grown into a substantial field of statistical research, with new methodology, numerous useful applications and interesting novel theoretical developments. In this brief note, we cannot even outline

D. Didericksen · P. Kokoszka (✉) · X. Zhang
Department of Mathematics and Statistics, Utah State University,
3900 Old Main Hill, Logan, UT 84322-3900, USA
e-mail: Piotr.Kokoszka@usu.edu

D. Didericksen
e-mail: didericksen@gmail.com

X. Zhang
e-mail: am.zhang@aggiemail.usu.edu

the central ideas, as the field has become very broad, so we merely mention comprehensive introductory expositions of Ramsay and Silverman (2002), Ramsay and Silverman (2005), and Ramsay et al (2009), and more theoretical works by Bosq (2000), Ferraty and Vieu (2006), Bosq and Blanke (2007) and Ferraty and Romain (2011).

The research summarized in this paper pertains to the functional autoregressive (FAR) process studied theoretically by Bosq (2000), and extensively used in both practical and theoretical studies since then, see Besse and Cardot (1996), Damon and Guillas (2002), Antoniadis and Sapatinas (2003), Horváth et al (2010), Hörmann and Kokoszka (2010), Gabrys et al (2010), among numerous other contributions. The FAR(1) model is given by the equation

$$X_{n+1} = \Psi(X_n) + \varepsilon_{n+1}, \tag{1}$$

in which the errors ε_n and the observations X_n are curves, and Ψ is a linear operator transforming a curve into another curve. Precise definitions and assumptions are stated in Sects 2. Model (1) has been introduced to predict curve–valued time series. In addition to Bosq (2000), an informative introduction and review of several prediction methods is given by Besse et al (2000).

Recently Kargin and Onatski (2008) proposed a sophisticated method of one step ahead prediction in model (1) based on predictive factors, and developed an advanced theory that justifies the optimality of their method, we provide a description in Sect. 2. The initial question that motivated this research was whether the method of Kargin and Onatski (2008) is superior in finite samples to the standard method described in Bosq (2000), which estimates of the operator Ψ and forecasts X_{n+1} by $\hat{\Psi}(X_n)$. We found that the predictive factors method never dominates the standard method, and in some cases it performs poorly. We also found that the standard method is almost perfect in a sense that its average prediction errors are typically, within a standard error, the same as if we had perfect knowledge of the operator Ψ. Thus it cannot be hoped that this method can be substantially improved. Surprisingly, this is the case even though the estimates $\hat{\Psi}$ of the operator Ψ are typically *very* poor. We found that it is possible to improve these estimates, we developed a simple algorithm to do it, but this improvement does not affect the quality of prediction. Finally, we realized some natural limits of predictions that can be expected from model (1); predictions with $\hat{\Psi}(X_n)$ are often not better than those with the mean function. We describe in this paper how we arrived at all these conclusions. It is hoped that this contribution will provide informative and useful insights into finite sample properties of estimators and predictors in the FAR(1) model, whose theoretical properties have already been studied in depth.

The paper is organized as follows. In Sect. 2, we describe the two prediction methods and state the assumptions for their validity. Before comparing them, we address in Sect. 4 the question of the estimation of Ψ, and show how better estimates can be constructed. Sects. 3 and 5 describe, respectively, the design of the simulation study and its outcomes. We conclude with Sect. 6 which discusses general properties of predictors derived from model (1).

2 Prediction methods

The theory of autoregressive and more general linear processes in Banach spaces is developed in the monograph of Bosq (2000), which also contains sufficient background. Hörmann and Kokoszka (2011) and Horváth and Kokoszka (2011+) also contain a suitable introduction. Here we state only the minimum of facts required to understand this paper. All functions are assumed to elements of the Hilbert space L^2 of real square integrable functions on the interval [0,1], equipped with the usual inner product $\langle f, g\rangle = \int f(t)g(t)dt$. The errors ε_n in (1) are iid mean zero random elements of L^2, which implies that the X_n also have mean zero. The operator Ψ acting on a function X is defined as

$$\Psi(X)(t) = \int \psi(t,s)X(s)ds,$$

where $\psi(t,s)$ is a bivariate kernel assumed to satisfy $||\Psi|| < 1$, where

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

The condition $||\Psi|| < 1$ ensures the existence of a stationary causal solution to FAR(1) equations.

Before describing the prediction methods, we note that prior to further analysis all simulated curves are converted to functional objects in `R` using 99 Fourier basis functions. We used the package `fda`, which also allows to compute the Functional Principal components (FPC's) of the observations X_n, see Ramsay et al (2009) for the details. We now describe the two methods, which we call "estimated kernel" and "predictive factors", for ease of reference.

