Original Article

Journal of Time Series Analysis

First version received May 2012

Published online in Wiley Online Library: 25 July 2012

(wileyonlinelibrary.com) DOI: 10.1111/j.1467-9892.2012.00816.x

Determining the order of the functional autoregressive model

Piotr Kokoszka[a] and Matthew Reimherr[b,*,†]

We propose a multistage testing procedure to determine the order p of a functional autoregressive process, FAR(p). At its core is the representation of the FAR(p) process as a fully functional linear model with dependent regressors. Estimating the kernel function in this linear model allows us to construct a test statistic which has, approximately, a chi–square distribution with the number of degrees of freedom determined by the number of functional principal components used to represent the data. The asymptotic justification relies on the concept of L^p–m–approximability which quantifies the temporal dependence of functional time series. The procedure enjoys very good finite sample properties, as confirmed by a simulation study and applications to functional time series derived from credit card transactions and Eurodollar futures data.

Keywords: Autoregressive order; functional data analysis; time series.

1. INTRODUCTION

Over the last two decades, functional data analysis has established itself as an important and dynamic area of statistics. It offers effective new tools and has stimulated new methodological and theoretical developments. The collection of Ferraty and Romain (2011) covers many recent developments in this field. The books of Ferraty and Vieu (2006) and Ramsay *et al.* (2009) focus, respectively, on nonparametric methods and computational issues. The monograph of Ramsay and Silverman (2005) offers an excellent and accessible introduction to to the central ideas of the field, while Bosq (2000) and Bosq and Blanke (2007) study its mathematical foundations.

This study focusses on an application of the ideas of functional data analysis to curves observed consecutively over time, which we may call a functional time series. The curves are often obtained by cutting a long temporal record, $Z(u), u > 0$, into natural consecutive intervals by setting

$$Z_i(t) = Z(i + t), \quad t \in [0, 1], \quad i = 0, 1, 2, \ldots$$

The interval [0,1] corresponds to a natural rescaled period of the record $Z(u), u > 0$; for example, to a year for climatic data, a week or day for pollution records. Functional time series need not however arise only in this way. The Eurodollar futures series discussed in Section 4 is not of this type. Hörmann and Kokoszka (2012) discuss several examples and review some recent developments in this field.

The most popular statistical model for data of this type is the functional autoregressive model of order 1, FAR(1), defined by the recursion

$$Z_i = \Phi(Z_{i-1}) + \varepsilon_i,$$

in which Φ is a linear operator acting on a fuction space, typically the Hilbert space of square integrable functions, and the ε_i are i.i.d. error functions in that space. In applications, Φ is assumed to be a kernel operator, so the FAR(1) equations take the form

$$Z_i(t) = \int \phi(t, s) Z_{i-1}(s) \mathrm{d}s + \varepsilon_i(t), \quad 0 \le t \le 1.$$

Unless specified otherwise, integration is over the interval [0,1]. The FAR(1) model has been applied and studied in many contexts, we list Besse *et al.* (2000), Damon and Guillas (2002), Antoniadis and Sapatinas (2003), Antoniadis *et al.* (2006), Kargin and Onatski (2008), Horváth *et al.* (2010) and Didericksen *et al.* (2012), among many others. The theory of autoregressive and more general linear processes in Hilbert and Banach spaces is developed in the monograph of Bosq (2000).

This study is concerned with determining the order p in the FAR(p) model

$$Z_i = \sum_{j=1}^{p} \Phi_j(Z_{i-j}) + \varepsilon_i. \tag{1}$$

[a]*Colorado State University*
[b]*University of Chicago*
**Correspondence to: Matthew Reimherr, Department of Statistics, University of Chicago, 5734 S. University Avenue, Ryerson N375, Chicago, IL 60637, USA.*
[†]*E-mail: mreimherr@uchicago.edu*

Order selection has been a central problem in time series analysis, and the resulting research has had a transforming impact on the application of time series models. The literature is very extensive, but we must mention the pioneering work of Akaike (1978), Hannan and Quinn (1979), Hannan (1980), Shibata (1980) and Hannan and Rissanen (1982). A comprehensive review is provided by Bhansali (1993), and a brief introduction in Section 9.3 of Brockwell and Davis (1991). Surprisingly, no serious attention has been devoted to the problem of order selection for functional time series, and this study's aim is to provide a practically applicable procedure which is suitable for functional data. In contrast to scalar autoregression, only very small values of p, say 0,1,2, would be of interest because the curves $Z_k(t)$ already consist of a large number of scalar observations, often hundreds, and the goal of functional data analysis is to replace all of these observations by a single functional object. Consequently, we do not attempt to develop analogues of the well-known information or prediction error-based order selection criteria, but propose an approach based on hypothesis testing. Our approach is specifically designed for functional data and fundamentally differs from the approaches in common use for scalar and vector valued time series. It relies on the observation that the union of intervals, $[0,1]\cup[1,2]\cup\cdots\cup[(p-1),p]$, is again an interval (which can be treated as a unit interval after rescaling), and on a multistage testing procedure rather than penalized likelihood.

The issue of determining an optimal order p can be approached in a problem-specific manner by checking if using the FAR(p) produces better results than using FAR($p-1$), in a sense defined by a statistical problem at hand. Nevertheless, an appropriate criterion may not be obvious, and we believe that a universal approach that can be applied to any such situation is useful. In this study we propose a suitable testing procedure. We focus on practical applicability, but we also provide a large sample justification. A somewhat related contribution was made by Gabrys and Kokoszka (2007) who considered testing the null hypothesis that the curves are i.i.d. but did not assume a specific alternative. Their test does not perform as well as the test proposed here when testing the i.i.d. hypothesis against the FAR(1) model.

The article is organized as follows. In Section 2, we state model assumptions and develop a representation and an estimation technique for the FAR(p) process suitable for the testing problem. Section 3 describes the testing procedure whose performance is assessed in Section 4 by a simulation study and application to credit card transaction and Eurodollar futures data. We conclude with Section 5 which contains the proofs of the asymptotic results which justify our procedure. These asymptotic results are of independent interest because they concern kernel estimation in the extensively used fully functional linear model without assuming the independence of the regressor/response pairs.

2. REPRESENTATION OF AN FAR(*P*) PROCESS AS A FUNCTIONAL LINEAR MODEL

By L^2 we denote the space of square intergrable functions on [0,1] with the usual inner product $\langle\cdot,\cdot\rangle$ and the norm $\|\cdot\|$. We will often work with the direct products

$$(x\otimes y)(t,s)=x(s)y(t),\quad x,y\in L^2,$$

which are elements of the space $L^2([0,1]\times[0,1])$. The inner product in the latter space will also be denoted by $\langle\cdot,\cdot\rangle$, as it will always be clear from the context what space the product is in.

We observe a sample of curves $Z_1(t),Z_2(t),\ldots,Z_N(t),t\in[0,1]$. We assume that these curves are a part–realization of an infinite sequence $\{Z_j\}$ which satisfies the following assumptions.

Assumption 1. *The operators Φ_j in* (1) *are Hilbert–Schmidt integral operators in L^2, i.e.*

$$\Phi_j(x)(t)=\int\phi_j(t,s)x(s)\mathrm{d}s,\quad \int\int\phi_j^2(t,s)<\infty. \tag{2}$$

The operator

$$\Phi'=\begin{bmatrix}\Phi_1 & \Phi_2 & \cdots & \Phi_{p-1} & \Phi_p\\ I & 0 & \cdots & 0 & 0\\ 0 & I & \cdots & 0 & 0\\ \vdots & \vdots & \vdots & \vdots & \vdots\\ 0 & 0 & \cdots & I & 0\end{bmatrix} \tag{3}$$

acting on the cartesian product $(L^2)^p$ satisfies

$$\|\Phi'\|_{\mathcal{L}}<1, \tag{4}$$

where $\|\cdot\|_{\mathcal{L}}$ is the operator norm in the cartesian product $(L^2)^p$.

Assumption 2. *The $\varepsilon_i\in L^2$ in (1) are i.i.d.*

Condition (4) and Assumption 2 imply that the Z_i form a stationary and ergodic sequence in L^2 such that ε_i is independent of $Z_{i-1}, Z_{i-2}, \ldots$, see Section 5.1 of Bosq (2000). For ease of reference, we state the following definition.

DEFINITION 1. *We say that the functional observations $Z_1, Z_2, \ldots, Z_N$ follow an FAR(p) process if Φ_p is not the zero operator, and Assumptions 1 and 2 hold.*

Sufficient conditions for (4) to hold are established in Chapter 5 of Bosq (2000). A condition analogous to the usual condition for the existence of a scalar AR(p) process is the following: if the operator

$$Q_p(z) = z^p I - \sum_{j=1}^{p} z^{p-j}\Phi_j$$

does not have a bounded inverse, then $|z| < 1$. A stronger condition is $\sum_{j=1}^{p} \|\Phi_j\| < 1$. These conditions are derived using a Markovian representation of the process (1) as an FAR(1) process in the cartesian product $(L^2)^p$. For the task of testing FAR(p) against FAR($p+1$), a different representation is useful. It directly uses the structure of the observations as curves, and of the kernels ϕ_j as surfaces, rather than treating them as elements of abstract Hilbert spaces.

