Inference following aggregate-level hypothesis testing in large scale genomic data

Ruth Heller

www.math.tau.ac.il/∼ruheller

Joint work with Nilanjan Chatterjee, Abba Krieger, and Jianxin Shi

Outline

1. A brief review of the multiple comparisons problem.
2. Inference following selection by aggregate level testing:
   - (i) Goal.
   - (ii) The conditional approach.
   - (iii) An existing alternative.
   - (iv) An empirical comparison.
   - (v) Conclusions.

The multiple comparisons problem

- A family of m null hypotheses are considered: $H_1, \ldots, H_m$.
- $P_1, \ldots, P_m$ are the p-values for testing $H_1, \ldots, H_m$, respectively.
- The hypotheses can be divided into two types:
  1. m_0 true null hypotheses : $P_i \sim U(0,1)$.
  2. $m_1 = m - m_0$ false null hypotheses: $P(P_i \leq x) \geq x, \forall x \in [0,1]$.
- A discovery is made if a null hypothesis is rejected.
- A false discovery is made if a true null hypothesis is rejected.

The two most common error rates

- R = the number of discoveries.
- V = the number of false discoveries.
- The familywise error rate (FWER) is $Pr(V > 0)$.
- The false discovery rate (FDR[1]) is $E\left(\frac{V}{\max(R,1)}\right)$.
- The two error rates coincide if $m_0 = m$.
- Procedures that control the FWER offer also FDR control:

$$E\left(\frac{V}{\max(R,1)}\right) \leq E(I[V > 0]) = Pr(V > 0).$$

[1] Benjamini and Hochberg, 1995. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing.

The Bonferroni Procedure

Reject H_i if $p_i \leq \alpha/m$.

Properties:

- FWER is controlled at level α:

$$Pr(V > 0) = Pr(\cup_{i \in I_0} P_i \leq \alpha/m) \leq \sum_{i \in I_0} Pr(P_i \leq \alpha/m) = m_0 \alpha/m \leq \alpha,$$

  where $I_0 \subseteq \{1, \ldots, m\}$ is the subset of true null hypotheses.

- The FWER error control is valid for any type of dependency across the p-values $P_1, \ldots, P_m$.

The BH procedure

1. Sort the p-values $p_{(1)} \leq \ldots \leq p_{(m)}$, with corresponding $H_{(1)}, \ldots, H_{(m)}$.
2. Find $R = \arg\max_{j=1,\ldots,m}\{p_{(j)} \leq \alpha j/m\}$.
3. Reject $H_{(1)}, \ldots, H_{(R)}$.

Properties:

- $FDR = \frac{m_0}{m}\alpha$ if the p-values are independent[1].
- $FDR \leq \frac{m_0}{m}\alpha$ if the p-values are positive dependent[2].
- $FDR \leq (1 + 1/2 + \ldots + 1/m)\frac{m_0}{m}\alpha \approx \log(m)\frac{m_0}{m}\alpha$ for any type of dependence across the p-values[2].

[1] Benjamini and Hochberg, 1995. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing.

[2] Benjamini and Yekutieli, 2001. The control of the false discovery rate in multiple testing under dependency.

The adjusted p-values

A multiple comparison procedure adjusted p-value for a hypothesis is the smallest nominal level at which the hypothesis would be rejected, given $p_1, \ldots, p_m$.

- The Bonferroni-adjusted p-value for H_i is

$$m \times p_i.$$

  The Bonferroni procedure at level α rejects H_i if and only if $m \times p_i \leq \alpha$.

- The BH-adjusted p-value for $H_{(i)}$ is

$$\min_{j \geq i} \left\{ \frac{m \times p_{(j)}}{j} \right\}.$$

  The BH procedure at level α rejects H_i if and only if $\min_{j \geq i} \left\{ \frac{m \times p_{(j)}}{j} \right\} \leq \alpha$.