Estimated Kernel (EK). Denote by $v_k,\ k = 1, 2, \ldots,$ the FPC's of the X_n, and by $\hat{v}_k,\ k = 1, 2, \ldots, p$, the estimated (or empirical) FPC's (EFPC's). The number p of EFPC's to be used is typically determined by the cumulative variance method, but other methods, including cross–validation or information criteria, can be used as well. Since the v_k form an orthonormal basis in L^2, the kernel ψ admits the expansion

$$\psi(t,s) = \sum_{k,\ell=1}^{\infty} \psi_{k\ell} v_k(t) v_\ell(s).$$

The empirical version of this relation, computed from the sample $X_1, X_2, \ldots, X_N$, is

$$\hat{\psi}_p(t,s) = \sum_{k,\ell=1}^{p} \hat{\psi}_{k\ell} \hat{v}_k(t) \hat{v}_\ell(s), \tag{3}$$

where

$$\hat{\psi}_{ji} = \hat{\lambda}_i^{-1}(N-1)^{-1}\sum_{n=1}^{N-1}\left\langle X_n, \hat{v}_i\right\rangle\left\langle X_{n+1}, \hat{v}_j\right\rangle. \tag{4}$$

Equation (4) is an empirical analog of the relation $\psi_{ji} = \lambda_i^{-1}E[\langle X_{n-1}, v_i\rangle\langle X_n, v_j\rangle]$, which is not difficult to derive; λ_i is the eigenvalue corresponding to v_i, and $\hat{\lambda}_i$ the eigenvalue corresponding to $\hat{v}_i$. Using the estimated kernel (3), we calculate the predictions as

$$\hat{X}_{n+1}(t) = \int \hat{\psi}_p(t,s)X_n(s)ds = \sum_{k=1}^{p}\left(\sum_{\ell=1}^{p}\hat{\psi}_{k\ell}\left\langle X_n, \hat{v}_\ell\right\rangle\right)\hat{v}_k(t). \tag{5}$$

Predictive Factors (PF). The estimator (3) is not directly justified by the problem of optimal prediction, it is based on FPC's, which may focus on the features of the data that are not relevant to prediction. In this section, we describe a technique known as predictive factors, which may (potentially) be better suited for forecasting. It finds directions most relevant to prediction, rather than explaining the variability, as the FPC's do. We describe only the general idea, as theoretical arguments developed by Kargin and Onatski (2008) are quite complex. One of the messages of this paper is that the PF method does not offer an advantage in finite samples, so we do not need to describe here all the details.

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

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

where

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

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

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

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

3 Design of a simulation study

Data generating processes. The FAR(1) series are generated according to model

$$X_{n+1}(t) = \int_0^1 \psi(t,s)X_n(s)ds + \varepsilon_{n+1}(t), \quad n = 1, 2, \ldots, N. \tag{6}$$

We use a burn–in period of 50 functional observations.

We consider three error processes $\varepsilon^{(1)}(t)$, $\varepsilon^{(2)}(t)$, and $\varepsilon^{(3)}(t)$ defined as follows:

$$\varepsilon^{(1)}(t) = BB(t) = W(t) - tW(1), \tag{7}$$

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

$$W\left(\frac{k}{K}\right) = \frac{1}{\sqrt{K}} \sum_{j=1}^{k} Z_j, \quad k = 0, 1, 2, \ldots, K,$$

where the Z_k are independent standard normals and $Z_0 = 0$.

$$\varepsilon^{(2)}(t) = \xi_1\sqrt{2}\sin(2\pi t) + \sqrt{\lambda}\sqrt{2}\xi_2\cos(2\pi t), \tag{8}$$

where ξ_1 and ξ_2 are independent standard normals and λ can be any constant (in the simulations we use $\lambda = 0.5$).

$$\varepsilon^{(3)}(t) = \varepsilon^{(2)}(t) + a\varepsilon^{(1)}(t), \tag{9}$$

where a can be any constant.

The errors $\varepsilon^{(1)}$ are Brownian bridges, they admit the Karhunen–Loéve expansion with infinitely many terms. In contrast, the errors $\varepsilon^{(2)}(t)$ have only two terms in this expansion. The errors $\varepsilon^{(3)}(t)$ have two dominant terms, and the degree of their dominance is controlled by the parameter a.

The eigenfunctions and the eignevalues of the covariance operator of the observations X_n can differ significantly from those of the errors because they depend on the kernel ψ. We use four kernels (defined for $(t, s) \in [0, 1]^2$):

$$\begin{aligned}
\text{Gaussian: } \psi(t,s) &= C\exp\{-(t^2+s^2)/2\},\\
\text{Identity: } \psi(t,s) &= C,\\
\text{Sloping plane (t): } \psi(t,s) &= Ct,\\
\text{Sloping plane (s): } \psi(t,s) &= Cs.
\end{aligned}$$

The normalizing constants C are chosen such that $||\Psi|| = 0.5$ or $||\Psi|| = 0.8$. To implement the numerical integration in (6), the kernels were evaluated on a grid of 200×200 points.

