Multiple hypothesis testing: the view using spacings

Debashis Ghosh

Departments of Statistics and Public Health Sciences
Penn State University

Outline

1. Introduction
   - Background
2. B-H procedure
3. Extensions
   - A generalized B-H procedure
   - Dependence
   - Empirical Null Hypothesis
4. Summary
   - Conclusions

Scientific Context

- High-dimensional data are now very commonplace in the medical and scientific literature
- Single nucleotide polymorphisms, next-generation sequencing, fMRI experiments
- Scientific goals: discovering new biology

Differential Expression

- A very common goal: identify genes that change between two (or more) conditions
- This has been termed differential expression in the genomics and statistics literature
- This has spawned a wealth of new statistical methods and new science ("gene expression profiles" in PubMed $\Rightarrow$ 7500+ citations since 1997)

Differential Expression in Statistics

- Reemergence of interest in the multiple comparisons problem
- In the past, much of the literature focused on control of the familywise error rate (FWER)
- More recent interest has focused on the false discovery rate (FDR)

False Discovery Rate: definition

	Accept	Reject	Total
True Null	U	V	n_0
True Alternative	T	S	n_1
	W	Q	n

- FDR $= E[V/Q|Q > 0]P(Q > 0)$; also FDR $= E(FDP)$, where FDP is referred to as the false discovery proportion
- Contrast with FWER $= P(V \geq 1)$
- Control of FDR leads to rejection of more hypotheses than FWER.

Problem Setup

- Assume that for each elementary hypothesis $H_1, \ldots, H_n$ only the p-values $p_1, \ldots, p_n$ are known
- Procedures: compare ordered p-values to cutoffs
- Initially, we will assume that $p_1, \ldots, p_n$ are independent

Procedure of Benjamini and Hochberg (1995)

(a) Let $p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(n)}$ denote the ordered, observed p-values.

(b) Find $\hat{k} = \max\{1 \leq i \leq n : p_{(i)} \leq \alpha i/n\}$.

(c) If $\hat{k}$ exists, then reject null hypotheses $p_{(1)} \leq \cdots \leq p_{(\hat{k})}$. Otherwise, reject nothing.

B-H procedure

- Benjamini and Hochberg (1995) show that using the procedure of Simes (1986) will control the FDR when the hypotheses are independent
- They later show that the procedure is valid under positive regression dependence (Benjamini and Yekutieli, 2001)
- In the case of arbitrary dependence, they propose a Bonferroni-style correction

B-H extensions

- Direct estimation of FDR (Storey, 2002)
- FDR control under dependence (Genovese and Wasserman, 2004; Storey et al., 2004, Efron, 2010, Schwartzman and Lin, 2010)
- Our insight is to cast the B-H procedure in terms of **spacings**

Spacings: a brief background

- The $p-$values are a random sample from $U(0,1)$
- The **spacings** are defined as

$$\tilde{p}_i = p_{(i)} - p_{(i-1)}$$

for $i = 1, \ldots, n+1$, where $\tilde{p}_0 = 0$ and $\tilde{p}_{n+1} = 1$.

- The spacings are dependent, but marginally, they have a Beta$(1, n)$ distribution

B-H procedure: a spacings view

(a) Let $p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(n)}$ denote the ordered, observed p-values.

(b) Define $\tilde{p}_j = p_{(j)} - p_{(j-1)}$, $j = 1, \ldots, n+1$, where $p_{(0)} = 1$ and $p_{(n+1)} = 1$.

(c) Find $\hat{k}$, where

$$\hat{k} = \max\{1 \leq i \leq n : i^{-1} \sum_{j=1}^{i} \tilde{p}_j \leq \alpha n^{-1}(n+1)E(\tilde{p}_1)\}.$$

(d) If $\hat{k}$ exists, then reject null hypotheses $p_{(1)} \leq \cdots \leq p_{(\hat{k})}$. Otherwise, reject nothing.

Remarks

- The expression in (c) compares the cumulative empirical averages of the spacings relative to the expected value, scaled by the FDR plus a factor that is approximately one
- The theoretical expectation is taken with respect to the (marginal) Beta distribution
- The cumulative sum in (c) can be thought of as a type of **scan statistic**
- What determines rejection of hypotheses in the B-H procedure is the clustering of the spacings

Further Extensions

- A generalized B-H procedure
- Dependence
- Empirical null distribution

Further Extensions

- A generalized B-H procedure
- Dependence
- Empirical null distribution

One generalization of the B-H procedure is to apply a monotonic function to the spacings: this leads to the following: **gBH procedure**: Reject $p_{(1)}, \ldots, p_{(\hat{k})}$, where

$$\hat{k} = \max[i : i^{-1} \sum_{j=1}^{i} g(\tilde{p}_j) \leq \alpha E\{g(\tilde{p}_1)\}], \tag{1}$$

g is some suitably chosen monotonic function, and $\hat{k} = 0$ if the set in (1) is empty.

gBH procedure (cont'd.)

