An Introduction on Multiple Testing:
False Discovery Control

Florent Chatelain

Gipsa-Lab / DIS, Grenoble INP

BasMatI, Porquerolles, 2015
Multiplicity problem and chance correlation

Lottery

- Winning probability for a given ticket is very low...
- But among the huge number of tickets, the probability that there is *at least one* winning ticket is quite high!

Paul the octopus

- Paul predicts eight of the 2010 FIFA World Cup matches with a perfect score!
- Does it really means that Paul is an Oracle?

Large-scale experiments: multiplying the comparisons dramatically increases the probability to obtain a good match by *pure chance*
Multiple Testing and FDR

Introduction

Multiplicity problem for statistical testing

- $T$ is the test statistics,
- $\mathcal{R}_\alpha$ is the region of rejection at level $\alpha$: if $H_0$ is true, $\Pr(T \in \mathcal{R}_\alpha) = \alpha$

Multiple testing issue

- $N$ independent statistics $T_1, \ldots, T_N$ obtained under the null $H_0$
- Probability to reject at least one of the $N$ null hypotheses:

$$
\Pr(\exists T_i \in \mathcal{R}_\alpha) = 1 - \Pr(T_1, \ldots, T_N \notin \mathcal{R}_\alpha) = 1 - \prod_{i=1}^{N} \Pr(T_i \notin \mathcal{R}_\alpha),
$$

$$
= 1 - \prod_{i=1}^{N} (1 - \alpha) = 1 - (1 - \alpha)^N
$$

- for a usual significative level $\alpha = 0.05$, performing $N = 20$ tests gives a probability 0.64 to find a ‘significative’ discovery by pure chance...

Pr ( at least one false positive ) $\gg$ Pr ( the $i$-th is a false positive )
Multiplicility problem in science

_The Economist, 2013, “Unreliable research”_

Many published research findings in top-ranked journals are not, or poorly, reproducible [Ioannidis, 2005]

- if the test power is only 0.4, 40 true positives in average for 45 false positives. Is this significant?
Large-Scale Hypothesis Testing [Efron, 2010]

Era of Massive Data Production

- “omics” revolution, e.g. microarrays measures expression levels of tens of thousands of genes for hundreds of subjects
- astrophysics, e.g. MUSE spectro-imager delivers cubes of $300 \times 300$ images for 3600 wavelengths: detecting faint sources leads to $N \approx 3 \times 10^8$ tests in a pixelwise approach

Large-Scale methodology

- statistical inference and hypothesis testing theory developed in the early 20th century (Pearson, Fisher, Neyman, ...) for small-data sets collected by individual scientist
- corrections are needed to assess significance in large-scale experiments
Outline

Multiple testing error control
  Basic statistical hypothesis testing concepts
  Family-Wise Error Rate FWER
  False Discovery Rate FDR

FDR control: Benjamini-Hochberg Procedure
  BH Procedure
  Bayesian interpretation of FDR
  Empirical Bayes interpretation of BH procedure

Variations on FDR control and BH Procedure
  Improving power
  Dependence
  Learning the null distribution
## Type I and Type II Errors

For an individual statistical hypothesis testing

<table>
<thead>
<tr>
<th>Actual</th>
<th>Decision</th>
<th>$H_0$ true</th>
<th>$H_0$ false</th>
</tr>
</thead>
<tbody>
<tr>
<td>$H_0$ true</td>
<td>True Negative (TN)</td>
<td>1 - $\alpha$</td>
<td>False Positive (FP)</td>
</tr>
<tr>
<td>$H_0$ false</td>
<td>False Negative (FN)</td>
<td>Type II Error $\beta$</td>
<td>True Positive (TP)</td>
</tr>
</tbody>
</table>

- False Positive ← *false alarm,*
- False Negative ← *miss-detection,*
- $\alpha = \Pr(\text{Type I Error})$ ← *significance level,*
- $\beta = \Pr(\text{Type II Error})$
- Power $\pi = \Pr(\text{True Positive}) = 1 - \beta$
**P-values : an universal language for hypothesis testing**

**Intuitive definition**

$p$-value $\equiv$ probability of obtaining a result as extreme or “more extreme” than the observed statistics, under $H_0$

**One-sided test example**

- $T$ is the test statistic, $t_{\text{obs}}$ an observed realization of $T$
- $H_0$ rejected when $t_{\text{obs}}$ is too large : $\mathcal{R}_\alpha = \{ t : t \geq \eta_\alpha \}$

$$p(t_{\text{obs}}) = \Pr_{H_0} (T \geq t_{\text{obs}})$$

**Mathematical definition**

Smallest value of $\alpha$ such that $t_{\text{obs}} \in \mathcal{R}_\alpha$

$$p(t_{\text{obs}}) = \inf_{\alpha} \{ t_{\text{obs}} \in \mathcal{R}_\alpha \}$$
Multiple Testing and FDR

Multiple testing error control

Basic statistical hypothesis testing concepts

Property of $p$-values

- Note that $p(t_{obs}) \leq u \iff t_{obs} \in \mathcal{R}_u$, for all $u \in [0, 1]$
- Let $P = p(T)$ be the random variable. If $H_0$ is true

$$\Pr_{H_0}(P \leq u) = \Pr_{H_0}(T \in \mathcal{R}_u) = u,$$

$p$-value $\equiv$ transformation of the test statistics to be uniformly distributed under the null (whatever the distribution of $T$)