Measures of quality of prediction and estimation. To measure the prediction error at time n, we use the quantities:

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

We use mean squared errors (MSE), averaged distance (AD) and ratio averaged distance (RAD) to measure the estimation of the kernel ψ:

$$MSE = \sqrt{\int_0^1 \int_0^1 \left(\hat{\psi}(t,s) - \psi(t,s)\right)^2 dsdt},$$

$$AD = \int_0^1 \int_0^1 |\hat{\psi}(t,s) - \psi(t,s)| dsdt,$$

$$RAD = \int_0^1 \int_0^1 \frac{|\hat{\psi}(t,s) - \psi(t,s)|}{|\psi(t,s)|} dsdt.$$

To present simulation results, we compute the appropriate averages, as explained in Sects. 4 and 5.

4 Improved estimation of the autoregressive kernel

We begin by illustrating the performance of estimator (3). Figure 1 shows the Gaussian kernel whose Hilbert–Schmidt norm (2) is 1/2, and three estimates which use $p = 2, 3, 4$. The innovations ε_n are generated as Brownian bridges. Such visual discrepancies are observed for other kernels and other innovation processes as well. Moreover, by all three measures, MSE, AD and RAD, the distance between ψ and $\hat{\psi}_p$ increases, as p increases. This is counterintuitive because by using more EFPC' $\hat{v}_j$, we would expect the approximation (3) to improve. For the FAR(1) used to produce Fig. 1, the sums $\sum_{j=1}^{p} \hat{\lambda}_j$ explain, respectively, 74, 83 and 87 percent of the variance for $p = 2, 3$ and 4, but (for the series length $N = 100$), the absolute deviation distances between ψ and $\hat{\psi}_p$ are 0.40, 0.44 and 0.55, see Table 1. As N increases, these distances decrease, but their tendency to increase with p remains. This problem is due in part to the fact that for many FAR(1) models, the estimated eigenvalues $\hat{\lambda}_j$ are very small, except $\hat{\lambda}_1$ and $\hat{\lambda}_2$, and so a small error in their estimation translates to a large error in the reciprocals $\hat{\lambda}_j^{-1}$ appearing in (3). We have experimented with many ways of remedying this, and found that a simple solution that gives a consistent

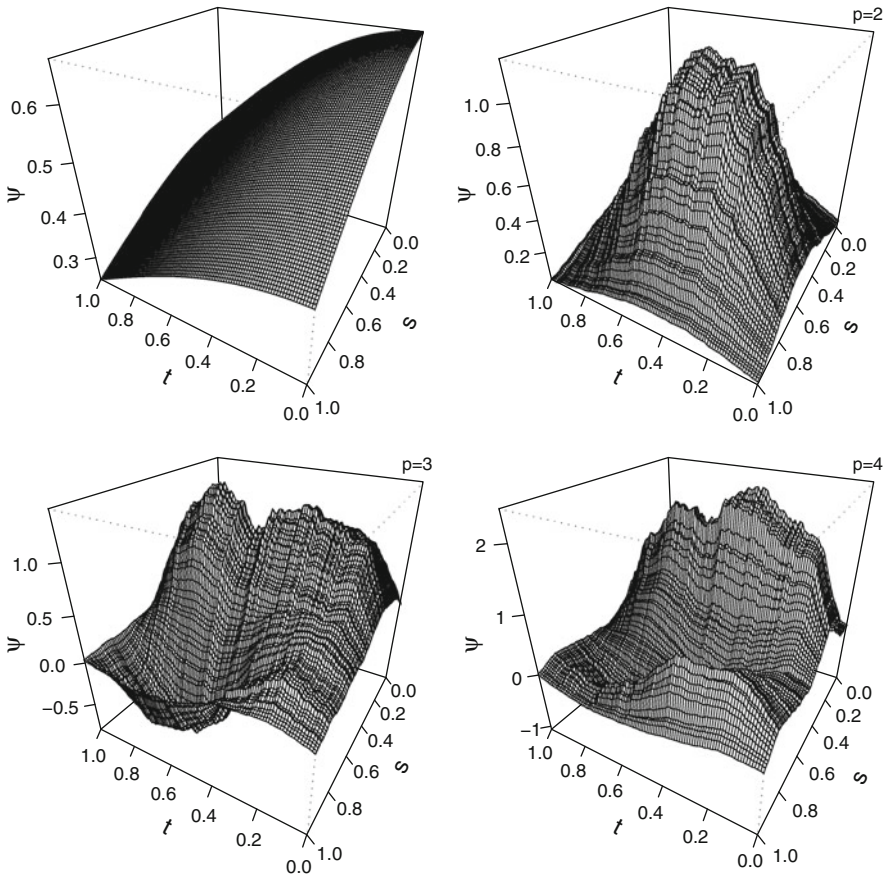

Fig. 1 The kernel surface $\psi(t, s)$ (*top left*) and its estimates $\hat{\psi}_p(t, s)$ for $p = 2, 3, 4$

