Discussion: Sun & Cai, JASA 2007

John Pura
BIOS900

September 30, 2016
Hypothesis Testing

- Well developed in single test scenario (e.g. Neyman-Pearson Lemma)
- Goal: maximize power (minimize Type II error) while controlling Type I error
- Procedure (frequentist): Compute test statistic (e.g. likelihood ratio) and compare to a single critical value, while maintaining Type I error
Multiple Hypothesis Testing

- Controlling Type I error more complicated when we have many tests
- Common problem in many big-data applications - e.g. imaging, genomics, Netflix: number of tests on the order of $10^3$-$10^6$
- Goal: separate the null from the nonnulls while still balancing Type I and Type II errors
- Power is critical in these applications because the most interesting effects are usually at the edge of detection.
- Rich literature on the topic spanning frequentist and Bayesian approaches

Discussion: Sun & Cai, JASA 2007
Data Structures in Large-Scale Studies

- \( m \) features (e.g. voxels, genes, movies) and \( k \) groups

Table 1: Common Data Matrix in Large Scale Studies

\[
\begin{array}{cccc}
X_1 & \cdots & X_{n_1} & \cdots & X_k & \cdots & X_{n_k} \\
X_{11} & \cdots & X_{n_11} & \cdots & X_{k1} & \cdots & X_{n_k1} \\
X_{12} & \cdots & X_{n_12} & \cdots & X_{k2} & \cdots & X_{n_k2} \\
\vdots & & \vdots & & \vdots & & \vdots \\
X_{1m} & \cdots & X_{n_1m} & \cdots & X_{km} & \cdots & X_{n_km} \\
\end{array}
\]
Suppose we perform $m$ simultaneous hypothesis tests of $H_{0,i}$ vs. $H_{1,i}$, $1 \leq i \leq m$ under a given procedure.

Summarize the outcomes of such a procedure as follows:

Table 2: Classification of tested hypotheses

<table>
<thead>
<tr>
<th></th>
<th>Claimed nonsignificant</th>
<th>Claimed significant</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Null</td>
<td>$N_{00}$</td>
<td>$N_{10}$</td>
<td>$m_0$</td>
</tr>
<tr>
<td>Nonnull</td>
<td>$N_{01}$</td>
<td>$N_{11}$</td>
<td>$m_1$</td>
</tr>
<tr>
<td>Total</td>
<td>$S$</td>
<td>$R$</td>
<td>$m$</td>
</tr>
</tbody>
</table>
Different ways of addressing Type I error

- Per comparison error rate (PCER) =
  \[ E\left(\frac{N_{10}}{m}\right) = Pr(\frac{N_{10,i}}{m} > 0) \leq \alpha, \ 1 \leq i \leq m \]
- Familywise error rate (FWER) = \[ Pr(N_{10} > 0) = 1 - (1 - \alpha)^m \]
- False-discovery rate (FDR) = \[ E\left(\frac{N_{10}}{R} \mid R > 0\right) Pr(R > 0) \] and its variants...
Some Notation

- \( \alpha = \) prespecified level, \( 0 < \alpha < 1 \)
- \( P_i = \) p-value for testing \( H_{0,i} \) vs. \( H_{1,i} \), \( 1 \leq i \leq m \)
Classical Procedures to Type I Error Control

- **FWER**
  - **Bonferroni**
    - single step procedure
    - Reject $H_{0,i}$ if $P_i \leq \alpha/m$
    - guarantees $FWER \leq \alpha$
  - **Holm (1979)**
    - stepwise procedure
    - Order p-values: $0 < P_1 \leq P_2 \leq \ldots \leq P_m$
    - Reject $H_{0,i}$ if $P_{(i)} \leq \alpha/(m - i + 1)$
    - more powerful than Bonferroni

