Functional time series

Łukasz Kidziński

CHILI Laboratory, École polytechnique fédérale de Lausanne,
RLC D1 740, CH-1015, Lausanne, Switzerland

February 26, 2015

Abstract

The continuous advances in data collection and storage techniques allow us to observe and record real-life processes in great detail. Examples include financial transaction data, fMRI images, satellite photos, earths pollution distribution in time etc. Due to the high dimensionality of such data, classical statistical tools become inadequate and inefficient. The need for new methods emerges and one of the most prominent techniques in this context is functional data analysis (FDA).

The main objective of this article is to present techniques of the analysis of temporal dependence in FDA. Such dependence occurs, for example, if the data consist of a continuous time process which has been cut into segments, days for instance. We are then in the context of so-called functional time series.

1 Introduction

In this article we introduce foundations of functional time series and the frequency–domain analysis in this context. We address the article to larger audience, assuming only elementary knowledge of probability theory and algebra. Although, we try to keep the text accurate, in some fragments we sacrifice detailed investigation for intuitive argumentation, referring advanced readers to appropriate sources.

The manuscript is divided into two parts. In the first part, we introduce concepts from statistics and functional data analysis (FDA). We built upon basic ideas about continuous functions and probability. In the second part, we present some state-of-the-art results in linear models for functional objects.

1.1 Motivation for statistics on functional data

The main concern of statistics is to obtain essential information from a sample of observations $X_1, X_2, ..., X_N$ from some space of objects. We are given a finite sample of size $N \in \mathbb{N}$, where $\{X_i\}_{i \in \mathbb{Z}}$ can be scalars (like heights of a sample of students in a school), vectors (like points on a
target after throwing several darts) or more complex objects, like genotypes, fMRI scans, images or frames of a video.

Functional data analysis deals with observations which can be naturally expressed as functions. Figures [1][2] and [3] present several cases, from various areas of science, which fit into the framework of functional data analysis.

When we deal with a physical process it is often natural to assume that it behaves in a continuous manner and that the observations do not oscillate significantly between the measurements. Although, in the Digital Age, we rarely record analog processes continuously, we often have enough datapoints, so that interpolation doesn’t cause a significant measurement error. Models incorporating this additional structure can lead to more precise and meaningful findings. In this context, FDA can be seen as a tool which embeds the continuity feature into the model.

Except for good approximation of a continuous process, FDA can also prove to be useful in a noisy, discontinuous case. Then, FDA can serve as a tool for denoising and smoothing the data and is beneficial whenever the underlying process is the main concern, like, for example, in finance.

From a pragmatic perspective, functional data can be seen simply as infinitely dimensional vectors, with extended notion of variance and mean, and thus we may be tempted to employ classical multivariate techniques. However, there are many practical and theoretical problems that need to be addressed and this approach is not advised, as we will see later.

The FDA approach is also useful in a parsimonious representation of the data by taking advantage of their smoothness. Instead of looking at a function as a dense vector of values, we can often represent it in a linear combination of a handful of (well–chosen) basis functions.

Practical applications of functional data analysis are spread across many areas of science and engineering. Panaretos et al. [11] use $[0, 1] \rightarrow \mathbb{R}^3$ closed curves to analyze the behavior of DNA microcircles, providing the testing methodology for the comparison of two classes of curves. Aston and Kirch [1] analyze the stationarity and change point detection for functional time series, with applications to fMRI data. Hadjipantelis et al. [5] analyze Mandarin language using functional principal components. Functional time series also naturally emerge in financial applications – Kokoszka and Reimherr [8] analyze predictability of the shape of intraday price curves. These works are only a fraction of the ongoing research and for a more accurate survey on applications and theory we refer to books [12], [4], [6] and [2].

1.2 Hilbert spaces

For most of the results presented in this work we only require a separable Hilbert space, a linear metric space with the norm function induced by the inner product and with a countable basis. It makes the setup very general, but for simplicity, and in order to give an intuitive example to each of the results, in most cases we will assume a concrete space of square-integrable functions on a bounded interval $[0, 1]$, to which we will refer to as $L_2$. A function $f : \mathbb{R} \rightarrow \mathbb{R}$ belongs to $L_2$ if and

\[\text{In this work we introduce Hilbert spaces in an elementary and accessible way, avoiding technical details. For a formal definitions and investigation of key properties refer to [13].}\]
Figure 1: Berkeley Growth Data: Heights of 20 girls taken from ages 0 through 18 (left). Growth process easier to visualize in terms of acceleration (right). Tuddenham and Snyder [14] and Ramsey and Silverman [12].