improvement is to replace in (4), $\hat{\lambda}_i$ by $\hat{\lambda}_i + \hat{b},\ i > 2$, where $\hat{b} = 1.5(\hat{\lambda}_1 + \hat{\lambda}_2)$. Adding the baseline $\hat{b}$ does not make the estimated surfaces $\hat{\psi}_p$ look much more similar to ψ, but the errors, MSE, AD and RAD, become smaller and do not increase with p. The latter property is important, because the quality of original estimates depends strongly on p, so an estimator with a weaker dependence on p offers an advantage. This is illustrated in Tables 2 and 3. Note that in Table 3, the original method gives best results for $p = 2$. This is because the FAR(1) model for which ψ has been especially chosen to have slowly decaying eignevalues $\hat{\lambda}_i$. For the innovations $\varepsilon^{(2)}$, the original method fails completely for $p > 2$ and some kernels, because the estimated eigenvalues $\hat{\lambda}_i, i > 2$ are practically zero. Table 2 shows the most typical picture.

We presented in this section only a very small selection of graphs and tables; an extensive presentation is given in Kokoszka and Zhang (2010). Our findings pertain to lengths N occurring in most applications, in which N does not exceed 200.

Table 1 Kernel estimation errors; Brownian bridge innovations $\varepsilon^{(1)}(t)$, Gaussian kernel

	MSE	AD	RAD
$N = 50$			
$p = 2$	0.67 (0.04)	0.56 (0.03)	0.79 (0.04)
$p = 3$	10.17 (0.06)	0.93 (0.05)	10.30 (0.07)
$p = 4$	10.70 (0.07)	10.32 (0.05)	10.83 (0.07)
$N = 100$			
$p = 2$	0.46 (0.01)	0.40 (0.01)	0.56 (0.01)
$p = 3$	0.53 (0.02)	0.44 (0.02)	0.61 (0.02)
$p = 4$	0.67 (0.02)	0.55 (0.02)	0.77 (0.03)
$N = 200$			
$p = 2$	0.44 (0.01)	0.38 (0.01)	0.53 (0.01)
$p = 3$	0.43 (0.01)	0.37 (0.01)	0.51 (0.02)
$p = 4$	0.52 (0.01)	0.42 (0.01)	0.58 (0.01)

In parentheses, standard errors based on 50 replications

Table 2 AD errors for original and improved kernel estimates; innovations $\varepsilon^{(1)}$, $N = 100$

	$p = 2$		$p = 3$		$p = 4$	
	Original	Improved	Original	Improved	Original	Improved
Gaussian	0.400 (0.010)	0.400 (0.010)	0.440 (0.020)	0.390 (0.010)	0.550 (0.020)	0.380 (0.010)
Identity	0.520 (0.010)	0.243 (0.005)	0.540 (0.020)	0.220 (0.006)	0.630 (0.020)	0.220 (0.006)
Ct	0.280 (0.010)	0.280 (0.005)	0.340 (0.010)	0.274 (0.005)	0.480 (0.020)	0.270 (0.006)
Cs	0.300 (0.010)	0.255 (0.004)	0.370 (0.010)	0.245 (0.004)	0.510 (0.020)	0.237 (0.005)

In parentheses, standard errors based on 50 replications

Table 3 AD errors for original and improved kernel estimates; innovations $\varepsilon^{(3)}$, $a = 3$, $N = 100$

	$p = 2$		$p = 3$		$p = 4$	
	Original	Improved	Original	Improved	Original	Improved
Gaussian	0.398 (0.008)	0.439 (0.004)	0.337 (0.009)	0.403 (0.004)	0.587 (0.020)	0.407 (0.003)
Identity	0.390 (0.007)	0.444 (0.003)	0.312 (0.009)	0.411 (0.003)	0.619 (0.020)	0.411 (0.003)
Ct	0.396 (0.008)	0.386 (0.004)	0.393 (0.010)	0.364 (0.004)	0.631 (0.020)	0.360 (0.002)
Cs	0.327 (0.007)	0.384 (0.004)	0.291 (0.008)	0.358 (0.003)	0.584 (0.030)	0.357 (0.003)

In parentheses, standard errors based on 50 replications

5 Comparison of prediction methods

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

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

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

EX (Exact) We set $\hat{X}_{n+1} = \Psi(X_n)$. This is not really a prediction method because the autoregressive operator Ψ is unknown. It is included to see if poor predictions might be due to poor estimation of Ψ (see Sect. 4).

