

Multivariate Data Analysis and Knowledge Extraction from Data

Ali Mohammad-Djafari
Jérôme Lapuyade, Ghazaleh Khodabandelou

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.

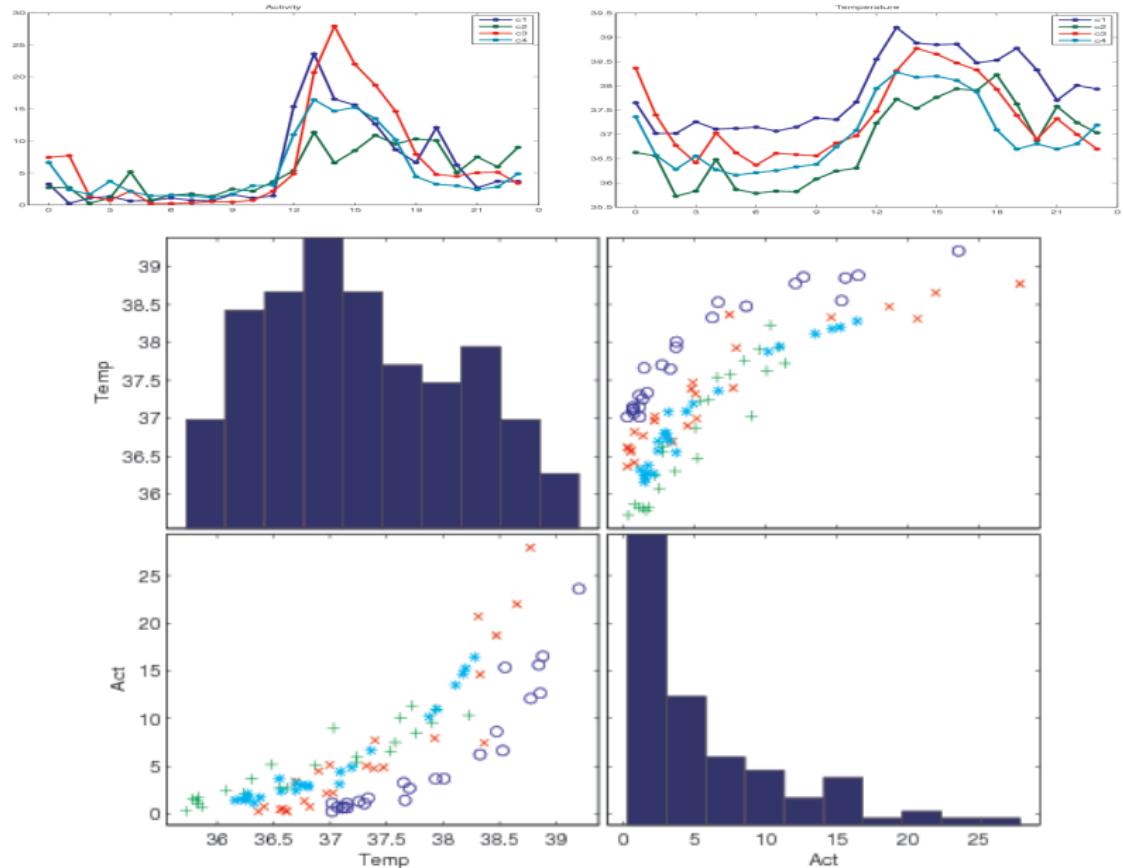
djafari@lss.supelec.fr
<http://djafari.free.fr>
<http://www.lss.supelec.fr>

Annual meeting of C5Sys_ERASysBio, Mars 30-31, 2011

Content

- ▶ Time series and their visualization
- ▶ Different Questions to Answer
- ▶ Dimensionality Reduction
- ▶ Discriminant Analysis and Classification
- ▶ Graph of links between variables
- ▶ Directed Graph of links between variables
- ▶ Questions and Discussion

Time series and their visualization

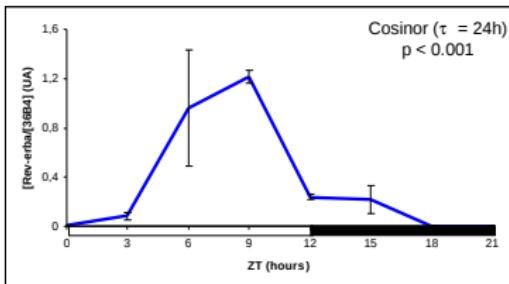


Genes expression time series

- ▶ B6D2F1 male mice synchronized with LD 12:12 for 3 weeks.
- ▶ Liver and colon mucosa sampled every 3 h for 24 h
- ▶ Circadian expression of mRNA quantified with RT-PCR:
 - ▶ Clock genes: Reverb α , Per2 and Bmal1 (liver & colon).
 - ▶ Cell cycle genes: Wee1, P21 (liver).
Wee1, Ccna2 and Ccnab2 (colon).
 - ▶ Apoptosis genes: Bcl2, Mdm2, Bax and P53 (liver & colon).