Final remarks

- The BH-adjusted p-values are at most as large as the Bonferroni adjusted p-values.
- Bonferroni provides simultaneous inference: the FWER guarantee is valid for any subset of $\{1, \ldots, m\}$.
- BH provide selective inference: the FDR guarantee is for the selected set of rejected hypotheses.
- More generally, with simultaneous inference the guarantee is for every possible subset, whereas with selective inference the guarantee is for the specific subset selected. Methods that assure simultaneous inference also assure selective inference, but not vice versa[3].

[3] Benjamini, 2010. Simultaneous and selective inference: Current Successes and future Challenges.

Outline

1. A brief review of the multiple comparisons problem.
2. Inference following selection by aggregate level testing:
   - (i) Goal.
   - (ii) The conditional approach.
   - (iii) An existing alternative.
   - (iv) An empirical comparison.
   - (v) Conclusions.

Multiple studies testing similar hypotheses

Examine m features in each of n studies. For feature (row) i:

- $H_{ij}, j = 1, \ldots, n$ are the n null hypotheses.
- $H_{iG} = \cap_{j=1}^{n} H_{ij}$ is the meta-analysis (global) null hypothesis.

We have $m \times n$ hypotheses for inference:

$$\left.\begin{array}{ccc|c} H_{11} & \ldots & H_{1n} & H_{1G} \\ \vdots & \ddots & \vdots & \vdots \\ H_{m1} & \ldots & H_{mn} & H_{mG} \end{array}\right.$$

Inference following aggregate level testing

- In meta-analysis, aggregate level hypotheses testing is performed for powerful identification of features with signal[1].
- A natural follow-up question is which studies contain signal within a discovered feature.
- Testing $H_{i1}, \ldots, H_{in}$ following rejection of H_{iG} without accounting for the fact that H_{iG} was rejected using an aggregate-level test statistic, will produce biased inference [2] .

[1] Bhattacharjee et al., 2012. A subset-based approach improves power and interpretation for the combined analysis of genetic association studies of heterogeneous traits.

[2] Bogomolov and Benjamini, 2014. Selective inference on multiple families of hypotheses.

Goal for inference

- Our goal is to develop multiple testing procedures that guarantee control of FWER/FDR conditional on the row being selected.
  - This type of false positive control is particularly important if a researcher conducts different follow-up studies for each selected row.
- A related goal: Controlling the average FWER/FDR over the selected[1].

[1] Bogomolov and Benjamini, 2014. Selective inference on multiple families of hypotheses.

A large scale genomic application

- Expression quantitative trait loci (eQTLs) are genomic regions with genetic variants that influence the expression level of genes.
- Gene regulation is tissue specific, but within a single tissue may lack power due to small sample size.
- The discovery power of eQTL SNPs predictive of gene expression across multiple tissues may be increased by aggregate testing across tissue types.
- For the n=17 tumor tissues in The Cancer Genome Atlas (TCGA) Project, we aggregate the 17 eQTL test statistics to select eQTL SNPs influencing gene expression in at least one tissue, out of $m = 7,732,750$ candidate cis-eQTL SNPs.
- We aim to discover the non-null tissues within selected eQTL SNPs.

Notation

- $\mathcal{S} \subseteq \{1, \ldots, m\}$ is the set of selected rows, e.g., all hypotheses rejected by Bonferroni/BH on the global null p-values.
- V_i = number of false discoveries for row i.
- R_i = number of discoveries for row i.
- The conditional FWER for row i is

$$E(I[V_i > 0] | i \in \mathcal{S}).$$

- The conditional FDR for row i is

$$E(V_i / \max\{R_i, 1\} | i \in \mathcal{S}).$$

Notation

For feature (row) i:

- P_{ij}, $j = 1, \ldots, n$ are the p-values.
- P_{iG} is the global null p-value. Examples[1]:

$$p_{iG} = Pr(\chi^2_{2n} \geq -2\sum_{j=1}^{n} \log p_{ij}).$$

$$p_{iG} = 2Pr\left[\chi^2_{2n} \geq \max\left\{-2\sum_{j=1}^{n} \log p^L_{ij}, -2\sum_{j=1}^{n} \log(1 - p^L_{ij})\right\}\right].$$

- Our data matrix for analysis is:

$$\left(\begin{array}{ccc|c} p_{11} & \cdots & p_{1n} & p_{1G} \\ \vdots & \ddots & \vdots & \vdots \\ p_{m1} & \cdots & p_{mn} & p_{mG} \end{array}\right)$$

[1] Owen, 2009. Karl Pearson's meta-analysis revisited.

Our approach for inference following row-selection

1. Compute the conditional p-values, conditional on being selected.
2. Apply a valid FWER/FDR controlling procedure on the conditional p-values.

Our approach for inference following row-selection

1. Compute the conditional p-values, conditional on being selected.
2. Apply a valid FWER/FDR controlling procedure on the conditional p-values.

Questions we address:

1. The row may contain both null and non-null p-values, so the probability of selection is not known even for the simplest rule $\{P_{iG} \leq \alpha/m\}$. How can the conditional p-values be computed?
2. Even though the original p-values in a row are independent, the conditional p-values will be dependent.
   What is a valid FDR controlling procedure?

The conditional p-value computation for a selected row

We compute the p-values conditional on the event that the row was selected, holding all other p-values fixed.

For example, for the first column:

$$p'_{i1} = p_{i1}/b_{i1}, \quad b_{i1} = \max\{p : p_{iG}(p, p_{i2}, \ldots, p_{in}) \leq \alpha/m\}.$$

This is a valid p-value, since:

- P_{i1} is independent of $P_{i2}, \ldots, P_{in}$.
- if H_{i1} is null, then

$$P_{i1} \mid P_{iG} \leq \alpha/m, P_{i2} = p_{i2}, \ldots, P_{in} = p_{in} \sim U(0, b_{i1}).$$

Properties of the conditional p-values

- If $P_{iG}(1, p_{i2}, \ldots, p_{in}) \leq \alpha/m$, there is no inflation, i.e., $p'_{i1} - p_{i1} = 0$.
- With Holm/BH on $p'_{i1}, \ldots, p'_{in}$, the conditional FWER/FDR is controlled.

Theorem

If $p_{iG} \leq t_i$, then the BH procedure at level α on $p'_{i1}, \ldots, p'_{in}$ controls the conditional FDR at level $\leq \frac{n_0(i)}{n}\alpha$.

The conditional p-values based on Fisher's global null

- For row i, the Fisher global null p-value is

$$p_{iG} = Pr\left(\chi^2_{2n} \geq -2\sum_{j=1}^{n} \log p_{ij}\right).$$

- The conditional p-value for column j, given $p_{iG} \leq \alpha/m$, is

$$p'_{ij} = \begin{cases} p_{ij} & \text{if } \Pi^n_{l=1,l\neq j} p_{il} \leq e^{-\frac{1}{2}\chi^2_{1-\alpha/m,2n}}, \\ \dfrac{\Pi^n_{l=1} p_{il}}{e^{-\frac{1}{2}\chi^2_{1-\alpha/m,2n}}} & \text{otherwise.} \end{cases} \qquad j = 1,\ldots,n$$

- If $p_{i1} \leq \ldots \leq p_{in}$, then $p'_{i1} \leq \ldots \leq p'_{in}$.