EK (Estimated Kernel) This method is described in Sect. 2.

EKI (Estimated Kernel Improved) This is method EK, but the $\hat{\lambda}_i$ in (4) are replaced by $\hat{\lambda}_i + \hat{b}$, as described in Sect. 4.

PF (Predictive Factors) This method is described in Sect. 2.

We produced boxplots of the errors E_n and R_n, $N - 50 < n < N$, defined in Sect. 3, for the innovations and kernels defined in Sect. 3. We considered $N = 50, 100, 200$, $||\Psi|| = 0.5$ and $||\Psi|| = 0.8$. Typical results are shown in Figs. 2 and 3, but for some choices of the innovations and the kernels, the relative placement of the boxplots changes. To assess the statistical significance of the results, we computed the averages of the E_n and R_n, $N - 50 < n < N$, and the standard errors of these averages. Typical examples are given in Tables 4 and 5. For example, in Table 5, the average R_n for kernel Ct with $||\Psi|| = .8$ is 0.30 for EK and 0.36 for PF. Given that the standard error is 0.02, the corresponding population averages are significantly different at 5% level. This agrees with the boxplots in Fig. 3. The standard errors like

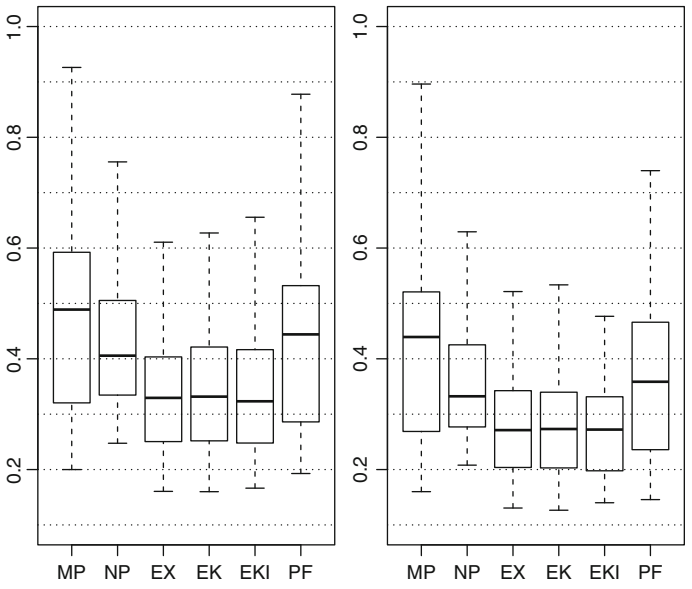

Fig. 2 Boxplots of the prediction errors E_n (*left*) and R_n (*right*); innovations: $\varepsilon^{(1)}$, kernel: sloping plane (t), $N = 100$, $p = 3$, $||\Psi|| = 0.5$

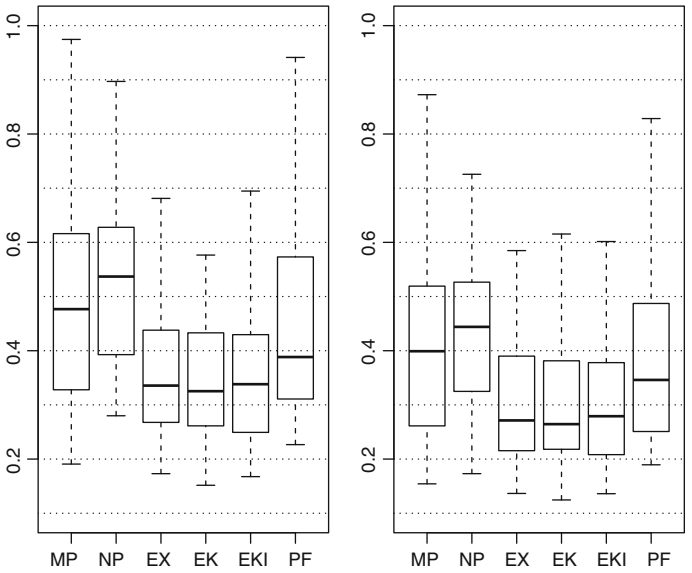

Fig. 3 Boxplots of the prediction errors E_n (*left*) and R_n (*right*); innovations: $\varepsilon^{(1)}$, kernel: sloping plane (t), $N = 100,\ p = 3,\ ||\Psi|| = 0.8$

Table 4 Averages of E_n and R_n and their standard errors; $\varepsilon^{(1)}(t)$, $p = 3$, $||\Psi|| = 0.5$