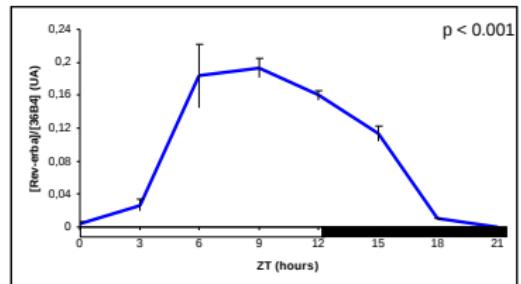
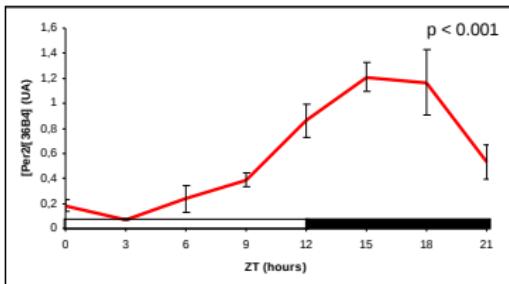
Clock

Rev-erba

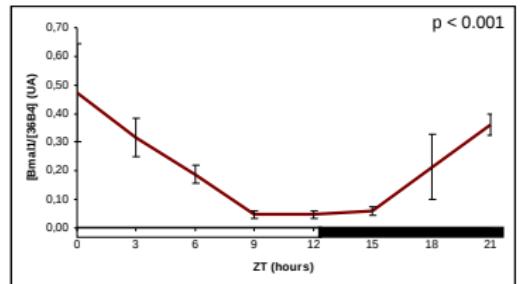
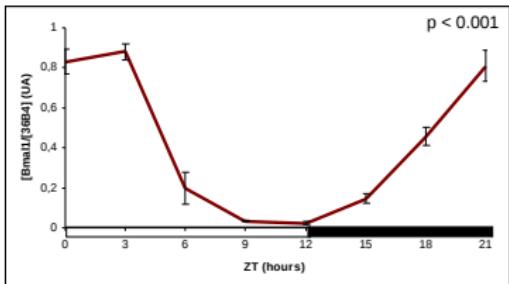


