Intrinsic bounds on the Benjamini Hochberg procedure

Pierre Neuvial[1,2]

[1]Laboratoire de Probabilités et Modèles Aléatoires — Paris VII University
[2]Bioinformatics group — Institut Curie, Paris

MSHT Workshop — May 15, 2007

Goal : investigate two different MSHT problems

D.L. Donoho and J. Jin. Higher criticism for detecting sparse heterogeneous mixtures.
The Annals of Statistics, 32(3) :962–994, 2004.

Z. Chi. On the performance of FDR control : constraints and a partial solution.
The Annals of Statistics, to appear.

Why study these MSHT problems ?

- highlight the limitations of the BH procedure for these problems
- connect these limitations to the behaviour of the *p*-value distribution near 0
- quantify these limitations in practical applications

Outline

1. Introduction
   - Context
   - FDR control
   - Intrinsic bounds
2. Criticality
   - Tails and criticality
   - Studentised statistics
3. Detection boundaries
   - Tails and detection boundary
   - Detection boundaries and criticality

Outline

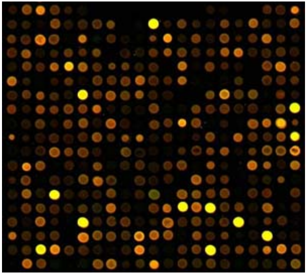
Motivation : DNA microarray analysis

Example : molecular analysis of cancer

DNA microarrays

High-throughput measurement of genes activity :

- m genes
- n samples (microarrays)
- $n << m$



Typical question : differential analysis of normal vs tumour samples

detection Do some genes behave differently between normal and tumour samples ?

multiple comparison Which of them ?

Such genes will be called *differentially expressed* (DE) genes

Mixture model

Settings

$(X_i, Y_i)_{1 \leqslant i \leqslant m}$ are identically independently distributed, with $Y_i \sim \mathcal{B}(\varepsilon)$ and

$$\begin{cases} X_i | Y_i = 1 \sim F^1 \\ X_i | Y_i = 0 \sim F^0 \end{cases}$$

- We observe a realisation of $(X_i)_{1 \leqslant i \leqslant m}$
- $(Y_i)_{1 \leqslant i \leqslant m}$ is hidden

Illustration from differential analysis of microarrays

- ε : proportion of DE genes
- $Y_i = \mathbf{1}_{\text{gene } i \text{ is DE}}$
- X_i : test statistic for gene i (built up from n samples)

Multiple comparison (MC) and detection (D) problems

Detection problem

Is ε equal to 0 ?

$$\begin{cases} \mathcal{H}_0^D: & (X_i)_i \overset{\text{iid}}{\sim} F^0 \\ \mathcal{H}_1^D: & (X_i)_i \overset{\text{iid}}{\sim} (1-\varepsilon)F^0 + \varepsilon F^1 \end{cases}$$

a binary testing problem

Multiple comparison problem

Which X_i come from F^1 ?

$$\begin{cases} \mathcal{H}_0^{MC}: & X_i \sim F^0 \\ \mathcal{H}_1^{MC}: & X_i \sim F^1 \end{cases}$$

a simultaneous test of m hypotheses

FDR for the multiple comparison problem

Possible outputs of a multiple comparison procedure

	accepted	rejected	
null	U	V	$m(1-\varepsilon)$
non null	S	T	$m\varepsilon$
	$m-R$	R	m

False Discovery Proportion

$$FDP = V/R$$

False Discovery Rate

$$FDR = E(FDP)$$

expected fraction of false discoveries

FDR for the multiple comparison problem

Possible outputs of a multiple comparison procedure

	accepted	rejected	
null	U	V	$m(1-\varepsilon)$
non null	S	T	$m\varepsilon$
	$m-R$	R	m

