Towards More Significant Discoveries in High Dimensional Data Analysis

Multiple Testing with Heterogeneous Multinomial Data

Joshua Habiger[1] David Watts[2] Michael Anderson[3]
Oklahoma State University

[1]Assistant Professor, Department of Statistics
[2]PhD student, Department of Statistics
[3]Associate Professor, Department of Plant and Soil Science

January, 2016

The Main Idea

- Data: large number of attributes p, small sample size n
  - fMRI analysis, GWAS, "omics", ...
- Objective: Discover reproducible **and** interesting attributes
- Standard method:
  1. Test statistic (*p*-values / post. probs.) computed for each attribute
  2. Apply multiple testing procedure ⇒ identify *"significant"* attributes
- Problem:
  - Many *significant* attributes not interesting
  - Many interesting attributes not *significant*

Overview

1. **Motivation**
   - Rhizosphere
   - Motivating Study
   - Data Analysis
   - Problem
2. Clfdr Procedure
   - Oracle Procedure
   - Adaptive Procedure
3. Assessment
   - Application
   - Thresholding Effect
4. Remarks

Rhizosphere and Rhizobacteria

What is the rhizosphere?

- Soil near the roots of a plants (plant stomache)
- Millions of unknown species of bacteria: **rhizobacteria**

Why do we care?

- Rhizosphere composition associated with plant health / productivity
- Understand rhizosphere ⇒ manipulate rhizosphere ⇒ increase productivity

Typical Wheat Rhizosphere Studies

- Standard objective: **Who's there**?
  - Method: Rhizosphere sample(s) + RNA sequencing technology ⇒ identify abundant species of rhizobacteria
  - Called *core microbiome*
- Assumption: **"Abundance = association hypothesis"**
  - Most abundant rhizobacteria are associated with productivity

Question

Core microbiome vs. core productivity-associated microbiome

Illustration

Abundance = association hypothesis?

Study

- Objective of Anderson and Habiger (2012): Identify productivity associated microbiome
- Data collection:
  1. 5 wheat rhizosphere soil samples: Average shoot biomass (g) among wheat plants in each sample measures productivity

x_1	x_2	x_3	x_4	x_5
0.86	1.34	1.81	2.37	3.00

  2. 16s rRNA software: # DNA copies of $m = 1, 2, ..., 778$ species in each sample (abundance)

Species m	y_{1m}	y_{2m}	y_{3m}	y_{4m}	y_{5m}	Total (n_m)
1	0	1	1	0	5	7
2	9	2	0	0	3	14
⋮	⋮	⋮	⋮	⋮	⋮	⋮
778	16	10	29	18	13	81

- Refined objective: Which bacteria are associated with productivity?

Statistical Analysis

Step 1: Compute test statistics / p-values

- Models: $Y_{nm} \sim Pois(\mu_{nm})$, $log(\mu_{nm}) = \alpha_m + \beta_m x_n$
- Null hypotheses: $H_m : \beta_m = 0$
- Z-scores: $Z_m = \frac{\hat{\beta}_m}{S.E.(\hat{\beta}_m)}$
- p-Values: $P_m = \Pr(|Z_m| \geq |z_m|)$

Step 2: Define rejection threshold

- Question: Reject H_m if $P_m \leq 0.05$ or $|Z_m| \geq 1.96$?

Error Rates

Common Error Rates

Error Rate	Properties	Uses
$FDR = E\left[\frac{V}{\max\{R,1\}}\right]$	liberal	large # tests
$FWER = \Pr(V > 0)$	conservative	small # tests

- V = # false discoveries (false rejections)
- R = # discoveries (rejections)

Remark: Many other error rates

Classical FDR Procedures

- Benjamini and Hochberg (1995) procedure:
  - Implementation:
    1. Order $P_{(1)} \leq P_{(2)} \leq \ldots \leq P_{(M)}$
    2. Reject $k = \max\{m : P_{(m)} \leq \alpha m/M\}$ null hypotheses
  - Properties: $FDR \leq \pi_0 \alpha \leq \alpha$ under positive dependence
- Adaptive BH procedures: Storey et. al. (2004), Liang and Nettleton (2012), ...
  - Implementation:
    1. Estimate π_0
    2. Apply BH at $\alpha/\hat{\pi}_0$
  - Properties: $FDR \leq \alpha$[1] under weak dependence

