Hierarchical Testing of Families of Hypotheses

Yoav Benjamini

Tel Aviv University

(visiting Stanford and Berkeley)

based on joint work with

Marina Bogomolov

The Technion

Presented at The San Francisco Chapter of ASA

February 29, 2012

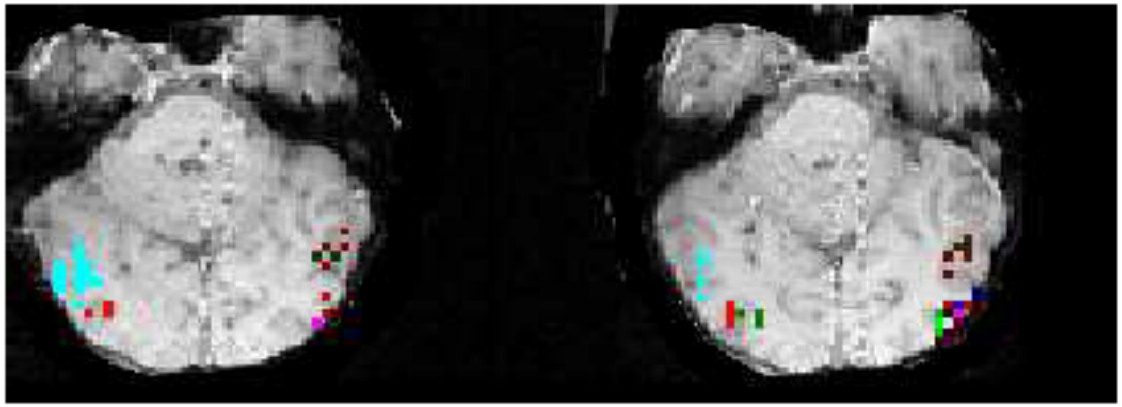
Outline of Talk

- Families of hypotheses
- Combined and separate testing
- Selective inference problem
- A simple selection adjusted testing of families
- The TICE analysis

Ex 1. Clusters of Voxels in Brain Regions

- Divide area to spatially contiguous clusters (ideally of similar response)
- The voxels in each cluster form a family
- The family of clusters is a family of families

Sample slices 9-10: each detected cluster in a different color.



Heller &YB (‘08)

Ex. 2: Voxelwise GenomeWise Association tudy

(Stein et al.'10)

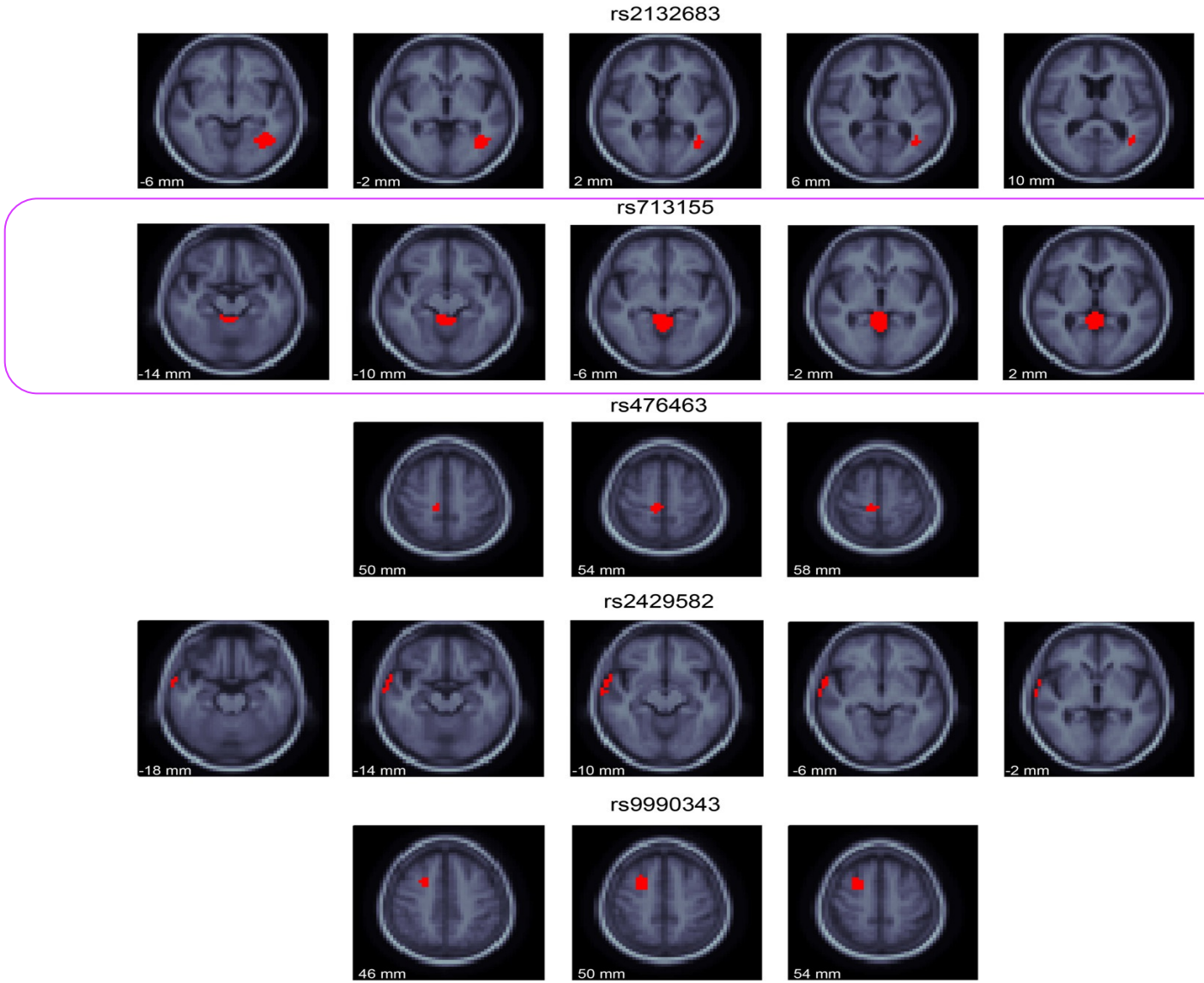
- Alzheimer's Disease Neuroimaging Initiative (ADNI) study: 2003-2008
- Goal: determine biological markers of Alzheimer's disease by searching for associations between volume changes at voxels with genotype Data on: 173 Alzheimer 361 Mild cognitive impairement 206 normals
- Method: Correlate between volume difference and number of minor alleles after adjusting for gender & age: a test for each pair of one of 31622 voxels and one of 448293 SNPs (adjusting for age and gender)

Results for the Alzheimer's disease

- For q=0.05, BH no SNP is found to be associated with voxel size change
- For q=0.5, there are 2 SNPs
- They select 5 top SNPs for further research.
- The display shows first a selection of SNPs,

  then a selection of voxels within each SNP
- The analysis conducted started from voxels defining the families (indirectly), assures control at the level of family but not within the brain region displayed per SNP.

