Summary of multicomponent/multivariate data and time series signal analysis tools developed during C5Sys EraSysBio 2010-2013

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 concortium meeting, April 24-27, 2013, Florence, Italy

Summary

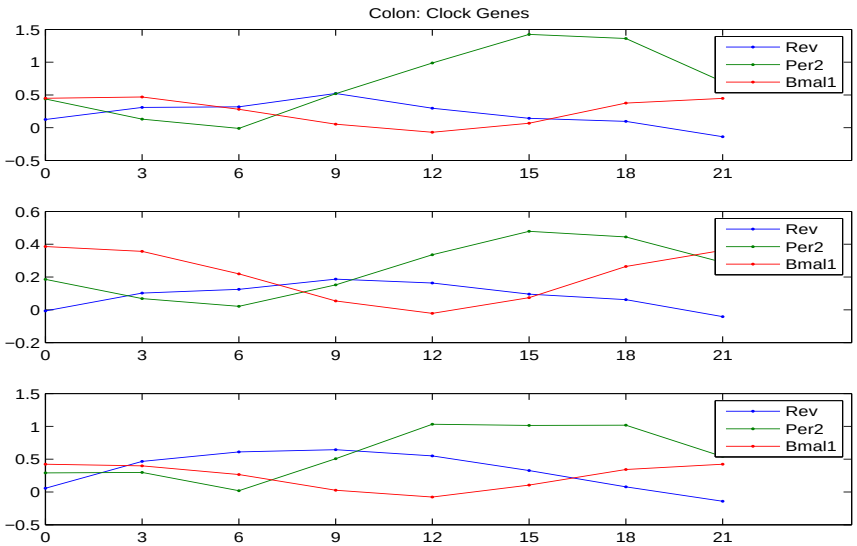
- Visualization tools
  - Time domain
  - Transformed domain: Fourier, Wavelets, Splines, ...
  - Scatter plots, histograms, ...
- Modeling time series
  - Parametric:
    Superposition of sinusoids (COSINOR),
    Superposition of Gaussians shapes, ...
  - Non Parametric:
    Fourier, Wavelets, ...
  - Probabilistic:
    Moving Average (MA), Autoregressive (AR), ARMA,
    Markovian models, ...
- Modeling the relation between data/signals
  - Linear / Non linear
  - Training and test data

Summary

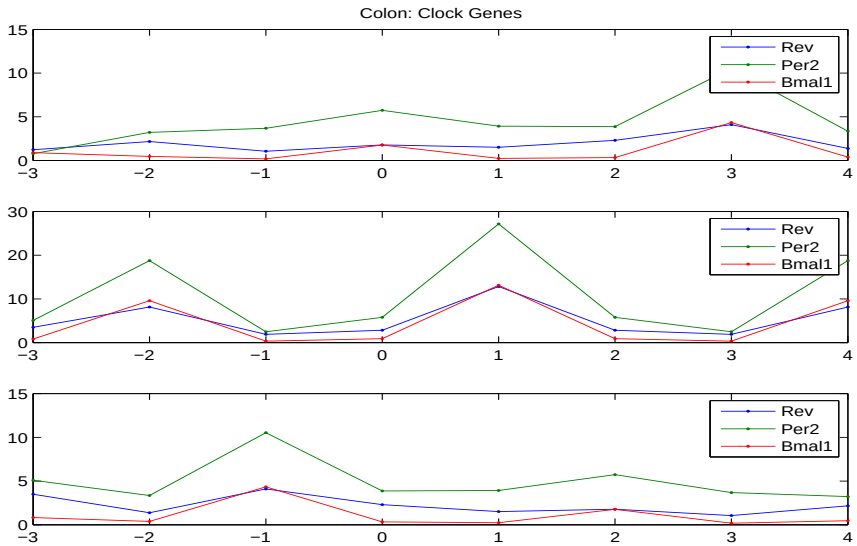
- Simple Analysis:
  Computing spectra, Estimating periods, ...
- Multicomponent/Multivariate data analysis:
  Dimensional Reduction
  PCA, FA, ICA, Sparse PCA for dimensional reduction and main factors extraction
- Multicomponent/Multivariate Discriminant Analysis with classification:
  LDA, EDA, RDA,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

Visualization and simple analysis example: Genes Clock time series and their FT

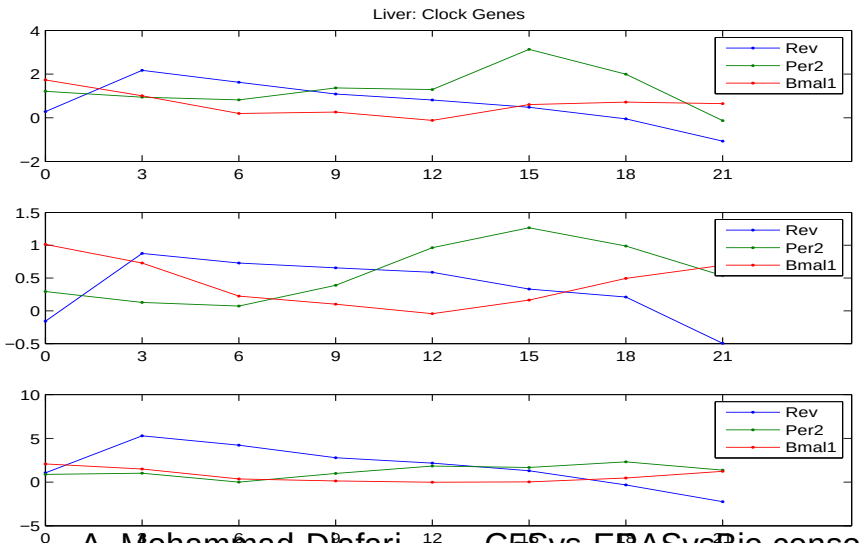
Colon: Time Series



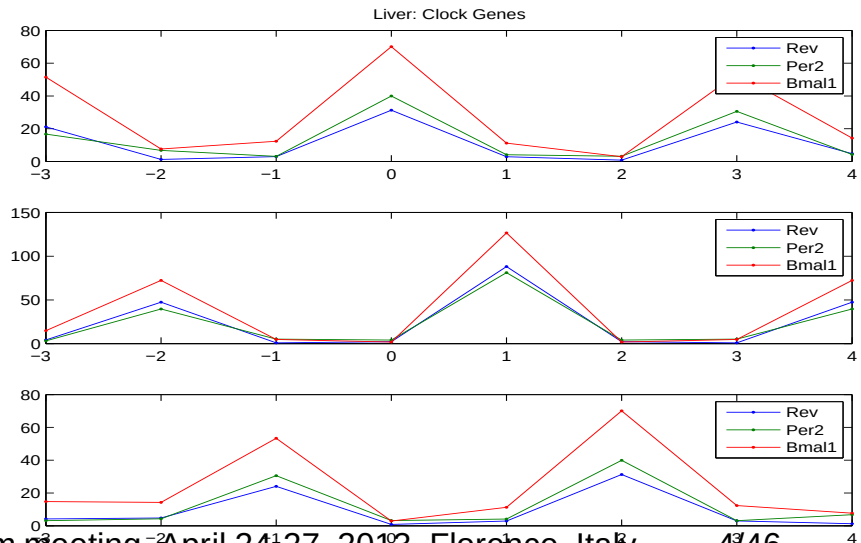
Fourier Transform



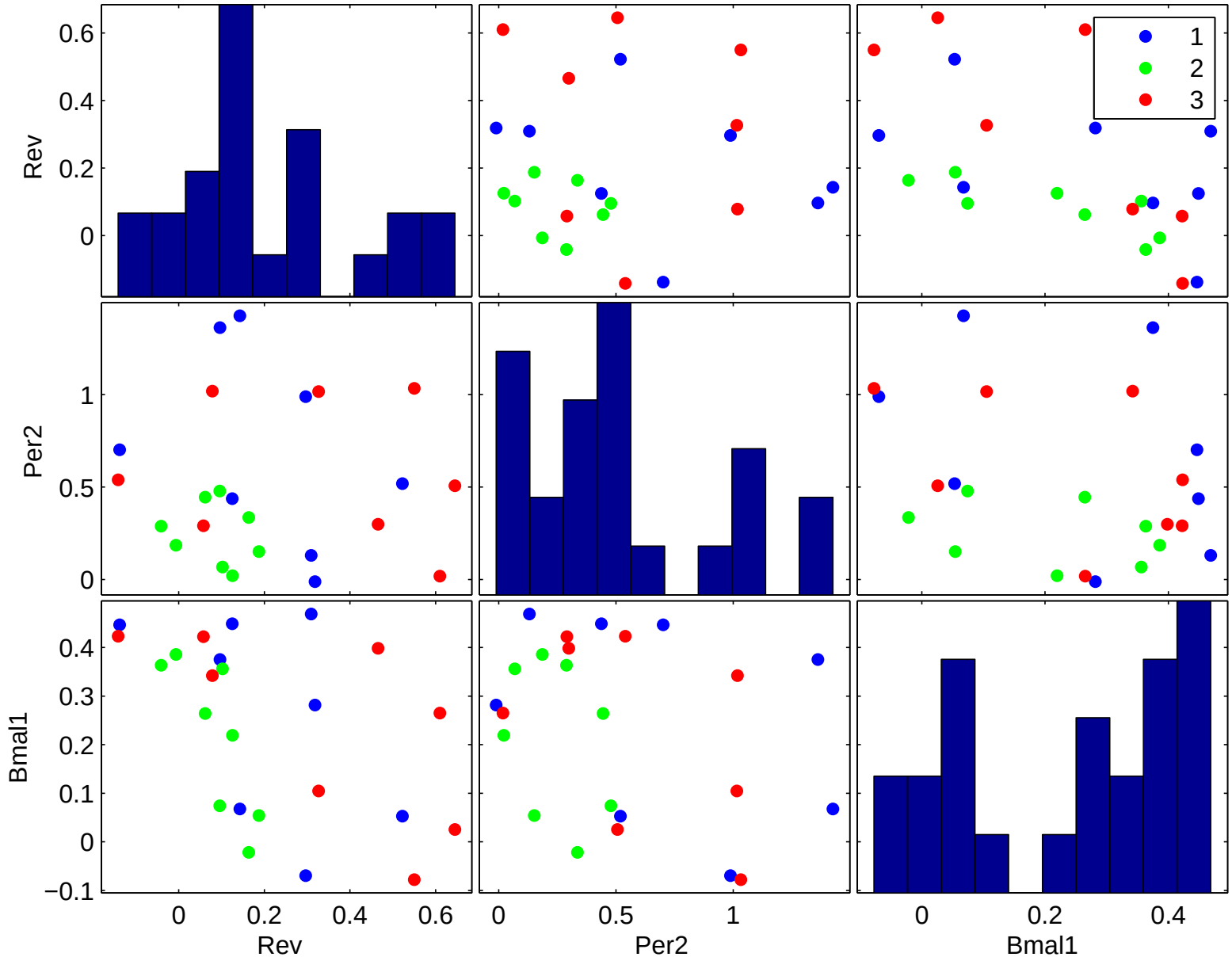
Liver: Time Series



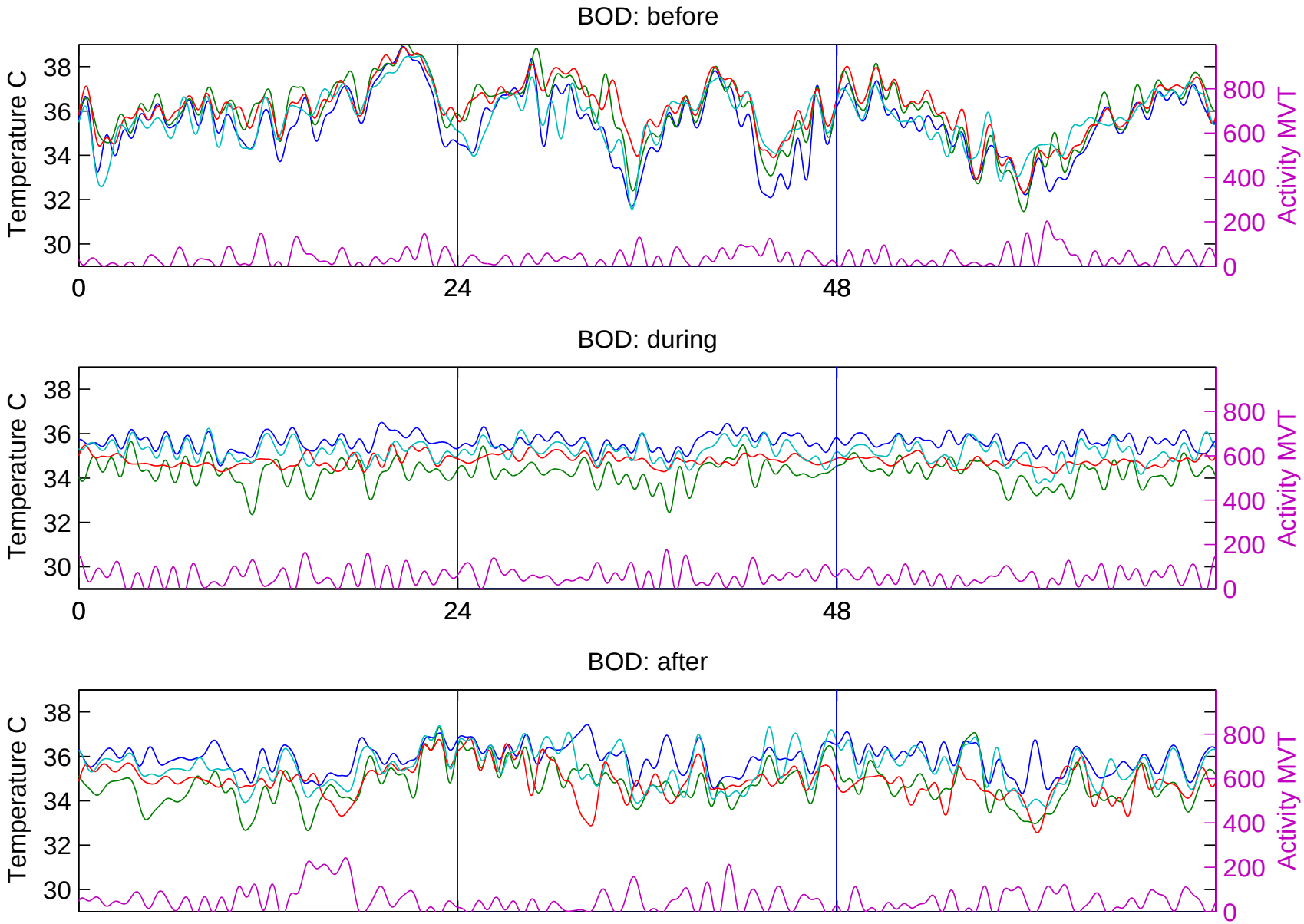
Fourier Transform



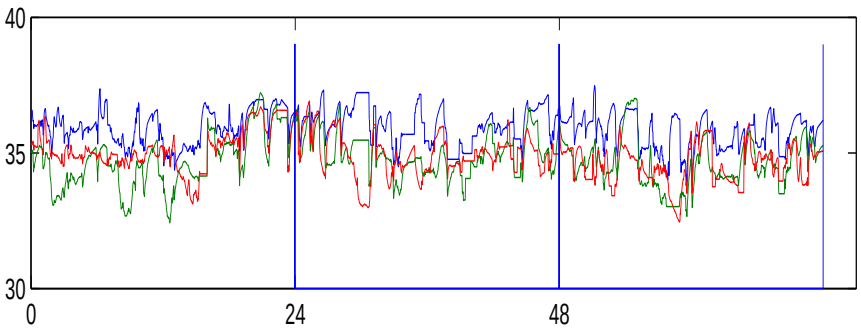
Scatterplot of Colon clock genes time series

