Factor Analysis for Multiple Testing : an R package for large-scale significance testing under dependence

Maela Kloareg, Chloé Friguet & David Causeur

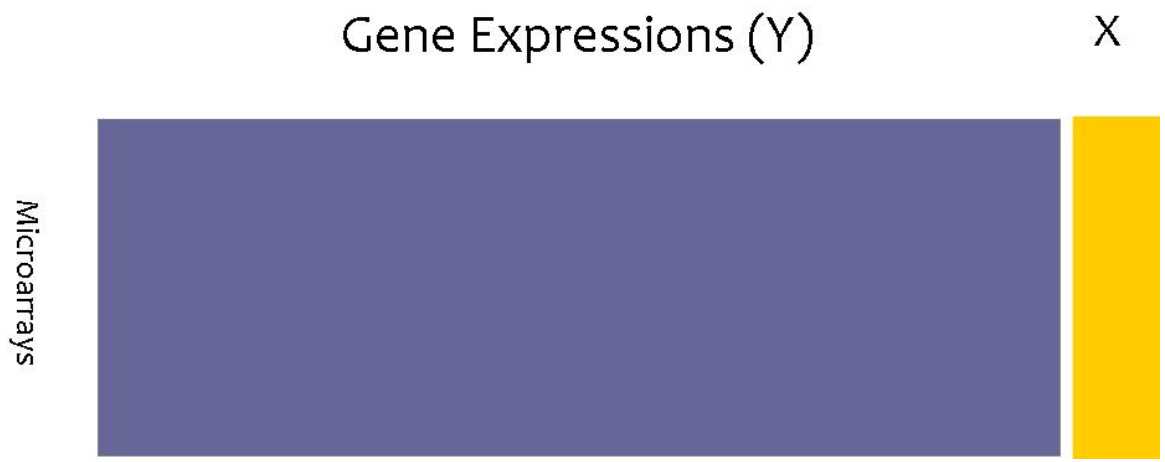
Applied mathematics department
Agrocampus Ouest, Université Européenne de Bretagne

The UseR! Conference, July 2009
Agrocampus Ouest, France

Outline

Impact of dependence in multiple testing

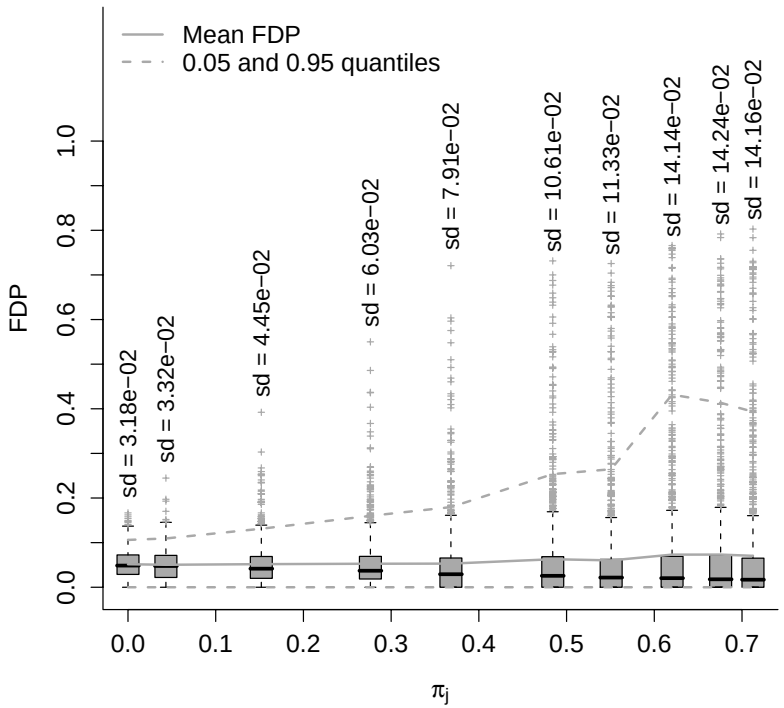
Multiple testing: to point out genes which expressions (Y) significantly depend on the experimental condition (X)
High dimension: a few microarrays and a huge number of gene expressions



A major concern: the biological links among genes and the high dimensional setting generates a large-scale correlation structure, which induces high instability in multiple testing procedures.