The locations of associated voxels per SNP, for the 5 most associated SNPs



Research Questions - and related families

1. What genes exhibit some association with brain volume ?

A single family with one (intersection) hypothesis per gene; or

2. For our most promising genes where in the brain (voxels) can we detect association?

A family of genes each one having a (sub)-family of voxels associations with it.

(?) What voxels exhibit some association with genes ?

A single family with one (intersection) hypothesis per voxel; or

(??) For our most promising genes ….

More familiy of families

- The genes whose expression is under study are divided into sets (families) that belong to the same pathways (e.g. GO terms)
- Multiple contrasts per gene
- More traditional: Multifactor ANOVA

  For each significant factor test its family of pairwise comparisons

Last example: from Efron(2008)

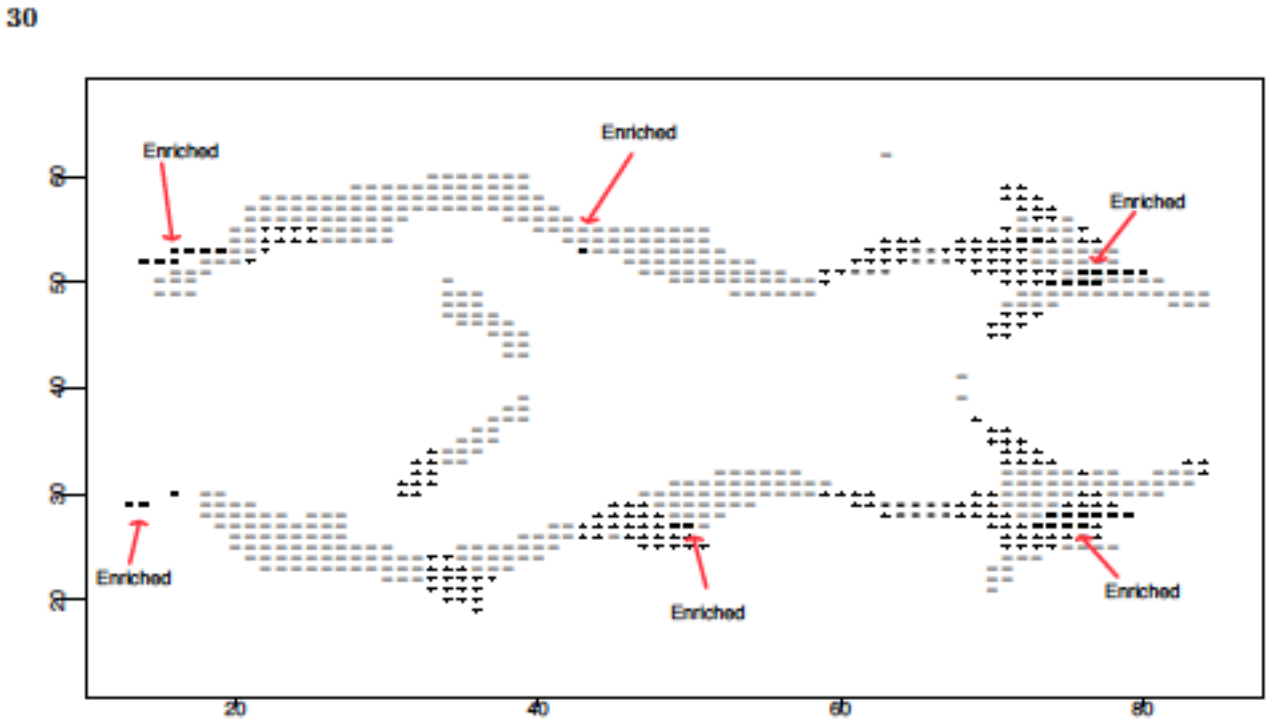


Fig 10: *Enrichment analysis of Imaging data, Panel D of Figure 1; z-value for original 15445 voxels have been averaged over "gene-sets" of neighboring voxels with city-block distance* ≤ 2. *Coded as "-" for* $\bar{z}_i < 0$, *"+" for* $\bar{z}_i \geq 0$; *solid rectangles, labeled as "Enriched", show voxels with* $\widehat{\text{fdr}}(\bar{z}_i) \leq 0.2$, *using empirical null.*

Notations for multiple families

m families of null hypotheses tested;

each family has m_i hypotheses m_{0i} of which are true.

R_i hypotheses are rejected in family i, V_i of them are rejected in error. $i=1,2,\ldots, m_i$.

$$V = \sum_{i=1}^{m} V_i \quad R = \sum_{i=1}^{m} R_i$$

Over-all FDR involves $Q = V / R$

$= 0$ if $R = 0$

Within family FDR involves $Q_i = V_i / R_i$

$= 0$ if $R_i = 0$

Combined FDR testing of all hypotheses

+ Offers over-all FDR control

+ Offers approximate within family FDR control,
  if the families are similar (i.e. $m_{0i}/m_i \sim$ constant<1)

- If the families are not similar inference may be distorted within the family:

Family 1: H_{0j} true for all j Family 2: H_{0j} false for all j

Using the BH FDR controlling procedure the cutoff is

$p_{(i)} \leq cq$ for some c, $0 < c < 1$.