We start by expressing $\Phi_j(Z_{i-j})$ as an integral over the interval $((j-1)/p, j/p]$. Setting $x := (s + j - 1)/p$, a change of variables yields

$$[\Phi_j(Z_{i-j})](t) = \int_0^1 \phi_j(t,s)Z_{i-j}(s)\mathrm{d}s = \int_{(j-1)/p}^{j/p} \phi_j(t, xp-(j-1))Z_{i-j}(xp-(j-1))p\ \mathrm{d}x.$$

Denoting by I_j the indicator function of the interval $((j-1)/p, j/p]$, we obtain

$$\sum_{j=1}^{p}[\Phi_j(Z_{i-j})](t) = \int_0^1 \sum_{j=1}^{p} I_j(x)\phi_j(t, xp-(j-1))Z_{i-j}(xp-(j-1))p\ \mathrm{d}x.$$

Next we define

$$X_i(s) = \sum_{j=1}^{p} Z_{i-j}(sp-(j-1))I_j(s) \tag{5}$$

and

$$\psi(t,s) = p\sum_{j=1}^{p} \phi_j(t, sp-(j-1))I_j(s). \tag{6}$$

Setting $Y_i = Z_i$, we have

$$Y_i = \Psi(X_i) + \varepsilon_i, \tag{7}$$

where Ψ is an integral Hilbert–Schmidt operator with the kernel ψ, i.e.

$$Y_i(t) = \int \psi(t,s)X_i(s)\mathrm{d}s + \varepsilon_i(t). \tag{8}$$

Thus, if we can estimate Ψ, then we can estimate each of the Φ_j. The FAR($p-1$) model will be rejected in favour of FAR(p) if the resulting estimate of $\hat{\Phi}_p$ is large in a sense established in section 3. We now turn to the estimation of the operator Ψ.

Let $\{\hat{v}_k, 1 \le k \le N\}$ be an orthonormal basis of L^2 (for each N), constructed from the eigenfunctions of the covariance operator

$$\hat{C}_X(t,s) = \frac{1}{N}\sum_{i=1}^{N}(X_i(t) - \bar{X}_N(t))(X_i(s) - \bar{X}_N(s)),$$

ordered by the corresponding eigenvalues $\hat{\lambda}_k$. To construct the test statistic, we will use only the first q_x eigenfunction/eigenvalue pairs $(\hat{v}_k, \hat{\lambda}_k)$. While we will use $\{\hat{v}_k\}$ in projecting the regressors, we allow for a separate basis in projecting the response variables. Define $\{\hat{u}_j\}$ analogously to $\{\hat{v}_k\}$ for the response functions. The tuning parameter q_y can be chosen to explain about 80–90 % of the variance of the Y_i. One can take q_x analogously to q_y, however since X_i is constructed by piecing together p of the Z_i one could also justify taking $q_x = q_y p$ (since $Z_i = Y_i$) to better capture the variability of each piece. While the latter results in a much larger q_x, our procedure will involve a truncation step that will bring it back in line with q_y. In our experience, taking $q_x = q_y p$ results in a slightly more powerful procedure, though both approaches are valid. We take $q_x = q_y p$ for the simulations and applications presented in this study.

We estimate ψ projected onto the random subspace

$$\hat{H}_{q_x,q_y} := \mathrm{span}\{\hat{v}_1, \ldots, \hat{v}_{q_x}\} \times \mathrm{span}\{\hat{u}_1, \ldots, \hat{u}_{q_y}\}.$$

Let $\hat{\pi}_{q_x,q_y}$ denote the projection operator onto $\hat{H}_{q_x,q_y}$. Then we wish to estimate $\hat{\pi}_{q_x,q_y}(\psi)$. We should mention that this differs sharply from an analogous multivariate problem. While we wish to estimate ψ, we can only estimate ψ projected onto a finite dimensional subspace. Furthermore, that subspace is actually random since the space we choose depends on the random operators $\hat{C}_X$ and $\hat{C}_Y$. We devolop an asymptotic framework that handles these issues in Section 5.

To construct a least squares estimator, we define for $i = 1,\ldots,N$, $j = 1,\ldots,q_y$, and $k = 1,\ldots,q_x$

$$\mathbf{Y}(i,j) = \langle Y_i, \hat{u}_j\rangle, \quad \mathbf{X}(i,k) = \langle X_i, \hat{v}_k\rangle, \tag{9}$$

$$\boldsymbol{\psi}(k,j) = \langle \psi, \hat{v}_k \otimes \hat{u}_j\rangle = \int\int \psi(t,s)\hat{v}_k(s)\hat{u}_j(t)\mathrm{d}t\mathrm{d}s.$$

For ease of reference, we list the dimensions of the matrices introduced above

$$\mathbf{Y}(N \times q_y), \quad \mathbf{X}(N \times q_x), \quad \boldsymbol{\psi}(q_x \times q_y).$$

Using these matrices, we now reduce model (7) to a finite dimensional linear model. The precision of this finite dimensional approximation will be reflected in the structure of its random errors. Observe that

$$\mathbf{Y}(i,j) = \langle Y_i, \hat{u}_j\rangle = \langle \Psi(X_i) + \varepsilon_i, \hat{u}_j\rangle = \langle \Psi(X_i), \hat{u}_j\rangle + \langle \varepsilon_i, \hat{u}_j\rangle.$$

Since Ψ has a kernel ψ and $\{\hat{v}_k\}$ forms a basis for L^2, we have

$$\begin{aligned}
&\langle \Psi(X_i), \hat{u}_j\rangle + \langle \varepsilon_i, \hat{u}_j\rangle = \langle \psi, X_i \otimes \hat{u}_j\rangle + \langle \varepsilon_i, \hat{u}_j\rangle \\
&= \left\langle \psi, \sum_{k=1}^{\infty}\langle X_i, \hat{v}_k\rangle \hat{v}_k \otimes \hat{u}_j \right\rangle + \langle \varepsilon_i, \hat{u}_j\rangle \\
&= \sum_{k=1}^{\infty}\langle X_i, \hat{v}_k\rangle\langle \psi, \hat{v}_k \otimes \hat{u}_j\rangle + \langle \varepsilon_i, \hat{u}_j\rangle \\
&= \sum_{k=1}^{q_x}\langle X_i, \hat{v}_k\rangle\langle \psi, \hat{v}_k \otimes \hat{u}_j\rangle + \langle \varepsilon_i, \hat{u}_j\rangle + \sum_{k=q_x+1}^{\infty}\langle X_i, \hat{v}_k\rangle\langle \psi, \hat{v}_k \otimes \hat{u}_j\rangle \\
&= \sum_{k=1}^{q_x}\mathbf{X}(i,k)\boldsymbol{\psi}(k,j) + \langle \varepsilon_i, \hat{u}_j\rangle + \sum_{k=q_x+1}^{\infty}\langle X_i, \hat{v}_k\rangle\langle \psi, \hat{v}_k \otimes \hat{u}_j\rangle.
\end{aligned}$$

Therefore, the projections lead to the multivariate relation

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\psi} + \boldsymbol{\varepsilon}',$$

The $N \times q_y$ matrix $\boldsymbol{\varepsilon}$ has absorbed the error we made in projecting onto a finite dimensional space, and is given by

$$\boldsymbol{\varepsilon}'(i,j) = \langle \varepsilon_i, \hat{u}_j\rangle + \sum_{l>q_x}\langle \mathbf{X}_i, \hat{v}_l\rangle\langle \psi, \hat{v}_l \otimes \hat{u}_j\rangle.$$

Observe also that the matrix $\boldsymbol{\psi}$ is not a population parameter, it is a projection of an unknown kernel function ψ onto a random subspace. It is therefore a random matrix. We can nevertheles compute the usual least squares estmator

$$\hat{\boldsymbol{\psi}} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{Y} \tag{10}$$

and use its entries to $\hat{\boldsymbol{\psi}}(k,j), k \le q_x, j \le q_y$, to construct a test statistic, as described in Section 3. The asymptotic properties of the estimator $\hat{\boldsymbol{\psi}}$ are established in Section 5.

3. DETERMINING THE ORDER OF AN FAR PROCESS

We begin by describing a multistage testing procedure to determine the order of an FAR process. It consists of a sequence of tests of the hypotheses

$$H_p : \{Z_i\} \text{ are FAR}(p).$$

We start by testing

Null hypothesis: $=$ $H_0 : \{Z_i\}$ are i.i.d. vs. Alternative hypothesis $:=$ $H_1 : \{Z_i\}$ are FAR(1).

If we accept H_0, then we conclude that the observations can be assumed to be i.i.d. If we reject H_0, then we make H_1 our new null hypothesis and H_2 our new alternative. We continue until we accept a null hypothesis. We then conclude that the process is of the corresponding order. As explained in the Introduction, the number of the individual tests will, in practice, be very small, one or two,

so we are not concerned with problems arising in testing a large number of hypotheses. Our goal is consequently to construct a statistic to test the null hypothesis H_{p-1} against the alternative H_p.

We now describe how we have constructed such a test statistic. As will be clear from the exposition that follows, some other variants are possible, but we focus on only one that seems most direct to us and leads to a test with very good finite sample properties. The test algorithm is summarized at the end of this section.

