

Functional Data Analysis

Prof S. Khare, Ayush Shukla

July 2015

1 Introduction

(The following text are key points of Chapter 1,2 and 3, from the book by J.O. Ramsay and B.W. Silverman) This is a very generic and remarkably versatile topic whose application includes in variety of fields such as
Bio-medical data - Psychometric, Medical tests, Drug discovery
Electronics - testing electronic chips, sensing and monitoring equipments etc.
Geography - Seismology, Meteorology
and many more fields...

Analysing the data helps to highlight various characteristic and explain variation in outcome. Learning the pattern & variation among the data and integrating it in various application aspects could be the next cutting edge technology. After approximating the data points by a function f , sometimes finding its differential operator may be of interest.

2 Steps in Data Analysis

1. Data representation : involves smoothing and interpolation 2. Feature alignment 3. Data display 4. Plotting derivatives graph : It may expose effects inconspicuous in the original function. The plot of second derivative vs first is called phase plane.

3 Tools to explore functional data

In functional data, we use the notion of scalars, vectors, derivatives, inner product space, composite function etc to do the analysis. For the statistics of functional data - means, variances, covariances and correlation functions are used. To get the relation between two variables, it is convenient to draw the graph wrt time on x-axis & y-axis and correlation on the z-axis. These are often visualized better by contour plots. For example, we draw the plots of correlation and cross correlation function for temperature and log precipitation. We read the plot as - autumn weather is highly correlated with spring weather as both temperature and precipitation correlation is high on either side of summer.

- looking at cross correlation, midwinter temperature and midwinter precipitation is high.

3.1 Function anatomy

Function features : Looking at the graph, following features may be of our interest : peak and valleys, derivative pattern (like the value at which acceleration is zero), location, amplitude, widths of peak.

3.2 Data resolution

High resolution data means the data set is very close and can pin down to small event. Function are infinite dimensional. Some functions are so erratic that no information is contained in $x(t)$ about $x(t+\Delta)$.

We draw the phase-plane plots to show how derivatives relate to each other, which could help analysing in some cases.

4 Function data to smooth function

It is easy to make analysis on smooth function. By smoothness, we mean that the function x possesses one or more derivative

$$Dx, D^2x, \dots, D^m x$$

The actual data may not be smooth due to presence of noise or measurement error. It may also be the case that the data shows high frequency variation. Practically we chose to ignore variation in small levels.

Suppose we have the function x and data vector as

$$y = (y_1, \dots, y_n)$$

we write the points as

$$y_j = x(t_j) + e_j$$

where the last term on right side of the equation is the roughness of the raw data. In vector form we can write equation as

$$y = x(t) + e$$

We also make the assumption that the error terms are independently distributed with mean zero and variance

$$(\sigma)^2$$

. Mathematically,

$$Var(y) = (\sigma)^2 I$$

4.1 Data Resolution

Resolution of data gives the idea of functional data analysis. The *curvature* of a function is measured by second derivative. The key observation after looking at certain results tells that the high sampling rate causes differencing to greatly magnify the influence of noise. This is caused by taking differences between extremely close values that magnifies the error.

4.2 Representing functions by basis function

A basis function is the set of function independent of each other and have the property that we can approximate any function by a weighted sum or linear combination of a sufficiently large K of these function. Like the power series

$$1, t, t^2, t^3, \dots, t^k, \dots$$

or fourier series

$$1, \sin(wt), \cos(wt), \sin(2wt), \cos(2wt), \dots$$

Mathematically we can write them as

$$x(t) = \sum c_k f_k(t)$$

The smaller the K is, the better the basis function reflects certain characteristic of data.

To approximate a function by basis function we use Fourier basis for periodic data and a B-spline for non periodic data. We also find that we use fourier basis when there are no strong local features, curvature is of same order and function is stable.

In spline functions, we define the degree of freedom as the order of the polynomial plus the number of interior break points. By increasing the number of breakpoints we can gain flexibility. Other useful basis systems are :

Wavelets, exponential and power bases, polynomial bases