Colon

Per2

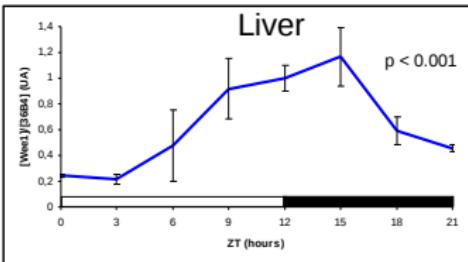


Bmal1

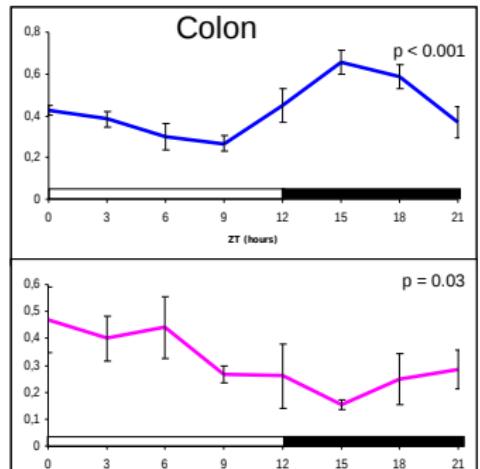


Cell Cycle

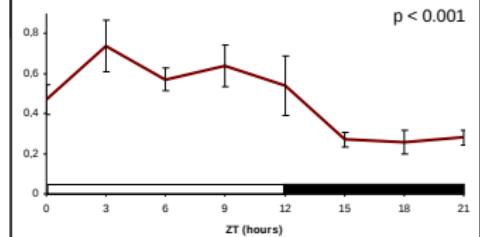
Wee1



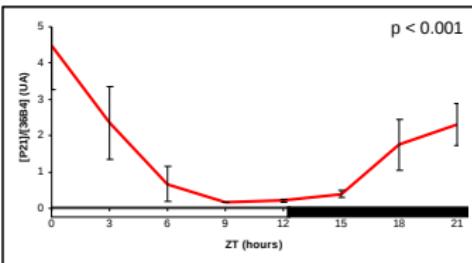
Ccna2



Ccnb2

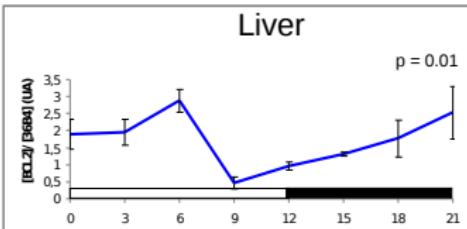


P21

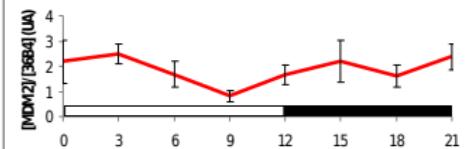


Apoptosis

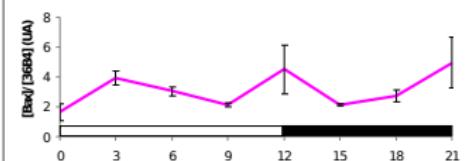
Bcl2



Mdm2



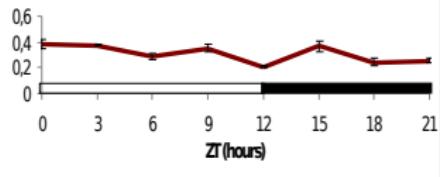
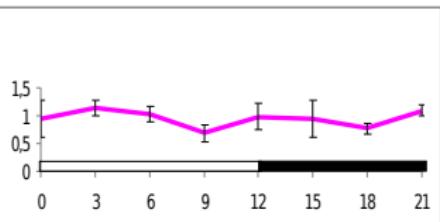
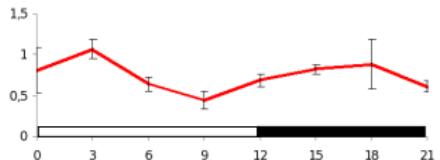
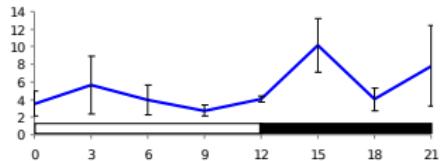
Bax



P53



Colon

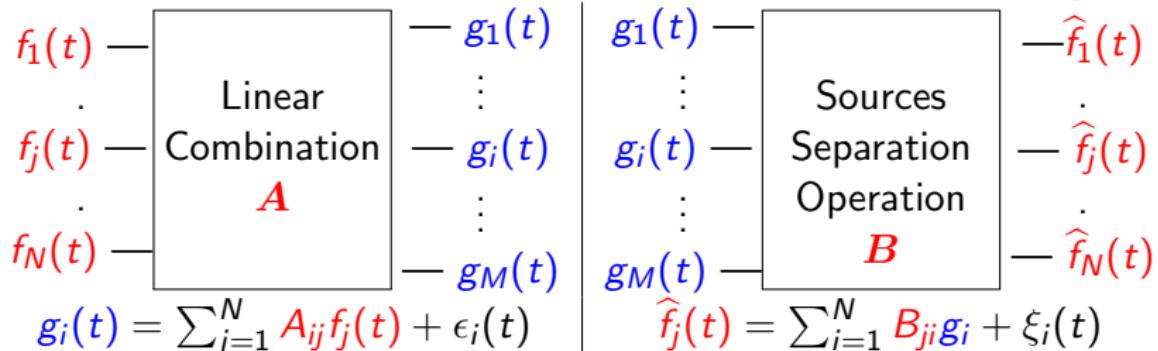


Different Questions to Answer

- ▶ What are the most important variables?
 - ▶ Principal Component Analysis (PCA) and Factor Analysis (FA)
 - ▶ Independent Component Analysis (ICA)
- ▶ What are the most discriminant variables?
 - ▶ Linear Discriminant Analysis (LDA)
 - ▶ Quadratic Discriminant Analysis (QDA), EDA, RDA, XDA, SVM, ...
 - ▶ Different classification and clustering methods (supervised or unsupervised)
- ▶ What are the most important links between variables?
 - ▶ Pearson Correlation measure ρ
 - ▶ Spearman Correlation measure ρ_s
 - ▶ Kendall Correlation measure τ
- ▶ What are the most important directed links between variables?
 - ▶ Directed Graph
 - ▶ Bayesian Network
 - ▶ Copula based Networks