- One choice of g: for $\lambda \geq 0$

$$g_\lambda(z) = \begin{cases} z^\lambda & \lambda > 0, \\ \log(z) & \lambda = 0. \end{cases} \tag{2}$$

By exploiting the fact that $\tilde{p}_1$ has a Beta distribution, it is easy to show that

$$E\{g_\lambda(\tilde{p}_1)\} = \begin{cases} B(1+\lambda, n)/B(1, n) & \lambda > 0, \\ \psi(1) - \psi(n+2) & \lambda = 0 \end{cases},$$

where $B(u, v) = \Gamma(u)\Gamma(v)[\Gamma(u+v)]^{-1}$, and

$$\psi(x) \equiv \frac{d}{dx} \log \Gamma(x) = \frac{\Gamma'(x)}{\Gamma(x)}$$

is the digamma function.

gBH procedure: Technical Results

- Assuming independence of p-values, can show exact control of FDR using martingale theory
- Asymptotic FDR control:
  - **Condition A:** Empirical convergence of g-transformed spacings: $\|F_n - F_0\| = \sup_{-\infty<x<\infty} \|F_n(x) - F_0(x)\| \to_p 0$ as $n \to \infty$, where F_n is the empirical cdf of the transformed spacings.
  - Can apply a argmax continuous mapping type result from empirical process theory to show that gBH procedure maintains asymptotic FDR control.

Simulation example

- $n = 300$ hypotheses
- Test statistics are multivariate normal with correlation 0 or 0.3
- $n_1 = 20, 60, 100$
- For true null hypotheses, $\mu = 0$, while for true alternatives, $\mu = 2$
- Apply all procedures at FDR $= 0.05$.
- Study FDR and $1 -$ NDR, where NDR is the non-discovery rate

FDR results of simulation studies

	Independent			Dependent		
Method/n_1	20	60	100	20	60	100
BH	0.04	0.04	0.03	0.001	0.005	0.004
Proposed, $\lambda = 2$	0.002	0.03	0.016	0.04	0.025	0.017
Proposed, $\lambda = 4$	0.01	0.04	0.019	0.03	0.06	0.06
Proposed, $\lambda = 16$	0.04	0.04	0.02	0.02	0.08	0.007

Power ($1 - NDR$) results of simulation studies

	Independent			Dependent		
Method/n_1	20	60	100	20	60	100
BH	0.60	0.96	0.99	0.34	0.63	0.71
Proposed, $\lambda = 2$	0.94	1.00	1.00	0.73	0.89	0.93
Proposed, $\lambda = 4$	0.89	0.96	0.99	0.79	0.91	0.94
Proposed, $\lambda = 16$	0.17	0.36	0.49	0.22	0.32	0.47

Further Extensions

- A generalized B-H procedure
- Dependence
- Empirical null distribution

Dependence

- Intuitively, if p-values have positive correlation, the spacings will exhibit clustering smaller than in the independence case.
- Use ideas from stochastic majorization theory (Nevius et al., 1977)

Dependence (cont'd.)

- Theorem: Assume that the FDP operation is monotonic in the number of rejections. If the spacings corresponding to the joint distribution of **p** is stochastically majorized by the joint distribution of spacings for n independent Uniform(0,1) random variables, and then the gBH procedure will provide exact control of the FDR.

Dependence (cont'd.)

- Example:
  1. Random effects correlation between p-values
  2. $\tilde{p}_i \overset{d}{=} a(p_{(i)} - p_{(i-1)})$, where a is the random effect that is shared by all p-values
  3. Positive correlation does not change the relative rankings of ordered hypotheses
- This structure is also implied by positive regression dependence

Dependence (cont'd.)

- The spacings interpretation suggests that the following correlation structures are problematic for FDR control:

  1. Negative correlation
  2. Long-range dependence between order statistics that does not decay

Further Extensions

- A generalized B-H procedure
- Dependence
- Empirical null distribution

Empirical Null Distribution

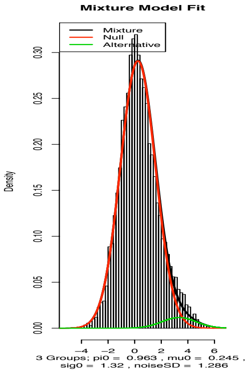
- A device introduced by Efron (2004)
- The idea: in standard practice, for a **single** hypothesis, we use a standard null distribution for a test statistic (Normal, χ^2, F, t,etc.)
- Efron (2004) argues that this is an implausible one for large-scale simultaneous inference problems

Empirical Null (cont'd.)