Distribution of error rates in multiple tests

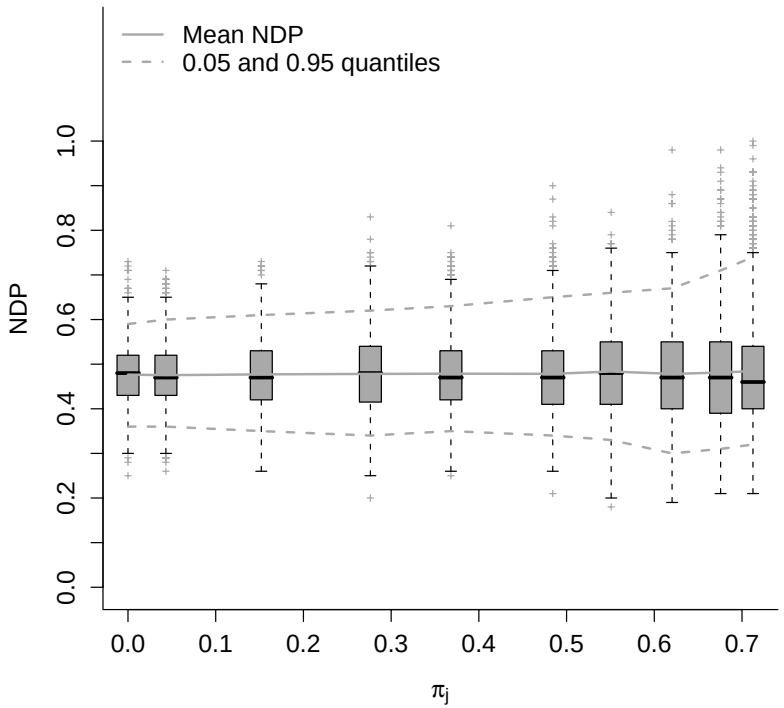
Distribution of False Discovery Proportion (V_t/R_t) on 1.000 simulated datasets/scenario (Friguet *et al.*, 2009, *JASA*)



	Declared H0	*Declared H1*	*Total*
True H0	$\mathbf{U_t}$	$\mathbf{V_t}$	*m0*
True H1	$\mathbf{T_t}$	$\mathbf{S_t}$	*m1*
	$m\text{-}R_t$	R_t	*m*

Distribution of error rates in multiple tests

Distribution of Non-Discovery Proportion ($T_t/m1$) on 1.000 simulated datasets/scenario (Friguet *et al.*, 2009, *JASA*)



	Declared H0	*Declared H1*	*Total*
True H0	$\mathbf{U_t}$	$\mathbf{V_t}$	*m0*
True H1	$\mathbf{T_t}$	$\mathbf{S_t}$	*m1*
	$m\text{-}R_t$	R_t	*m*

Outline

Factor Analysis for Multiple Testing

The common information shared by all the variables (m) is modeled by a factor analysis structure.
The common factors Z : small number ($q << m$) of latent variables (Friguet *et al.*, 2009, *JASA*)

Specific variability (uniqueness) Common variability

$$\Sigma = \Psi + BB'$$

$$Y^{(k)} = \beta_0^{(k)} + x'\beta^{(k)} + BZ + \varepsilon^{(k)}$$

Common factors

$$Z \sim N\left(0; I_q\right), V(\varepsilon) = \Psi$$

Factor Analysis for Multiple Testing

The common information shared by all the variables (m) is modeled by a factor analysis structure.
The common factors Z : small number ($q << m$) of latent variables (Friguet *et al.*, 2009, *JASA*)

Similar idea : Surrogate Variable Analysis method, Leek and Storey, 2007, 2008.

Factor-adjusted test statistics

The adjusted test statistics are conditionally centered and scaled version of usual test statistics

Conditional distribution of the usual test statistic $T^{(k)}$

$$\mathbb{E}(T^{(k)} \mid Z) = \tau_k + \frac{b_k'}{\sigma_k}\tau(Z), \quad \mathrm{Var}(T^{(k)} \mid Z) = \frac{\psi_k^2}{\sigma_k^2}.$$

Conditional centering and scaling

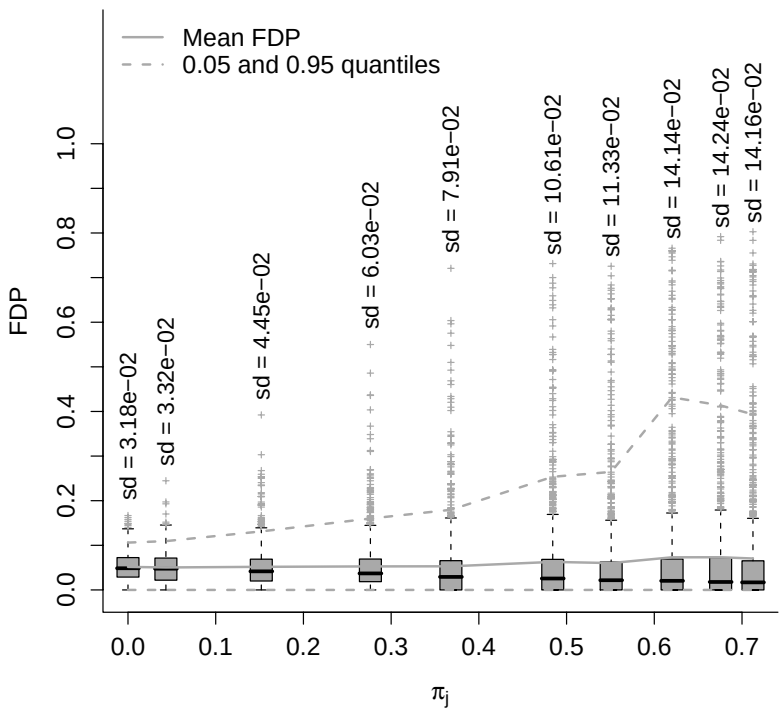
$$T_z^{(k)} = \frac{\sigma_k}{\psi_k}\left[T^{(k)} - \frac{b_k'}{\sigma_k}\tau(Z)\right].$$

with $\mathbb{E}(T_z^{(k)}) = \frac{\tau_k}{\sqrt{1-h_k^2}}$ and $\mathrm{Var}(T_z) = I_m$.