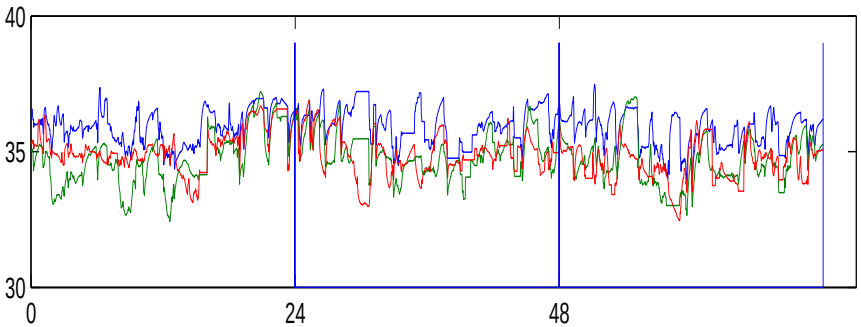


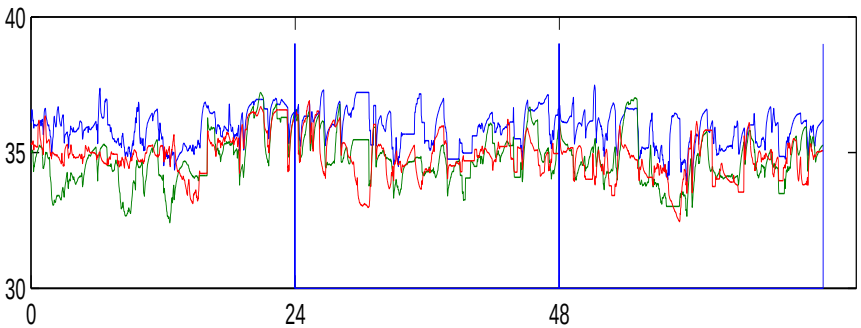
Temperature and activity Time series before, during and after some treatment



