A brief introduction to Functional Data Analysis

Gabriel Graciano Herrera

Mentor: Lucia Tabacu

1 Introduction

Over the past decade the humanity generates more data than the last 5000 years, the analysis of this data proves to be a good source of knowledge to take important decisions. Exists multiple approaches for this depending on the type of data, one of them is Functional Data Analysis (FDA). FDA is useful to analyze data from a variety of sensors, medical equipment like electrocardiogram machines, accelerometers, etc, data from climate, weather and financial among others.[1, 2]

The functional data (FD) are "the variables or units of interest in a dataset that can view as a smooth curve or function[1]"[4], a more technical definition "Observations on subjects i that you can imagine as arising from the evaluation of a subj-random curve $X_i(\cdot)$ defined on $[a, b]$ at Finite grid of points, $t_{i1}, ..., t_{im_i}$. Often such evaluations are contaminated with noise; so we observe $X_i(t_{ij}) + Noise_{ij}$" [5]. FD is not only about curves or functions, we can have FD in surfaces or anything else varying over a continuum, "the continuum is often time but can be spatial location, wavelength, etc".[6]

The functional data has most commons types than others, because its complexity we can't list all of them, the most commons are [2] types that depends on the type of data and the number of observations, this classifications have their own sub-classifications. For the type that depends on the type of data, their sub-classifications are the follows: **Independent replications of the same event**, **single long record of the same event**, **input of an event that affects the output of another element** and **multivariate events**. [7]

The type that depends the number of observations is simpler, it has 2 sub-types. **Sparse**, when the dataset has only a few observations available for each function. In the other spectrum is **dense**, FD referred as dense is when the data has a lot of observations[3].[4]

2 Examples

To understand better each type of functional data the next examples are provided with context about the data and visualizations of them.[4]

2.1 Growth of girls and Non-durable goods in USA

These two cases of study will help us to explain the **Independent replications** and **Single long record** types respectively. The first dataset correspond from a Berkely Growth Study where it was measured the height of 10 girls 31 times on different ages. The second correspond for non-durable goods manufacturing index in USA, the measure were taken per month from January 1991 to 2000.[7]

[1] A smooth function has continuous derivatives up to some desired order over a domain. The number of continuous derivatives can vary from two to infinity[3]

[2] This classifications aren't exclusive, indeed a dataset belong both classifications

[3] There isn't an obvious number of observations we need to classified in one type or the other, it's depends in the study case.

[4] The data and the code for plotting them were taken from *fda package* in R, provide by the book authors of [7]

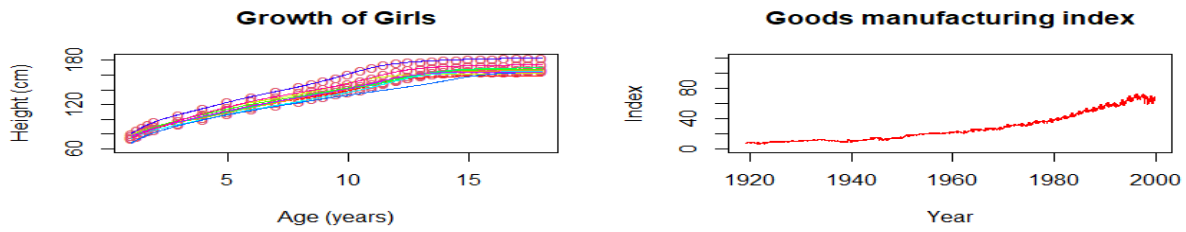

Figure 1: Height of 10 girls since 1 to 18 years old and tendency of non-durable goods manufacturing index in USA

In the growth of girls plot we have ten **indepedent replications** of the same event, each sample represent the height of each girl, also we could consider this dataset as a **sparse** type. In contrast, for the goods index there is **one long sample** of the same event; this could be a **dense** dataset because for the left plot there are 31 observations for each sample and this has 972 observations, the contrast of observations help us to identify the sparse and dense types.

