Multicomponent Multivariate data and signal analysis and processing in Cancer Biology

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

ISSSMA, June 3-4, 2013, IHP, Paris, France

June 5, 2013

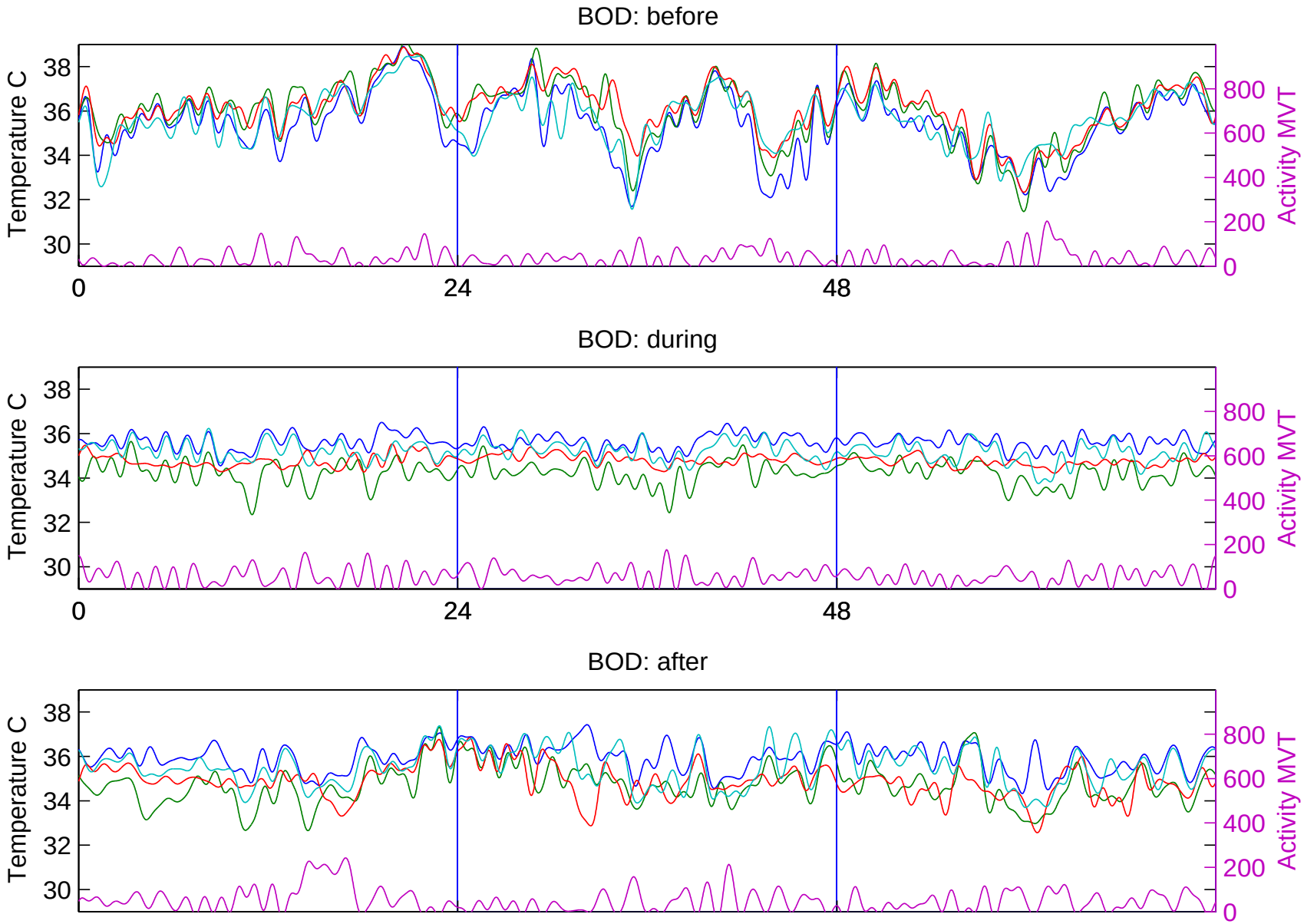
Summary

- Different experiences
  - Individual cells, Population of cells, Small animals, Human
  - In vitro and In Vivo
- A great number of data, variables, time series, signals, images, ... $\longrightarrow$ multivariate and multicomponent
- Need for Visualization tools
  - Time domain
  - Transformed domain: Fourier, Wavelets, Time-Frequency...
  - Scatter plots, histograms, statistics, ...
- Modeling time series
  - Parametric:
    Superposition of sinusoids, Gaussians shapes, ...
  - Non Parametric:
    Fourier, Wavelets, Time-frequency, time-scale,...
  - Probabilistic:
    Moving Average (MA), Autoregressive (AR), ARMA, Markovian models, ...
- Modeling the relation between data/signals
  - Linear / Non linear
  - Training and test data

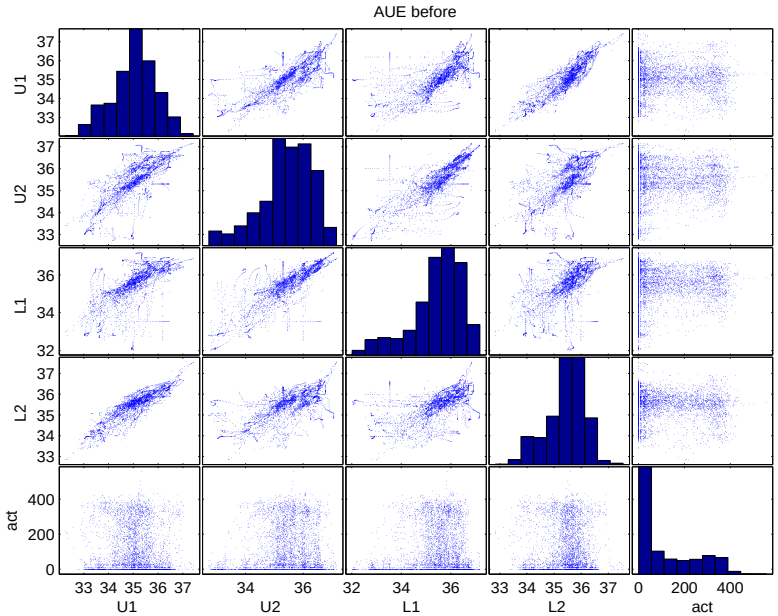
Summary

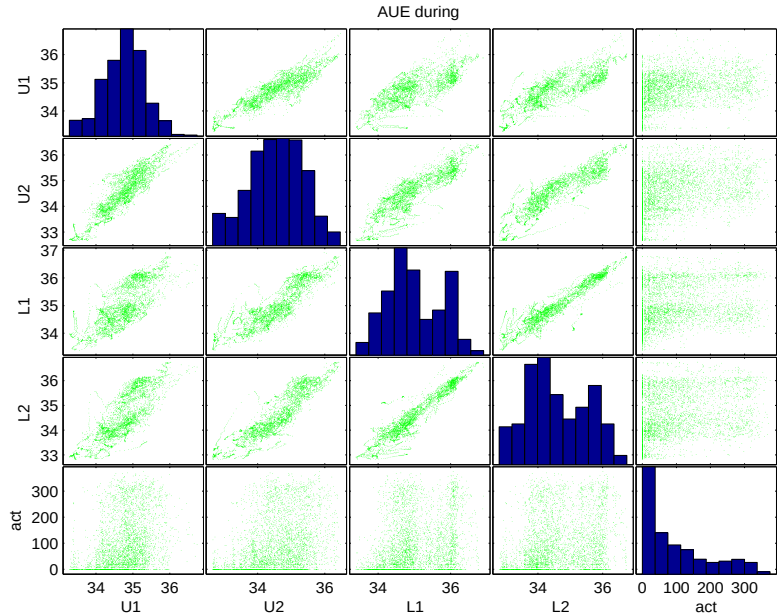
- Simple Analysis:
  Estimating periods, Computing harmonic components, spectra, , ...
- Multicomponent/Multivariate data analysis:
  Dimensional Reduction
  PCA, FA, ICA, Sparse PCA for dimensional reduction and main factors extraction
- Multicomponent/Multivariate Discriminant Analysis with classification:
  LDA, EDA, RDA, Sparse LDA for finding the most discriminant factors
- Blind sources separation
- Correlation (Pearson or Spearman) computation and dependency graph visualization
- Modelling input-output relations

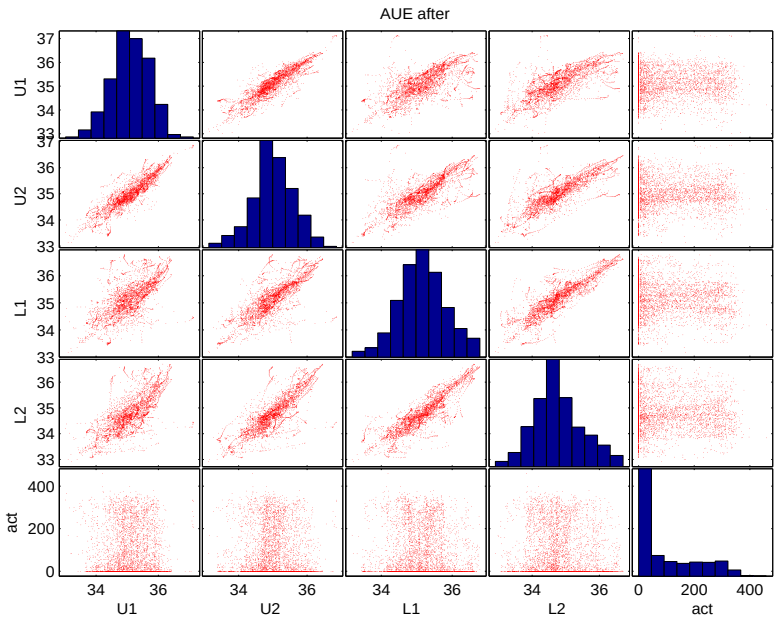
Temperature and activity Time series before, during and after some treatment