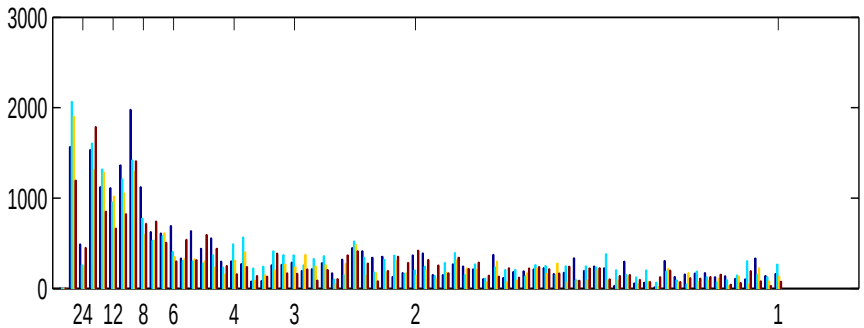
Temperatures Fourier domaine analysis

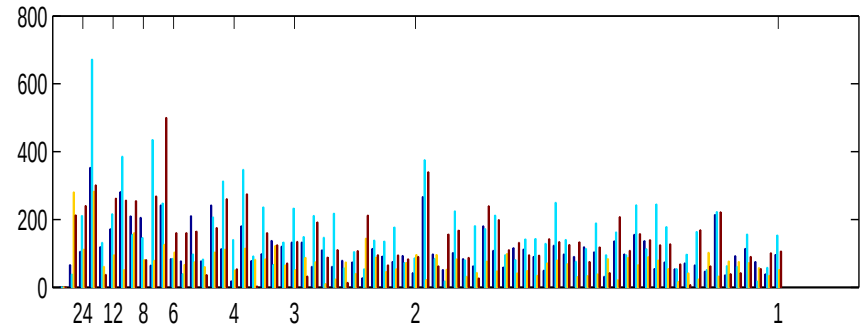


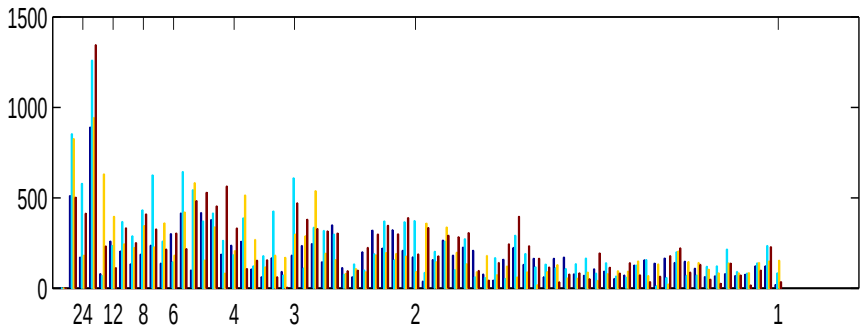




Time series

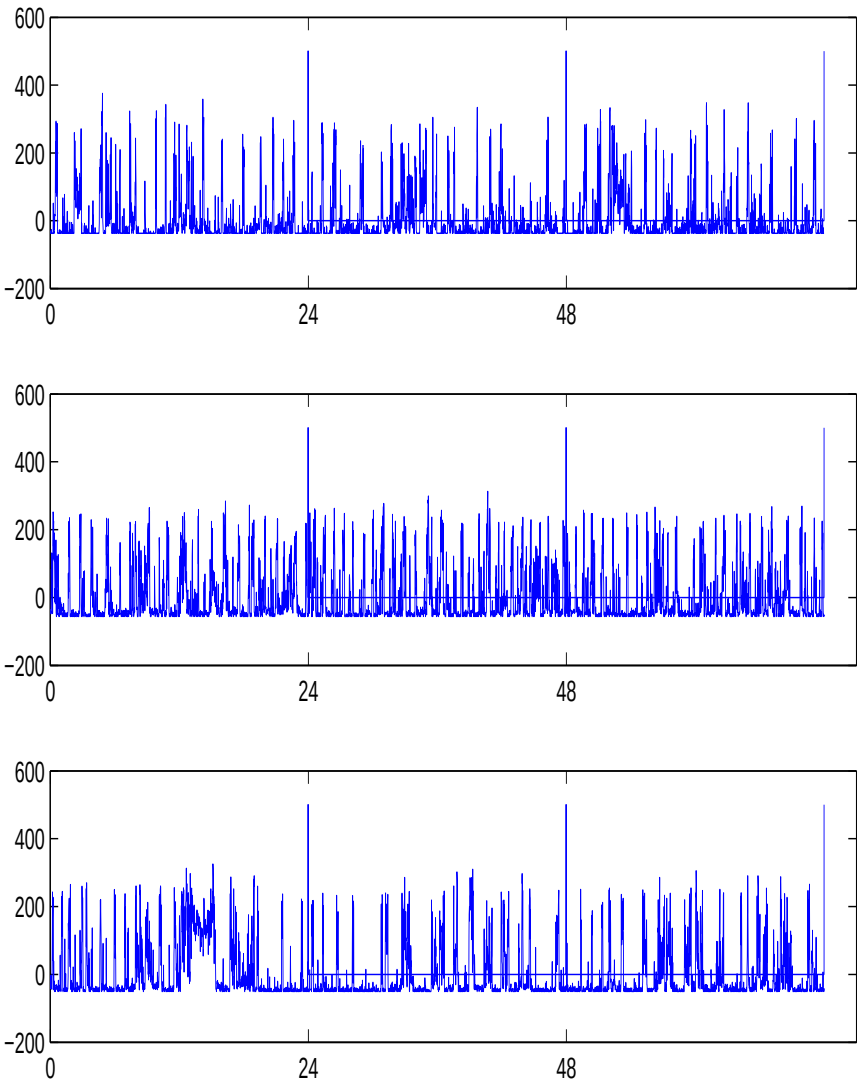




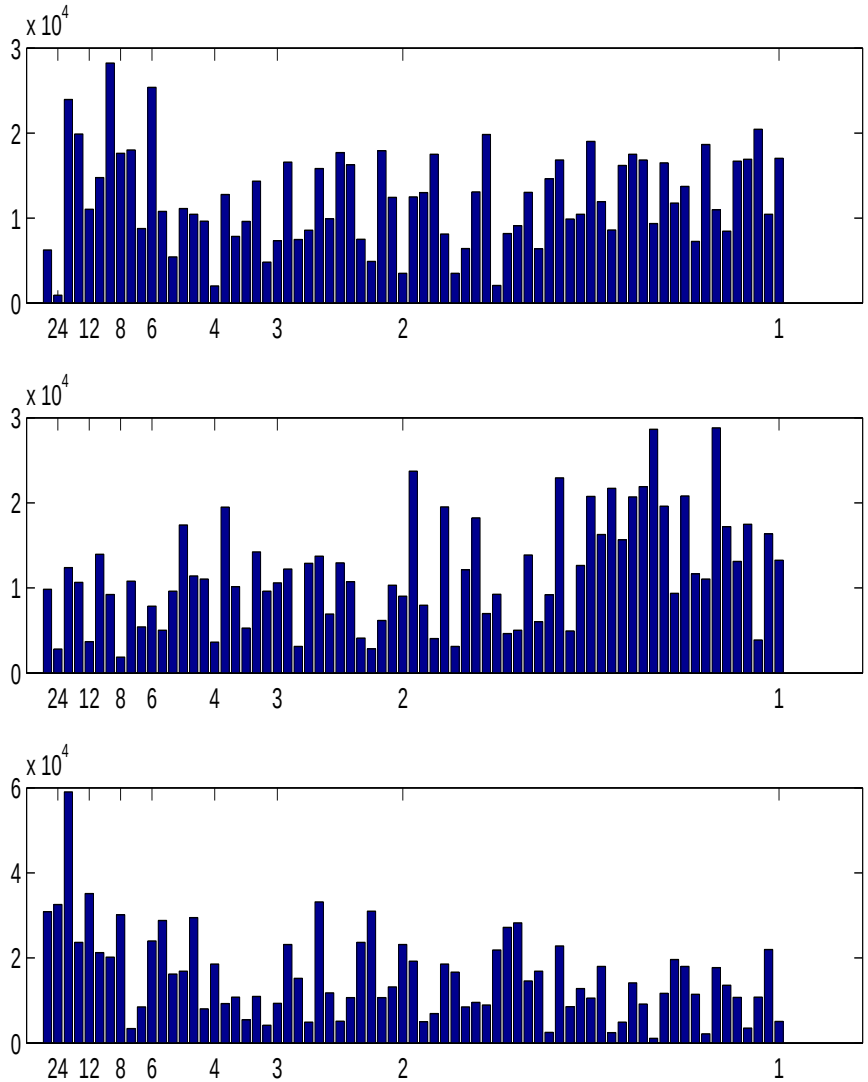


Spectra

Activity Fourier domaine analysis

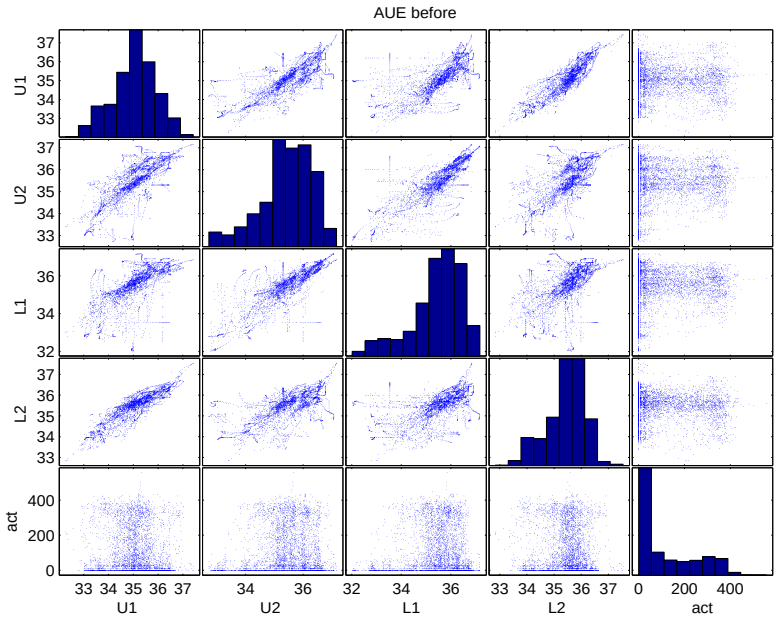


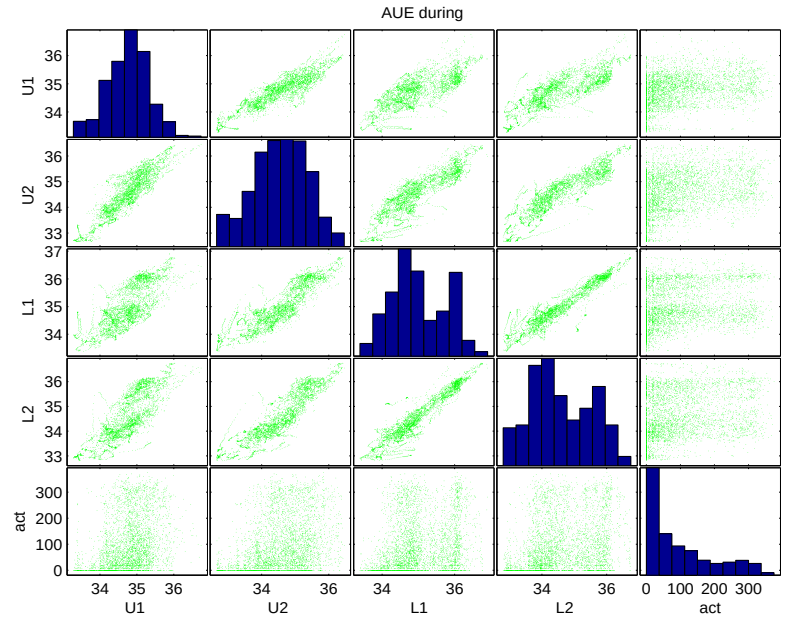
Time series

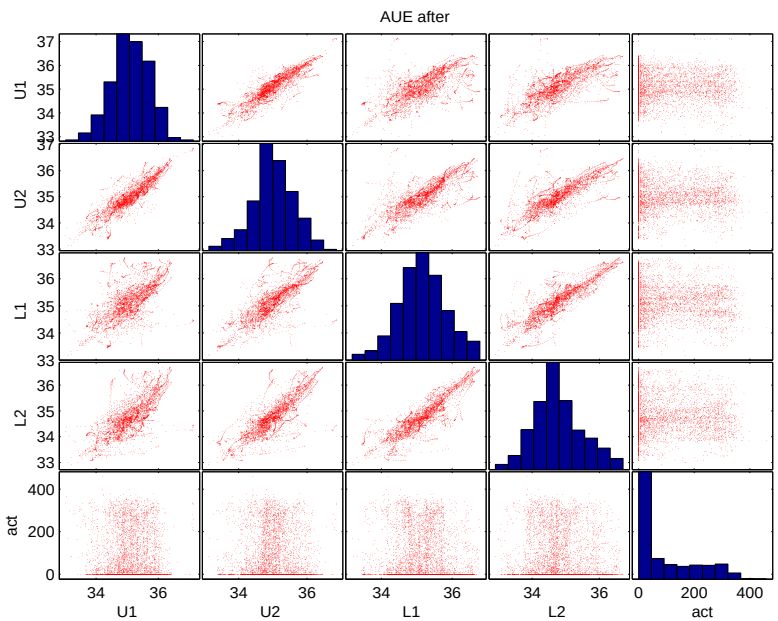


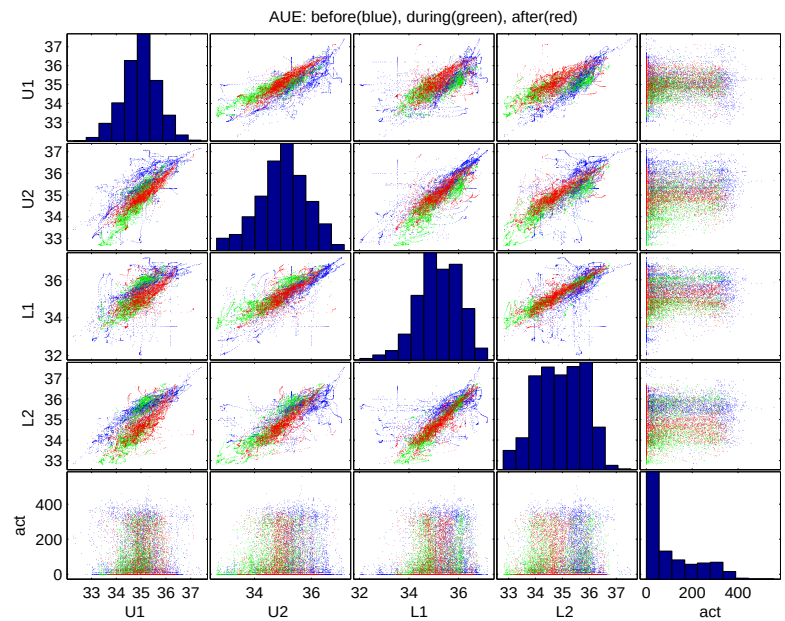
Spectra

Temperatures, before, during and after changes

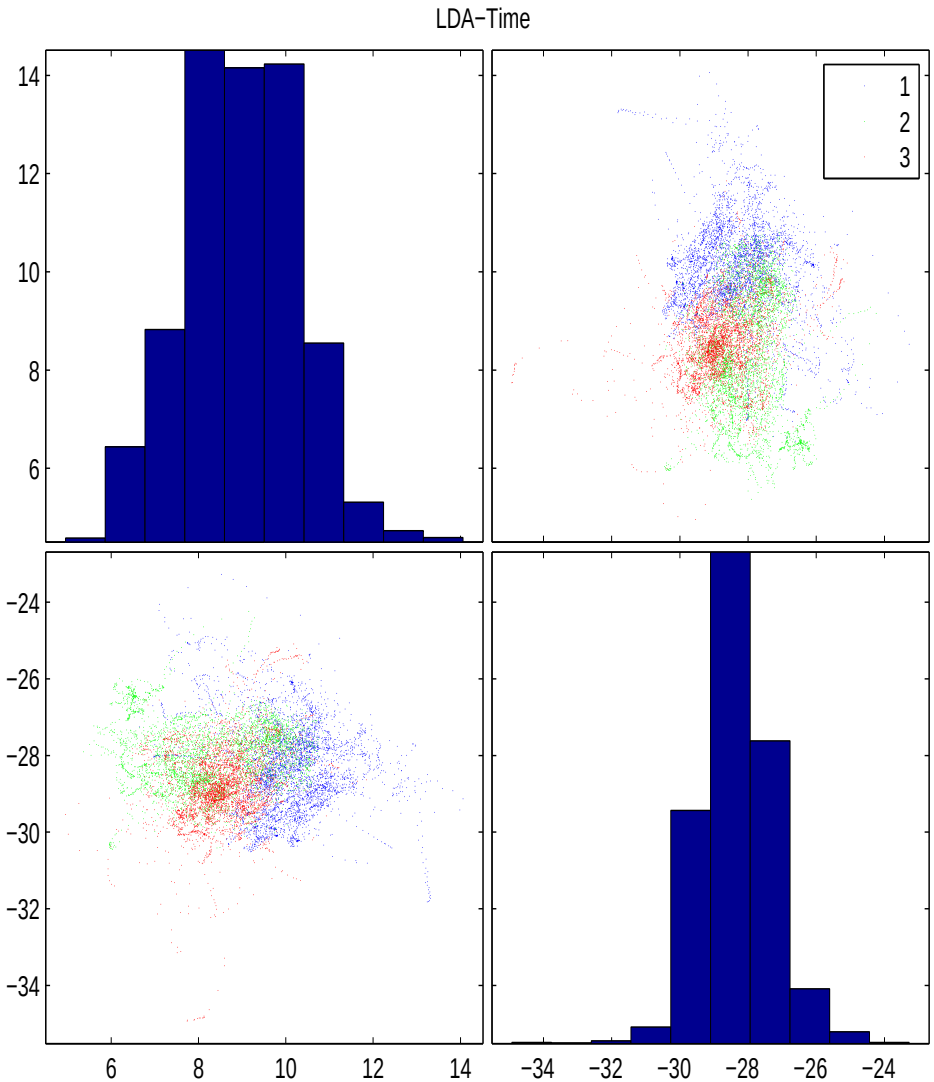


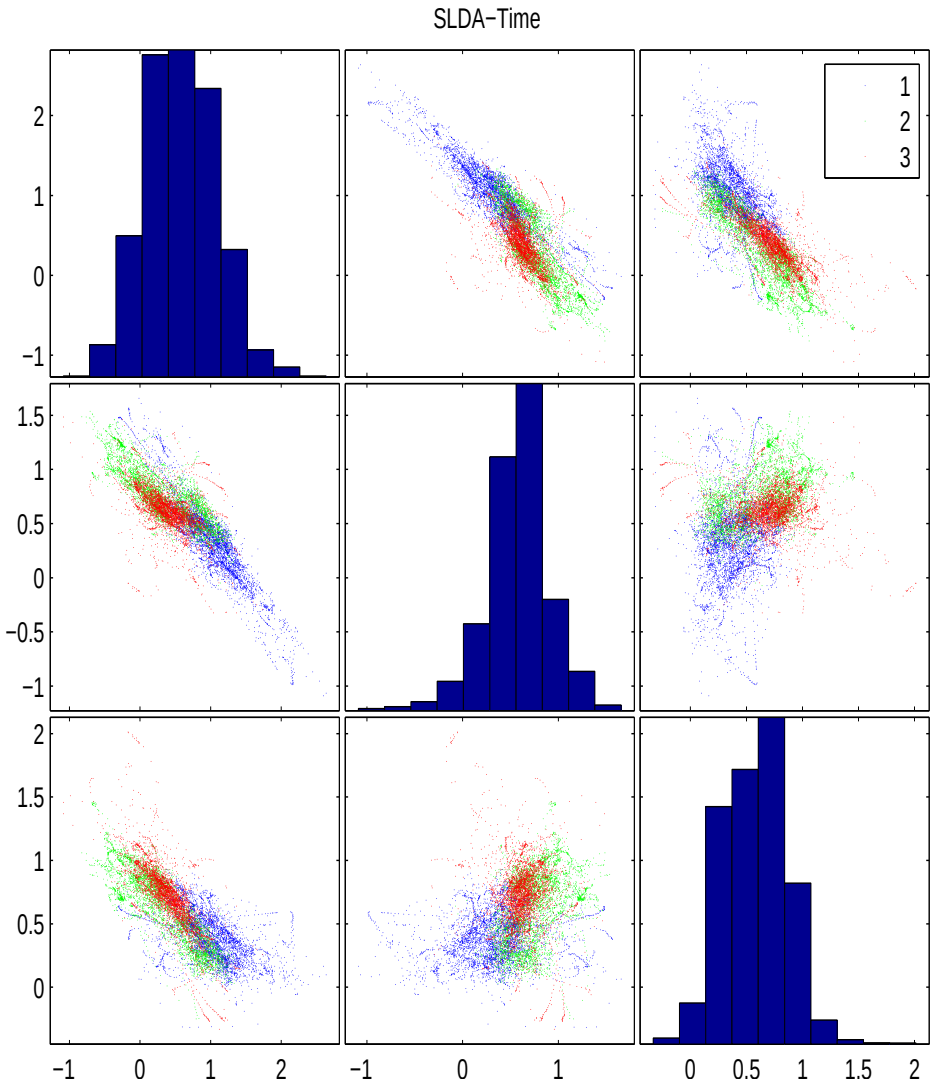






Temperatures, before, during and after changes: LDA and SLDA components space

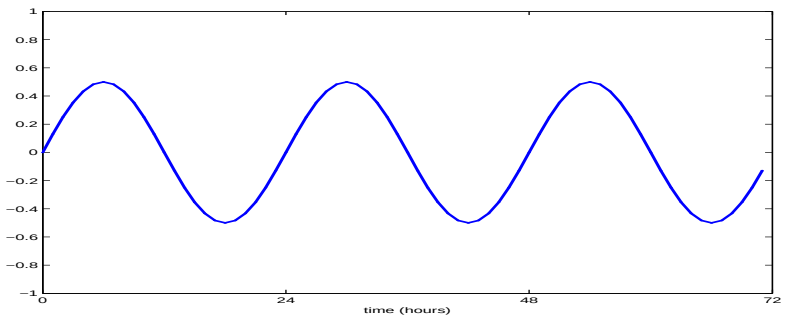


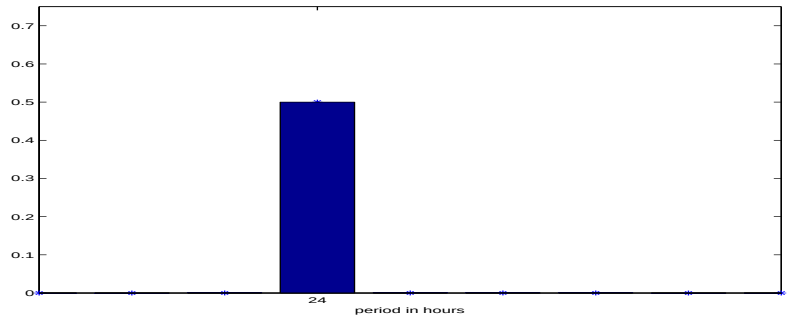


Simple Analysis tools: Period estimation.

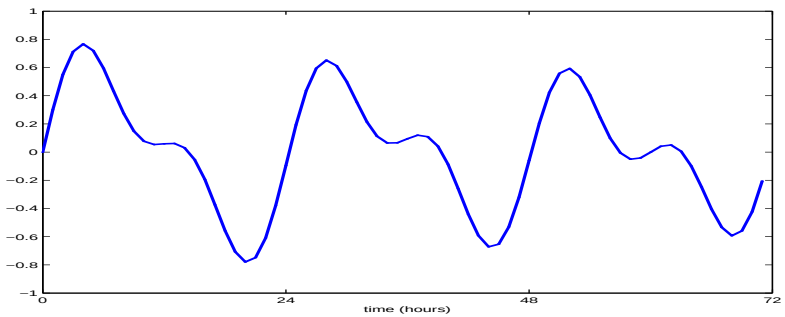
What do we mean by period ?

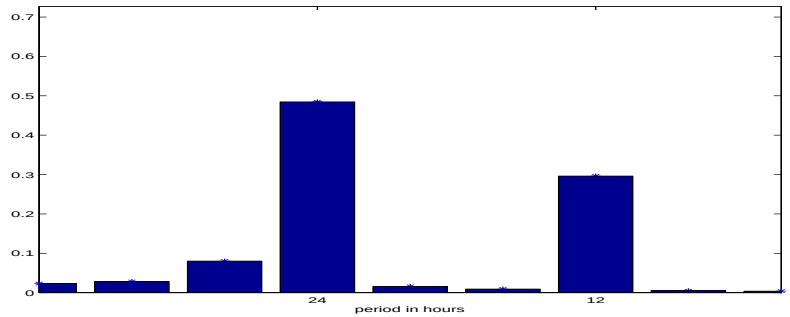
Case of 3 sinusoids



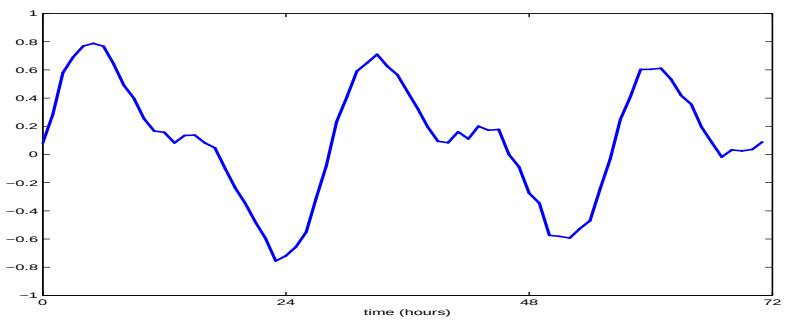


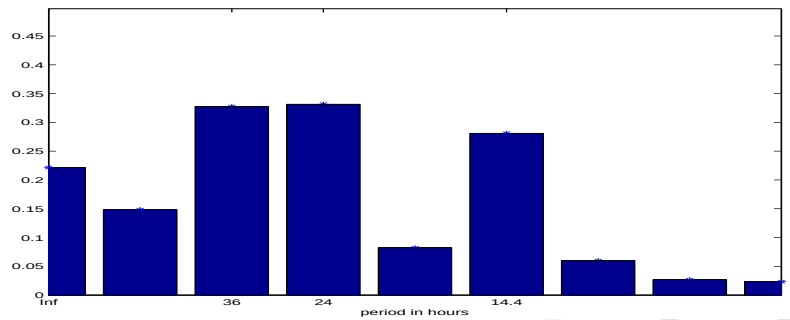
Case of 3 sinusoids+noise





Case of few sinusoids+noise

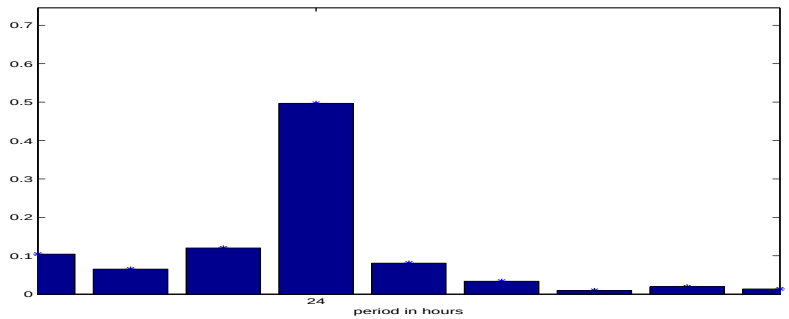


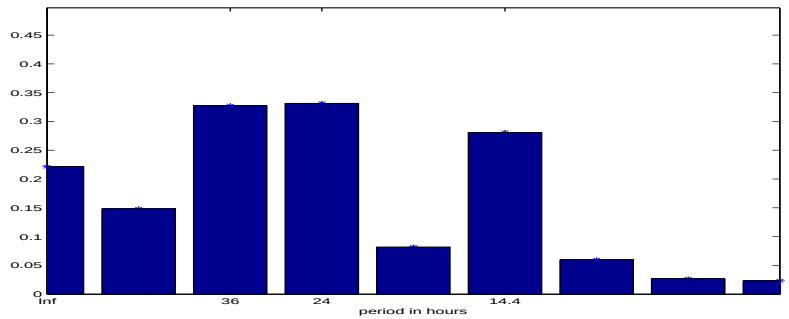


How to define a period from Spectra $S(\omega)$?