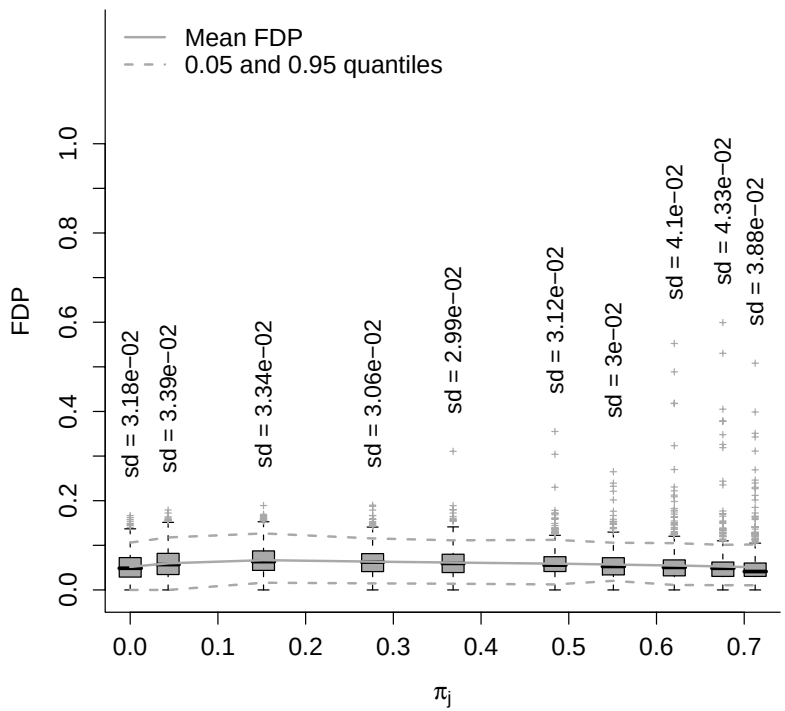
Distribution of error rates in multiple tests

Distribution of False Discovery Proportion on 1.000 simulated datasets/scenario (Friguet *et al.*, 2009, *JASA*)