Statistical hypothesis test based on $p$-value

$H_0$: $p$-value has a uniform distribution on $[0, 1]: P \sim \mathcal{U}([0, 1])$

$H_1$: $p$-value is stochastically lower than $\mathcal{U}([0, 1])$:

$$\Pr_{H_1}(P \leq u) = \Pr_{H_1}(T \in \mathcal{R}_u) > u,$$

$p$ the smaller is $p \equiv p(t_{obs})$, the more decisively is $H_0$ rejected

- for a given $\alpha$, $H_0$ is rejected at level $\alpha$ if $p \leq \alpha$
Counting the errors in multiple testing

- $N$ hypothesis tests with a common procedure

<table>
<thead>
<tr>
<th></th>
<th>$H_0$ retained</th>
<th>$H_0$ rejected</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>$H_0$ true</td>
<td>$V$</td>
<td>$U$</td>
<td>$N_0$</td>
</tr>
<tr>
<td>$H_0$ false</td>
<td>$S$</td>
<td>$T$</td>
<td>$N_1$</td>
</tr>
<tr>
<td>Total</td>
<td>$N - R$</td>
<td>$R$</td>
<td>$N$</td>
</tr>
</tbody>
</table>

- $N_0 = \#$ true nulls, $N_1 = \#$ true alternatives
- $U = \#$ False Positives ← Type I Errors
- $T = \#$ True Positives,
- $R = \#$ Rejections

How to define, and control, a global Type I Error rate/criterion?
Family-Wise Error Rate FWER

Multiple testing settings for $N$ tests

- $H_0^1, H_0^2, \ldots, H_0^N \equiv$ family of null hypotheses
- $p_1, p_2, \ldots, p_N \equiv$ corresponding p-values

Definition

- The familywise error rate is

  \[ \text{FWER} \equiv \Pr \left( \text{Reject at least one true } H_0^i \right) = \Pr \left( U > 0 \right) \]

- A FWER control procedure inputs a family of p-values $p_1, p_2, \ldots, p_N$ and outputs the list of rejected null hypotheses with the constraint

  \[ \text{FWER} \leq \alpha \]

for any preselected $\alpha$
Bonferroni’s correction and FWER control

Bonferroni’s correction
Reject the null hypotheses $H_0^i$ for which $p_i \leq \frac{\alpha}{N}$, ($N$ is the number of tests)

FWER control
Let $I_0$ be the indexes of the true null hypotheses, and $N_0 = \#I_0$

$$\text{FWER} = \Pr \left( \bigcup_{i \in I_0} p_i \leq \frac{\alpha}{N} \right) \leq \sum_{i \in I_0} \Pr \left( p_i \leq \frac{\alpha}{N} \right),$$

$$= N_0 \frac{\alpha}{N} \leq \alpha,$$

where the first inequality is the Boole’s inequality $\Pr (\bigcup_i A_i) \leq \sum_i \Pr (A_i)$.

- Bonferroni’s does not require that the tests be independent (the $p_i$ can be dependent)
- Šidák correction ’improves’ Bonferroni for independent tests by rejecting the $H_0^i$ for which $p_i \leq 1 - (1 - \alpha)^{1/N}$ ← equivalent for small $\alpha/N$ to Bonferroni : no real improvement.
Bonferroni’s correction and FWER control

Bonferroni’s correction
Reject the null hypotheses $H_0^i$ for which $p_i \leq \frac{\alpha}{N}$, ($N$ is the number of tests)

FWER control
Let $I_0$ be the indexes of the true null hypotheses, and $N_0 = \#I_0$

$$\text{FWER} = \Pr \left( \bigcup_{i \in I_0} p_i \leq \frac{\alpha}{N} \right) \leq \sum_{i \in I_0} \Pr \left( p_i \leq \frac{\alpha}{N} \right),$$

$$= N_0 \frac{\alpha}{N} \leq \alpha,$$

where the first inequality is the Boole’s inequality $\Pr (\bigcup_i A_i) \leq \sum_i \Pr (A_i)$.

- Bonferroni’s does not require that the tests be independent (the $p_i$ can be dependent)
- Šidák correction 'improves' Bonferroni for independent tests by rejecting the $H_0^i$ for which $p_i \leq 1 - (1 - \alpha)^{1/N} \leftarrow$ equivalent for small $\alpha/N$ to Bonferroni : no real improvement.
Stepwise FWER control procedures

Ordered p-values $p(1) \leq p(2) \leq \ldots \leq p(N)$ and associated null hypotheses $H_0^{(1)}, \ldots, H_0^{(N)}$

**Step-down procedures**

- Reject $H_0^{(k)}$ when $p(j) \leq t_{\alpha,j}$ for $j = 1, \ldots, k$

  Learning from the other experiments idea

- Reject $H_0^{(1)}, \ldots, H_0^{(\hat{k}_{\text{max}})}$ where $\hat{k}_{\text{max}}$ is the largest index satisfying (1)

  Global threshold $\hat{t}_\alpha \equiv t_{\alpha,\hat{k}_{\text{max}}}$ ← “testimation” problem