- 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: $p_{\text{L}},\ p_{\text{U}}$

Fourier Transform, Autocorrelation and Spectra

- Monovariate time series:
  - Time serie: $g(t)$
  - Aucorrelation function: $\gamma(\tau)$
  - Fourier transform: $f(\omega)$
- Spectral density function definitions: $S(\omega)$
  - Deterministic:

$$f(\omega) = \int g(t) \exp[-j\omega t] \, \mathrm{d}t \longrightarrow S(\omega) = |f(\omega)|^2 \tag{1}$$

  - Probabilistic:

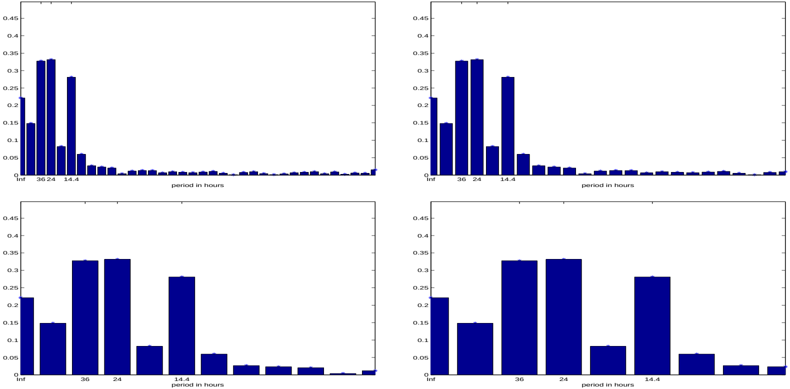
$$\gamma(\tau) = \mathrm{E}\{g(t)g(t+\tau)\} \longrightarrow S(\omega) = \int \gamma(\tau) \exp[-j\omega\tau] \, \mathrm{d}\tau \tag{2}$$

  - $S(\omega) = S(2\pi\nu) = S(2\pi/p)$

- Multivariate time series
  - $g_1(t), \cdots, g_N(t)$
  - Estimating common factors spectra

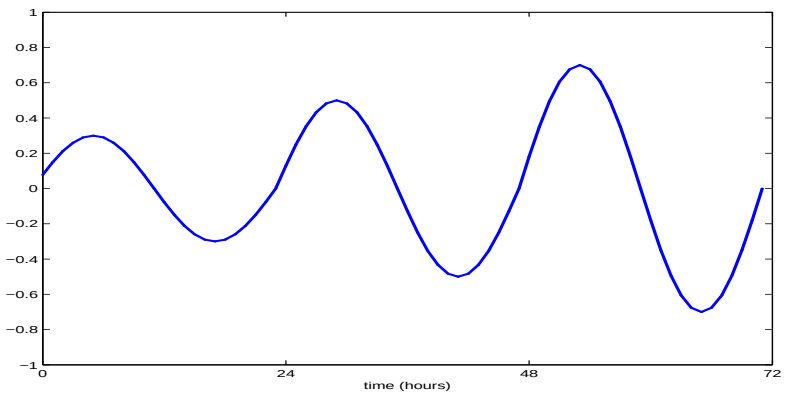
How to estimate Spectra $S(\omega)$?

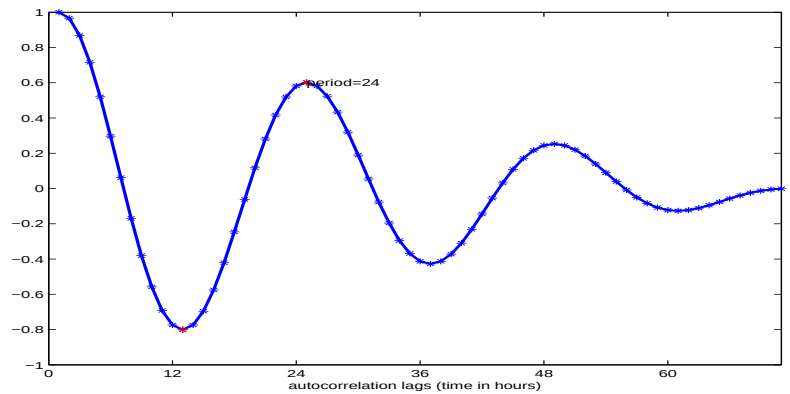
- Fast Fourier Transformn (FFT):
  $g(t) \longrightarrow FFT \longrightarrow f(\omega) \longrightarrow S(\omega) = |f(\omega)|$
  - Advantages: Well-known and understood, fast
  - Drawbacks: linear in frequencies ν,
    but not equidistance in periods
    $\nu = [0, \cdots, N-1] \longrightarrow p = [\infty, N, N/2, \cdots, N/(N-1)]$

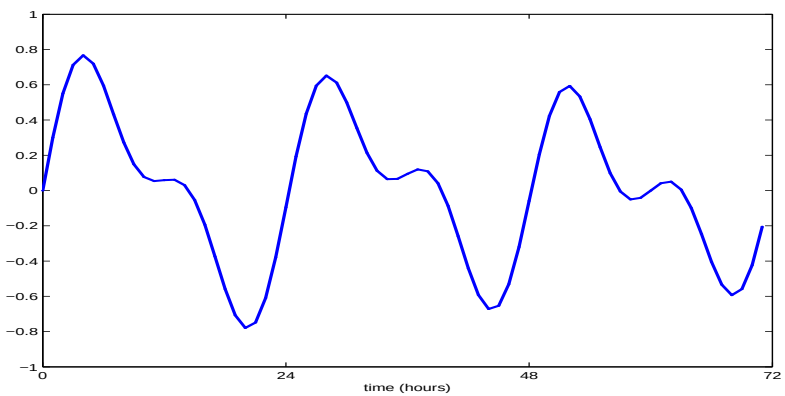


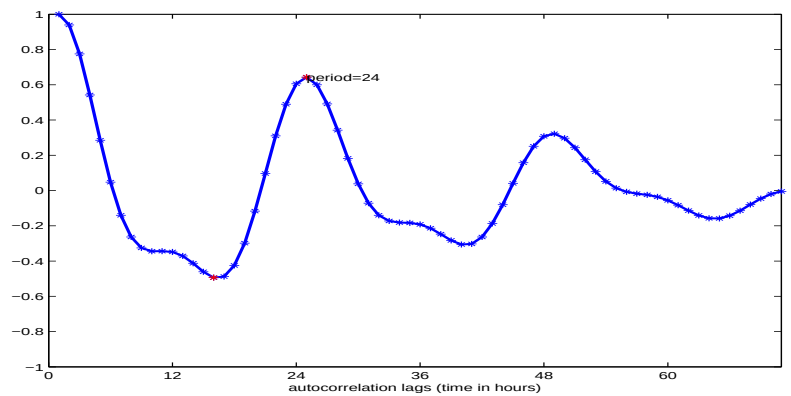
How to estimate Spectra $S(\omega)$?

- Autocorrelation function: $\gamma(\tau)$
  - If $g(t)$ is periodic, then $\gamma(\tau)$ is also periodic, but much smoother
  - $\gamma(0) = 1\ \gamma(\tau) \leq \gamma(\tau), \forall \tau$
  - Distance between $\gamma(0)$ and the next maximum gives the main period



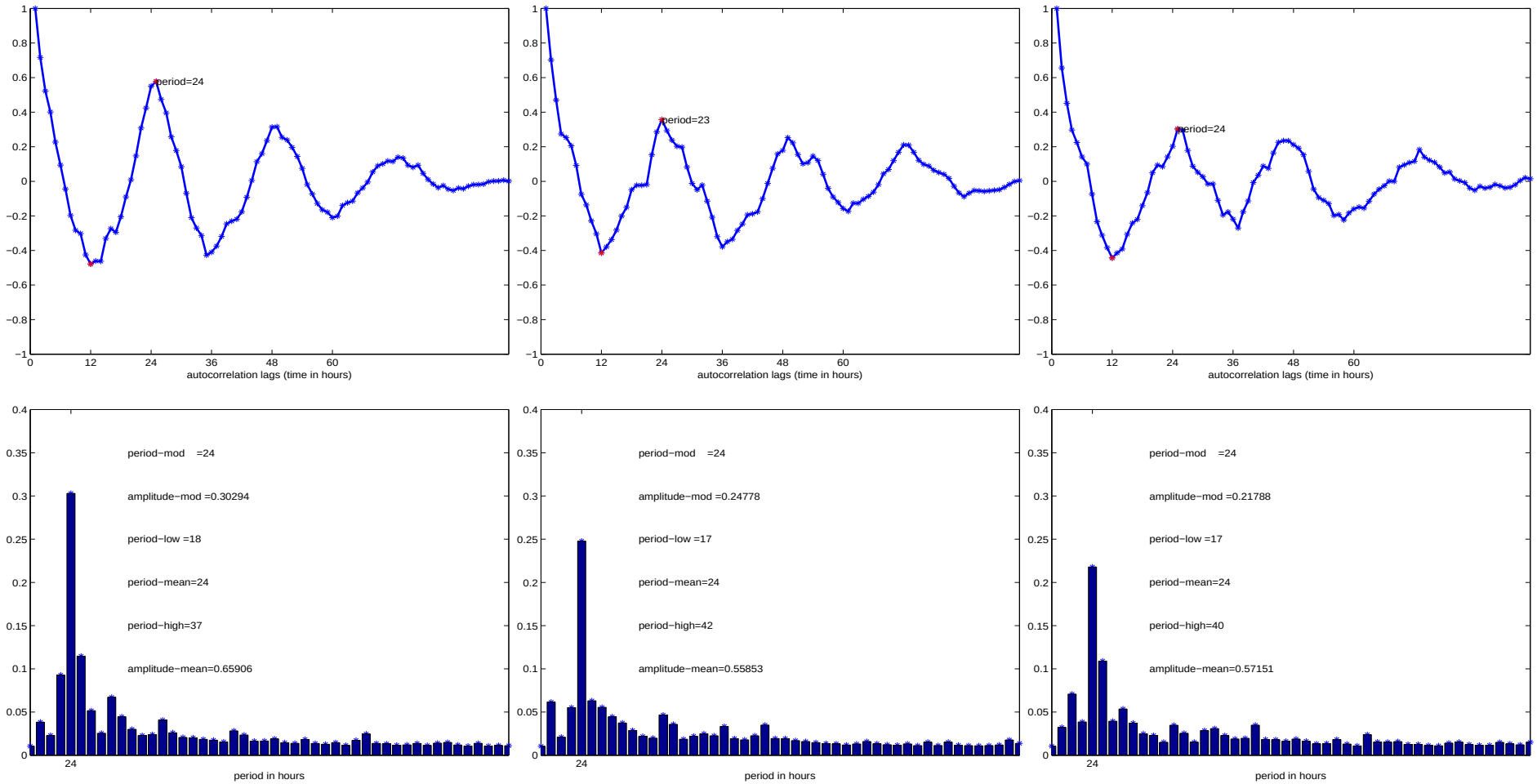






$g(t)$ $\gamma(\tau)$

Period estimation



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

- Fast Fourier Transform (FFT):
  $g(t) \longrightarrow FFT \longrightarrow f(\omega) \longrightarrow S(\omega) = |f(\omega)|$
  - 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
  - $\gamma(0) = 1\ \gamma(\tau) \leq \gamma(\tau), \forall \tau$
  - Distance between $\gamma(0)$ and the next maximum gives the main period
- Autocorrelation function and FT: $\gamma(\tau) \longrightarrow$ FFT $S(\omega)$
- Inverse problem approach:
  (Q: Compute spectra for given values of periods)
  - $f(p) \longrightarrow g(t)$ is a linear forward operation $\longrightarrow$
    $\boldsymbol{f} \longrightarrow \boldsymbol{H} \longrightarrow \boldsymbol{g}$
  - $\boldsymbol{g} = \boldsymbol{H}\boldsymbol{f} + \boldsymbol{\epsilon} \longrightarrow \widehat{\boldsymbol{f}}$

How to estimate Spectra $S(\omega)$?
Inverse Problem Approach

- $f(p) \longrightarrow g(t)$ is a linear forward operation

$$g(t_m) = \sum_n f(\omega_n) \exp\left[j(2\pi/p_n)t_m\right] \longrightarrow \boldsymbol{g} = \boldsymbol{H}\boldsymbol{f} + \boldsymbol{\epsilon},$$

$$H_{mn} = \exp\left[j(2\pi/p_n)t_m\right], \quad \begin{matrix} m = 1, \cdots M \\ n = 1, \cdots N \end{matrix}$$

- Discrete (Fast) Fourier Transform (DFT/FFT): $\boldsymbol{g} = \boldsymbol{H}\boldsymbol{f}$
  If $M = N$ and if p_m are chosen such that $\omega_n = 2\pi/p_n = [0 : N-1]\omega_0$ with $\omega_0 = 2\pi/\delta t$, then $\boldsymbol{H}$ is the DFT matrix and

$$\boldsymbol{H}'\boldsymbol{H} = \boldsymbol{I} \longrightarrow \widehat{\boldsymbol{f}} = \boldsymbol{H}^{-1}\boldsymbol{g} = \boldsymbol{H}'\boldsymbol{g}$$

- General case: For example when we want to compute $f(\omega)$ for equidistance valued periods. Then, $\boldsymbol{H}'\boldsymbol{H} \neq \boldsymbol{I}$ and even the numbers data and unknowns different

How to estimate Spectra $S(\omega)$?
Inverse Problem Approach

- General case $\boldsymbol{g} = \boldsymbol{H}\boldsymbol{f}$
- Generalized inverse or pseudoinverse

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

$$\widehat{\boldsymbol{f}} = \arg\min_{\boldsymbol{f}} \left\{ \|\boldsymbol{f}\|^2 \right\} \text{ s.t. } \boldsymbol{H}\boldsymbol{f} = \boldsymbol{g} \longrightarrow \widehat{\boldsymbol{f}} = \boldsymbol{H}'(\boldsymbol{H}\boldsymbol{H}')^{-1}\boldsymbol{g}$$

- Better if we account for ill-conditioning of the $\boldsymbol{H}$
- Still better, if we account for errors and uncertainties

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

How to estimate Spectra $S(\omega)$?
Inverse Problem Approach

- Regularization:
  Better if we account for ill-conditioning of the $\boldsymbol{H}$

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

- Still better, if we account for errors and uncertainties
  $\boldsymbol{g} = \boldsymbol{H}\boldsymbol{f} + \boldsymbol{\epsilon}$
- Bayesian approach:
  - Assign the Likelihood : $p(\boldsymbol{g}|\boldsymbol{f})$
  - Assign the prior law: $p(\boldsymbol{f})$
  - Use the Bayes rule : $p(\boldsymbol{f}|\boldsymbol{g}) = \frac{p(\boldsymbol{g}|\boldsymbol{f})\, p(\boldsymbol{f})}{p(\boldsymbol{g})}$
  - Use this posterior law to infer on $\boldsymbol{f}$.
  - For example MAP:

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

Bayesian estimation of spectra

- Bayesian approach: $p(\boldsymbol{f}|\boldsymbol{g}) \propto p(\boldsymbol{g}|\boldsymbol{f})\,p(\boldsymbol{f})$
- 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,...)