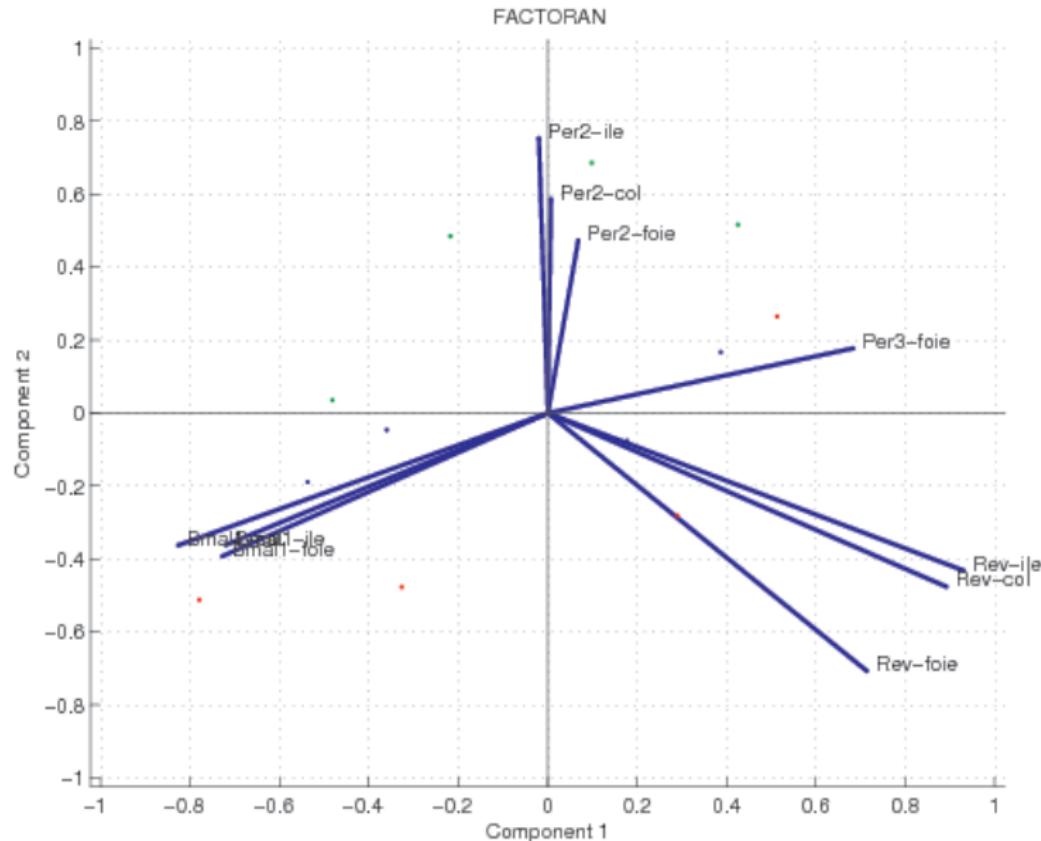
What are the most important variables?

Basic idea: the data $g_i(t)$ can be obtained by a linear combination of some basic sources (factors) $f_j(t)$

