

Bayesian Multiple Hypothesis Testing

Sudipto Banerjee

Division of Biostatistics
School of Public Health
University of Minnesota

May 1, 2008

Many Hypothesis Tests

- Technological advances in many fields have resulted in massive amounts of data being processed.
- Statisticians increasingly encounter massive multiple comparison situations which involve simultaneous screening of many (hundreds or thousands) of hypotheses to determine whether we have 'noise' or 'signals'.
- A typical example is in gene expression (microarrays), when many genes are tested for differential expression among different treatments.
- Yet another example is in spatial epidemiology: are adjacent regions truly significantly different from one another?

- Assume $\mathbf{y} = (y_1, y_2, \dots, y_M)$. We want to perform M tests of hypotheses:

$$H_{0i} : y_i \sim f_{0i} \text{ vs. } H_{1i} : y_i \sim f_{1i}.$$

Here y_i can be a vector. Often y_i is a test statistic, and f_{1i} and f_{0i} involve unknown parameters.

- Each hypothesis corresponds to a model.
- Let $\gamma = (\gamma_1, \dots, \gamma_M)$ be a set of “model indicators”:

$$\gamma_i = 1 \text{ if } H_{1i} \text{ is true}$$

$$\gamma_i = 0 \text{ if } H_{0i} \text{ is true.}$$

- The multiple testing problem can thus be formulated as a model selection problem: to choose among the 2^M models indexed by all possible values of γ .

- Frequentists usually address this problem using generalizations of hypothesis testing tools. Model selection tools are used rarely, if ever.
- If the M tests are independent and each is tested at level α , then, even when $\gamma = \mathbf{0}$, we expect αM rejections. In simultaneous testing, this is perceived as ‘too many’, unduly masking detection of incorrect null.
- From a Bayesian point of view, *and in contrast with most problems of model choice*, the posterior probability of the models is not the object of main interest, but rather the probability that each γ_i is non-zero (inclusion probabilities).

An Example: Scott and Berger (*JSPI*; 2006)

- Observe $y_i \sim N(\mu_i, \sigma^2)$, $i = 1, \dots, M$, (σ^2 unknown).
- To determine which μ_i 's are non-zero, we have M independent tests:

$$H_{0i} : \mu_i = 0 \quad vs. \quad H_{1i} : \mu_i \neq 0.$$

- $p_0 = P(\gamma_i = 0)$ is the prior probability that H_{0i} is true.
- **IMPORTANT:** the Bayesian will let the data estimate p_0 .

- Scott and Berger (2006) adopt the following hierarchical model:

1. $y_i | \mu_i, \sigma^2 \stackrel{iid}{\sim} N(\gamma_i \mu_i, \sigma^2)$
2. $\mu_i | \tau^2 \stackrel{iid}{\sim} N(0, \tau^2); \quad \gamma_i | p_0 \stackrel{iid}{\sim} Ber(p_0)$
3. $(\sigma^2, \tau^2) \sim \pi(\sigma^2, \tau^2); \quad p_0 \sim \pi(p_0).$

- Any proper prior for σ^2 and τ^2 will work.
- Scott and Berger derive an improper “objective” prior:

$$\pi(\sigma^2, \tau^2) = \pi(\tau^2 | \sigma^2)\pi(\sigma^2) = \frac{1}{(\tau^2 + \sigma^2)\sigma^2}$$

- The usual objective prior $\pi(\sigma^2, \tau^2) = (\tau^2 + \sigma^2)^{-1}$ is improper and will lead to improper posteriors since τ^2 does not occur in all models.

- Priors for p_0 could be *Unif*, *Beta* or other informative priors.
- The quantities of interest here are:
 - The probabilities that each of the hypothesis is true (i.e. $\gamma_i = 1$)
 - The distributions of the 'signals' if the corresponding null is not true.
 - The posterior probability of models (i.e. $p(\gamma | \mathbf{y})$) are not of main interest.

- Consider the following tabulation of our M hypothesis tests:

	Accept H_0	Reject H_0	
H_0 is true ($\gamma_i = 0$)	U	V	M_0
H_1 is true ($\gamma_i = 1$)	T	S	M_1
observed	W	R	M

- $M_1 = \sum_{i=1}^M \gamma_i$ and $M_0 = M - M_1$ – both are unobserved.
- R is the total number of rejections (discoveries)
- V is the total number of “false discoveries”.
- Frequentists almost never concern themselves about T in the non-Bayesian literature – this is controversial from a decision-theoretic perspective.

Family-wise Error Rate (FWER)

- This is the classical solution to deal with multiplicity
- Based on total number of false rejections (discoveries) V
- FWER is defined as:

$$\text{FWER} = P(V \geq 1) = P\left(\bigcup_{i=1}^M H_{1i} \mid H_{0i}\right) \leq \sum_{i=1}^M P(H_{1i} \mid H_{0i}).$$

- Bonferroni: controls at level $\leq \alpha$ by testing each H_{0i} at level $\frac{\alpha}{M}$.
- Results in very conservative tests which can result in very low power.

False Discovery Rate (FDR)

- Benjamini and Hochberg (1995) argued that the interesting quantity to control is the % of false discoveries (erroneous rejections) among the rejected hypotheses.
- Ideally based upon $\frac{V}{R}$. **Caution:** This is not defined for $R = 0$ (all M nulls accepted)
- Several modifications suggested:

$$\text{FDR} = E \left[\frac{V}{\max\{R, 1\}} \right] = E \left[\frac{V}{R} \mid R > 0 \right] \times P(R > 0)$$

$$\text{pFDR: Positive FDR} = E \left[\frac{V}{R} \mid R > 0 \right]$$

$$\text{PFP: Proportion of False Positives} = \frac{E[V]}{E[R]}.$$

- Benjamini and Hochberg (1995) was the pioneering work that started the interest in the FDR-type error rates. They argue that, when all the nulls are true, ($M_0 = M$):

$$\text{pFDR} = E \left[\frac{V}{R} \mid R > 0 \right] = 1 = \frac{E[V]}{E[R]} = \text{PFP}.$$

- Thus, neither pFDR nor PFP can be controlled, so they choose FDR as a 'good' error rate which can be controlled.
- $\text{FDR} \leq \text{FWER}$ with $\text{FDR} = \text{FWER}$ when all nulls are true. So, controlling FDR leads to less conservative tests than FWER (e.g. Bonferroni).

- Mueller, Parmigiani and Rice (2006) develop a Bayesian decision-theoretic framework FDR.
- Let $\delta_i(\mathbf{y}) = 1$ if the i -th null is rejected and $\delta_i(\mathbf{y}) = 0$ otherwise. Here $\delta_i(\mathbf{y})$ is a decision rule that does not depend upon parameters.
- Then $V = \sum_{i=1}^M \delta_i(1 - \gamma_i)$ and $R = \sum_{i=1}^M \delta_i$.
- The the posterior expected FDR is:

$$\begin{aligned}
 E \left[\frac{V}{R} \mid \mathbf{y} \right] &= E \left[\frac{1}{\sum_{i=1}^M \delta_i} \sum_{i=1}^M \delta_i(1 - \gamma_i) \mid \mathbf{y} \right] \\
 &= \frac{\sum_{i=1}^M \delta_i (1 - E[\gamma_i \mid \mathbf{y}])}{\sum_{i=1}^M \delta_i}
 \end{aligned}$$

- The final decision to identify signal can be developed using simple threshold rules or even loss functions.