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

  - Gaussian $\Omega(\boldsymbol{f}) = \sum_j |f_j|^2$
  - Generalized Gaussian $\Omega(\boldsymbol{f}) = \sum_j |f_j|^\beta, \quad 2 \leq \beta \leq 2$
  - Cauchy $\Omega(\boldsymbol{f}) = \sum_j \ln\left(1 + |f_j|^2\right)$

Bayesian estimation of spectra with priors enforcing sparsity

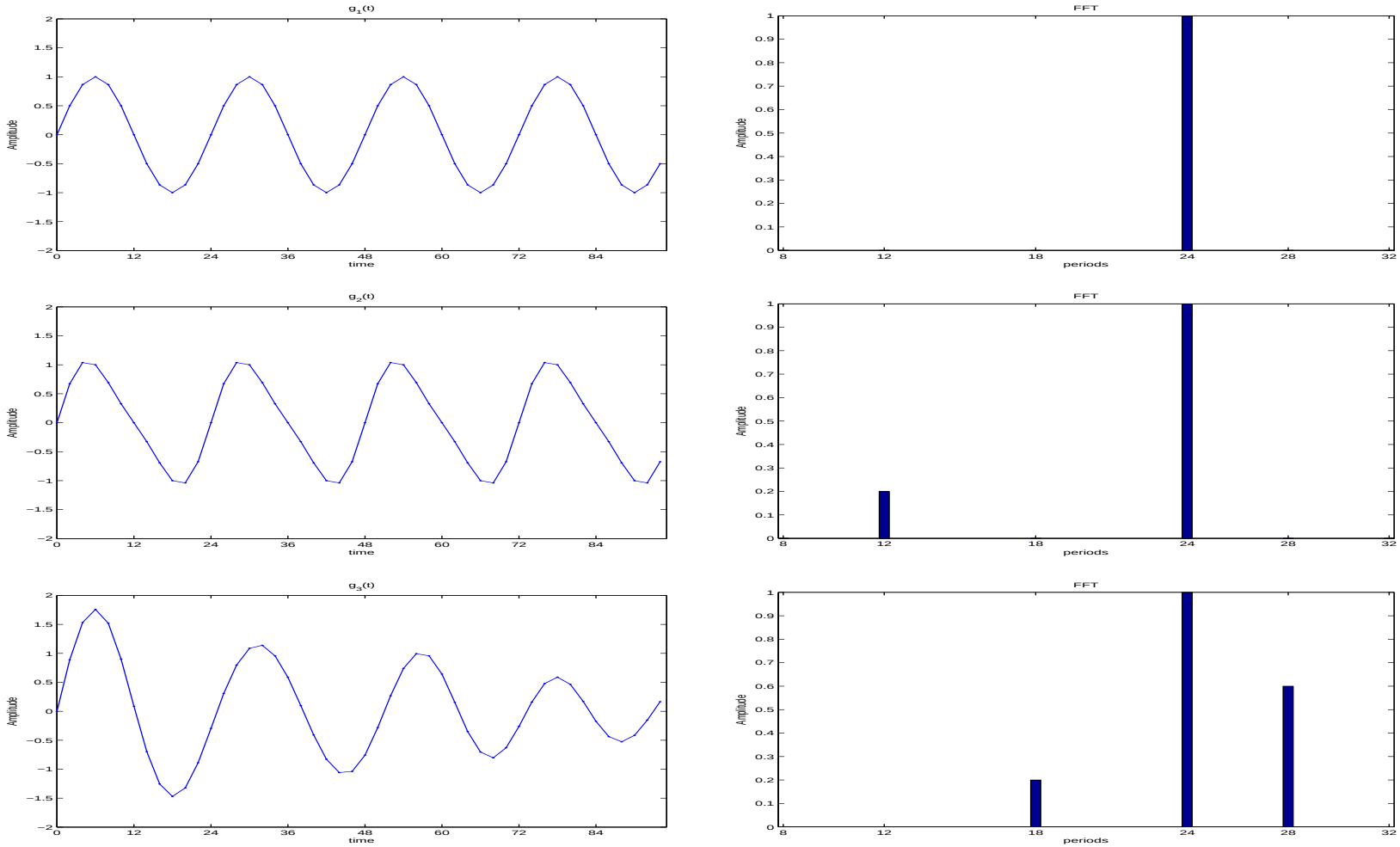
- Sparsity: For any periodic signal, the spectrum is a set of Diracs
- Biological signals related to clocks: a few independent oscillators
- Spectrum has a few non zero elements in any given interval

- How to translate this information ?
- Use a heavy tailed prior law like Double exponential or Cauchy
- Use a hierarchical prior with hidden variables
- See my paper in Eurasip journal of Advances in signal processing

Available tools for spectral and period estimation

- Spectral estimation:
  f=spectral_estimate(t,g,method,periods)
  methods:
  - FFT (range of periods is imposed)
  - Autocorr+FFT (range of periods is imposed)
  - IP:Gaussian (range of periods can be provided as desired)
  - IP:GG (range of periods can be provided as desired)
  - IP:Cauchy (range of periods can be provided as desired)
- Period estimation:
  [p_mod,a_mod,f,p_l,p_u,p_mean,a_l,a_u,a_mean]=
  period_estimate(t,g,method,periods)
  methods:
  - Autocorr maxima
  - FFT
  - Autocorr+FFT
  - IP:Gaussian (range of periods can be provided as desired)
  - IP:GG (range of periods can be provided as desired)
  - IP:Cauchy (range of periods can be provided as desired)

Multicomponent period estimation



$$\boldsymbol{g}_k = \boldsymbol{H}\boldsymbol{f}_k + \epsilon_k$$

$\boldsymbol{f}_k$ have some common spectra.

Dimension reduction, PCA, Factor Analysis, ICA

- PCA, Factor Analysis and ICA try to answer to the question:
  How many Principal components (Factors, Independent Components) can describe the observed data?

$$\boldsymbol{g}(t) = \boldsymbol{A}\boldsymbol{f}(t) + \epsilon(t), \quad \begin{cases} \boldsymbol{A} : (M \times N) \text{ Loading matrix }, N \leq M \\ \boldsymbol{f}(\text{t}) \ : \ \textit{factors}, \textit{sources} \end{cases}$$

- 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:

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

Deterministic and Bayesian Factor Analysis

- 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:

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

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

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

  $\beta_1 = 1$ and $\beta_2 = 1$ leads to sparse solutions

- These analysis can be done either directly on time series or on FT amplitudes.

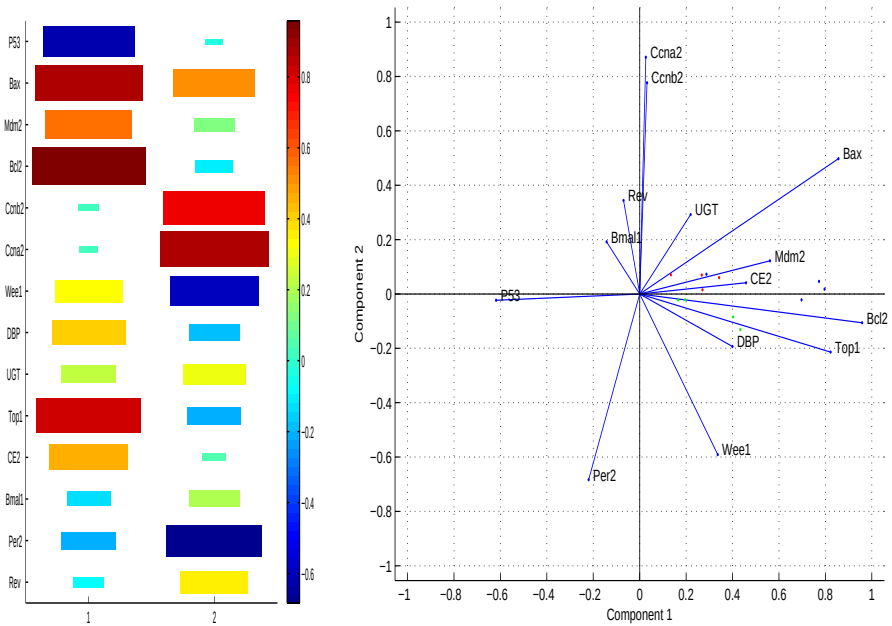
How to determine the number of factors

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

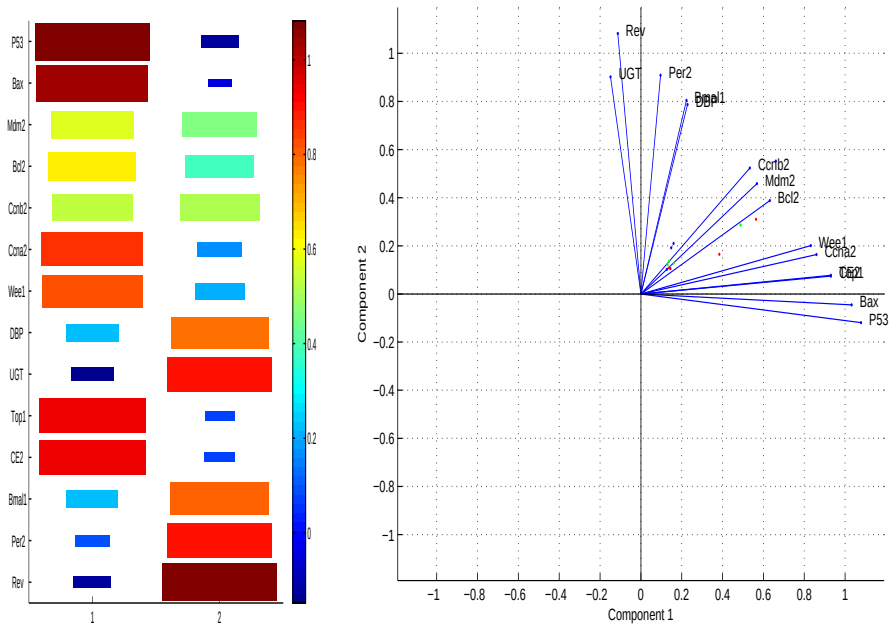
- To determine the number of factors we do the analyze with different N factors and use two criteria:
  - -log likelihood $-\ln p(\boldsymbol{g}|\boldsymbol{A}, N)$ of the observations and
  - DFE: Degrees of freedom error $(N-M)^2 - (N+M))/2$ related to AIC or BIC model selection criteria.

Factor Analysis: 2 factors: Colon

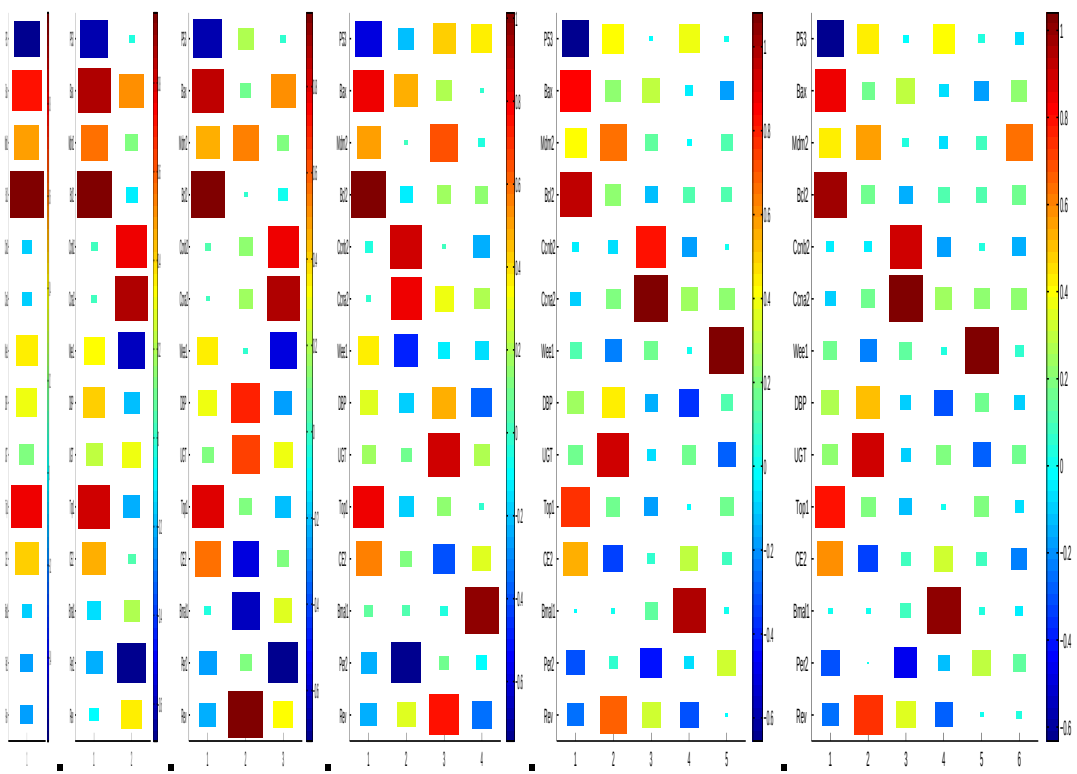
Time series

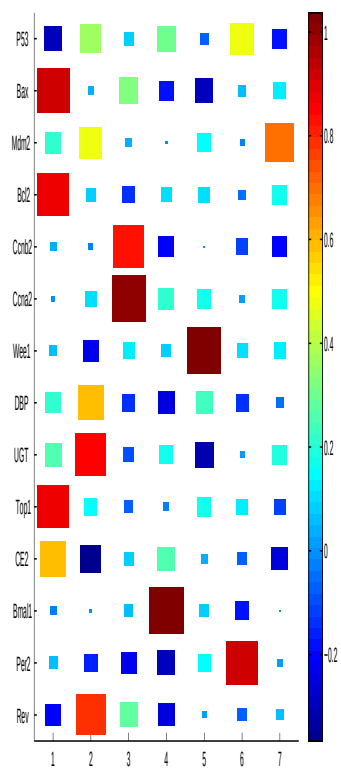


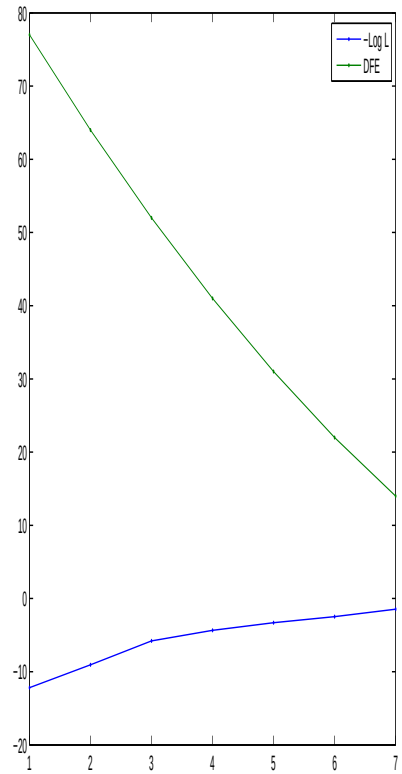
FT Amplitudes



Factor Analysis: Time series, colon



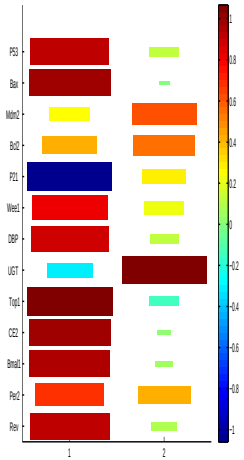




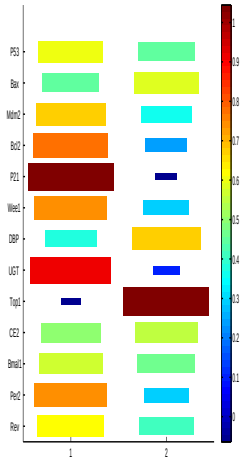
Factor Analysis for each class: FT, Liver

Two Factors:

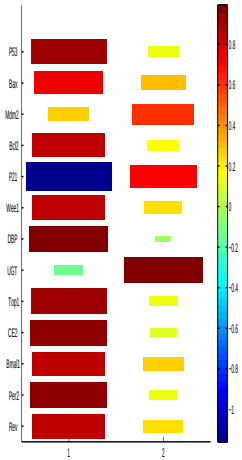
Class 1



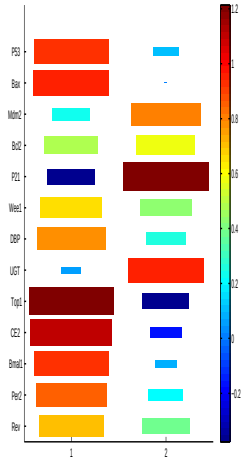
Class 2



Class 3

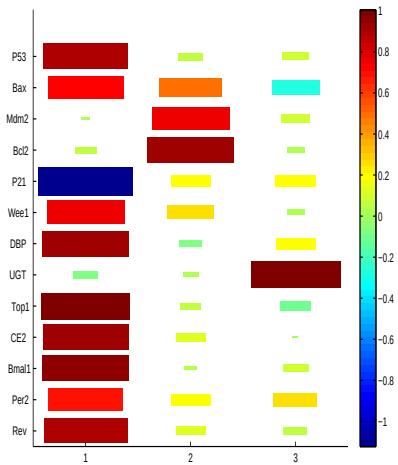


All Classes

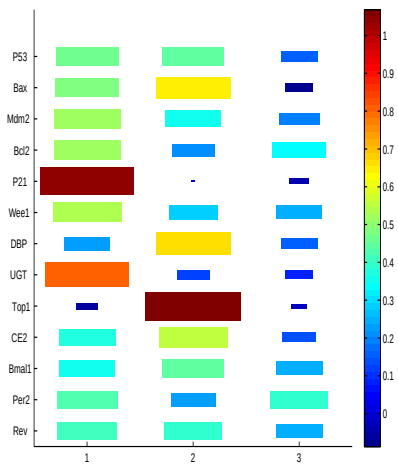


Three Factors:

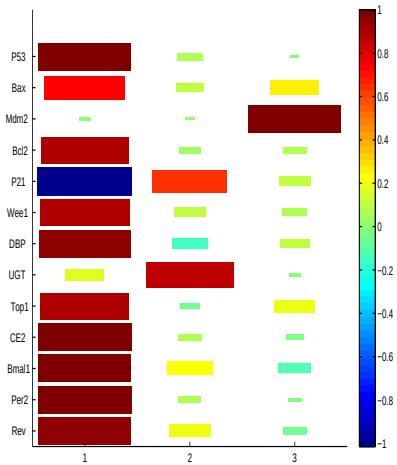
Class 1



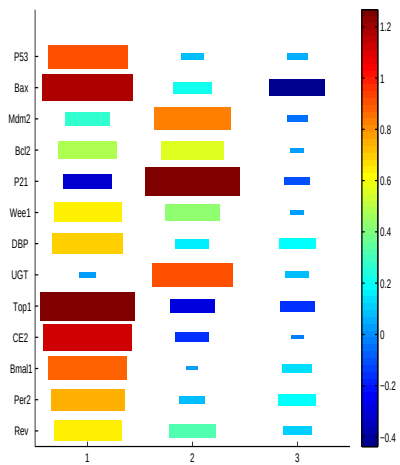
Class 2



Class 3



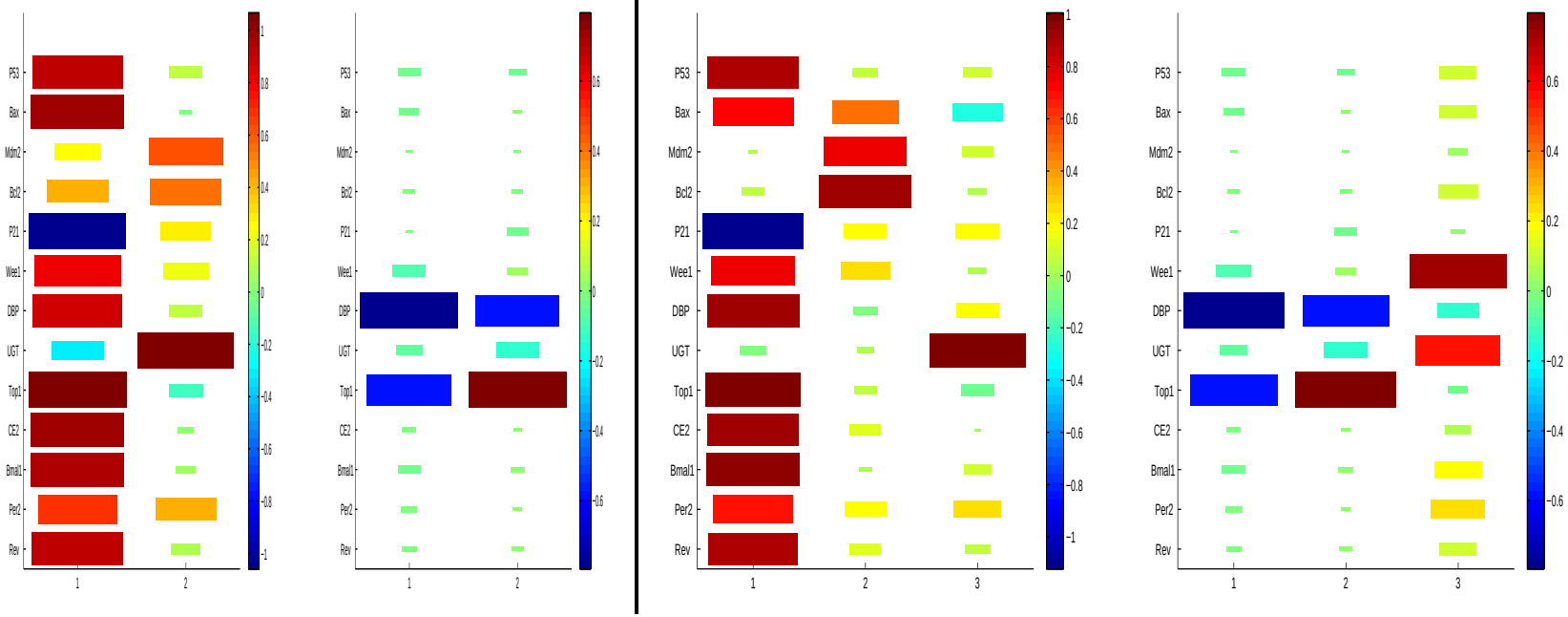
All Classes



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.

