Mutivariate Analysis Tools for Time Series Data and Knowledge Extraction

Ali Mohammad-Djafari

Groupe Problèmes Inverses
Laboratoire des Signaux et Systèmes
UMR 8506 CNRS - SUPELEC - Univ Paris Sud 11
Supélec, Plateau de Moulon, 91192 Gif-sur-Yvette, FRANCE.

djafari@lss.supelec.fr
http://djafari.free.fr
http://www.lss.supelec.fr

C5Sys, ERASysBio, CNRS, Villejuif, France, Jan 11-12, 2013

March 24, 2013

Summary

- We always have many time series to analyse:
  - Genes expressions,
  - Proteins, Physiological quantities, ...
- We want to visualize them
  - Time domaine
  - Transformed domain: Fourier, Wavelets, Splines, ...
- We want to model them to summarize their information content
  - Parametric: Fourier series, Superposition of Gaussians shapes, ...
  - Non parametric: Markovian models
- We want to model the relations between them
  - Linear / Non linear
  - Training and test data
- We want to extract knowledge from them.

Summary

We developped easy running tools for:

- Visualization:
  Time series, spectra, histograms, scatterplots, ...
- Simple Analysis:
  Computing spectra, Estimating periods, ...
- Mutivariate Analysis: Dimensional Reduction
  PCA, FA, ICA, Sparse PCA for dimensional reduction and main factors extraction
- Mutivariate Discriminant Analysis
  LDA, EDA, RDA, Sparse LDA for finding the most discriminant factors
- Correlation (Pearson or Spearman) computation and dependancy graph visualization
- Modelling input-output relations

Deterministic and Bayesian Factor Analysis

- Gaussian case: $p(\boldsymbol{g}|\boldsymbol{A}, N) = \mathcal{N}(\boldsymbol{0}, \boldsymbol{A}\boldsymbol{\Sigma}_f\boldsymbol{A}^t + \boldsymbol{\Sigma}_\epsilon)$
- Deterministic methods:

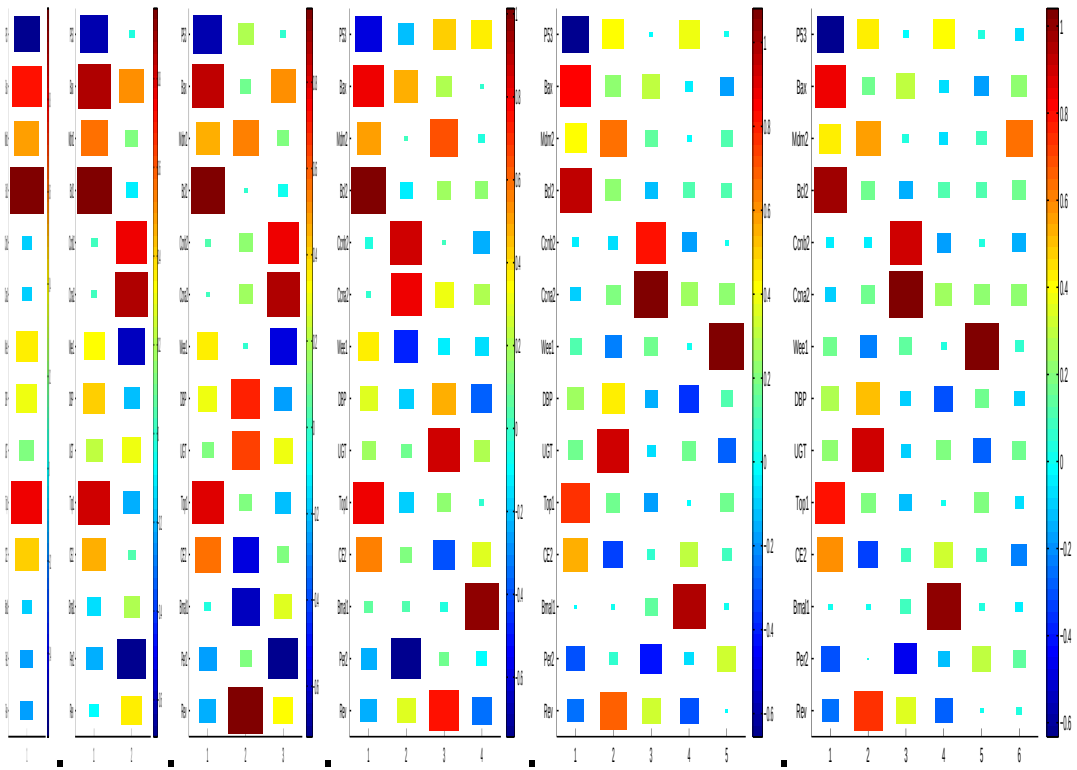
$$(\widehat{\boldsymbol{A}}, \widehat{\boldsymbol{f}}) = \arg\min_{(\boldsymbol{A},\boldsymbol{f}} \left\{ \|\boldsymbol{g} - \boldsymbol{A}\boldsymbol{f}\|^2 \right\} \text{ s.t. constraints on } \boldsymbol{A} \text{ and } \boldsymbol{f}$$

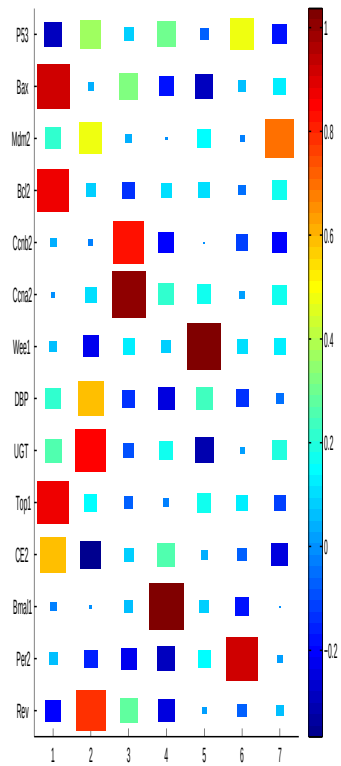
  Uncorrelated (PCA), Independent (ICA)
- Bayesian methods:

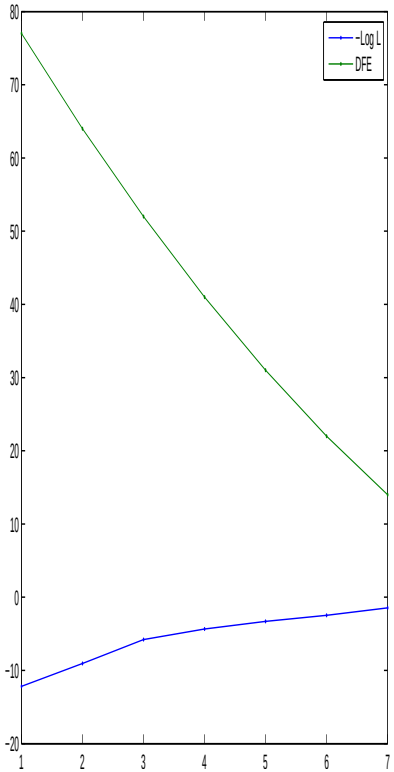
$$(\widehat{\boldsymbol{A}}, \widehat{\boldsymbol{f}}) = \arg\min_{(\boldsymbol{A},\boldsymbol{f})} \left\{ \|\boldsymbol{g} - \boldsymbol{A}\boldsymbol{f}\|^2 + \lambda_1\|\boldsymbol{A}\|^{\beta_1} + \lambda_2\|\boldsymbol{f}\|^{\beta_2} \right\}$$

  $\beta_1 = 1$ and $\beta_2 = 1$ leads to sparse solutions
- To determine the number of factors we do the analyse with different N factors and use two criteria:
  -log likelihood $-\ln p(\boldsymbol{g}|\boldsymbol{A}, N)$ of the observations and
  DFE: Degrees of freedom error $(N - M)^2 - (N + M))/2$
  related to AIC or BIC model selection criteria.
- These analysis can be done either directly on time series or on FT amplitudes.

Factor Analysis: Time series, colon



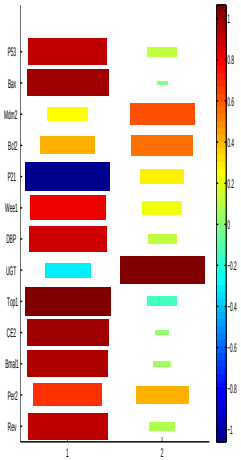




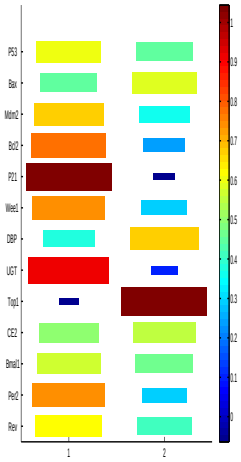
Factor Analysis for each class: FT, Liver

Two Factors:

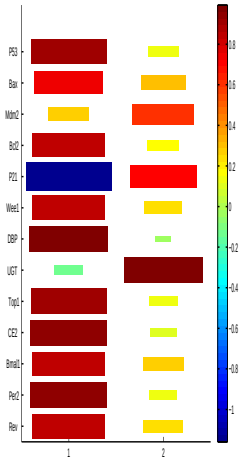
Class 1



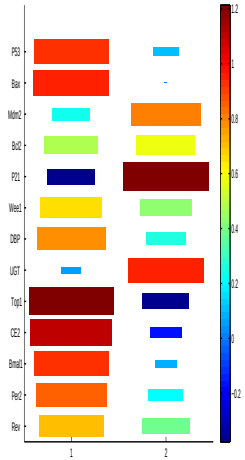
Class 2



Class 3

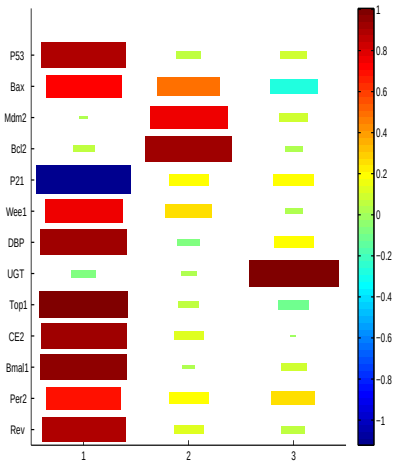


All Classes

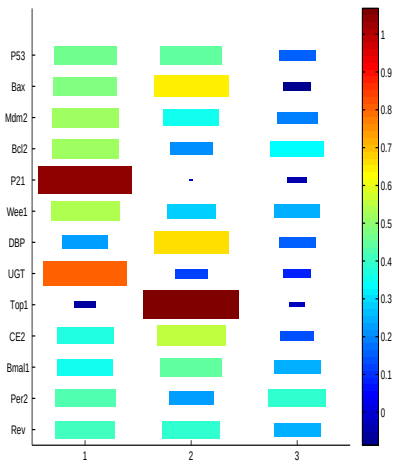


Three Factors:

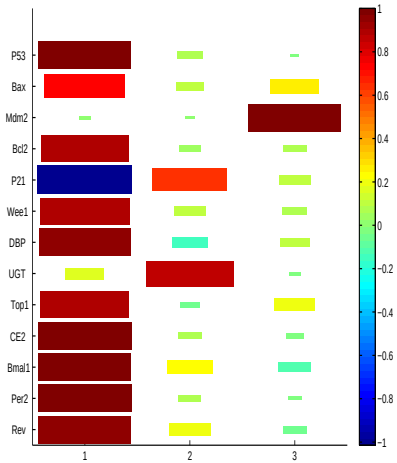
Class 1



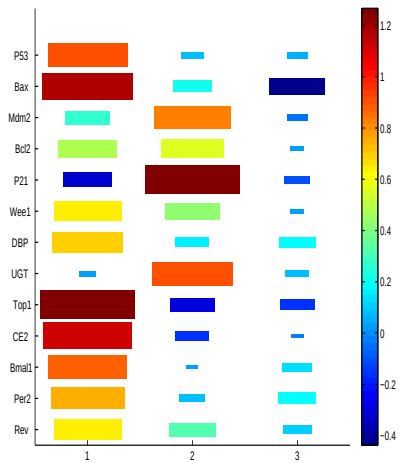
Class 2



Class 3



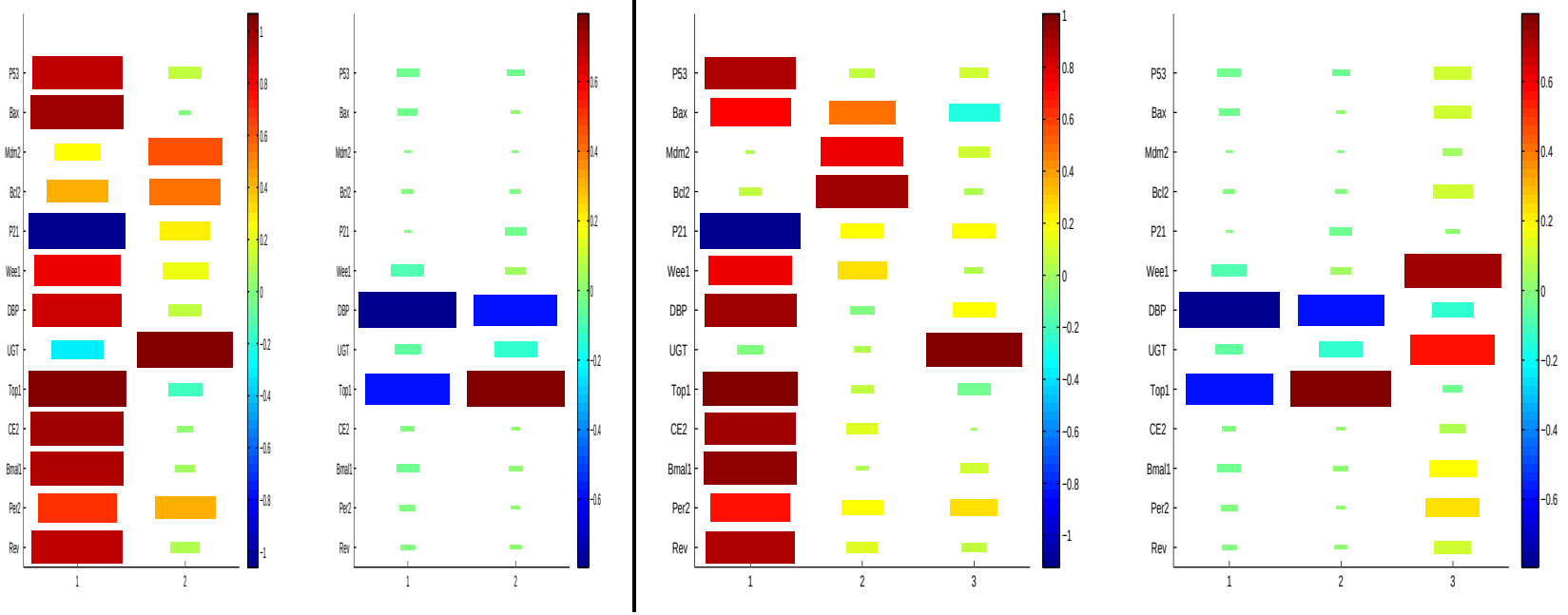
All Classes



Sparse PCA

- In classical PCA, FA and ICA, one looks to obtain principal (uncorrelated or independent) components.
- In Sparse PCA or FA, one looks for sparsest components. This leads to least variables selections.