Results for the cross-tissue eQTL analysis in TCGA

Table : The original two-sided p-values, conditional two-sided p-values, and BH-adjusted conditional two-sided p-values for each tissue, for three eQTL SNPs that differ in the number of post-selection discoveries.

	rs10896016-CTSW p-values			rs1437891-ASNSD1 p-values			rs13066873-LARS2 p-values		
	p_{ij}	p'_{ij}	$BH^{adj}\, p'_{ij}$	p_{ij}	p'_{ij}	$BH^{adj}\, p'_{ij}$	p_{ij}	p'_{ij}	$BH^{adj}\, p'_{ij}$
BLCA	0.01259	0.29510	0.38590	0.45523	0.45523	0.64491	0.00199	0.00199	**0.00484**
BRCA	0.73273	0.73273	0.83043	0.00030	0.00804	**0.02278**	0.00026	0.00026	**0.00147**
COAD	0.26604	0.29510	0.38590	0.00231	0.00231	**0.02278**	0.00099	0.00099	**0.00362**
GBM	0.36091	0.29510	0.38590	0.90232	0.90232	0.90232	0.00716	0.00716	**0.01353**
HNSC	0.92247	0.92247	0.98012	0.54711	0.54711	0.66435	0.54393	0.54393	0.54393
KIRC	0.00743	0.29510	0.38590	0.00000	0.00804	**0.02278**	0.01362	0.01362	**0.01781**
KIRP	0.99577	0.99577	0.99577	0.51974	0.51974	0.66435	0.00834	0.00834	**0.01418**
LAML	0.02349	0.29510	0.38590	0.77827	0.77827	0.82691	0.00345	0.00345	**0.00733**
LGG	0.13963	0.29510	0.38590	0.00005	0.00804	**0.02278**	0.00107	0.00107	**0.00362**
LIHC	0.01575	0.29510	0.38590	0.34415	0.34415	0.64491	0.01007	0.01007	**0.01426**
LUAD	0.00004	0.29510	0.38590	0.00078	0.00804	**0.02278**	0.00000	0.00000	**0.00000**
LUSC	0.12911	0.29510	0.38590	0.30344	0.30344	0.64481	0.04074	0.04074	**0.04827**
OV	0.06658	0.29510	0.38590	0.16256	0.16256	0.39479	0.00961	0.00961	**0.01426**
PAAD	0.25674	0.25674	0.38590	0.64167	0.64167	0.72723	0.04259	0.04259	**0.04827**
PRAD	0.14091	0.29510	0.38590	0.00495	0.00804	**0.02278**	0.06407	0.06407	0.06807
SKCM	0.01577	0.29510	0.38590	0.41503	0.41503	0.64491	0.00018	0.00018	**0.00147**
UCEC	0.59226	0.59226	0.71917	0.42909	0.42909	0.64491	0.00167	0.00167	**0.00473**
p_{iG}	3×10^{-9}			2×10^{-10}			$< 10^{-20}$		

An existing alternative approach[1]

The BB selection adjusted procedure: apply an FWER/FDR controlling procedure within selected rows at level $\frac{|\mathcal{S}|}{m}\alpha$.

Theorem (based on Theorem 3 in Benjamini and Bogomolov, 2014)

If for each column, the set of p-values is PRDS on the subset of p-values corresponding to true null hypotheses, the selection is by fixed thresholding/BH on the global null p-values, and the procedure used for testing each selected row is level α (a) Bonferrnoi or (b) BH, then the select-adjusted procedure guarantees in case (a)

$$E\left(\frac{\sum_{i\in\mathcal{S}} I[V_i > 0]}{\max\{|\mathcal{S}|, 1\}}\right) \leq \alpha,$$

and in case (b)

$$E\left(\frac{\sum_{i\in\mathcal{S}} V_i/\max\{R_i, 1\}}{\max\{|\mathcal{S}|, 1\}}\right) \leq \alpha.$$

.

[1] Bogomolov and Benjamini, 2014. Selective inference on multiple families of hypotheses.

