T-61.6030: Special Course in Computer and Information Science III: Introductory Elements of Functional Data Analysis
Project Work, Spring 2007
Helsinki University of Technology

For the project you are expected to familiarize with the Functional Data Analysis Toolbox either using R or Matlab and, run experiments on a set of given datasets. The deadline for submission is May, 20. Send via email (corona,lendasse,elia@cis.hut.fi) copies of your report in PDF format. Use the provided LateX template for writing down the report.

0) Download the data from the course web page. The three provided datasets (lab.txt, pilot.txt and full.txt) consists of input spectral measurements and the corresponding scalar output. The domain arguments are common to all the datasets (wavenumber.txt). Each datasets is formatted in such a way that the rows corresponds to observations and the columns to the variables; for each dataset, the output is in the last column. From the three datasets provided, choose only two.

1) For both datasets, implement the code that calls the toolbox functions to perform the following tasks:

   a.) Smooth functional data by least squares (Chap. 4, Ramsay's book) using 4-th order B-splines. Select the "optimal"number K of basis functions using a Leave-One-Out criterion averaged over all curves and monitor the norm of the 2-nd derivative. In order to limit the computational time, the optimal number could be selected using an appropriate integer spacing for the grid, e.g. k=[kmin:10:kmax];

   b.) Smooth functional data with a roughness penalty (Chap. 5, Ramsay's book) using 4-th order B-splines and penalizing the 2-nd derivative. Select the "optimal"penalty term λ using the generalized cross-validation criterion and monitor the norm of the 2-nd derivative. In order to limit the computational time, the optimal lambda could be selected using an appropriate logarithmic spacing for the grid, e.g. $\lambda = 10^{([-7:1:+7])}$;

2) Using the smooth curves from 1b.), perform the following tasks:

   a.) Functional Principal Components Analysis (Chap. 8, Ramsay's book, without regularization for the components) with one of the datasets;

b.) Functional Canonical Components Analysis (Chap. 11, Ramsay's book) with both datasets. The choice of the basis/regularization for the weight functions can be done in an intuitive way to get fairly smooth curves;

3) Choose only one dataset, divide it in learning and testing subsets (e.g., 2/3 of the observations for learning, and the rest for testing). Then, perform Functional Linear Regression for scalar responses (Chap. 15, Ramsay's book) using crossvalidation for choosing the amount of regularization for the weight function. The number of basis functions used for the weight function can be reduced to speed-up the computation if necessary.

4) Report and discuss on the obtained results and provide a brief description on the used methods. The expected length for the report is from 8 to 10 pages. Do not hesitate to include figures and your comments in the report.

   This project should require no more than 15 hours for coding and reporting. Calculation times are subject to the computing resources that you have at your hands; hence, start early!