Usual t-tests



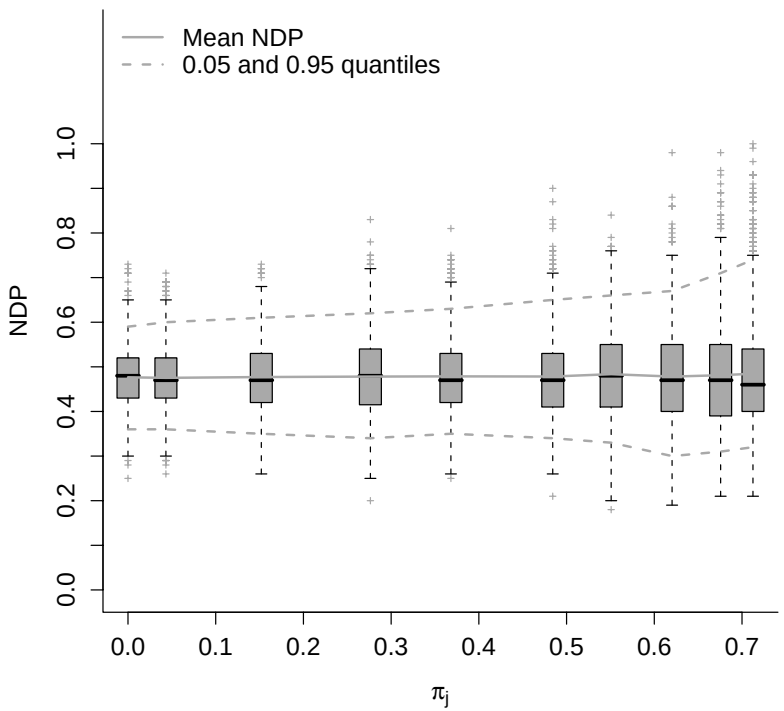
Factor-adjusted t-tests



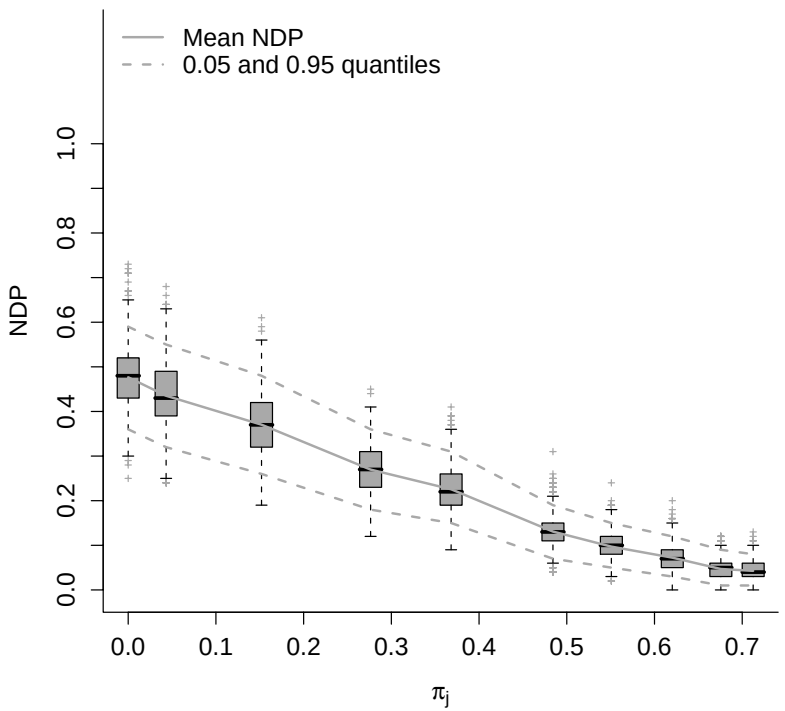
Distribution of error rates in multiple tests

Distribution of Non-Discovery Proportion on 1.000 simulated datasets/scenario (Friguet *et al.*, 2009, *JASA*)

Usual t-tests



Factor-adjusted t-tests



Outline

The FAMT package steps

1. Estimation of the number of factors
2. Factor Analysis model (using $\widehat{\mathcal{M}_0} = \{k,\ P_k \geq \alpha\}$)
3. Multiple testing : conditional statistics and p-values
   $\widehat{\mathcal{M}_0}$ updated, step 1 to 3 are done twice
4. Estimation of the proportion of null hypotheses
5. Benjamini and Hochberg's procedure to control the FDR

The FAMT package steps

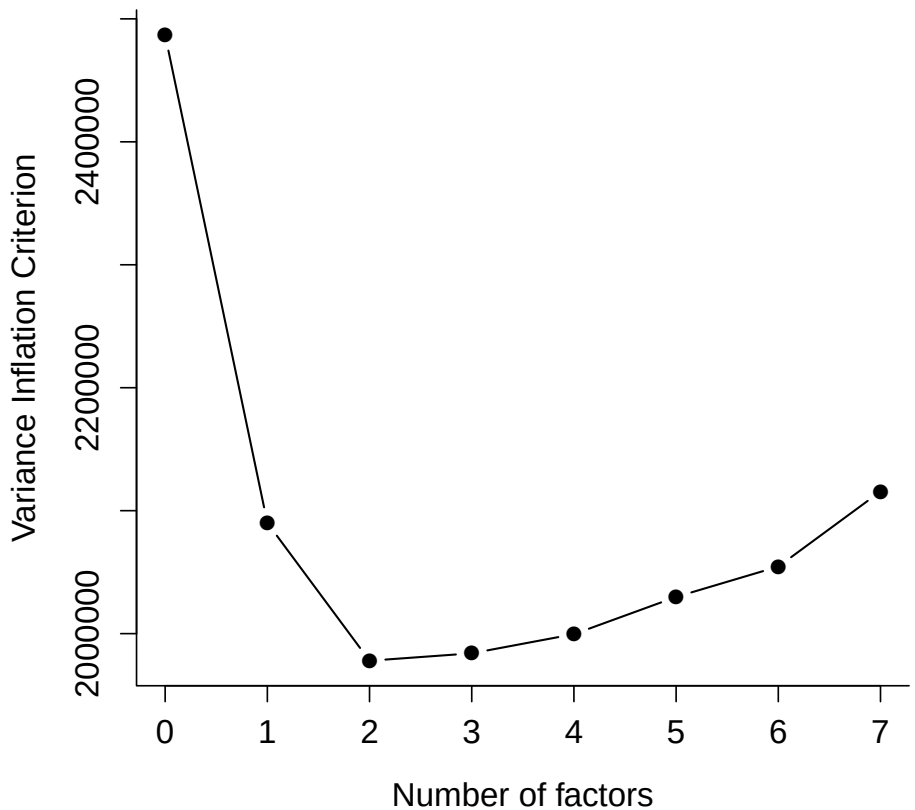
1. Estimation of the number of factors
2. Factor Analysis model (using $\widehat{\mathcal{M}}_0 = \{k,\ P_k \geq \alpha\}$)
3. Multiple testing : conditional statistics and p-values
   $\widehat{\mathcal{M}}_0$ updated, step 1 to 3 are done twice
4. Estimation of the proportion of null hypotheses
5. Benjamini and Hochberg's procedure to control the FDR

Illustration on the Lymphoma dataset (Alizadeh *et al.* 2000)

- 32 samples : 2 classes of B cell-like diffuse large cell lymphoma (DLCL) : germinal center B cell-like DLCL (18 samples) and active B cell-like DLCL (14 samples)
- Expression levels of 10295 genes

1/ Estimation of the number of factors

The number of factors is chosen to reduce the variance of the number of false positives in multiple tests.