Results for the cross-tissue eQTL analysis in TCGA

The BB selection adjusted procedure applies the BH procedure on the original p-values at level $\frac{19{,}690}{7{,}732{,}750}0.05 = 0.00013$. With BB: no discoveries are made for the first two eQTL SNPs; a single discovery is made for the third eQTL SNP.

	rs10896016-CTSW p-values			rs1437891-ASNSD1 p-values			rs13066873-LARS2 p-values		
	p_{ij}	p'_{ij}	$BH^{adj}\,p'_{ij}$	p_{ij}	p'_{ij}	$BH^{adj}\,p'_{ij}$	p_{ij}	p'_{ij}	$BH^{adj}\,p'_{ij}$
BLCA	0.01259	0.29510	0.38590	0.45523	0.45523	0.64491	0.00199	0.00199	**0.00484**
BRCA	0.73273	0.73273	0.83043	0.00030	0.00804	**0.02278**	0.00026	0.00026	**0.00147**
COAD	0.26604	0.29510	0.38590	0.00231	0.00231	**0.02278**	0.00099	0.00099	**0.00362**
GBM	0.36091	0.29510	0.38590	0.90232	0.90232	0.90232	0.00716	0.00716	**0.01353**
HNSC	0.92247	0.92247	0.98012	0.54711	0.54711	0.66435	0.54393	0.54393	0.54393
KIRC	0.00743	0.29510	0.38590	0.00000	0.00804	**0.02278**	0.01362	0.01362	**0.01781**
KIRP	0.99577	0.99577	0.99577	0.51974	0.51974	0.66435	0.00834	0.00834	**0.01418**
LAML	0.02349	0.29510	0.38590	0.77827	0.77827	0.82691	0.00345	0.00345	**0.00733**
LGG	0.13963	0.29510	0.38590	0.00005	0.00804	**0.02278**	0.00107	0.00107	**0.00362**
LIHC	0.01575	0.29510	0.38590	0.34415	0.34415	0.64491	0.01007	0.01007	**0.01426**
LUAD	0.00004	0.29510	0.38590	0.00078	0.00804	**0.02278**	0.00000	0.00000	**0.00000**
LUSC	0.12911	0.29510	0.38590	0.30344	0.30344	0.64481	0.04074	0.04074	**0.04827**
OV	0.06658	0.29510	0.38590	0.16256	0.16256	0.39479	0.00961	0.00961	**0.01426**
PAAD	0.25674	0.25674	0.38590	0.64167	0.64167	0.72723	0.04259	0.04259	**0.04827**
PRAD	0.14091	0.29510	0.38590	0.00495	0.00804	**0.02278**	0.06407	0.06407	0.06807
SKCM	0.01577	0.29510	0.38590	0.41503	0.41503	0.64491	0.00018	0.00018	**0.00147**
UCEC	0.59226	0.59226	0.71917	0.42909	0.42909	0.64491	0.00167	0.00167	**0.00473**
p_{iG}	3×10^{-9}			2×10^{-10}			$< 10^{-20}$		

Simulations with block dependence

We consider 100 blocks of 11 rows, where the signal within a non-null blocks is $N_{11}(\vec{\mu}, \Sigma)$ and the signal within a null blocks is $N_{11}(\vec{0}, \Sigma)$, where

$$\vec{\mu} = \begin{pmatrix} \rho^5\mu \\ \vdots \\ \rho\mu \\ \mu \\ \rho\mu \\ \vdots \\ \rho^5\mu \end{pmatrix}, \quad \Sigma = \begin{pmatrix} 1 & \rho & \rho^2 & \dots & \rho^{B-1} \\ \rho & 1 & \rho & \dots & \rho^{B-2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho^{B-1} & \rho^{B-2} & \rho^{B-3} & \dots & 1 \end{pmatrix},$$