	MP	NP	EX	EK	EKI	PF
E_n						
Gaussian	0.41 (0.03)	0.47 (0.02)	0.38 (0.02)	0.37 (0.02)	0.37 (0.02)	0.37 (0.02)
Identity	0.39 (0.03)	0.49 (0.03)	0.35 (0.02)	0.35 (0.02)	0.34 (0.02)	0.37 (0.02)
Ct	0.44 (0.02)	0.57 (0.03)	0.42 (0.03)	0.42 (0.03)	0.42 (0.03)	0.44 (0.03)
Cs	0.39 (0.02)	0.48 (0.03)	0.37 (0.02)	0.36 (0.02)	0.36 (0.02)	0.36 (0.02)
R_n						
Gaussian	0.35 (0.03)	0.38 (0.02)	0.31 (0.02)	0.31 (0.02)	0.31 (0.02)	0.31 (0.02)
Identity	0.32 (0.02)	0.41 (0.03)	0.29 (0.02)	0.29 (0.02)	0.29 (0.02)	0.30 (0.02)
Ct	0.38 (0.02)	0.48 (0.03)	0.36 (0.02)	0.36 (0.02)	0.36 (0.02)	0.37 (0.02)
Cs	0.33 (0.02)	0.40 (0.02)	0.30 (0.02)	0.30 (0.02)	0.30 (0.02)	0.30 (0.02)

those in Tables 4 and 5 are computed assuming that the E_n and R_n are uncorrelated. This is confirmed by the examination of the ACF plots, and can be expected because prediction errors are close to the iid model errors.

Since we cannot present all 32 sets of boxplots and 32 sets of tables, we report only the general conclusions:

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

Table 5 Averages of E_n and R_n and their standard errors; $\varepsilon^{(1)}(t)$, $p = 3$, $||\Psi|| = 0.8$

	MP	NP	EX	EK	EKI	PF
E_n						
Gaussian	0.46 (0.02)	0.52 (0.03)	0.40 (0.02)	0.40 (0.02)	0.40 (0.02)	0.55 (0.02)
Identity	0.40 (0.03)	0.48 (0.03)	0.36 (0.02)	0.36 (0.02)	0.36 (0.02)	0.36 (0.02)
Ct	0.46 (0.02)	0.50 (0.02)	0.37 (0.02)	0.36 (0.02)	0.37 (0.02)	0.42 (0.02)
Cs	0.49 (0.03)	0.48 (0.03)	0.38 (0.02)	0.36 (0.02)	0.36 (0.02)	0.36 (0.02)
R_n						
Gaussian	0.39 (0.02)	0.43 (0.02)	0.33 (0.02)	0.33 (0.02)	0.33 (0.02)	0.48 (0.04)
Identity	0.33 (0.02)	0.39 (0.02)	0.30 (0.02)	0.30 (0.02)	0.30 (0.02)	0.30 (0.02)
Ct	0.39 (0.02)	0.41 (0.02)	0.30 (0.02)	0.30 (0.02)	0.31 (0.02)	0.36 (0.02)
Cs	0.43 (0.03)	0.39 (0.02)	0.32 (0.02)	0.30 (0.02)	0.30 (0.02)	0.30 (0.02)

2. None of the Methods EX, EK, EKI uniformly dominates the other. In most cases method EK is the best, or at least as good at the others.
3. In some cases, method PF performs visibly worse than the other methods, but always better than NP.
4. Using the improved estimation described in Sect. 4 does not generally reduce prediction errors.

We also applied all prediction methods to mean corrected precipitation data studied in Besse et al (2000). For this data set, the averages of the E_n and the R_n are not significantly different between the first five methods, method PF performs significantly worse than the others. We should point out that method PF depends on the choice of the parameters α and k. It is possible that its performance can be improved by better tuning these parameters. On the other hand, our simulations show that method EK essentially reaches the limit of what is possible, it is comparable to the theoretically perfect method EX.

6 Limits of prediction quality

In this section we provide some discussion of the empirical findings reported in Sect. 5.

Our simulation study shows that while taking into account the autoregressive structure of the observations does reduce prediction errors, the boxplots in Figs. 2 and 3 suggest that many prediction errors are comparable to those of the trivial MP method. To analyze this observation further, we present in Fig. 4 six consecutive trajectories of the FAR(1) process with Gaussian kernel, $||\Psi|| = 0.5$, and Brownian bridge innovations together with EK predictions. Predictions obtained with other methods look similar. We see that the predictions look much smoother than the observations, and their range is much smaller. If the innovations ε_n are smooth, like the $\varepsilon_n^{(2)}$, the observations and their predictions are also smooth, but the predicted curves have a visibly smaller range than the observations. This is further illustrated in Fig. 5 which shows centered precipitation curves studied by Besse et al (2000) together with their EK