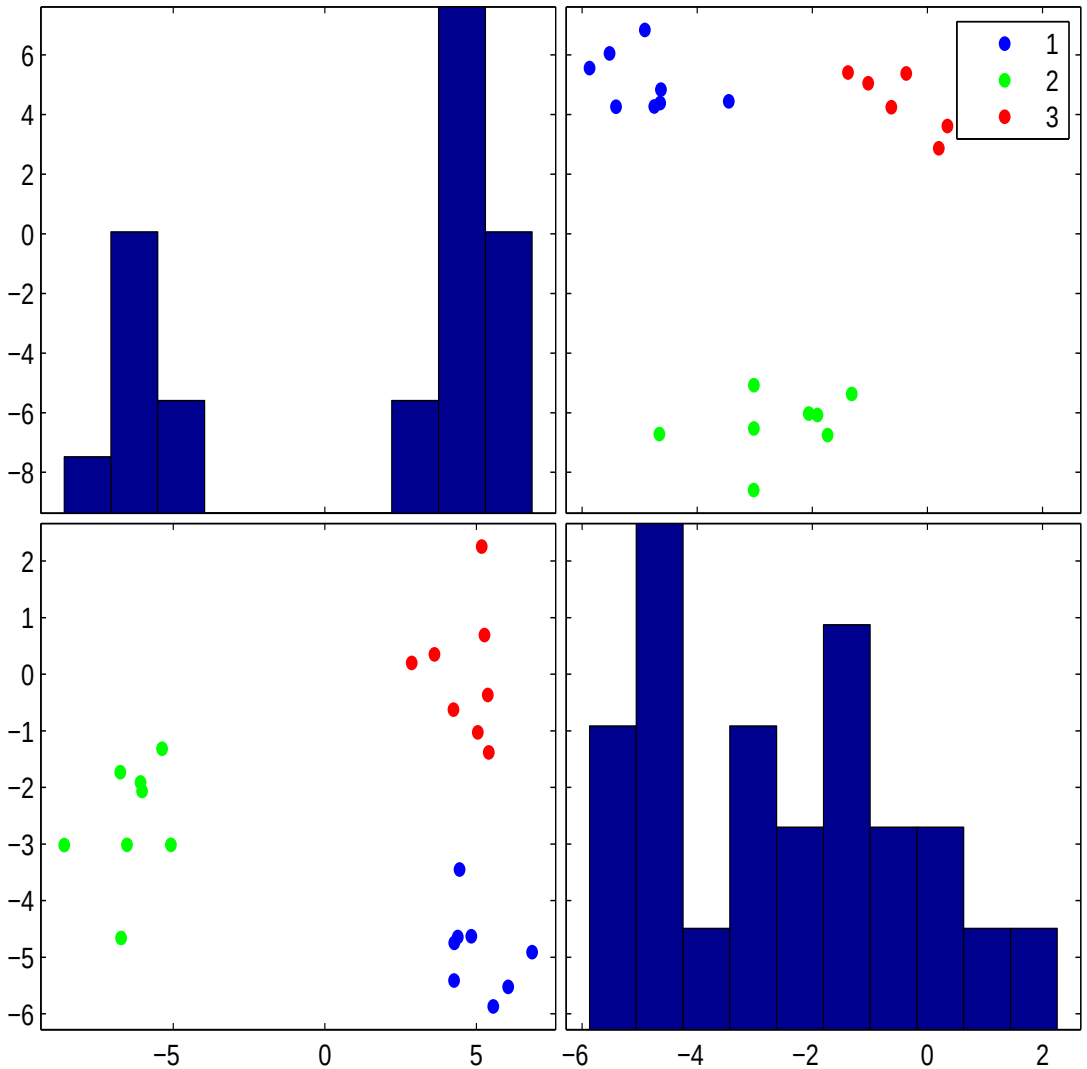
PCA SPCA PCA SPCA

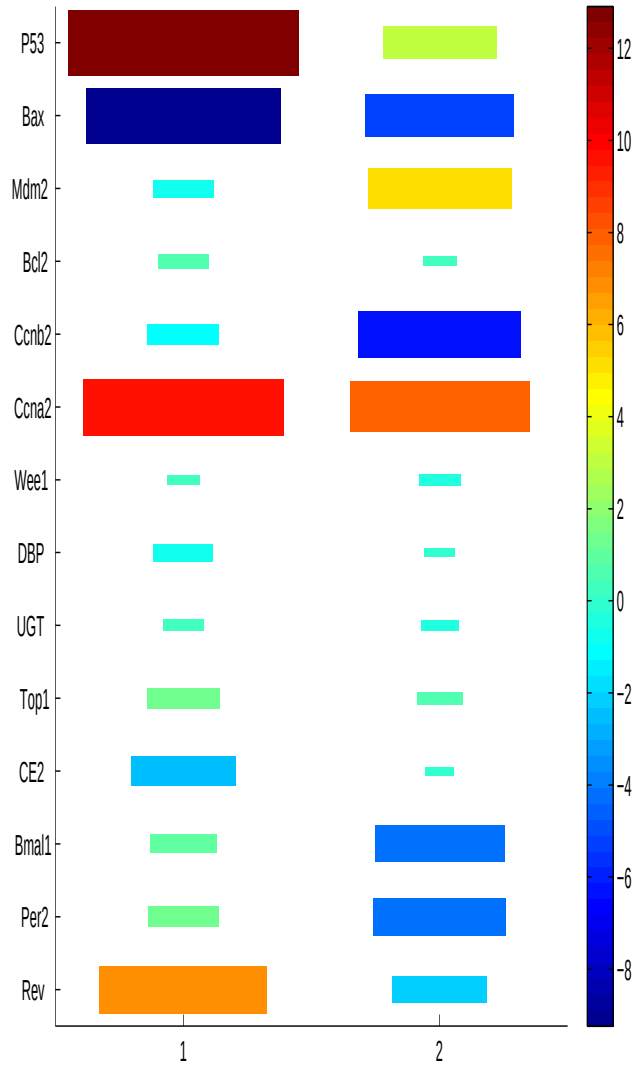


Discriminant Analysis

- When we have data and classes, the question to answer is:
  What are the most discriminant factors?
- There are many variants:
  - Linear Discriminant Analysis (LDA),
  - Quadratic Discriminant Analysis (QDA),
  - Exponential Discriminant Analysis (EDA),
  - Regularized LDA (RLDA), ...
- One can also ask for Sparsest Linear Discriminant factors (SLDA)
- Deterministic point of view (Geometrical distances)
- Probabilistic point of view (Mixture densities)
- Mixture of Gaussians models:
  Each classe is modelled by a Gaussian pdf