2.2 How children walk

The motion Analysis Laboratory at Children´s Hospital, San Diego, CA, collect data from a group of 39 children over each child´s gait cycle [7]. The gait cycle represent the way a person walk, describe the movements that limbs do until the person's destination is reached[8].In the gait cycle are involved the knee and hip, "the cycle begins and ends at the point where the heel of the limb under observation strikes the ground".[7]

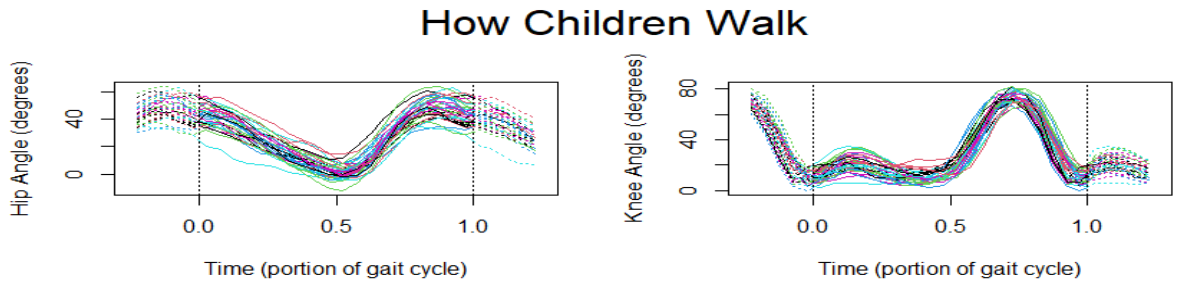

Figure 2: Hip and knee angle of 39 childs when they walked, the dotted lines delimited one gait cycle

The data type in this example is **multivariate event** and **dense**. The knee represent one variable and the hip the other, the two variables involved in the way a person walk.

3 How to build functions?

The first is building the functions that roughly matches the data. We have a set of functions to use as templates to build the function, we call this set *basis functions*, also we need an array of *coefficients* to express the functions as a linear combination of *basis functions*. The equation that summarize the above is as follows: [7]

$$x(t) = \sum_{k=1}^{K} c_k \phi_k(t) \tag{1}$$

This equation represent the function $x(t)$ called *basis function expansion*. c_k represents the *coefficients* $(c_1, c_2, ..., c_k)$ and $\phi_k(t)$ represents the *basis functions* $(\phi_1(t), \phi_2(t), ..., \phi_k(t)$, the sum of each one build the function of the data.

Sometimes we want to represent more than one sample for the data, for example when we have multiple samples of the same event like in the examples **Growth of Girls** (2.1) or **How Children Walk** (2.2).

$$x(t) = \sum_{k=1}^{K} c_{ik} \phi_k(t) = \mathbf{C}\phi(t) \tag{2}$$
$$where\ i{=}1{,}2{,}...{,}N\ and\ N{=}number\ of\ samples$$

The equation is similar to 1 the difference falls in the coefficients, the C stand for a matrix of size $N \times K$, N means the number of samples and denote the rows and K represents the columns.

4 Basis functions

We mention before that we have a set of templates, called *basis functions*, to build the *basis function expansion*. Exists plenty basis functions, some of them are very basics and others are very complex, for the purpose of an introduction are listed the most commonly used: **Monomial basis**, **Fourier basis**, **B-Spline basis**.[7]

4.1 Monomial basis

The monomial basis includes straight lines, quadratic polynomials, cubic polynomials and so on, i,e the monomial basis denotes the power of t: $1, t, t^2, ..., t^n$.[]

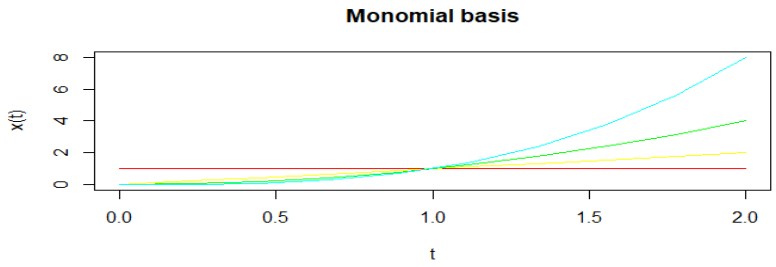

Figure 3: Monomial basis of order 4, this orderc goes from 1 to t^3

The monomial basis includes the **constant basis**, the horizontal line in the plot, and can form the **polynomial basis** doing a expressing it as a linear combination of monomial basis functions.