In family1 the threshold for discovery will be too high

In family 2 " too low

Efron's comment (2008)

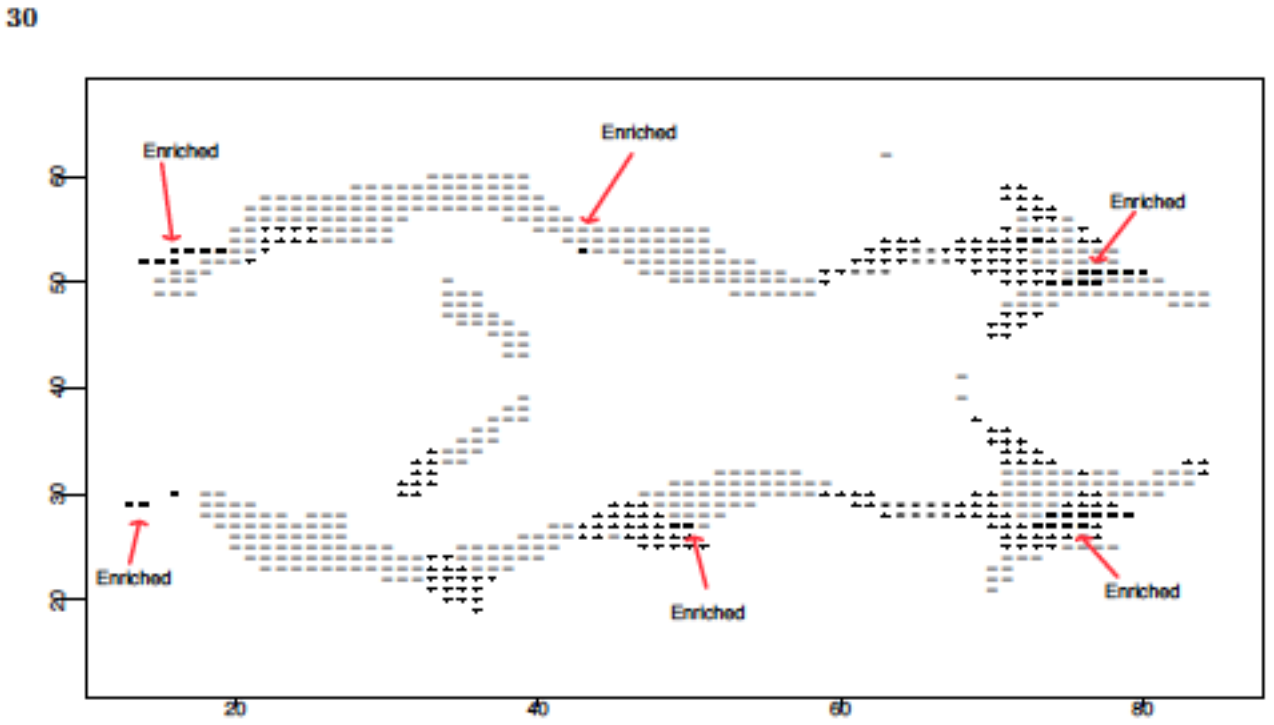


Fig 10: *Enrichment analysis of Imaging data, Panel D of Figure 1; z-value for original 15445 voxels have been averaged over "gene-sets" of neighboring voxels with city-block distance* ≤ 2. *Coded as "-" for* $\bar{z}_i < 0$, *"+" for* $\bar{z}_i \geq 0$; *solid rectangles, labeled as "Enriched", show voxels with* $\widehat{\text{fdr}}(\bar{z}_i) \leq 0.2$, *using empirical null.*

Separate FDR testing

+ The multiplicity problem even further reduced

+ The inference within each family legitimate

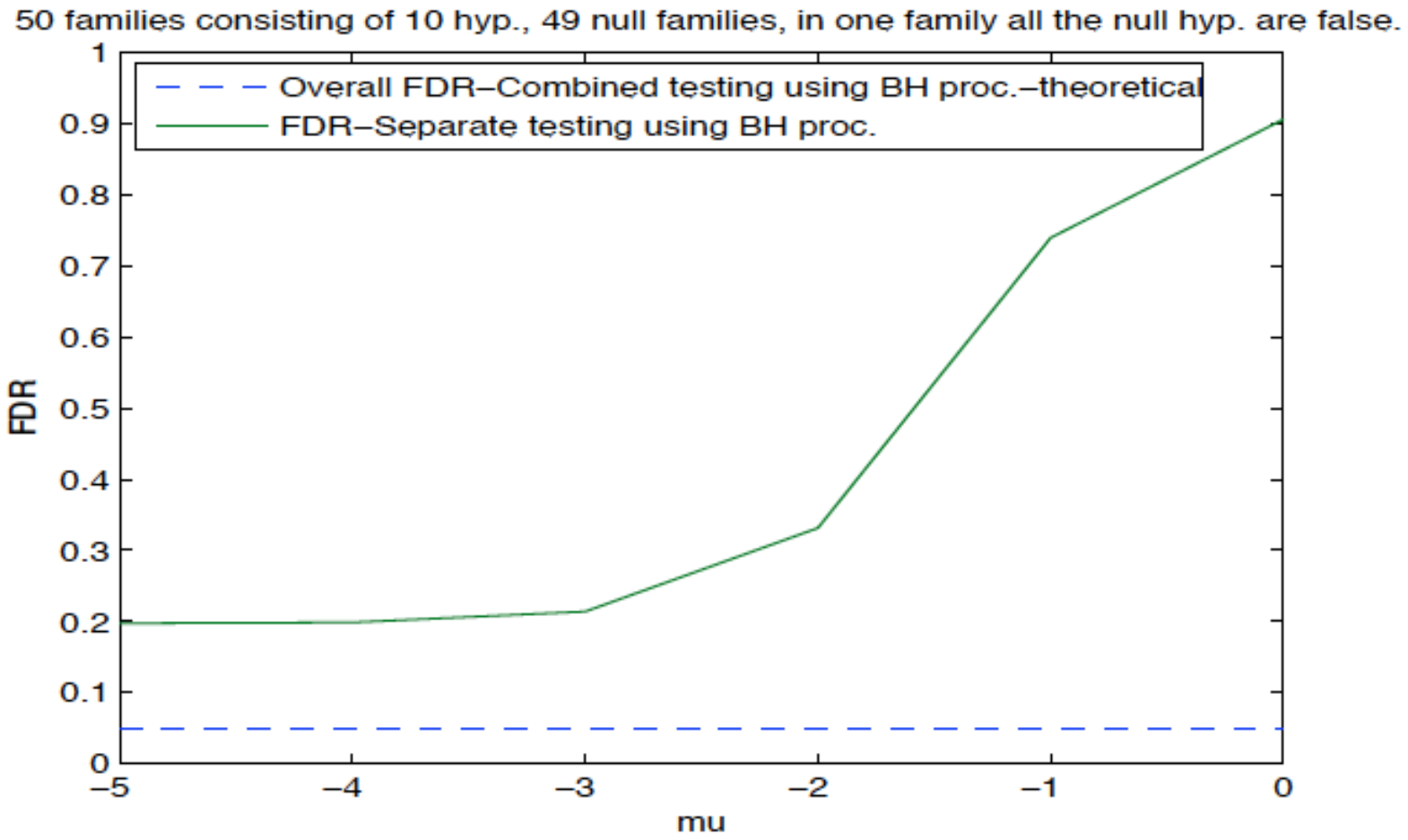
What about the overall FDR control?