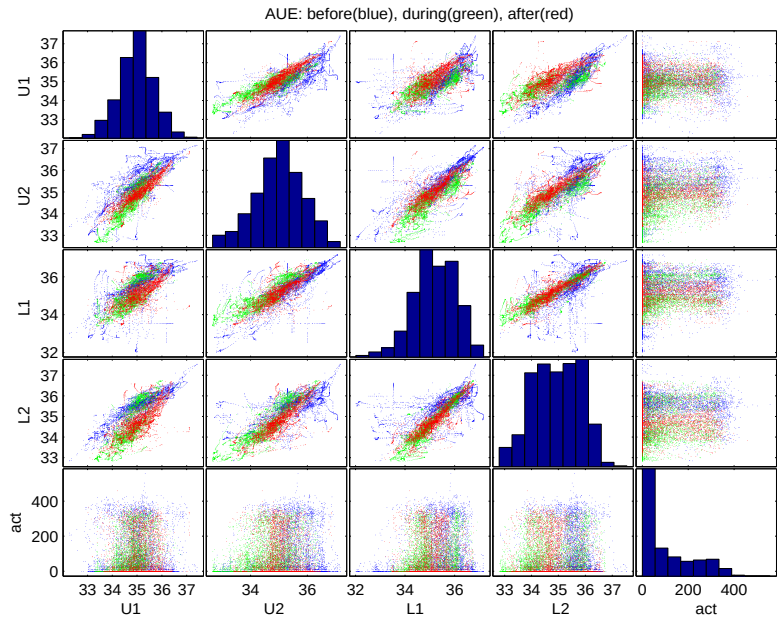


Temperatures, before, during and after changes





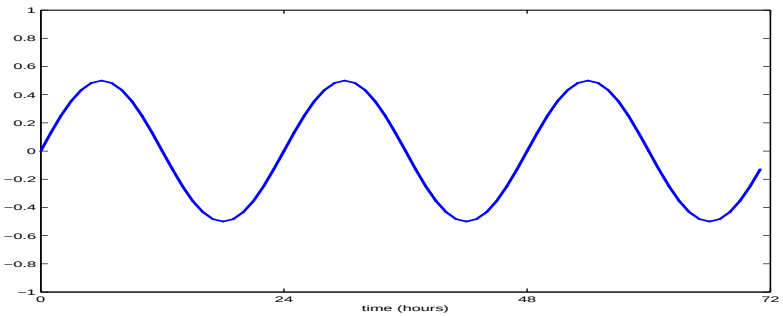


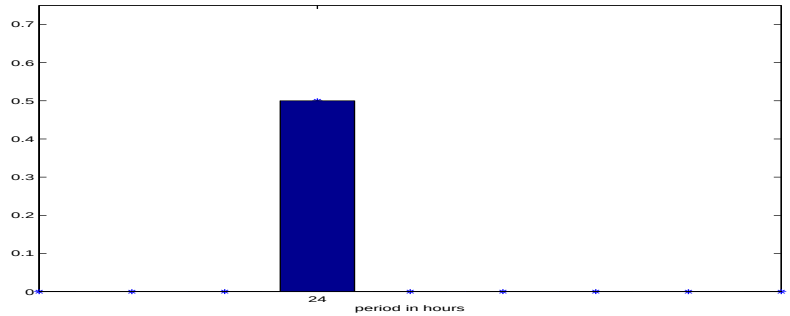


Simple Analysis tools: Period estimation.

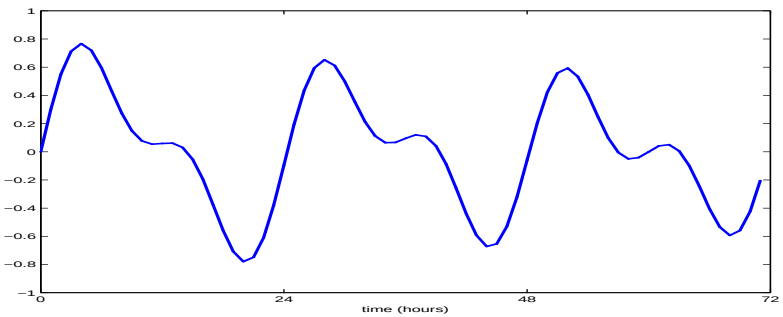
What do we mean by period ?

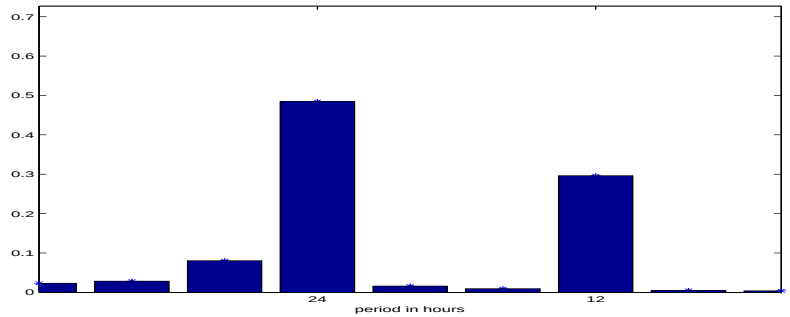
Case of 3 sinusoids



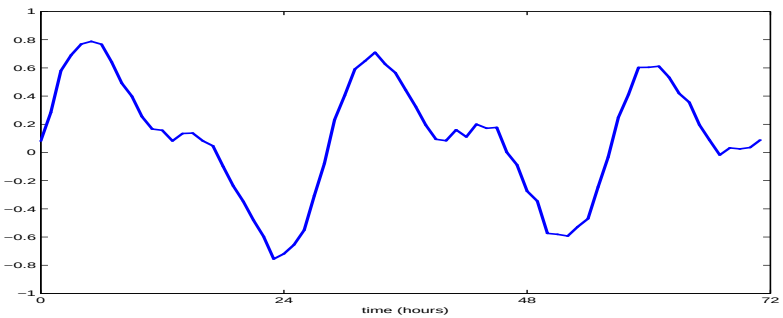


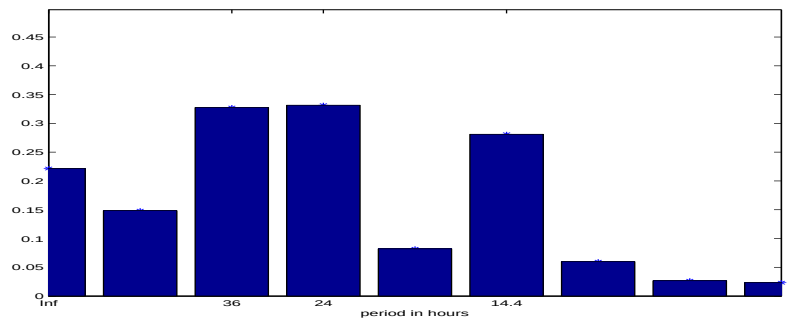
Case of 3 sinusoids+noise





Case of few sinusoids+noise

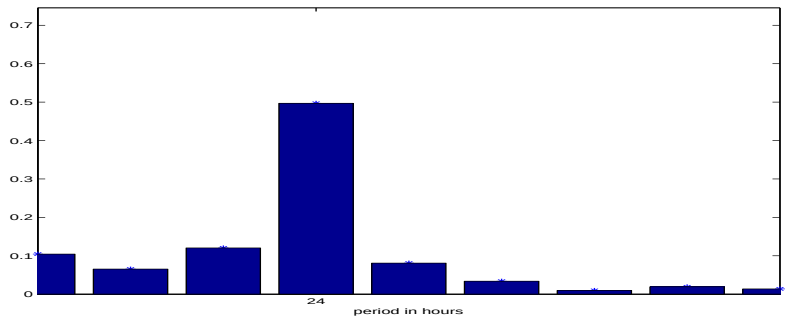


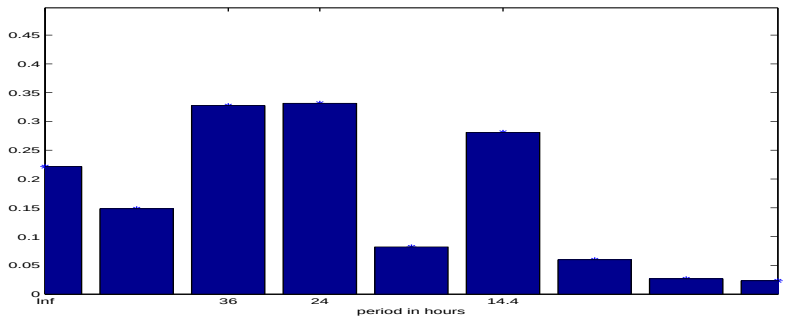


How to define a period from Spectra $S(\omega)$?

- Consider $S(\omega)$ as a distribution

  (normailise such that $\int S(\omega)\, \mathrm{d}\omega = 1$)





- principal harmonic: $\omega_{\text{mod}} = \arg\max_{\omega} \{S(\omega)\}$
- mean harmonic: $\omega_{\text{mean}} = \int \omega\, S(\omega)\, \mathrm{d}\omega$
- Lower and upper limits: $\omega_{\text{L}},\ \omega_{\text{U}}$

- principal period: $p_{\text{mod}} = \arg\max_{p} \{S(p)\}$
- mean period: $p_{\text{mean}} = \int p\, S(p)\, \mathrm{d}p$
- Lower and upper limits: $p_{\text{L}},\ p_{\text{U}}$

How to estimate Spectra $S(\omega)$?