Figure 2: Lower lip movement (top), acceleration (middle) and EMG of a facial muscle (bottom) of a speaker pronouncing the syllable “bob” for 32 replications. Malfait, Ramsay, and Froda [9].

Figure 3: Projections of DNA minicircles on the planes given by the principal axes of inertia (three panels on the left side: TATA curves, right: CAP curves). Mean curves are plotted in white. Panaretos, Kraus and Maddocks [11].
only if
\[ \int_0^1 f^2(x)dx < \infty. \]

In this section we present elementary properties of Hilbert spaces.

With a space \( L_2 \), we associate an inner product, a bilinear operator \( L_2 \times L_2 \to \mathbb{R} \). For two functions \( f, g \in L_2 \), we define the inner product as

\[ \langle f, g \rangle = \int_0^1 f(x)g(x)dx. \]

We define the norm of the element \( f \in L_2 \) as \( \|f\| = \sqrt{\langle f, f \rangle} \). Since \( L_2 \) is a space of square-integrable functions, each element \( f \in L_2 \) has a finite norm. We say that \( f \) and \( g \) are orthogonal if \( \langle f, g \rangle = 0 \). If, additionally, \( \|f\| = \|g\| = 1 \), we say that \( f \) and \( g \) are orthonormal. Both definitions are related to vector spaces: a norm corresponds to the "distance" from 0 function, whereas orthonormal elements behave as perpendicular vectors.

Hilbert space is called separable if we can find a series of pairwise orthonormal elements \( e_1, e_2, e_3, \ldots \) such that each element in \( e \in L_2 \) can be expressed as a weighted sum of elements \( e_1, e_2, e_3, \ldots \). This series \( \{e_i\}_{1 \leq i} \) is called a basis, and each \( e_i \) is a basis function.

One can show, that given the basis \( \{e_i\}_{1 \leq i} \), the representation of \( f \in L_2 \) is uniquely given by

\[ f = \sum_{i=1}^{\infty} \langle e_i, f \rangle e_i, \quad (1) \]

where scalars \( \langle e_i, f \rangle \) are called coefficients in the basis \( \{e_i\}_{1 \leq i} \).

We may find infinitely many basis of \( L_2 \). As an example, often used in practice, consider the Fourier series, defined as

\[ e_i(x) = \begin{cases} \sin(k\pi x), & \text{if } i = 2k + 1 \\ \cos(k\pi x), & \text{if } i = 2k \end{cases} \]

where \( k \in \mathbb{Z} \) and \( i \geq 1 \). Several first elements are presented in Figure 4. A proof that elements are orthonormal is a simple exercise. The proof that each element in \( f \) can be uniquely expressed as a linear combination \( 1 \) is more complicated and an interested reader is referred to [13].

We will use, one of the fundamental equations in separable Hilbert spaces, a Parseval’s identity.

**Lemma 1** (Parseval’s identity). Let \( f \in L_2 \) and let \( \{e_i\}_{1 \leq i} \) be an orthonormal basis in \( L_2 \). Then

\[ \|f\|^2 = \sum_{i=1}^{\infty} |\langle e_i, f \rangle|^2. \]
Figure 4: First 5 Fourier basis functions in the order black, red, green, blue and light blue (left). Step function approximated by a weighted sum of 1, 2, 3 and 4 basis functions (right).

Proof. It follows from the definition of the norm

\[ \|f\|^2 = \langle f, f \rangle = \left( \sum_{i=1}^{\infty} \langle e_i, f \rangle e_i, \sum_{i=1}^{\infty} \langle e_i, f \rangle e_i \right) \]

\[ = \sum_{i=1}^{\infty} \langle e_i, f \rangle \left( e_i, \sum_{j=1}^{\infty} \langle e_j, f \rangle e_j \right) \]

\[ = \sum_{i=1}^{\infty} \sum_{j=1}^{\infty} \langle e_j, f \rangle \langle e_i, e_j \rangle \]

\[ = \sum_{i=1}^{\infty} |\langle e_i, f \rangle|^2 , \]

because \( \langle e_j, f \rangle \langle e_i, e_j \rangle = \langle e_i, f \rangle \) if \( i \neq j \) and 0 otherwise, due to orthonormality of the series \( \{e_i\}_{i \geq 1} \).