False Discovery Proportion

$$FDP = V/R$$

False Discovery Rate

$$FDR = E(FDP)$$

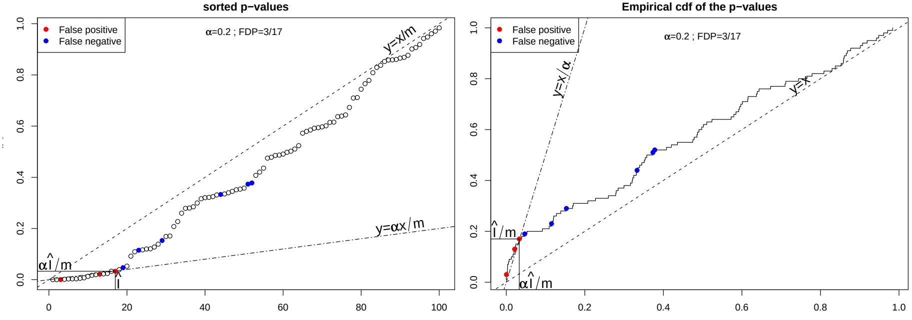
expected fraction of false discoveries

BH procedure for the multiple comparison problem

A *step-up* method providing strong control of the FDR (Benjamini & Hochberg, 1995)

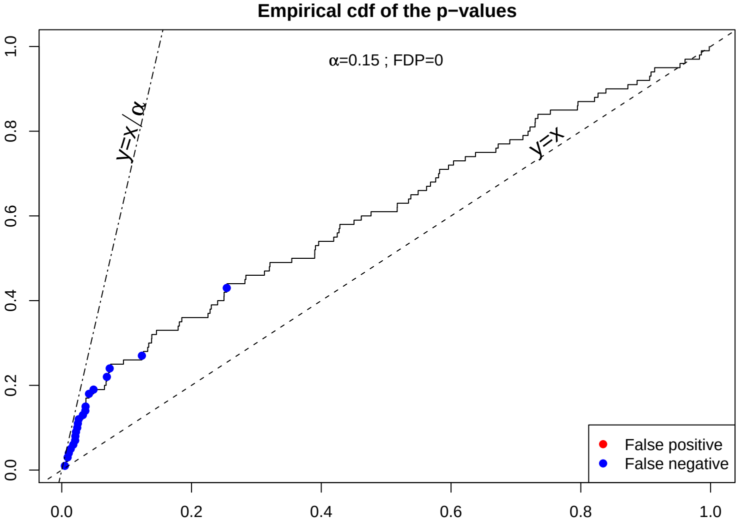
The BH procedure at level α

1. Sort the m p-values : $P_{(1)} \leqslant \ldots \leqslant P_{(m)}$ $\qquad P_i = 1 - F^0(X_i)$
2. Calculate $\widehat{l} = Max\left\{k | P_{(k)} \leqslant \alpha \frac{k}{m}\right\}$
3. Reject all p-values smaller than $= \alpha \widehat{l}/m$



Criticality of the multiple comparison problem

Chi (2007), Chi and Tan (2007)



Gaussian detection boundaries

BH detection boundary for sparse Gaussian mixtures Donoho and Jin (2004)

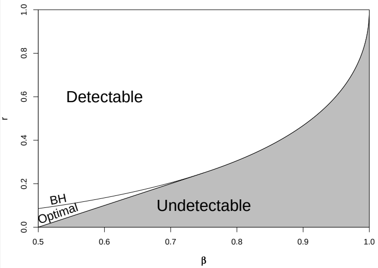
BH^D : BH as a detection procedure

- Reject $\mathcal{H}_0^D$ iff BH (α) rejects at least one hypothesis
- This procedure has level at most α for the detection problem

Gaussian mixtures

$$\begin{cases} \mathcal{H}_0^D : & (X_i)_i \overset{\text{iid}}{\sim} \mathcal{N}(0,1) \\ \mathcal{H}_1^D : & (X_i)_i \overset{\text{iid}}{\sim} (1-\varepsilon_m)\mathcal{N}(0,1) + \varepsilon_m \mathcal{N}(\mu_m, 1) \end{cases}$$

- sparsity : $\varepsilon_m = m^{-\beta}$, $\frac{1}{2} < \beta < 1$
- magnitude : $\mu_m = \sqrt{2r \log m}, 0 < r < 1$



