Special Course in Computer and Information Science III L

Introductory Elements of Functional Data Analsysis

Francesco Corona, Amaury Lendasse and Elia Liitiäinen

The project, in brief

Familiarize with the basic methods of FDA

- hands on approach

Run experiments on given datasets:

- using an existing toolbox

Write down a scientific report:

- brief description of the methods
- discussion on performed experiments
- figures and comments on results

A toolbox for FDA (Matlab, R and S-PLUS)

Download and install:

- ftp://ego.psych.cghill.ca/pub/ramsay/FDAfuns/

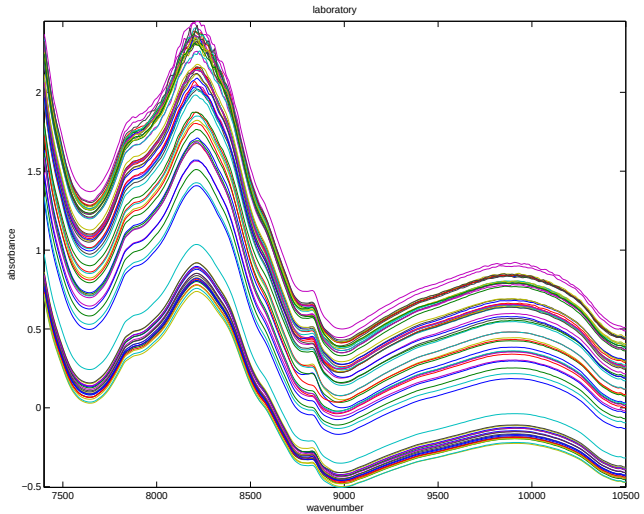
Use the R's FDA manual for reference:

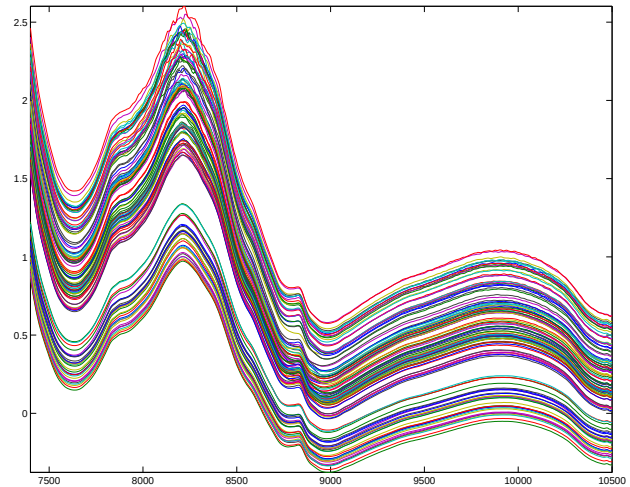
- http://ftp.sunet.se/pub/lang/CRAN/

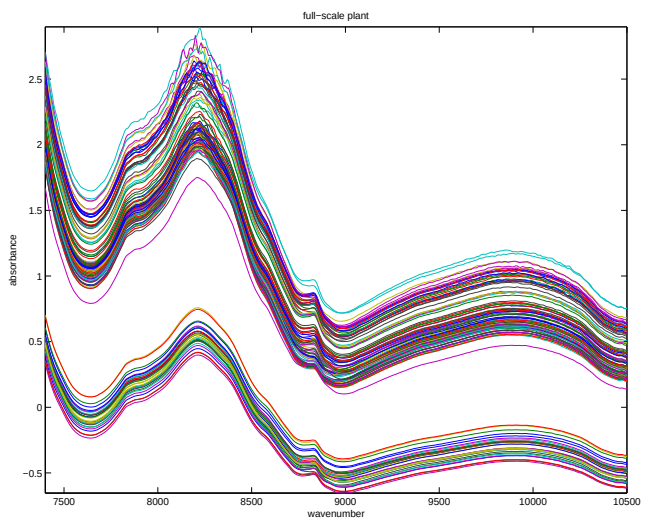
On methods and experiments:

- Ramsays' books and examples
- http://www.functionaldata.org

Three datasets (Pharmaceutical industry)







In the data

Laboratory data:

- 70 functional observations (Y) and the scalar output (z)

Pilot plant data:

- 120 functional observations (Y) and the scalar output (z)

Full-scale plant data:

- 120 functional observations (Y) and the scalar output (z)

The domain is common:

- 404 arguments (x, or t)

Download and select two

Project data:

- http://www.cis.hut.fi/Opinnot/T-61.6030/

Original data:

- http://www.models.kvl.dk/research/data/Tablets/index.asp

Data description:

- Dyrby, M, S. B. Engelsen, N. Nörgaard, M. Bruhn, L. Lundsberg-Nielsen (2002). Chemometric quantification of the active substance (containing $C \equiv N$) in a pharmaceutical near-infrared (NIR) transmittance tablet using NIR FT-Raman spectra. *Applied Spectroscopy* **56**, 579-85.

The project, in detail

After downloading/installing the toolbox and the data

- Represent the data as functions

1. smooth by least squares
2. penalized smoothing

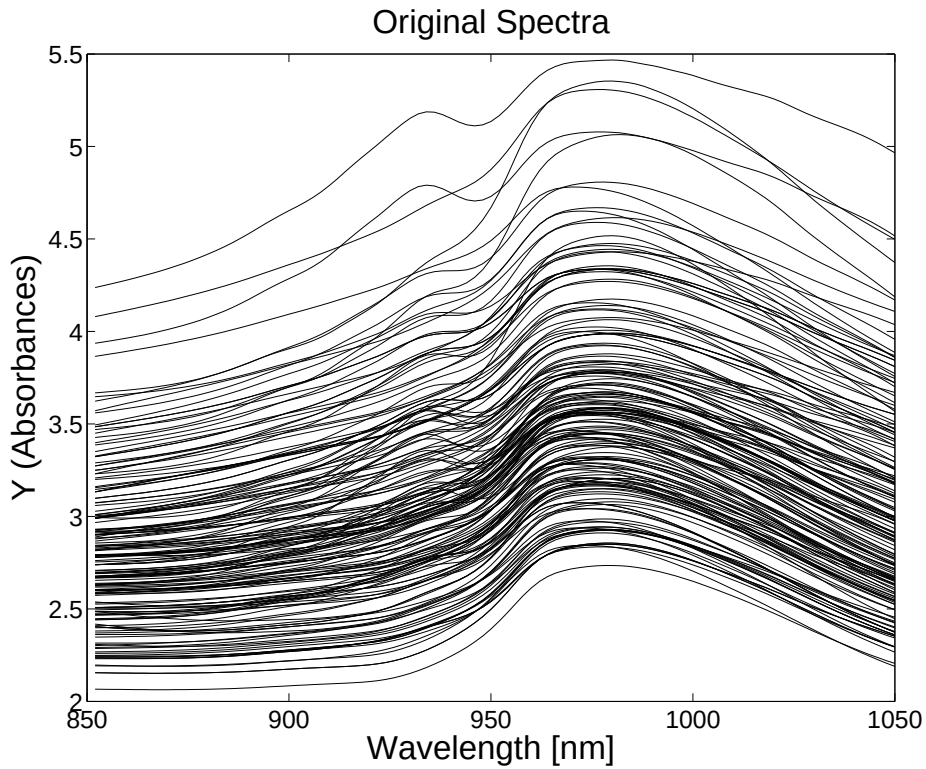
- Perform some basic analysis

1. principal components analysis
2. canonical correlation analysis

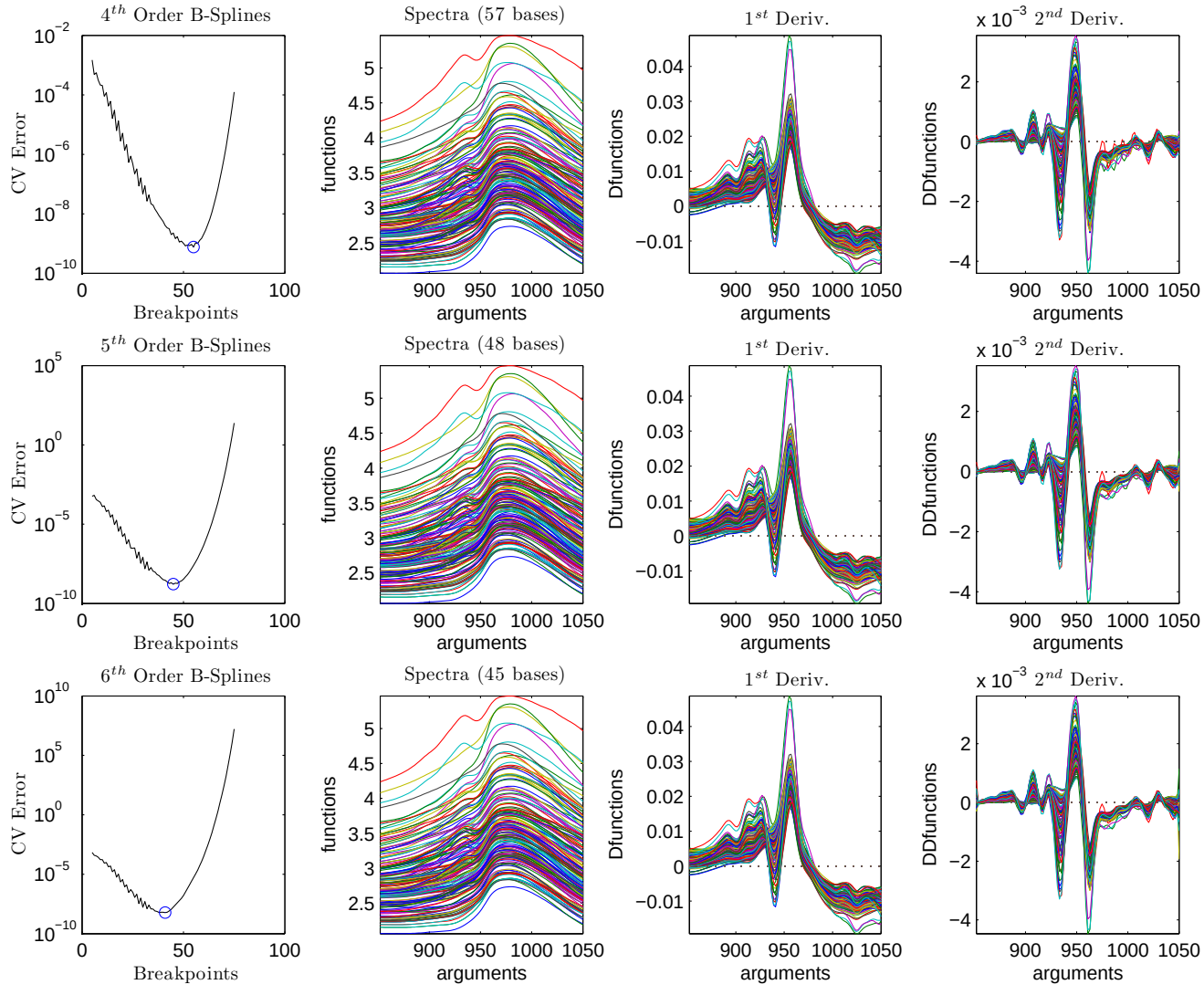
- Build a regression model

1. functional regression for scalar output
2. test the results with independent data

An example


Original Spectra
Y (Absorbances)
Wavelength [nm]
5.5
5
4.5
4
3.5
3
2.5
2
850
900
950
1000
1050

Smoothing by least squares - selecting K



Smoothing by least squares - selecting K (cont.)