4.2 Fourier basis

Often we need to build functions that are periodic, the Fourier basis usually is the choice to build this type of functions because these repeat themselves over a certain period T. The Fourier basis are based in the Fourier series: [7]

$$\phi_1(t) = 1, \phi_2(t) = sin(\omega t), \phi_3(t) = cos(\omega t), \phi_4(t) = sin(2\omega t), \phi_5(t) = cos(2\omega t), \ldots, \phi_{m-1}(t) = sin(m\omega t), \phi_m(t) = cos(m\omega t) \quad (3)$$
$$where\ \omega = \frac{2\pi}{T}$$

To define a Fourier basis system we need two parameters. One is K that represents the number of basis functions, usually a Fourier basis comes in pairs of sin and cos therefore $K = 1 + 2m$. The other parameter is T, stands for the period, usually the period is the range of the values of t.

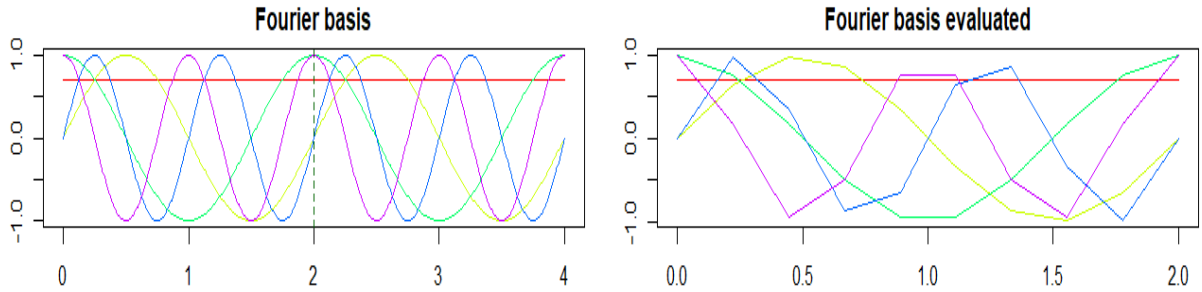

Figure 4: Fourier basis of period $T = 2$, $m = 2$ and $K = 5$

In the left panel we can see that the functions repeat themselves at point two that is the value of the period. The right panel shows the functions in the top panel evaluated at values $t = [0, 0.22, 0.44, ..., 2]$ these functions look like a trend over time, this one example where we can use Fourier basis.

4.3 B-Spline basis

The most complex basis listed, the B-Spline basis is useful to build non-periodic functions. B-Spline is a type of Spline basis, there are more types like Cubic Splines, V-Splines.

The Spline series are defined by four parameters: **break points**, **knots**, **order** and **degree**. A spline is constructed by dividing the whole record into subintervals, this subintervals are identified by **break points** at the end of each subinterval. Each subinterval is a polynomial function of a fixed degree or order, the **degree** is the highest power of the polynomial and the **order** is one higher than the **degree**. The last parameter is the **knots**, they have the same value as a break point but we can have multiple **knots** in one break point; the knots determine the number of derivatives of each break point.[7]

