Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing and Prediction

Ch.2 Large-Scale Hypothesis Testing

Huei-Lun Siao

Szu-Han Lin

January 4, 2018

Outline

1. Two-Groups Model: A Microarray Example
2. Bayesian Approach
3. Empirical Bayes Estimates
4. $\overline{\mathrm{Fdr}}(\mathcal{Z})$ as a Point Estimate
5. Independence versus Correlation

1. Two-Groups Model: A Microarray Example
2. Bayesian Approach
3. Empirical Bayes Estimates
4. $\overline{\mathrm{Fdr}}(\mathcal{Z})$ as a Point Estimate
5. Independence versus Correlation

Two-Groups Model: A Microarray Example

There is a microarray example, the *prostate data*.

Goal: To discover genes whose expression levels differ between the prostate and normal subjects.

- $N = 6033$ genes
- 50 normal control subjects and 52 prostate cancer patients
- Data matrix

	the normal contral (1,2,...,50)	the cancer patients (51,53,...,102)
gene: N=6033	x_{ij} = level for gene i on patient j,	

Hypothesis Testing

- H_{0i} : gene i is "null"
- The two-sample t-statistic for testing gene i

$$t_i = \frac{\bar{x}_i(2) - \bar{x}_i(1)}{s_i},$$

  where

  - $\bar{x}_i(1), \bar{x}_i(2)$: the averages of x_{ij} for the normal controls and for the cancer patients.
  - $$s_i^2 = \frac{\Sigma_1^{50}(x_{ij} - \bar{x}(1))^2 + \Sigma_{51}^{102}(x_{ij} - \bar{x}(2))^2}{100} \times \left(\frac{1}{50} + \frac{1}{52}\right)$$

- The usual α rejection criterion (α=5%)
- Based on normal theory reject H_{0i}, if $|t_i| > t_{100}(\alpha)$

Using z-values instead of t-values

- $t_i \sim t_\nu$ (here $\nu = 100$)
- We transform t_i to

$$z_i = \Phi^{-1}(F_\nu(t_i))$$

  where Φ and F_ν are the cumulative distribution functions for standard normal and t_ν distributions
- $z_i \sim \mathcal{N}(0, 1)$

Rewriting Hypothesis Testing

- H_{0i} : gene i is "null"
  - The two-sample t-statistic for testing gene i

$$t_i = \frac{\bar{x}_i(2) - \bar{x}_i(1)}{s_i},$$

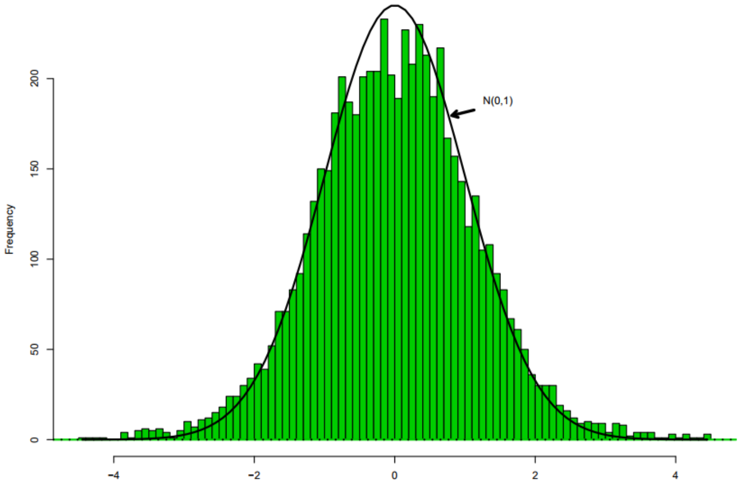
  - We transform t_i to

$$z_i = \Phi^{-1}(F_{100}(t_i))$$

- H_{0i} : $z_i \sim \mathcal{N}(0, 1)$
- The usual two-sided 5% test

$$\text{rejects } H_{0i} \text{ for } |z_i| > 1.96.$$

$N = 6033$ genes



Multiple testing

- H_0: all of the genes were "null"
- The Bonferroni bound approach:

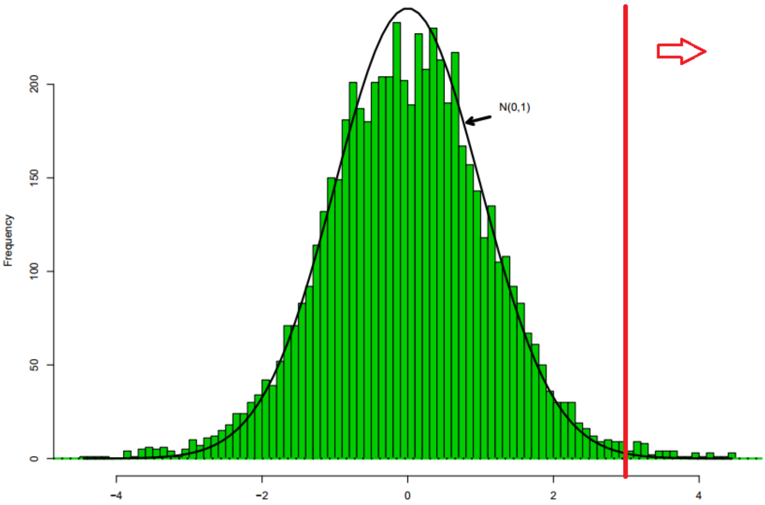
The rejection level for each test from 0.05 to 0.05/6033.

- H_{0i} : gene i is "null"

$$|z_i| > 4.31$$

Problem: $\mathcal{Z} = (-\infty, -4.31) \cup (4.31, \infty)$ seems overly cautious

Set rejection region $\mathcal{Z} = (3, \infty)$, we observe 49 z_i values in $\mathcal{Z}$.



Problem: Is every gene really null?

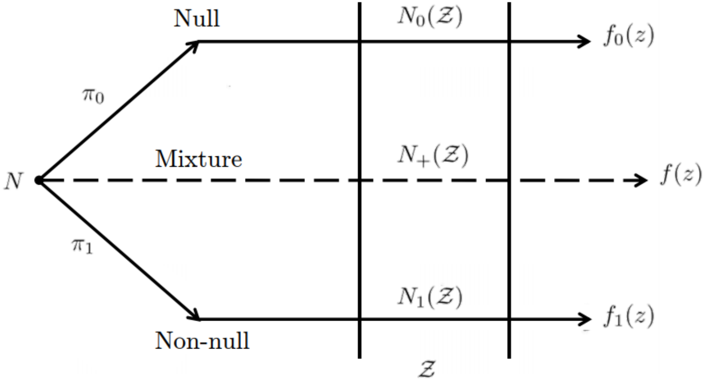
Bayesian Approach

- We suppose that the N cases are each either null or non-null with prior probability π_0 or $\pi_1 = 1 - \pi_0$,

$$\begin{aligned} \pi_0 &= \Pr\{\text{null}\} & f_0(z) &= \text{density if null} \\ \pi_1 &= \Pr\{\text{non-null}\} & f_1(z) &= \text{density if non-null} \end{aligned} \tag{1}$$

- π_0 will be much bigger than π_1, say

$$\pi_0 \geq 0.9$$



- The mixture density: $f(z) = \pi_0 f_0(z) + \pi_1 f_1(z)$
- If $\mathcal{Z} = (3, \infty)$, $N_+(\mathcal{Z}) = 49$.

Multiple testing by Bayesian Approach

- H_0: all of the genes were "null"
- Given rejection region $\mathcal{Z}$
- We would like to know, but can't observe, the *false discovery proportion*

$$\text{Fdp}(\mathcal{Z}) = \frac{\color{red}{N_0(\mathcal{Z})}}{N_+(\mathcal{Z})}$$

- If $\text{Fdp}(\mathcal{Z})$ is small, reject H_0.

Some Notation

- Assume $H_{0i}: \ z_i \ \sim \mathcal{N}(0, 1)$,
  - $f_0(z) = \varphi(z) = e^{-\frac{1}{2}z^2}/\sqrt{2\pi}$
  - $f_1(z)$ might be some alternative density yielding z-values further away from 0.
- For any subset $\mathcal{Z}$ of the real line,

$$F_0(\mathcal{Z}) = \int_{\mathcal{Z}} f_0(z)dz \quad \text{and} \quad F_1(\mathcal{Z}) = \int_{\mathcal{Z}} f_1(z)dz$$

- The mixture density: $f(z) = \pi_0 f_0(z) + \pi_1 f_1(z)$
- The mixture probability distribution: $F(\mathcal{Z}) = \pi_0 F_0(\mathcal{Z}) + \pi_1 F_1(\mathcal{Z})$