Discriminant Analysis: Time series, Colon

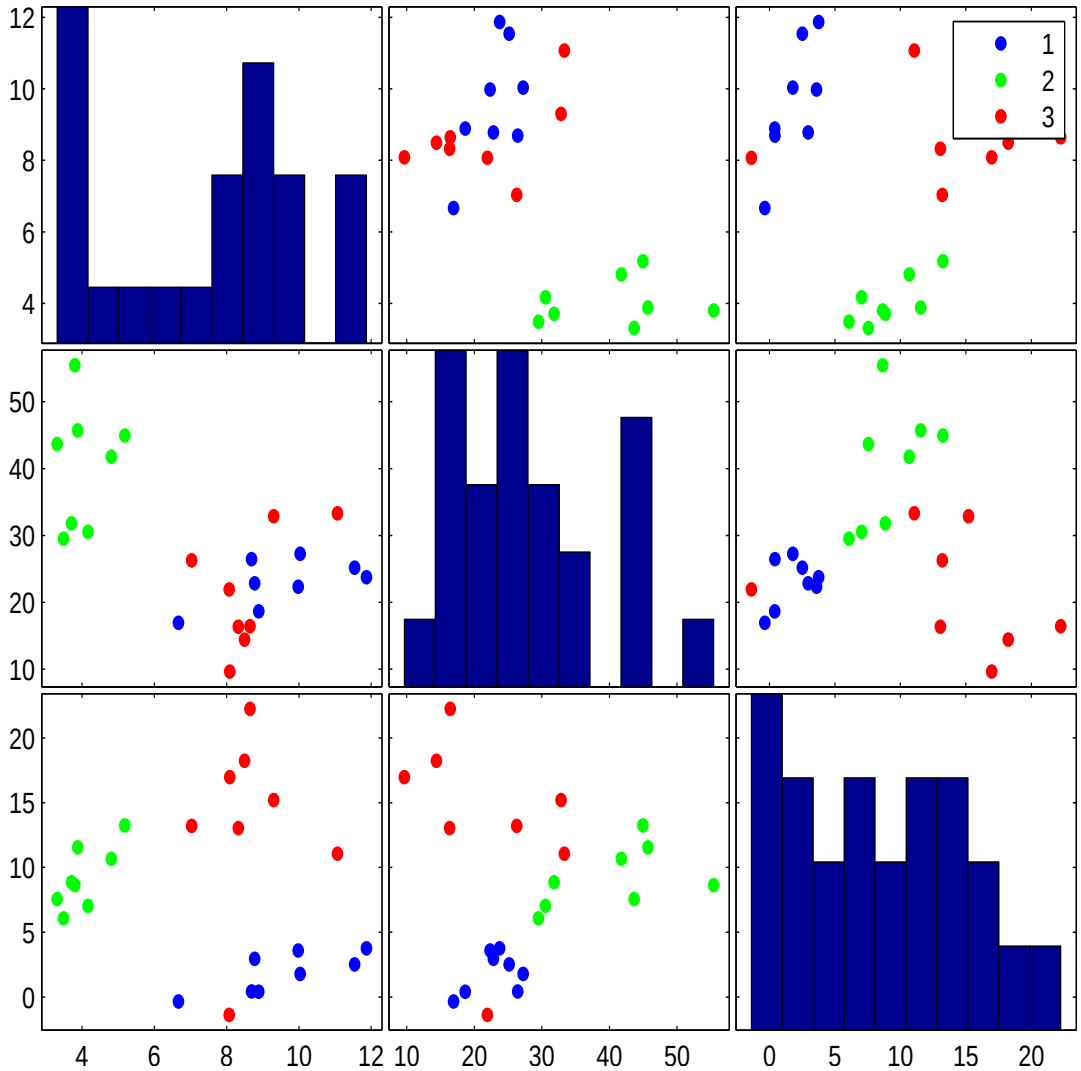


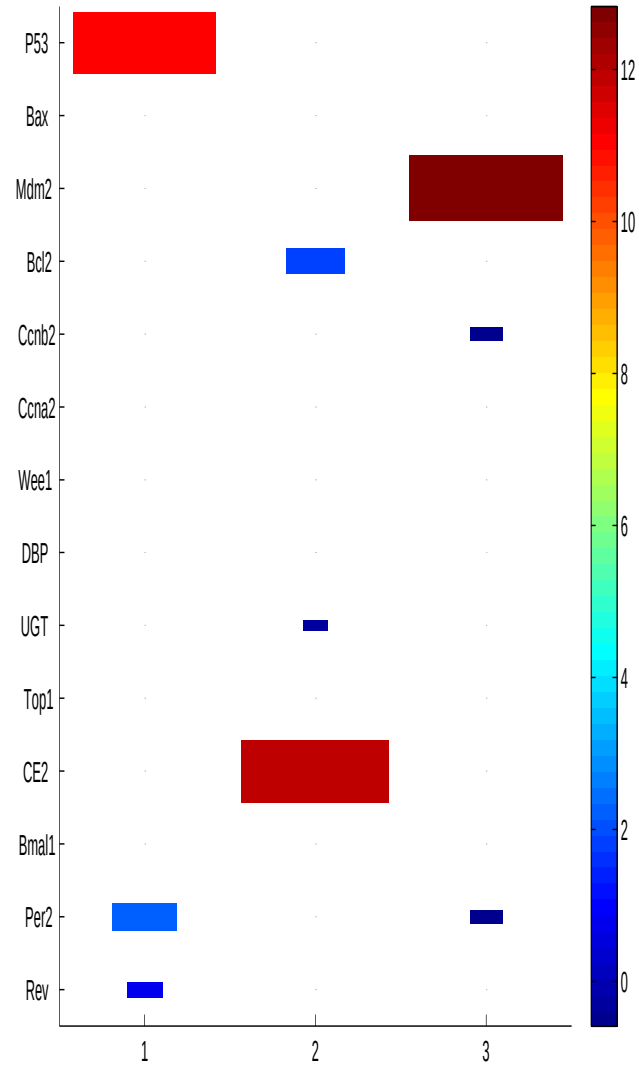


Sparse Discriminant Analysis

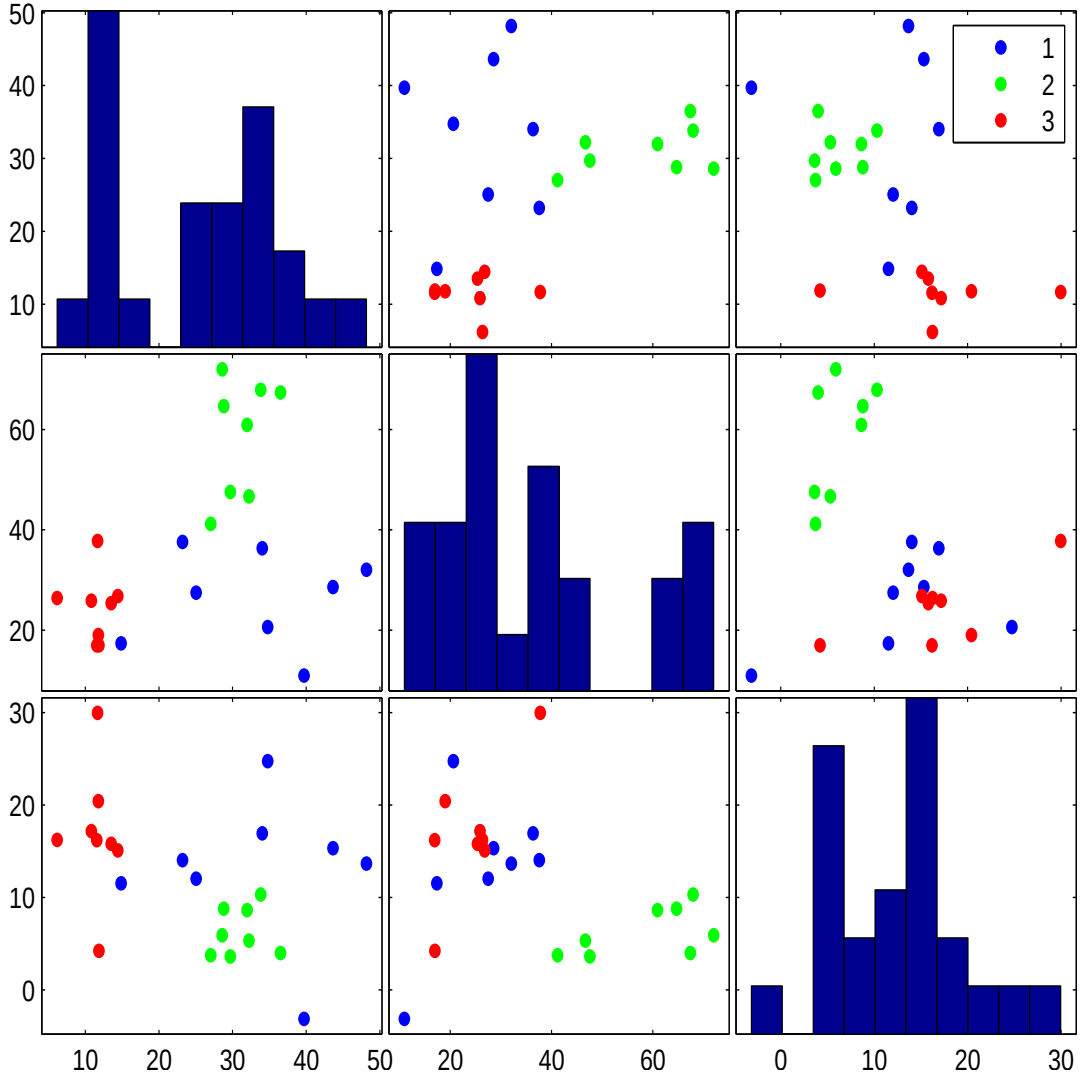
- The question to answer here is:
  What are the sparsest discriminant factors?