- Fast Fourier Transform (FFT):
  $g(t) \longrightarrow FFT \longrightarrow f(\omega) \longrightarrow S(\omega) = |f(\omega)|^2$
  - Advantages: Well-known and understood, fast
  - Drawbacks: linear in frequencies ν,
    but not equidistance in periods
    $\nu = [0, \cdots, N-1] \longrightarrow p = [\infty, 1, \cdots, 1/(N-1)]$
- Autocorrelation function: $\gamma(\tau)$
  - If $g(t)$ is periodic, then $\gamma(\tau)$ is also periodic,
    but much smoother
  - $\gamma(0) = 1\ \gamma(\tau) \leq \gamma(0), \forall \tau$
- Autocorrelation function and FT: $\gamma(\tau) \longrightarrow$ FFT $S(\omega)$
- Inverse problem approach:
  Compute spectra for a given interval of possible periods)
  - $f(p) \longrightarrow g(t)$ is a linear forward operation

$$g(t) = \sum_n f(2\pi/p_n) \exp[j2\pi nt/p_n]$$

  - $\boldsymbol{g = Hf + \epsilon \longrightarrow \widehat{f}}$

How to estimate Spectra? Inverse Problem Approach

$$\boldsymbol{g} = \boldsymbol{H}\boldsymbol{f} + \boldsymbol{\epsilon}$$

- Regularization:

$$\widehat{\boldsymbol{f}} = \arg\min_{\boldsymbol{f}} \left\{ \|\boldsymbol{g} - \boldsymbol{H}\boldsymbol{f}\|^2 + \lambda \|\boldsymbol{f}\|^2 \right\} \rightarrow \widehat{\boldsymbol{f}} = (\boldsymbol{H}'\boldsymbol{H} + \lambda \boldsymbol{I})^{-1} \boldsymbol{H}'\boldsymbol{g}$$

- Bayesian approach:
  - Assign the Likelihood : $p(\boldsymbol{g}|\boldsymbol{f})$
  - Assign the prior law: $p(\boldsymbol{f})$
  - Use the Bayes rule : $p(\boldsymbol{f}|\boldsymbol{g}) \propto p(\boldsymbol{g}|\boldsymbol{f})\, p(\boldsymbol{f})$
  - Use this posterior law to infer on $\boldsymbol{f}$.
  - For example MAP:

$$\widehat{\boldsymbol{f}} = \arg\max_{\boldsymbol{f}} \{p(\boldsymbol{f}|\boldsymbol{g})\} = \arg\min_{\boldsymbol{f}} \{J(\boldsymbol{f})\}$$

    but there are other possibilities: Posterior mean, median, ...
  - Assuming Gaussian noise and Gaussian prior

$$J(\boldsymbol{f}) = \|\boldsymbol{g} - \boldsymbol{H}\boldsymbol{f}\|^2 + \lambda \|\boldsymbol{f}\|^2$$

  - Different priors (Gaussian, Generalized Gaussian, Cauchy,...) $J(\boldsymbol{f}) = \|\boldsymbol{g} - \boldsymbol{H}\boldsymbol{f}\|^2 + \lambda \Omega(\boldsymbol{f})$

Bayesian estimation with priors enforcing sparsity

- Sparsity: For any periodic signal, the spectrum is a set of Diracs
- Biological signals related to clocks: a few independent oscillators
- Spectrum has a few non zero elements in any given interval
- To translate this information use a heavy tailed prior law like Double exponential or Cauchy with its hierarchical structure and hidden variables

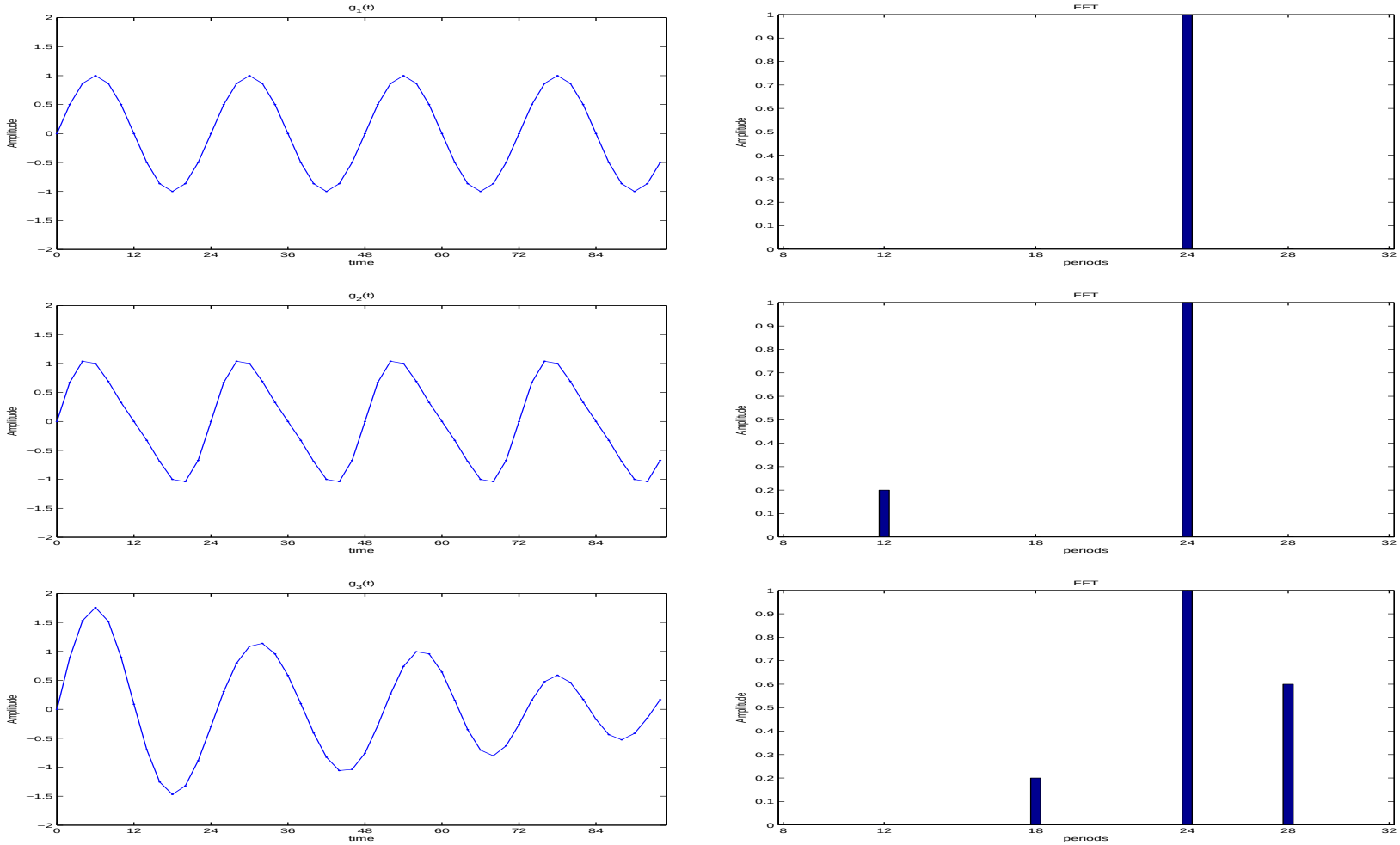
$$(\boldsymbol{f}|\nu) = \prod_j \mathcal{S}t(f_j|\nu) \text{ with } \mathcal{S}t(f_j|\nu) = \int_0^\infty \mathcal{N}(f_j|0,1/\tau_j)\,\mathcal{G}(\tau_j|\nu/2,\nu/2)\,\mathrm{d}\tau_j$$

- Hierarchical structure with hidden variables τ_j:

$$\begin{cases} p(f_j|\tau_j) &= \mathcal{N}(f_j|0,1/\tau_j) \propto \exp\left[-\frac{1}{2}\tau_j f_j^2\right] \\ p(\tau_j|\alpha,\beta) &= \mathcal{G}(\tau_j|\alpha,\beta) \propto \tau_j^{(\alpha-1)} \exp\left[-\beta\tau_j\right] \text{ with } \alpha=\beta=\nu/2 \end{cases}$$

- Posterior $p(\boldsymbol{f},\boldsymbol{\tau}|\boldsymbol{g}) \propto p(\boldsymbol{g}|\boldsymbol{f})\,p(\boldsymbol{f}|\boldsymbol{\tau})\,p(\boldsymbol{\tau})$

Multicomponent period estimation



$$g_k = H f_k + \epsilon_k$$

f_k have some common spectra.

Dimension reduction, PCA, Factor Analysis, ICA

- M variables $\boldsymbol{g}(t)$ are observed. They are redundant.
  Can we express them with $N \leq M$ factors $\boldsymbol{f}$?
  How many factors (Principal Components, Independent Components) can describe the observed data?

$$\begin{cases} g_i(t) = \sum_{j=1}^{N} a_{ij} f_j(t) + \epsilon_i(t) \\ \boldsymbol{g}(t) = \boldsymbol{A}\boldsymbol{f}(t) + \boldsymbol{\epsilon}(t) \end{cases} \quad \begin{cases} \boldsymbol{A} : (M \times N) \text{ Loading matrix }, N \leq M \\ \boldsymbol{f} \text{ (t) } : \quad \textit{factors}, \textit{sources} \end{cases}$$

- How to find both $\boldsymbol{A}$ and factors $\boldsymbol{f}(t)$?
- Deterministic methods:

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

- Bayesian methods:

$$(\widehat{\boldsymbol{A}}, \widehat{\boldsymbol{f}}) = \arg\max_{(\boldsymbol{A},\boldsymbol{f})} \{p(\boldsymbol{A}, \boldsymbol{f}|\boldsymbol{g})\} = \arg\min_{(\boldsymbol{A},\boldsymbol{f})} \{\ln p(\boldsymbol{g}|\boldsymbol{A}, \boldsymbol{f}) - \ln p(\boldsymbol{A}) - \ln p(\boldsymbol{f}$$