- Holm’s procedure: $t_{\alpha,j} = \frac{\alpha}{N-j+1}$ ensures FWER control at level $\alpha$ (not requiring independence) ← uniformly more powerful than Bonferroni

**Step-up procedures**

- Hochberg’s procedure: $\hat{t}_\alpha = t_{\alpha,\hat{k}_{\text{max}}}$ where $\hat{k}_{\text{max}}$ is the largest index satisfying $p(k) \leq \frac{\alpha}{N-k+1}$

  Unif. more powerful than Holm but requires the tests to be independent
Practical limits of FWER

- FWER is appropriate to guard against *any* false positives
- In many applications, this appears to be too stringent: we can accept several false positives if their number is still much “lower” than the number of true positives...
- More liberal variants

\[ k - \text{FWER} \equiv \Pr \left( \text{Reject at least } k \text{ true } H_0^i \right) = \Pr \left( U \geq k \right), \]

but how to preselect a relevant \( k \) for a given problem?

- Need to define a less stringent global Type I Error rate criterion, more useful in many applications
False Discovery Rate FDR [Benjamini and Hochberg, 1995]

“Discovery” terminology

- \( R \equiv \# \) Discoveries (Rejections)
- \( U \equiv \# \) False Discoveries (False Positives) \leftarrow \text{Type I errors},
- \( T \equiv \# \) True Discoveries (True Positives),

\[
\begin{array}{c|c|c|c}
\text{Actual} & H_0 \text{ retained} & H_0 \text{ rejected} & \text{Total} \\
\hline
H_0 \text{ true} & V & U & N_0 \\
H_0 \text{ false} & S & T & N_1 \\
\hline
\text{Total} & N - R & R & N \\
\end{array}
\]

Definition

\[
\text{FDP} \equiv \frac{U}{R \vee 1}, \quad \text{where} \quad R \vee 1 \equiv \max (R, 1) \leftarrow \text{False Discovery Proportion}
\]

\[
\text{FDR} \equiv E[\text{FDP}] = E \left[ \frac{U}{R \vee 1} \right] \leftarrow \text{False Discovery Rate}
\]

- Single test errors, or power, are calculated horizontally in the table
- False Discovery Rate is calculated vertically (Bayesian flavor)
False Discovery Rate FDR [Benjamini and Hochberg, 1995]

“Discovery” terminology

- \( R \equiv \# \) Discoveries (Rejections)
- \( U \equiv \# \) False Discoveries (False Positives) ← Type I errors,
- \( T \equiv \# \) True Discoveries (True Positives),

<table>
<thead>
<tr>
<th>Decision</th>
<th>( H_0 ) retained</th>
<th>( H_0 ) rejected</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>( H_0 ) true</td>
<td>( V )</td>
<td>( U )</td>
<td>( N_0 )</td>
</tr>
<tr>
<td>( H_0 ) false</td>
<td>( S )</td>
<td>( T )</td>
<td>( N_1 )</td>
</tr>
<tr>
<td>Total</td>
<td>( N - R )</td>
<td>( R )</td>
<td>( N )</td>
</tr>
</tbody>
</table>

Definition

- FDP \( \equiv \frac{U}{R \lor 1} \), where \( R \lor 1 \equiv \max (R, 1) \) ← False Discovery Proportion
- FDR \( \equiv E \left[ \text{FDP} \right] = E \left[ \frac{U}{R \lor 1} \right] \) ← False Discovery Rate

- Single test errors, or power, are calculated horizontally in the table
- False Discovery Rate is calculated vertically (Bayesian flavor)
False Discovery Rate FDR [Benjamini and Hochberg, 1995]

“Discovery” terminology

- $R \equiv \# \text{ Discoveries (Rejections)}$
- $U \equiv \# \text{ False Discoveries (False Positives)} \leftarrow \text{Type I errors,}$
- $T \equiv \# \text{ True Discoveries (True Positives),}$

<table>
<thead>
<tr>
<th>Decision</th>
<th>$H_0$ retained</th>
<th>$H_0$ rejected</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>$H_0$ true</td>
<td>$V$</td>
<td>$U$</td>
<td>$N_0$</td>
</tr>
<tr>
<td>$H_0$ false</td>
<td>$S$</td>
<td>$T$</td>
<td>$N_1$</td>
</tr>
<tr>
<td>Total</td>
<td>$N - R$</td>
<td>$R$</td>
<td>$N$</td>
</tr>
</tbody>
</table>

Definition

FDP $\equiv \frac{U}{R \lor 1}$, where $R \lor 1 \equiv \max(R, 1) \leftarrow \text{False Discovery Proportion}$

FDR $\equiv E[\text{FDP}] = E\left[\frac{U}{R \lor 1}\right] \leftarrow \text{False Discovery Rate}$

- single test errors, or power, are calculated horizontally in the table
- False Discovery Rate is calculated vertically (Bayesian flavor)
FDR control is more liberal than FWER

**FWER control procedure controls FDR**

\[
\text{FDR} = E \left[ \frac{U}{R \lor 1} \right] = E \left[ \frac{U}{R \lor 1} \mid U = 0 \right] \Pr (U = 0) + E \left[ \frac{U}{R \lor 1} \mid U > 0 \right] \Pr (U > 0),
\]

\[
= E \left[ \frac{U}{R} \right] \Pr (U > 0), \quad \text{where } 0 \leq \frac{U}{R} \leq 1,
\]

\[
\leq \Pr (U > 0) = \text{FWER}
\]