Outline

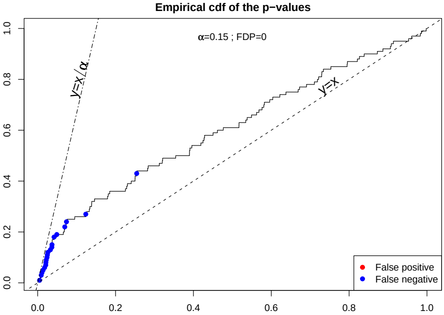
Criticality of the multiple comparison problem

Definition and interpretation

Multiple comparison problem

$$\begin{cases} \mathcal{H}_0^{MC}: & X_i \sim F^0 \\ \mathcal{H}_1^{MC}: & X_i \sim F^1 \end{cases}$$

p-values	$P_i = 1 - F^0(X_i)$
cdf	$G(u) = \varepsilon G^1(u) + (1-\varepsilon)(u)$
density	$g(u) = \varepsilon g^1(u) + (1-\varepsilon)$



Critical value (Chi, 2007)

$$\alpha^\star = \inf_{u \in [0,1]} \frac{u}{G(u)}$$

Interpretation of $\alpha^\star$

$$\alpha^\star = \lim_{u \to 0} \frac{u}{G(u)} = \frac{1}{g(0)}$$

Criticality of the multiple comparison problem

Properties and relationship to the likelihood ratio

Properties (Chi, 2007 and Chi and Tan, 2007)

For $\alpha < \alpha^\star$:

- the number of correct rejections made by BH (α) is asymptotically bounded as $m \to +\infty$
- BH (α) has asymptotically null power as $m \to +\infty$

Relationship to g^1 and $\frac{f^1}{f^0}$

- $\alpha^\star = \frac{1}{g(0)} = \frac{1}{\varepsilon g^1(0)+1-\varepsilon}$
- $g^1(u) = \frac{f^1}{f^0}\left(q^0(u)\right)$, where $q^0(u) = \left(F^0\right)^{-1}(1-u)$
- criticality occurs iff $\frac{f^1}{f^0}$ has a finite limit at $+\infty$

Criticality of the multiple comparison problem

Properties and relationship to the likelihood ratio

Properties (Chi, 2007 and Chi and Tan, 2007)

For $\alpha < \alpha^\star$:

- the number of correct rejections made by BH (α) is asymptotically bounded as $m \to +\infty$
- BH (α) has asymptotically null power as $m \to +\infty$

Relationship to g^1 and $\frac{f^1}{f^0}$

- $\alpha^\star = \frac{1}{g(0)} = \frac{1}{\varepsilon g^1(0)+1-\varepsilon}$
- $g^1(u) = \frac{f^1}{f^0}\left(q^0(u)\right)$, where $q^0(u) = \left(F^0\right)^{-1}(1-u)$
- criticality occurs iff $\frac{f^1}{f^0}$ has a finite limit at $+\infty$

Gaussian multiple comparison problem

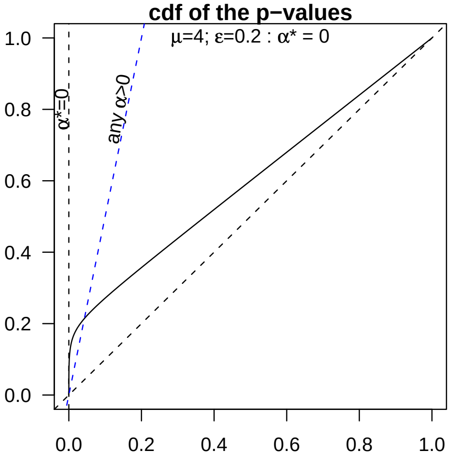
A simple example with no criticality phenomenon

Gaussian tails

$$\frac{f^1}{f^0}(t) = \exp\left[-\frac{1}{2}(t-\mu)^2 + \frac{1}{2}t^2\right]$$
$$= \exp\left[-\frac{\mu^2}{2} + \mu t\right]$$