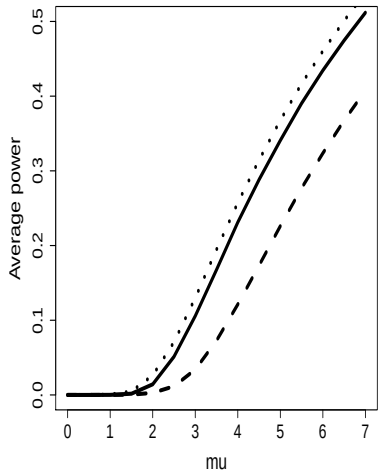
In n_1 studies there was one non-null block, and the remaining $n - n_1$ studies where all null:

$$\begin{pmatrix} N_{11}(\vec{\mu}, \Sigma) & \dots & N_{11}(\vec{\mu}, \Sigma) & N_{11}(\vec{0}, \Sigma) & \dots & N_{11}(\vec{0}, \Sigma) \\ N_{11}(\vec{0}, \Sigma) & \dots & N_{11}(\vec{0}, \Sigma) & N_{11}(\vec{0}, \Sigma) & \dots & N_{11}(\vec{0}, \Sigma) \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \end{pmatrix},$$

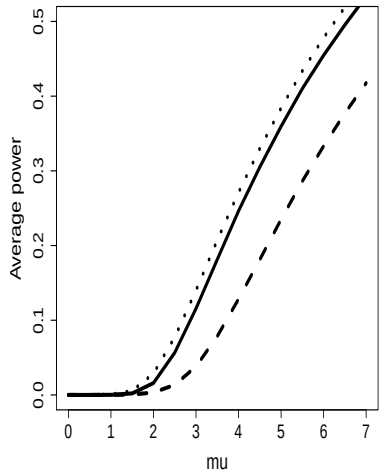
Results on power: conditional approach (solid), BB (dashed), naive (dotted)

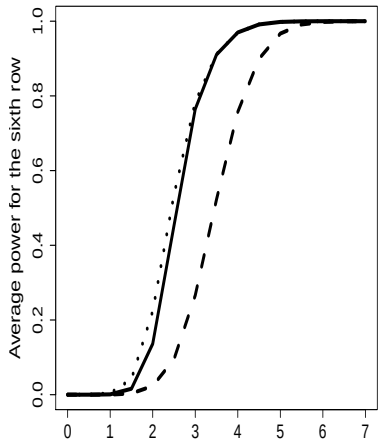
$(n, n_1) = (21, 7)$, Row Selection by:

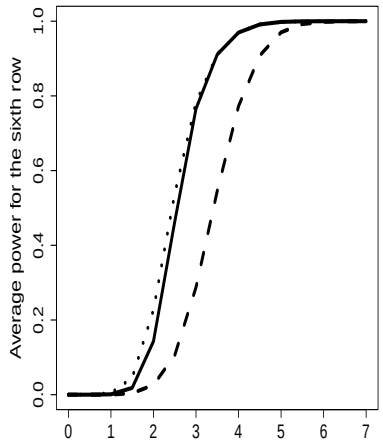
Bonferroni



BH

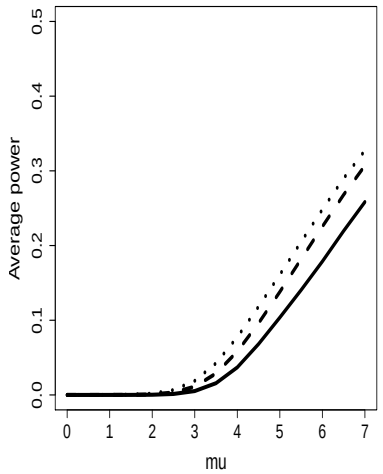




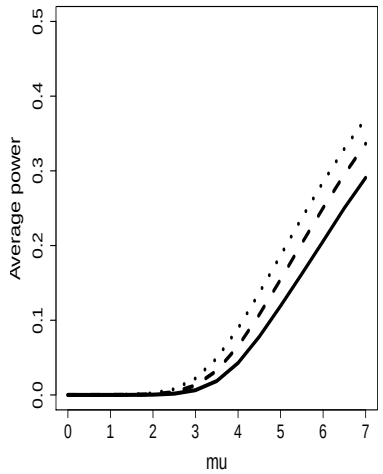


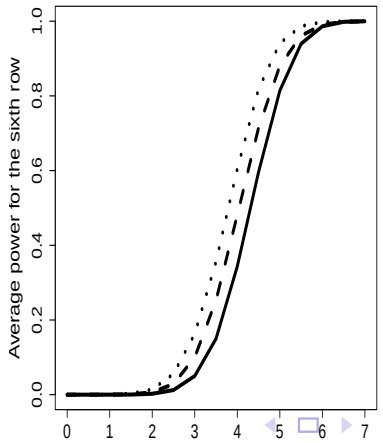
$(n, n_1) = (10, 2)$, Row Selection by:

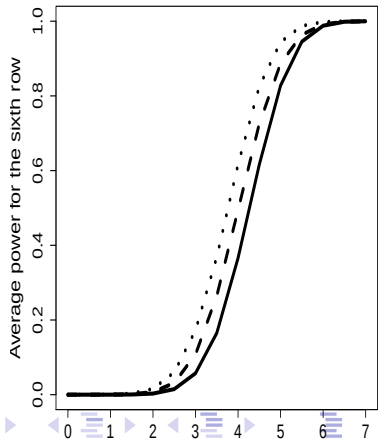
Bonferroni



BH



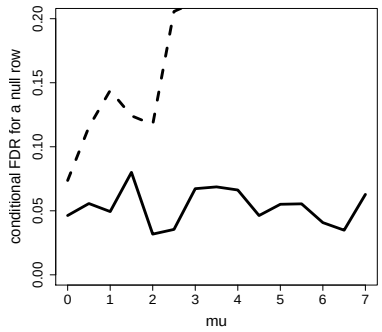




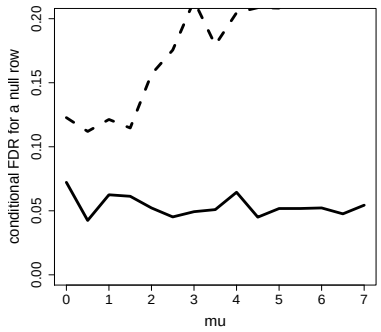
Results on error control: conditional approach (solid), BB (dashed), naive (dotted)

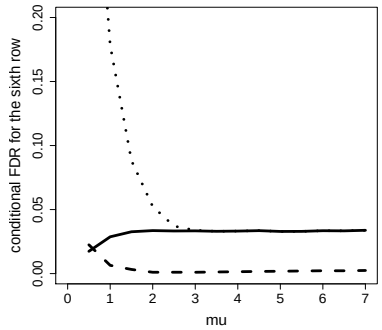
$(n, n_1) = (21, 7)$, Row Selection by:

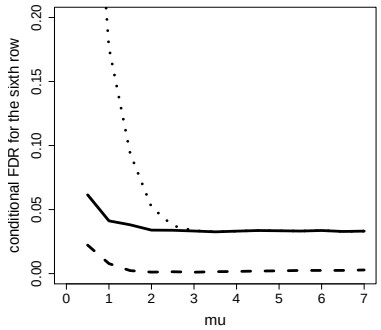
Bonferroni

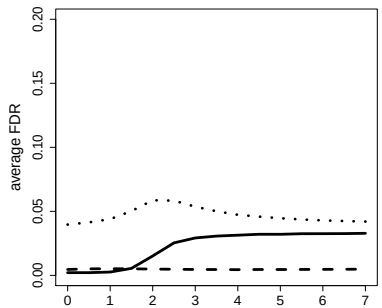


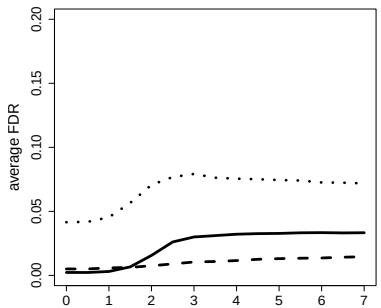
BH





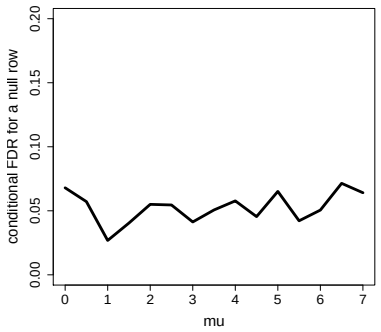




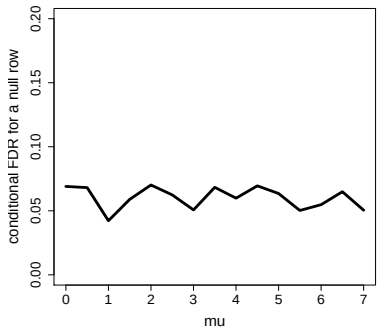


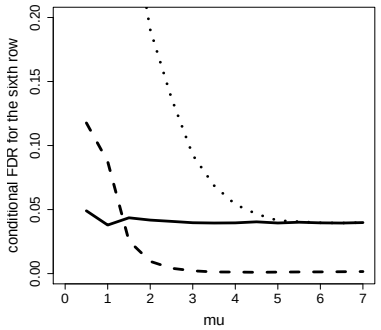
$(n, n_1) = (10, 2)$, Row Selection by:

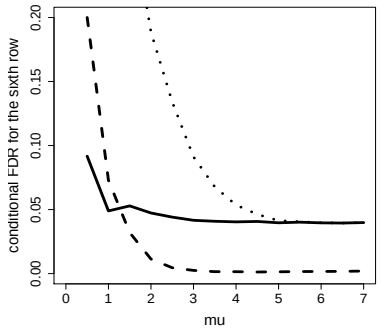
Bonferroni

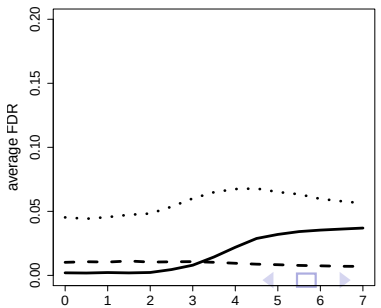


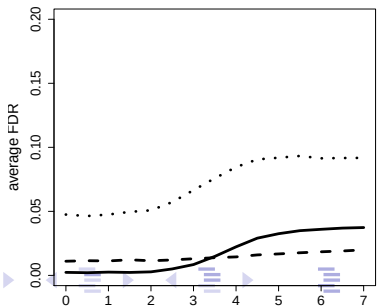
BH











Summary

- Following row-selection, we presented a valid and powerful selection adjusted method for identification of columns/studies that drive the signal in the row.
- A comparison with the method of Benjamini and Bogomolov, 2014, suggests that although it is less general, when the columns are independent the power gain can be very large.

Summary

- Two-way structured hypotheses provide an exciting opportunity for novel procedures with more than one error guarantee.

	row level	within a selected row	over all the selected	column level	within a selected column
Benjamini and Bogomolov[1]	✓		✓		
Heller et al. [2]	✓	✓	✓		
Foygel Barber and Ramdas[3]	✓		✓	✓	
Liu et al. [4]		✓	✓		

[1] Bogomolov and Benjamini, 2014. Selective inference on multiple families of hypotheses.

[2] Heller, Chatterjee, Krieger, and Shi, 2016. Post-selection inference following aggregate level hypotheses testing in large scale genomic data.

[3] Foygel Barber and Ramdas, 2016. The p-filter: multi-layer FDR control for grouped hypotheses.

[4] Liu, Sarkar, and Zhao, 2016. A new approach to multiple testing of grouped hypotheses.