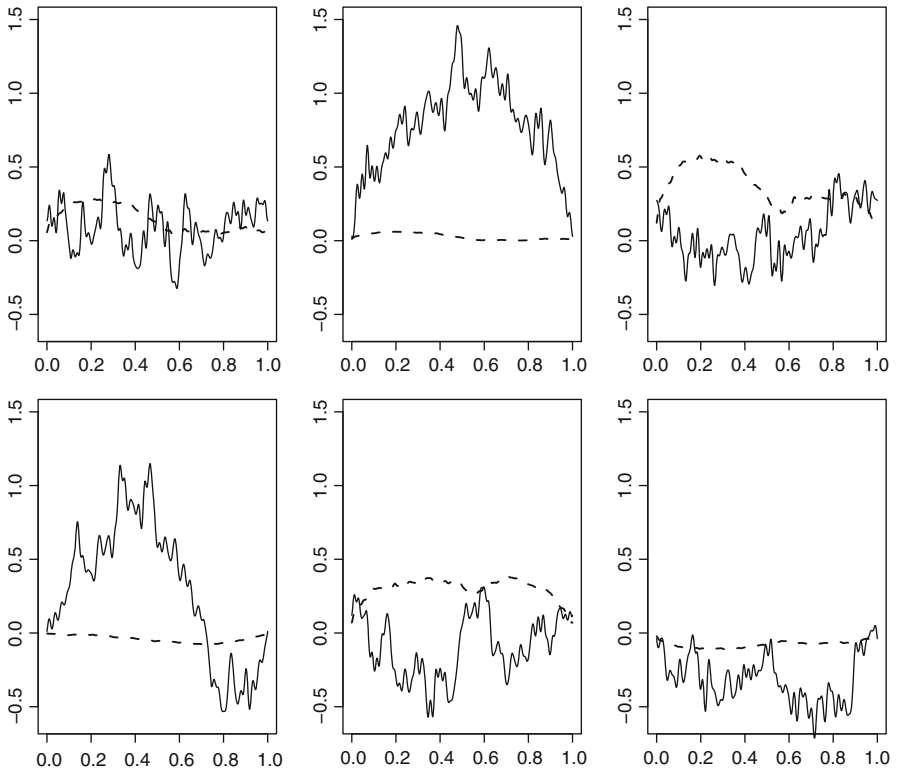

Fig. 4 Six consecutive trajectories of the FAR(1) process with Gaussian kernel, $||\Psi|| = 0.5$, and Brownian bridge innovations. *Dashed lines* show EK predictions with $p = 3$

predictions. We estimate the mean function for precipitation data first, subtract from the curves, and then do the forecasting for the centered data.

The smoothness of the predicted curves follows from representation (5), which shows that each predictor is a linear combination of a few EFPC's, which are smooth curves themselves. The smaller range of the the predictors is not peculiar to functional data, but is enhanced in the functional setting. For a mean zero scalar AR(1) process $X_n = \psi X_{n-1} + \varepsilon_n$, we have $\mathrm{Var}(X_n) = \psi^2 \mathrm{Var}(X_{n-1}) + \mathrm{Var}(\varepsilon_n)$, so the variance of the predictor $\hat{\psi} X_{n-1}$ is about ψ^{-2} times smaller than the variance of X_n. In the functional setting, the variance of $\hat{X}_n(t)$ is close to $\mathrm{Var}[\int \psi(t,s) X_n(s) ds]$. If the kernel ψ admits the decomposition $\psi(t,s) = \psi_1(t)\psi_2(s)$, as all the kernels we use do, then