No criticality

- $\lim_{t\to+\infty} \frac{f^1}{f^0}(t) = +\infty$
- $\lim_{u\to 0} g(u) = +\infty$
- $\alpha^\star = 0$



Laplace multiple comparison problem

A simple example with a criticality phenomenon

Laplace (double exponential) test statistics

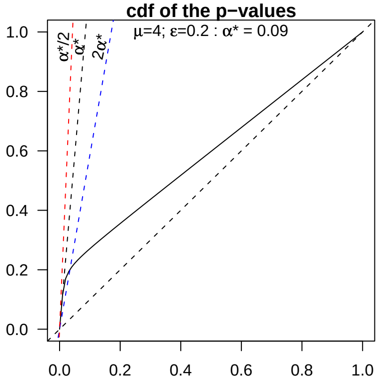
$$\begin{cases} \mathcal{H}_0^{MC} : X_i \sim \mathcal{E}^0 & f^0(t) = \frac{1}{2} e^{-|t|} \\ \mathcal{H}_1^{MC} : X_i \sim \mathcal{E}^\mu & f^1(t) = \frac{1}{2} e^{-|t-\mu|} \end{cases}$$

Heavier tails

$$\frac{f^1}{f^0}(t) = \begin{cases} e^{2t-\mu} & \text{if } t \leqslant \mu \\ e^{\mu} & \text{if } t > \mu \end{cases}$$

Criticality

- $\alpha^\star = \frac{1}{\varepsilon e^{\mu} + (1-\varepsilon)}$
- BH (α) has asymptotically null power for $\alpha < \alpha^\star$



Student multiple comparison problem

A problem of practical interest

Likelihood Ratio

$$\frac{f^1}{f^0}(t) \quad = \quad \exp\left[-\frac{\delta^2}{2}\,\frac{1}{1+\frac{t^2}{k}}\right]\frac{Hh_k\left(-\frac{\delta t}{\sqrt{k+t^2}}\right)}{Hh_k(0)}$$

with

$$Hh_k(z) = \int_0^{+\infty} \frac{x^k}{k!} e^{-\frac{1}{2}(x+z)^2} dx$$

Parameters of the model

- δ : non-centrality parameter
- k : number of degrees of freedom

Critical value of the Student MC problem

Criticality

- $\alpha^\star = \frac{1}{\varepsilon \frac{Hh_k(-\delta)}{Hh_k(0)} + (1-\varepsilon)}$
- BH (α) has asymptotically null power for $\alpha < \alpha^\star$

Whan can we do then ?

- k is an increasing function of sample size
- for fixed $\delta > 0$, $\lim_{k \to +\infty} \frac{Hh_k(-\delta)}{Hh_k(0)} = +\infty$

Theorem (Criticality vanishes as sample size increases)

$$\begin{cases} \mathcal{H}_0^{MC}: & X_i \sim t_0(k) \\ \mathcal{H}_1^{MC}: & X_i \sim t_\delta(k) \end{cases}$$

Let $k = k_m \to +\infty$ as $m \to +\infty$, then $\lim_{m \to +\infty} \alpha_m^\star = 0$

Outline

Detecting sparse heterogeneous mixtures

Detection problem

$$\begin{cases} \mathcal{H}_0^D : & (X_i)_i \overset{\text{iid}}{\sim} F_m^0 \\ \mathcal{H}_1^D : & (X_i)_i \overset{\text{iid}}{\sim} (1-\varepsilon_m)F_m^0 + \varepsilon_m F_m^1 \end{cases}$$

- p-values : $P_i = 1 - F_m^0(X_i)$
- g_m : density of the p-values under $\mathcal{H}_1^D$

Example : location problems

- $F_m^1(t) = F_m^0(t - \mu_m)$
- $\mu_m \to +\infty$, $\varepsilon_m \to 0$

For which (μ_m, ε_m) $\mathcal{H}_0^D$ is asymptotically correctly rejected by a given detection procedure ?