How to determine the number of factors

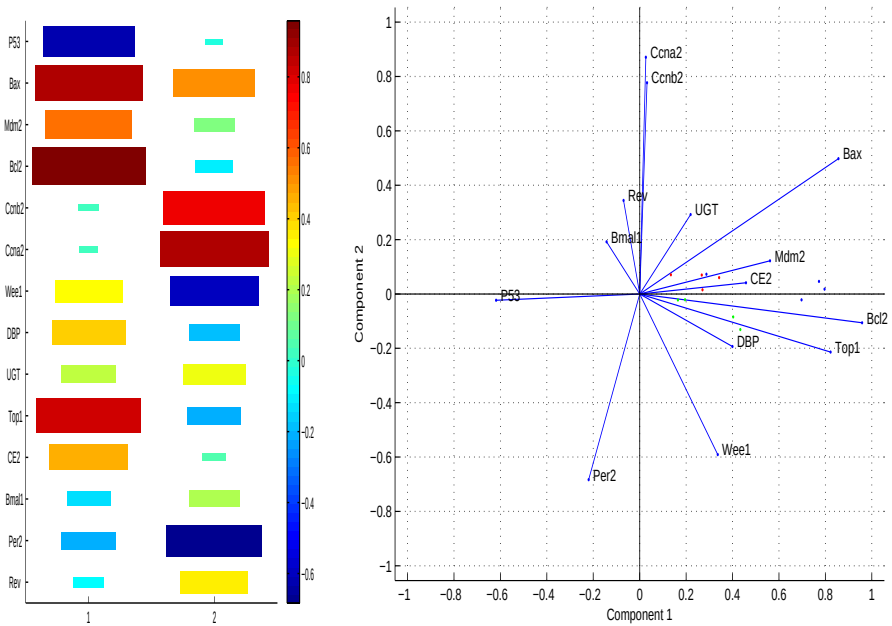
- When N is given:

$$p(\boldsymbol{A}, \boldsymbol{f}|\boldsymbol{g}) \propto p(\boldsymbol{g}|\boldsymbol{A}, \boldsymbol{f})\, p(\boldsymbol{A})\, p(\boldsymbol{f})$$

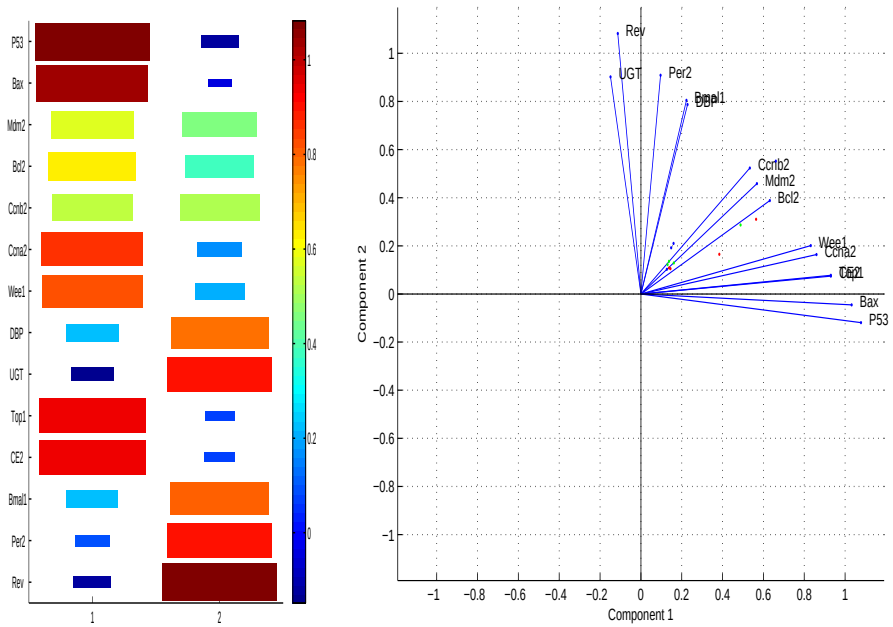
  Different choices for $p(\boldsymbol{A})$ and $p(\boldsymbol{f})$ and
  Different methods to estimate both $\boldsymbol{A}$ and $\boldsymbol{f}$:
  JMAP, EM, Variational Bayesian Approximation
- When N is not known:
  - Model selection
  - Bayesian or Maximum likelihood methods
  - To determine the number of factors we do the analyze 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.