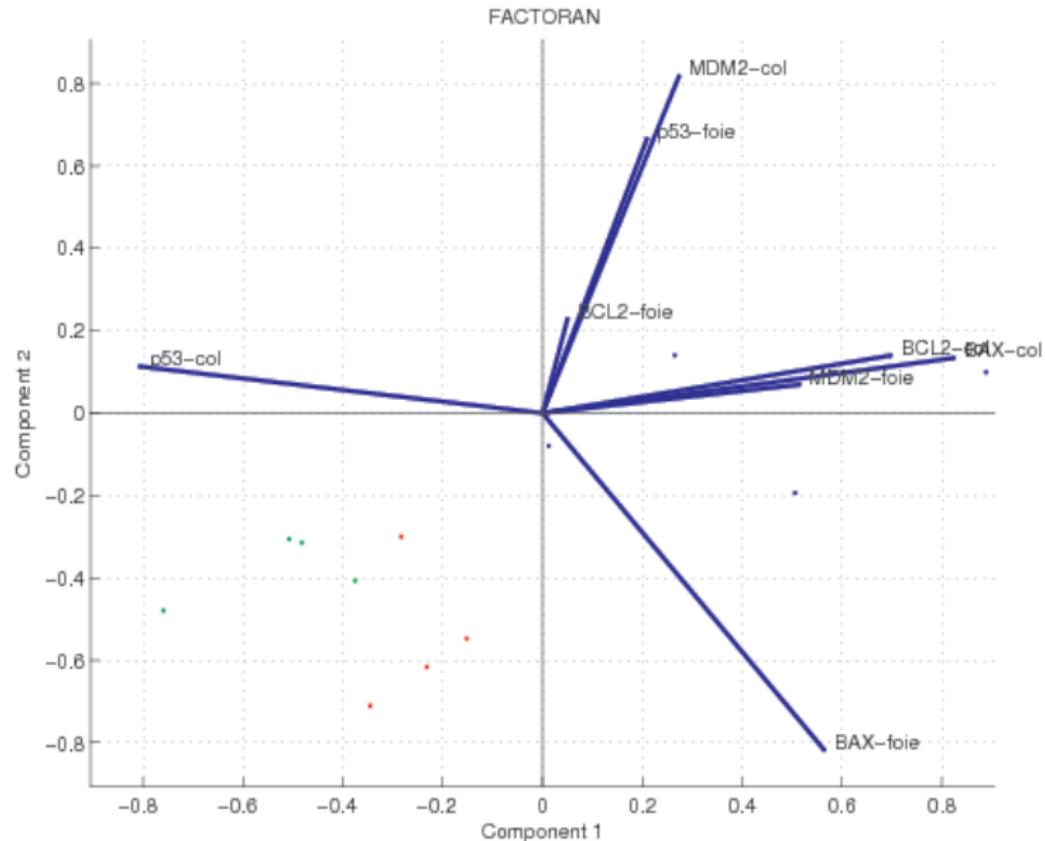


- ▶ A is called **Mapping** or **Mixing Matrix**
- ▶ Columns of A are called **Factors Loading**.
- ▶ PCA and FA are the tools to obtain: $\Rightarrow A, f(t)$
- ▶ B is called **Separation Matrix**
- ▶ ICA methods focus on determining B
- ▶ Sources Separation methods try to estimate B and $f(t)$