2/ Factor Analysis model

To deal with high-dimension, the model parameters are estimated with an EM-algorithm (Rubin and Thayer, 1982) :

- E step : estimation of Z
- M step : estimation of B and Ψ

Specific variability (uniqueness) Common variability

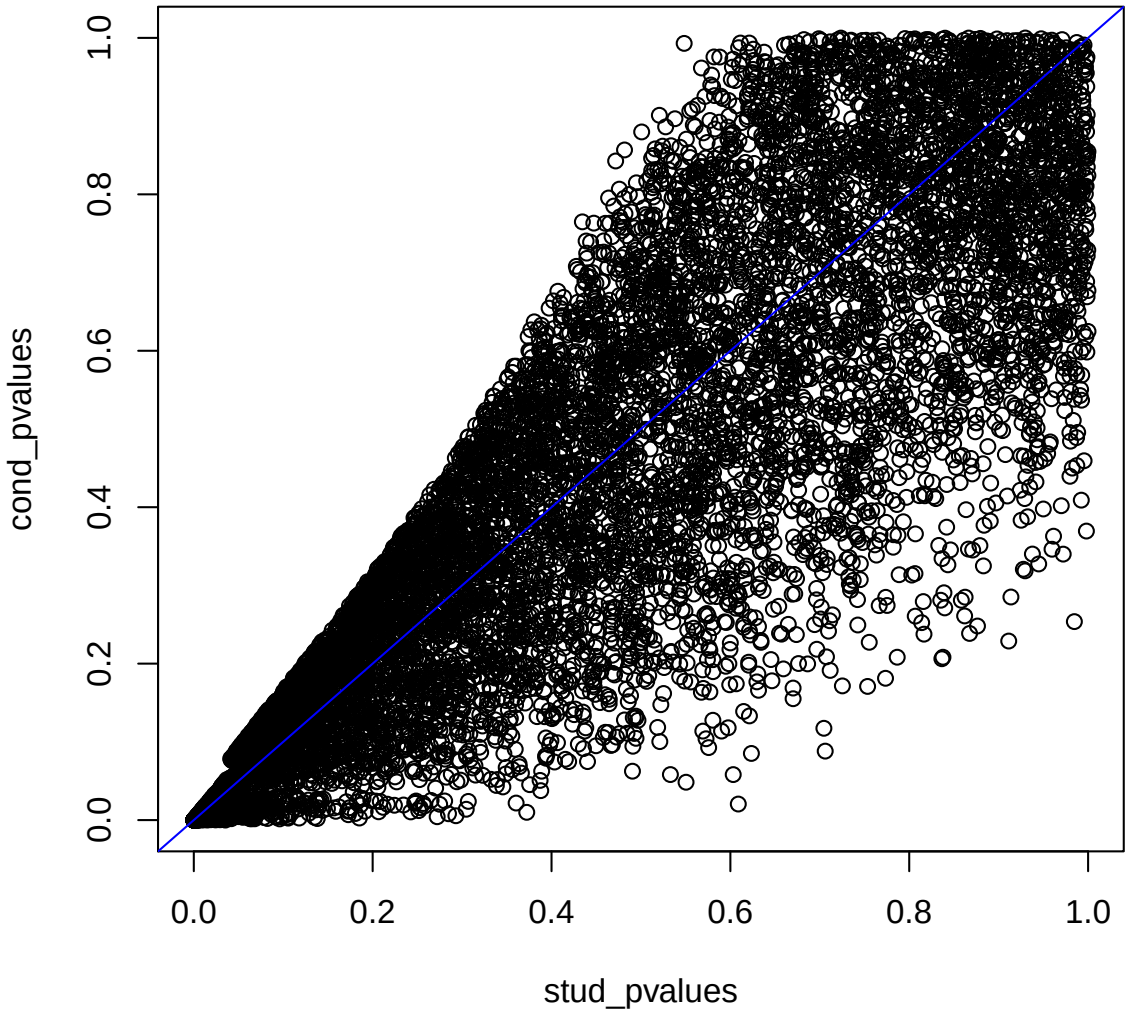
$$\Sigma = \Psi + BB'$$

$$Y^{(k)} = \beta_0^{(k)} + x'\beta^{(k)} + BZ + \varepsilon^{(k)}$$

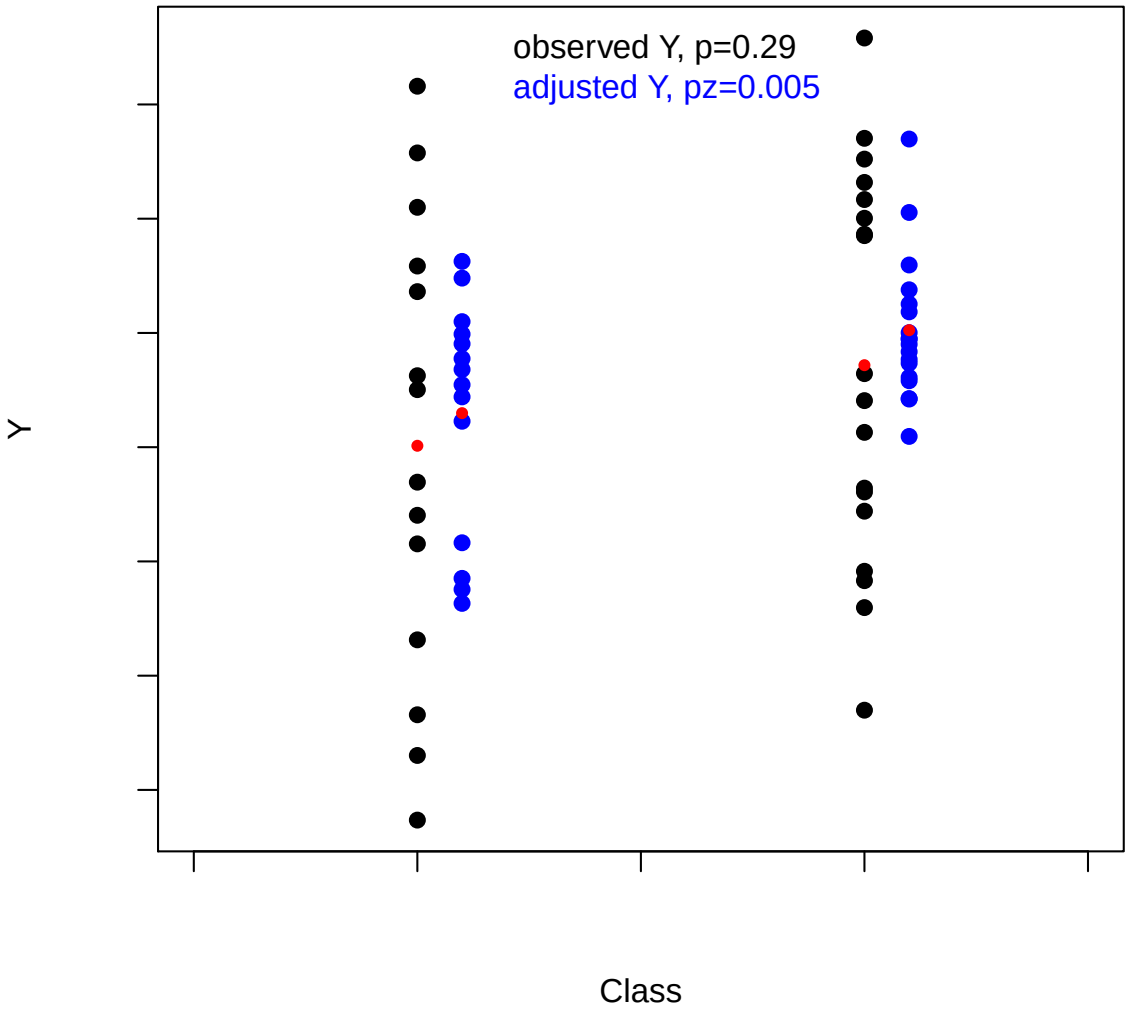
Common factors

$$Z \sim N(0; I_q), V(\varepsilon) = \Psi$$

3/ Multiple testing (conditional p-values)