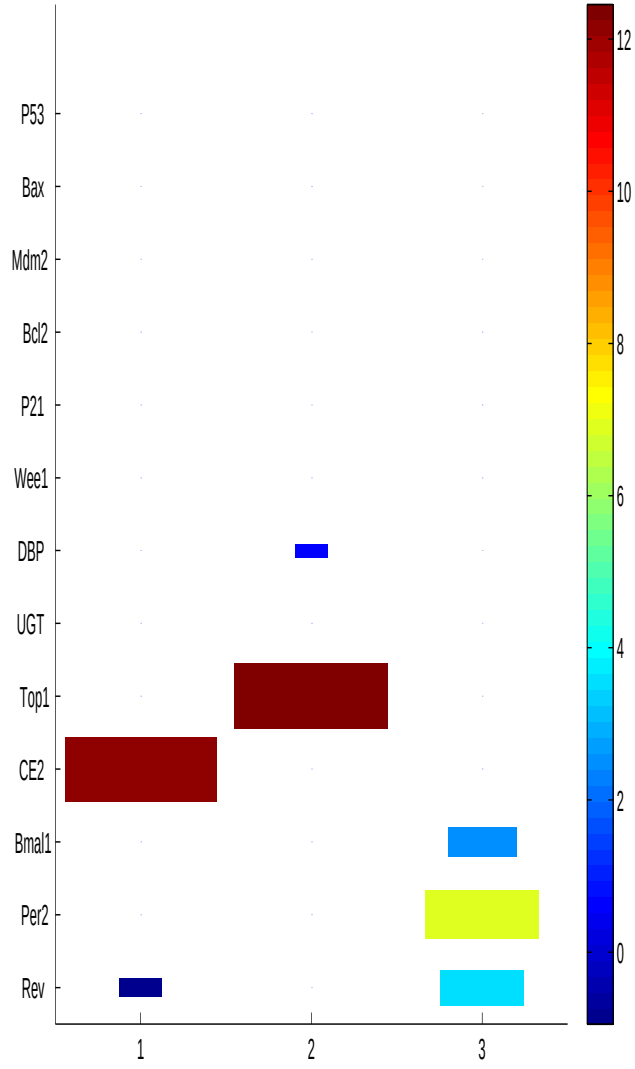
Sparse Discriminant Analysis: Time series, colon





Sparse Discriminant Analysis: Time series, Liver

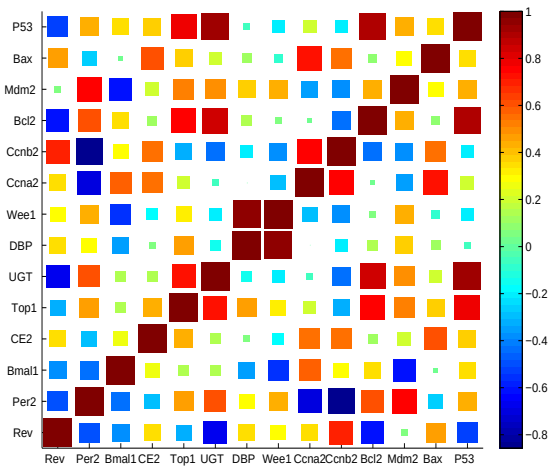




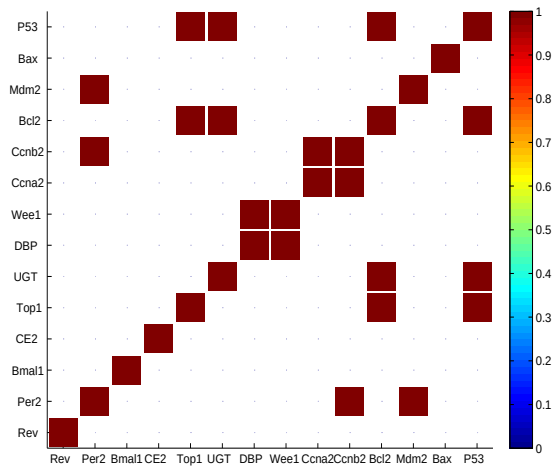
Dependancy graphs

- The main objective here is to show the dependencies between variables
- Three different measures can be used: Pearson ρ, Spearman ρ_S and Kendall τ
- In this study we used ρ_S
- A table of 2 by 2 mutual ρ_S are computed and used in different forms: Hinton, Adjacency table and Graphical network representation