Discussion: Sun & Cai, JASA 2007
Classical Procedures to Type I Error Control

- FDR
    - For $m$ independent tests, of which $m_0$ are true, order p-values in decreasing order: $P_m \geq P_{m-1} \geq \ldots \geq P_1 > 0$
    - For a fixed $\alpha \in (0, 1)$, find $k = \max_i \{ P_i \leq m_0 m \alpha \}$
    - Reject $H_{0,i}$ for all $i \leq k$
    - Guarantees $FDR \leq \alpha$
    - More powerful than FWER control procedures (greater number of rejections)
    - Equivalent to $FWER$-control when all $m$ null hypotheses are true ($m_0 = m$)
    - Works well for sparse situations ($m_0 \approx m$)

Discussion: Sun & Cai, JASA 2007
Additional Notation

- Data: \( X_i | \theta_i \sim (1 - \theta_i)F_0 + \theta_i F_1 \)
- Target of inference: \( \theta_1, \ldots, \theta_m \) \( \overset{\text{indep}}{\sim} \) Bern\((p)\)
  - null: \( \theta_i = 0 \), nonnull: \( \theta_i = 1 \)
- Test statistic: \( T(x) = [T_i(x): 1 \leq i \leq m] \)
- Z-values, \( Z_i = \Phi^{-1}(F(T_i)) \) and corresponding p-values, \( P_i \), \( 1 \leq i \leq m \)
- Decision rule: \( \delta = (\delta_1, \ldots, \delta_m) \in I = \{0, 1\}^m \)
- \( mFDR = E(N_{01})/E(R) \), \( mFNR = E(N_{01})/E(S) \)
- \( p = Pr(\theta_i = 1) \) \( 1 \leq i \leq m \)

John Pura BIOS900

Discussion: Sun & Cai, JASA 2007
Compound decision view of multiple testing problem:

- Find $\delta$ that minimizes mFNR (efficient) while controlling mFDR at level $\alpha$ (valid) among all SMLR decision rules.

- Procedures based on thresholding ordered p-values are not generally efficient (e.g. $\delta(p(\cdot), c) = \{I[p(X_i) < c] : 1 \leq i \leq m\}$)

- **Remark 1.** Copas (1974): if a symmetric (aka permutation-invariant) classification rule for dichotomies is admissible in a weighted classification problem, then it must be ordered by the likelihood ratios.
Sun and Cai’s Approach

• Spoiler Alert!
  • Optimal solution to multiple testing problem equivalent to optimal solution to weighted classification problem among symmetric/permutation-invariant decision rules
  • The local FDR (Lfdr) is a fundamental quantity which can be used directly for optimal FDR control.
  • An adaptive Lfdr procedure based on z-values unifies the goals of global error control and individual case interpretation
Suppose we have \( m \) tests of \( H_{0,i} : \mu = 0 \) vs. \( H_{1,i} : \mu \neq 0 \), \( 1 \leq i \leq m \) with corresponding \( z \)-values \( z_1, \ldots, z_m \sim f(z) \),
\[
f(z) = p_0 \phi(z) + p_1 \phi(z - \mu_1) + p_2 \phi(z - \mu_2)
\]
\( p_0 = 0.8 \), \( p_1 + p_2 = 0.2 \), \( mFDR = 0.1 \)

Compare the oracle procedure (more on this in a bit) based on \( p \)-value (Genovese & Wasserman, 2002) to one based on \( z \)-value
Oracle Setting - Example 1

Discussion: Sun & Cai, JASA 2007
Remarks on Example 1

- p-value oracle procedure (Disclaimer: for two-sided alternatives) is dominated by z-value oracle procedure, with greater difference in performance as $p_1 > p_2$.
- The Lfdr oracle procedure based on z-values may accept more extreme (i.e. larger absolute z-value or equivalently lower p-value) observations while rejecting less extreme ones $\implies$ rejection region is asymmetric
  - $z = -2$ (Lfdr = 0.227, p-value = 0.046) is rejected, but $z = 3$ (Lfdr = 0.543, p-value = 0.003) is accepted
- This is due to the inherent properties of the Lfdr statistic

Discussion: Sun & Cai, JASA 2007
Local FDR (Lfdr)

- Inherently Bayesian idea
- \( Lfdr = Pr(null|z) = (1 - p)f_0(z)/f(z) \)
- Allows for *local* inference, in which outcomes such as \( z = 3 \) are "judged on their own terms and not with respect to hypothetical possibility of more extreme results" (Efron, 2010)
  - In other words, our interest in an outlying \( z_i \) is judged relative to the local feature \( f_0/f \) about \( z_i \), rather than relative to features of the theoretical null distribution. See *Remark 4*. 