Factor Analysis: 2 factors: Colon

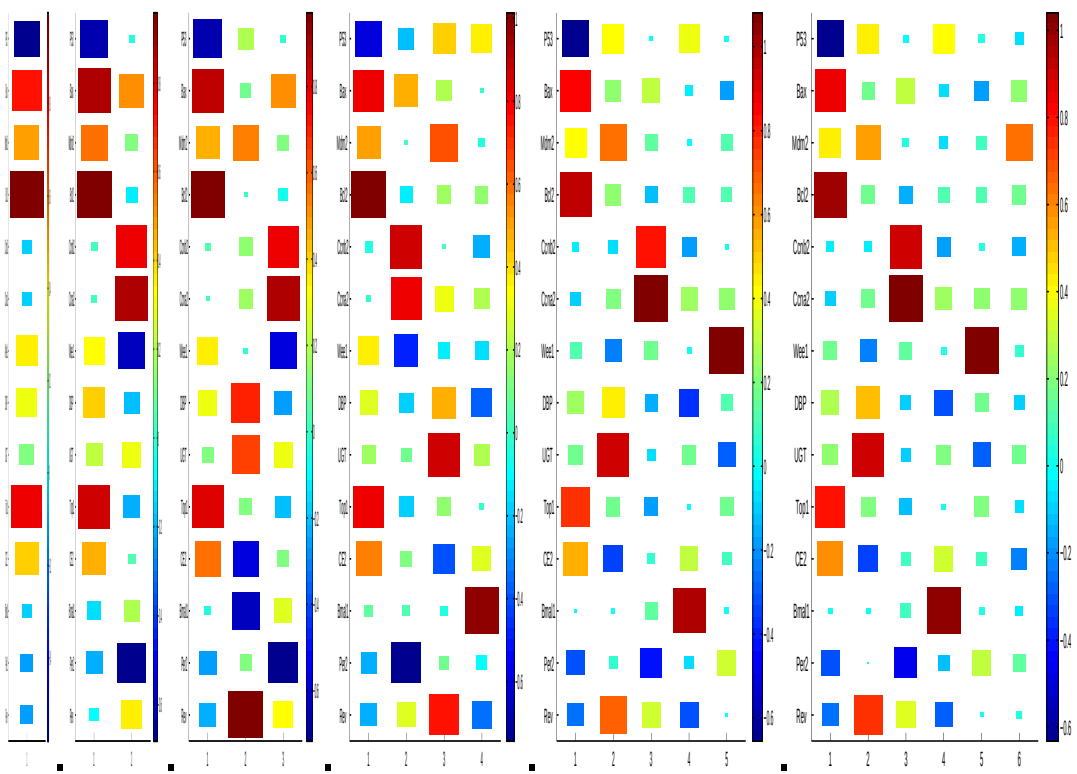
Time series

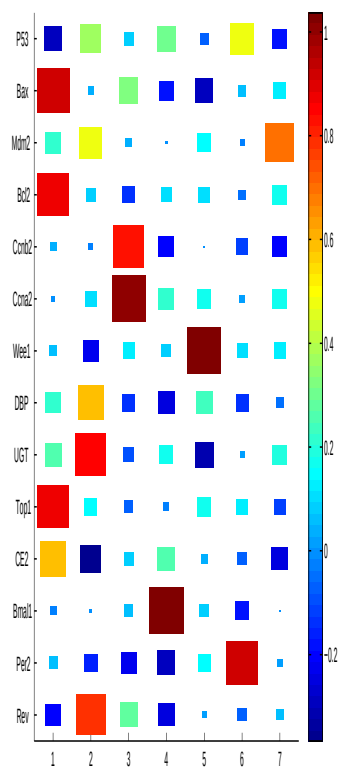


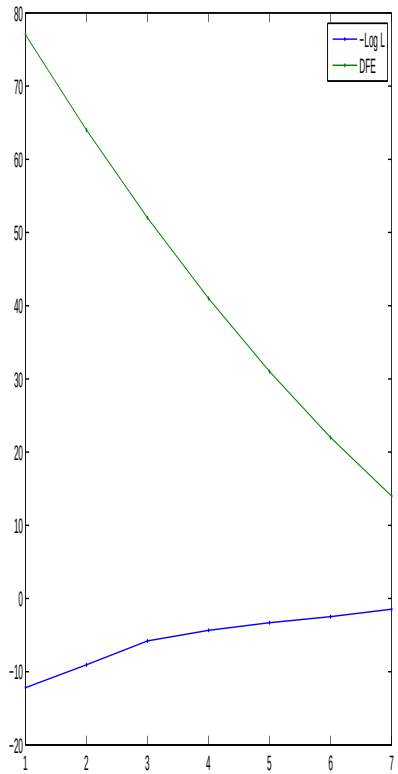
FT Amplitudes



Factor Analysis: Time series, colon

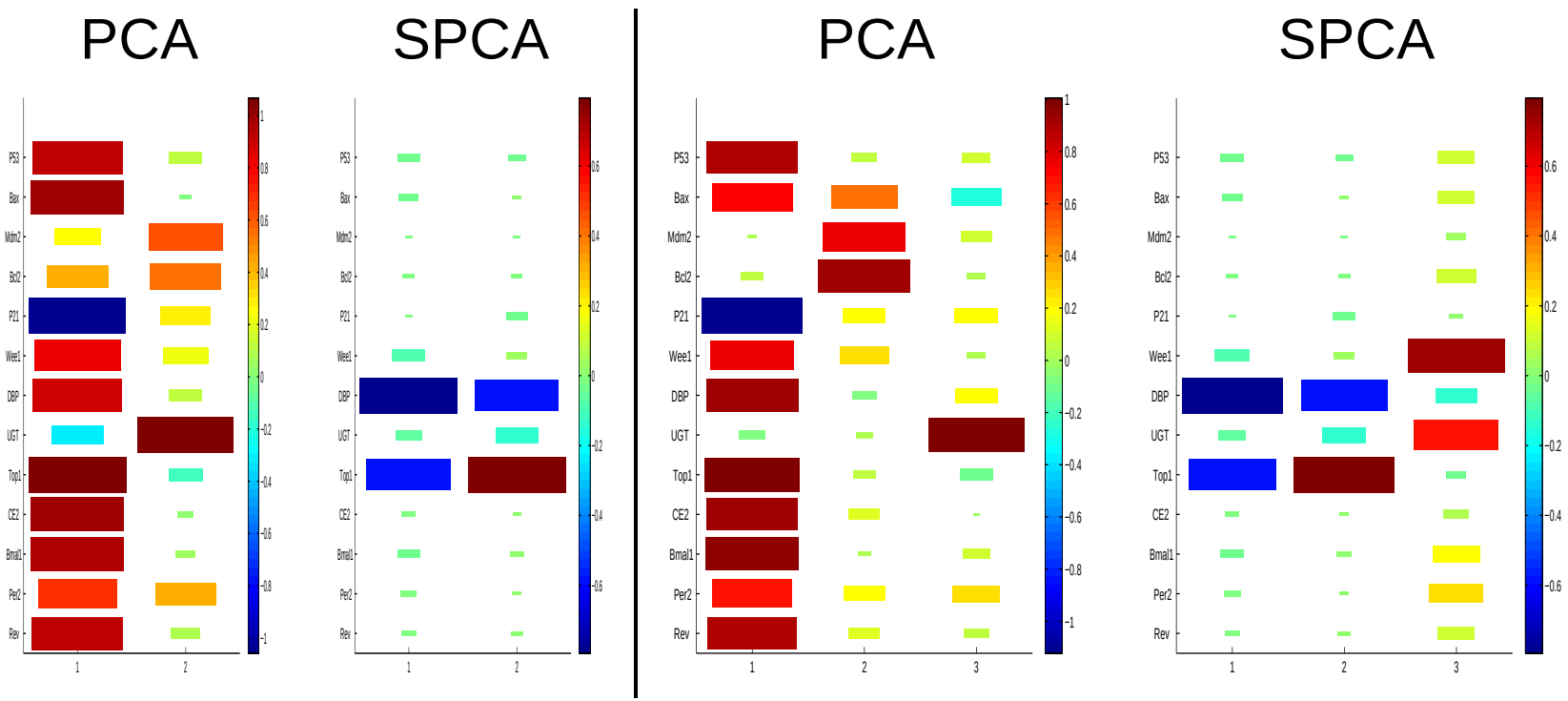






Sparse PCA

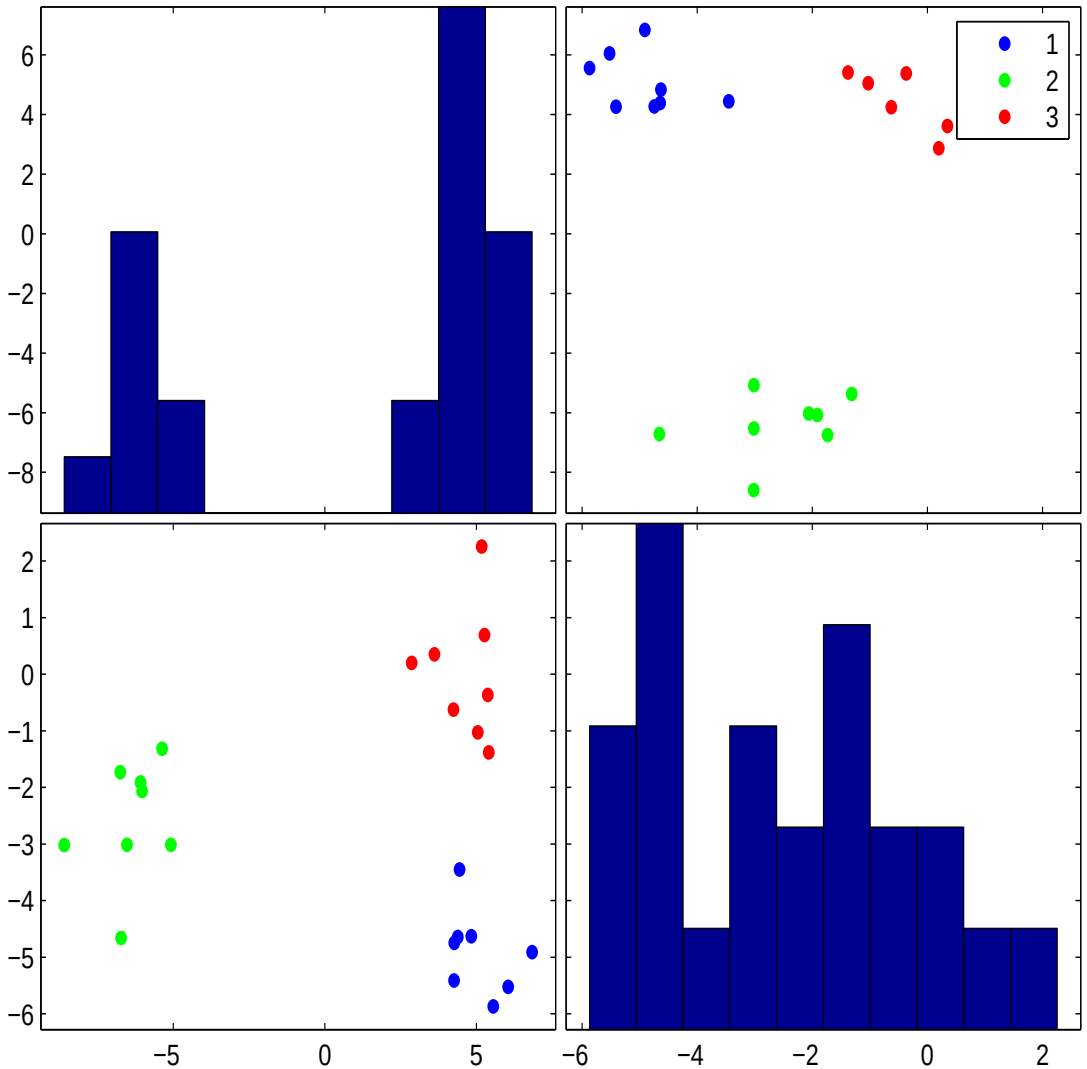
- In classical PCA, FA and ICA, one looks to obtain principal (uncorrelated or independent) components.
- In Sparse PCA or FA, one looks for the loading matrix A with sparsest components.
- This can be imposed via the prior $p(A)$. This leads to least variables selections.