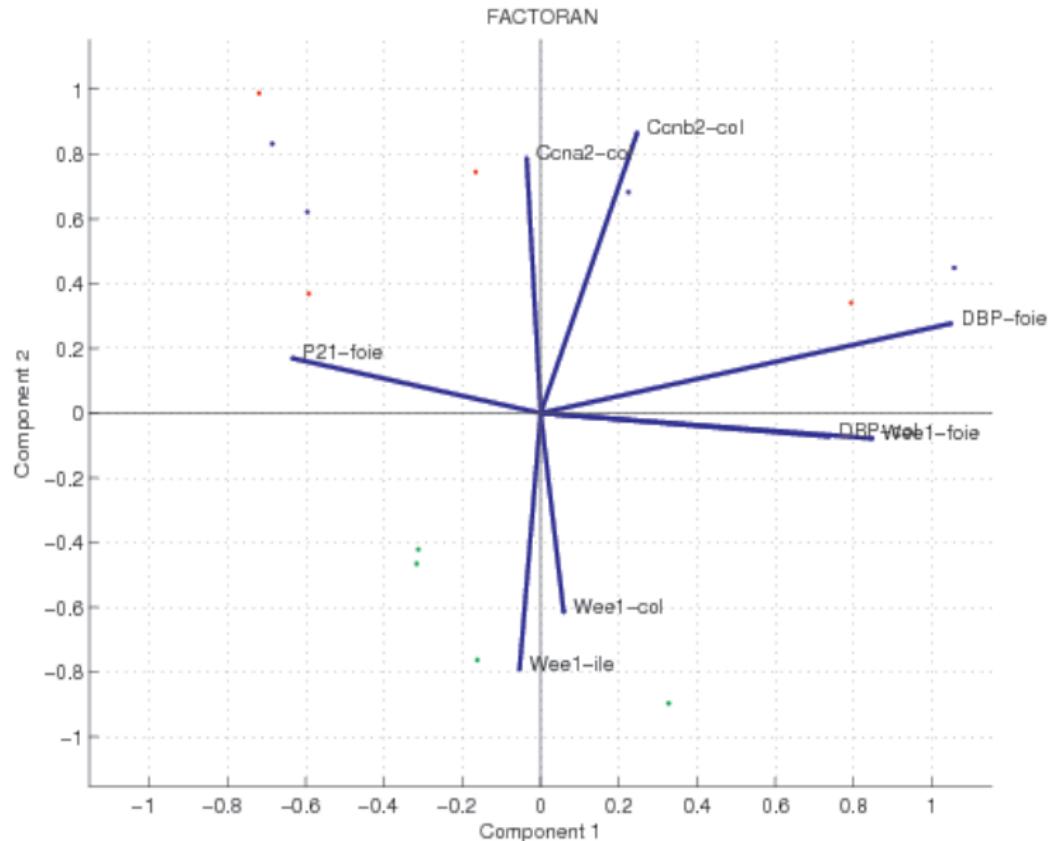
Most important Clock Genes



Most important Apoptosis Genes



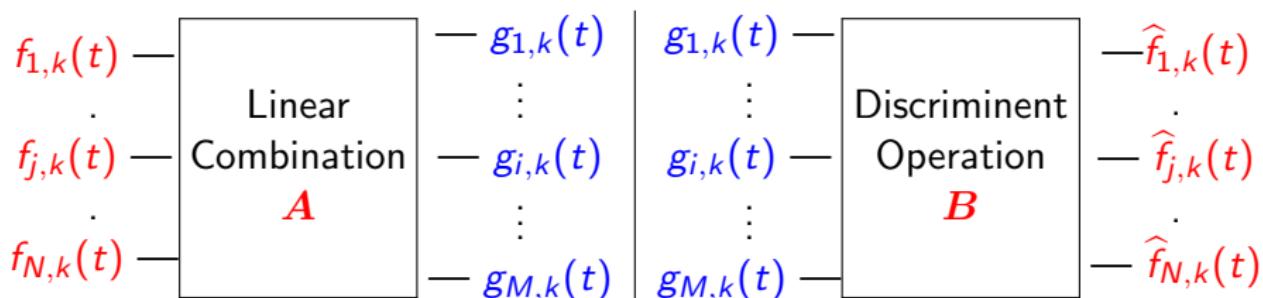
Most important Cell Cycle Genes



Discriminant Analysis

- ▶ Very often the data are gathered for different cases (classes)
- ▶ Discriminant Analysis try to find the minimum number of factors which are the most discriminant
- ▶ The data are

$$g_{i,k}(t), \quad i = 1, \dots, M, \quad k = 1, \dots, K, \quad t = 1, \dots, T$$



$$g_{i,k}(t) = \sum_{j=1}^N A_{ij} f_{j,k}(t) + \epsilon_{i,k}(t)$$

$$\hat{f}_{j,k}(t) = \sum_{i=1}^M B_{ji} g_{i,k} + \xi_{j,k}(t)$$

- ▶ Discriminant Analysis methods try to find **B**

Discriminant Analysis and Classification

- DA methods try to obtain the most discriminant factors via the matrix \mathbf{B} .
- Linear Discriminant Analysis (LDA) tries to find hyperplanes equations between any two classes. Need the inter Σ_b and intra Σ_w classes covariance matrices

$$\bar{\mathbf{g}} = \frac{1}{MK} \sum_i \sum_k \mathbf{g}_{i,k}$$

global mean

$$\bar{\mathbf{g}}_k = \frac{1}{M} \sum_i \mathbf{g}_{i,k}$$

means of each class

$$\Sigma_w = \frac{1}{MK-K} \sum_k \sum_i (\mathbf{g}_{i,k} - \bar{\mathbf{g}}_k)(\mathbf{g}_{i,k} - \bar{\mathbf{g}}_k)'$$

within

$$\Sigma_b = \frac{K}{K-1} \sum_k (\bar{\mathbf{g}}_k - \bar{\mathbf{g}})(\bar{\mathbf{g}}_k - \bar{\mathbf{g}})'$$

between

- The solution is described as

$$\hat{\mathbf{B}} = \arg \max_{\mathbf{B}} \left\{ \frac{|\mathbf{B}' \Sigma_b \mathbf{B}|}{|\mathbf{B}' \Sigma_w \mathbf{B}|} \right\}$$

which is obtained via SVD: $[\Sigma_w]^{-1} \Sigma_b \mathbf{B}_{*i} = \lambda_i \mathbf{B}_{*i}$

- Many extensions via Mixture Modelling: LDA, QDA, RDA, EDA, ...

Discriminant Analysis and Classification via Mixture Modelling

- ▶ Mixture modelling:

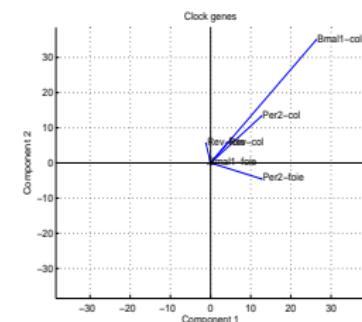
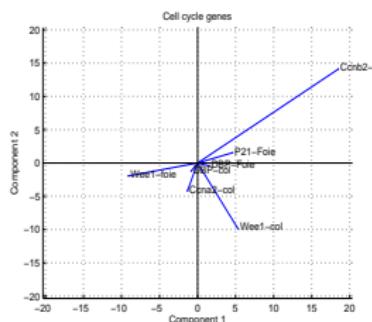
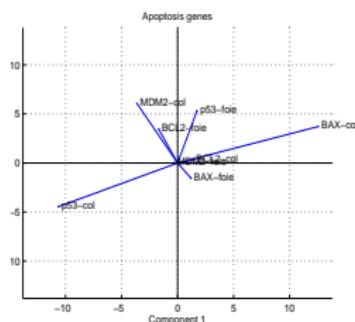
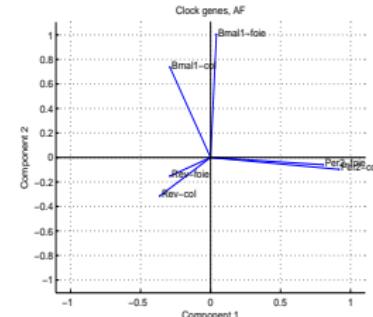
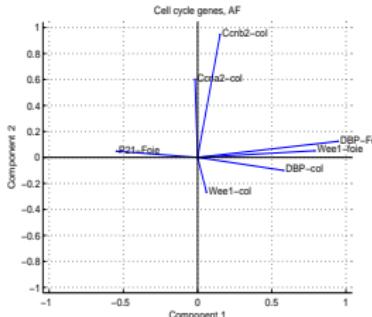
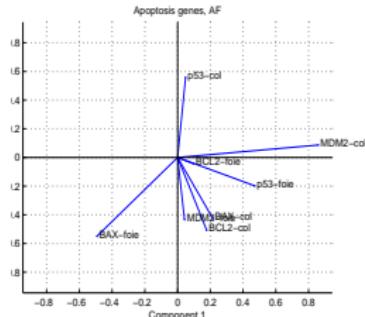
$$p(\mathbf{g}|C = k), \ p(C = k) \longrightarrow p(\mathbf{g}) = \sum_{k=1}^K p(C = k) p(\mathbf{g}|C = k)$$