Using the estimator 10, we obtain an estimator of the kernel ψ given by

$$\hat{\psi}(t,s) = \sum_{k \le q_x, j \le q_y} \hat{\psi}(k,j)\hat{v}_k(s)\hat{u}_j(t). \tag{11}$$

By (6), we can estimate the kernel ϕ_p by

$$\hat{\phi}_p(t,s) = \frac{1}{p}\hat{\psi}\left(t, \frac{s+p-1}{p}\right) = \frac{1}{p}\sum_{k \le q_x, j \le q_y} \hat{\psi}(k,j)\hat{v}_k\left(\frac{s+p-1}{p}\right)\hat{u}_j(t).$$

Testing the nullity of ϕ_p is thus equivalent to checking if the sum

$$\sum_{k \le q_x, j \le q_y} \hat{\psi}(k,j)\hat{v}_k(x)\hat{u}_j(t), \quad \frac{p-1}{p} \le x \le 1, \quad 0 \le t \le 1 \tag{12}$$

is close to zero. The key element is the range of the argument x of $\hat{v}_k$, which reflects the part of ψ whose nullity we want to test. Based on the above representation, we want to find linear combinations of the $\hat{\psi}(k,j)$ which make the sum (12) small. Clearly, we do not want to test if all $\hat{\psi}(k,j)$ are small because that would mean that the whole kernel ψ and so all of the $\phi_j, 1 \le j \le p$, vanish. For further discussion, it is convenient to set

$$\hat{v}_{k,p}(s) = \hat{v}_k\left(\frac{s+p-1}{p}\right), \quad 0 \le s \le 1,$$

so that

$$\hat{\phi}_p(t,s) = \frac{1}{p}\sum_{k \le q_x, j \le q_y} \hat{\psi}(k,j)\hat{v}_{k,p}(s)\hat{u}_j(t), \quad 0 \le s, \quad t \le 1.$$

The idea behind the construction of the test statistic is to replace the $\hat{v}_{k,p}$ by a smaller set of functions that optimally describe the space spanned by them, and so, in a sense, by the $\hat{v}_k(x), x \ge (p-1)/p$. In other words, we test the nullity of ϕ_p only in the most significant orthogonal directions of the $\hat{v}_{k,p}$. We orthogonalize them as

$$\hat{w}_{k,p}(s) = \sum_{i=1}^{q_x} \hat{\alpha}_{i,k}\hat{v}_{i,p}(s)$$

with the vectors

$$\hat{\boldsymbol{\alpha}}_k = [\hat{\alpha}_{1,k}, \hat{\alpha}_{2,k}, \ldots, \hat{\alpha}_{q_x,k}]^{\mathsf{T}}$$

such that $\|\hat{\boldsymbol{\alpha}}_k\| = 1$. To accomplish this, we construct the $q_x \times q_x$ matrix $\hat{\mathbf{V}}$ whose entries are the inner products

$$\hat{V}(k,k') = \langle \hat{v}_{k,p}, \hat{v}_{k',p} \rangle. \tag{13}$$

Since the matrix $\hat{\mathbf{V}}$ is positive–definite and symmetric, we define the $\hat{\boldsymbol{\alpha}}_k$ as its orthonormal eigenvectors ordered by their eigenvalues, that is, we have

$$\hat{\mathbf{V}}\hat{\boldsymbol{\alpha}}_k = \hat{\gamma}_k\hat{\boldsymbol{\alpha}}_k, \quad 1 \le k \le q_x, \tag{14}$$

where

$$\hat{\gamma}_1 \ge \hat{\gamma}_2 \ge \cdots \ge \hat{\gamma}_{q_x}.$$

A direct verification shows that

$$\langle \hat{w}_{k,p}, \hat{w}_{k',p} \rangle = \hat{\gamma}_k\delta_{k,k'},$$

where $\delta_{k,k'}$ is Dirac's delta.

Next we project $\hat{\phi}_{p+1}$ onto the functions $\{\hat{w}_{k,p} \otimes \hat{u}_j\}$. However, we will only include $\hat{w}_{k,p}$ whose norms are above a certain threshold, as the larger the value of $\|\hat{w}_{k,p}\|$ the greater its role in estimating ϕ_{p+1}. We obtained very good empirical performance by setting

$$q_\star = \max\{k \in \{1, \ldots, q_x\} : \|\hat{w}_{k,p}\|^2 \ge 0.9p\}.$$

What happens for both simulated and real data is that a few $\hat{w}_{k,p}$ have norms close to p, and the remaining norms are significantly smaller. An approximate upper bound of p, holds because

$$\|\hat{w}_{k,p}\| \leq \sum_{i=1}^{q_x} |\hat{\alpha}_{i,k}| \|\hat{v}_{i,p}\| \leq p\|\boldsymbol{\alpha}_k\|_1,$$

where we see $\|\hat{v}_{k,p}\| \leq p$ by the change of variables

$$\|\hat{v}_{k,p}\| = p \int_{(p-1)/p}^{1} \hat{v}_k^2(x)\mathrm{d}x.$$

Since $\int_0^1 \hat{v}_k^2(x)\mathrm{d}x = 1$, $\|\hat{v}_{k,p}\|$ will generally not be very close to p, unless most of the mass of $\hat{v}_k$ is concentrated on the interval $[(p-1)/p,1]$.

We thus want to determine if the coefficients

$$\langle \hat{\phi}_p, \hat{w}_{k,p} \otimes \hat{u}_j \rangle, \quad k = 1, \ldots, q_\star, \quad j = 1, \ldots, q_y$$

are collectively small. Observe that

$$\begin{aligned}
p\langle \hat{\phi}_p, \hat{w}_{k,p} \otimes \hat{u}_j \rangle &= p \int\int \hat{\phi}_p(t,s)\hat{w}_{k,p}(s)\hat{u}_j(t)\mathrm{d}s\mathrm{d}t \\
&= \int\int \left(\sum_{k',j'} \hat{\psi}(k',j')\hat{v}_{k',p}(s)\hat{u}_{j'}(t) \right) \hat{w}_{k,p}(s)\hat{u}_j(t)\mathrm{d}s\mathrm{d}t \\
&= \int \sum_{k',j} \hat{\psi}(k',j)\hat{v}_{k',p}(s)\hat{w}_{k,p}(s)\mathrm{d}s \\
&= \int \sum_{k',j} \hat{\psi}(k',j)\hat{v}_{k',p}(s) \left(\sum_i \hat{\alpha}_{i,k}\hat{v}_{i,p}(s) \right) \mathrm{d}s \\
&= \sum_{k',i} \hat{\psi}(k',j)\hat{V}(k',i)\hat{\alpha}_{i,k} \\
&= \sum_{k'} \hat{\psi}(k',j)[\hat{\mathbf{V}}\hat{\boldsymbol{\alpha}}_k](k') \\
&= \sum_{k'} \hat{\psi}(k',j)\hat{\delta}_k\hat{\alpha}_{k',k} = \hat{\gamma}_k[\boldsymbol{\alpha}_k^{\mathsf{T}}\hat{\boldsymbol{\psi}}](j).
\end{aligned}$$

The above calculation shows that the coefficients $\langle \hat{\phi}_{p+1}, \hat{w}_{k,p} \otimes \hat{u}_j \rangle$ are small if the matrices $\hat{\gamma}_k\boldsymbol{\alpha}_k^{\mathsf{T}}\hat{\boldsymbol{\psi}}$ have small entries. As explained above, $\hat{\gamma}_k = \|\hat{w}_{k,p}\|^2 \geq 0.9p$, so we reject H_p if the entries of the matrices $\boldsymbol{\alpha}_k^{\mathsf{T}}\hat{\boldsymbol{\psi}}$ are collectively large. To derive a test statistic, consider the following matrices (with their dimensions in parentheses)

$$\hat{\mathbf{A}}_\star = [\hat{\boldsymbol{\alpha}}_1, \ldots, \hat{\boldsymbol{\alpha}}_{q_\star}] \quad (q_x \times q_\star), \quad \hat{\mathbf{A}}_\star^{\mathsf{T}}\hat{\boldsymbol{\psi}} \quad (q_\star \times q_y). \tag{15}$$

