Introduction to Functional Data Analysis

Elia Liitiäinen (eliitiai@cc.hut.fi)

Time Series Prediction Group
Adaptive Informatics Research Centre
Helsinki University of Technology, Finland

January 30, 2007

Introduction

- Functional data occurs for example in time series analysis, chemometry and econometry.
- In many cases the amount of samples available is small.
- Taking the structure of the inputs into account improves results of statistical inference.
- FDA is a framework that provides tools for this purpose.

Outline

1. General Considerations
2. Correlation analysis
3. Interpolation

Goal of Functional Data Analysis

- Exploratory data analysis: Data provides new information and sheds light on known features.
- Confirmatory analysis: Hypothesis testing.
- Prediction: Prediction of the future.

Functional Data

- Real world phenomena are usually continuous at small enough time scale.
- The worst-case dimension of functional data is infinite (white noise).
- For smooth functions with bounded derivative the instrinsic dimension is finite.
- Typically for smooth functions the practical dimension is 10-20.

Noise

- Typically in function data there is noise.
- In mathematical terms

$$x_i(t) = y_i(t) + \epsilon(t). \tag{1}$$

- To make things worse, often $\mathrm{Cov}(\epsilon(t_2), \epsilon(t_1)) \neq 0$ for $t_2 \neq t_1$.

Data Representation

- The form of the curve is important.
- The first step in FDA is transformation of the inputs to remove noise.
- Basic tools include smoothing and interpolation.

Derivatives

- Derivatives are important.
- Numerical differentiation amplifies noise.
- Interpolation or smoothing helps in this regard.

Covariance and Variance Functions

- $\{x_i(t)\}_{i=1}^{N}$ is a sample of functions.
- Mean:

$$\bar{x}(t) = N^{-1} \sum_{i=1}^{N} x_i(t). \tag{2}$$

- Variance function:

$$\text{var}_X(t) = (N-1)^{-1} \sum_{i=1}^{N} [x_i(t) - \bar{x}(t)]^2. \tag{3}$$

- Covariance Function

$$\text{cov}_X(t_1, t_2) = (N-1)^{-1} \sum_{i=1}^{N} \{x_i(t_1) - \bar{x}_i(t_1)\}\{x_i(t_2) - \bar{x}_i(t_2)\}. \tag{4}$$

Correlation

- Correlation function:

$$\text{corr}_X(t_1, t_2) = \frac{\text{cov}_X(t_1, t_2)}{\sqrt{\text{var}_X(t_1)\text{var}_X(t_2)}}. \tag{5}$$

It is often useful to examine the plot of cross-correlation.

Cross-correlation

- Now we have pairs of functions (x_i, y_i).
- Cross-covariance:

$$\text{cov}_{X,Y}(t_1, t_2) = (N-1)^{-1} \sum_{i=1}^{N} \{x_i(t_1) - \bar{x}(t_1)\}\{y_i(t_1) - \bar{y}(t_1)\}. \tag{6}$$

- Cross-correlation:

$$\text{corr}_X(t_1, t_2) = \frac{\text{cov}_{X,Y}(t_1, t_2)}{\sqrt{\text{var}_X(t_1)\text{var}_Y(t_2)}}. \tag{7}$$

Case Study: Tecator Data

- 240 samples of absorbance spectrums.
- In addition to the absorbance spectrums we have fat content as output.
- The cross-correlation with the output can be misleading.