- If the families are similar (m_{0i}/m_i ~ constant < 1) the overall FDR ≤ q – scalability

(asymptotic result)

But if not:

A highly non-homogeneous example



- m=50 m_i=10 m_{0i}/m_i=1 for i-1,…,49 $m_{0,50}/m_{50}$=0

eBayes Justification for separate FDR testing

Efron ('08, '10 Ch. 10): Under the mixture model:

If Bayesian false discovery rate is controlled at q for every subset

then unconditionally $Fdr = q$

Which is often achievable in large subsets (but we'll return to it at the end).

Similar justification in the frequentist framework:

Consider $Q_i = V_i / R_i$

Justification for separate testing ?

- Control $E(Q_i)$ separately for each family *i* and get for free control of

  the average over all families!

$$E\left(\frac{\sum_i Q_i}{m}\right) = \frac{\sum_i E(Q_i)}{m} \le \frac{mq}{m} = q$$

- But if only some of the families are selected (or highlighted for presentation or to be acted upon) based on the same data, control on the average over the selected ones is not guaranteed
- and selection is the more common situation

Selection is common

1) Screen for genes where significant differences between strains exist (using Analysis of Variance) as was done before ….

Then test within the selected genes what brain region is different from the average of the others

Sometimes called ’ Template matching’ (Pavlidis)

(see also Smyth et al, in LIMMA)

2) Show top 5 / top 10 / top 100 genes, then study a family of contrasts or association for each

The multi-family selective inference problem

We select interesting/significant/promising families

The uninteresting families loose importance

and are dropped/ignored from the reported results

(or hidden in the available database/online appendix)

We wish to infer on the selected families

- test hypotheses within
- set confidence interval
- estimate

A problem for other within family error-rated as well

Example: lack of control over selected

- Error-rate is FWER i.e. $E(C_i)$, where $C_i = I_{\{V_i \geq 1\}}$
- Select family *i* if $min_j(P_{ij}) \leq q$
- Use Bonferroni at level 0.05 at each family

m	m_i	$E(C_S)$
20	100	0.049
100	20	0.076
100	10	0.122
100	2	0.506

The more severe the selection the worse the control

A variety of error-rates

Unadjusted inference	$E(V/m) \leq \alpha$
False Discovery Rate	$E(V/R) \leq \alpha = q$
Strong control of FWER	$Pr(V \geq 1) \leq \alpha$
Per family Error-rate	$E(V) \leq \alpha$

A variety of error-rates

Unadjusted inference	$E(V/m) \leq \alpha$
False Excedance Rate	$Pr\ (V/R \geq q) \leq \alpha$
k- FDR	$E(\ (V-k)_+/R\ ) \leq q$
False Discovery Rate	$E(\ V/R\ ) \leq \alpha = q$
k-FWER	$P(\ V \geq k\ ) \leq \alpha$
Strong control of FWER	$Pr\ (\ V \geq 1\ ) \leq \alpha$
Per family Error-rate	$E(\ V\ ) \leq \alpha$

A variety of error-rates

Unadjusted inference	$E(V/m) \leq \alpha$
False Excedance Rate	$Pr\ (V/R \geq q) \leq \alpha$
k- FDR	$E(\ (V-k)_+/R\ ) \leq q$
False Discovery Rate	$E(\ V/R\ ) \leq \alpha = q$
k-FWER	$P(\ V \geq k\ ) \leq \alpha$
Strong control of FWER	$Pr\ (\ V \geq 1\ ) \leq \alpha$
Per family Error-rate	$E(\ V\ ) \leq \alpha$

All above are of the form $E(\mathcal{C})$

But not Fdr = E(V)/E(R); local fdr(z); positive FDR

Selection adjusted separate testing of families

Let P_i be the p-values for the hypotheses in family i,

$P=\{P_1,P_2,\ldots P_m\}$. $I=\{1,2,\ldots m\}$.

Any data based selection procedure of families yields $S(P)$ in I. Let $|S(P)|$ be the (random) number of families selected.

The control of error E(C) (FDR, FWER, and others) on the average over the selected families means

$$E\left(\frac{\sum_{i\in S(P)} C_i}{|S(P)|}\right) \le q$$

Selection adjusted separate testing

Theorem:

For any "simple" selection procedure $S(P)$, and for any error-rate of the form $E(C_i)$, if the P_i across families are independent,

controlling $E(C_i) \le q|S(P)|/m$ for all i,

assures control on the average over the selected at level q.

$$E(C_i) \le \frac{|S(P)|}{m} q$$

Note 1: if only one selected - amounts to q/m;

if all selected no adjustment needed

…more notes

- Note 2: P_{ij} within a family need not be independent - the testing procedure should be valid under their dependency
- Note 3: "Simple" is not that simple and includes many natural selection rules:
  - Thresholding on any function of the family's p-values
- In particular multiple testing procedure can be used
  - Stepwise FWER and FDR controlling procedures
- The "Simple" does not include adaptive (plug-in) methods. Generalizing |S(P)| makes it work - as in selective Confidence Intervals Yekutieli & BY ('06)

The general case:

1) Apply the selection criterion $S(\boldsymbol{P})$

2) For each $i \in S(\boldsymbol{P})$, partition $\boldsymbol{P}$ to $\boldsymbol{P}_i$ and $\boldsymbol{P}^{(i)}$

Let $R_{min} = \min_p \{ |S(\boldsymbol{P}^{(i)}, P_i=p)| \mid i \in S(\boldsymbol{P}^{(i)}, P_i=p \}$

3) Continue as before

Actually, a simple selection procedure is one in which

$R_{min} = |S(\boldsymbol{P})|$

Note 4: The adjustment is closely connected to the False Confidence-statement Rate for selective Confidence Intervals in YB&Yekutiely (2005)

- There was no restriction on the selection rule
- In particular for each family calculate a p-value for the intersection hypothesis $H_{0,i} = \bigcap_j H_{0,i,j}$
- test across families

It has the desired properties

Within family FDR ,

Average FDR over selected,

Across families FDR (or any other error-rate).