Detection boundary of the BH^D procedure

Connection with the *p*-value distribution

$\mathrm{BH}^D_{\alpha_m}$: the BH procedure for detection, with target *FDR* level α_m.

Theorem (Detection boundary of the BH^D procedure)

1. Let $\alpha_m \to 0$. For each m, $\mathrm{BH}^D_{\alpha_m}$ has level at most α_m, and

$$\lim_{m \to +\infty} \mathbb{P}_{\mathcal{H}_0^D}\left(\mathrm{BH}^D_{\alpha_m} \text{ rejects } \mathcal{H}_0^D\right) = 0$$

2. Let $\alpha_m \to 0$ slowly enough, if $\lim_{m \to +\infty} g_m\left(\frac{1}{m}\right) = +\infty$, then $\mathrm{BH}^D_{\alpha_m}$ has asymptotically full power for separating $\mathcal{H}_1^D$ from $\mathcal{H}_0^D$:

$$\lim_{m \to +\infty} \mathbb{P}_{\mathcal{H}_1^D}\left(\mathrm{BH}^D_{\alpha_m} \text{ rejects } \mathcal{H}_0^D\right) = 1$$

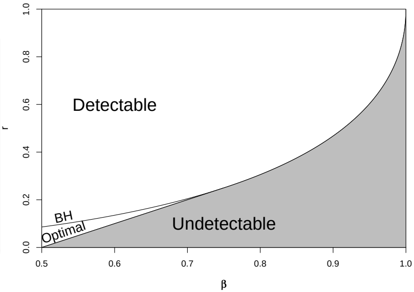
Application to the Gaussian detection problem

Sparse Gaussian mixtures

$$F_m = (1-\varepsilon_m)\mathcal{N}(0,1) + \varepsilon_m \mathcal{N}(\mu_m, 1)$$

$$\varepsilon_m = m^{-\beta} \quad \tfrac{1}{2} < \beta < 1$$

$$\mu_m = \sqrt{2r\log m} \quad 0 < r < 1$$



Gaussian detection boundaries (Donoho and Jin, 2004)

$$\rho^\star(\beta) = \begin{cases} \beta - \frac{1}{2} & \text{if } 1/2 < \beta \leqslant 3/4 \\ (1-\sqrt{1-\beta})^2 & \text{if } 3/4 < \beta < 1 \end{cases} \quad \text{(optimal)}$$

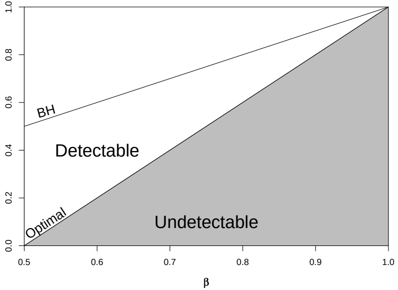
$$\rho^{\mathrm{BH}}(\beta) = (1-\sqrt{1-\beta})^2 \text{ for } 1/2 < \beta < 1 \quad \text{(BH)}$$

Application to the Laplace detection problem

Sparse Laplace mixtures

$$F_m = (1-\varepsilon_m)\mathcal{E}(0) + \varepsilon_m \mathcal{E}(\mu_m)$$

$$\varepsilon_m = m^{-\beta} \quad \tfrac{1}{2} < \beta < 1$$
$$\mu_m = r \log m \quad 0 < r < 1$$



Laplace Detection boundaries (Donoho and Jin, 2004)

$$\rho^{\star}(\beta) = 2\left(\beta - \tfrac{1}{2}\right) \quad \text{(optimal)}$$
$$\rho^{\mathrm{BH}}(\beta) = \beta \quad \text{(BH)}$$

Take-home message

Two problems related to multiple hypothesis testing

1. a detection problem : *Is ε null ?*
2. a multiple comparison problem : *Which X_i come from F^1 ?*

New connexions between these problems

1. existence of intrinsic bounds to the BH procedure
2. tight connexion between these bounds and the *p*-value distribution

Result of practical interest : sample size and criticality

For Studentised test statistics, criticality is asymptotically cancelled when sample size grows to $+\infty$.