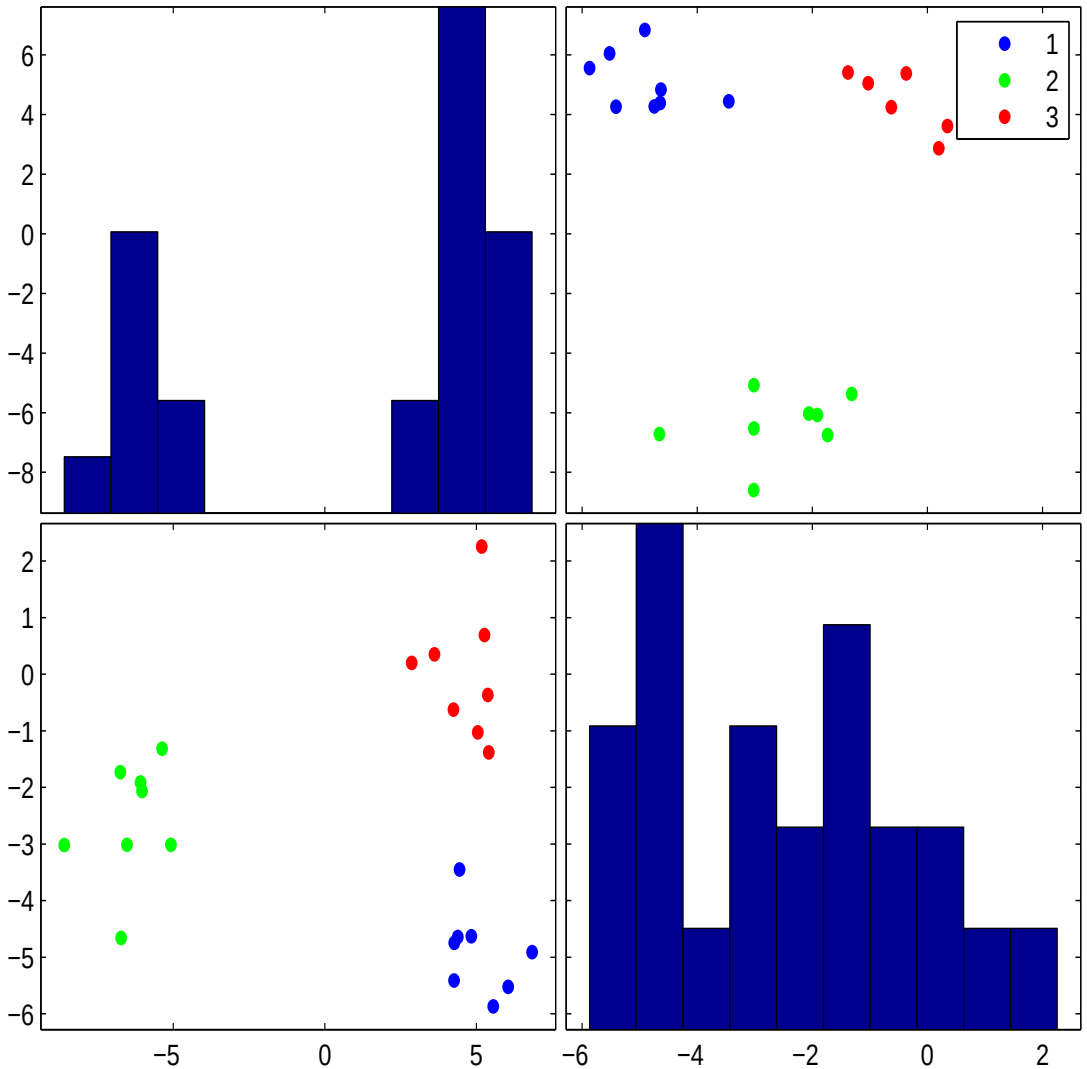
PCA SPCA PCA SPCA

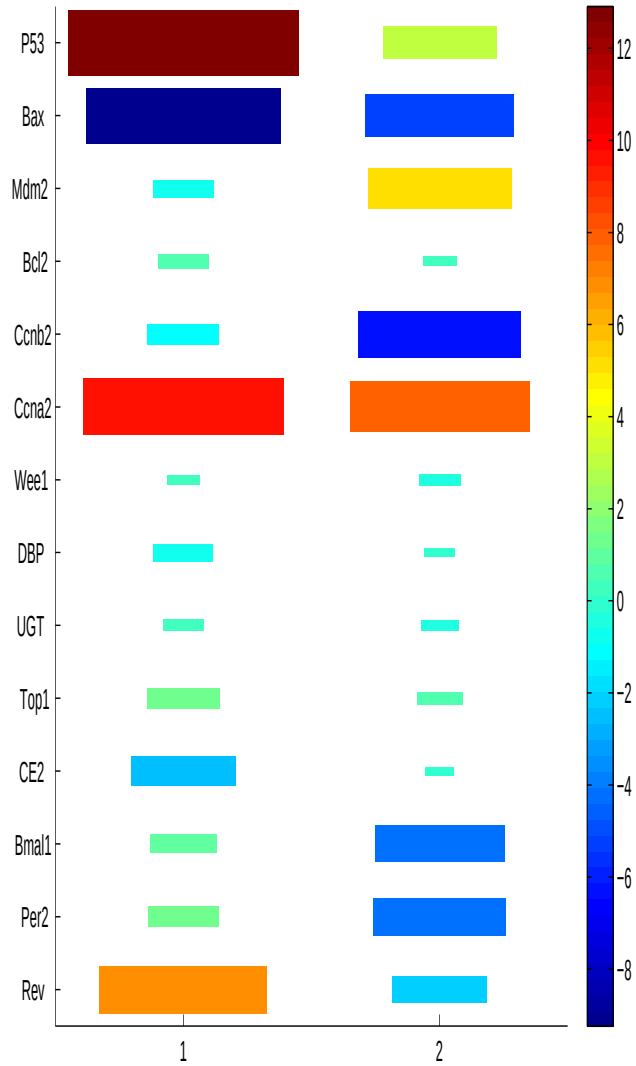


Discriminant Analysis

- 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

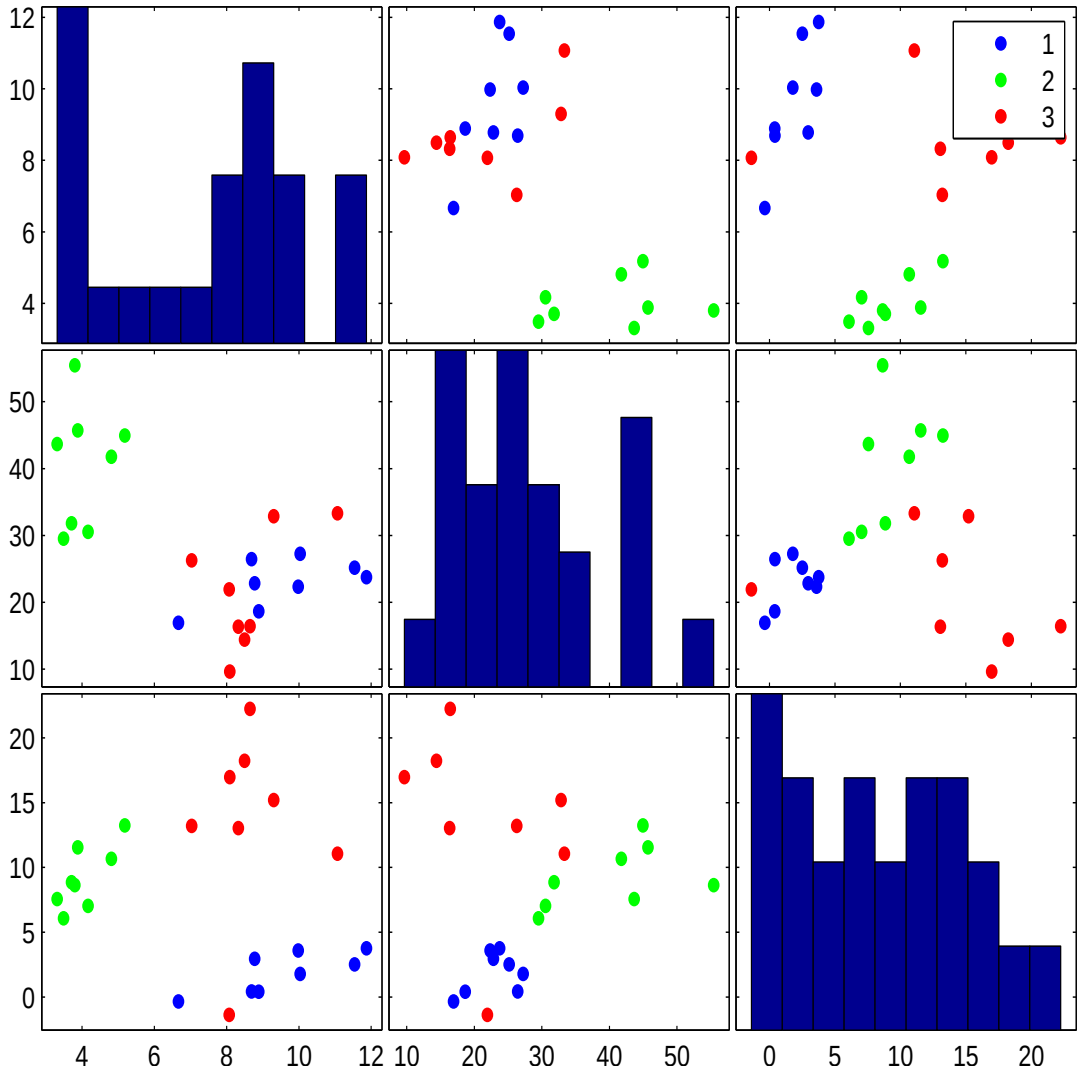


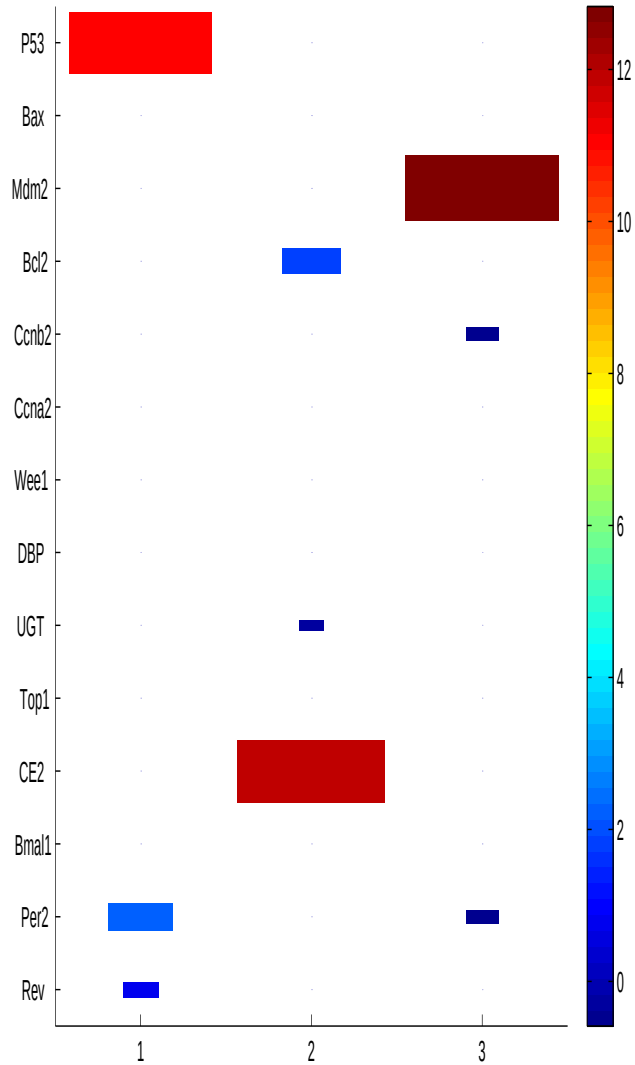


Sparse Discriminant Analysis

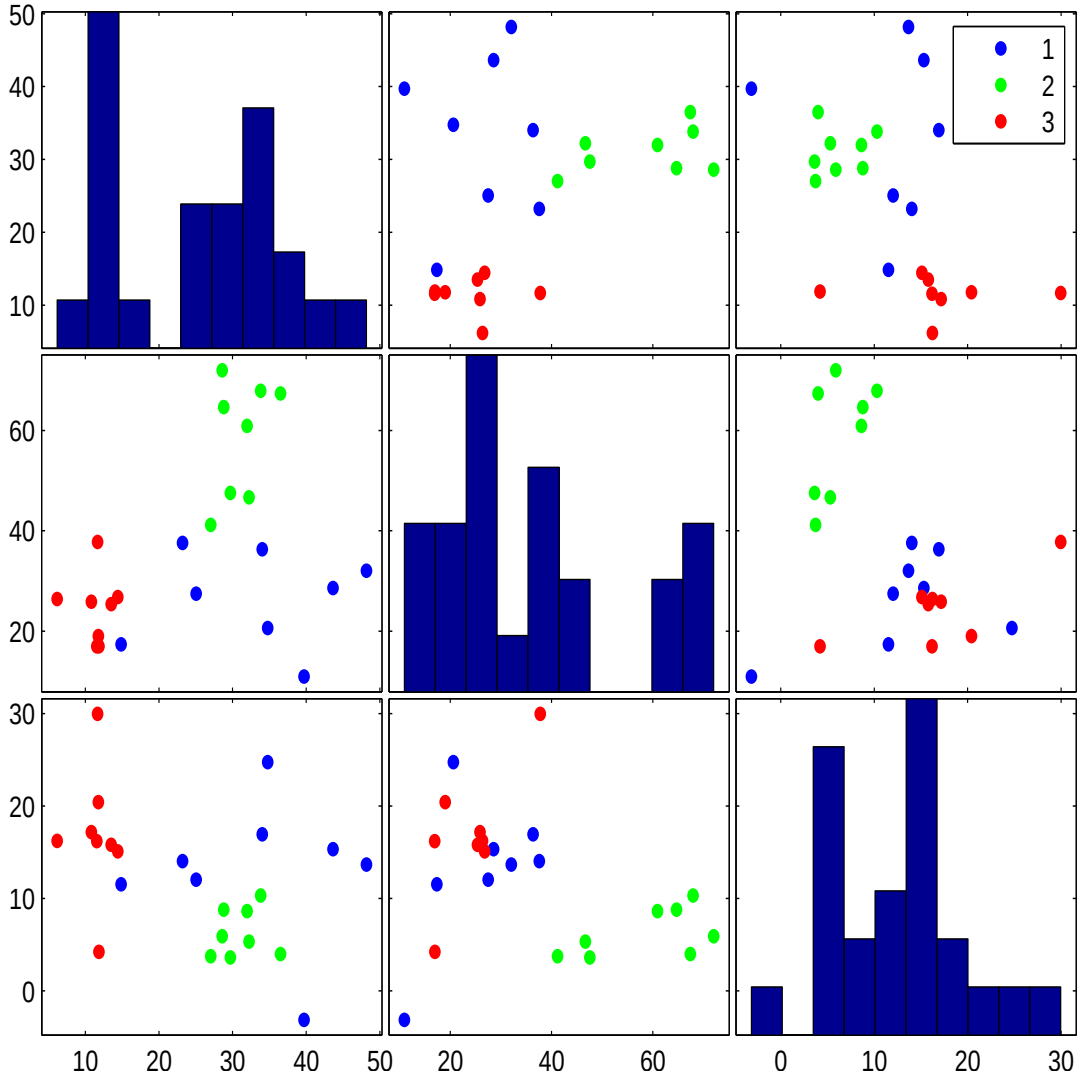
- The question to answer here is:
  What are the sparsest discriminant factors?

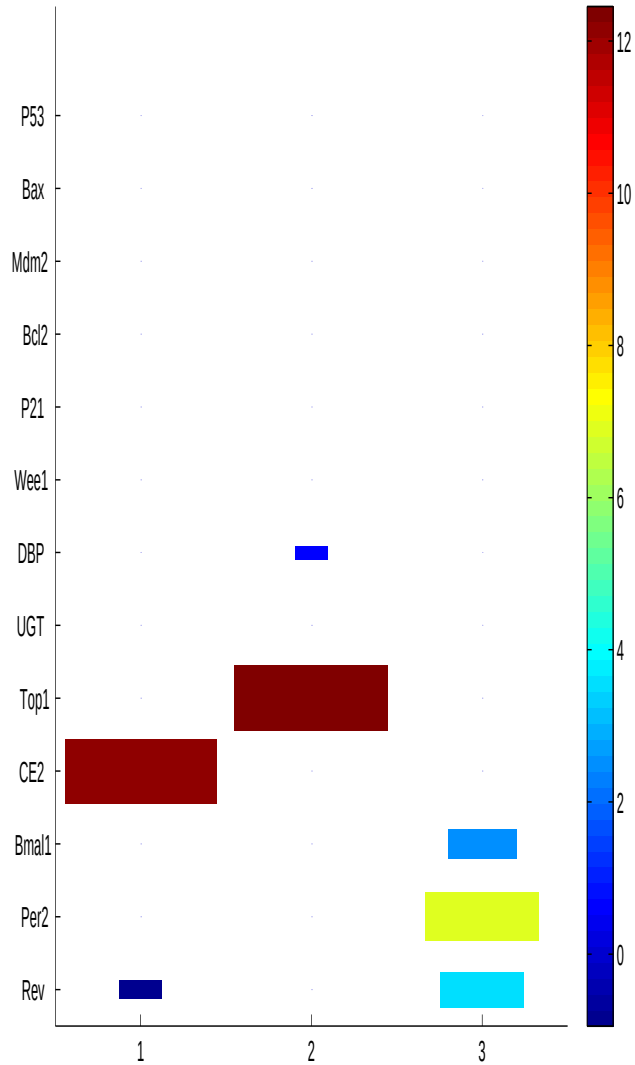
Sparse Discriminant Analysis: Time series, colon



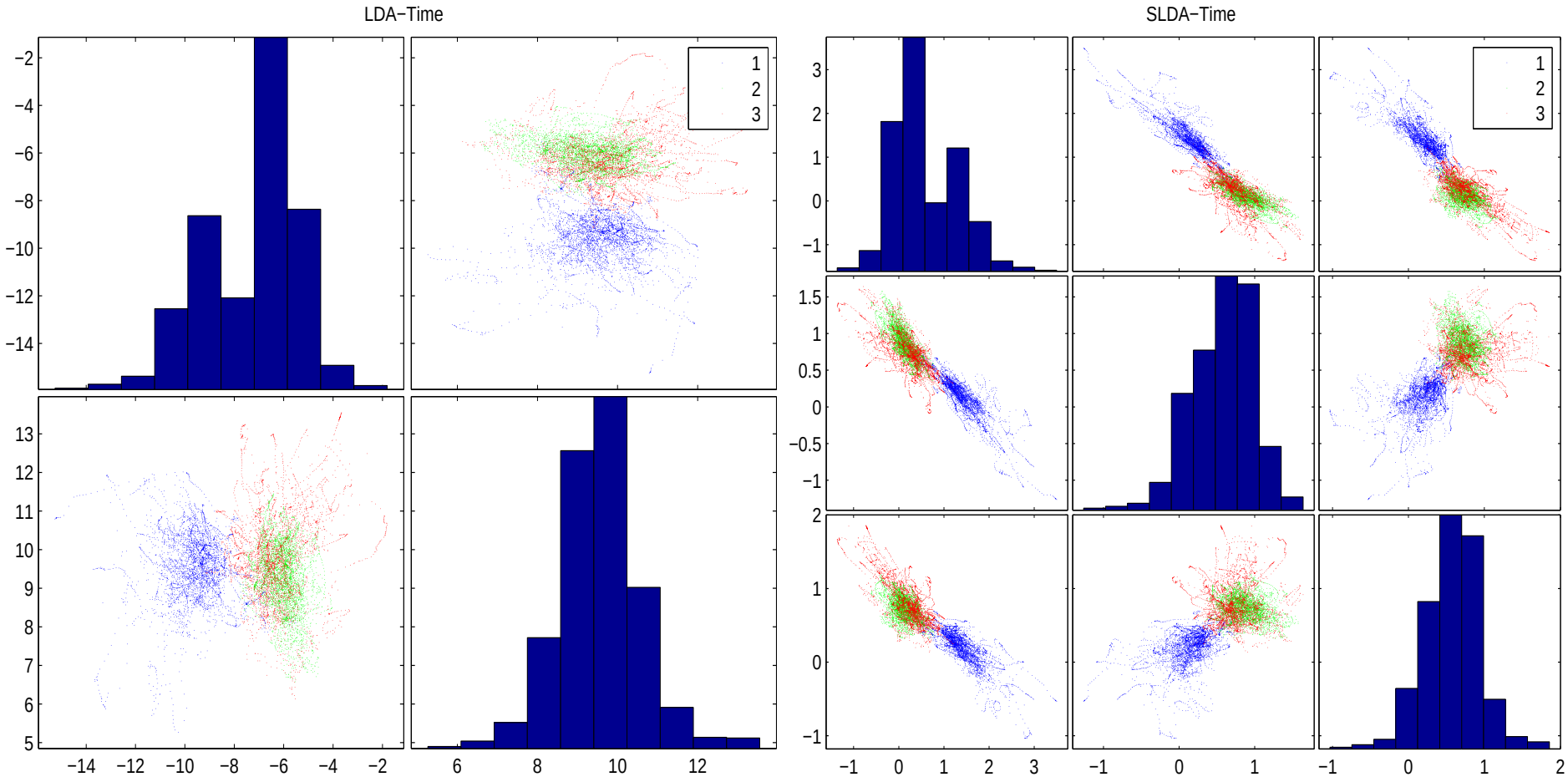


Sparse Discriminant Analysis: Time series, Liver





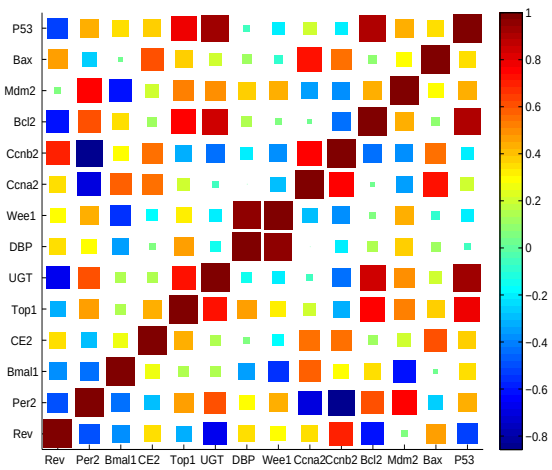
LDA and SLDA study on time serie: 1:before, 2:during, 3:after