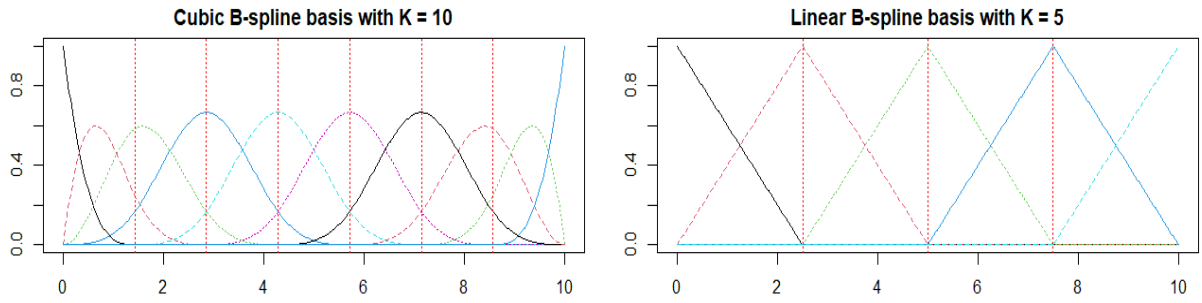

Figure 5: The left panel is a B-Spline which order equal to 4 and in the right panel a B-Spline of order 2

In the above figure each dotted vertical line represents the break points of each subinterval and one knot [5], equally spaced, per break point. The K represents the number of basis functions [6].

5 How to build a function using basis functions

The next figure shows 2 different functions built using basis functions.

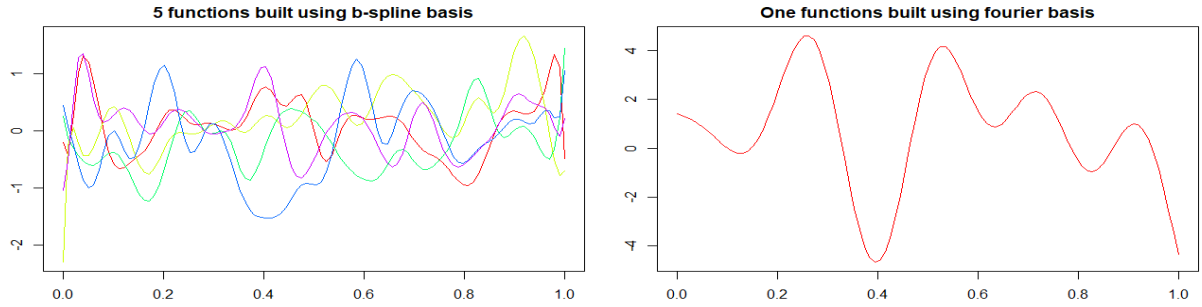

Figure 6: Functions built using basis functions

The left plot show 5 samples, each sample is built using bspline basis of $order = 5$, the number of basis functions is $K = 30$ and for the coefficients, each value is a random number, we need a matrix because we are sampling more than one function, its size is $5X30$ because we sampled 5 functions and used 30 basis functions per sample. The second panel shows just 1 sample, for this plot we use 40 Fourier basis functions of period equal to 4. In this case we don't need a matrix of coefficients instead we use a vector of length equal to 40, each value of the vector is a random number.

[5] Usually we only use the interior knot, this means the knots not placed either at the beginning or end of the function's domain

[6] To calculate K we use the next equation $K = order + numberofinteriorknots$

5.1 Using real data to build a function

We are going to use the **Canadian Weather** dataset provide in **fda** package in R. The dataset contains information related to the weather in some Canadian cities like Vancouver, Sydney, Quebec among others. Contains 3 types of measures: Temperature of each city measured in C°, daily average precipitation measured in mm and logarithm base 10 of the previous measure. The following plot shows the logarithm base 10 of daily average precipitation in mm of Sydney.