- The *Bayes false discovery rate* for $\mathcal{Z}$

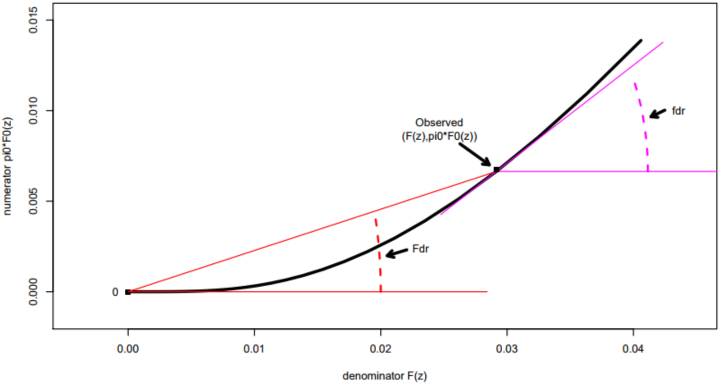
$$\phi(\mathcal{Z}) \equiv \Pr\{\text{null}|z \in \mathcal{Z}\} = \frac{\pi_0 F_0(\mathcal{Z})}{F(\mathcal{Z})} = \text{Fdr}(\mathcal{Z})$$

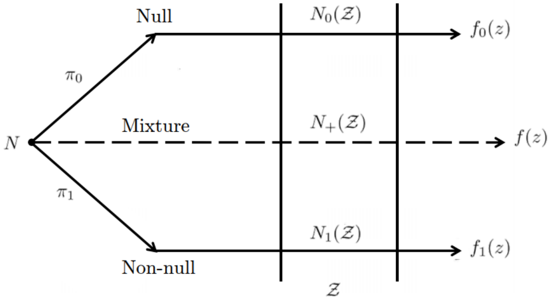
- The *local Bayes false discovery rate*

$$\phi(z_0) \equiv \Pr\{\text{null}|z = z_0\} = \frac{\pi_0 f_0(z_0)}{f(z_0)} = \text{fdr}(z_0)$$

- Let $\mathcal{Z} = (-\infty, z)$,
  - $\phi((-\infty, z)) \equiv \text{Fdr}(z) = \pi_0 F_0(z)/F(z)$
  - $\phi(z) \equiv \text{fdr}(z) = \pi_0 f_0(z)/f(z)$

Relationship between Fdr(z) and fdr(z)





null	π_0	$F_0(\mathcal{Z})$	$N_0(\mathcal{Z})$	$e_0(\mathcal{Z}) = E(N_0(\mathcal{Z}))$
non-null	π_1	$F_1(\mathcal{Z})$	$N_1(\mathcal{Z})$	$e_1(\mathcal{Z}) = E(N_1(\mathcal{Z}))$
mixture		$F(\mathcal{Z})$	$N_+(\mathcal{Z})$	$e_+(\mathcal{Z}) = E(N_+(\mathcal{Z}))$

$N_+(\mathcal{Z}) = \#\{z_i \in \mathcal{Z}\}$, $e_0(\mathcal{Z}) = N\pi_0 F_0(\mathcal{Z})$, and $\bar{F}(\mathcal{Z}) = N_+(\mathcal{Z})/N$

null	π_0	$F_0(\mathcal{Z})$	$N_0(\mathcal{Z})$	$e_0(\mathcal{Z}) = E(N_0(\mathcal{Z}))$
non-null	π_1	$F_1(\mathcal{Z})$	$N_1(\mathcal{Z})$	$e_1(\mathcal{Z}) = E(N_1(\mathcal{Z}))$
mixture		$\bar{F}(\mathcal{Z})$	$N_+(\mathcal{Z})$	$e_+(\mathcal{Z}) = E(N_+(\mathcal{Z}))$

- Estimate false discovery rate

$$\overline{\mathrm{Fdr}}(\mathcal{Z}) \equiv \bar{\phi}(\mathcal{Z}) = \frac{\pi_0 F_0(\mathcal{Z})}{\bar{F}(\mathcal{Z})} = \frac{e_0(\mathcal{Z})}{N_+(\mathcal{Z})}$$

- The false discovery proportion $\mathrm{Fdp}(\mathcal{Z})$ is still unknown.

$$\mathrm{Fdp}(\mathcal{Z}) = \frac{N_0(\mathcal{Z})}{N_+(\mathcal{Z})}$$

Example: The prostate data

- The prostate data has $N_+(\mathcal{Z}) = 49$ z_i values in $\mathcal{Z} = (3, \infty)$,

$$e_0(\mathcal{Z}) = 6.033 \;\cdot\; \pi_0 \cdot (1 - \Phi(3))$$

- The upper bound $\pi_0 = 1$ gives $e_0(\mathcal{Z}) = 8.14$ and

$$\overline{\mathrm{Fdr}}(\mathcal{Z}) = 8.14/49 = 0.166$$

$\overline{\mathrm{Fdr}}(\mathcal{Z})$ as a Point Estimate

- There are three quantities to consider,

$$\overline{\mathrm{Fdr}}(\mathcal{Z}) = \frac{e_0(\mathcal{Z})}{N_+(\mathcal{Z})}, \quad \phi(\mathcal{Z}) = \frac{e_0(\mathcal{Z})}{e_+(\mathcal{Z})}, \quad \text{and} \quad \mathrm{Fdp}(\mathcal{Z}) = \frac{N_0(\mathcal{Z})}{N_+(\mathcal{Z})}$$