We will also use the notion of linear operators and Hilbert–Schmidt operators. A linear operator \( F : L_2 \to L_2 \) is a function such that for any pair of scalars \( a, b \) and elements \( v, w \in L_2 \), \( F(av + bw) = aF(v) + bF(w) \). A Hilbert Schmidt–operator \( F \) is a linear operator, such that

\[ \sum_{i=1}^{\infty} \|F(e_i)\|^2 < \infty , \]

where \( \{e_i\}_{i \geq 1} \) is an orthonormal basis of \( L_2 \). An operator is symmetric if \( F(v) = F(-v) \) for each \( v \in L_2 \).
1.3 Representation and fit

In practice we are often given just a sample of observations from a curve and we need to interpolate, i.e. draw a curve between these points which is most likely to be close to the underlying process.

In this work, we follow the ideas popularized by Ramsey and Silverman \[12\], based on a basis function expansion (1). Note that, given the representation 1 by the Parseval’s identity, for any \( \varepsilon > 0 \), there exists \( d \), such that

\[
\sum_{i=d}^{\infty} |\langle e_i, f \rangle|^2 < \varepsilon.
\]

We can therefore approximate the function with arbitrary precision \( \varepsilon > 0 \) using only the first \( d \) basis elements.

In practice, inner products, which are typically obtained by integration, will be themselves approximated by corresponding sums. Then, a discretized sample curve \((x(t_j): 1 \leq j \leq n)\) can be transformed into a (finite dimensional) curve \( y(t) \) through

\[
y(t) := \sum_{i=1}^{d} \left( \sum_{j=1}^{n} f(t_j)e_i(t_j)(t_j - t_{j-1}) \right) e_i(t),
\]

for some grid \( 0 \leq t_1 < t_2 < ... < t_n \leq 1 \), where the expression in brackets accounts for the approximation of the inner product between \( f \) and \( e_i \).

Fitting and representation of functional data is an important and intensively studied topic on its own, however, in this article we assume that datapoints are \emph{fully observed}, i.e. we observe the whole curves, instead of just a sample of points. For more information on fitting we refer to \[12\].

1.4 Statistics in Hilbert spaces

Random function is the key concept used in the sequel. We can think of it as an extension of a random value or a random vector from \( \mathbb{R}^2 \). Instead of drawing a random point from a plane, we draw a whole function from all functions in \( L_2 \). In this section we introduce a mean function and a covariance operator. We will extend the concept of expected value of a scalar variable to vectors and functions.

In the space \( \mathbb{R}^2 \), we refer to a mean as an expected value on each coefficient, for example, having a random vector \( \mathbf{X} = (X_1, X_2) \), we define a mean of this variable as \( EX = (EX_1, EX_2) \).

Similarly we can look at a random function \( X \). We can define it’s mean as a mean on each coefficient. Let’s take a fixed basis \( \{e_i\}_{1 \leq i} \) and an expansion of \( X \), given by

\[
X = \sum_{i=1}^{\infty} \langle e_i, X \rangle e_i.
\]
Then, random variables \( \langle e_i, X \rangle \) correspond to coefficients, so a mean function can be defined as

\[
EX = \sum_{i=1}^{\infty} (E\langle e_i, X \rangle)e_i,
\]

where we defined the expectation of random function as a sum of scalar expectations \( E\langle e_i, X \rangle \). Note that, as in scalar case \( E\langle e_i, X \rangle \) may not be finite and then also \( EX \) will not exist. Moreover, in order to have \( EX \in L_2 \), we need the series \( (E\langle e_i, X \rangle)_{1 \leq i} \) to be square summable.