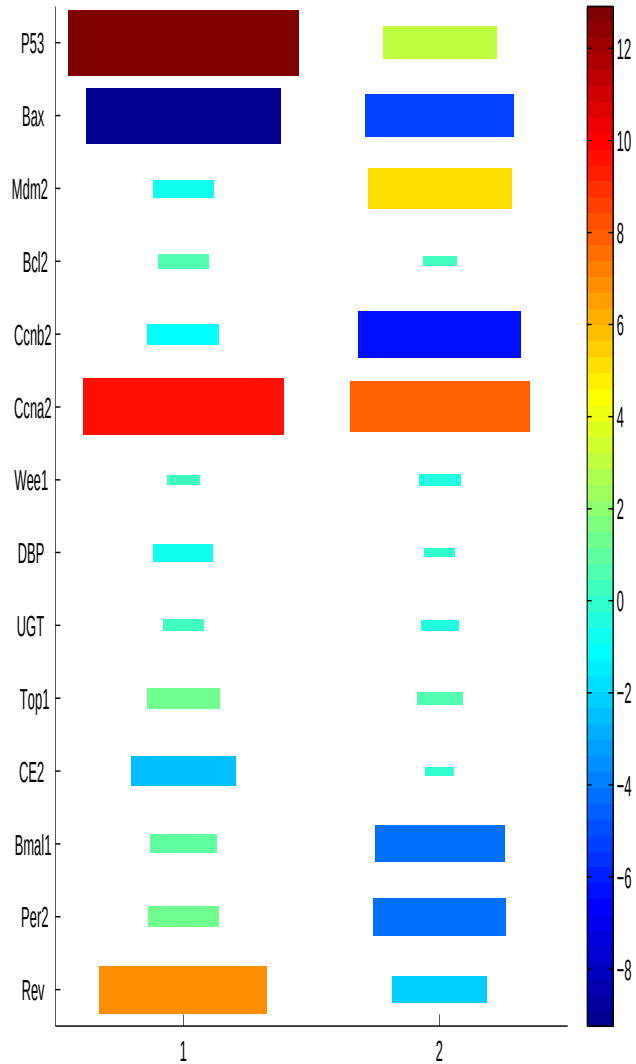


Discriminant Analysis

- 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

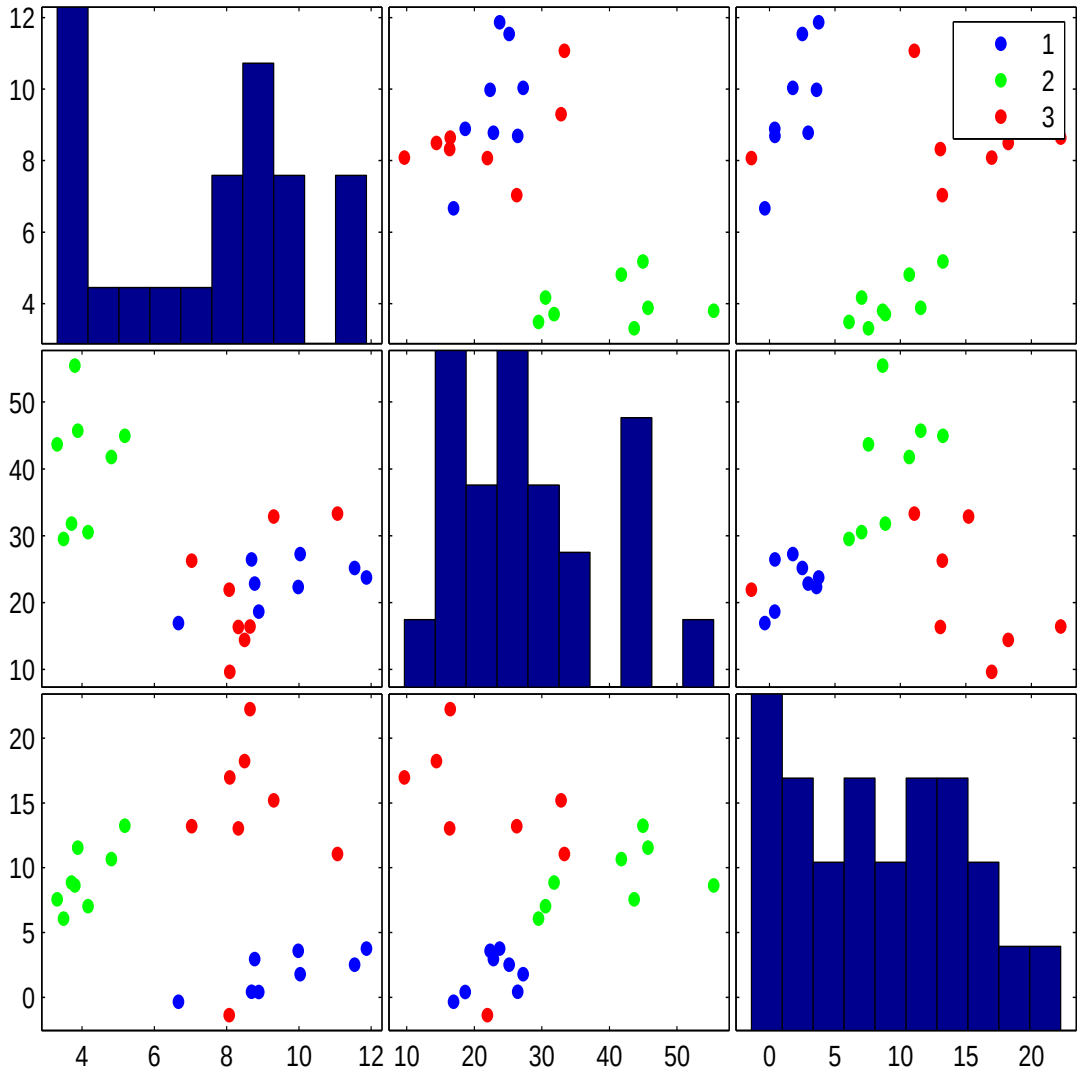
Discriminant Analysis: Time series, Colon

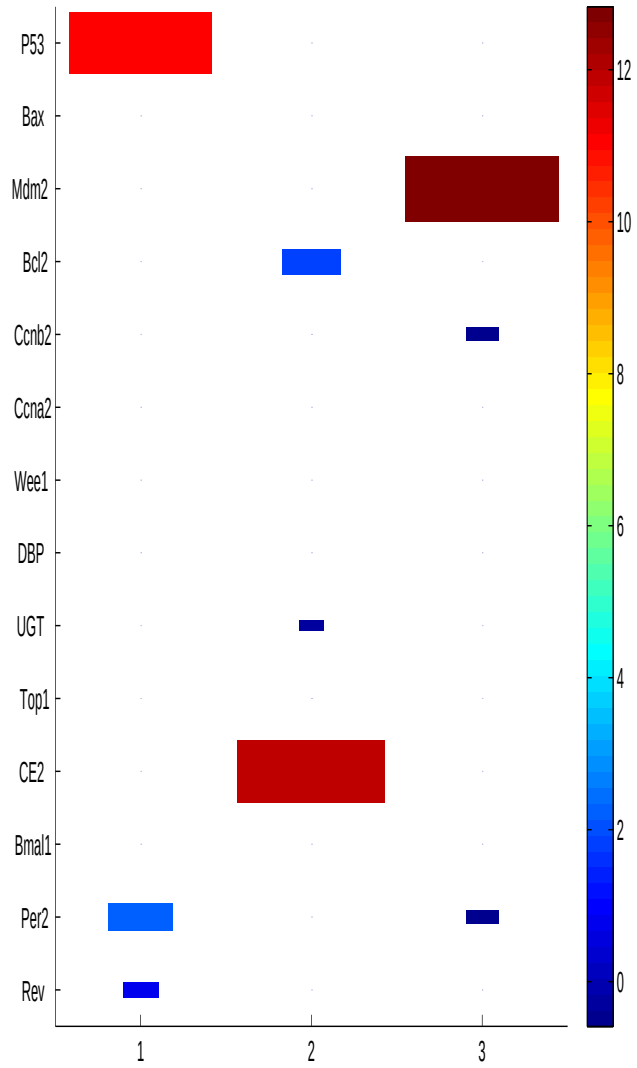




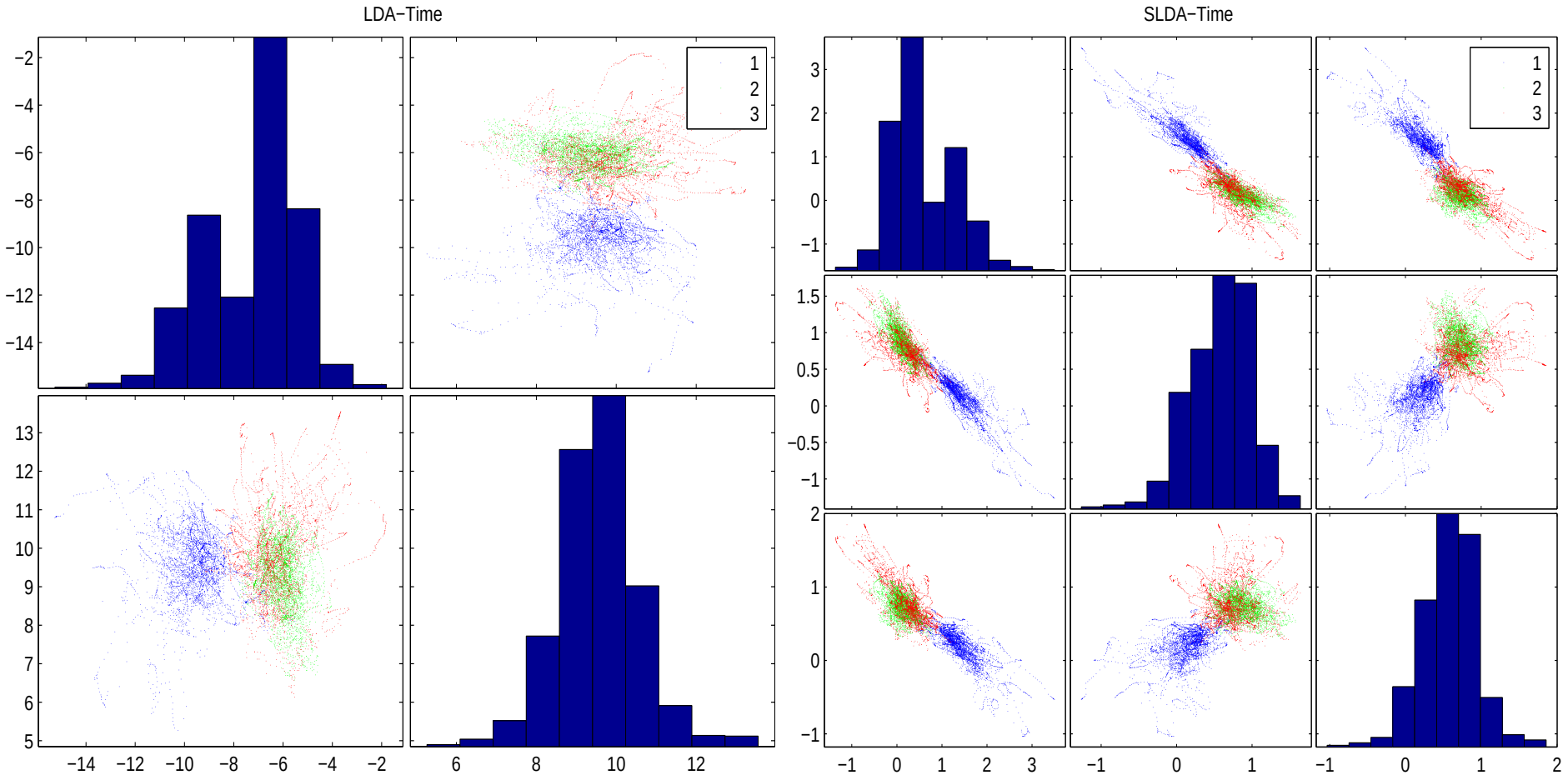
Sparse Discriminant Analysis: Time series, colon

What are the sparsest discriminant factors?