- Idea: fit a two-component mixture model to test statistics, assume normality for both component distributions, i.e.

$$T_i \overset{iid}{\sim} \pi_0 N(\mu_0, \sigma_0^2) + (1 - \pi_0) N(\mu_1, \sigma_1^2). \tag{3}$$

- Compute $\text{locfdr}(T) = P(H = 0|T)$ using the above mixture model and reject hypotheses for small values.
- Idea is to be more conservative (or more generally, data-adaptive) in selecting "interesting" hypotheses
- No direct way of incorporating empirical null into the original B-H procedure


Mixture Model Fit
Mixture
Null
Alternative
Density
3 Groups; pi0 = 0.963 , mu0 = 0.245 ,
sig0 = 1.32 , noiseSD = 1.286
Effect Size Estimates
delta.hat
z
fdr/FDR curves
fdr/FDR
fdr
FDR 2 Sided
z

Our proposal

- Recall that we earlier expressed the B-H procedure as the following: reject $p_{(1)}, \ldots, p_{(\hat{k})}$, where

$$\hat{k} = \max\{i : i^{-1} \sum_{j=1}^{i} \tilde{p}_j \leq \alpha(n+1)n^{-1} E(\tilde{p}_1)\}, \qquad (4)$$

  where the expectation is taken with respect to a Beta distribution.

- Our proposal: model the spacings with respect to some other distribution.

Empirical Null Distribution - Proposal # 1

- Fit a two-component Beta mixture to the spacings (in practice, use nonzero spacings)
- Compute the mean based on the mixture model, averaged over components, plug in (5)
- New rule: reject $p_{(1)}, \ldots, p_{(\hat{k})}$, where

$$\hat{k} = \max\{i : i^{-1} \sum_{j=1}^{i} \tilde{p}_j \leq \alpha(n+1)n^{-1} E_{\hat{G}}(\tilde{p}_1)\},$$

where $\hat{G}$ is the mixture model

Empirical Null Distribution - Proposal # 2

- Bayesian flavor
- Requires another dataset

1. Given the original set of p-values, construct an estimator of $E(\tilde{p}_1)$; an obvious estimator would be the empirical average of the spacings.
2. Compute Y, the observed number of spacings in the second less than the estimator from the previous step.
3. Compute $E(\tilde{p}_1 | Y = y)$.
4. Apply the following methodology to the original set of p-values: reject $p_{(1)}, \ldots, p_{(\hat{k})}$, where

$$\hat{k} = \max\{i : i^{-1} \sum_{j=1}^{i} \tilde{p}_j \leq \alpha(n+1)n^{-1} E(\tilde{p}_1 | y)\}, \qquad (5)$$

Empirical Null Distribution and FDR control

- We can use stochastic ordering ideas to show FDR control
- $X <_{s.t.} Y$ if $P(X > x) < P(Y > x)$ for all x
- If V, the distribution for spacings, is stochastically smaller than a Beta$(1, n)$ r.v., then FDR control is achieved.
- This can be either asymptotic or exact

Real Data Example: Mixed lineage leukemia

- This is a type of cancer affecting red blood cells
- Data from `http://www.broadinstitute.org/cgi-bin/cancer/datasets.cgi`
- Some preprocessing steps taken
- 14240 genes
- On test statistic scale, mixture model approach of Efron (2004) does not converge
- By contrast, no problem for our proposed methods

Real Data Example: Mixed lineage leukemia (cont'd.)

- BH using theoretical null: 10685 genes found "interesting"
- BH using empirical null with two-component mixture model: 10777 genes found "interesting"
- BH using Bayesian empirical null:
  1. Suppose there was a second study with the same number of genes and number of spacings less than emprical average was $0 \Rightarrow 10238$ genes interesting.
  2. Suppose there was a second study with the same number of genes and number of spacings less than emprical average was $1 \Rightarrow 10668$ genes interesting.
- Empirical null is restricted to $[0, 1]$

Conclusions and Future Work

- The proposed spacings justification for Benjamini-Hochberg gives new insight on multiple comparisons problem.
- Ideas from scan statistics and clustering can be brought to bear to this problem
- Understanding dependence
- Innovative empirical null idea
- Lots of possible extensions
  1. Looking at higher-order gaps to account for dependence
  2. Multivariate definitions of spacings

Future genomic applications

- Multivariate data integration problems
- Heterogeneity in genomic meta-analyses
- Periodicity of genes
- Multigroup differential expression
- Genomic Outlier Profile Analysis (Ghosh and Chinnaiyan, 2009)

References

- Ghosh, D. (2011). Generalized Benjamini-Hochberg procedures using spacings, technical report.
- Ghosh, D. (2011). Dependence and the empirical null hypothesis within the B-H procedure, in preparation.
- Software in preparation
- First manuscript is available, second manuscript will be available at `http://works.bepress.com/debashis_ghosh/`