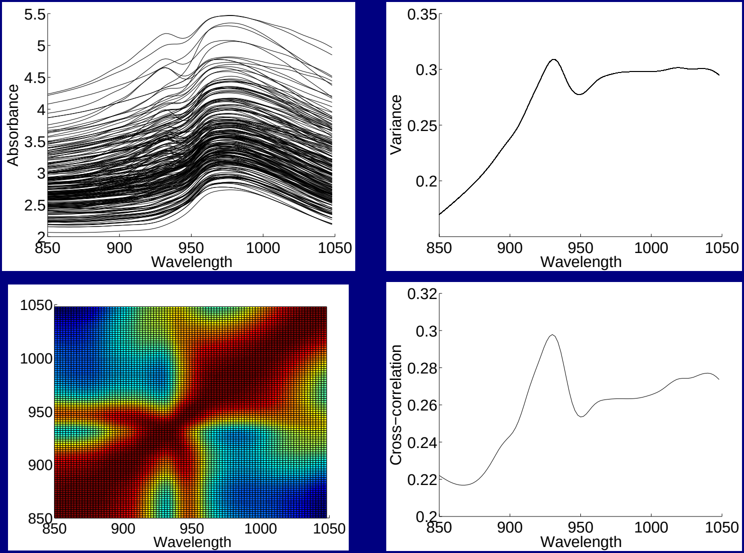

Figure: From left to right: the inputs, the variance function, the correlation function and the cross-correlation with the scalar output.

Function Basis

- A basis is a linearly independent set of function $\{\omega_i\}_{i=1}^{\infty}$ that spans the function space.
- Example: the set of monomials $\{t^i\}_{i=0}^{\infty}$.
- Basis expansion: the functional inputs $\{x_i(t)\}_{i=1}^{N}$ are approximated as (for some finite $K > 0$)

$$x_i(t) \approx \sum_{k=1}^{K} c_k \omega_k(t). \tag{8}$$

- The weights are solved by minimizing some cost function.

Why to use basis expansions?

- Dimension reduction.
- Reduces computational demand in later stages of analysis.
- Noise removal.

Fourier Basis

- Fourier basis on $[0,1]$ is $\{\sin 2\pi jt, \cos 2\pi jt\}_{j=1}^{\infty}$.
- Sometimes good for periodic data.
- Lack of locality.
- Computational complexity $O(N \log N)$.

Wavelets

- Under some conditions, the functions

$$\psi_{jk}(t) = 2^{j/2}\psi(2^j t - k) \tag{9}$$

  form a basis.
- Wavelets are local.
- Fast computation.

Splines (1)

- Consider the interval $[0, 1]$ and the breakpoints $\tau = \{\tau_l\}_{l=0}^{L}$ with $\tau_0 = 0$ and $\tau_L = 1$.
- A spline is piecewise polynomial with degree K.
- At the breakpoints it is required that the values of the polynomials and derivatives up to $K - 1$ agree.
- Thus a spline is K-1 times differentiable.
- For $K = 1$, spline is a piecewise linear function.

Splines (2)

- The number of intervals: L.
- Degrees of freedom:

$$LK - (L-1)(K-1) = K + L - 1, \tag{10}$$

  that is, the number of interior knots plus the order.
- It is not necessary to require same smoothnes in all the knots.

Spline Basis

- Splines can be represented using a basis expansion

$$S(t) = \sum_{k=1}^{K+L-1} c_k B_k(t). \tag{11}$$

- The basis is not orthonormal the locality being determined by K (complexity grows linearly with respect to the number of data).
- The coefficients can be used in regression and data analysis.

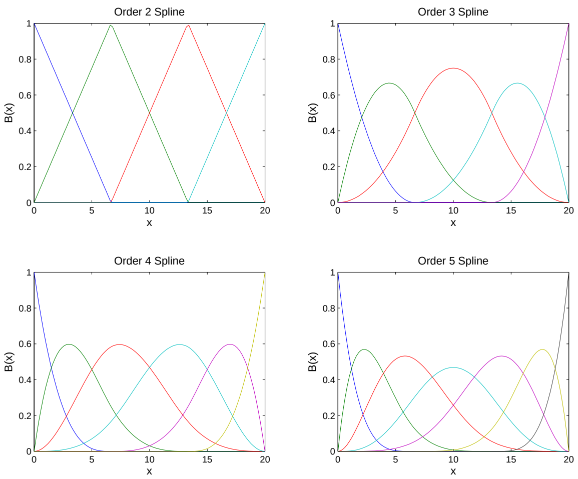

Figure: Spline basis for different orders.

Conclusion

- Functional data occurs in real world.
- Important tools include correlation plots, derivatives and basis expansions.
- Removal of noise is needed.