What is the optimal procedure and how does it behave if we know $p$, $f_0$, and $f_1$?

Assume SMLR property for a given test statistic. See Proposition 1. Why is this useful?

The optimal solution to the classification problem is also optimal in the multiple testing problem. See Theorem 1.

Oracle rule:

$$
\delta^\lambda(\Lambda, 1/\lambda) = \left[ I\{\Lambda(x_i) < 1/\lambda\}, 1 \leq i \leq m \right] = \frac{L_{fdr}(\cdot)}{1-L_{fdr}(\cdot)}, \text{ where } \Lambda(\cdot) = \frac{(1-p)f_0(\cdot)}{pf_1(\cdot)}
$$

Bayes rule that minimizes the classification risk (always symmetric, admissible, and ordered)
Oracle Procedure based on z-values

- Test statistic: \( T_{OR}(Z_i) = (1 - p)f_0(Z_i)/f(Z_i) \sim G_{OR} \), where \( G_{OR} = (1 - p)G^0_{OR} + pG^1_{OR} \)
- mFDR = \( (1 - p)G^0_{OR}/G_{OR} \)
- Threshold: \( \lambda_{OR} = \sup_{t \in (0, 1)} \{ Q_{OR}(t) \leq \alpha \} \)
- Testing rule: \( \delta(T_{OR}, \lambda_{OR}) = [I\{ T_{OR}(Z_i) < \lambda_{OR} \}, 1 \leq i \leq m] \)
- The Lfdr oracle procedure ranks the relative importance of the test statistics according to their likelihood ratios
  - Rankings generally different from p-values, unless alternative distribution is symmetric about null hypothesis
Another Example

Discussion: Sun & Cai, JASA 2007
Remarks - Example 2

- When $|\mu_1| = |\mu_2|$, mFNR is constant for p-value oracle procedure; equivalence only under symmetric alternative (a)
- Lfdr oracle procedure utilizes the distributional information of the nonnulls (i.e. adapts to asymmetry), but the p-value oracle procedure does not (a)
- Under large degree of asymmetry in alternative distribution, oracle procedures differ the greatest when $|\mu_1| < \mu_2$ and $p_1 > p_2$ (b,c)
- mFNR decreases with mFDR for a given range (d)
Adaptive Procedure

Suppose we have consistent estimators $\hat{p}_0$, $\hat{f}_0$, $\hat{f}$ (e.g. Jin & Cai, 2007)

Define $\hat{T}_{OR}(z_i) = \min([((1 - \hat{p})\hat{f}_0(z_i)/\hat{f}(z_i)], 1)$

Order the $\hat{T}_{OR}(z_i)$ as $L\hat{fdr}(1) \leq \ldots \leq L\hat{fdr}(m)$

Let $k = \max_{1 \leq i \leq m} \{ \frac{1}{i} \sum_{j=1}^{i} L\hat{fdr}(j) \leq \alpha \}$

Reject $H_{0,(i)}$ for all $i \leq k \iff \delta(\hat{T}_{OR}, \hat{\lambda}_{OR}) = [I\{\hat{T}_{OR}(z_i) < \hat{\lambda}_{OR}\}]$

Asymptotically efficient: asymptotically controls mFDR at level $\alpha$ and attains the mFNR level achieved by the Lfdr oracle procedure. See Theorems 4 & 5.
Yet, More Examples

Discussion: Sun & Cai, JASA 2007
Conclusion

- The adaptive Lfdr procedure based on z-values is asymptotically optimal, particularly under asymmetric alternative distribution.
- The Lfdr can be viewed as ranking the relative importance of the LR, allowing for local inference and efficient separation of nulls and nonnulls.