Procedure controlling FWER at level \( \alpha \) controls FDR at level \( \alpha \)

FDR control procedure controls the FWER in the *weak* sense

If all the nulls \( H_0^1, \ldots, H_0^N \) are true then \( U = R \) and

\[
\text{FDR} = E \left[ \frac{U}{R} \mid U > 0 \right] \Pr (U > 0) = 1 \times \Pr (U > 0) = \text{FWER}
\]

Procedure controlling FDR at level \( q \) controls FDR at level \( q \) only when all null hypotheses are true
FDR control is more liberal than FWER

FWER control procedure controls FDR

\[
FDR = E \left[ \frac{U}{R \vee 1} \right] = E \left[ \frac{U}{R \vee 1} \mid U = 0 \right] \Pr (U = 0) + E \left[ \frac{U}{R \vee 1} \mid U > 0 \right] \Pr (U > 0),
\]

\[
= E \left[ \frac{U}{R} \mid U > 0 \right] \Pr (U > 0), \quad \text{where } 0 \leq \frac{U}{R} \leq 1,
\]

\[
\leq \Pr (U > 0) = \text{FWER}
\]

Procedure controlling FWER at level \( \alpha \) controls FDR at level \( \alpha \)

FDR control procedure controls the FWER in the weak sense

If all the nulls \( H_0^1, \ldots, H_0^N \) are true then \( U = R \) and

\[
FDR = E \left[ \frac{U}{R} \mid U > 0 \right] \Pr (U > 0) = 1 \times \Pr (U > 0) = \text{FWER}
\]

Procedure controlling FDR at level \( q \) controls FDR at level \( q \) only

when all null hypotheses are true
Outline

Multiple testing error control
  - Basic statistical hypothesis testing concepts
  - Family-Wise Error Rate FWER
  - False Discovery Rate FDR

FDR control: Benjamini-Hochberg Procedure
  - BH Procedure
  - Bayesian interpretation of FDR
  - Empirical Bayes interpretation of BH procedure

Variations on FDR control and BH Procedure
  - Improving power
  - Dependence
  - Learning the null distribution
Canonical example

Source detection (oversimplified) problem
Statiscal linear model (source + noise)

\[
\begin{pmatrix}
X_1 \\
X_2 \\
\vdots \\
X_N
\end{pmatrix}
= \mu
\begin{pmatrix}
r_1 \\
r_2 \\
\vdots \\
r_N
\end{pmatrix}
+ \begin{pmatrix}
\epsilon_1 \\
\epsilon_2 \\
\vdots \\
\epsilon_N
\end{pmatrix}
\]

- $\mu > 0 \leftarrow$ source response
- $r_i \in \{0, 1\} \leftarrow$ absence ($r_i = 0$) or presence ($r_i = 1$) of source for $i$th location
- $\epsilon_i, 1 \leq i \leq N$, are iid with $\mathcal{N}(0, 1)$ distribution $\leftarrow$ gaussian noise
- $X_i$ is the $i$th observation
Canonical example (cont’d)

Multiple testing problem
for each $i$

- $H_0$: null hypothesis $\equiv$ absence of signal, i.e. $r_i = 0$
- $H_1$: alternative hypothesis $\equiv$ presence of signal, i.e. $r_i = 1$

Test statistics
for each $i$

- $X_i$ is the test statistics
- $p_i = 1 - \Phi(X_i)$, where $\Phi$ is the standard normal cdf, is the associated p-value

How to choose a good threshold $t$ to reject the tests s.t. $p_i \leq t$?
Ordered p-values plot for $N = 50$, $N_0 = 40$, $\mu = 2$, $\alpha = 0.1$

Ordered p-values $p_1 \leq p_2 \leq \cdots \leq p_N$ vs theoretical quantiles $1/N, 2/N, \ldots, 1$ under the null
Ordered p-values plot for $N = 100$, $N_0 = 80$, $\mu = 3$, $\alpha = 0.1$

Try something between Bonferroni and one test control: choose $t_i = q \frac{i}{N}$ (here $q = \alpha = 0.1$)
Benjamini-Hochberg (BH) procedure

Ordered p-values $p(1) \leq p(2) \leq \ldots \leq p(N)$ and associated null hypotheses $H_0^{(1)}, \ldots, H_0^{(N)}$, let $p(0) = 0$ by convention

Step-up BH procedure
For a preselected control level $0 \leq q \leq 1$, BH$_q$ procedure rejects $H_0^{(1)}, \ldots, H_0^{(\hat{k})}$ where

$$\hat{k} = \max \left\{ 0 \leq k \leq N : p(k) \leq q \frac{k}{N} \right\}$$

$\Leftrightarrow$ region of rejection $R^{BH} = \{ p \leq \hat{t}_q \}$ with $\hat{t}_q = q \frac{\hat{k}}{N}$

learning from the other experiments idea

“testimation problem” : blurs the line between testing and estimation
FDR control of BH procedure