We want to construct a quadratic form which is large when some entries of $\hat{\mathbf{A}}_\star^{\mathsf{T}}\hat{\boldsymbol{\psi}}$ are large, and which has an approximately parameter free distribution. We will exploit the approximation $\mathbf{Z}^{\mathsf{T}}(\mathrm{var}\mathbf{Z})^{-1}\mathbf{Z} \xrightarrow{d} \chi^2_{\dim(\mathbf{Z})}$, which holds for an asymptotically normal vector $\mathbf{Z}$. To this end, we form the column vector $\mathrm{vec}(\hat{\mathbf{A}}_\star^{\mathsf{T}}\hat{\boldsymbol{\psi}})$ by stacking the columns of $\hat{\mathbf{A}}_\star^{\mathsf{T}}\hat{\boldsymbol{\psi}}$, a process known as *vectorization*. Using the property, $\mathrm{vec}(ABC) = (C^{\mathsf{T}} \otimes A)\mathrm{vec}(B)$, where $\otimes$ now denotes the Kronecker product of matrices, see for example, Chapter 4 of Horn and Johnson(1985), we obtain

$$\mathrm{vec}(\hat{\mathbf{A}}_\star^{\mathsf{T}}\hat{\boldsymbol{\psi}}) = (\mathbf{I}(q_y) \otimes \hat{\mathbf{A}}_\star^{\mathsf{T}})\mathrm{vec}(\hat{\boldsymbol{\psi}}),$$

where $\mathbf{I}(q_y)$ is the $q_y \times q_y$ identity matrix. We use the above identity to determine the approximate covariance matrix of $\mathrm{vec}(\hat{\mathbf{A}}_\star^{\mathsf{T}}\hat{\boldsymbol{\psi}})$. Applying the formula $\mathrm{var}(\mathbf{QZ}) = \mathbf{Q}\mathrm{var}(\mathbf{Z})\mathbf{Q}^{\mathsf{T}}$, and treating the matrix $\hat{\mathbf{A}}_\star$ as deterministic, we obtain

$$\mathrm{var}[\mathrm{vec}(\hat{\mathbf{A}}_\star^{\mathsf{T}}\hat{\boldsymbol{\psi}})] \approx ((\mathbf{I}(q_y) \otimes \hat{\mathbf{A}}_\star^{\mathsf{T}})\mathrm{var}(\mathrm{vec}(\hat{\boldsymbol{\psi}}))((\mathbf{I}(q_y) \otimes \hat{\mathbf{A}}_\star^{\mathsf{T}}),$$

where we used the property $(A \otimes B)^{\mathsf{T}} = A^{\mathsf{T}} \otimes B^{\mathsf{T}}$. We will see in Section 5, see Theorem 5.4, that

$$N\mathrm{var}(\mathrm{vec}(\hat{\boldsymbol{\psi}})) \approx \hat{\mathbf{C}}_\varepsilon \otimes \hat{\boldsymbol{\Lambda}},$$

where

$$\hat{\boldsymbol{\Lambda}} = \mathrm{diag}\{\hat{\lambda}_1, \ldots, \hat{\lambda}_{q_x}\}, \quad \hat{\mathbf{C}}_\varepsilon = N^{-1}(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\psi}})^{\mathsf{T}}(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\psi}}).$$

Combining these results, we arrive at the test statistic

$$\hat{\Delta}_p := N(\mathrm{vec}[\hat{\mathbf{A}}_\star\hat{\boldsymbol{\psi}}])^{\mathsf{T}}[(\mathbf{I}(q_y) \otimes \hat{\mathbf{A}}_\star)(\hat{\mathbf{C}}_\varepsilon \otimes \hat{\boldsymbol{\Lambda}})(\mathbf{I}(q_y) \otimes \hat{\mathbf{A}}_\star)]^{-1}\mathrm{vec}(\hat{\mathbf{A}}_\star\hat{\boldsymbol{\psi}}). \tag{16}$$

The statistic $\hat{\Delta}_{p+1}$ has an approximately chi–square distribution with $q_y q_\star$ degrees of freedom. In Section 4 we evaluate the quality of this approximation. We conclude this section with an algorithmic description of the test procedure.

Test algorithm (H_{p-1} against H_p).

1. Subtract the sample mean from the functional observations. Continue to work with the centred data.
2. Construct the regressors X_i according to (5), and set $Y_i = Z_i$.
3. Determine q_y such that the first q_y eigenfunctions of the covariance operator $\hat{C}_Y$ explain between 80 % and 90 % of the variance.
   (a) Set $q_y = q_x p$ or
   (b) take q_x analogous to q_y.
4. Construct the matrices **Y** and **X** according to (9).
5. Calculate the $q_x \times q_y$ matrix $\hat{\boldsymbol{\psi}}$ according to (10).
6. Calculate the $q_x \times q_x$ matrix $\hat{\mathbf{V}}$ according to (13), and its eigenvectors $\hat{\boldsymbol{\alpha}}_k$ and eigenfunctions $\hat{\gamma}_k$ defined in (14).
7. Determine $q_\star$ such that the first $q_\star$ eigenvalues $\hat{\gamma}_k$ are $>$ 0.9p. (The procedure is not sensitive to the cut–off value of 0.9, taking 0.5 produced the same conclusions in data examples and simulations.)
8. Construct the matrices $\hat{\mathbf{A}}_\star$ and $\hat{\mathbf{A}}_\star^{\mathrm{T}}\hat{\boldsymbol{\psi}}$ defined in (15) and compute the test statistic $\hat{\Delta}_p$ defined in (16).
9. Compute the p-value using the chi–square density with $q_y q_\star$ degrees of freedom.

4. FINITE SAMPLE PERFORMANCE AND APPLICATION TO FINANCIAL DATA

We first evaluate the performance of the test using simulated data, then we turn to the application to two financial data sets.

4.1 Simulated data

The data are generated according to an FAR model, the choice of the autoregressive operators specifies the order. We consider two models

$$Z_i = c_1 Z_{i-1} + c_2 Z_{i-2} + \varepsilon_i, \tag{17}$$

where the $c_j \in [0,1)$ are scalars, and

$$Z_i = \Phi_1(Z_{i-1}) + \Phi_2(Z_{i-2}) + \varepsilon_i, \tag{18}$$

where the kernel of Φ_i is given by

$$\phi_i(t,s) = \frac{c_i}{0.7468} \mathrm{e}^{-(t^2+s^2)/2}.$$

The L^2 norm of ϕ_i is approximately c_i.

The ε_i in both models are standard Brownian bridges. We used a burn–in period of 200 functional observations. The rejection rates are based on one thousand replications, so the standard errors for the empirical size are about 0.009, 0.006 and 0.003 respectively for the nominal sizes of 0.10, 0.05 and 0.01. To speed up the simulations, we used fixed values $q_y = 3$, which explain about 85% of the variance of the Y_{ii} and $q_x = 3p$.

The results of the simulation study are displayed in Tables 1 and 2. For $n \geq 100$, the sample sizes are generally within two standard errors off the nominal sizes. The power is practically 100% for testing the null hypothesis of the iid model against the alternative of an FAR(p) model with some $p = 1$ or $p = 2$. The power is also very high when testing the null hypothesis of the FAR(1) model against FAR(2) model, but lower than for testing the iid hypothesis. For $n = 300$, the power is 100% for all cases we considered.

We now apply our multistage test procedure to two financial data sets that have recently received some attention in functional data analysis research. The first set consists of daily credit card transactions, the second of curves of Eurodollar futures prices.

Table 1. Empirical size and power for model (17)

Null hyp	$p = 0$	$p \leq 1$	$p = 0$	$p \leq 1$	$p = 0$	$p \leq 1$
Alt hyp	$p \geq 1$	$p \geq 2$	$p \geq 1$	$p \geq 2$	$p \geq 1$	$p \geq 2$
	$c_1 = 0$	$c_1 = 0$	$c_1 = 0.5$	$c_1 = 0.5$	$c_1 = 0.5$	$c_1 = 0.5$
Sig. level	$c_2 = 0$	$c_2 = 0$	$c_2 = 0$	$c_2 = 0$	$c_2 = 0.3$	$c_2 = 0.3$
$n = 100$						
0.10	0.115	0.122	1	0.112	1	0.831
0.05	0.070	0.068	1	0.060	1	0.753
0.01	0.022	0.015	1	0.016	1	0.558
$n = 200$						
0.10	0.117	0.120	1	0.105	1	0.986
0.05	0.054	0.062	1	0.058	1	0.968
0.01	0.012	0.013	1	0.010	1	0.925

Table 2. Empirical size and power for model (18)

Null hyp	$p = 0$	$p \leq 1$	$p = 0$	$p \leq 1$	$p = 0$	$p \leq 1$
Alt hyp	$p \geq 1$	$p \geq 2$	$p \geq 1$	$p \geq 2$	$p \geq 1$	$p \geq 2$
	$c_1 = 0$	$c_1 = 0$	$c_1 = 0.5$	$c_1 = 0.5$	$c_1 = 0.5$	$c_1 = 0.5$
Sig. level	$c_2 = 0$	$c_2 = 0$	$c_2 = 0$	$c_2 = 0$	$c_2 = 0.3$	$c_2 = 0.3$
$n = 100$						
0.10	0.108	0.105	0.996	0.112	1	0.807
0.05	0.059	0.054	0.995	0.066	1	0.724
0.01	0.014	0.012	0.987	0.019	0.999	0.549
$n = 200$						
0.10	0.107	0.105	1	0.116	1	0.979
0.05	0.057	0.051	1	0.063	1	0.961
0.01	0.016	0.012	1	0.009	1	0.925

4.2 Credit card transactions

The data available for this analysis consist of all transactions completed using credit cards issued by Vilnius Bank, Lithuania. Details of every transaction are documented, but here we are interested only in modelling the daily pattern of the volume of transactions. The transaction volume data set was studied by Laukaitis and Rackauskas (2002), Gabrys and Kokoszka (2007) and Horváth *et al.* (2010). Denote by $D_n(t_i)$ the number of credit card transactions in day n,$n = 1, \ldots$,200 (3/11/2000 – 10/2/2001) between times t_{i-1} and t_i, where $t_i - t_{i-1} = 8\text{min}, i = 1, \ldots, 128$. We thus have $n = 200$ daily curves, which we view as individual functional observations. The transactions are normalized to have time stamps in the interval [0,1], which thus corresponds to one day. Some smoothing is applied to construct the functional objects, the details are explained in Gabrys and Kokoszka (2007).