Heller & YB ('08), Sun & Wei ('10+) False Sets Rate

The hierarchical testing tool for families

- Calculate a p-value for the intersection hypothesis for the family.
- Use a testing procedure Proc_1 at some level q_1
- In each selected families test with procedure Proc_2 at some level q_2

  We denote this procedure $(\text{Proc}_1\text{-}q_1, \text{Proc}_2\text{-}q_2)$

- Let's look closer at (BH-q,BH-qR/m)

(BH-q, BH-q) - hierarchical testing

- If we further use $p_{i,(1)}^{BH}$ to test the intersection

Theorem

Selection adjusted FDR for (BH-q , BH-qR/m) holds under positive regression dependency

(also for correlated two-sided Gaussians)

Re-analysis of SNP-voxel data for Alzhiemer

- Family = the set of all association hypotheses for a specific SNP and all voxels (~34K)

  (So selection of families = selection of SNPs)

  Calculate p-value per SNP-family

- Test SNPs while controlling FDR over SNPs
- Test voxels within familes of selected SNPs, assuring FDR control on the average over the selected.

Practically

- Calculate p-value for association between a single SNP and a single voxel. ~1.43 billion p-values

  Can't keep easily: store p-values < .1

- Calculate p-value for intersection hypothesis per SNP-family using Simes test

- Test SNPs while controlling FDR over SNPs: 35 SNPs

- Test voxels within familes of selected SNPs, assuring FDR control on the average over the selected – using BH at level .05*35/34,000

  For most SNPs ≤ 50 voxels; the max 400.

Transcriptomics in Cancer Epidemiology

Using the NOWAC sample of Norwegian Women

- Filled life-style questionnaire
- Donated blood sample that was frozen and stored

In TICE: A breast cancer patient who belongs to NOWAC sample is matched by age and follow-up time with a healthy NOWAC woman

- Gene expression is evaluated for both (same chip)
- Study offers unique opportunity to compare differences in gene expression between case and control as a function of time prior to diagnosis and of risk factors at that time.

Transcriptomics in Cancer Epidemiology TICE

Hence, opportunity to study

Risk factors =>>

Change in gene expression =>>

Breast Cancer

- Expression diff in gene ~ risk factors + background factors + … + follow-up time
- For each gene – a family of tests

Study led by Eiliv Lund, Tromso Univ.

Statistical group: Rosenhalc, Thalabard, Plancade, Neil, Bovestad, Ferro, Gorfine, Heller, MB & YB

TICE Pilot Results

- Data:150 case-control pairs; limited set of variables

Gene expression diff. ~ menopause+ smoking+ BMI,
+ age + HRT+ follow-up time

- For each gene – a family of tests
- Selecting genes (families) by p-values of F-tests for regression we found 81 genes at FDR level 0.05
- Within each family: one-rejection in 24 families (22 BMI, 2 smoking), all others 0, at FDR level 0.05.
- No significant time dependency yet (later 600)

(time dependency not expected to be linear)

(BH-q', BH-q') - hierarchical testing

The (BH-q', BH-q' R/m) procedure, with $q' = 1-(1-q)^{1/2}$, offers overall FDR control at q

It has thus all the desired properties

Within family FDR ,

Selection adjusted FDR,

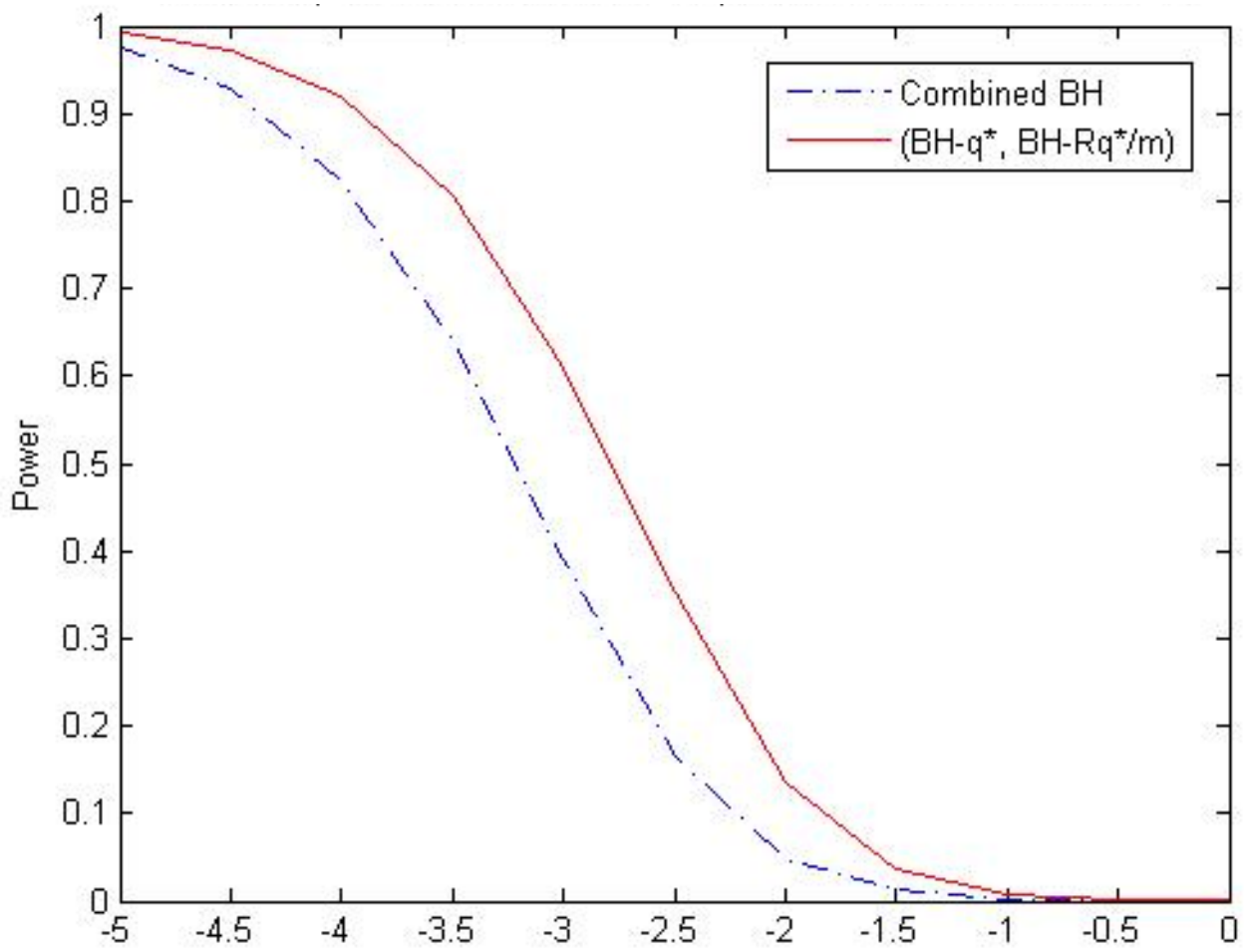
Across families FDR

Over-all FDR.