Lemma

Suppose $e_0(\mathcal{Z}) = N\pi_0 F_0(\mathcal{Z})$ *is the same as the conditional expectation of* $N_0(\mathcal{Z})$ *given* $N_1(\mathcal{Z})$. *Then the conditional expectations of* $\overline{Fdr}(\mathcal{Z})$ *and* $Fdp(\mathcal{Z})$ *given* $N_1(\mathcal{Z})$ *satisfy*

$$E\{\overline{Fdr}(\mathcal{Z})|N_1(\mathcal{Z})\} \geq \phi_1(\mathcal{Z}) \geq E\{Fdr(\mathcal{Z})|N_1(\mathcal{Z})\}$$

where

$$\phi_1(\mathcal{Z}) = \frac{e_0(\mathcal{Z})}{e_0(\mathcal{Z}) + N_1(\mathcal{Z})}.$$

Lemma

Let $\gamma(\mathcal{Z})$ *indicate the squared coefficient of variation of* $N_+(\mathcal{Z})$,

$$\gamma(\mathcal{Z}) = var\{N_+(\mathcal{Z})\}/e_+(\mathcal{Z})^2.$$

Then $\overline{Fdr}(\mathcal{Z})/\phi(\mathcal{Z})$ *has approximate mean and variance*

$$\frac{\overline{Fdr}(\mathcal{Z})}{\phi(\mathcal{Z})} \dot{\sim} (1 + \gamma(\mathcal{Z}), \gamma(\mathcal{Z})).$$

Independence Assumption

- Each z_i follows (1) independently.
- Then $N_+(\mathcal{Z}) \sim \mathrm{Bi}(N, F(\mathcal{Z}))$ with squared coefficient of variation

$$\gamma(\mathcal{Z}) = \frac{1 - F(\mathcal{Z})}{NF(\mathcal{Z})} = \frac{1 - F(\mathcal{Z})}{e_+(\mathcal{Z})}$$

- Giving $\gamma(\mathcal{Z}) \doteq 1/e_+(\mathcal{Z})$,

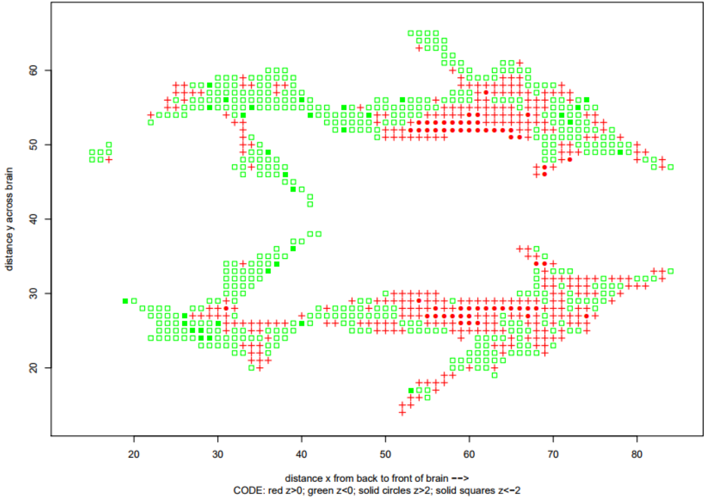
$$\overline{\mathrm{Fdr}}(\mathcal{Z})/\phi(\mathcal{Z}) \dot{\sim} (1 + 1/e_+(\mathcal{Z}), 1/e_+(\mathcal{Z}))$$

- For the prostate data,
  - $\overline{\mathrm{Fdr}}(\mathcal{Z}) = 0.166 = \phi(\mathcal{Z})$
  - variation about 0.14
  - A rough 95% confidence interval for $\phi(\mathcal{Z})$ is

$$0.166 \cdot (1 \pm 2 \cdot 0.14) = (0.12, 0.21)$$

Independence versus Correlation

- There is DTI (Diffusion Tensor Imaging) data.
- The study comparing brain activity of six dyslexic children versus six normal controls.
- Two-sample tests
  - $N = 15443$ voxels
  - Each $z_i \sim \mathcal{N}(0, 1)$
  - H_0: no difference between the dyslexic and normal children



red: $z_i > 0$; green: $z_i < 0$; solid circles $z_i > 2$; solid squares $z_i < 2$