# *Introduction to Nonparametric Functional Data Analysis*

Elia Liitiäinen (eliitiai@cc.hut.fi)

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

March 27, 2007

# *Introduction*

- The functional context allows various non-classical tools.
- The infinite dimension of the data poses a challenge for nonparametric methods.
- In this presentation basic concepts for understanding functional data are introduced.

# *Outline*

# *Finite-dimensional space*

- In the finite-dimensional vector space $\Re^n$, we may define the norms
$$\|x\|_p = (\sum_{i=1}^{n} x(i)^p)^{1/p}. \tag{1}$$
- The norm generates a metric $d(x, y) \geq 0$ with the properties
  - $d(x, y) = d(y, x)$
  - $d(x, z) \leq d(x, y) + d(y, z)$
  - $d(x, y) = 0$ if and only if $x = y$.
- The definition of metric generalizes to a more general class of spaces.

# *Functional Space*

- In functional data analysis instead of vectors, a set of functions $(f_i)_{i=1}^{M}$ is available.
- The functions can be considered as points in an infinite dimensional space $X$ with a metric $d$.
- The $L^2$-norm is a common choice ($I$ is the domain, for example a range of frequencies):

$$\|f_i\| = (\int_I |f_i(t)|^2 dt)^{1/2} \tag{2}$$

# *Semimetrics*

- A semimetric satisfies the properties of metric, except that $d(f,g) = 0$ may hold for $f \neq g$.
- Often an useful choice is using derivatives:

$$d_q(f,g) = \left(\int_I |f^{(q)}(t) - g^{(q)}(t)|^2 dt\right)^{1/2}. \tag{3}$$

- Many classical techniques like PCA can be implented with respect to a semimetric.

# *PCA as a semimetric*

- Denote by $v_1, \ldots, v_q$ the principal components in data.
- Then an useful seminorm can be defined by

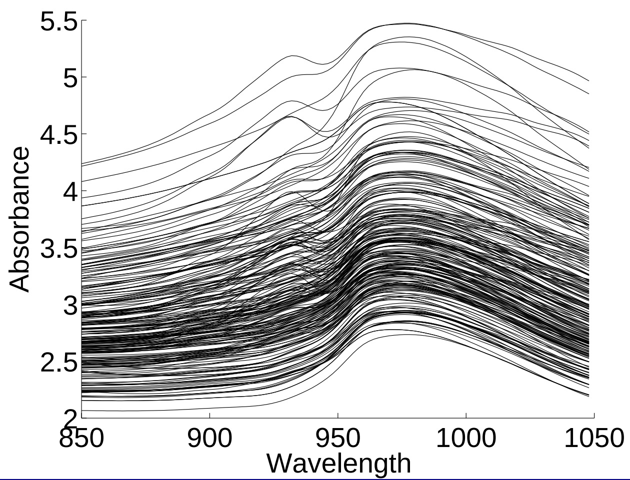
$$\|f\|^{\mathrm{PCA}} = (\sum_{i=1}^{q}(\int_I f(t)v_i(t)dt)^2)^{1/2}. \tag{4}$$

- Thus PCA offers an useful way to build a semimetric.

# *Curse of Dimensionality*

- The effective dimensionality of functional data is often relatively low.
- Typically the first few principal components explain most variability in the data, a phenomen with big industrial applications.

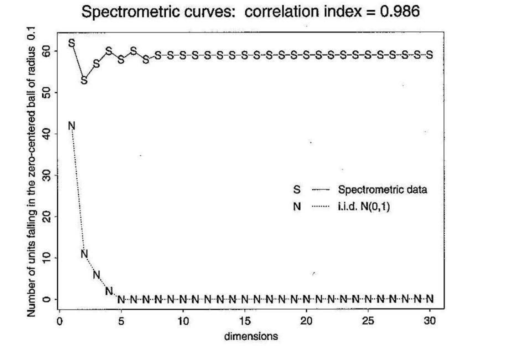
# *Tecator Data*



# *Experiment on the dimensionality of the Tecator Data Set*

- For $p = 1, \ldots, 30$, we discretize the Tecator data set with a grid of p points.
- After that the squares of the Euclidean norms of the resulting vectors are calculated.
- The resulting set of scalar is scaled so that all values are between 0 and 1.
- Finally the number of those points that fall to $(0, 0.1)$ are calculated.
- The result is plotted together with the same experiment for Gaussian i.i.d vectors.

# *Result*




- Average correlation/PCA reveals the reason behind the result: a large part of the variance in the data can be explained with just one variable.
- Instrinsic dimensionality estimation gives $\approx 2$.

# *Functional Kernels*

- Consider the functional i.i.d. random variables $(f_i)_{i=1}^{M}$.
- A kernel is a positive function on $\Re$.
- With the semimetric $d$, local weighting is given by

$$\delta_i(g) = \frac{K(d(g, f_i)/h)}{E[K(d(g, f_i)/h)]}. \tag{5}$$

- The expectation can be approximated with empirical mean.

# *Classification of Kernels*

- We always assume that $\int K = 1$.
- Type I:

$$C_1 1_{[0,1]} \leq K \leq C_2 1_{[0,1]} \tag{6}$$

- Type II kernel has the support [0,1] and a non-positive derivative with

$$C_2 \leq K' \leq C_1 < 0. \tag{7}$$

- An example of type II kernel is $K(u) = 2(1-u)I_{[0,1]}(u)$.

# *Small Ball Probabilities*

- The small ball probability is a probabilistic concept related to dimensionality defined by

$$\phi_g(h) = P(f_i \in B(g, h)). \tag{8}$$

- Under realistic assumptions, for type I and type II kernels,

$$C\phi_g(h) \leq E[K(d(g, f_i)/h)] \leq C'\phi_g(h) \tag{9}$$

for some constants $C, C'$.

# *Conclusion*

- In this presentation formal tools for functional data analysis were presented.
- Analysis of the properties of functional data offers relatively unexplored possiblities both for applications and basic research.