Multiple Hypothesis Testing with Groups

James X. Hu

Yale University
Jointly with Hongyu Zhao and Harrison H. Zhou

May 15, 2009

Group information is valuable

- In large-scale multiple hypothesis testing, the false discovery rate (FDR) has been widely adopted to replace traditional methods.
- In many cases, the hypotheses can be divided into groups based on additional information.
  - Gene Ontology for gene expression data.
  - Phenotypes for genome-wide association studies.
- Clustering is an alternative when the group structure is less apparent.
- Use group information via the proportion of true null hypotheses, $\pi_{g,0}$, which represents the relative importance of each group.

The Group Benjamini-Hochberg (GBH) procedure

The GBH procedure (for known $\pi_{g,0}$'s)

For n p-values that can be divided into K groups.

1. For each $P_{g,i}$ in group g, compute $P_i^w = \dfrac{\pi_{g0}}{1-\pi_{g0}} P_{g,i}$.
2. Let $P_{(1)}^w \leq \ldots \leq P_{(n)}^w$ be the ordered pooled p-values. Compute

$$k = \max\left\{ i : P_{(i)}^w \leq \frac{i\alpha^w}{n} \right\}, \quad \alpha^w = \frac{\alpha}{1-\bar{\pi}_0},$$

   where $\bar{\pi}_0$ is the overall proportion of true nulls among n hypotheses.
3. If such a k exists, reject the k hypotheses associated with $P_{(1)}^w, \ldots, P_{(k)}^w$; otherwise do not reject any of the hypotheses.

Theorem

Assume $\pi_{g,0}$ is known for each group, if the p-values are independent, the GBH procedure controls the FDR at level α.

The adaptive procedure

The adaptive GBH procedure

1. Estimate the $\pi_{g,0}$ in each group by $\hat{\pi}_{g,0}$.
2. Apply the GBH procedure with $\pi_{g,0}$ replaced by $\hat{\pi}_{g,0}$.

- The adaptive GBH procedure does not depend on how one estimates $\pi_{g,0}$.

Theorem

The adaptive GBH procedure controls the FDR asymptotically for p-values under weak dependence.

Applications

- A well known breast cancer data (van't Veer et al, 2002), 24,184 genes.
- Grouping using k-means clustering with 6 clusters.