- ▶ Mixture of Gaussians (MoG) modelling:

$$p(\mathbf{g}) = \sum_{k=1}^K \alpha_k \mathcal{N}(\mathbf{g}|\mathbf{m}_k, \boldsymbol{\Sigma}_k)$$

- ▶ Learning: Estimation of the parameters $\boldsymbol{\theta} = \{\alpha_k, \mathbf{m}_k, \boldsymbol{\Sigma}_k\}$
Classical or Bayesian EM algorithms
- ▶ Classification:
 - ▶ Supervised: $p(C = k|\mathbf{g}, \boldsymbol{\theta}, K)$
 - ▶ Semi-Supervised: $p(C = k|\mathbf{g}, \boldsymbol{\theta})$
 - ▶ Unsupervised: $p(C = k|\mathbf{g})$

Factor Analysis Vs. Discriminant Analysis



Most important Dependency Graph between variables

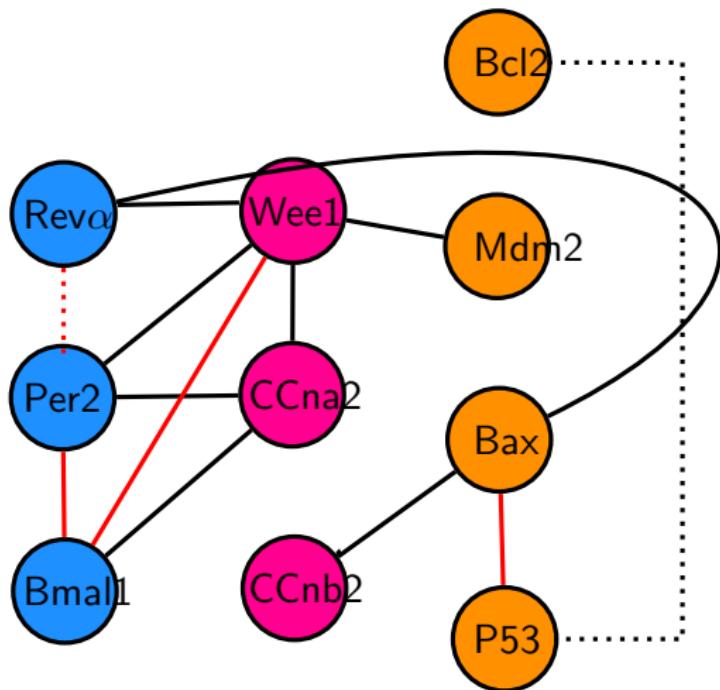
- ▶ Question: How to measure the dependency between two variables X and Y ?
- ▶ Classical Pearson correlation ρ