Hu, Zhao & Zhou ('10) FDR control with groups, set to get improved power by using groups and weighting. As in eBayes FDR problem with all null families

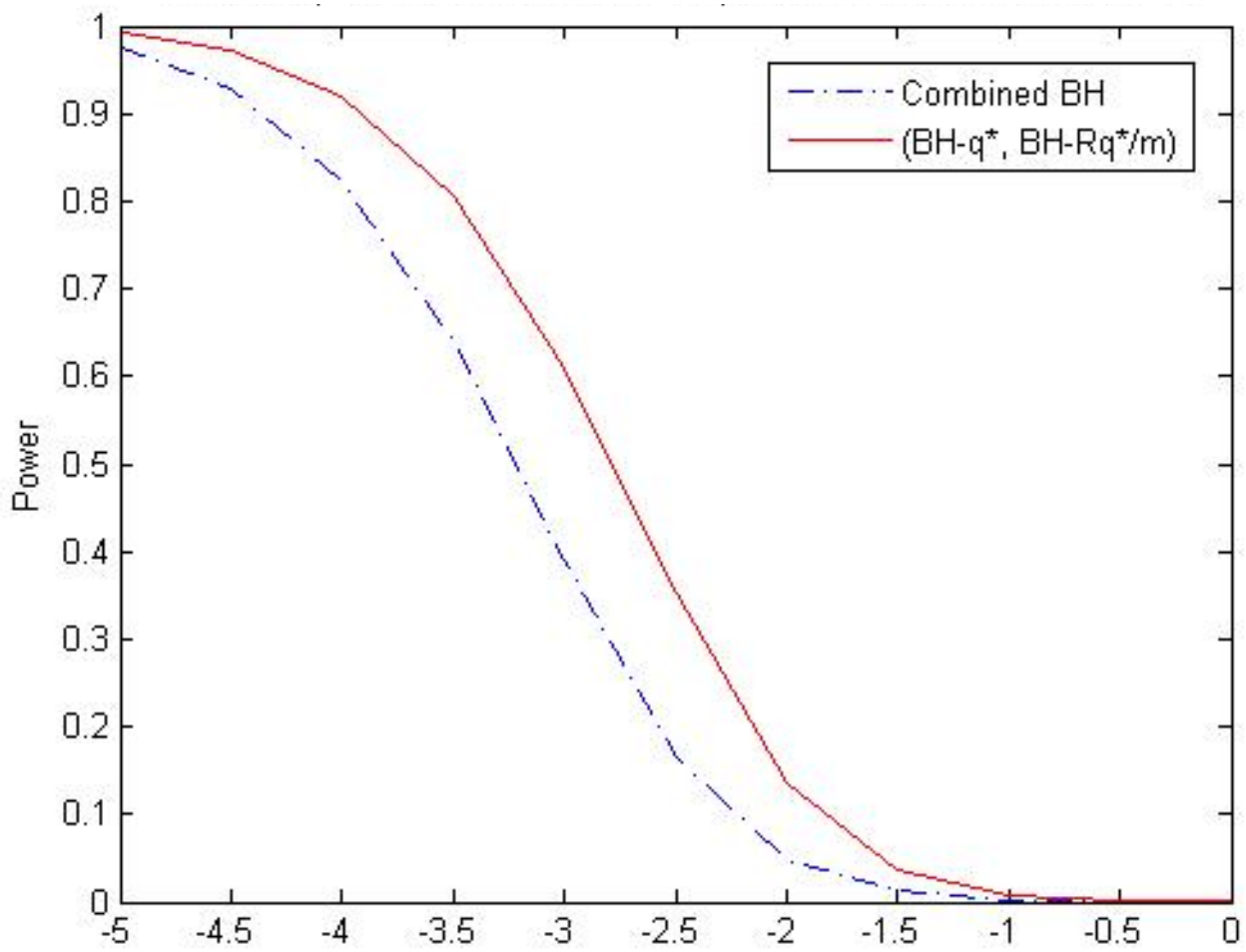
Power gain

8 families of 100 all null; 2 families of 10 all at a



Power gain

8 families of 100 all null; 2 families of 10 all at a



But not always a

(BH-q', BH-q') - hierarchical testing

The (BH-q', BH-q' R/m) procedure, with $q' = 1-(1-q)^{1/2}$, offers overall FDR control at q