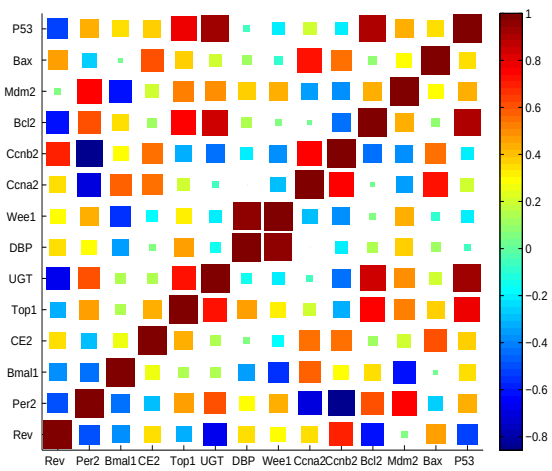
LDA and SLDA study on time serie: 1:before, 2:during, 3:after



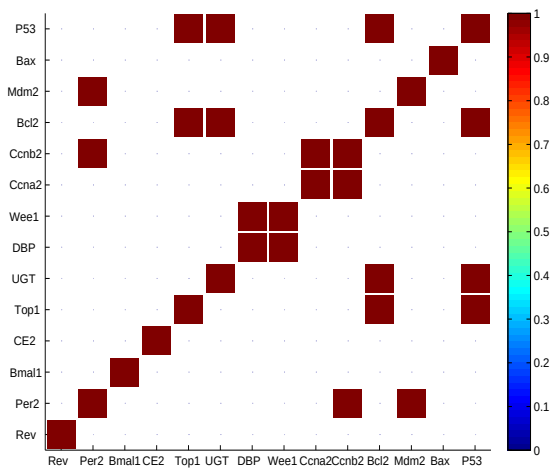
Dependancy graphs

- 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

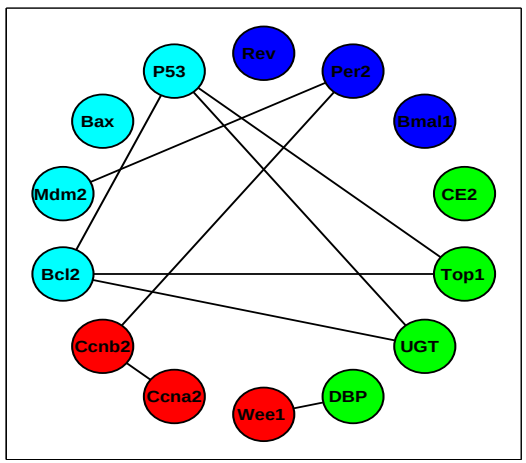
Hinton



Adjacency

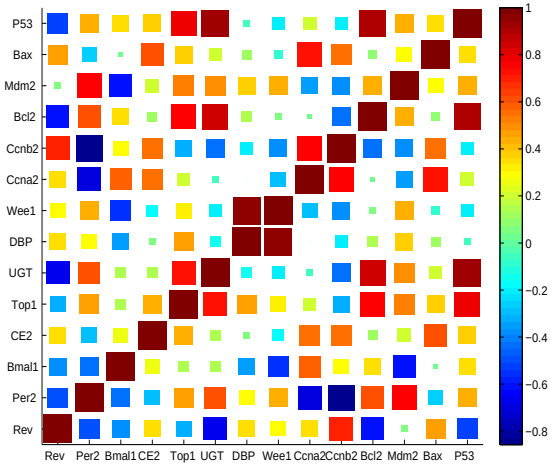


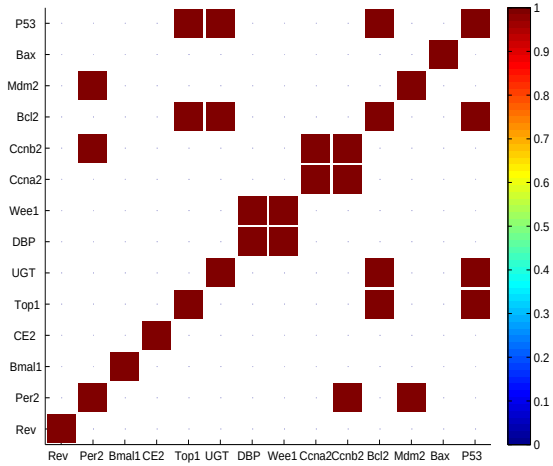
Network

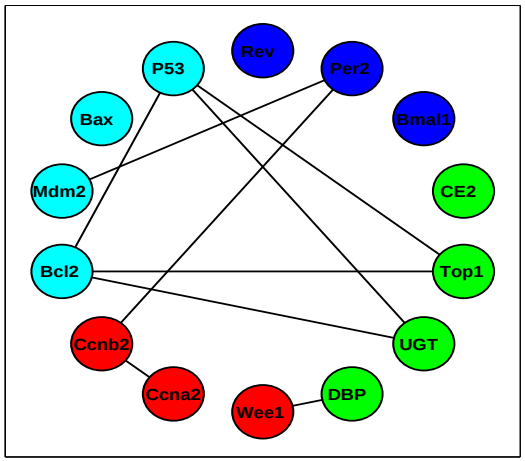


Graph of Dependancies: Colon, Class 1

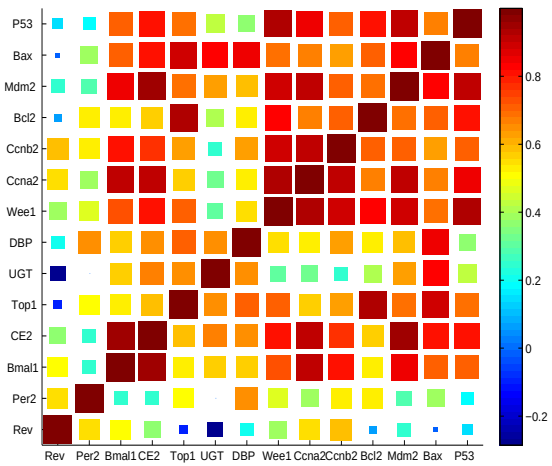
Time series

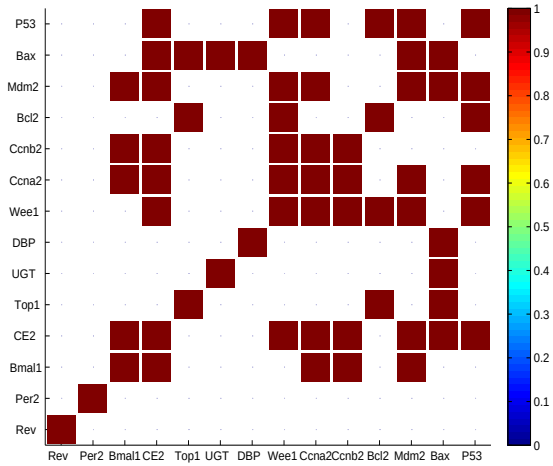


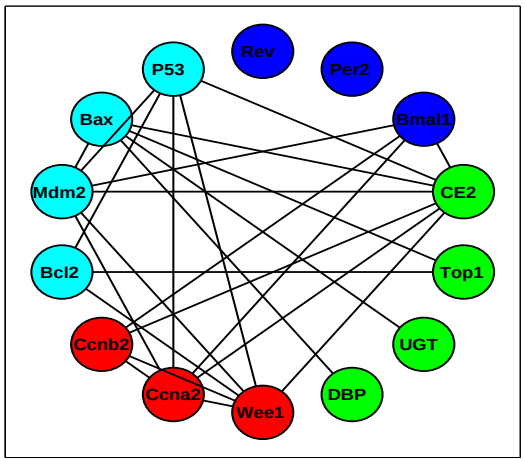




FT amplitudes

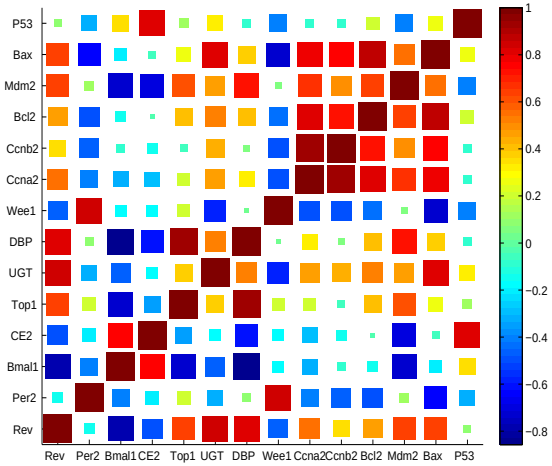


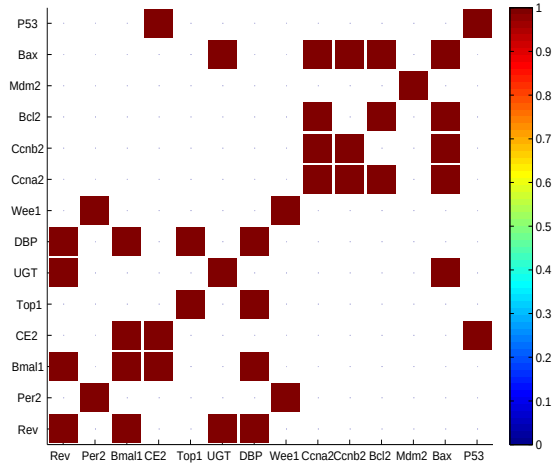


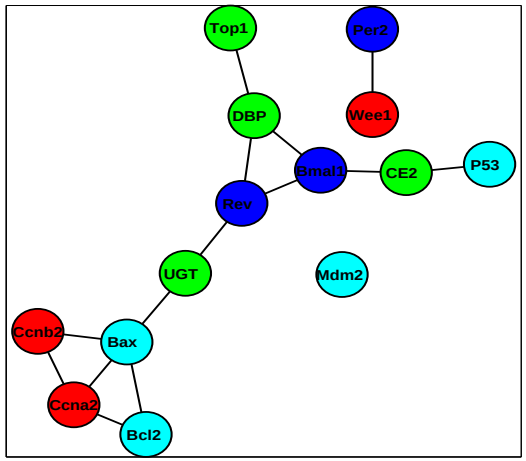


Graph of Dependancies: Colon, Class 3

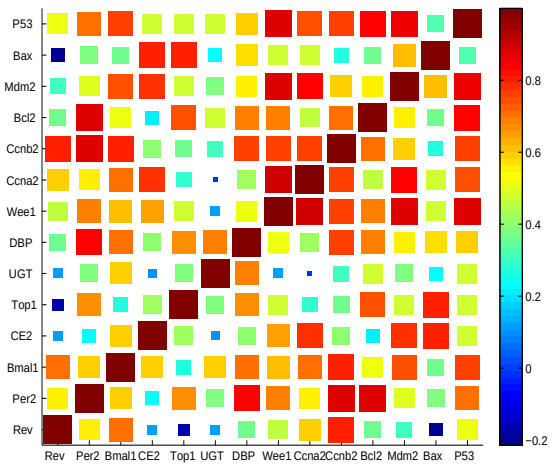
Time series

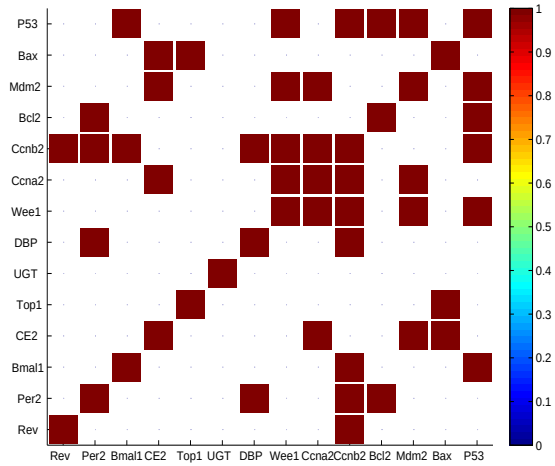


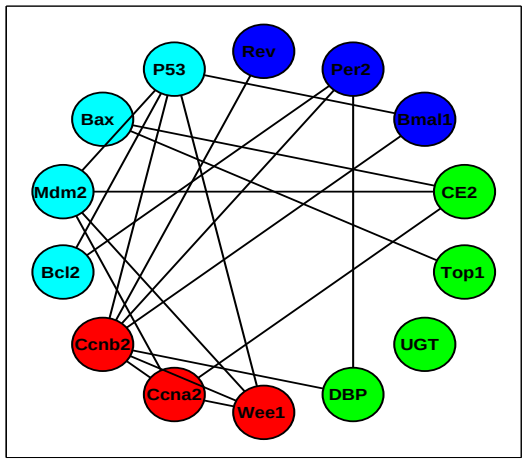




FT amplitudes



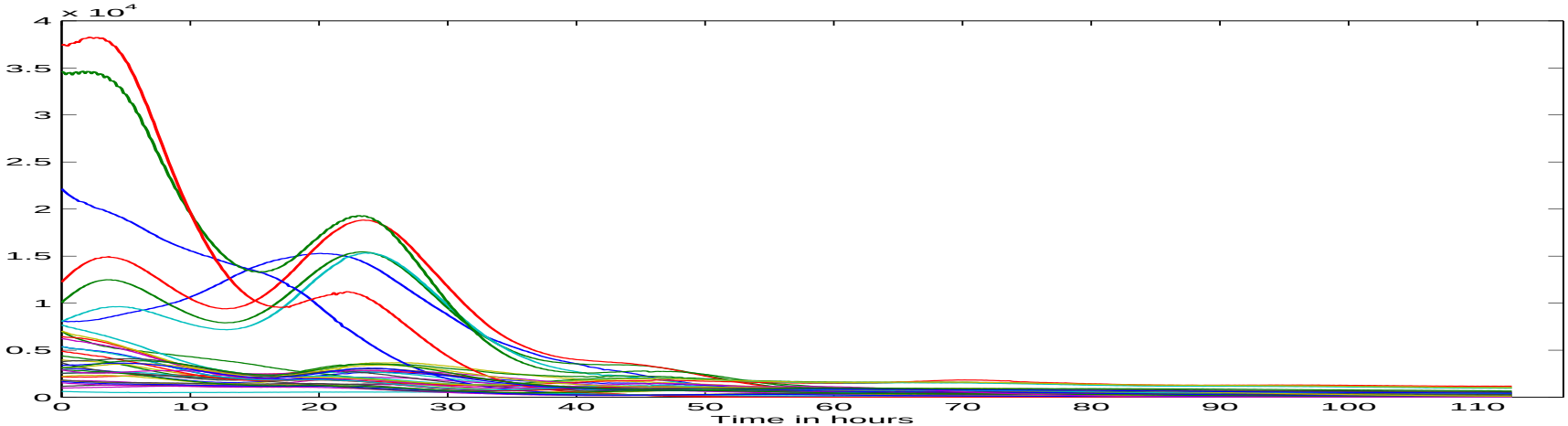


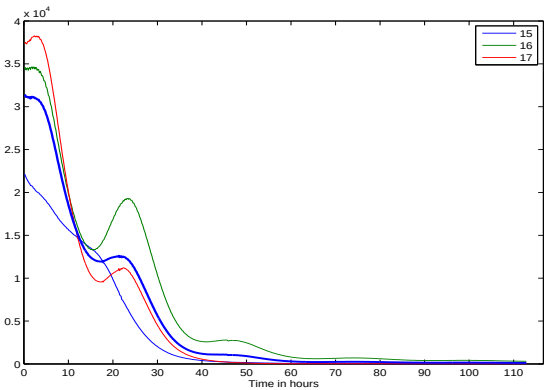


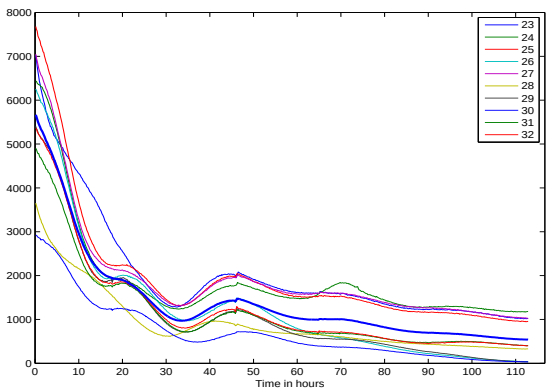
Classification tools

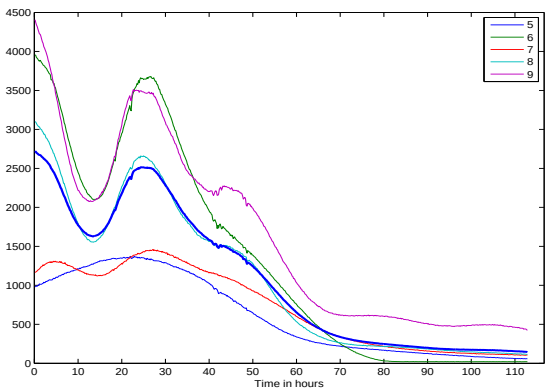
- Supervised classification
  - K nearest neighbors methods
  - Needs Training sets data
  - Must be careful to measure the performances of the classification on a different set of data (Test set)
- Unsupervised classification
  - Mixture models
  - Expectation-Maximization methods
  - Bayesian versions of EM
  - Bayesian Variational Approximation (VBA)

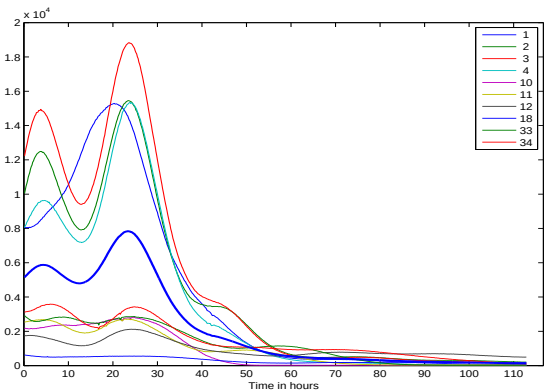
Classification tools

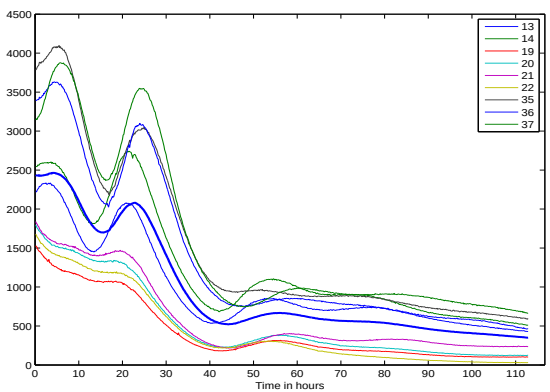