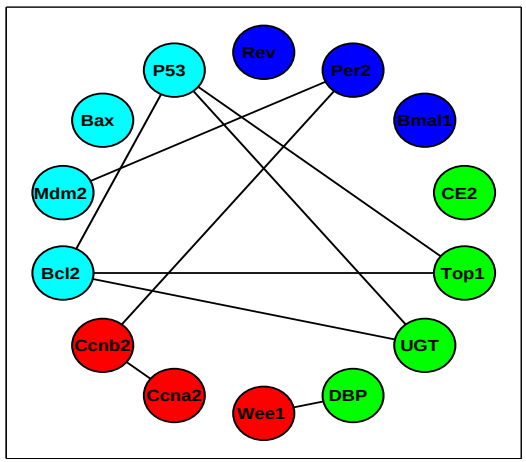
Hinton



Adjacency

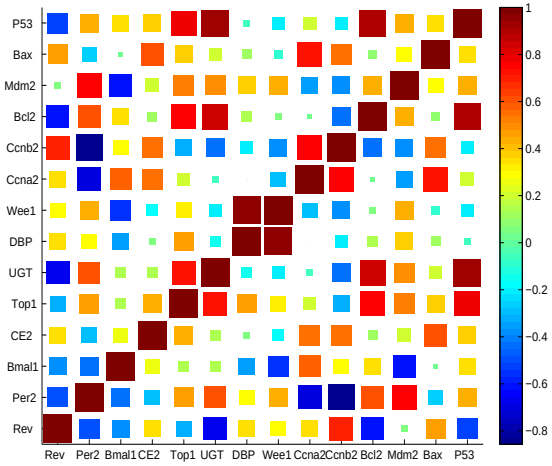


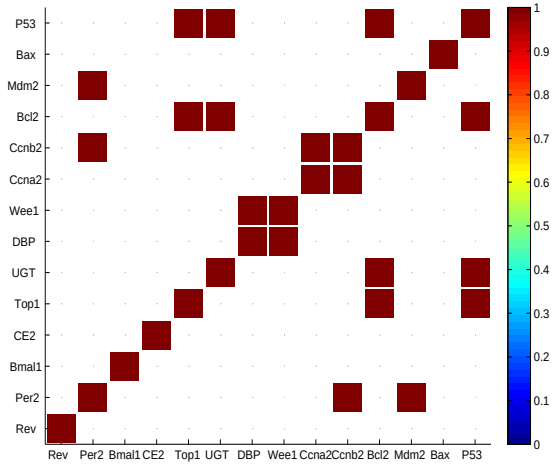
Network

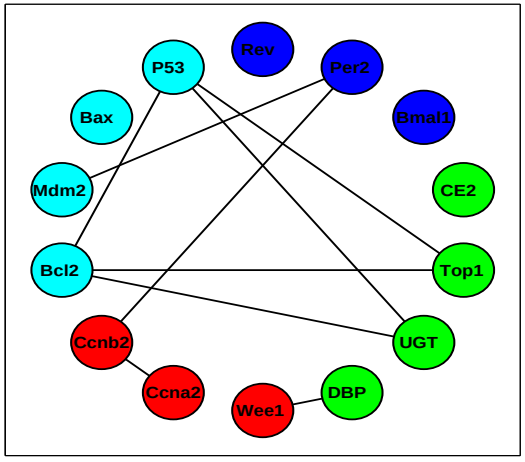


Graph of Dependancies: Colon, Class 1

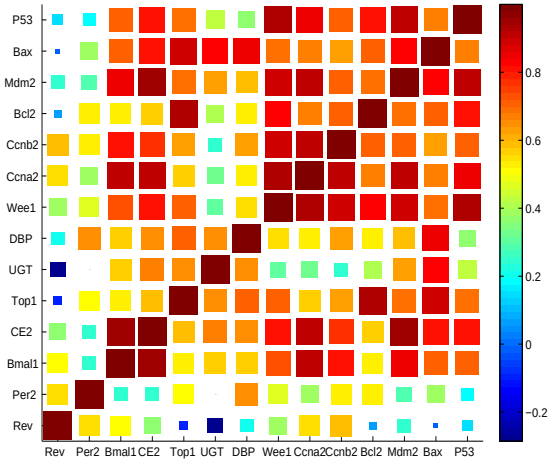
Time series

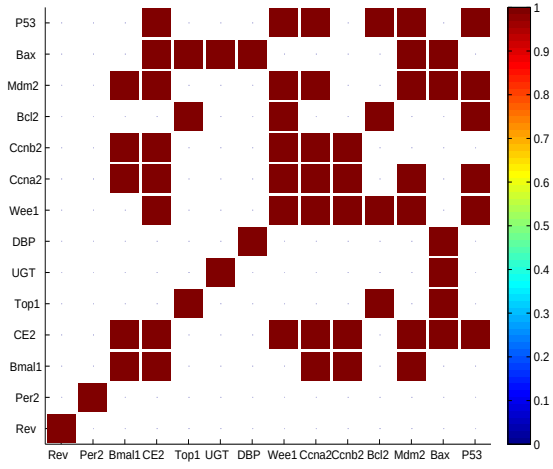


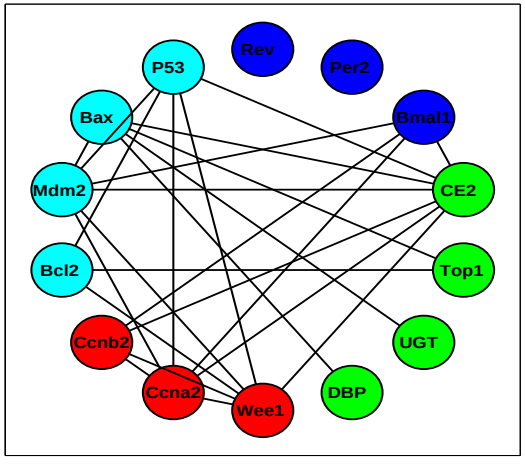




FT amplitudes

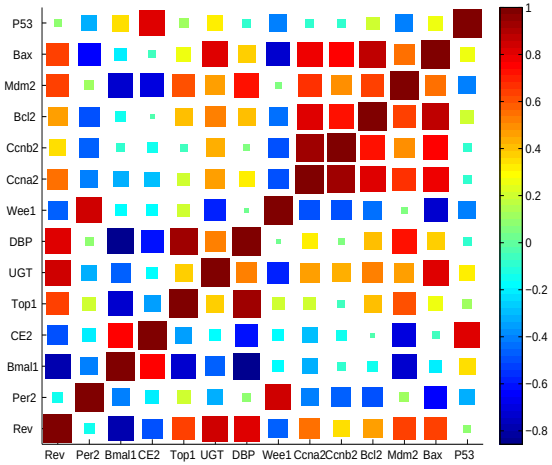


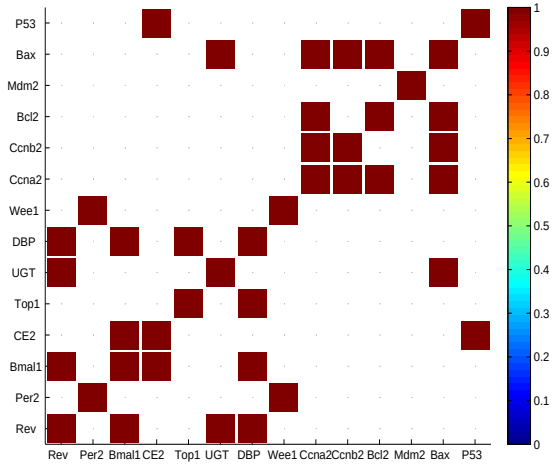


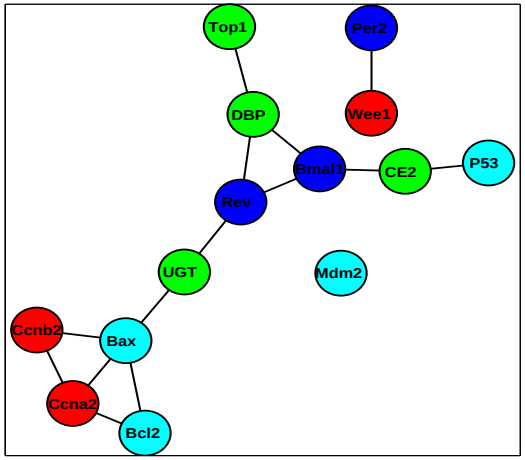


Graph of Dependancies: Colon, Class 3

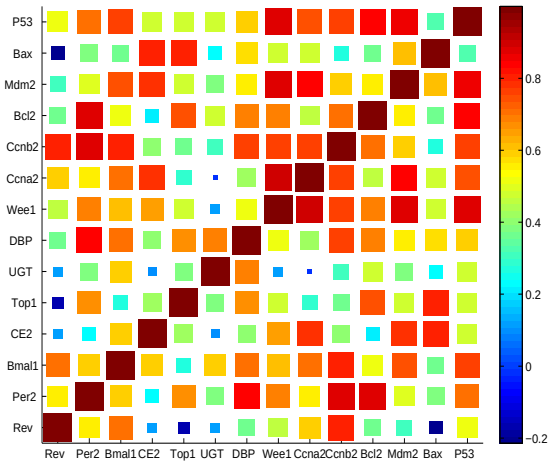
Time series

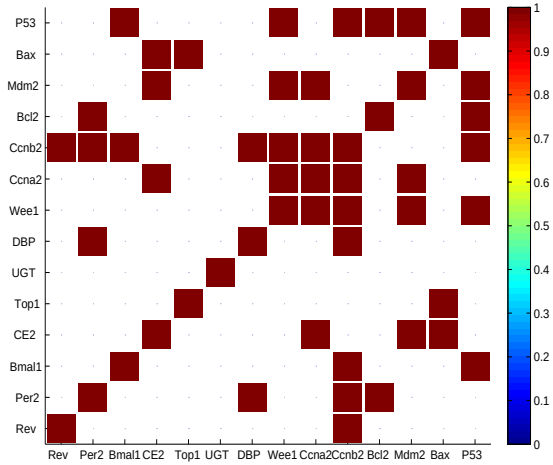


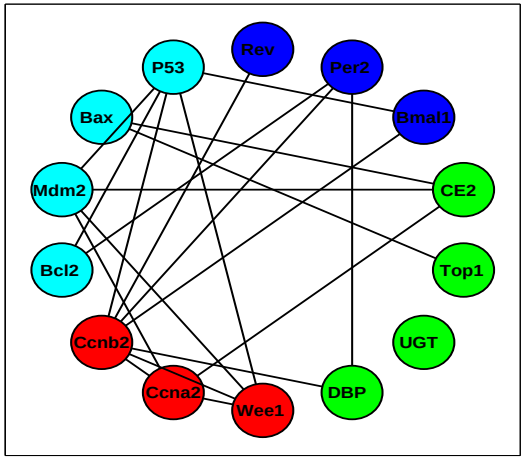




FT amplitudes







Future

- Application to C5Sys data
  - Some difficulties but plannification is done with Franck and Céline
- Modeling and model parameter estimation
  - With Jean and Frédérique, we are going to re-examine the estimation of the parameters of a Gamma pdf
- Inverse problems
  - With Jean, we are going to re-examine the estimation of the parameters of a Gamma pdf
- Input-Output modeling using training and test data
  - With Mircea, Xiaome and Francis, we processed some data relating two genes expressions and toxicity
- Causalities
  - Theoretical studies

Application to C5Sys data

- Classification of data
  - outputs of the cell cycle tracking: 3 curves per cell
  - First classification based only on clock activity
  - More general classification base on all variables
  - When classification is done, then we can study the relation between CC and clock
  - Discrimination parameters between classes
- Analysing data before and after some clocks knockdown
- We are applying these techniques on temperature-activity data before, during and after some treatment for Chronotherapy

Publications in relation with C5Sys

- J. Lapuyade-Lahorgue and A. Mohammad-Djafari,
  Nearest neighbors and correlation dimension for dimensionality estimation. Application to factor analysis of real biological time series data,
  in ESANN 2011 Proceedings, Gent, Belgium.
  ISBN 978-2-87419-044-5
- A. Mohammad-Djafari, G. Khodabandelou and J. Lapuyade-Lahorgue,
  A Matlab toolbox for data reduction, visualization, classification and knowledge extraction of complex biological data,
  BIOCOMP2011, Las Vegas, USA
- A. Mohammad-Djafari,
  Bayesian approach with prior models which enforce sparsity in signal and image processing,
  Review paper accepted for publication in European Association for Signal, Speech, and Image Processing (EURASIP) special issue on Sparsity in signal and image processing.