It has thus all the desired properties

Within family FDR ,

Selection adjusted FDR,

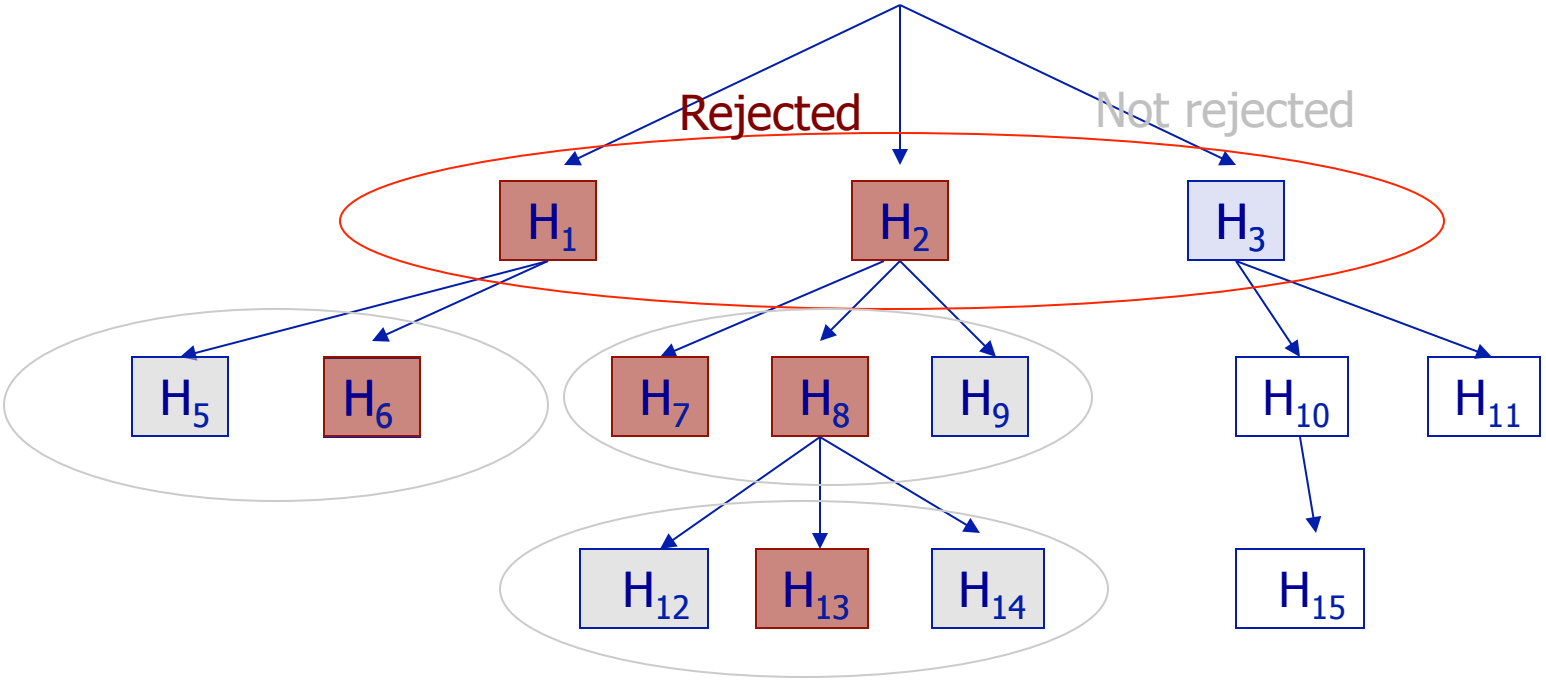
Across families FDR

Over-all FDR.

Why do I call such methods hierarchical?

Hierarchical FDR testing of tree of hypotheses

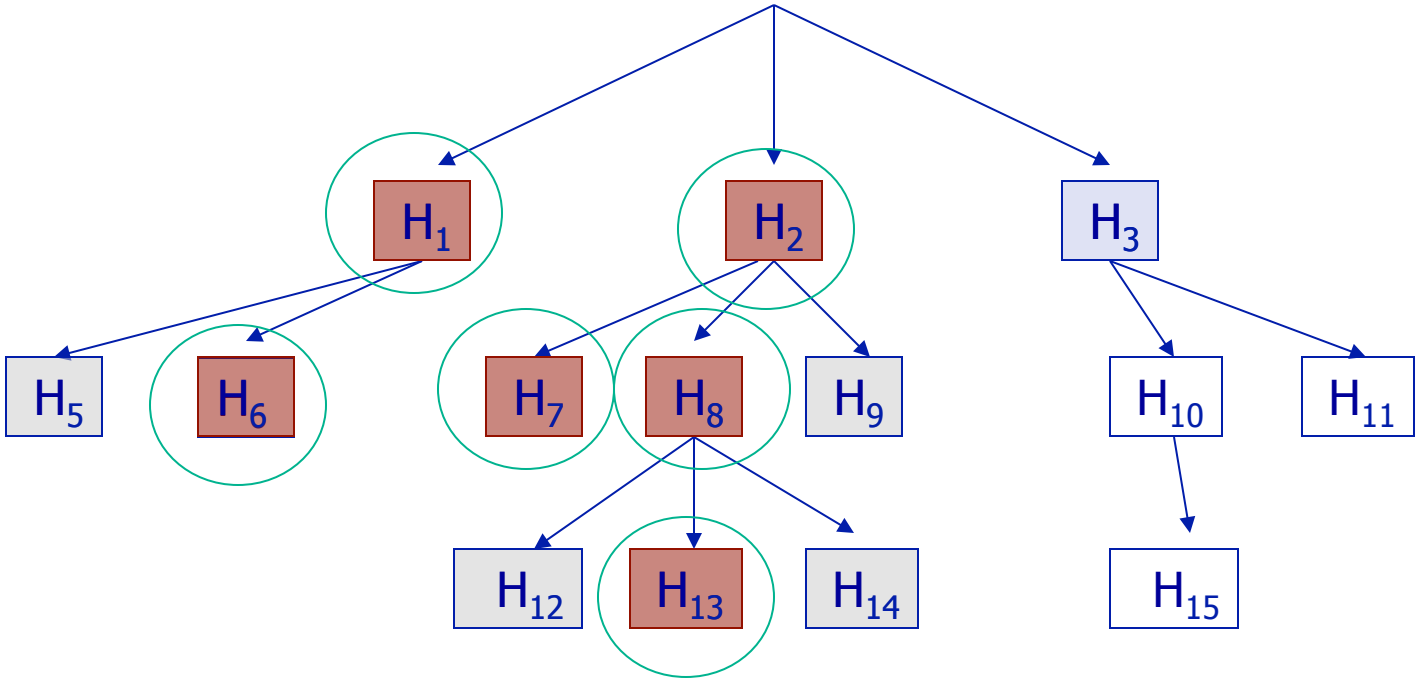
(Yekutieli et al `06, Yekutieli ‘08)



1. Arrange hypotheses in sub-families corresponding to a single parent hypotheses

2. Test sub-family of a rejected parent hypothesis by the procedure in BH at q

Full tree FDR



Theoretical results: independent test statistics FDR upper bound (any sized tree) $FDR_{full} < 2 \times \delta^* \times q$, where $\delta^* < 1.44$

In more realistic settings and in simulations: FDR $\approx$ q (depending on # sub-trees visited < # of discoveries)

Differences

In the previous work

independent between parent and child hypotheses

all hypotheses considered jointly

Here we were interested with controlling FDR

Average over the selected families,

At level 1 (across families) and

Within each family, and

At level 2 (overall FDR) All at the same time

The challenge is higher - but so is its importance

Additional and Future work:

We currently study procedures such as

(BH-q', BH-q'')