The curves thus obtained have non–zero mean and exhibit strong weekly periodicity. By computing the differences $Z_n(t) = Y_n(t) - Y_{n-7}(t), n = 8, 9, \ldots, 200$, we can remove both. We refer to this method of obtaing the Z_i for further analysis as *differencing*. Another way to remove the weekly periodicity and the mean is to centre the observations according to their day of the week. We refer to this method as *centring*.

The P–values are displayed in Tables 3 and 4. The stationary process obtained by differencing can be modelled as FAR(1). This agrees with the conclusions reached by Laukaitis and Rackauskas (2002), Gabrys and Kokoszka (2007) and Horváth *et al.*(2010) who evaluated/tested the suitability of the FAR(1) model using its predictive performance and formal significance tests against error correlations and change points. Centering by week days leads to a more complex structure, which can be approximately captured by the FAR(2) model. That differencing and centring give slightly different results is not too surprising because they transform the original data in different ways.

4.3 Eurodollar futures

We now turn to the application of our procedure to the data set consisting of Eurodollar futures contract prices studied by Kargin and Onatski (2008). The seller of a Eurodollar futures contract takes on an obligation to deliver a 3-month deposit of one million US dollars to a bank account outside the US i months from today. The price the buyer is willing to pay for this contract depends on the prevailing interest rate. These contracts are traded at the Chicago Mercantile Exchange, and provide a way to lock in an interest rate. They are liquid assets responsive to Federal Reserve policy, inflation, and economic indicators. The data we study consist of 114 points per day; point i corresponds to the price of a contract with closing date i months from today. We work with centred data, that is the mean function has been subtracted from all observations. Fifty centred functions are shown in Figure 1.

The P–values displayed in Table 5 indicate that the FAR(1) model is not suitable for modelling the whole data set, but the FAR(2) model is acceptable. This conclusion agrees with the analysis presented in Horváth and Kokoszka (2012) where a change point test was applied to these data. Horath and Kokoszka (2012) report that the FAR(1) model is not suitable for the whole data set, merely for shorter subintervals. The present analysis shows that a slightly more complex FAR(2) model captures the stochastic structure of the whole data set.

Table 3. *P*-Values for the test applied to credit card data transformed by *differencing*

Null hyp	$p = 0$	$p \leq 1$
Alt hyp	$p \geq 1$	$p \geq 2$
P-Value	0.000	0.427

Table 4. *P*-Values for the test applied to credit card data transformed by *centring*

Null hyp	$p = 0$	$p \leq 1$	$p \leq 2$
Alt hyp	$p \geq 1$	$p \geq 2$	$p \geq 3$
P-Value	0.000	0.00	0.161

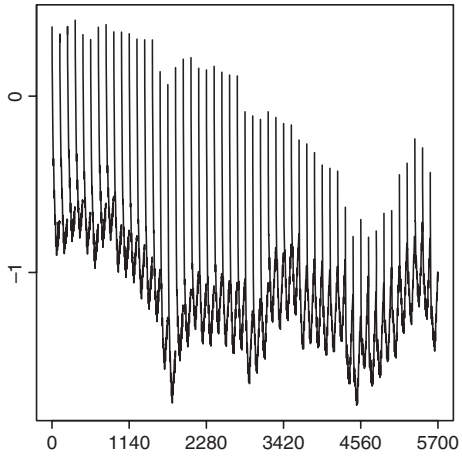

Figure 1. Centred Eurodollar futures curves on 50 consecutive days. (The mean of all curves, not just the 50 shown, was subtracted.)

Table 5. *P*-values of the test applied to Eurodollar futures curves.

Null hyp	$p = 0$	$p \leq 1$	$p \leq 2$
Alt hyp	$p \geq 1$	$p \geq 2$	$p \geq 3$
P-Value	0.000	0.000	0.731

4.4 Conclusions

The findings reported in this section show that the test has very good empirical size and power. When applied to real data, our procedure leads to conclusions that agree with previous studies and support modelling these functional time series as FAR(p) processes.

5. ESTIMATION IN THE FUNCTIONAL LINEAR MODEL WITH DEPENDENT REGRESSORS

The asymptotic justification of the chi–square approximation to the distribution of statistic (16) requires the asymptotic normality of the matrix $\hat{\psi}$, and the convergence in probability of the matrices $\hat{A}_*$ and $\hat{\mathbf{C}}_\varepsilon$. We address in detail the central issue of the asymptotic normality of $\hat{\psi}$, and provide comments on the asymptotic behaviour of $\hat{\mathbf{C}}_\varepsilon$. Its exact large sample distribution is not directly applicable, so we focus only on the approximations we use to construct the test. A convenient and powerful framework applicable to functional data uses the notion of L^p–m–approximability, see Hormann and Kokoszka(2010). We begin by stating the definition and several simple results which will be used in the proofs of the main results.

For $p \geq 1$, we let L_H^p be the space of $H = L^2$ valued random functions X such that

$$v_p(X) = (E\|X\|^p)^{1/p} = \left(E \int |X(t)|^p \mathrm{d}t\right)^{1/p} < \infty. \tag{17}$$

Definition 2. *A sequence $\{X_i\} \in L_H^p$ is called L^p–m–approximable if each X_i admits the representation*

$$X_i = f(\varepsilon_i, \varepsilon_{i-1}, \ldots), \tag{18}$$

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

$$X_i^{(m)} = f(\varepsilon_i, \varepsilon_{i-1}, \ldots, \varepsilon_{i-m+1}, \varepsilon_{i-m}^{(i)}, \varepsilon_{i-m-1}^{(i)}, \ldots), \tag{19}$$

we have

$$\sum_{m=1}^{\infty} v_p(X_m - X_m^{(m)}) < \infty. \tag{20}$$

The idea behind Definition 2 is that the impact of the shocks ε_{i-m} in the nonlinear moving average (18) is so small for large m that they can be replaced by different, stochastically equivalent, shocks.

The arguments that follow rely on the L^4–m–approximability of the sequence $\{X_i\}$ defined by (5). To formulate a sufficient condition for this to hold we use the Markovian representation of the FAR(p) process developed in Section 5.1 of Bosq (2000). It represents the FAR(p) process as an FAR(1) process in the cartesian product $(L^2)^p$ whose autoregressive operator is Φ' defined by (3). If $\|\Phi'\| < 1$, then by Lemma 5.1 of Bosq (2000) and Example 2.1 of Hormann and Kokoszka (2010), the sequence $\{[Z_i, Z_{i-1}, \ldots, Z_{i-p+1}]^T\}$ is L^p–m–approximable, provided $E\|\varepsilon_0\|^p < \infty$. The sequence $\{X_i\}$ is obtained from the above sequence by a bounded linear transformation, so by Lemma 2.1 of Hormann and Kokoszka (2010) it is also L^p–m–approximable. The above arguments established the following Lemma.

LEMMA 1. *If Assumptions 1 and 2 hold and $E\|\varepsilon_0\|^p < \infty$, then the sequence $\{X_i\}$ defined by (5) is L^p–m–approximable.*

We will also use the following theorem established in Horváth *et al.* (2012).

THEOREM 1. *If $\{Y_i\}$ is a mean zero L^2–m–approximable sequence, then*

$$N^{-1/2}\sum_{i=1}^{N} Y_i \xrightarrow{d} G \text{ in } L^2,$$

where G is a Gaussian process with

$$EG(t) = 0 \quad \text{and} \quad E[G(t)G(s)] = c(t,s);$$
$$c(t,s) = EY_0(t)Y_0(s) + \sum_{i\geq 1} EY_0(t)Y_i(s) + \sum_{i\geq 1} EY_0(s)Y_i(t).$$

Using Lemma 1 and Theorem 1, we obtain the following result

THEOREM 2. *If Assumptions 1 and 2 hold and $E\|\varepsilon_0\|^2 < \infty$, then*

$$N^{-1/2}\sum_{i=1}^{N} \varepsilon_i \otimes X_i \xrightarrow{d} G_1, \tag{21}$$

where G_1 is a mean zero Gaussian random function taking values in $L^2([0,1]\times[0,1])$. (Recall that X_i is defined by (5) and $(\varepsilon_i \otimes X_i)(t,s) = \varepsilon_i(s)X_i(t)$.)