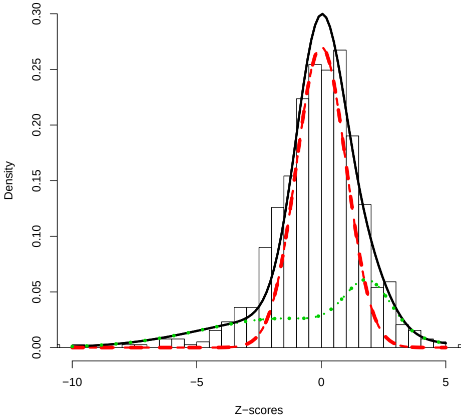
[1] $FDR = \alpha$ for any π_0 under least favorable p-Value configuration - Habiger (2014)

Local FDR / Bayesian Procedures

- Local FDR (lFDR) - Efron (2010)
  - Mixture model: $Z_m \sim f(z) = \pi_0 f_0(z) + (1 - \pi_0) f_1(z)$
  - Local FDR: $lFDR(z) = \frac{\pi_0 f(z)}{f(z)} = \Pr(H_m \text{ true } | Z_m = z)$
  - Local FDR statistics: $lFDR_m = lFDR(Z_m)$
  - Adaptive: $\hat{\pi}_0, \hat{f}_1 \rightarrow \widehat{lFDR}_m$
- Adaptive lFDR procedure - Sun and Cai (2007)
  1. Order $\widehat{lFDR}_{(1)} \leq \widehat{lFDR}_{(2)} \leq ... \leq \widehat{lFDR}_{(M)}$
  2. Reject $k = \max\left\{m : \sum_{i=1}^{m} \widehat{lFDR}_{(i)} \leq \alpha m\right\}$ null hypotheses
- Properties:
  - $FDR \rightarrow \alpha$
  - Asymptotically "optimal"

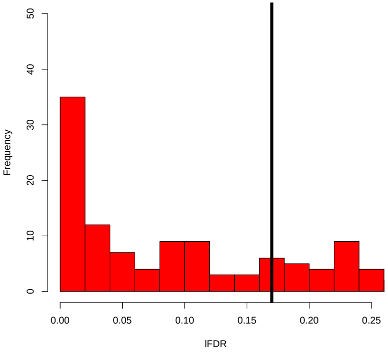
Estimated Mixture Model

$$f(z) = 0.67\phi(z; 0, 1) + 0.33 f_1(z)$$



lFDR Procedure: $\alpha = 0.05$

$$lFDR_m = lFDR(z_m) = \frac{0.67\phi(z_m)}{f(z_m)}$$



85 discoveries ⇒ **productivity-associated microbiome**?

"Significant"

Question: Which species is discovered?

m	Y_{1m}/n_m	Y_{2m}/n_m	Y_{3m}/n_m	Y_{4m}/n_m	Y_{5m}/n_m	$\hat{\beta}_m$	n_m	$\widehat{lFDR}_m$	Discover
1	0.36	0.50	0.00	0.07	0.07	?	?	?	?
2	0.15	0.13	0.28	0.25	0.19	?	?	?	?
Null	0.20	0.20	0.20	0.20	0.20	0	—	1	x

"Significant"

Question: Which species is discovered?

m	Y_{1m}/n_m	Y_{2m}/n_m	Y_{3m}/n_m	Y_{4m}/n_m	Y_{5m}/n_m	$\hat{\beta}_m$	n_m	$\widehat{lFDR}_m$	Discover
1	0.36	0.50	0.00	0.07	0.07	**-1.09**	?	?	?
2	0.15	0.13	0.28	0.25	0.19	**0.19**	?	?	?
Null	0.20	0.20	0.20	0.20	0.20	0	—	1	x

"Significant"

Question: Which species is discovered?

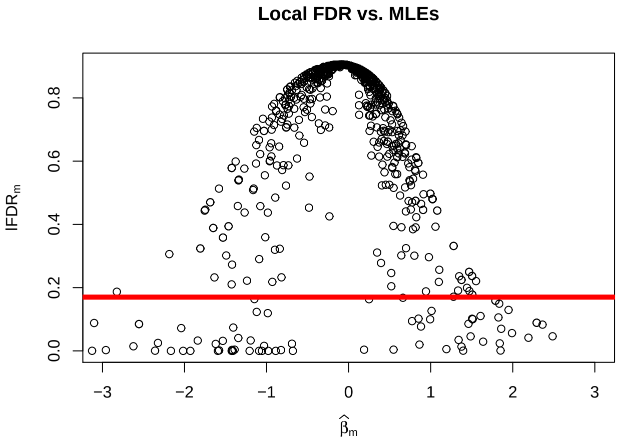
m	Y_{1m}/n_m	Y_{2m}/n_m	Y_{3m}/n_m	Y_{4m}/n_m	Y_{5m}/n_m	$\hat{\beta}_m$	n_m	$\widehat{lFDR}_m$	Discover
1	0.36	0.50	0.00	0.07	0.07	-1.09	**11**	?	?
2	0.15	0.13	0.28	0.25	0.19	0.19	**911**	?	?
Null	0.20	0.20	0.20	0.20	0.20	0	—	1	x