Defining covariance in functional space is more complicated and, in order to show a direct relation, we will recall a non-standard representation of a covariance in multidimensional space. One of the ways to look at a covariance of a random vector \( X \) in \( \mathbb{R}^2 \), is to see it as a linear function \( \mathbb{R}^2 \to \mathbb{R}^2 \), defined as

\[
C(v) = E\langle X, v \rangle X,
\]

where \( v \in \mathbb{R}^2 \). We have \( C(v) = (EXX')v \) (and we refer to the variance as \( C = EXX' \)). Note, that since \( \langle X, v \rangle \) is a scalar, \( \langle X, v \rangle X \) is a random vector and therefore we just use the definition of the mean vector.

In the functional space, we will define a covariance operator as an operator \( L_2 \to L_2 \). For a random element \( X \) in \( L_2 \), similarly to the vector case, we take the expectation of \( \langle X, v \rangle X \), i.e.

\[
C_X(f) = E\langle X, f \rangle X, \text{ where } f \in L_2.
\]

We call \( \langle X, \cdot \rangle X \), an outer product and write \( X \otimes X = \langle X, \cdot \rangle X \). Again, we consider only such random functions \( X \) that \( C_X \) exists. One can show that the operator \( C_X \) is a positively definite, symmetric Hilbert-Schmidt operator and therefore it can be inversed.

## 2 Functional Time Series

In many practical situations functions are naturally ordered in time. For example, when we deal with daily observations of the stock market or with sequences of tumor scans. Then, we are in the context of so-called functional time series (FTS).

As a motivating example consider Figure 5. Here, the assumption of independence can be too strong – values at the beginning of each day are highly correlated with those at the end of the preceding day. Moreover, we see that big jumps are often followed by significant drops.

These, and similar features, may indicate significant temporal dependence not just within a subject, but also between different subjects (e.g. days). In this section we discuss possible frameworks which allow to quantify, and use this additional information.

### 2.1 Stationarity

Many physical processes are known to have time-invariant distribution. This motivates the frequentist approach to time series, where we assume that the structure does not change in time and we
Figure 5: Horizontal component of the magnetic field measured in one minute resolution at Honolulu magnetic observatory from 1/1/2001 00:00 UT to 1/7/2001 24:00 UT. 1440 measurements per day.

Let \{X_t\} be a series of random functions. We say that \{X_t\} is stationary in the strong-sense if for any \(h \in \mathbb{Z}, k \in \mathbb{N}\) and any sequence of indices \(t_1, t_2, ..., t_k\) vectors \((X_{t_1}, ..., X_{t_k})\) and \((X_{h+t_1}, ..., X_{h+t_k})\) are identically distributed.

We also define weak stationary by looking only on the second order structure of the series. We say that \{X_t\} is weakly stationary if \(E\|X_t\|^2 < \infty\) and

1. \(EX_t = EX_0\) for each \(t \in \mathbb{Z}\) and
2. \(EX_t \otimes X_s = EX_{t-s} \otimes X_0\) for each \(t, s \in \mathbb{Z}\).

Additionally, we will assume that a series in the sequel are weakly dependent, which intuitively means that observations from the far past have little to no effect on the present. Many frameworks were suggested to quantify this behavior, for a survey on most popular ones refer to [7]. For simplicity in this work we just assume a very strong condition, of a weakly dependent series as a stationary series for which

\[
\sum_{t=0}^{\infty} E\|X_t \otimes X_0\|^2 < \infty,
\]

meaning that the covariance between far elements decays very fast.

### 2.2 Functional linear regression

One of the most popular frameworks in classical statistics is the linear regression, where we try to quantify the linear dependence between two (possibly multivariate) variables \(X\) and \(Y\). The problem of finding the relation of this type can be also addressed in FDA. As an illustrative example we can think of the relation between some farm’s income during the year (a function of time defined on a yearly interval) and precipitation over a year.
We assume the model

\[ Y_t = A(X_t) + \varepsilon_t, \quad t \geq 1, \]  

(2)

where \( A \) is a linear Hilbert-Schmidt operator from \( L_2 \to L_2 \) and \( \varepsilon_t \) is a sequence of independent identically distributed functions white noise sequence, independent from \( (X_t) \).

As we are concerned with functional time series, we will assume that \( X_t, Y_t \) are weakly stationary and weakly dependent. Classical case of iid \( X_t \) is of great scientific interest and the interested reader is referred to [12] and [15].