As well as multi-stage procedure,

with an eye on all properties

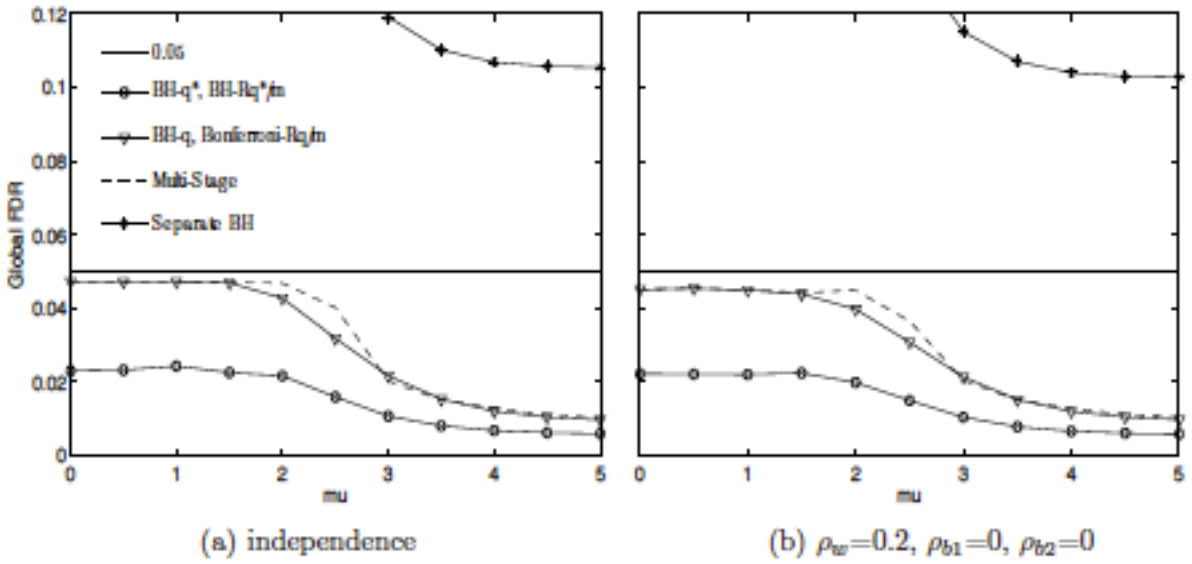
I discussed exclusively

selective inference across families,

simultaneous inference across families is available

Additional and Future work:

If all we care is global FDR – when is "large" large enough?

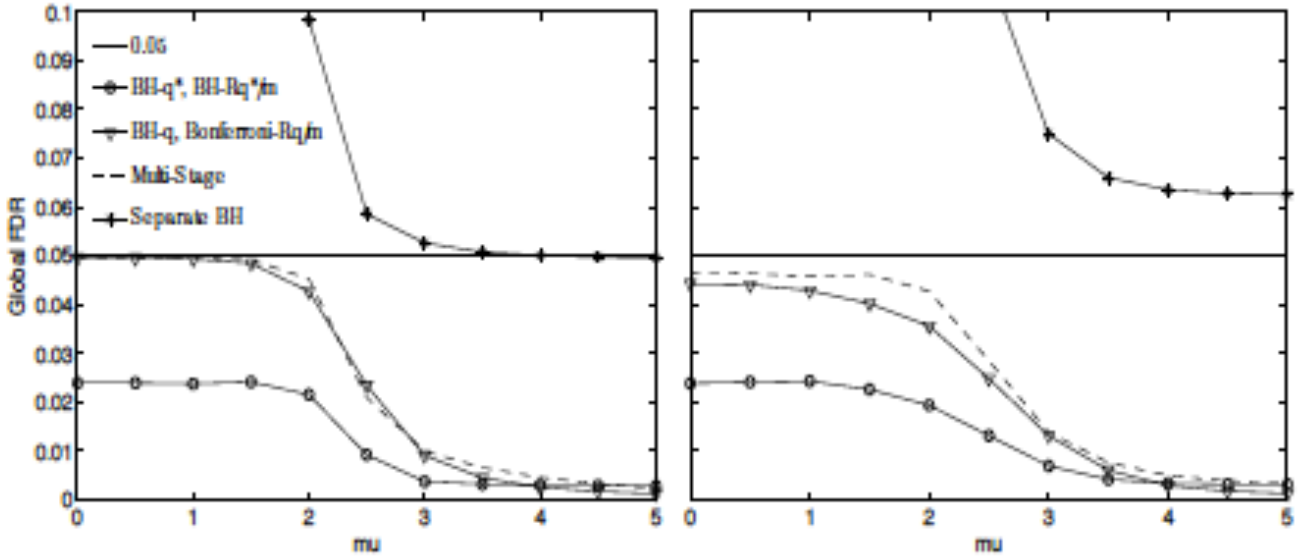


← Separate BH

Figure 6.3: Global FDR of different procedures where the settings are as follows. There are 1000 families, 10 hypotheses in each. There are 900 families containing only true null hypotheses, 90 families containing 5 false null hypotheses and 10 families containing 9 false null hypotheses (configuration ELS2). The simulations were conducted for four dependency

Additional and Future work:

If all we care is global FDR – when is “large” large enough?



(a) independence (b) ρ_w=0.2, ρ_{b1}=0, ρ_{b2}=0

Separate BH

Figure 6.4: Global FDR of different procedures where the settings are as follows. There are 10 families, 1000 hypotheses in each. There are 9 families containing only true null hypotheses, one family contains 100 false null hypotheses (configuration ESL1). The simulations were

Final Practical Comments: Recognizing families

Family; family of families; but what is a family?

A family should best be defined by the danger of selective inference that is being faced:

A family is the smallest set of items of inference in an analysis from which selection of results for presentation and highlighting was made

Different researchers can have different goals and thus define differently the families – still decisions can be defendable and with no arbitrariness.

Final Practical Comments

Once families are defined decide:

Do you select inferences

- Only on all elements with no reference to the ‘family tag’?
- On elements in each family as part of that family?
- On the families and their elements

In each level, how strict control is needed:

FDR? FWER? Other?

Final Practical Comments: creating families

- When pooling together related hypotheses

For maximum benefit, generated families should be as dissimilar as possible :

containing either most true hypotheses or most false

- Clustering based on preliminary data
- Clustering based on previous information
- Clustering based on same data on which testing is done: great prospects - serious additional difficulties

As always, watch out for hidden selection.

Thanks