"Significant"

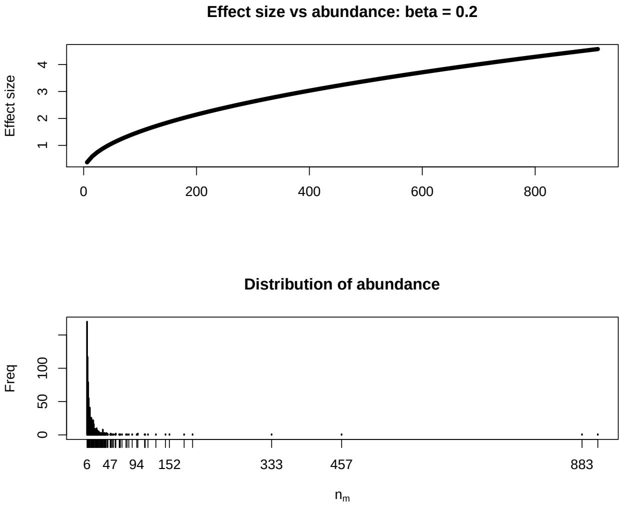
Question: Which species is discovered?

m	Y_{1m}/n_m	Y_{2m}/n_m	Y_{3m}/n_m	Y_{4m}/n_m	Y_{5m}/n_m	$\hat{\beta}_m$	n_m	$\widehat{lFDR}_m$	Discover
1	0.36	0.50	0.00	0.07	0.07	-1.09	11	0.29	**x**
2	0.15	0.13	0.28	0.25	0.19	0.19	911	0.003	✓
Null	0.20	0.20	0.20	0.20	0.20	0	—	1	x

Illustration



Negligible Associations Detected if Abundant Enough



Consequence

	Abundant	Rare
Strong association	✓	o
Weak association	✓	o

✓ = discovered as "associated with productivity"

- **Abundance = association hypothesis** RETAINED!

Illustration

"Statistics show that Fred is associated with productivity"

1. Motivation
   - Rhizosphere
   - Motivating study
   - Data analysis
   - Problem
2. **Clfdr procedure**
   - Oracle procedure
   - Adaptive procedure
3. Assessment
   - Application
   - Thresholding effect
4. Remarks

Multinomial Mixture Model

- Under log-linear model $\boldsymbol{Y}_m | N_m = n_m \sim Multinomial(n_m, \boldsymbol{p}(\beta_m))$
  - $p_n(\beta_m) = \frac{\exp\{\beta_m x_n\}}{\sum_{n=1}^{N} \exp\{\beta_m x_n\}}$
  - pmf notation: $p(\boldsymbol{y}_m | n_m; \beta_m)$
- Prior $\Pr(\beta_m = \gamma_k) = \pi_k$ for $k = 0, 1, ..., K$
  - Null prior: Take $\gamma_0 = 0 \Rightarrow \Pr(\beta_m = 0) = \Pr(H_m \text{ true }) = \pi_0$
- Mixture of Multinomial pmfs:

$$p(\boldsymbol{y}_m | n_m; \boldsymbol{\gamma}, \boldsymbol{\pi}) = \pi_0 p(\boldsymbol{y}_m | n_m; 0) + \pi_1 p(\boldsymbol{y}_m | n_m; \gamma_1) + ... + \pi_K p(\boldsymbol{y}_m | n_m; \gamma_K)$$

Oracle clFDR Procedure

1. Compute clFDRs :

$$clFDR_m \equiv \frac{\pi_0 p(\boldsymbol{y}_m | n_m; \gamma_0)}{p(\boldsymbol{y}_m | n_m; \boldsymbol{\gamma}, \boldsymbol{\pi})} = \Pr(\beta_m = 0 | \boldsymbol{y}_m, n_m; \boldsymbol{\gamma}, \boldsymbol{\pi})$$

2. Rank clFDRs: $clFDR_{(1)} \leq clFDR_{(2)} \leq \ldots \leq clFDR_{(M)}$

3. Reject k nulls with smallest clFDR:

$$k = \max\left\{ m : \sum_{i=1}^{m} clFDR_{(i)} \leq \alpha m \right\}$$

FDR control

Theorem

If each $(\boldsymbol{Y}_m, \beta_m)$ is generated according to the Multinomial mixture model, then the clFDR procedure has FDR $\leq \alpha$ regardless of $(n_1, n_2, ..., n_M)$.