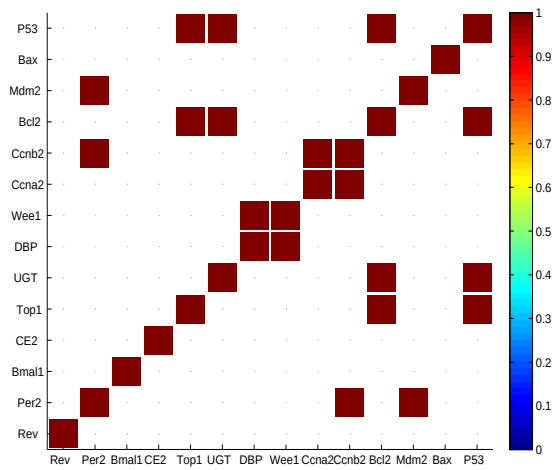
Dependancy graphs

- 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

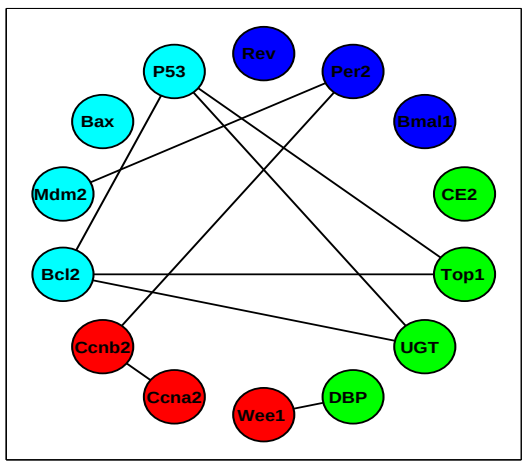
Hinton



Adjacency

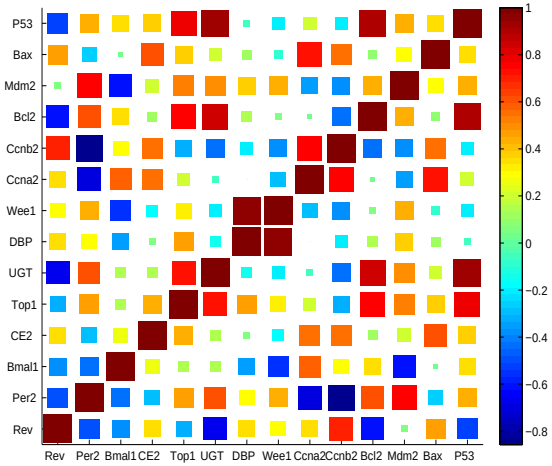


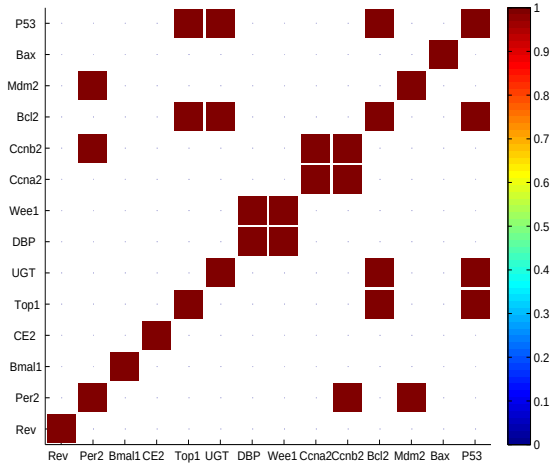
Network

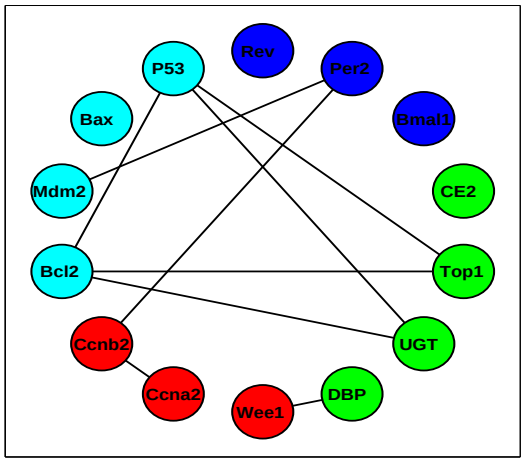


Graph of Dependancies: Colon, Class 1

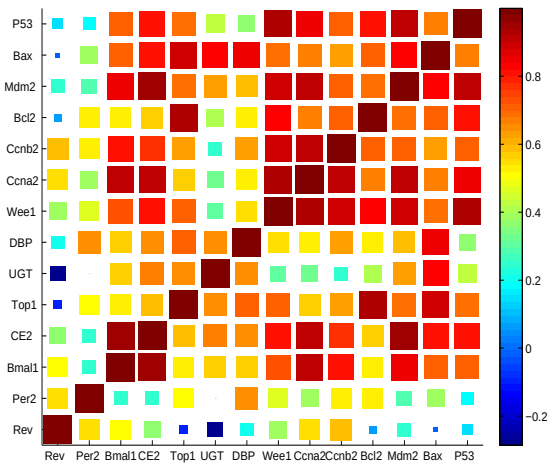
Time series

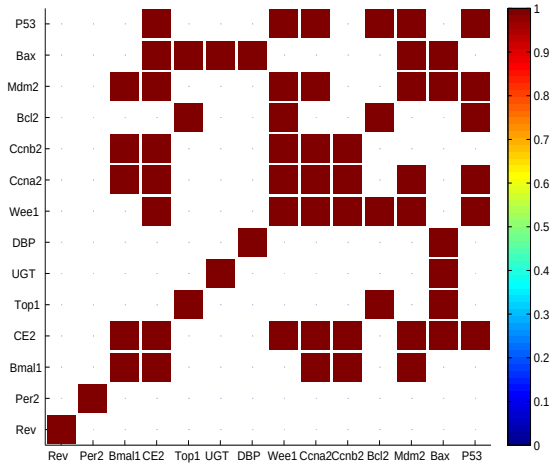


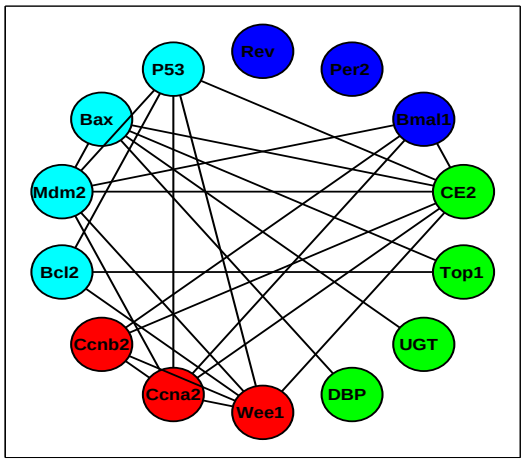




FT amplitudes

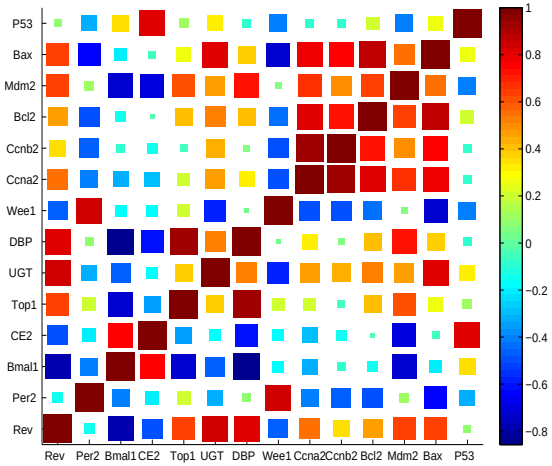


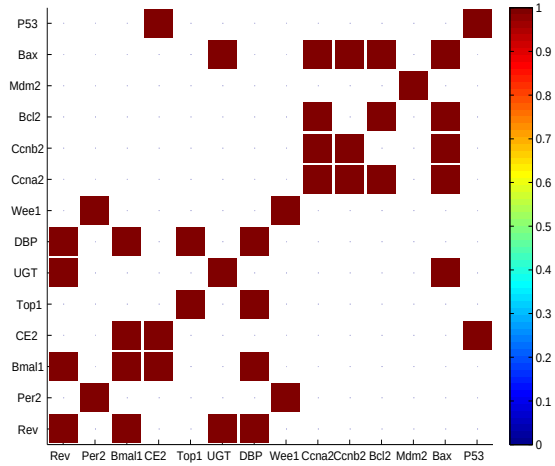


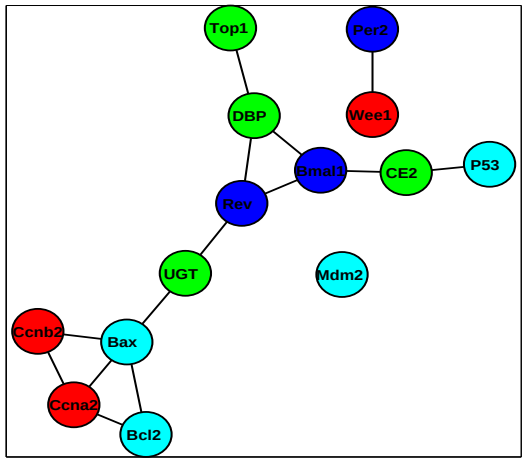


Graph of Dependancies: Colon, Class 3

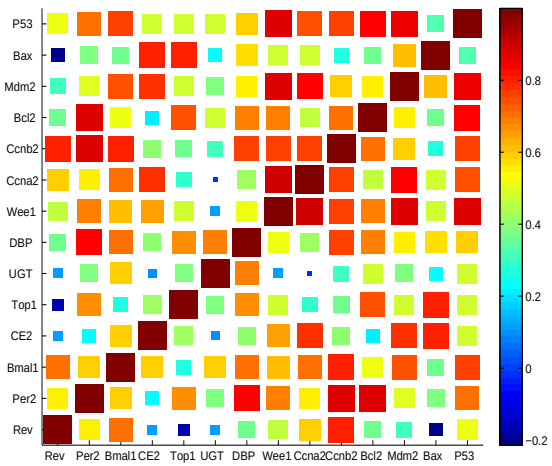
Time series

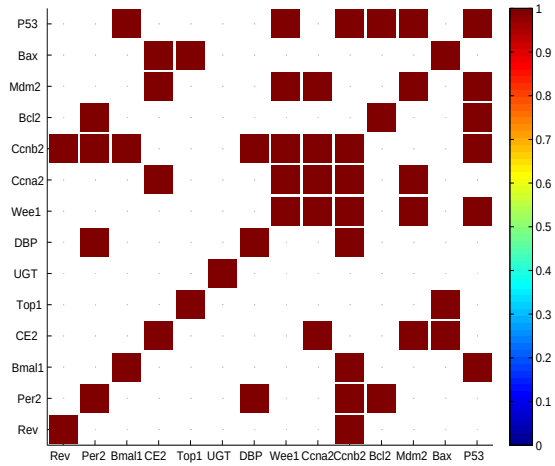


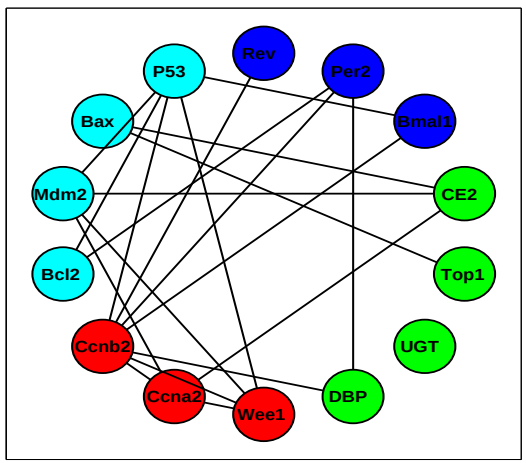




FT amplitudes



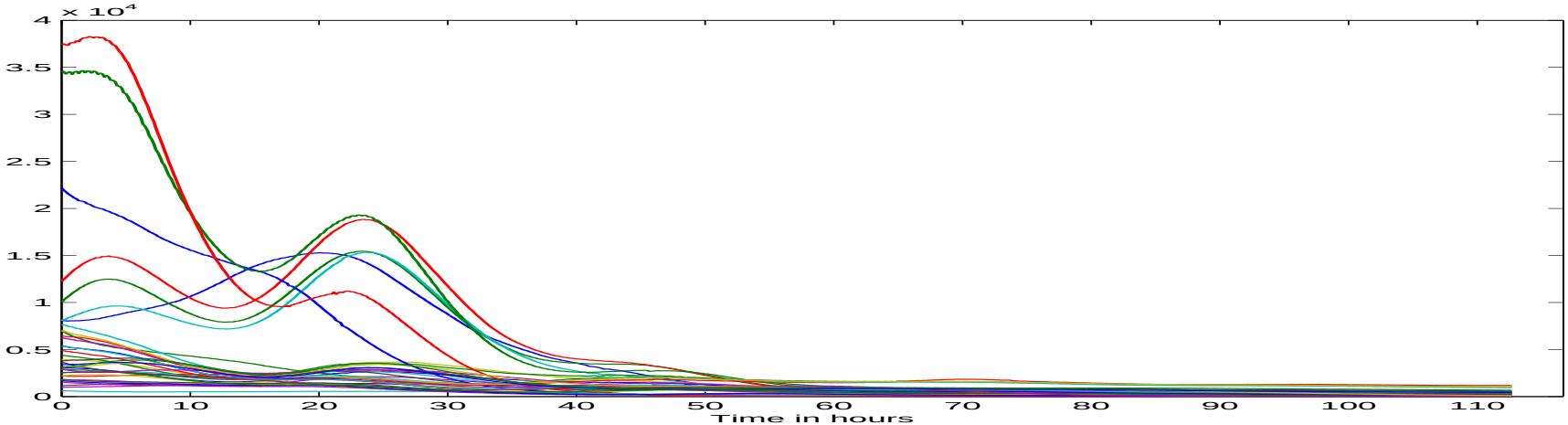


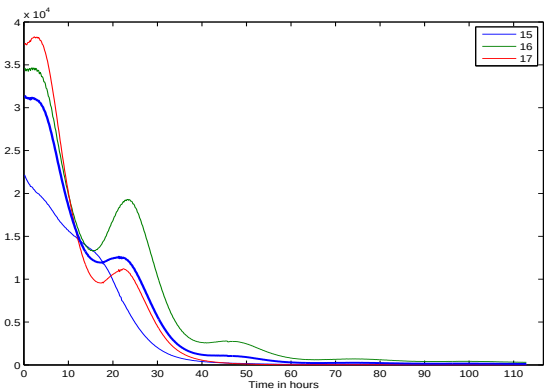


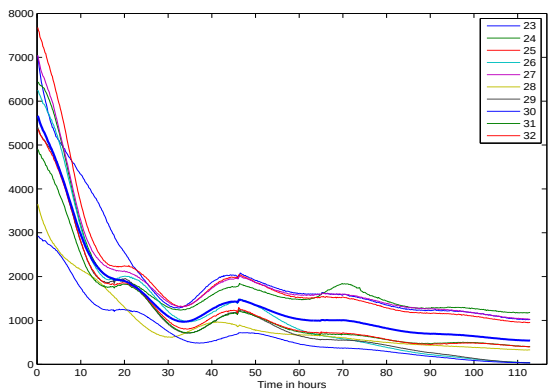
Classification tools

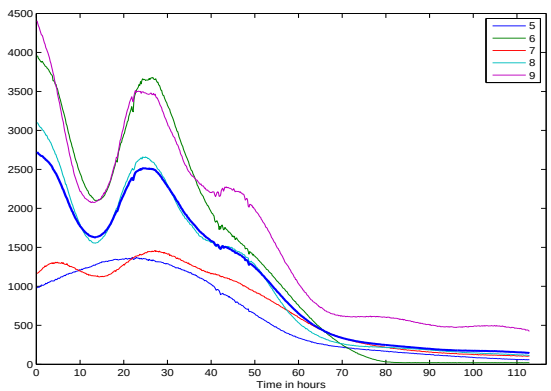
- Supervised classification
  - K nearest neighbors methods
  - Needs Training sets data
  - Must be careful to measure the performances of the classification on a different set of data (Test set)
- Unsupervised classification
  - Mixture models
  - Expectation-Maximization methods
  - Bayesian versions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

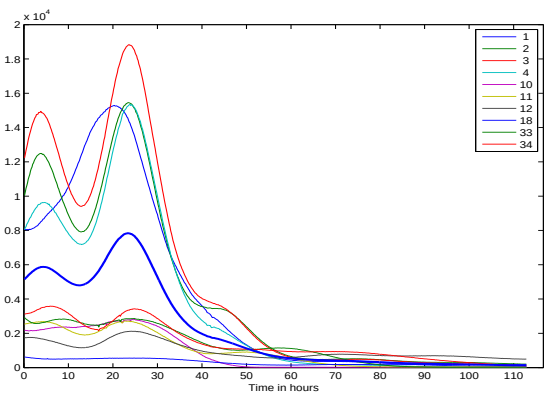
Classification tools

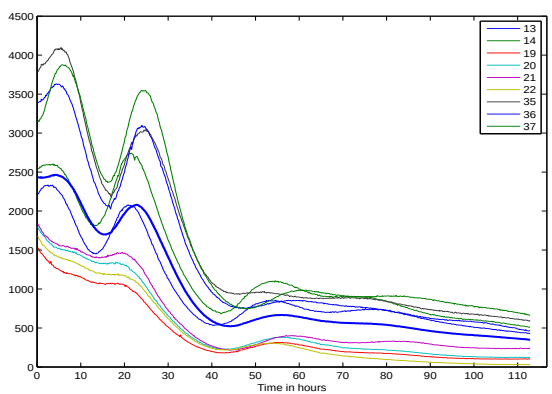












Input-Output modeling using training data and test data

- Linear models
- Bayesian framework, MAP estimation with hyperparameter estimation
- Careful identification and learning conditions
- See work of Mircea Dumitru et al for weight loss prediction from the two genes expressions of Bmal1 and Rev-erb-alpha

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
- Analyzing data before and after some clocks knockdown
- We are applying these techniques on temperature-activity data before, during and after some treatment for Chronotherapy

Future

- Application to other C5Sys data
  - Some difficulties but it will 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

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.