Theorem [Benjamini and Hochberg (1995)]
Under the independence assumption among the tests, BH$_q$ procedure control the FDR at level:

$$\text{FDR} \leq \frac{N_0}{N} q \leq q,$$

where $N_0$ is the number of true null hypotheses

- in practice $N_0$ is unknown and bounded by $N$ ($\pi_0 \equiv \frac{N_0}{N} \leq 1$)
- BH procedure control can be extended beyond independence for special cases of positive dependence [Benjamini and Yekutieli (2001)]
- Typical value of $q$: no real conventional choice in the literature, though $q = 0.1$ seems to be popular
Popularity of FDR and BH procedure

Historical context and citations of the seminal paper [Benjamini and Hochberg, 1995] (many thanks to Marine Roux for the picture)

FDR for Big Data
Large-scale hypothesis testing in many fields
- DNA microarray, genomics, fMRI data, . . . .
- Several works with astronomical imaging applications since the early 2000s
Outline

**Multiple testing error control**
- Basic statistical hypothesis testing concepts
- Family-Wise Error Rate FWER
- False Discovery Rate FDR

**FDR control : Benjamini-Hochberg Procedure**
- BH Procedure
- Bayesian interpretation of FDR
- Empirical Bayes interpretation of BH procedure

**Variations on FDR control and BH Procedure**
- Improving power
- Dependence
- Learning the null distribution
Mixture model

Denote by $X$ one test statistic, and $\Gamma$ any subset of the real line.

**Two Group Mixture model**

$N$ tests statistics are either null or non-null with prior probability

- $\pi_0 \equiv \Pr(H_0)$ (in practice, $\pi_0$ will be often close to 1),
- $\pi_1 \equiv \Pr(H_1) = 1 - \pi_0$,

and respective distributions,

- $F_0(\Gamma) \equiv \Pr(X \in \Gamma|H_0) \leftarrow$ null case
- $F_1(\Gamma) \equiv \Pr(X \in \Gamma|H_1) \leftarrow$ non-null case

The distribution of any $X$ is the mixture with distribution

$$F(\Gamma) = \pi_0 F_0(\Gamma) + \pi_1 F_1(\Gamma)$$
Bayesian Fdr

Classification problem

- We observe $x \in \Gamma$, does it corresponds to the null group?
- Applying the Bayes rule yields the posterior of the null

$$
\Pr (H_0 | X \in \Gamma) = \frac{\pi_0 F_0(\Gamma)}{F(\Gamma)}
$$

Bayesian false discovery rate [Efron (2004,2010)]

- $\Gamma$ is now the region of rejection of the null
- *Bayesian false discovery rate* defined as

$$
\text{Fdr}(\Gamma) \equiv \Pr (H_0 | X \in \Gamma) = \frac{\pi_0 F_0(\Gamma)}{F(\Gamma)},
$$
Bayesian Fdr and positive FDR [Storey (2003)]

Positive FDR : \( \text{pFDR} \equiv E \left[ \frac{U}{R} \middle| R > 0 \right] \)

\( \text{FDR} = \text{pFDR} \times \Pr(R > 0) \)

Theorem [Storey (2003)]

- \( R \equiv R(\Gamma) = \# \text{ discoveries for the region of rejection } \Gamma \)
- \( U \equiv U(\Gamma) = \# \text{ false discoveries for the region of rejection } \Gamma \)

If the \( X_i \) are independent and distributed according to the mixture model,

\[ \text{Fdr}(\Gamma) \equiv \Pr(H_0|X \in \Gamma) = E \left[ \frac{U(\Gamma)}{R(\Gamma)} \middle| R(\Gamma) > 0 \right] \leftarrow \text{Positive FDR} \]

- proof relies on \( U(\Gamma) \mid R(\Gamma) = k \sim \text{binomial distribution } B(k, \text{Fdr}(\Gamma)) \)
- interpretation of a frequentist concept as a Bayesian one
Empirical Bayes Fdr estimate [Efron (2004, 2010)]

- $F$, $F_0$ and $F_1$ denote now the cdf of the mixture, null and non-null
- the test can be assumed to be left-sided: $\Gamma = (-\infty, t]$ and $p_i = F_0(x_i)$

Estimation of $\text{Fdr}(t) = \pi_0 F_0(t)/F(t)$

- $F_0$, assumed to be known,
- $\pi_0$, unknown but usually close to 1,
- $F_1$, unlikely to be known in large-scale inference

However $F = \pi_0 F_0 + \pi_1 F_1$ can be estimated by its empirical distribution:

$$\overline{F}(t) = \#\{x_i \leq t\}/N$$

- does not require to specify $H_1$ : robust to alternative miss-specifications
- empirical Bayes : prior on $F$ estimated from the observations
- empirical Bayes Fdr estimate : $\overline{\text{Fdr}}(t) = \pi_0 F_0(t)/\overline{F}(t)$
Empirical Bayes Fdr estimate [Efron (2004, 2010)]

- $F$, $F_0$ and $F_1$ denote now the cdf of the mixture, null and non-null
- the test can be assumed to be left-sided: $\Gamma = (-\infty, t]$ and $p_i = F_0(x_i)$

Estimation of $\text{Fdr}(t) = \pi_0 F_0(t)/F(t)$

- $F_0$, assumed to be known,
- $\pi_0$, unknown but usually close to 1,
- $F_1$, unlikely to be known in large-scale inference