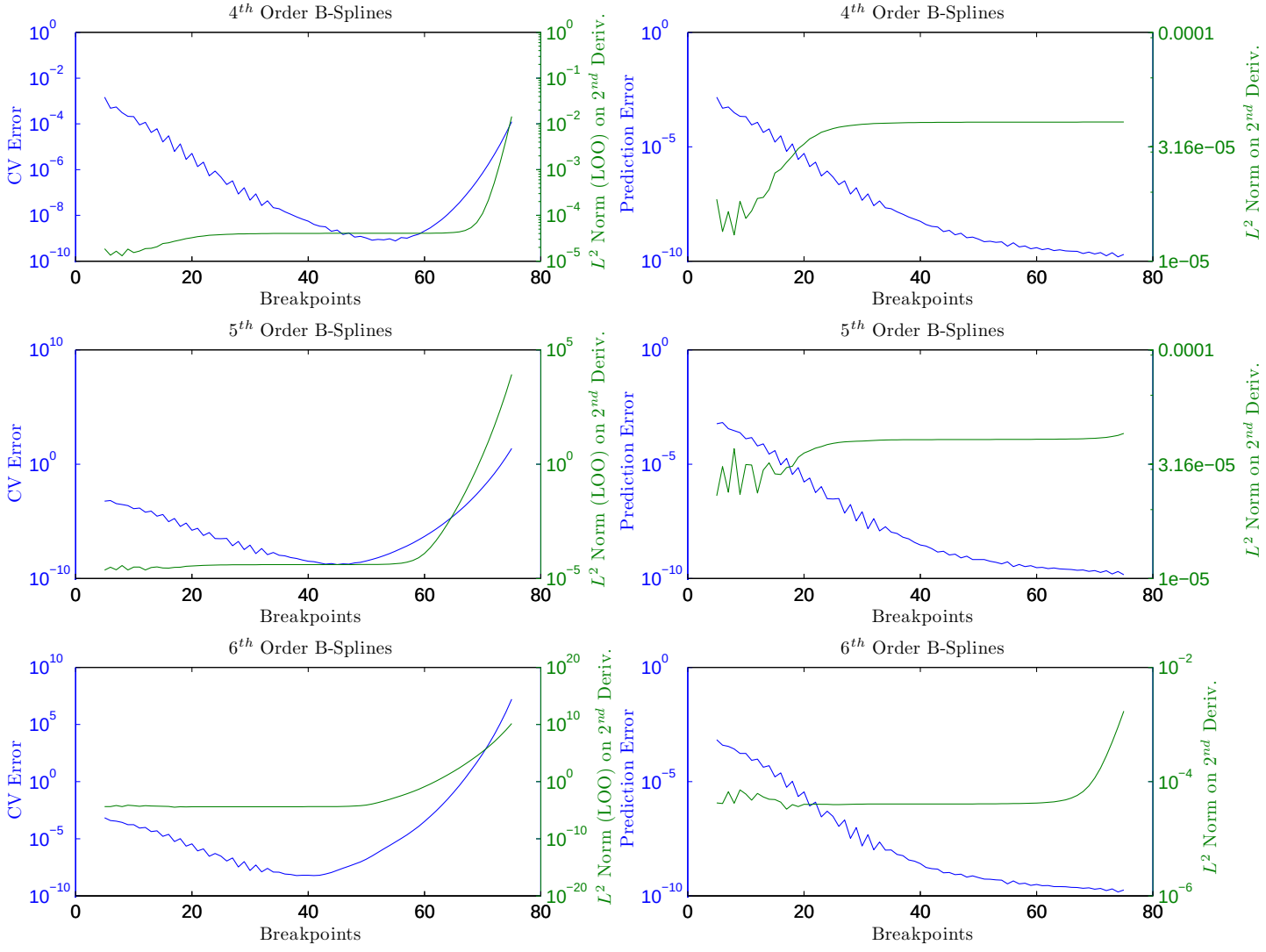
Use 4:th order B-splines and the LOO criterion