Problem: $\boldsymbol{\pi}, \boldsymbol{\gamma}$ unknown.

Idea

- Adaptive procedure plugs in **consistent** estimates for $\boldsymbol{\pi}$ and $\boldsymbol{\gamma}$
- Maximum likelihood estimation:
  - Under conditional independence get log likelihood

$$l(\boldsymbol{\gamma}, \boldsymbol{\pi}) = \sum_{m=1}^{M} \log \left(\sum_{k=0}^{K} \pi_k p(\boldsymbol{y}_m | n_m; \gamma_k) \right).$$

  - Use EM algorithm to get MLE

EM Algorithm

- E step: $\hat{z}_{km} = \frac{\pi_k^{old} p(\boldsymbol{y}_m | n_m; \gamma_k^{old})}{\sum_{k=0}^{K} \pi_k^{old} p(\boldsymbol{y}_m | n_m; \gamma_k^{old})}$.
- M step: Maximize $Q(\gamma, \boldsymbol{\pi})$ s.t. $\sum_k \pi_k = 1$

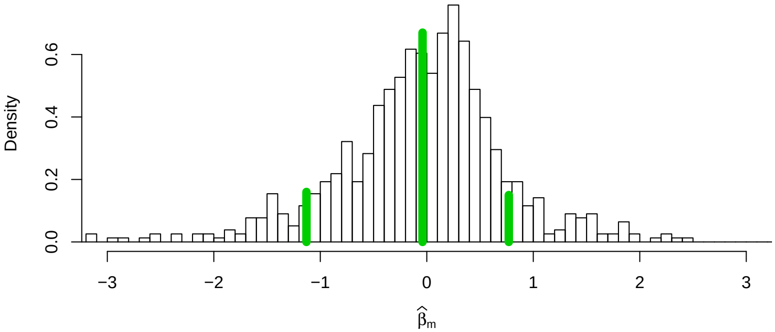
$$Q(\gamma, \boldsymbol{\pi}) \equiv \sum_{m=1}^{M} \sum_{k=0}^{K} \hat{z}_{km} \log(\pi_k p(\boldsymbol{y}_m | n_m; \gamma_k))$$

$$= \sum_{m=1}^{M} \sum_{k=0}^{K} \hat{z}_{km} \log(\pi_k) + \sum_{m=1}^{M} \sum_{k=0}^{K} \hat{z}_{km} \log p(\boldsymbol{y}_m | n_m; \gamma_k)$$

  - 1st quantity + contraint $\Rightarrow \hat{\pi}_k^{new} = \frac{1}{M} \sum_m \hat{z}_{km}$
  - 2nd quantity + tweeked optim() $\Rightarrow \hat{\gamma}_k^{new}$

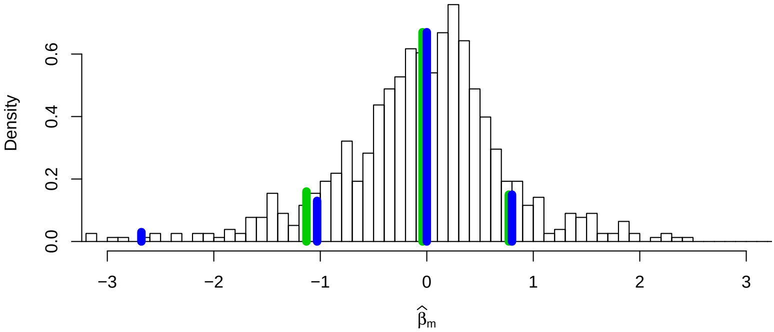
1. Motivation
    - Rhizosphere
    - Motivating study
    - Data Analysis
    - Problem
2. Clfdr Procedure
    - Oracle Procedure
    - Adaptive Procedure
3. **Assessment**
    - Application
    - Thresholding Effect
4. Remarks