However $F = \pi_0 F_0 + \pi_1 F_1$ can be estimated by its empirical distribution:

$$\overline{F}(t) = \#\{x_i \leq t\}/N$$

- does not require to specify $H_1$: robust to alternative miss-specifications
- empirical Bayes: prior on $F$ estimated from the observations
- empirical Bayes Fdr estimate: $\overline{\text{Fdr}}(t) = \pi_0 F_0(t)/\overline{F}(t)$
Equivalence between Empirical Bayes Fdr control and BH procedure

Ordered observations $x(1) \leq x(2) \leq \ldots \leq x(N)$

- $F_0(x(i)) = p(i)$, and $\overline{F}(x(i)) = i/N$

$\Rightarrow \overline{\text{Fdr}}(x(i)) = \pi_0 \frac{N}{i} p(i)$

**Fdr control at level $\pi_0 q$**

Given a preselected $q$, find $\hat{t} = \max \{ t : \overline{\text{Fdr}}(t) \leq \pi_0 q \}$

$\Leftrightarrow t = \max_i x(i)$ s.t. $p(i) \leq q \frac{i}{N}$

$\Leftrightarrow$ reject $H_0^{(1)}, \ldots, H_0^{(\hat{k})}$ where $\hat{k}$ is the largest index s.t. $p(k) \leq q \frac{k}{N}$

$\Leftrightarrow$ BH$_q$ procedure

**Fdr control and dependence**

- $\overline{F}(t)$ is an unbiased estimator of $F(t)$ even under dependence,
- $\overline{\text{Fdr}}$ is a rather slightly upward biased estimate of FDR even under dependence [Efron (2010)],
- price of dependence is the variance of the estimator $\overline{\text{Fdr}}(t)$
Equivalence between Empirical Bayes Fdr control and BH procedure

Ordered observations \( x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(N)} \)

\[ F_0(x_{(i)}) = p(i), \quad \text{and } \overline{F}(x_{(i)}) = i/N \]

\[ \Rightarrow \overline{\text{Fdr}}(x_{(i)}) = \pi_0 \frac{N}{i} p(i) \]

Fdr control at level \( \pi_0 q \)

Given a preselected \( q \), find \( \hat{t} = \max \{ t : \overline{\text{Fdr}}(t) \leq \pi_0 q \} \)

\[ \Leftrightarrow t = \max_i x_{(i)} \text{ s.t. } p(i) \leq q \frac{i}{N} \]

\[ \Leftrightarrow \text{reject } H_0^{(1)}, \ldots, H_0^{(\hat{k})} \text{ where } \hat{k} \text{ is the largest index s.t. } p(k) \leq q \frac{k}{N} \]

\[ \Leftrightarrow \text{BH}_q \text{ procedure} \]

Fdr control and dependence

\[ \overline{F}(t) \text{ is an unbiased estimator of } F(t) \text{ even under dependence,} \]

\[ \overline{\text{Fdr}} \text{ is a rather slightly upward biased estimate of FDR even under dependence} \quad \text{[Efron (2010)]}, \]

\[ \text{price of dependence is the variance of the estimator } \overline{\text{Fdr}}(t) \]
Outline

Multiple testing error control
- Basic statistical hypothesis testing concepts
- Family-Wise Error Rate FWER
- False Discovery Rate FDR

FDR control: Benjamini-Hochberg Procedure
- BH Procedure
- Bayesian interpretation of FDR
- Empirical Bayes interpretation of BH procedure

Variations on FDR control and BH Procedure
- Improving power
- Dependence
- Learning the null distribution
Estimation of the proportion $\pi_0$ of true $H_0$

Adaptive BH procedures [Benjamini et al. (2006), Storey et al. (2004)]

- BH procedure overcontrols FDR: $\text{FDR}(\text{BH}_{q}) = \pi_0 q$, where $\pi_0 = \frac{N_0}{N}$
- an upward bias estimator $\hat{\pi}_0$ of $\pi_0$ can be plugged to improve power

Adaptive BH procedure: BH procedure at control level $q/\hat{\pi}_0$ to obtain a FDR control at nominal level $q$

Storey’s $\pi_0$ estimator [Storey et al. (2004)]

- Survival function $G(t) = 1 - F(t)$ of the p-values

$$G(\lambda) = \pi_0 G_0(\lambda) + \pi_1 G_1(\lambda) \geq \pi_0 G_0(\lambda) = \pi_0 (1 - \lambda)$$

for large enough $\lambda$, $G_1(\lambda) \approx 0$, thus $\pi_0 \approx G(\lambda)/(1 - \lambda)$

Based on the empirical survival function, the modified Storey’s estimator is

$$\hat{\pi}_0(\lambda) = \frac{\#\{p_i > \lambda\} + 1}{N(1 - \lambda)}, \text{ for a given } \lambda \in (0, 1),$$
Estimation of the proportion $\pi_0$ of true $H_0$

Adaptive BH procedures [Benjamini et al. (2006), Storey et al. (2004)]

- BH procedure overcontrols FDR: $\text{FDR}(\text{BH}_q) = \pi_0 q$, where $\pi_0 = \frac{N_0}{N}$
- an upward bias estimator $\hat{\pi}_0$ of $\pi_0$ can be plugged to improve power

Adaptive BH procedure: BH procedure at control level $q/\hat{\pi}_0$ to obtain a FDR control at nominal level $q$

Storey’s $\pi_0$ estimator [Storey et al. (2004)]

