Intrinsic bounds on the Benjamini Hochberg procedure

Pierre Neuvial\textsuperscript{1,2}

\textsuperscript{1}Laboratoire de Probabilités et Modèles Aléatoires — Paris VII University
\textsuperscript{2}Bioinformatics group — Institut Curie, Paris

MSHT Workshop — May 15, 2007
Goal: investigate two different MSHT problems


Z. Chi. On the performance of FDR control: constraints and a partial solution.


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
Introduction

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
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
Introduction

Context

Mixture model

Settings

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

\[
\begin{align*}
X_i | Y_i = 1 & \sim F^1 \\
X_i | Y_i = 0 & \sim F^0
\end{align*}
\]

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

Illustration from differential analysis of microarrays

- \(\varepsilon\) : proportion of DE genes
- \(Y_i = 1\) gene \(i\) is DE
- \(X_i\) : test statistic for gene \(i\) (built up from \(n\) samples)
Multiple comparison (MC) and detection (D) problems

Detection problem
Is $\epsilon$ equal to 0?

\[
\begin{align*}
\mathcal{H}_0^D &: \quad (X_i)_i \overset{iid}{\sim} F^0 \\
\mathcal{H}_1^D &: \quad (X_i)_i \overset{iid}{\sim} (1 - \epsilon)F^0 + \epsilon F^1
\end{align*}
\]

a binary testing problem

Multiple comparison problem
Which $X_i$ come from $F^1$?

\[
\begin{align*}
\mathcal{H}_0^{MC} &: \quad X_i \sim F^0 \\
\mathcal{H}_1^{MC} &: \quad X_i \sim F^1
\end{align*}
\]

a simultaneous test of m hypotheses
FDR for the multiple comparison problem

Possible outputs of a multiple comparison procedure

<table>
<thead>
<tr>
<th>null</th>
<th>accepted</th>
<th>rejected</th>
</tr>
</thead>
<tbody>
<tr>
<td>$U$</td>
<td>$V$</td>
<td>$m(1 - \varepsilon)$</td>
</tr>
<tr>
<td>$S$</td>
<td>$T$</td>
<td>$m\varepsilon$</td>
</tr>
<tr>
<td>$m - R$</td>
<td>$R$</td>
<td>$m$</td>
</tr>
</tbody>
</table>

False Discovery Proportion

$$FDP = \frac{V}{R}$$

False Discovery Rate

$$FDR = E(FDP)$$

expected fraction of false discoveries
FDR for the multiple comparison problem

Possible outputs of a multiple comparison procedure

<table>
<thead>
<tr>
<th></th>
<th>accepted</th>
<th>rejected</th>
</tr>
</thead>
<tbody>
<tr>
<td>null</td>
<td>$U$</td>
<td>$V$</td>
</tr>
<tr>
<td>non null</td>
<td>$S$</td>
<td>$T$</td>
</tr>
<tr>
<td>$m - R$</td>
<td>$R$</td>
<td>$m$</td>
</tr>
</tbody>
</table>

$FDP = \frac{V}{R}$

$FDR = E(FDP)$

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 $\alpha$

1. Sort the $m$ $p$-values: $P_{(1)} \leq \ldots \leq P_{(m)}$

2. Calculate $\hat{l} = \text{Max} \left\{ k \mid P_{(k)} \leq \alpha \frac{k}{m} \right\}$

3. Reject all $p$-values smaller than $\frac{\alpha \hat{l}}{m}$

$P_i = 1 - F^0(X_i)$

Intrinsic bounds on the BH procedure

---

P. Neuvial (Inst. Curie & Univ. Paris VII)
Critically of the multiple comparison problem
Chi (2007), Chi and Tan (2007)
Gaussian detection boundaries
BH detection boundary for sparse Gaussian mixtures

\[ \mathcal{H}_0^D: (X_i)_{i \sim i.i.d.} \sim \mathcal{N}(0, 1) \]
\[ \mathcal{H}_1^D: (X_i)_{i \sim i.i.d.} \sim (1 - \varepsilon_m)\mathcal{N}(0, 1) + \varepsilon_m\mathcal{N}(\mu_m, 1) \]

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

\( \mathbf{BH}^D: \mathbf{BH}^D \) as a detection procedure
- Reject \( \mathcal{H}_0^D \) iff \( \mathbf{BH}^D(\alpha) \) rejects at least one hypothesis
- This procedure has level at most \( \alpha \) for the detection problem
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
### Multiple comparison problem

\[
\begin{align*}
\mathcal{H}_0^{\text{MC}} : & \quad X_i \sim F^0 \\
\mathcal{H}_1^{\text{MC}} : & \quad X_i \sim F^1
\end{align*}
\]

- **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)

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

### Interpretation of \( \alpha^* \)

\[
\alpha^* = \lim_{u \to 0} \frac{u}{G(u)} = \frac{1}{g(0)}
\]

---

**Empirical cdf of the p-values**

- \( \alpha = 0.15 \); FDP=0

- False positive
- False negative

\( \alpha^* = \lim_{u \to 0} \frac{u}{G(u)} = \frac{1}{g(0)} \)
## Criticality of the multiple comparison problem

### Properties and relationship to the likelihood ratio

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

For $\alpha < \alpha^*$:

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

#### Relationship to $g^1$ and $f_1^{f_0}$

- $\alpha^* = \frac{1}{g(0)} = \frac{1}{\varepsilon g^1(0) + 1 - \varepsilon}$
- $g^1(u) = \frac{f_1^{f_0}}{f_0} (q^0(u))$, where $q^0(u) = (F^0)^{-1} (1 - u)$
- Criticality occurs iff $\frac{f_1^{f_0}}{f_0}$ 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 $\alpha < \alpha^*$:
- the number of correct rejections made by BH ($\alpha$) is asymptotically bounded as $m \to +\infty$
- BH ($\alpha$) has asymptotically null power as $m \to +\infty$

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

- $\alpha^* = \frac{1}{g(0)} = \frac{1}{\varepsilon g^1(0) + 1 - \varepsilon}$
- $g^1(u) = \frac{f^1}{f^0} (q^0(u))$, where $q^0(u) = (F^0)^{-1}(1 - u)$
- 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

- \( \lim_{t \to +\infty} \frac{f_1}{f_0}(t) = +\infty \)
- \( \lim_{u \to 0} g(u) = +\infty \)
- \( \alpha^* = 0 \)
Laplace multiple comparison problem
A simple example with a criticality phenomenon

### Laplace (double exponential) test statistics

$$\begin{align*}
H_0^{MC} : X_i &\sim \mathcal{E}^0 \\
H_1^{MC} : X_i &\sim \mathcal{E}^\mu
\end{align*}$$

$$f^0(t) = \frac{1}{2} e^{-|t|}$$

$$f^1(t) = \frac{1}{2} e^{-|t-\mu|}$$

### Heavier tails

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

### Criticality

- \(\alpha^* = \frac{1}{\epsilon e^\mu + (1-\epsilon)}\)
- BH (\(\alpha\)) has asymptotically null power for \(\alpha < \alpha^*\)
Student multiple comparison problem
A problem of practical interest

Likelihood Ratio

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

with

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

Parameters of the model

- \( \delta \): non-centrality parameter
- \( k \): number of degrees of freedom
Critical value of the Student MC problem

Criticality

\[ \alpha^* = \frac{1}{\varepsilon \frac{Hh_k(-\delta)}{Hh_k(0)}} + (1 - \varepsilon) \]

BH (\( \alpha \)) has asymptotically null power for \( \alpha < \alpha^* \)

When 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 \( \lim_{m \to +\infty} \alpha^*_m = 0 \)
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
Detecting sparse heterogeneous mixtures

Detection problem

\[
\begin{align*}
\mathcal{H}_0^D : & \quad (X_i)_i \overset{iid}{\sim} F_m^0 \\
\mathcal{H}_1^D : & \quad (X_i)_i \overset{iid}{\sim} (1 - \varepsilon_m)F_m^0 + \varepsilon_m F_m^1
\end{align*}
\]

- \textit{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 \((\mu_m, \varepsilon_m)\ \mathcal{H}_0^D\) is asymptotically correctly rejected by a given detection procedure?
Detection boundary of the BH\textsuperscript{D} procedure

Connection with the \( p \)-value distribution

\( \text{BH}_{\alpha_m}^D \): the BH procedure for detection, with target \( FDR \) level \( \alpha_m \).

**Theorem (Detection boundary of the BH\textsuperscript{D} procedure)**

1. Let \( \alpha_m \to 0 \). For each \( m \), \( \text{BH}_{\alpha_m}^D \) has level at most \( \alpha_m \), and

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

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

\[
\lim_{m \to +\infty} \mathbb{P}_{\mathcal{H}_1^D} \left( \text{BH}_{\alpha_m}^D \text{ rejects } \mathcal{H}_0^D \right) = 1
\]
Application to the Gaussian detection problem

Sparse Gaussian mixtures

\[ F_m = (1 - \varepsilon_m)N(0, 1) + \varepsilon_m N(\mu_m, 1) \]

\[ \varepsilon_m = m^{-\beta} \quad \frac{1}{2} < \beta < 1 \]

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

Gaussian detection boundaries (Donoho and Jin, 2004)

\[ \rho^*(\beta) = \begin{cases} 
\beta - \frac{1}{2} & \text{if } 1/2 < \beta \leq 3/4 \\
(1 - \sqrt{1 - \beta})^2 & \text{if } 3/4 < \beta < 1 
\end{cases} \quad \text{(optimal)} \]

\[ \rho^{BH}(\beta) = (1 - \sqrt{1 - \beta})^2 \quad \text{for } 1/2 < \beta < 1 \quad \text{(BH)} \]
Application to the Laplace detection problem

Sparse Laplace mixtures

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

\[ \varepsilon_m = m^{-\beta} \quad \frac{1}{2} < \beta < 1 \]

\[ \mu_m = r \log m \quad 0 < r < 1 \]

Laplace Detection boundaries (Donoho and Jin, 2004)

\[ \rho^* (\beta) = 2 \left( \beta - \frac{1}{2} \right) \quad \text{(optimal)} \]

\[ \rho^{BH} (\beta) = \beta \quad \text{(BH)} \]
Two problems related to multiple hypothesis testing

1. a detection problem: \( \text{Is } \epsilon \text{ null?} \)
2. a multiple comparison problem: \( \text{Which } X_i \text{ come from } F^1? \)

New connections 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 \( +\infty \).