Model 1 and Results



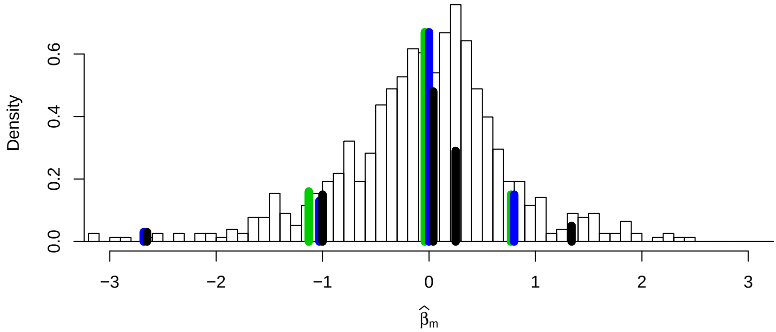
K	$\hat{\pi}_0$	$(\hat{\pi}_1, \hat{\gamma}_1)$	$(\hat{\pi}_2, \hat{\gamma}_2)$	$(\hat{\pi}_3, \hat{\gamma}_3)$	$(\hat{\pi}_4, \hat{\gamma}_3)$	AIC	BIC	Disc.
2	0.69	(0.16, -1.13)	(0.15,0.78)	NA	NA	1222	1224	99
3								
4								

Model 2 and Results



K	$\hat{\pi}_0$	$(\hat{\pi}_1, \hat{\gamma}_1)$	$(\hat{\pi}_2, \hat{\gamma}_2)$	$(\hat{\pi}_3, \hat{\gamma}_3)$	$(\hat{\pi}_4, \hat{\gamma}_3)$	AIC	BIC	Disc.
2	0.69	(0.16, -1.13)	(0.15,0.78)	NA	NA	1222	1224	99
3	0.69	(0.03, -2.68)	(0.13, -1.03)	(0.15, 0.79)	NA	1214	1217	97
4								

Model 3 and Results



K	$\hat{\pi}_0$	$(\hat{\pi}_1, \hat{\gamma}_1)$	$(\hat{\pi}_2, \hat{\gamma}_2)$	$(\hat{\pi}_3, \hat{\gamma}_3)$	$(\hat{\pi}_4, \hat{\gamma}_3)$	AIC	BIC	Disc.
2	0.69	(0.16, -1.13)	(0.15,0.78)	NA	NA	1222	1224	99
3	0.69	(0.03, -2.68)	(0.13, -1.03)	(0.15, 0.79)	NA	1214	1217	97
4	0.48	(0.03, -2.68)	(0.15, -1.03)	(0.29, 0.25)	(0.05,1.34)	1211	1215	114

"Significant"

Question: Now which species is discovered?

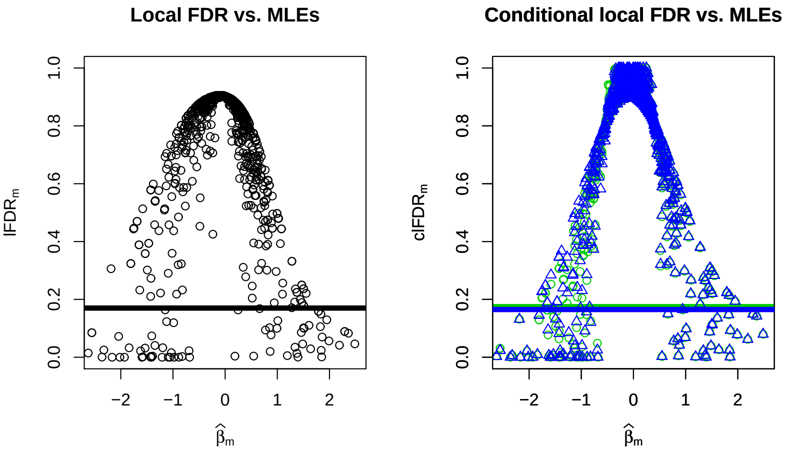
Local FDR procedure

m	Y_{1m}/n_m	Y_{2m}/n_m	Y_{3m}/n_m	Y_{4m}/n_m	Y_{5m}/n_m	$\hat{\beta}_m$	n_m	$\widehat{lFDR}_m$	Disc.
1	0.36	0.50	0.00	0.07	0.07	-1.09	11	0.29	**x**
2	0.15	0.13	0.28	0.25	0.19	0.19	911	0.003	✓
Null	0.20	0.20	0.20	0.20	0.20	0	—	1	x

