Correlation and Large-Scale Simultaneous Significance Testing

Bradley Efron

Ref Efron (2005)
www-stat.stanford.edu/brad/papers/Correlation.pdf
and Efron (2004), *JASA*, 96-104.

• With Thanks to Trevor Hastie and Rob Tibshirani.

Large-Scale Significance Testing

Cases ("genes") $1, 2, 3, \ldots, N$

Null Hypotheses $H_1, H_2, H_3, \ldots, H_N$

Test Statistics $z_1, z_2, z_3, \ldots z_N$

- $H_i : z_i \sim \mathcal{N}(0, 1)$ (Not necessarily independent)
  "theoretical null"
- *Question* How much correlation is there,
  and how does it affect inferences?
- *Answer* Even if theoretical null is marginally correct,
  correlation can make it wrong simultaneously

BRCA Study

- *15 Subjects* 7 BRCA1, 8 BRCA2
- *15 Microarrays* Same 3226 genes
- *Two-Sample t-statistic* "t_i" comparing BRCA2's with BRCA1's
- *Transformation* $z_i = \Phi^{-1}(F_{13}(t_i))$

so $z_i \sim \mathcal{N}(0,1)$ under H_0 if F_{13} correct null for t_i.

False Discovery Rates

- $Y_1 = \#\{\text{Null genes with } z_i \geq 2.5\}$

and

$T_1 = \#\{\text{genes with } z_i \geq 2.5\}$ $(= 112)$

- *Actual Fdr* $= Y_1/T_1$
- "*Expected*" *Fdr* $= E\{Y_1\}/T_1$

Question What is $E\{Y_1\}$?

Theoretical Null $E\{Y_1\} = 20$

Empirical Null $E\{Y_1\} = 118$

Permutation Calculations

- *Permute* the 15 microarrays:

  $X^* \rightarrow \mathbf{z}^* \rightarrow \mathbf{y}^*$

- *4000 Permutations* $\mathbf{y}^{*1}, \mathbf{y}^{*2}, \ldots, \mathbf{y}^{*4000}$;

- *Estimate cov* $(\mathbf{y})$ by

$$\boxed{C_{\text{perm}} = \sum (\mathbf{y}^{*b} - \mathbf{y}^{*\cdot})(\mathbf{y}^{(*b} - \mathbf{y}^{*\cdot})'/3999}$$

$$\left(\mathbf{y}^{*\cdot} = \sum_{b=1}^{4000} \mathbf{y}^{*b}/B \quad \dot{\sim} N(0,1) \right)$$

- Permutations preserve correlation structure

  But nullify genuine BRCA1-BRCA2 differences, so

- C_{perm} estimates *null* covariance of $\mathbf{y}$

Tail Counts and Central Counts

- *Tail Counts* $Y_1^* = \#\{z_i^* \geq 2.5\}$

  $\sim (20.0, 14.9^2)$ [indep : $(20, 4.5^2)$]

- *Central Counts* $Y_0^* = \#\{z_i^* \text{ in } [-1, 1]\}$

- *Main Idea* Can predict tail counts from central counts; use to get better estimates of Fdr numerator "$E\{Y_1\}$."

- Easy to see for the 4000 perms

  ("simulation study" for predicting actual Y_1)

Explanation (1)

- *Typical* $\mathbf{z}^*$: $\boxed{z_i^* \dot{\sim} N(0, \sigma_0^{*2})}$

  (next slide)

- *Central Estimate* of σ_0^*:

  $\widehat{\sigma}_0^* = 1/\Phi^{-1}\left(\frac{1+Y_0^*/N}{2}\right)$

  (like interquartile range)

- Y_0^* small $\Rightarrow \widehat{\sigma}_0^*$ big

  $\Rightarrow z_i^*$'s more dispersed $\Rightarrow Y_1^*$ big

Estimating the Null Tail Counts

- $Y_1 = \#\{\text{Null } z_i \geq 2.5\}$ is unobservable
- But central null count Y_0 "almost observable"
- $Y_0 \to \widehat{\sigma}_0 \to$ conditional estimate of Y_1
- *BRCA*: $E\{Y_1|\widehat{\sigma}_0\} \doteq 118$, Fdr $\doteq 118/112$
- *Empirical Null* gives $E\{Y_1|\widehat{\mu}_0, \widehat{\sigma}_0\}$

Bivariate Normal Theory

- Assume all $z_i \sim N(0,1)$
- Also all pairs (z_i, z_j) bivariate

  normal, correlation ρ_{ij}
- Let $g(\rho)$ be empirical density of

  all $\binom{N}{2}$ correlations
- *BRCA* $g(\rho) \sim N(0, \alpha^2),\ \alpha = 0.153$

  $\Rightarrow \text{cov}(\mathbf{y}) =$ "C_{norm}" $(\doteq C_{\text{perm}})$
- *Note* Columns X standardized $\Rightarrow E\{\rho\} \doteq 0$

Wing-Shaped Function

- $E\{\mathbf{y}\} = \mathbf{V}$ with $V_k \doteq N\Delta\varphi(x_k)$

[Δ = bin width, x_k = bin midpoint,

$\varphi(x) = \exp\{-x^2/2\}/\sqrt{2\pi}$]

- *Wing-shaped function* $\mathbf{W}$ proportional φ'',

$W_k = N\Delta\varphi(x_k)\frac{x_k^2-1}{\sqrt{2}}$

- $\boxed{C_{\text{norm}} \doteq C_0 + \alpha^2\mathbf{W}\mathbf{W}'}$

where $C_0 = \text{cov}(\mathbf{y})$ under independence

- α summarizes entire correlation structure

Poisson Model

- *Notation* $\mathbf{y} \sim Po(\mathbf{u})$ means $y_k \overset{\text{ind}}{\sim} Po(u_k)$

- *Hierarchical Poisson Model*

(1) $\mathbf{U} = \mathbf{V} + A\mathbf{W}$ where $A \sim (0, \alpha^2)$

(2) $\mathbf{y}|\mathbf{u} \sim Po(\mathbf{u})$

Independence Case $\alpha = 0$ (all $\rho_{ij} = 0$)

$\Rightarrow \mathbf{y} \sim Po(\mathbf{V})$

(exact if N Poisson)

Explanation of $(\mathbf{Y_0}, \mathbf{Y_1})$ Relationship (2)

- $\mathbf{y}|\mathbf{u} \sim Po(\mathbf{u})$ with $\mathbf{u} = \mathbf{V} + A\mathbf{W}$
- $A > 0$: $\mathbf{u}$ depressed in center (smaller Y_0)
  raised in tails (bigger Y_1)
- Conversely for $A < 0$
- *Approximation* $\mathbf{u} \propto N(0, \sigma_A^2)$ with

$$\sigma_A^2 = 1 + \sqrt{2}\, A$$

(σ_A like $\widehat{\sigma}_0^*$)

Summary

• Even if theoretical null $N(0, 1)$ is individually correct, correlations between z-values can make effective null distribution wider or narrower

• Center of z histogram estimates the "empirical null", taking into account correlation and other effects

• Empirical null gives more realistic estimates of null tail counts (better Fdr's)

• *BRCA* Nothing Significant