Optimal Power Prediction in Large Scale Multiple Testing:

A Fourier Approach

Avranil Sarkar

Department of Statistics, Carnegie Mellon University

May 16, 2009

Problem

- In many settings of large-scale multiple testing, signal are faint.
- Example: gene microarray
- Even optimal procedures lack power
- **Solution**: need more samples
- **Goal**: decide sample size economically

Equivalently: predict power from present data

Predicting Testing Power

- given data in **current** stage
- in a **future** follow-up study, we plan to enlarge sample size by n times
- how to predict (**future**)-testing power using **present** data?

Gaussian Hierarchical Model

We are testing p hypotheses

$H_i : \mu_i = 0 \qquad 1 \leq i \leq p.$

Each hypothesis has prob. ϵ to be false

- current z-scores:
  - $X_i \sim N(\mu_i, 1)$

$$\mu_i = \begin{cases} 0, & \text{when } H_i \text{ is true} \\ \text{samples from } F, & \text{otherwise} \end{cases}$$

  - Marginally,

$$X_i \overset{iid}{\sim} (1-\epsilon)\phi(x) + \epsilon \int \phi(x-u)dF(u).$$

- future z-scores (sample size enlarged by n times):

$$\mu_i \qquad \longrightarrow \qquad \sqrt{n} \cdot \mu_i$$

Conclusion

We propose a method to predict the power for testing with enlarged sample size based on the current $z-$scores.