PROOF. The functions $\varepsilon_i \otimes X_i$ are centred because $E\varepsilon_i = 0$ and ε_i is independent of X_i. The last property also implies that the sequence $\{\varepsilon_i \otimes X_i\}$ is L^2–m–approximable because the sequence $\{X_i\}$ is so. Indeed, since $\varepsilon_m^{(m)} = \varepsilon_m$, we have

$$\begin{aligned} v_2^2(\varepsilon_m \otimes X_m - \varepsilon_m^{(m)} \otimes X_m^{(m)}) &= E\int\int (X_m(t)\varepsilon_m(s) - X_m^{(m)}(t)\varepsilon_m(s))^2 dtds \\ &= E\int (X_m(t) - X_m^{(m)}(t))^2 dt \int \varepsilon_m^2(s) ds \\ &= v_2^2(X_m - X_m^{(m)})E\|\varepsilon_0\|^2. \end{aligned}$$

Consequently,

$$\sum_{m=1}^{\infty} v_2(\varepsilon_m \otimes X_m - \varepsilon_m^{(m)} \otimes X_m^{(m)}) \leq (E\|\varepsilon_0\|^2)^{1/2} \sum_{m=1}^{\infty} v_2(X_m - X_m^{(m)}) < \infty,$$

by Lemma 1. It remains to apply Theorem 1 to complete the proof. ∎

We are now prepared to turn to the estimation of the kernel ψ in the functional linear model (8). Notice that we cannot assume that the pairs (X_i, ε_i) (or (Y_i, X_i)) are independent. The last assumption is made in most contributions studying the functional linear model, see Yao *et al.* (2005a, 2005b), Müller and Stadtmüller (2005), Cai and Hall (2006), Li and Hsing (2007), Horvath *et al.* (2009), Mckeague and Sen (2010) and Li *et al.* (2010), to name only a handful of recent studies. Consequently, we cannot use any of the asymptotic theory developed by these authors. The asymptotic theory developed by Hormann and Kokoszka (2010) does not apply either because while they allow the regressors X_i to be dependent, they assume that the sequences $\{X_i\}$ and $\{\varepsilon_i\}$ are independent. This is clearly not the case in our setting because X_i is a function of $\varepsilon_{i-1}, \varepsilon_{i-2}, \ldots$. We therefore establish new asyptotic results which essentially use only the convergence (21) and the convergence of the empirical eigenfunctions.

We begin by establishing the convergence of the empirical eigenfunctions and eigenvalues. The X_i and the Y_i in (8) have mean zero, so their population covariance operators are

$$C_X(\cdot) = E[\langle X_1, \cdot\rangle X_1], \quad C_Y(\cdot) = E[\langle Y_1, \cdot\rangle Y_1].$$

Denote by (v_k, λ_k) the eigenfunction/eigenvalue pairs of the operator C_X ordered by decresing eigenvalues λ_k, and introduce analogously defined pairs (u_k, μ_k) for the operator C_Y. Assume that for the integers q_x and q_y,

$$\lambda_1 > \lambda_2 > \cdots > \lambda_{q_x} > \lambda_{q_{x+1}} \quad \text{and} \quad \mu_1 > \mu_2 > \cdots \mu_{q_y} > \mu_{q_{y+1}}. \tag{22}$$

Since the convergence of the empirical eigenfunctions holds only up to a sign, it is convenient to define

$$\hat{c}_k = \text{sign}\langle \hat{v}_k, v_k \rangle, \quad \hat{d}_j = \text{sign}\langle \hat{u}_j, u_j \rangle.$$

In the sequel, we will use the following Lemma.

LEMMA 2. *Suppose relation (8) holds with any L^4–m–approxmable sequence $\{X_i\}$ and $E\|\varepsilon_0\|^4 < \infty$. If condition (22) holds, then*

$$\max_{1\le k\le q_x} \|\hat{c}_k \hat{v}_k - v_k\| = O(N^{-1/2}); \quad \max_{1\le k\le q_x} \|\hat{\lambda}_k - \hat{\lambda}_k\| = O(N^{-1/2}) \tag{23}$$

and

$$\max_{1\le j\le q_y} \|\hat{d}_j \hat{u}_k - u_j\| = O(N^{-1/2}); \quad \max_{1\le j\le q_y} \|\hat{\mu}_j - \mu_j\| = O(N^{-1/2}). \tag{24}$$

PROOF. By Theorem 3.2 of Hormann and Kokoszka (2010), relation follows from the assumed L^4–m–approxmability of the sequence $\{X_i\}$. Since Ψ is a bounded linear operator, the $\Psi(X_i)$ form an L^4–m–approximable sequence, and thus so do the $Y_i = \Psi(X_i) + \varepsilon_i$. Lemma 2.1 of Hormann and Kokoszka (2010) states that bounded linear transformations and sums preserve L^p–m–approximability. Thus, the sequence $\{Y_i\}$ also satisfies the assumptions of Theorem 3.2 of Hörmann and Kokoszka (2010),and so (24) holds. ■

To formulate the main theorem of this section, we introduce the matrices

$$\mathbf{C}_N := \begin{bmatrix} \hat{c}_1 & 0 & 0 & \cdots & 0 \\ 0 & \hat{c}_2 & 0 & \cdots & 0 \\ \vdots & & \ddots & & \vdots \\ 0 & \cdots & & & \hat{c}_{q_x} \end{bmatrix} \quad \text{and} \quad \mathbf{D}_N := \begin{bmatrix} \hat{d}_1 & 0 & 0 & \dot{s} & 0 \\ 0 & \hat{d}_2 & 0 & \cdots & 0 \\ \vdots & & \ddots & & \vdots \\ 0 & \cdots & & & \hat{d}_{q_y} \end{bmatrix}.$$

Note that $\mathbf{C}_N$ and $\mathbf{D}_N$ are just diagonal matrices of ± 1,and are needed to ensure that the emprical eigenfunctions converge to their population counterparts.

THEOREM 3. *Suppose Assumptions 1 and 2 hold and $E\|\varepsilon_0\|^4 < \infty$. If $\hat{\boldsymbol{\psi}}$ is the least squares estimator of $\boldsymbol{\psi}$ and $\hat{\psi}$ is the corresponding estimator of $\hat{\pi}_{q_x,q_y}(\psi)$ given by*

$$\hat{\psi}(t,s) = \sum_{k\le q_x, j\le q_y} \hat{\boldsymbol{\psi}}(k,j)\hat{v}_k(s)\hat{u}_j(t),$$

then there exists a Gaussian random matrix, $\mathbf{G}$, and a Gaussian random function, G_2, taking values in $L^2(T\times T)$ such that

$$\mathbf{C}_N N^{1/2}(\hat{\boldsymbol{\psi}} - \boldsymbol{\psi})\mathbf{D}_N \xrightarrow{d} \mathbf{G} \quad \text{and} \quad N^{1/2}(\hat{\psi} - \hat{\pi}_{q_x,q_y}(\psi)) \xrightarrow{d} G_2,$$

where $\xrightarrow{d}$ denotes the convergence in distribution in their respective spaces.

PROOF. Consider the difference

$$\hat{\boldsymbol{\psi}} - \boldsymbol{\psi} = (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{Y} - \boldsymbol{\psi} = (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\boldsymbol{\varepsilon}'.$$

We express $\boldsymbol{\varepsilon}'$ as a sum of the error from model (7), $\boldsymbol{\varepsilon}(i,j) := \langle \varepsilon_i, \hat{u}_j \rangle$, and an error from the projections

$$\boldsymbol{\eta}(i,j) := \sum_{l>q_x} \langle X_i, \hat{v}_l \rangle \langle \psi, \hat{v}_l \otimes \hat{u}_j \rangle,$$

so we have

$$\hat{\boldsymbol{\psi}} - \boldsymbol{\psi} = (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\boldsymbol{\varepsilon} + (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\boldsymbol{\eta}. \tag{25}$$

Since the $\hat{v}_i$ are the eigenfunctions of $\widehat{C}_X$, $N^{-1}\mathbf{X}^\mathsf{T}\mathbf{X}$ is a $q_x \times q_x$ diagonal matrix of the empirical eigenvalues $\hat{\lambda}_k$. Examining the first term of (25), we thus have

$$[(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\boldsymbol{\varepsilon}](k,j) = \frac{1}{N\hat{\lambda}_k}[\mathbf{X}^{\mathrm{T}}\boldsymbol{\varepsilon}](k,j) = \frac{1}{N\hat{\lambda}_k}\sum_{i=1}^{N}\langle X_i, \hat{v}_k\rangle\langle \varepsilon_i, \hat{u}_j\rangle.$$

We move one inner product inside the other to obtain

$$\frac{1}{N\hat{\lambda}_k}\sum_{i=1}^{N}\langle X_i, \hat{v}_k\rangle\langle \varepsilon_i, \hat{u}_j\rangle = \frac{1}{N\hat{\lambda}_k}\left\langle \sum_{i=1}^{N} X_i \otimes \varepsilon_i, \hat{v}_k \otimes \hat{u}_j \right\rangle.$$

Next, we multiply by $\hat{c}_k\hat{d}_jN^{1/2}$, and conclude, by Theorem 2 and Lemmas 1 and 2, that

$$\begin{aligned}&\hat{c}_k\hat{d}_jN^{1/2}\frac{1}{N\hat{\lambda}_k}\left\langle \sum_{i=1}^{N} X_i \otimes \varepsilon_i, \hat{v}_k \otimes \hat{u}_j \right\rangle \\ &= N^{1/2}\frac{1}{N\hat{\lambda}_k}\left\langle \sum_{i=1}^{N} X_i \otimes \varepsilon_i, \hat{c}_k\hat{v}_k \otimes \hat{d}_j\hat{u}_j \right\rangle \xrightarrow{d} \frac{1}{\lambda_k}\langle G_1, v_k \otimes u_j\rangle,\end{aligned}$$

where G_1 is a Gaussian random function of $L^2(T\times T)$ defined in (21). Notice that multiplying by $\hat{c}_k$ and $\hat{d}_j$ coordinate wise is equivalent to multiplying matrix wise by the matrices $\mathbf{C}_N$ and $\mathbf{D}_N$ from the left and right respectively.

