Multiple Testing with Heterogeneous Data

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

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

August, 2017

Main References

- Habiger, J., D. Watts, and M. Anderson (2017). Multiple testing with heterogeneous multinomial distributions. *Biometrics 73*(2), 562 – 570.
- Habiger, J. (2017). Adaptive False Discovery Rate Control for Heterogeneous Data. *Statistica Sinica (in press)*

Outline

- Can We Ignore Heterogeneity?
- Proposed Procedure
- Assessment
- Comments

Background

- Background:
  - Rhizosphere: Area of the soil near roots
  - Rhizosphere microbiome: Microorganisms / bacteria in the rhizosphere
  - Millions of bacteria per gram of soil
  - Standard rhizosphere microbiome study: **Who's there / abundant**?
  - If we know who's there we can intervene
- Research question (Anderson and Habiger; 2012):
  - Who's there vs. who's **relevant** (associated with plant health/productivity)?
  - Is the abundance = association hypothesis true?

Illustration of Research Question

Fred is *abundant*. Is he *"productive"*?

Study

- 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

- Remark: $6 \leq n_m \leq 911$

Classical Benjamini and Hochberg (1995)

Step 1: Compute Z-scores / p-values

- Model: $Y_{nm} \sim Pois(\mu_{nm})$, $log(\mu_{nm}) = \alpha_m + \beta_m x_n$
- 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 to control FDR

- Reject k null hypotheses for $k = \max\{i : P_{(i)} \leq \alpha \frac{i}{m}\}$

Remark: Much work on adaptive BH procedure: Storey et. al (2004), Nettleton and Liang (2012)

Bayes - Sun and Cai (2007), Efron (2010)

Step 1: Compute / estimate posterior null probability

- 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: $\hat{\pi}_0, \hat{f}_1 \rightarrow \widehat{lFDR}_m$

Step 2: Define a rejection threshold

- Reject k null hypotheses for $k = \max\left\{m : \sum_{i=1}^{m} \widehat{lFDR}_{(i)} \leq \alpha m\right\}$

"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	$\widehat{IFDR}_m$	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"

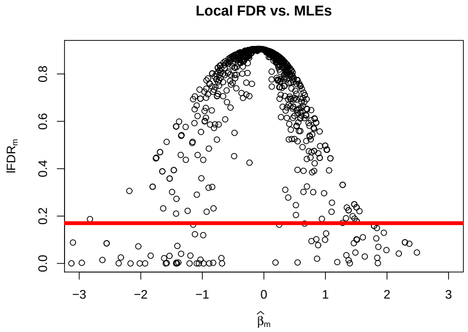
Question: Which species is discovered?

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	$\widehat{IFDR}_m$	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

Remarks:

- f_1 is a mixture of normals. Results same for 2,3,4 component densities
- BH procedure behaves similarly

Illustration



- What's happening:
  1. $Lfdr(z_m) \to 0$ as n_m/abundance $\to \infty$ if $\beta_m \neq 0$.
  2. Recall $6 \leq n_m \leq 911$
- Consequence: **Abundance = association hypothesis** RETAINED INCORRECTLY!
- See also Sun and McLain (2012)→ Berger and Selke (1987) → Berkson (1938).

Illustration

"Statistics show that Fred is productive"

Finite Multinomial Mixture Model

- Under log-linear model $\boldsymbol{Y}_m | N_m = n_m \sim \mathit{Multinomial}(n_m, \boldsymbol{p}(\beta_m))$
  - $p_n(\beta_m) = \frac{\exp\{\beta_m x_n\}}{\sum_{n=1}^{N} \exp\{\beta_m x_n\}}$
  - $H_m : \beta_m = 0 \Rightarrow p_1 = p_2 = \ldots = p_N = 1/N$
  - pmf notation: $p(\boldsymbol{y}_m | n_m; \beta_m)$
- Prior $\Pr(\beta_m = \gamma_k) = \pi_k$ for $k = 0, 1, \ldots, K$
  - Null prior: Take $\gamma_0 = 0 \Rightarrow \Pr(\beta_m = 0) = \Pr(H_m \text{ true }) = \pi_0$
- Mixture of Multinomial pmfs:

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

Oracle and Adaptive clFDR Procedure

Oracle 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. Reject k nulls with smallest clFDR:

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

Adaptive Procedure:

- Plug in ML estimates of $\pi_0, \pi_1, \ldots \gamma_1, \gamma_2, \ldots$
- EM algorithm - M step requires iterative procedure
  - Can update $\hat{\gamma}_1, \hat{\gamma}_2, \ldots$ one at a time - Newton-Raphson or optim()

"Significant"

Question: Now which species is discovered?