- average over all functional observations

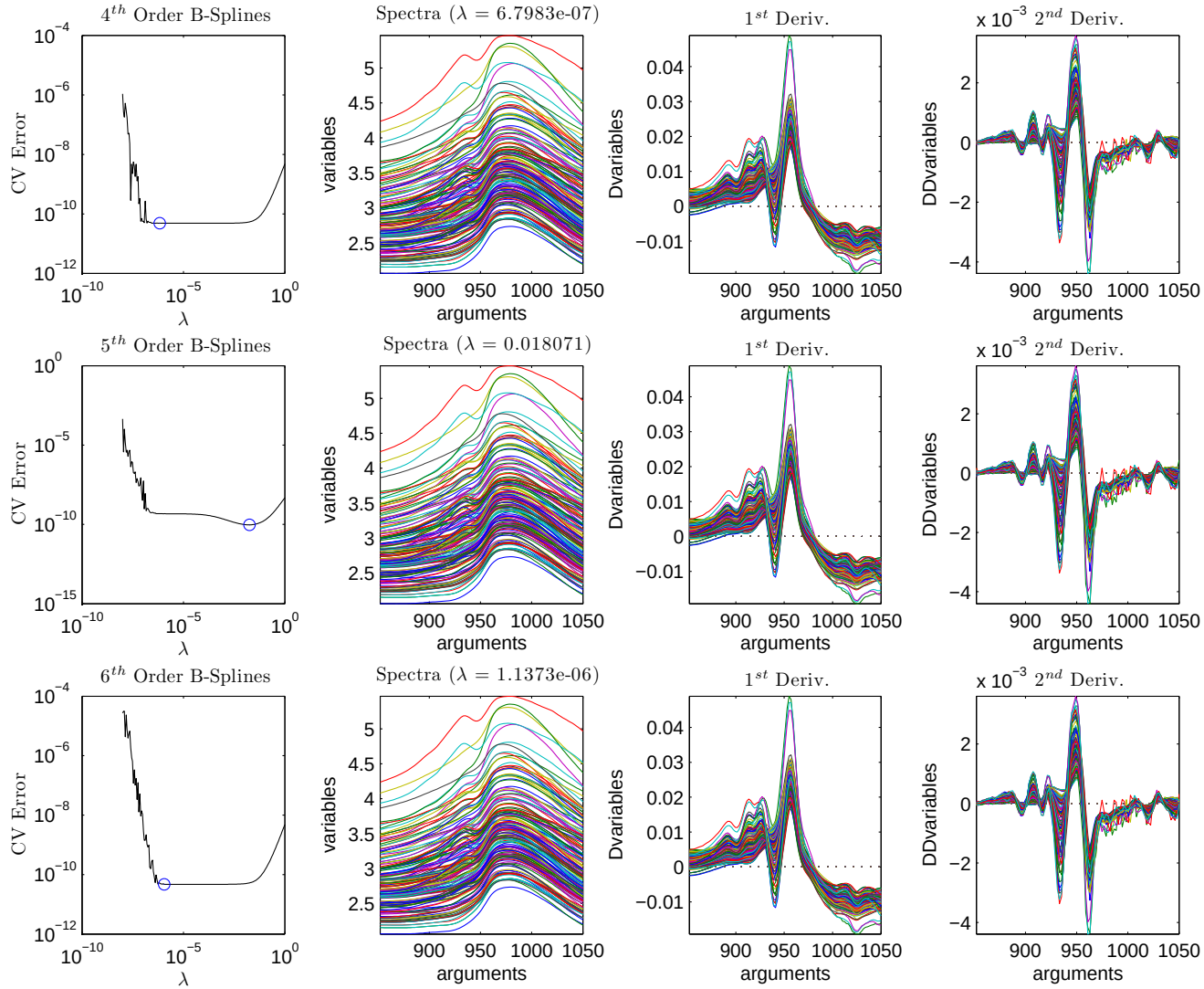
Can be time consuming:

- k=[kmin:10:kmax]

Smoothing by least squares (monitoring the 2:nd derivative)



Penalized smoothing - selecting λ



Smoothing by least squares- selecting λ (cont.)