Turning to the second term in (25), we have

$$[(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\boldsymbol{\eta}](k,j) = \frac{1}{N\hat{\lambda}_k}[\mathbf{X}^{\mathrm{T}}\boldsymbol{\eta}](k,j) = \frac{1}{N\hat{\lambda}_k}\sum_{i=1}^{N}\langle X_i, \hat{v}_k\rangle\sum_{l>q_x}\langle X_i, \hat{v}_l\rangle\langle \psi, \hat{v}_l \otimes \hat{u}_j\rangle.$$

We can run into numerous problems when we try and use an arbitrary basis or the population principal components. However, when using the empirical principal components, things simplify greatly. We move around the inner products and use the identity $\hat{C}_X(\hat{v}_k) = \hat{\lambda}_k\hat{v}_k$ to obtain

$$\begin{aligned}\frac{1}{N\hat{\lambda}_k}\sum_{i=1}^{N}\langle X_i, \hat{v}_k\rangle\sum_{l>q_x}\langle X_i, \hat{v}_l\rangle\langle \psi, \hat{v}_l \otimes \hat{u}_j\rangle &= \hat{\lambda}_k{}^{-1}\sum_{l>q_x}\langle \psi, \hat{v}_l \otimes \hat{u}_j\rangle N^{-1}\sum_{i=1}^{N}\langle X_i, \hat{v}_k\rangle\langle X_i, \hat{v}_l\rangle \\ &= \hat{\lambda}_k{}^{-1}\sum_{l>q_x}\langle \psi, \hat{v}_l \otimes \hat{u}_j\rangle\left\langle \hat{C}_X(\hat{v}_k), \hat{v}_l \right\rangle \\ &= \sum_{l>q_x}\langle \psi, \hat{v}_l \otimes \hat{u}_j\rangle\langle \hat{v}_k, \hat{v}_l\rangle = 0.\end{aligned}$$

Since $k \le q_x < l$, the products $\langle \hat{v}_k, \hat{v}_l\rangle$ vanish, and so the second term vanishes. Furthermore, the random function G_1 is the same for every component, so we have the joint convergence

$$\mathbf{C}_N N^{1/2}(\hat{\boldsymbol{\psi}} - \boldsymbol{\psi})\mathbf{D}_N \xrightarrow{d} \mathbf{G},$$

where $\mathbf{G}$ is Gaussian random matrix. We conclude that

$$\begin{aligned}N^{1/2}(\hat{\pi}_{q_x,q_y}(\psi) - \hat{\psi}) &= \sum_{k\le q_x, j\le q_y} N^{1/2}(\hat{\boldsymbol{\psi}}(k,j) - \boldsymbol{\psi}(k,j))\hat{v}_k \otimes \hat{u}_j \\ &= \sum_{k\le q_x, j\le q_y} \hat{c}_k\hat{d}_jN^{1/2}(\hat{\boldsymbol{\psi}}(k,j) - \boldsymbol{\psi}(k,j))(\hat{c}_k\hat{v}_k) \otimes (\hat{d}_j\hat{u}_j) \\ &\xrightarrow{d} \sum_{k\le q_x, j\le q_y} \frac{1}{\lambda_k}\langle G_1, v_k \otimes u_j\rangle v_k \otimes u_j := G_2,\end{aligned}$$

by Slutsky's lemma. The random function G_2 is a Gaussian process in $L^2([0,1]\times[0,1])$ since $\{\langle G_1, v_k \otimes u_j\rangle\}$ are jointly normal. ■

A key element of the proof, and conclusion, of Theorem 3 is that the second term on the right–hand side of (25) vanishes. This follows from the construction of the estimator $\hat{\boldsymbol{\psi}}$ (10) using the empirical functional principal components. Other orthonormal bases would not yield the convergence with the rate $N^{-1/2}$.

The parametric rate is due to the fact that we estimate only a finite dimensional projection of ψ. When it is desirable to estimate ψ without such a restriction, the available results, see Yao *et al.* (2005b) and Hormann and Kokoszka (2010), claim only the convergence in probaility of an appropriately defined $\hat{\psi}$ to ψ, without any specific rate. The rates are however implicit in these results and would depend on the unknown rates at which the eigenvalues λ_k and μ_j, and the distances between them decay to zero. Results of this type would not be useful for the problem of hypothesis testing which requires the computation of a test statistic, and so a known rate of estimation.

The construction of the test statistic is based on the covariance matrix of the vectorized estimator $\hat{\boldsymbol{\psi}}$. This matrix can be computed using Theorem 3. We discuss its estimation after stating the result.

THEOREM 4. *Suppose the assumptions of Theorem 3 hold. Then*

$$N\text{cov}((\hat{\psi}-\psi)(k,j),(\hat{\psi}-\psi)(k',j')) \to \lambda_k^{-1}\delta_{kk'}\langle C_\varepsilon, u_j\otimes u_{j'}\rangle,$$

where C_ε is the covariance operator $E[\langle\varepsilon_0,\cdot\rangle\varepsilon_0]$, and $\delta_{kk'}$ is Dirac's delta.

PROOF. Recall that

$$\{\hat{c}_k\hat{d}_j N^{1/2}(\hat{\psi}-\psi)(k,j)\} \xrightarrow{d} \left\{\frac{1}{\lambda_k}\langle G_1, v_k\otimes u_j\rangle\right\}.$$

and

$$N^{-1/2}\sum_{i=1}^{N} X_i\otimes\varepsilon_i \xrightarrow{d} G_1.$$

Since ε_i is the error term for Z_i and X_i is a function of $Z_{i-1}, Z_{i-2},\ldots$, it follows that ε_i, and X_i are independent for any FAR(p) model. Thus the summands above are uncorrelated elements of $L^2([0,1]\times[0,1])$. Since they are also equal in distribution, the covariance function of G_1 is given by

$$E[G_1\otimes G_1] = E[(X_1\otimes\varepsilon_1)\otimes(X_1\otimes\varepsilon_1)].$$

So the asymptotic variance of $N^{1/2}(\hat{\psi}-\psi)(k,j)$ is given by

$$\begin{aligned}\lambda_k^{-2}E\langle G_1, v_k\otimes u_j\rangle^2 &= \lambda_k^{-2}E\langle G_1, v_k\otimes u_j\rangle\langle G_1, v_k\otimes u_j\rangle\\ &= \lambda_k^{-2}E\langle G_1\otimes G_1, v_k\otimes u_j\otimes v_k\otimes u_j\rangle\\ &= \lambda_k^{-2}E\langle X_1\otimes\varepsilon_1\otimes X_1\otimes\varepsilon_1, v_k\otimes u_j\otimes v_k\otimes u_j\rangle\\ &= \lambda_k^{-2}E\left[\langle X_1\otimes X_1, v_k\otimes v_k\rangle\langle\varepsilon_1\otimes\varepsilon_1, u_j\otimes u_j\rangle\right].\end{aligned}$$

Using the independence of X_i and ε_i and the linearity of the inner product we have

$$\begin{aligned}\lambda_k^{-2}E\langle G_1, v_k\otimes u_j\rangle^2 &= \lambda_k^{-2}E\left[\langle X_1\otimes X_1, v_k\otimes v_k\rangle\langle\varepsilon_1\otimes\varepsilon_1, u_j\otimes u_j\rangle\right]\\ &= \lambda_k^{-2}\langle E(X_1\otimes X_1), v_k\otimes v_k\rangle\langle E(\varepsilon_1\otimes\varepsilon_1), u_j\otimes u_j\rangle\\ &= \lambda_k^{-2}\langle C_X, v_k\otimes v_k\rangle\langle C_\varepsilon, u_j\otimes u_j\rangle\\ &= \lambda_k^{-1}\langle C_\varepsilon, u_j\otimes u_j\rangle.\end{aligned}$$

Similarly, the asymptotic covariance between the (k,j) and (k',j') elements is

$$\begin{aligned}&\lambda_k^{-1}\lambda_{k'}^{-1}E\langle G_1, v_k\otimes u_j\rangle\langle G_1, v_{k'}\otimes u_{j'}\rangle\\ &= \lambda_k^{-1}\lambda_{k'}^{-1}E\langle G_1\otimes G_1, v_k\otimes u_j\otimes v_{k'}\otimes u_{j'}\rangle\\ &= \lambda_k^{-1}\lambda_{k'}^{-1}E\langle X_1\otimes\varepsilon_1\otimes X_1\otimes\varepsilon_1, v_k\otimes u_j\otimes v_{k'}\otimes u_{j'}\rangle\\ &= \lambda_k^{-1}\lambda_{k'}^{-1}E\langle X_1\otimes X_1, v_k\otimes v_{k'}\rangle\langle\varepsilon_1\otimes\varepsilon_1, u_j\otimes u_{j'}\rangle\\ &= \lambda_k^{-1}\delta_{kk'}\langle C_\varepsilon, u_j\otimes u_{j'}\rangle.\end{aligned}$$