$$\rho = \frac{E\{X\} E\{Y\}}{\sqrt{E\{(X - m_x)^2\} E\{(Y - m_y)^2\}}}$$

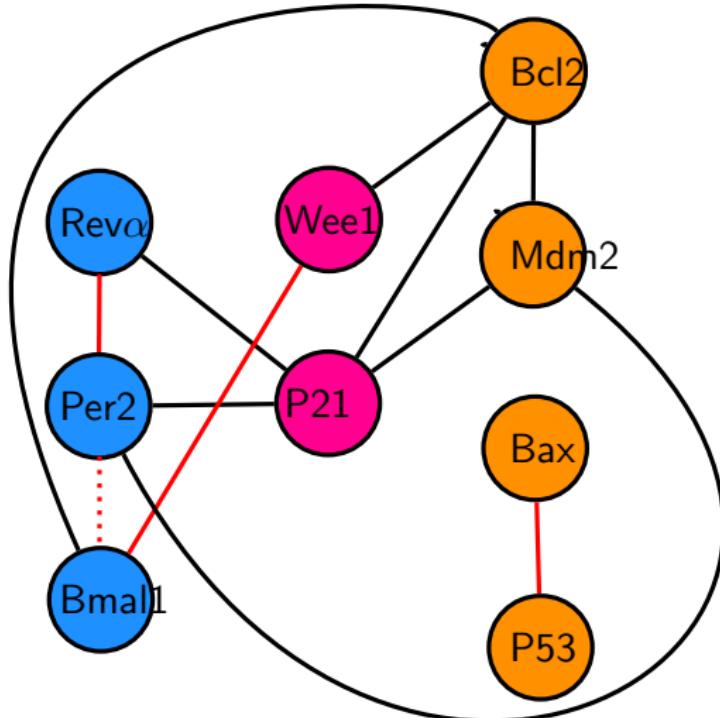
measures only the linear dependency

- ▶ Two other measures are: Spearman ρ_s and Kendall τ
- ▶ When $\rho = 0$ does not mean that the variables are not dependent !
- ▶ We used a combination of these measures to create a graph of dependencies between the variables.
- ▶ We followed two directions:
 - ▶ For each tissue (Liver & Col) and each class, create separate graphs and then combine the results (Decision fusion)
 - ▶ Use the whole data and create one common graph

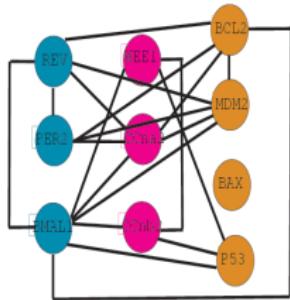
Directed Graph of links between variables



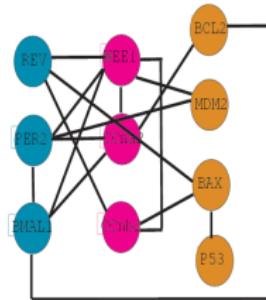
Directed Graph of links between variables



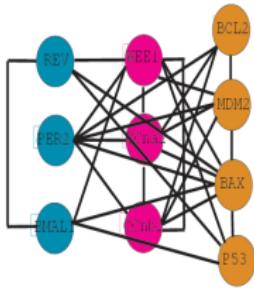
Most important Dependency Graph between variables



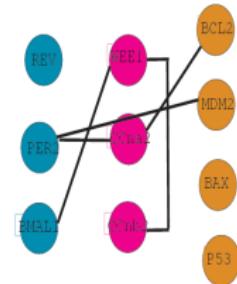
Class 1



Class 2

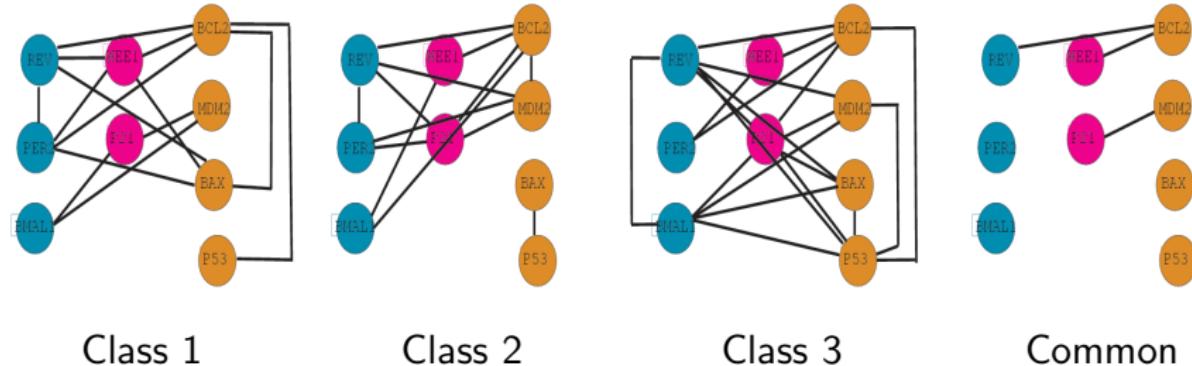


Class 3



Common

Most important Dependency Graph between variables



Most important Dependency Graph between variables



Class 1

Class 2

Class 3

Directed Graph of links between variables

- ▶ Question: Which genes are affecting the others ?
- ▶ Directed graphs
- ▶ Bayesian networks
- ▶ Needs to have more data
- ▶ We plan to have more dense time series data (1h) and longer (48h), in normal and blocked genes situations.
- ▶ Hope to be able to answer the question: which variables are the causes and which expressions are the effects

Questions and Discussion