Conditional local FDR procedure

m	Y_{1m}/n_m	Y_{2m}/n_m	Y_{3m}/n_m	Y_{4m}/n_m	Y_{5m}/n_m	$\hat{\beta}_m$	n_m	$\widehat{clFDR}_m$	Disc.
1	0.36	0.50	0.00	0.07	0.07	-1.09	11	0.10, 0.12	✓
2	0.15	0.13	0.28	0.25	0.19	0.19	911	1, 1	x

Illustration: lFDR vs clFDR



Setting for Theoretical Study

- Model: $\Pr(\beta_m = 0) = \pi_0$, $\Pr(\beta_m = \gamma_1) = (1 - \pi_0)$, $\gamma_1 > 0$
- Z-score: $Z_m = \frac{T_m - E[T_m|\beta_m=0, N_m=n_m]}{\sqrt{Var(T_m|\beta_m=0, N_m=n_m)}}$
- Conditional lFDR procedure
  - $f(z|N_m = n) = \pi_0\phi(z;0,1) + (1-\pi_0)\phi(z;\mu(\gamma_1, n), \sigma^2(\gamma_1))$
  - $clFDR(z, n) = \pi_0\phi(z;0,1)/f(z|N_m = n)$
  - $[clFDR(z, n) \leq \lambda] = [z \geq a(n)]$
- lFDR procedure
  - $f(z) = \pi_0\phi(z;0,1) + (1-\pi_0)\sum_{n\in\mathcal{N}} \phi(z;\mu(\gamma_1, n), \sigma^2(n))p(n)$
  - $lFDR(z) = \pi_0\phi(z;0,1)/f(z)$
  - $[lFDR(z) \leq \lambda] = [z \geq b]$

Thresholding Effect

Theorem

Under $f(z|N_m = n)$, the rejection threshold $a(n)$ is increasing in n whenever

$$\mu(n,\gamma_1)^2 > 2\log\left(\sigma(\gamma_1)\frac{\pi_0(1-\lambda)}{(1-\pi_0)\lambda}\right). \quad (1)$$

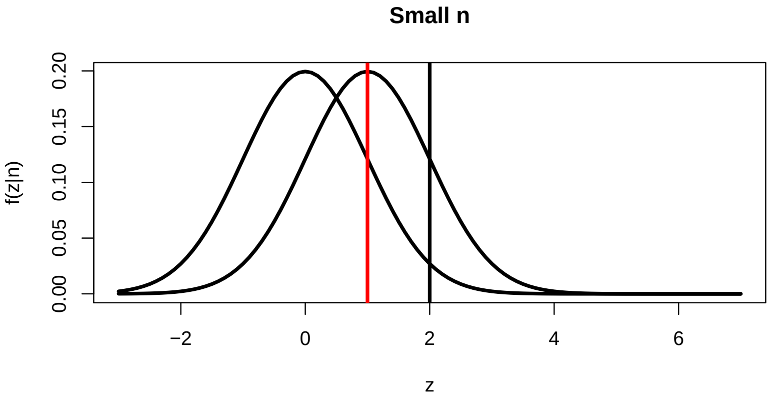
for any $\gamma_1 > 0, \lambda > 0$ and $\pi_0 \in (0,1)$.

Important points:

- Eq. (1) satisfied for all large enough n: $\mu(n,\gamma_1) \nearrow n$
- Safeguard against $\gamma_1 \approx 0$ and large n
- No such safeguard for lFDR procedure