- Survival function $G(t) = 1 - F(t)$ of the p-values
  \[ G(\lambda) = \pi_0 G_0(\lambda) + \pi_1 G_1(\lambda) \geq \pi_0 G_0(\lambda) = \pi_0 (1 - \lambda) \]
  for large enough $\lambda$, $G_1(\lambda) \approx 0$, thus $\pi_0 \approx G(\lambda)/(1 - \lambda)$

Based on the empirical survival function, the modified Storey’s estimator is
\[
\hat{\pi}_0(\lambda) = \frac{\# \{ p_i > \lambda \} + 1}{N (1 - \lambda)}, \quad \text{for a given } \lambda \in (0, 1),
\]
Adaptive BH procedures: Storey’s $\pi_0$ estimator

Properties of adaptive BH procedure with modified Storey’s estimator $\hat{\pi}_0(\lambda)$

- exact control of FDR at nominal level $q$ for independent tests,
- asymptotic control of FDR in case of weak dependence

Typical values of $\lambda$

Based on various simulations [Blanchard et al. (2009)]

- $\lambda = \frac{1}{2}$ $\leftarrow$ “uniformly” more powerful than other adaptive procedures, but not robust to strong dependences (e.g. equicorrelation of the test statistics)
- $\lambda = q$ $\leftarrow$ powerful and quite robust to long memory dependences
Extension of BH procedure to dependent tests

Positive Regression Dependence on a Subset PRDS [Benjamini et al. (2001)]
BH procedure still controls FDR at nominal value \( q \) when the test statistics vector obey the PRDS property: e.g. for one-sided tests

- Gaussian vector with positive correlations,
- Studentized gaussian PRDS vector for \( q \leq 0.5 \)

Universal bound [Benjamini and Yekutieli (2001)]
For any dependence structure, BH procedure still controls FDR at level

\[
\text{FDR}(\text{BH}_q) \leq \pi_0 q c,
\]

where \( \pi_0 = \frac{N_0}{N} \) and \( c = \sum_{i=1}^{N} \frac{i}{N} \approx \log(N) \)

- too conservative to be useful in practice (more conservative than Bonferroni when the number of rejected test \( \hat{k} \) is lower than \( c \))
Knockoff filter for dependent tests [Barber and Candes (2015)]

Statistical linear model

\[ y = X\beta + \epsilon \]

- \( y \in \mathbb{R}^n \) is the response vector
- \( \epsilon \in \mathbb{R}^n \) is a white gaussian noise vector
- \( X \in \mathbb{R}^{n \times p} \) is a deterministic matrix of the \( p \) column predictors
- \( \beta \in \mathbb{R}^p \) is the weight vector

Multiple testing problem: predictors associated with the response?

- \( H_0^i : \beta_i = 0 \), for \( 1 \leq i \leq p \)

- Knockoff construction to control FDR based on any model selection procedure

- Application to large-scale hypothesis testing, and/or strong local dependences?
Null hypothesis specification diagnosis

- BH procedure requires so little: only the choice of the test statistics and its specification when the null hypothesis is true
- crucial to check that the null is correctly specified before!

Graphical diagnosis

qq-plot of the p-values must be linear for large enough values

correctly specified $H_0$
Null hypothesis specification diagnosis

- BH procedure requires so little: only the choice of the test statistics and its specification when the null hypothesis is true
- crucial to check that the null is correctly specified before!

Graphical diagnosis

qq-plot of the p-values must be linear for large enough values
Learning the null distribution

Deviation from the theoretical null

- theoretical null hypothesis usually derived in an idealized framework, does not account for sample correlations,...
- unlikely to be correctly specified in large-scale testing!
- possibility to detect and correct possible miss-specification of the null hypothesis

Empirical null distribution [Efron 2010]

Parametric $H_0$ estimation: based on the observations that are the most likely under theoretical $H_0$, estimates of the null parameters (and $\pi_0$)

- central matching
- maximum likelihood

Non-parametric $H_0$ estimation

- permutation null distribution
Concluding remarks

- FDR is a very useful global error criterion that allows one to control a trade-off between Type I error and Power

- BH procedure is a very simple and quite robust procedure to control FDR

- Main important problem and challenges still concerns the dependence: how to explicitly account for dependence?
Selected references

Selected references with Astrophysics applications


Great talks on multiple testing / FDR available on the web

Genovese, C. R. “A Tutorial on False Discovery Control,”
http://www.stat.cmu.edu/~genovese/talks/hannover1-04.pdf


Benjamini Y. (2005) “Discovering the False Discovery Rate,”
Supplementary materials

Some proofs of the BH procedure FDR control

- FDR control under independence
- FDR control under PRDS property
- universal FDR bound for an arbitrary dependence structure
proof of the FDR control of BH procedure

