Factor Analysis for Multiple Testing

A general approach for differential analysis of genome-scale dependent data

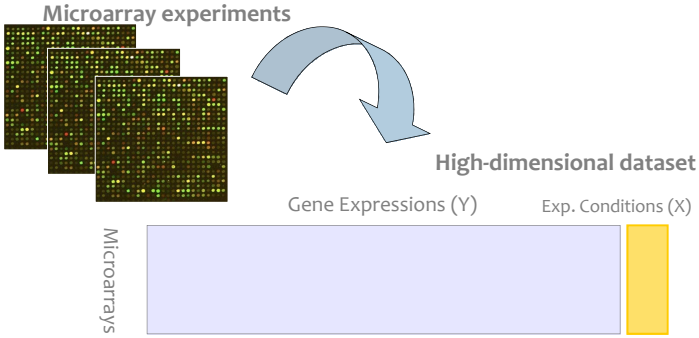
Differential analysis

Biological issue Identify the genes which expressions are significantly linked to the experimental condition thanks to microarray biotechnology

Statistical solution Multiple Testing

For each gene k: test of the null hypothesis H_0 of no association between its expression level Y_k and the environmental covariate X

- Huge number of simultaneous tests, usually several thousands
- High dimensional setting « *small n, large p* »
- Large-scale correlation structure, due to biological links among genes



Factor Analysis

Explain the dependence among a huge set of variables thanks to a small number of latent variables Z called the **common factors**

- Number of factors *q* chosen to reduce the variance of the number of false positives in multiple tests
- Estimation of the model parameters with an EM-algorithm to deal with high-dimension

$$Y = \beta_0 + x'\beta + BZ + \varepsilon$$

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

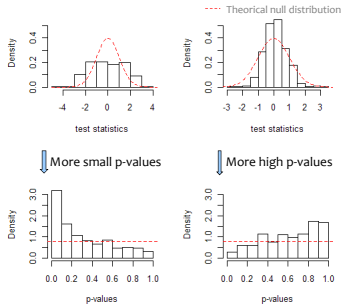
Specific variability (uniqueness) — Common variability

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

Ajusted test statistics and p-values

$$T_k \Rightarrow T_k^Z = \frac{\sigma_k}{\psi_k}(T_k - \tau_Z)$$

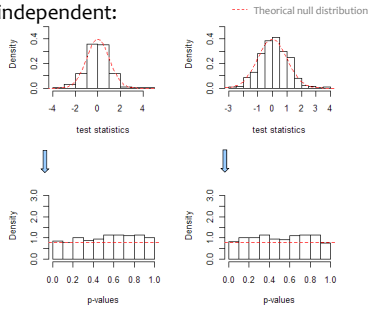
- Effect of correlation on usual test statistics and p-values distribution under the true null hypothesis:
  - Widen distribution
  - Narrow distribution



P-values histograms dissent from independent case (*U[0; 1]*)

- Adjusted test statistics: conditionally centered and scaled version of usual test statistics

Considering the FA model, they are independent:



Distribution of p-values = U[0; 1]

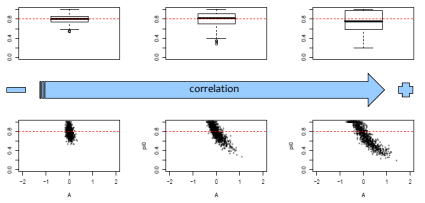
True null hypotheses proportion

Proportion of true-null hypotheses

$$\pi_0(t) = \frac{\#(p_k > t)}{1 - t}$$

Key parameter of most Multiple Testing Procedures (MTP)

- High variability of π_0 estimation as its variance depends on the correlation
- Most estimation methods rely on the behaviour of the p-values density near 1: under or over estimation of π_0 in presence of correlation

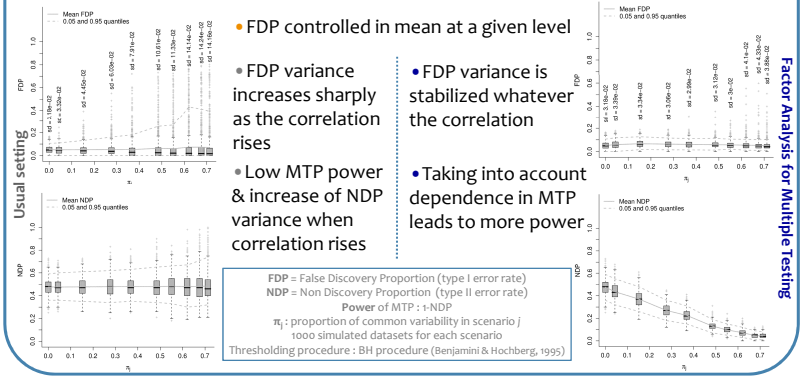


For each scenario 1000 datasets are simulated (true value of π_0=0,8)

A is Efron's criterion that scores the effect of correlation on test statistics distribution, the more correlation the more A's variability around 0:
A=0 : Tests statistics are independent
A<0 : Histogram of tests statistics is narrowed compared to the theorical null distribution
A>0 : Histogram is widened

Using the factor structure to define a conditional estimator induces an accurate estimation of π_0 and therfore increases the power of MTP

Power of Multiple Testing Procedures



- FDP controlled in mean at a given level
- FDP variance increases sharply as the correlation rises
- Low MTP power & increase of NDP variance when correlation rises
- FDP variance is stabilized whatever the correlation
- Taking into account dependence in MTP leads to more power

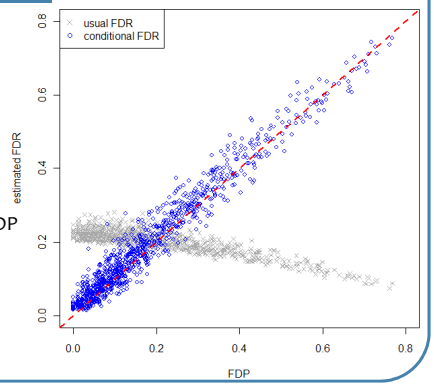
FDP = False Discovery Proportion (type I error rate)
NDP = Non Discovery Proportion (type II error rate)
Power of MTP : 1-NDP
π_i : proportion of common variability in scenario *j*
1000 simulated datasets for each scenario
Thresholding procedure : BH procedure (Benjamini & Hochberg, 1995)

FDR estimation

False Discovery Rate (FDR) : expected False Discoveries Proportion (FDP) among the rejected hypotheses (type-I error rate)

Considering the **usual FDR** leads to misleading estimation of the actual FDP in presence of correlation

Conditional FDR corrects FDP estimation from dependence effects



References

- **B. Efron (2007)** *Correlation and large-scale simultaneous significance testing – JASA*
- **C. Friguet, M. Kloareg & D. Causeur (2009)** *A factor model approach to multiple testing under dependence – JASA*
- **M. Langaas, B.H. Lindqvist & E. Ferkingstad (2005)** *Estimating the proportion of true null hypotheses with application to DNA microarray data – JRSS.B*
- **J.T. Leek & J.D. Storey (2008)** *A general framework for multiple testing dependence – PNAS*

R package : FAMT

http://www.agrocampus-ouest.fr/math/FAMT

Chloé FRIGUET ~ Maela KLOAREG ~ David CAUSEUR

Rannvro BREIZH

AGRO CAMPUS OUEST — Applied Mathematics Department — Rennes France