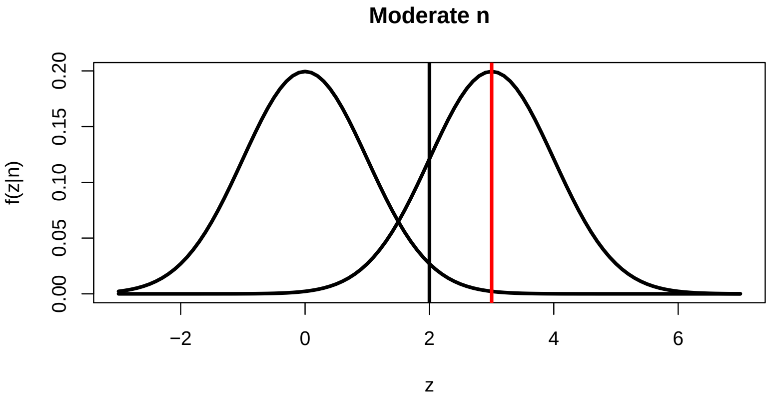
Thresholding Illustration

lFDR threshold vs. clFDR threshold: fixed γ_1



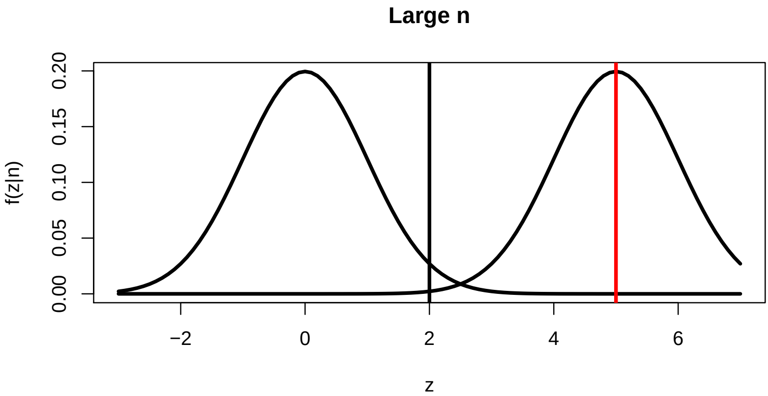
Thresholding Illustration

lFDR threshold vs. clFDR threshold: fixed γ_1



Thresholding Illustration

lFDR threshold vs. clFDR threshold: fixed γ_1



1. Motivation
   - Rhizosphere
   - Motivating Study
   - Data Analysis
   - Problem
2. Clfdr Procedure
   - Oracle Procedure
   - Adaptive Procedure
3. Assessment
   - Application
   - Thresholding effect
4. **Remarks**

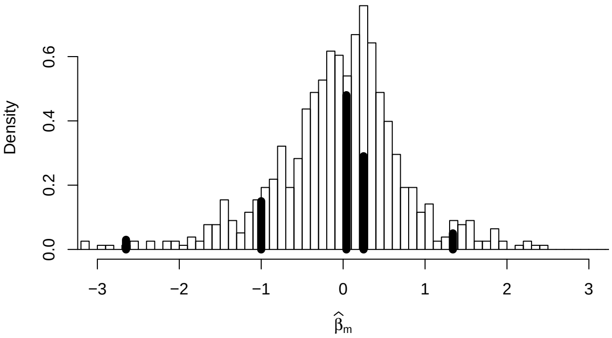
What We Did

- **Standard objective: Maximize # discoveries** s.t. FDR controlled
  - Method: Rank attributes according to $lFDR_m$
  - Problem: $lFDR_m \to 0$ as $n_m \to \infty$ if $\beta_m \neq 0$
    - **Statistical significance does not imply practical significance**

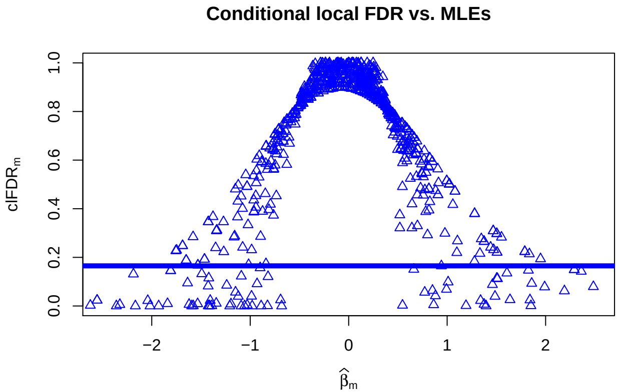
- **Better objective: Maximize # interesting discoveries** s.t. FDR controlled
  - Method: Given n_m, rank attributes according to $clFDR_m$
  - Solution: $clFDR_m \to 1$ as $n_m \to \infty$ if $\beta_m \in \mathcal{N}(0)$
    - **Statistical significance does imply practical significance**

Future work 1

Empirical vs. theoretical null: Efron (2004) and Bickel (2012)

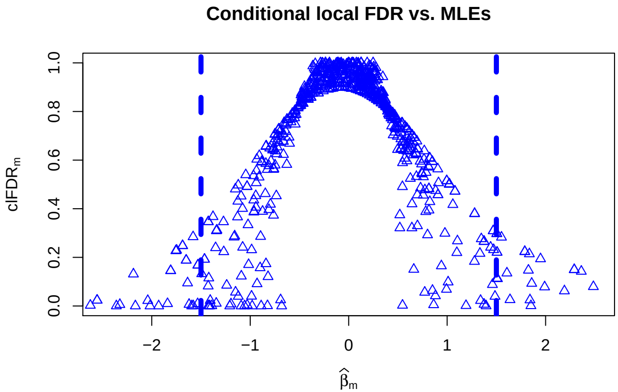


Future Work 2



Should we use this rejection region?