- $\mathcal{R} \equiv$ set of discoveries, $\mathcal{H}_0 \equiv$ set of $N_0$ true null hypotheses
- indicator trick: for discrete random variables $E[A|B = b] \Pr(B = b) = E[A \times 1_{B=b}]$

$$\text{FDR} = \sum_{k=1}^{N} E \left[ \frac{|\mathcal{R} \cap \mathcal{H}_0|}{k} \Bigg| |\mathcal{R}| = k \right] \Pr(|\mathcal{R}| = k) = \sum_{k=1}^{N} \frac{1}{k} E \left[ |\mathcal{R} \cap \mathcal{H}_0| \times 1_{|\mathcal{R}| = k} \right],$$

$$= \sum_{k=1}^{N} \frac{1}{k} E \left[ \sum_{i \in \mathcal{H}_0} 1_{p_i \leq q \frac{k}{N}} \times 1_{|\mathcal{R}| = k} \right] = \sum_{i \in \mathcal{H}_0} \sum_{k=1}^{N} \frac{1}{k} \Pr \left( \hat{k} = k, p_i \leq q \frac{k}{N} \right),$$

$$= \frac{q}{N} \sum_{i \in \mathcal{H}_0} \sum_{k=1}^{N} \Pr \left( \hat{k} = k \bigg| p_i \leq q \frac{k}{N} \right),$$

where the last equality comes from $p_i \sim U_{[0,1]}$ under the null (this becomes an inequality $\leq$ if $p_i$ is assumed to be stochastically greater than $U_{[0,1]}$ under the null).
proof of the FDR control of BH procedure (cont’d)

Independent case

\[ \hat{k}^i \equiv \text{number of discoveries except the } i\text{th test (r.v in } \{0, \ldots, N-1\}) \]

\[ \Pr(\hat{k} = k \mid p_i \leq q \frac{k}{N}) = \Pr(\hat{k}^i = k - 1 \mid p_i \leq q \frac{k}{N}) = \Pr(\hat{k}^i = k - 1) \]

\[ \text{FDR} = \frac{q}{N} \sum_{i \in \mathcal{H}_0} \sum_{k=1}^{N} \Pr(\hat{k} = k \mid p_i \leq q \frac{k}{N}) = \frac{q}{N} \sum_{i \in \mathcal{H}_0} \sum_{k=0}^{N-1} \Pr(\hat{k}^i = k - 1) = \frac{q}{N} \sum_{i \in \mathcal{H}_0} 1 = \frac{N_0}{N} q \]

PRDS case

\[ \Pr(\hat{k} \leq k - 1 \mid p_i \leq q \frac{k-1}{N}) \]

\[ \sum_{k=1}^{N} \Pr(\hat{k} = k \mid p_i \leq q \frac{k}{N}) = \sum_{k=1}^{N} \Pr(\hat{k} \leq k \mid p_i \leq q \frac{k}{N}) - \Pr(\hat{k} \leq k - 1 \mid p_i \leq q \frac{k}{N}) \]

\[ \leq \sum_{k=1}^{N} \Pr(\hat{k} \leq k \mid p_i \leq q \frac{k}{N}) - \Pr(\hat{k} \leq k - 1 \mid p_i \leq q \frac{k-1}{N}) \]

\[ = \Pr(\hat{k} \leq N \mid p_i \leq q) - \Pr(\hat{k} \leq 1 \mid p_i \leq q \frac{1}{N}) + \Pr(\hat{k} = 1 \mid p_i \leq q) = 1, \]

thus \( \text{FDR} \leq \frac{N_0}{N} q \)
proof of the FDR control bound for arbitrary dependence

Arbitrary dependence

\[
\frac{1}{k} = \frac{1}{k-1} - \frac{1}{k(k-1)}
\]

\[
\text{FDR} = \sum_{i \in \mathcal{H}_0} \sum_{k=1}^{N} \left[ \frac{1}{k} \Pr \left( \hat{k} \leq k, p_i \leq q \frac{k}{N} \right) - \frac{1}{k} \Pr \left( \hat{k} \leq k-1, p_i \leq q \frac{k}{N} \right) \right],
\]

\[
\leq \sum_{i \in \mathcal{H}_0} \sum_{k=1}^{N} \frac{1}{k} \Pr \left( \hat{k} \leq k, p_i \leq q \frac{k}{N} \right) - \sum_{i \in \mathcal{H}_0} \sum_{k=2}^{N} \frac{1}{k} \Pr \left( \hat{k} \leq k-1, p_i \leq q \frac{k-1}{N} \right),
\]

\[
= \sum_{i \in \mathcal{H}_0} \sum_{k=1}^{N} \frac{1}{k} \Pr \left( \hat{k} \leq k, p_i \leq q \frac{k}{N} \right) - \sum_{i \in \mathcal{H}_0} \sum_{k=2}^{N} \frac{1}{k-1} \Pr \left( \hat{k} \leq k-1, p_i \leq q \frac{k-1}{N} \right),
\]

\[
+ \sum_{i \in \mathcal{H}_0} \sum_{k=2}^{N} \frac{1}{k(k-1)} \Pr \left( \hat{k} \leq k-1, p_i \leq q \frac{k-1}{N} \right),
\]

\[
\leq \sum_{i \in \mathcal{H}_0} \frac{1}{N} \Pr (p_i \leq q) + \sum_{i \in \mathcal{H}_0} \sum_{k=2}^{N} \frac{1}{k(k-1)} \Pr \left( p_i \leq q \frac{k-1}{N} \right),
\]

\[
= \frac{N_0}{N} q + \frac{N_0}{N} q \left( \frac{1}{2} + \ldots + \frac{1}{N} \right) = \frac{N_0}{N} q \sum_{j=1}^{N} \frac{1}{j}.
\]