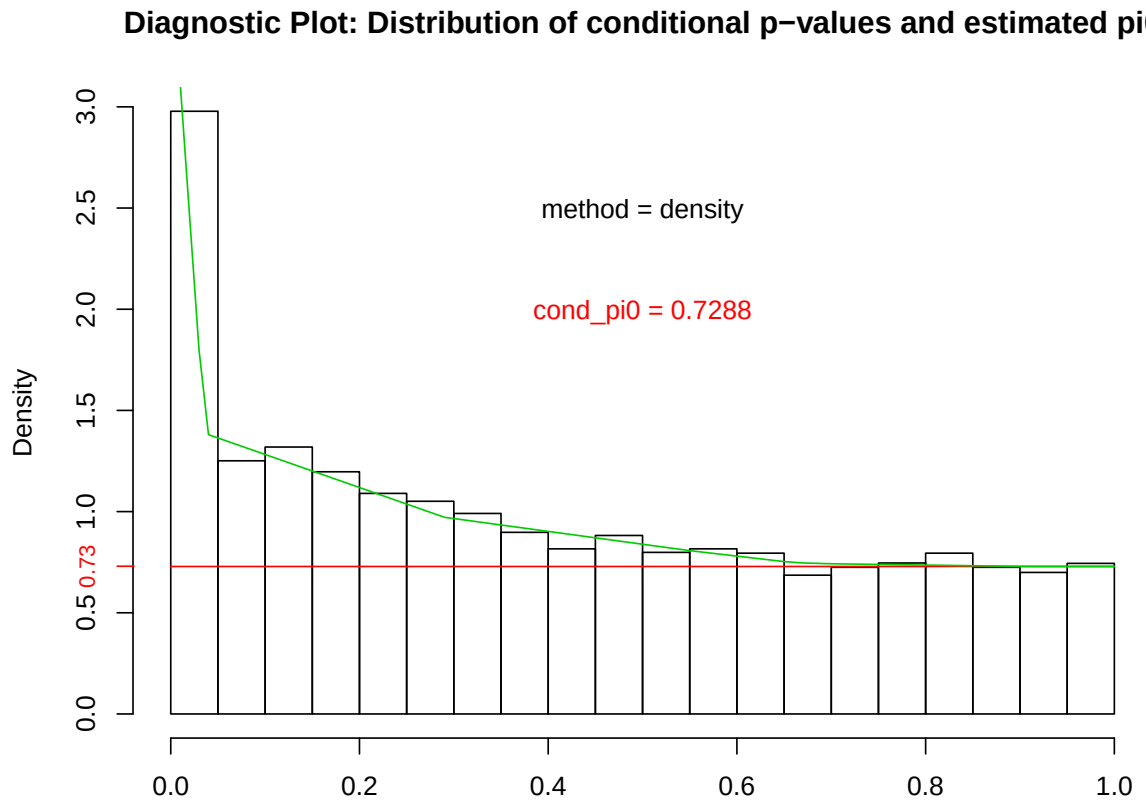
3/ Multiple testing (conditional p-values)



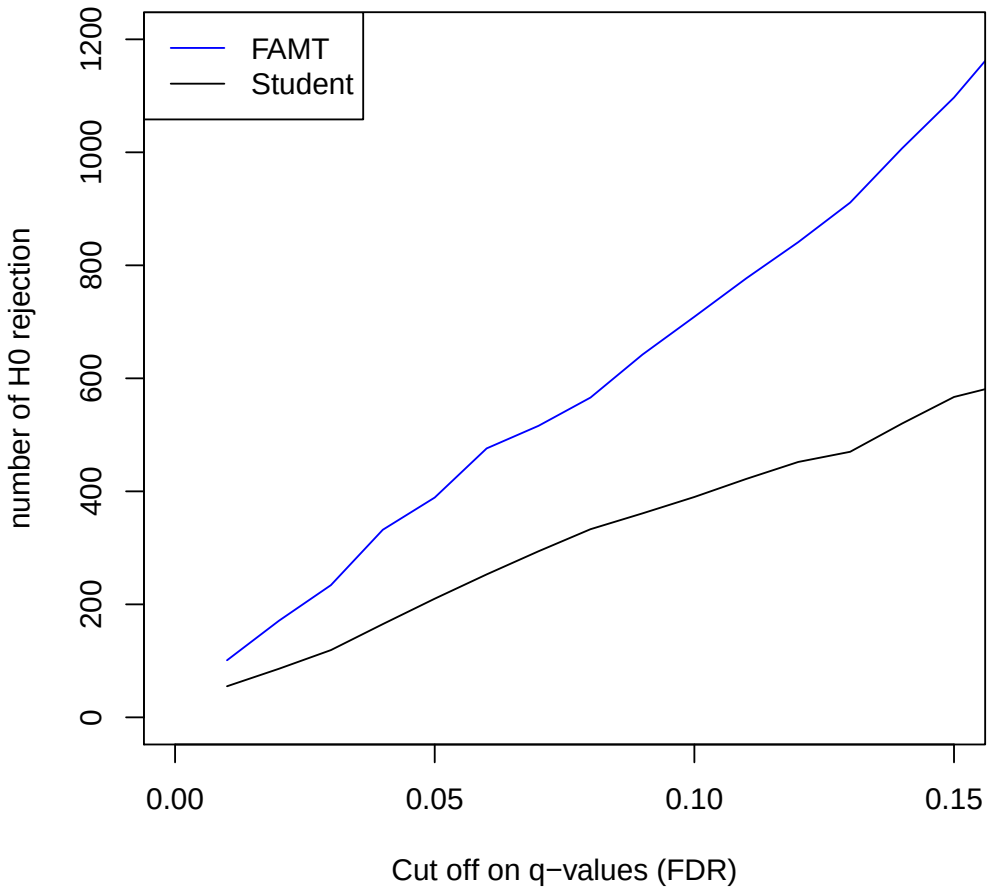
4/ Estimation of the proportion of null hypotheses

Key parameter to control the error rates.
FAMT provides 2 estimation algorithms :

- one based on the density of the conditional p-values
- the other uses a modified smoothing spline approach (based on Storey and Tibshirani, 2003).