Penalize 2:nd order derivatives and use GCV:

- average over all observations
- use K=kmax

Can also be time consuming:

- λ=10^[-8:1:+8]

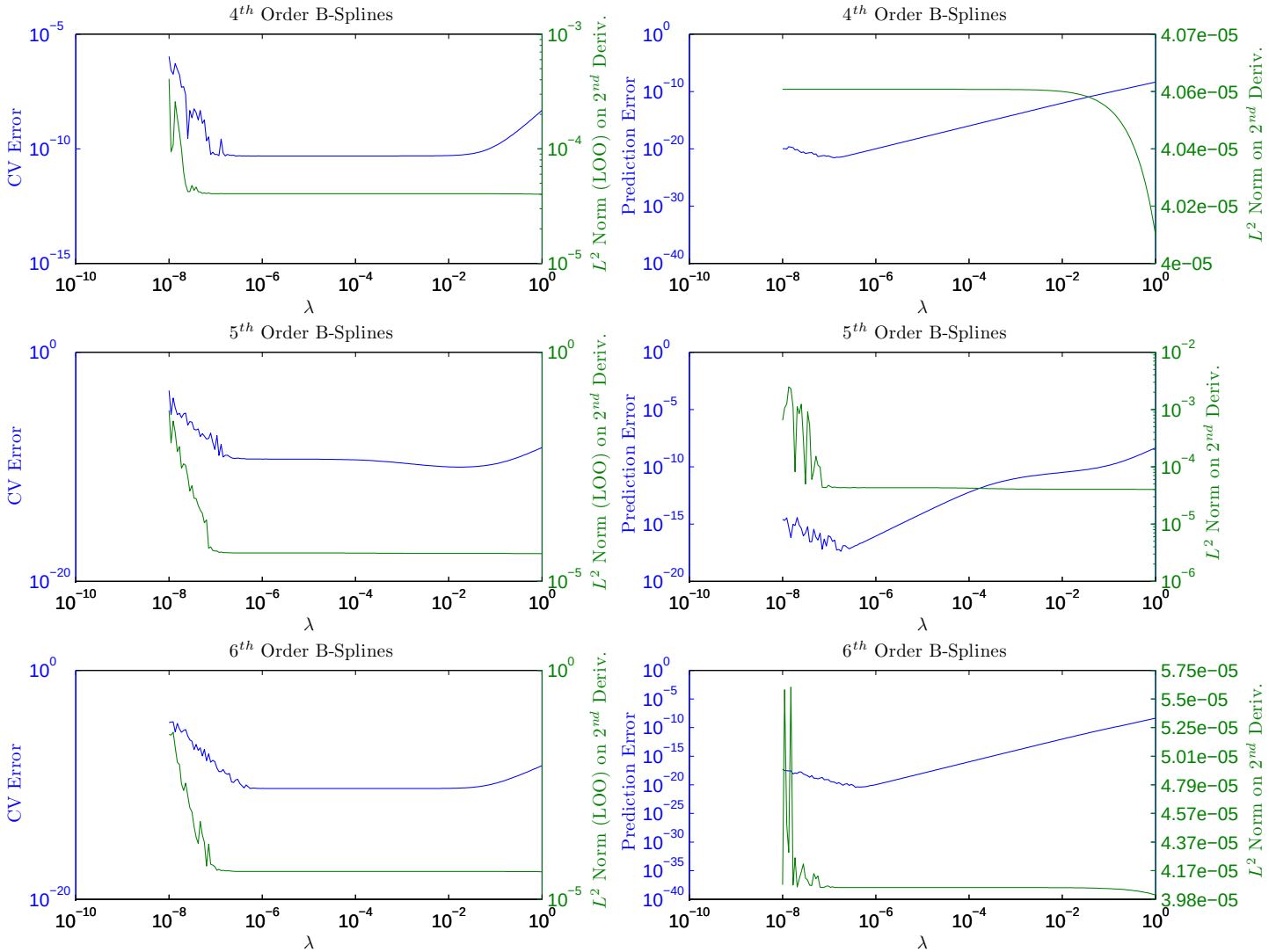
If it still too long:

- reduce K (e.g. K=1/2 kmax)

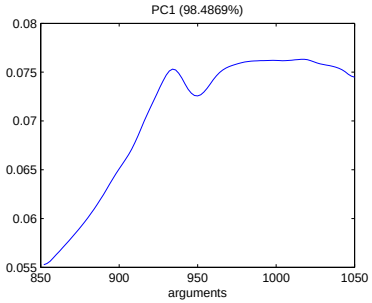
Compare GCV with LOO:

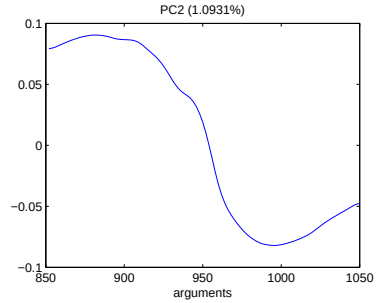
- not required

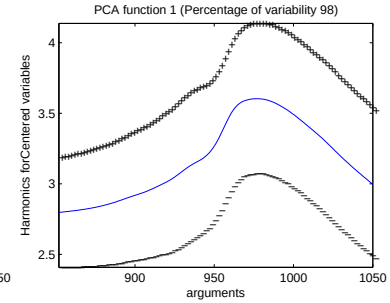
Penalized smoothing (monitoring the 2:nd derivative)