Input-Output modeling using training data and test data

- Linear models: $\boldsymbol{g}_k = \boldsymbol{A}\boldsymbol{f}_k + \boldsymbol{\epsilon}_k, \quad k = 1, \cdot, K$
- Bayesian framework, MAP estimation with hyperparameter estimation

$$p(\boldsymbol{A}|\{\boldsymbol{g}_k, \boldsymbol{f}_k\}) \propto \prod_k p(\boldsymbol{g}_k|\boldsymbol{A}, \boldsymbol{f}_k) p(\boldsymbol{A})$$

- Gaussian priors for ϵ_k and for $\boldsymbol{A}$ and MAP solution: $\widehat{\boldsymbol{A}} = \arg\max_{\boldsymbol{A}} \{p(\boldsymbol{A}|\{\boldsymbol{g}_k, \boldsymbol{f}_k\})\}$

$$\widehat{\boldsymbol{A}} = \left(\sum_k \boldsymbol{g}_k \boldsymbol{f}_k'\right) \left(\sum_k \boldsymbol{f}_k \boldsymbol{f}_k' + \lambda \boldsymbol{I}\right)^{-1}$$

- Other priors to enforce sparsity or bloc-sparsity of the prediction matrix $\boldsymbol{A}$
- See the poster of Mircea Dumitru et al for weight loss prediction from the two genes expressions of Bmal1 and Rev-erb-alpha

Conclusions

- A lot to do to answer the questions of biologists
- Forward modeling and Bayesian inference are natural tools to answer these questions
- Very often the questions are ill-posed inverse problems which need prior knowledge
- Appropriate translation of prior knowledge to prior laws is very important
- Carefull computational algorithms have to be developped
- Carfeful presentation and interpretation of the inference results are very important
- Constant dialogue between ”biologists” and ”Data and Signal processors” is of great importance.

Publications

- 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.