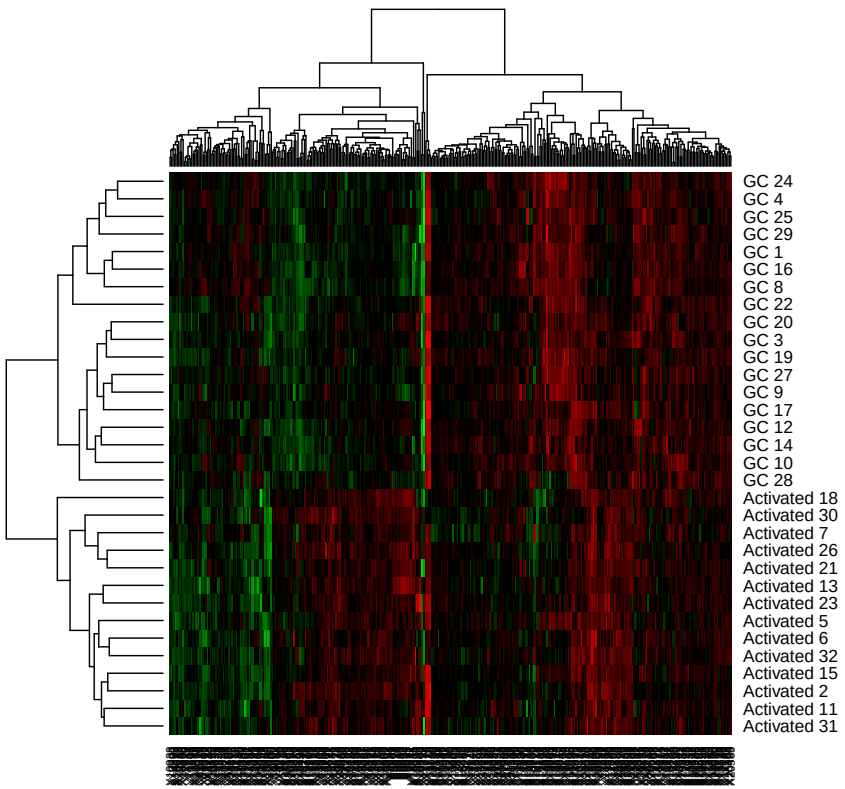
5/ Benjamini and Hochberg's procedure (q-values)



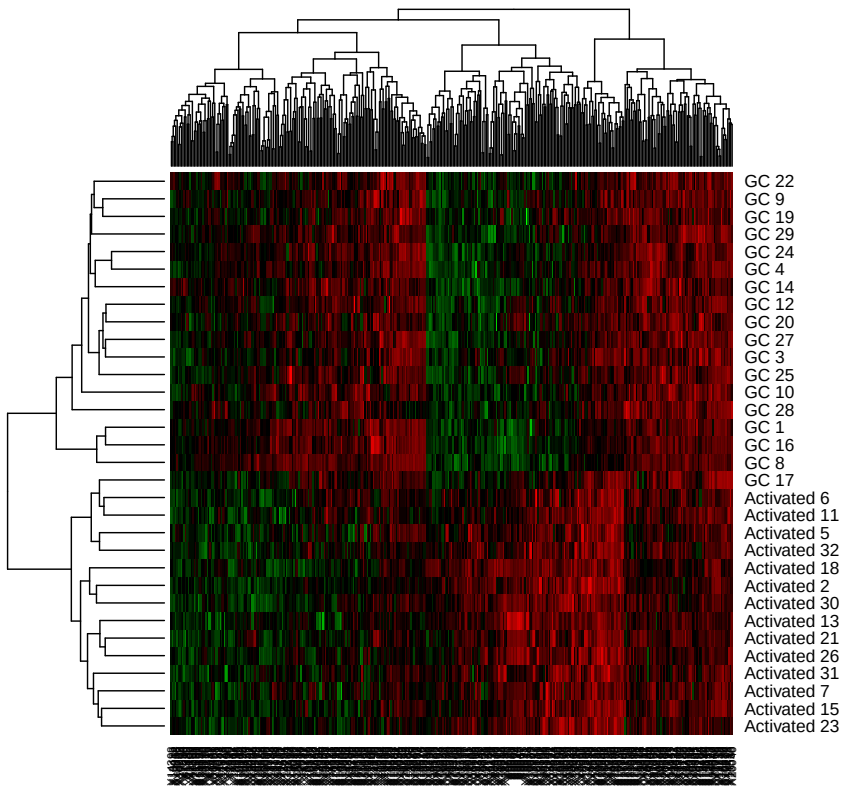
Heat maps

Cut off on the adjusted q-values : 5% FDR control level (389 genes)

Observed values



Factor-adjusted values



Outline

Concluding comments

- FAMT procedure : large improvements in multiple testing procedures regarding the FDR control and the power (decreasing the non-discovery proportion)
- The interpretation of the factors can be useful for biologists
- The factor-adjustment of test statistics also decreases misclassification rates and improves stability of model selection in supervised classification
- FAMT  package available at *http://www.agrocampus-ouest.fr/math/FAMT*

Interpretation of the factors