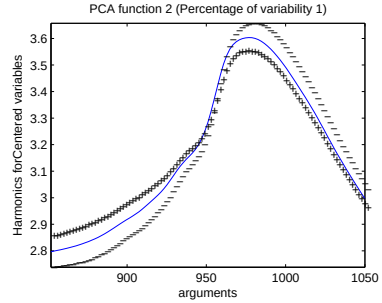


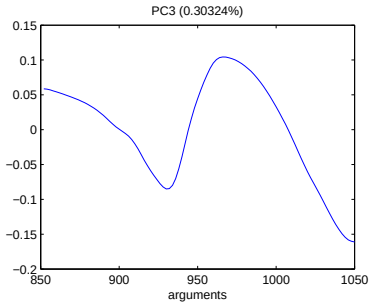
Functional projection - PCA

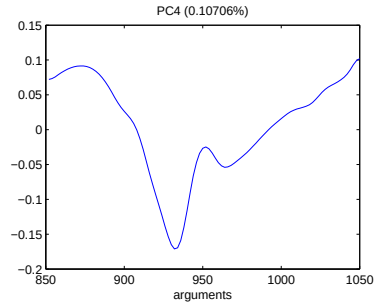


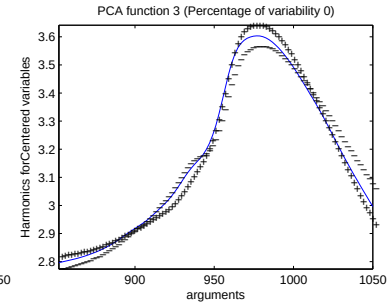


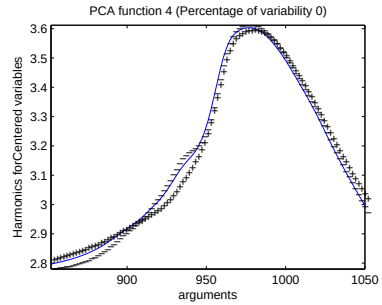


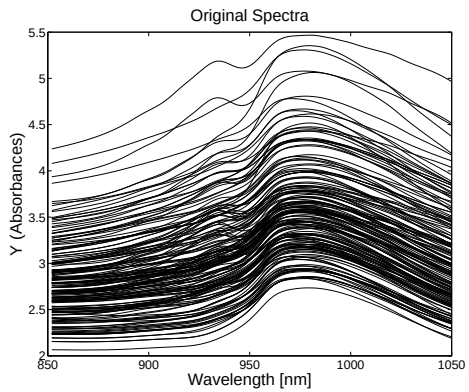




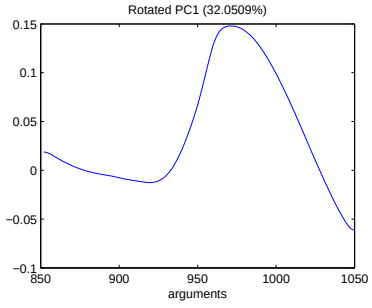


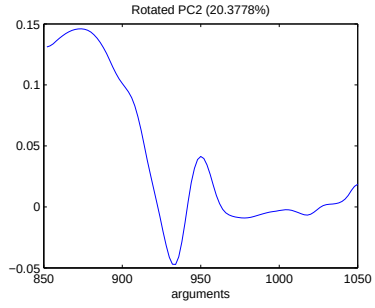


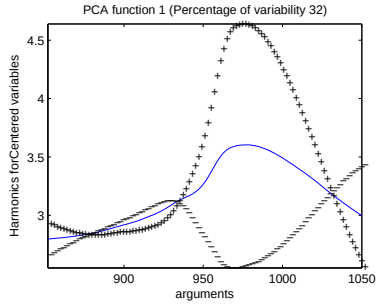


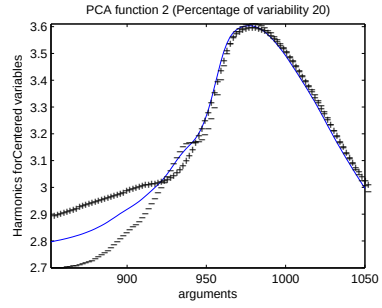


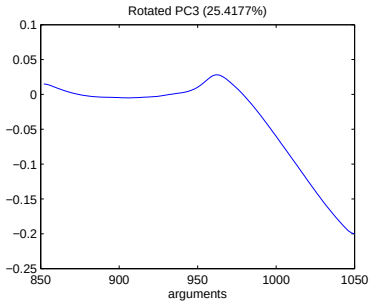
Functional projection - PCA + VariMax

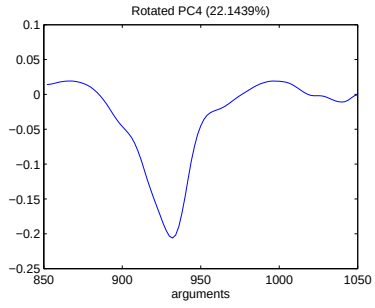


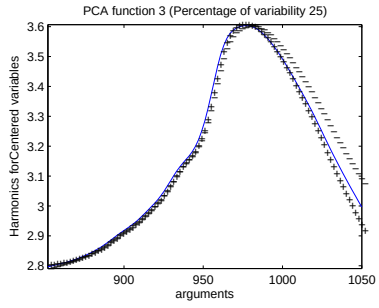


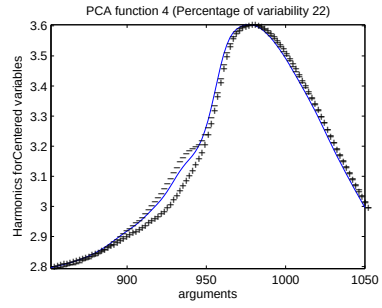












Functional projection and regression

Perform regularized PCA +/- VariMax rotation

- try different penalties
- not required

Perfom CCA using the two selected dataset:

- the number of observation must be same
- pilot + full should do
- but we'd like to have lab

Perfom regression using only one dataset:

- cross-validate the pernalty

May require a fixed function

- ask, if needed

Submission

Use the LaTeXtemplate (8-10 pages, 15 at most):

- http://www.cis.hut.fi/Opinnot/T-61.6030

Send a copy of the report in PDF format:

- {elia,lendasse,corona}@cis.hut.fi
- May, 20!