■

We see that in order to estimate the asymptotic variances and covariances of the entries of $N^{1/2}(\hat{\psi}-\psi)$, we need an estimator of the matrix

$$\mathbf{C}_\varepsilon(j,j') = \langle C_\varepsilon, u_j\otimes u_{j'}\rangle.$$

In multivariate regression, $\mathbf{C}_\varepsilon$ is estimated via the residual sum of squares:

$$\hat{\mathbf{C}}_\varepsilon = N^{-1}(\mathbf{Y}-\mathbf{X}\hat{\psi})^{\mathsf{T}}(\mathbf{Y}-\mathbf{X}\hat{\psi}).$$

Notice that in our functional setting

$$\hat{\mathbf{C}}_\varepsilon = N^{-1}(\mathbf{X}\psi+\varepsilon'+\eta-\mathbf{X}\hat{\psi})^{\mathsf{T}}(\mathbf{X}\psi+\varepsilon+\eta-\mathbf{X}\hat{\psi}),$$

where$\varepsilon = \varepsilon+\eta$ separates the error into the model error ε and the projection error η. As we showed in the proof of Theorem 3, $\mathbf{X}^{\mathsf{T}}\eta = 0$. Thus the residual sum of squares becomes

$$N^{-1}\varepsilon^{\mathsf{T}}\varepsilon + N^{-1}\eta^{\mathsf{T}}\eta + o_P(1),$$

as $\mathbf{X}^{\mathsf{T}}\varepsilon = O_P(N^{1/2})$, $(\psi-\hat{\psi})^{\mathsf{T}}\mathbf{X}^{\mathsf{T}}\mathbf{X}(\psi-\hat{\psi}) = O_P(1)$, and $\varepsilon^{\mathsf{T}}\eta = O_P(N^{1/2})$. Clearly the term from the model errors converges to what we want, that is

$$N^{-1}\boldsymbol{\varepsilon}^{\mathsf{T}}\boldsymbol{\varepsilon} \xrightarrow{a.s} \mathbf{C}_{\varepsilon}.$$

Examining the projection error we have

$$\begin{aligned}
N^{-1}(\boldsymbol{\eta}^{\mathsf{T}}\boldsymbol{\eta})(j,j') &= N^{-1}\sum_{i=1}^{N}\sum_{l>q_x}\langle X_i,\hat{v}_l\rangle\langle\psi,\hat{v}_l\otimes\hat{u}_j\rangle\sum_{r>q_x}\langle X_i,\hat{v}_r\rangle\langle\psi,\hat{v}_r\otimes\hat{u}_{j'}\rangle \\
&= N^{-1}\sum_{l>q_x}\sum_{r>q_x}\langle\psi,\hat{v}_l\otimes\hat{u}_j\rangle\langle\psi,\hat{v}_r\otimes\hat{u}_{j'}\rangle\sum_{i=1}^{N}\langle X_i,\hat{v}_l\rangle\langle X_i,\hat{v}_r\rangle \\
&= N^{-1}\sum_{l>q_x}\sum_{r>q_x}\langle\psi,\hat{v}_l\otimes\hat{u}_j\rangle\langle\psi,\hat{v}_r\otimes\hat{u}_{j'}\rangle N\hat{\lambda}_l\delta_{lr} \\
&= \sum_{l>q_x}\langle\psi,\hat{v}_l\otimes\hat{u}_j\rangle\langle\psi,\hat{v}_l\otimes\hat{u}_{j'}\rangle\hat{\lambda}_l.
\end{aligned}$$

Thus, in using the residuals, we pick up a bias term for our covariance estimates. However, we can still use those estimates since the bias takes the form of a nonnegative definite matrix. Thus the bias term would actually make our procedures more conservative. However, q_x is chosen such that the eigenvalues after q_x are all very small, thus it is reasonable to believe that the bias is generally small and will have a minimal effect on power. This is confirmed by the very good empirical performance of the test.

REFERENCES

Akaike, H.(1978) A Bayesian analysis of the minimum AIC procedure. *Annals of the. Institute of Statistical. Mathematics* **30A**, 9–14.
Antoniadis, A., Paparoditis, E. and Sapatinas, T. (2006) A functional wavelet–kernel approach for time series prediction. *Journal of the Royal Statistical Society Series B* **68**, 837–57.
Antoniadis, A. and Sapatinas, T. (2003) Wavelet methods for continuous time prediction using hilbert–valued autoregressive processes. *Journal of Multivariate Analysis* **87**, 133–58.
Besse, P., Cardot, H. and Stephenson, D. (2000) Autoregressive forecasting of some functional climatic variations. *Scandinavian Journal of Statistics* **27**, 673–687.
Bhansali, R. J. (1993) Order selection for linear time series models: a review. In *Developments in Time Series Analysis, London* (ed. T. Subba Rao), London, UK: Chapman and Hall, pp. 50–6.
Bosq, D. (2000) *Linear Processes in Function Spaces*. New York: Springer.
Bosq, D. and Blanke, D. (2007) *Inference and Prediction in Large Dimensions*. Chichester, UK: Wiley.
Brockwell, P.J. and Davis, R.A. (1991) *Time Series: Theory and Methods*. New York: Springer.
Cai, T. and Hall, P. (2006) Prediction in functional linear regression. *The Annals of Statistics* **34**, 2159–179.
Damon, J. and Guillas, S. (2002) The inclusion of exogenous variables in functional autoregressive ozone forecasting. *Environmetrics* **13**, 759–74.
Didericksen, D., Kokoszka, P. and Zhang, X. (2012) Empirical properties of forecasts with the functional autoregressive model. *Computational Statistics* **27**, 585–98.
Ferraty, F. and Romain, Y. (eds) (2011) *The Oxford Handbook of Functional Data Analysis*. New York: Oxford University Press.
Ferraty, F. and Vieu, P. (2006) *Nonparametric Functional Data Analysis: theory and practice*. New York: Springer.
Gabrys, R. and Kokoszka, P. (2007) Portmanteau test of independence for functional observations. *Journal of the American Statistical Association* **102**, 1338–48.
Hannan, E.J. (1980) The estimation of the order of an ARMA process. *The Annals of Statistics*, **8**, 1071–81.
Hannan, E.J. and Quinn, B.G. (1979) The determination of the order of an autoregression. *Journal of the Royal Statistical Society series, B* **41**, 190–5.
Hannan, E.J. and Rissannen, J.(1982) Recursive estimation of mixed autoregressive-moving average order. *Biometrika*, **69**,81–94; Correction (1983) 70,303.
Hormann, S. and Kokoszka, P. (2010) Weakly dependent functional data. *The Annals of Statistic* **38**, 1845–84.
Hormann, S. and Kokoszka, P. (2012). Functional time series. In *Time Series* (eds C.R. Rao and T. Subba Rao), Handbook of Statistics, volume 30. Oxford, UK: Elsevier, pp. 155–186.
Horn, R. A.and Johnson, C. R. (1985) *Matrix Analysis*. Cambridge, UK: Cambridge University Press.
Horváth, L., Hušková, M. and Kokoszka, P. (2010) Testing the stability of the functional autoregressive process. *Journal of Multivariate Analysis* **101**, 352–67.
Horváth, L. and Kokoszka, P. (2012) *Inference for Functional Data with Applications. Springer Series in Statistics*. New York: Springer.
Horváth, L., Kokoszka, P. and Reeder, R. (2012) Estimation of the mean of functional time series and a two sample problem. *Journal of the Royal Statistical Society (B)*,Forthcoming
Horváth, L., Kokoszka, P. and Reimherr, M. (2009) Two sample inference in functional linear models. *Canadian Journal of Statistics* **37**, 571–91.
Kargin, V. and Onatski, A. (2008) Curve forecasting by functional autoregression. *Journal of Multivariate Analysis* **99**, 2508–26.
Laukaitis, A. and Ravčkauskas, A. (2002) Functional data analysis of payment systems. *Nonlinear Analysis:Modeling and Control* **7**, 53–68.
Li, Y. and Hsing, T. (2007) On rates of convergence in functional linear regression. *Journal of Multivariate Analysis* **98**, 1782–04.
Li, Y., Wang, N. and Carroll, R.J. (2010) Generalized functional linear models with semiparametric single-index interactions. *Journal of the American Statistical Association* **105**, 621–33.
McKeague, I. and Sen, B. (2010) Fractals with point impacts in functional linear regression. *The Annals of Statistics* **38**, 2559–86.
Muller, H.G. and Stadtmuller, U. (2005) Generalized functional linear models. *The Annals of Statistics* **33**, 774–805.
Ramsay, J.,Hooker, G. andGraves, S. (2009) *Functional Data Analysis with R and MATLAB*. New York: Springer.
Ramsay, J.O. and Silverman, B.W. (2005) *Functional Data Analysis*. New York: Springer.
Shibata, R. (1980) Asymptotically efficient selection of the order of the model for estimating parameters of a linear process. *The Annals of Statistics* **8**, 147–64.
Yao, F., Muller, H.G. and Wang, J.L. (2005a) Functional data analysis for sparse longitudinal data. *Journal of the American Statistical Association* **100**, 577–90.
Yao, F., Muller, H.G. and Wang, J.L. (2005b) Functional linear regression analysis for longitudinal data. *The Annals of Statistics* **33**, 2873–903.