Local FDR Procedure

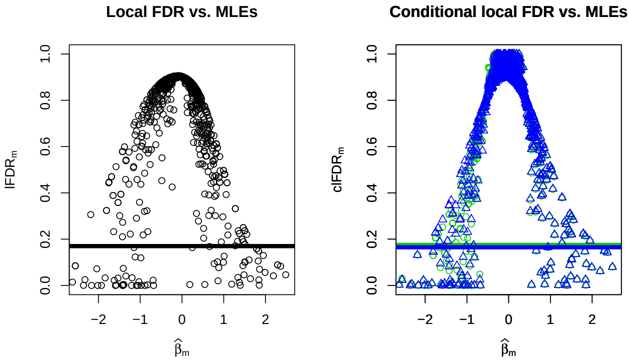
m	Y_{1m}/n_m	Y_{2m}/n_m	Y_{3m}/n_m	Y_{4m}/n_m	Y_{5m}/n_m	$\hat{\beta}_m$	n_m	$\widehat{lFDR}_m$	Disc.
1	0.36	0.50	0.00	0.07	0.07	-1.09	11	0.29	**x**
2	0.15	0.13	0.28	0.25	0.19	0.19	911	0.003	✓
Null	0.20	0.20	0.20	0.20	0.20	0	—	1	x

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

- 3 component pmfs
- 4 component pmfs

Illustration: lFDR vs clFDR

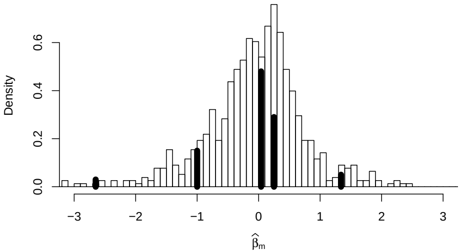


Theorem 1: FDR controlled - based on Sun and Cai(2009) proof

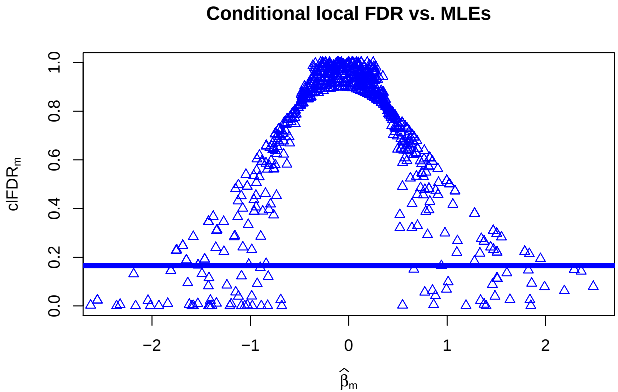
Theorem 2: $[clfdr(z,n) \leq \lambda] \searrow n$ for all $n \geq N$.

Advantages of Finite Mixture Model

- Computationally feasible / consistent parameter estimation
- Flexible: Over-dispersion
- Can *inspect for practical significance* rather than *specify it* apriori
  - Don't have *specify* ϵ in $H_m : \beta_m \in [-\epsilon, \epsilon]$
  - Facilitates follow-up classification analysis if H_m rejected
  - Facilitates power analysis / estimated effect size
- Can reconsider null hypothesis - Efron (2004). Warning: Bickel (2012)

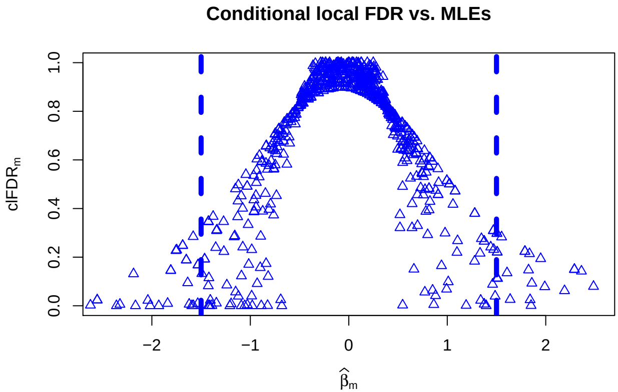


Why clFDR?



Q: Should we use this rejection region?

Why clFDR?



Q: Should we use this rejection region?

A: See Watts and Habiger (2017).

Weighted Adaptive FDR Control

Method:

1. Specify weights $w(n_1), w(n_2), ..., w(n_M)$
   - "Optimal" weights: $w(n_m) \downarrow n_m$ for large enough n_m
2. Compute weighted p-values $Q_m = P_m / w(n_m)$
3. Apply adaptive BH procedure to Q_ms - Storey et. al (2004)

Assessment:

- Finite FDR control and asymptotic FDP control (a.s. under weak dependence)
- Procedure is "α-exhaustive" - See Finner (2009)
- Optimal weights can be consistently estimated
- Simpler weights can be specified (robust)

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.

Cai, T and Sun, W. (2009). Simultaneous Testing of Grouped Hypotheses: Finding Needels in Multiple Haystacks. *Journal of the American Statistical Association 104*(488), 673–687.

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

Habiger, J., D. Watts, and M. Anderson (2017). Multiple testing with heterogeneous multinomial distributions. *Biometrics 73*(2), 562 – 570.

Habiger, J. (2017). Adaptive False Discovery Rate Control for Heterogeneous Data. *Statistica Sinica (in press)*

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.

Watts and Habiger (2017). A New Multiple Testing Protocol for Exploratory Data Analysis and the Local Misclassification Rate. *Communications in Statistics(in press)*