Although the functional linear regression shares many properties with its multivariate equivalent, there are important theoretical difficulties, which preclude direct extension of the results from the simpler setup. Especially, we note that the linear operator \( A: L_2 \to L_2 \) is infinitely dimensional, which considerably complicates the estimation. If we approach the problem in the classical way by multiplying both sides of (2) by \( X_t \) and taking the expectation, for \( t \geq 1 \) we have

\[ EY_t \otimes X_t = EA(X_t) \otimes X_t + E\varepsilon_t \otimes X_t = EA(X_t) \otimes X_t = A(EX_t \otimes X_t), \]

by independence of \( X_t \) and \( \varepsilon_t \). Now, for convenience, let’s denote this by

\[ C^{XY} = AC^X, \]  

(3)

where \( C^{XY} \) is the cross-covariance operator of \( X \) and \( Y \) and \( C^X \) is the covariance of \( X \). Now, the natural way to obtain \( A \) is to apply the inverse of \( C^X \) to both sides of the equation (3), which yields

\[ A = C^{XY}(C^X)^{-1}. \]

The main problem is that the operator \((C^X)^{-1}\) is no longer bounded. Indeed, the domain of \( C^X \) is only a subset \( D \), say, of \( L_2 \). To see this, note that formally, as the inverse of \( C^X \) is a linear operator, we may express \((C^X)^{-1}(x) = \sum_{k \geq 0} \lambda_k^{-1} \langle e_k, x \rangle e_k\), where \( \lambda_k \) and \( e_k \) are the eigenvalues (tending to zero) and eigenfunctions of \( C^X \). Hence, \( D = \{ x \in L_2: \sum_{k \geq 1} \langle x, e_k \rangle^2 \lambda_k^{-2} < \infty \} \). The problem can be approached by some regularization. E.g. one may replace \((C^X)^{-1}\) by a finite dimensional approximation of the form \( \sum_{k \leq K} \lambda_k^{-1} e_k \otimes e_k \), where \( K \) is a tuning parameter. This is still quite delicate, when applied to the sample version. Then for large values of \( K \), if we underestimate one of the small eigenvalues, its reciprocal explodes and will lead to very instable estimators. On the other hand, for small \( K \) we may get a very poor approximation of \( A \).

This difficulty was addressed by Bosq [2], who gives an extensive survey on the problem. However, proposed results are based on strong assumptions on the rate of convergence of eigenvalues, which are impossible to check in practice. Alternative, elementary data-driven approach, was suggested in [7].

Finally, note that exactly the same technique can be used for lagged linear regression, i.e. where the response \( Y_t \) depends linearly not only on the current observation \( X_t \) but also on the whole series
Consider

\[ Y_t = \sum_{k=0}^{m} A_k(X_{t-k}) + \varepsilon_t, \quad (4) \]

where \( m \in \mathbb{N} \). Again, as an example, we can think of it as an income during a given year based on precipitation in last 3 years.

Let \( m \) be the largest lag that we want to take into account and let us introduce \( Z_t = (X_t, X_{t-1}, \ldots, X_{t-m}) \in L_2^m \). One can easily show that the space \( L_2^m \) is a Hilbert space. Then, the model can be written as

\[ Y_t = BZ_t + \varepsilon_t, \quad (5) \]

where \( B : L_2^m \to L_2 \) is a linear operator such that \( BZ_t = \sum_{k=0}^{m} A_k(Z_{t}^{(k)}) \).

Now, for estimating \( B \) in (5), we can apply the same estimation procedures as in (2). This method of estimation in lagged regression models is efficient only for small dimensions and small \( m \), as opposed to the technique briefly introduced in the following section.

### 2.3 Frequency-domain methods

The lagged linear model (4) can be linked with the concept of linear filtering, popular in multivariate time series as well as in signal processing. For the theory and survey on applications in this context we refer to the classical book of Oppenheim and Schafer [10].

**Definition 1.** We say that \( \mathcal{A} = \{A_k\}_{k \in \mathbb{Z}} \) is a linear filter if for each \( k \in \mathbb{Z} \), \( A_k \in L_2 \to L_2 \) is a linear operator and \( \sum \|A_k\|^2 < \infty \).