Future Work 2



Should we use this rejection region?

Future Work 3

A general procedure:

1. Rank attributes using any measure of practical significance
   - $\hat{\beta}$, SSR, AIC, R^2, IMCR . . .
2. Choose threshold
   - Compute $Q_m = \Pr(H_m \text{ true}|\text{data})$
   - Let $\mathcal{R} \subseteq \{1, 2, ..., M\}$ index any arbitrary set of discoveries, say the set of R most practically significant attributes. If $\sum_{m \in \mathcal{R}} Q_m \leq \alpha|\mathcal{R}|$ then $FDR \leq \alpha$

Development:

- Parameter estimation effect?
- Dependence?
- FDR?
- Measures of practical significance?

Some References

Anderson, M. and J. Habiger (2012).
Characterization and identification of productivity-associated rhizobacteria in wheat.
Applied and Environmental Microbiology 78(12), 4434 – 444.

Efron, B. (2010).
Large-Scale Inference.
Cambridge: Cambridge University Press.

Habiger, J., D. Watts, and M. Anderson (2015).
Multiple testing with heterogeneous multinomial distributions.
arXiv:1511.01400.

Habiger, J. (2014).
Weighted adaptive multiple decision functions for false discovery rate control.
arXiv:1412.0645.

Ruppert, D., D. Nettleton, and J. T. Hwang (2007).
Exploring the information in p-values for the analysis and planning of multiple-test experiments.
Biometrics 63(2), 483–495.

Sun, W. and T. T. Cai (2007).
Oracle and adaptive compound decision rules for false discovery rate control.
Journal of the American Statistical Association 102(479), 901–912.

Sun, W. and A. C. McLain (2012).
Multiple testing of composite null hypotheses in heteroscedastic models.
Journal of the American Statistical Association 107(498), 673–687.

Overtime: The classical approach

Weighted Adaptive BH Procedure

1. Compute weights $w_m = w(n_m)$
2. Get weighted p-value: $Q_m = P_m / w_m$
3. Estimate π_0:

$$\hat{\pi}_0 = \frac{\sum_m I(Q_m \geq \lambda) + 1}{1 - \lambda}$$

4. Apply BH procedure to Q_ms at level $\alpha / \hat{\pi}_0$

Finite Sample Results

Theorem

If P_ms ind under H_ms and independent of other P_ms

$$FDR \leq \alpha \bar{w}_0 \frac{1-\lambda}{1-\lambda \bar{w}_0}$$

for $\bar{w}_0$ mean weight among true H_ms.

Corollaries for *FDR* control

- $\bar{w}_0 \leq 1$
- $\boldsymbol{w} = \mathbf{1}$ - Storey et. al (2004)
- Take $\alpha^* = \alpha \frac{1}{w_{(M)}} \frac{1-\lambda w_{(M)}}{1-\lambda}$

Asymptotic Results

Under weak dependence, as $M \to \infty$...

Theorem

The weighted adaptive BH procedure almost surely dominates its unadaptive counterpart in that it uses a larger rejection threshold.

Theorem

The weighted adaptive procedure has FDP $\leq \alpha$ almost surely if $\lim_{M\to} \bar{w}_0 = \mu_0 \leq 1$. Equality is achieved under least favorable configuration if $\mu_0 = 1$ (α-exhaustive).

Corollaries for FDP control:

- optimal weights for random effects model
- weights positively correlated with optimal weights
- $w_m \overset{i.i.d.}{\sim} E[W_m] = 1$
- Storey et. al. (2004) is α exhaustive

Weights

- For any fixed γ_ks + technical details $\Rightarrow w_m = \frac{Mt_m}{\sum_m t_m}$ where

$$t_m = 2\bar{\Phi}\left(0.5\bar{\Phi}^{-1}(\alpha/4)\left[\frac{\sqrt{n_m}}{\sqrt{n_.}/M} + \frac{\sqrt{n_.}/M}{\sqrt{n_m}}\right]\right)$$

- Main point: w_m is decreasing in n_m for all large enough n_m