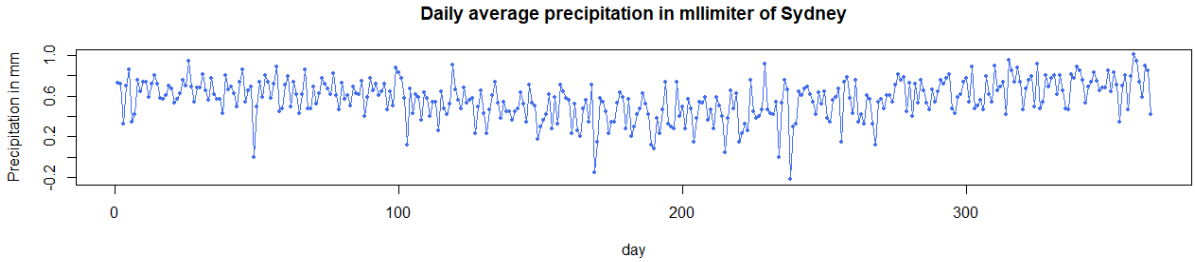

Figure 7: Each dot is the value associated to a day

In the plot we can see that some points differ significantly from other values, to minimize the difference we can *smooth* the observed points and build a function that match them. There are multiple approaches to smooth a function, we used the **linear regression** approach for smoothing. The result is the following.

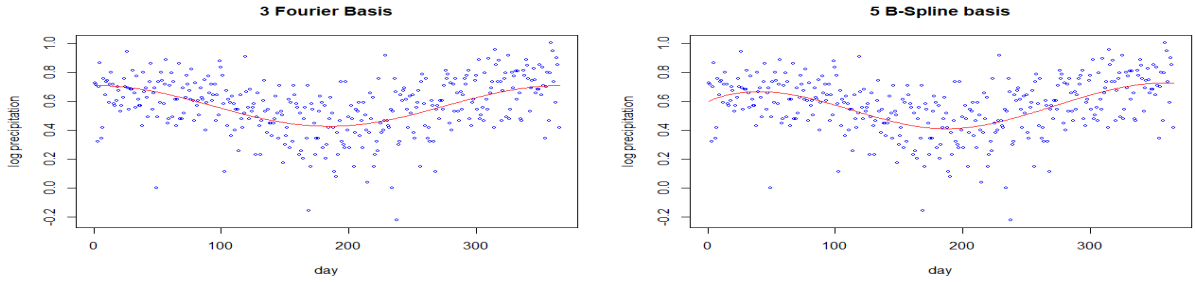

Figure 8: Function built using Fourier basis for the left panel and b-spline basis for the right

The left and right functions look quite similar, the process of building a function is try and error, often a Fourier basis can match very well a type of data and do a bad job in other, there are not a guide to build functions.

6 Conclusions

Have more approaches and methodologies to analyze data is important because increase of data generation and complexity of these is making more difficult their analysis. In practice, using the libraries available is not too difficult unlike knowing the basis and the mathematical theories that support this methodology .

References

[1] Jane-Ling Wang, Jeng-Min Chiou, and Hans-Georg Mueller. Review of functional data analysis. https://arxiv.org/abs/1507.05135, 2015. Accessed: 2022-08-04.

[2] Helle Sørensen, Jeff Goldsmith, and Laura M. Sangalli. An introduction with medical applications to functional data analysis. *Statistics in Medicine*, 32(30):5222–5240, 2013.

[3] Eric Weisstein. Smooth function. https://mathworld.wolfram.com/SmoothFunction.html, 2022. Accessed: 2022-08-04.

[4] Piotr Kokoszka and Reimherr Matthew. *Introduction to Functional Data Analysis*. CRC Press, New York, 2017.

[5] Staicu Ana-Maria and Park So Young. Short course on applied functional data analysis funding provided by nsf carrer.

[6] Jorge Mateu and Ramón Giraldo. *Geostatistical Functional Data Analysis*. Wiley, New York, 2022.

[7] J.O. Ramsay, Giles Hooker, and Spencer Graves. *Functional Data Analysis with R and Matlab*. Springer, New York, 2017.

[8] Ashutosh Kharb, Vipin Saini, Y Jain, Surender Dhiman, M Tech, and Scholar. A review of gait cycle and its parameters. *IJCEM Int J Comput Eng Manag*, 13, 01 2011.