In order to find a method for estimation of operators \( A_k \) in (4), in a more efficient way than in (2.2), we employ Fourier analysis and results from the seminal work of Brillinger [3].

The Fourier transform has two important properties which simplify analysis of the process (2). First, multiplication in the frequency domain is equivalent to convolution in the time domain. Second, the Fourier transform is a bijection, so results in frequency domain are equivalent to these in the time domain.

To illustrate the usage of these features let us multiply equation (2) by \( X_s \) for some \( s \in \mathbb{Z} \) and take the expectation. By linearity we have

\[ EY_t \otimes X_s = \sum_{k \in \mathbb{Z}} A_k EX_{t-k} \otimes X_s, \]

and by stationarity

\[ EY_u \otimes X_0 = \sum_{k \in \mathbb{Z}} A_k EX_{u-k} \otimes X_0, \]

where \( u = t - s \). Now, noting that on left we have \( C_u^{YX} \) and on right we have the convolution of \( A_k \) and \( C_u^{YX} \), the Fourier transform of both sides yields the so-called cross-spectral operator between
\{Y_t\}$ and \{X_t\}$ and can be obtained as

$$F^Y_X(\theta) = A(\theta) F^X_T,$$

where $A(\theta) = \sum_{k \in \mathbb{Z}} A_k e^{ik\theta}$ is the frequency response function of the series \{\(A_k\)\}_{k \in \mathbb{Z}}. \(F^Y_X(\theta)\) is the spectral density operator of \{\(Y_t\)\} and \{\(X_t\)\} and \(F^X_T(\theta)\) is the spectral density operator of \{\(X_t\)\}.

Having relation (6), again we can invert \(F^Y_X(\theta)\) and obtain a closed–form expression for \(A(\theta)\). Now, the inverse Fourier transform gives us coefficients of the model (4). Note that, although we employ more sophisticated tools than in linear regression (3), symbolically approach presented here is analogical, but developed in the frequency domain.

References

[1] J. A. D. Aston and C. Kirch. Evaluating stationarity via change-point alternatives with applications to fmri data. ArXiv e-prints, 2013.
[2] D. Bosq. Linear Processes in Function Spaces. Springer, 2000.
[3] D. R. Brillinger. Time Series: Data Analysis and Theory. Holt, New York, 1975.
[4] F. Ferraty and P. Vieu. Nonparametric Functional Data Analysis: Theory and Practice. Springer, 2006.
[5] P. Z. Hadjipantelis, J. A. D. Aston, and J. P. Evans. Characterizing fundamental frequency in mandarin: A functional principal component approach utilizing mixed effect models. The Journal of the Acoustical Society of America, 131(6):4651–4664, 2012.
[6] L. Horváth and P. Kokoszka. Inference for Functional Data with Applications. Springer, 2012.
[7] Łukasz Kidziński. Inference for stationary functional time series: dimension reduction and regression. PhD thesis, Université libre de Bruxelles, Bruxelles, Belgique, 10 2015.
[8] P. Kokoszka and M. Reimherr. Predictability of shapes of intraday price curves. The Econometrics Journal, 16(3):285–308, 2013.
[9] N. Malfait and J. O. Ramsay. The historical functional model. Canadian Journal of Statistics, 31:115–128, 2003.
[10] A.V. Oppenheim and R.W. Schafer. Discrete-time signal processing. Prentice-Hall signal processing series.
[11] V. M. Panaretos, D. Kraus, and J. H. Maddocks. Second-order comparison of gaussian random functions and the geometry of dna minicircles. Journal of the American Statistical Association, 105(490):670–682, 2010.
[12] J. O. Ramsay and B. W. Silverman. Functional Data Analysis. Springer, 2005.
[13] Walter Rudin. *Real and Complex Analysis, 3rd Ed.* McGraw-Hill, Inc., New York, NY, USA, 1987.

[14] R. D. Tuddenham and M. M. Snyder. Physical growth of california boys and girls from birth to eighteen years. *Publications in child development. University of California, Berkeley*, 1(2):183–364, 1953.

[15] F. Yao, H-G. Müller, and J-L. Wang. Functional linear regression analysis for longitudinal data. *The Annals of Statistics*, 33:2873–2903, 2005.