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

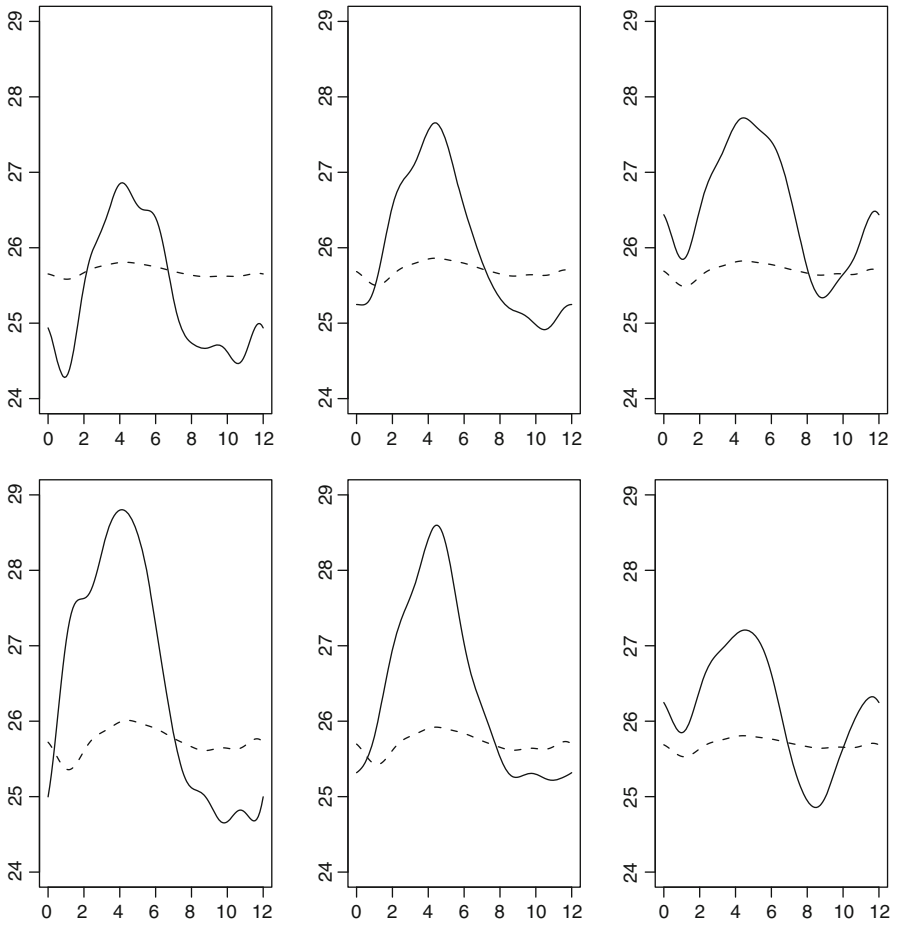

Fig. 5 Six consecutive trajectories (1989–1994) of centered pacific precipitation curves (*solid*) with their EK predictions (*dashed*)

If the function ψ_1 is small for some values of $t \in [0, 1]$, it will automatically drive down the predictions. If ψ_2 is small for some $s \in [0, 1]$, it will reduce the integral $\int_0^1 \psi_2(s) X_{n-1}(s)$. For the Gaussian kernel, $\psi_1 = \psi_2$ are small for arguments less than 1/2, so the predictions are very small, as seen in Fig. 4. The estimated kernels do not, in general admit a factorization of this type, but are always weighted sums of products of orthonormal functions (the EFPC's $\hat{v}_k$). A conclusion of this discussion is that the predicted curves will in general look smoother and "smaller" than the data. This somewhat disappointing performance is however, not due to poor prediction methods, but to a natural limit of predictive power of the FAR(1) model; the curves $\Psi(X_n)$ share the general properties of the curves $\hat{\Psi}(X_n)$, no matter how Ψ is estimated.

Acknowledgments This research was partially supported by NSF grants DMS-0804165 and DMS-0931948.

References

Antoniadis A, Sapatinas T (2003) Wavelet methods for continuous time prediction using Hilbert–valued autoregressive processes. J Multivar Anal 87:133–158
Besse P, Cardot H (1996) Approximation spline de la prévision d'un processus fonctionnel autorégressif d'ordre 1. Can J Stat 24:467–487
Besse P, Cardot H, Stephenson D (2000) Autoregressive forecasting of some functional climatic variations. Scand J Stat 27:673–687
Bosq D (2000) Linear processes in function spaces. Springer, New York
Bosq D, Blanke D (2007) Inference and prediction in large dimensions. Wiley, New Jersey
Damon J, Guillas S (2002) The inclusion of exogenous variables in functional autoregressive ozone forecasting. Environmetrics 13:759–774
Ferraty F, Romain Y (eds) (2011) The oxford handbook of functional data analysis. Oxford University Press, Oxford
Ferraty F, Vieu P (2006) Nonparametric functional data analysis: theory and practice. Springer, New York
Gabrys R, Horváth L, Kokoszka P (2010) Tests for error correlation in the functional linear model. J Am Stat Assoc 105:1113–1125
Hörmann S, Kokoszka P (2010) Weakly dependent functional data. Ann Stat 38:1845–1884
Hörmann S, Kokoszka P (2011) Consistency of the mean and the principal components of spatially distributed functional data. Tech. rep., Utah State University, Logan
Horváth L, Kokoszka P (2011+) Inference for functional data with applications. Springer Series in Statistics, Springer, forthcoming
Horváth L, Hušková M, Kokoszka P (2010) Testing the stability of the functional autoregressive process. J Multivar Anal 101:352–367
Kargin V, Onatski A (2008) Curve forecasting by functional autoregression. J Multivar Anal 99:2508–2526
Kokoszka P, Zhang X (2010) Improved estimation of the kernel of the functional autoregressive process. Tech. rep., Utah State University, Logan
Ramsay J, Hooker G, Graves S (2009) Functional data analysis with R and MATLAB. Springer, New York
Ramsay JO, Silverman BW (2002) Applied functional data analysis. Springer, New York
Ramsay JO, Silverman BW (2005) Functional data analysis. Springer, New York