False discovery proportion estimation by permutations: confidence for SAM

Jesse Hemerik, Jelle Goeman

Leiden University Medical Center, The Netherlands

Adaptive Designs and Multiple Testing Procedures Workshop 2016

LEIDEN UNIVERSITY MEDICAL CENTER

Main message

- SAM ("Significance Analysis of Microarrays") is a useful method for FDP estimation
- First paper about SAM (2001) cited 10,000 times
- SAM is only heuristic
- We provide confidence bound for FDP
- We use closed testing to decrease the bound

FDP

We test hypotheses $H_1, \ldots, H_m$

$$\mathcal{R} := \{1 \leq i \leq m : \ H_i \text{ is rejected}\}$$
$$\mathcal{N} := \{1 \leq i \leq m : \ H_i \text{ is true}\}$$

$V := \#(\mathcal{N} \cap \mathcal{R})$ number of false positives

$R := \#\mathcal{R}$

$$\boxed{FDP := \frac{V}{R}}$$

Setting of SAM

- Hypotheses $H_1, \ldots, H_m$
- Data X with any distribution
- Test statistics $T_1(X), \ldots, T_m(X)$
- G a finite *group* of transformations from and to the range of X
- Joint distr. of the $T_i(gX)$ with $i \in \mathcal{N}$, $g \in G$, is invariant under all transformations in G of the data X.

Output of SAM

1. User chooses a rejection region $D \subset \mathbb{R}$
2. SAM rejects the H_i with $T_i \in D$ and provides $\widehat{FDP}$

SAM's calculation of $\widehat{FDP}$

1. $R = R(X) = \#\{1 \leq i \leq m : \; T_i(X) \in D\}$
2. For each permutation g_j, calculate
   $R(g_jX) = \#\{1 \leq i \leq m : \; T_i(g_jX) \in D\}$
3. $\widehat{V} :=$ median of the values $R(g_jX), \; 1 \leq j \leq w$
4. $\widehat{FDP} := \frac{\widehat{V}}{R}$
5. $\widehat{FDP}' := \widehat{FDP} \cdot \widehat{\pi}_0 \qquad\qquad (\pi_0 = \frac{\#\mathcal{N}}{m})$

Part 2: our results

Results on $\widehat{FDP}$

Proven: $\widehat{FDP}$ is a *median controlling* estimator of FDP, i.e:

$$P(FDP \leq \widehat{FDP}) \geq \frac{1}{2}.$$

$\widehat{FDP}' = \widehat{FDP} \cdot \widehat{\pi}_0$ has unknown properties

Generalization

Choose:

- for each T_i any rejection region $D_i \subset \mathbb{R}$
- some $\alpha \in [0, 1]$

We provide:

a $(1 - \alpha)100\%$-confidence upper bound $\overline{FDP}$ for the FDP:

$$P(FDP \leq \overline{FDP}) \geq 1 - \alpha$$

Calculation of upper bound

The $(1-\alpha)100\%$-confidence upper bound is

$$\overline{FDP} := \frac{\overline{V}}{R},$$

where $\overline{V}$ is the $(1-\alpha)$-quantile of the values $R(g_j X),\ 1 \leq j \leq w$

Recall permutation test:

- Consider:
  - data X with any distribution
  - a group G of transformations from and to the range of X
  - a test statistic $T(X)$
- $\boxed{H_0\text{: } X \overset{d}{=} gX \text{ for all } g \in G.}$
- Let $T^{1-\alpha}$ be the $(1-\alpha)$-quantile of the values $T(gX)$, $g \in G$.
- Then under H_0, $P(T(X) > T^{1-\alpha}) \leq \alpha$.

Proof upper bound

First note:
$\mathcal{R}$, and V depend on the data. Write $\mathcal{R}(x)$, $V(x)$.
$\mathcal{N}$ does not depend on the data. Thus $V(x) = \#(\mathcal{R}(x) \cap \mathcal{N})$.

To show: $P(V > \overline{V}) \leq \alpha$.
Proof: Let $V^{1-\alpha}$ be the $(1-\alpha)$-quantile of the values

$$V(g_j X), \quad 1 \leq j \leq w.$$

By permutation principle:

$$P(V(X) > V^{1-\alpha}) \leq \alpha.$$

Finally note that $V^{1-\alpha} \leq \overline{V}$. $\square$

Data analysis

- Same data as in SAM paper (Tusher et al 2001)
- $\sim$ 7000 hypotheses $H_1, \ldots, H_m$
- H_i: Expression rate of gene i same for irradiated and unirradiated cells

Δ	R	$\widehat{FDP}$	$\overline{FDP}$ ($\alpha = 0.1$)
0.5	191	0.30	0.99
0.6	162	0.25	0.98
0.9	80	0.13	0.88
1.2	46	0.09	0.67
1.8	26	0.08	0.46
2.5	12	0.08	0.42
3	10	0.10	0.30
3.5	3	0	0.33

Conservativeness

When there are many false hypotheses, $\widehat{FDP}$ is conservative

SAM software therefore uses $\widehat{FDP}' := \widehat{FDP} \cdot \widehat{\pi}_0$

Unknown properties. No confidence

We want to decrease the bound without losing the property
$P(FDP \leq \overline{FDP}) \geq 1 - \alpha$

Part 3: Closed testing for improved bounds

General definition closed testing

Goal:
Want to test each intersection hypothesis $H_{\mathcal{I}} = \bigcap_{i \in \mathcal{I}} H_i$, $\mathcal{I} \subseteq \{1, \ldots, m\}$ such that $P(\text{no false positives}) \geq 1 - \alpha$

Closed testing:
For each $H_{\mathcal{I}}$, define a test of level α. (So $2^m - 1$ *local tests*)

C.t.p. rejects all $H_{\mathcal{I}}$ for which every $H_{\mathcal{J}}$ with $\mathcal{J} \supseteq \mathcal{I}$ is rejected by its local test

Deriving upper bounds using c.t.p.

For each $K \subseteq \{1, \ldots, m\}$ define

$$\overline{V}_{\mathrm{ct}}(\mathcal{K}) = \max\{\#\mathcal{I} : \ \mathcal{I} \subseteq \mathcal{K}, \ H_{\mathcal{I}} \text{ not rejected by c.t.p.}\}$$

By Goeman and Solari (2011):

$$\boxed{P\left[\bigcap_{\mathcal{K}\subseteq\{1,\ldots,m\}} \{\#\mathcal{K} \cap \mathcal{N} \leq \overline{V}_{\mathrm{ct}}(\mathcal{K})\}\right] \geq 1 - \alpha}$$

Our c.t.p.

In the SAM context, recall

$$\mathcal{R}(X) = \{1 \leq i \leq m : T_i(X) \in D_i\}.$$

For each $H_{\mathcal{I}}$ consider local test that rejects iff

$$\#\mathcal{I} \cap \mathcal{R}(X) > R_{\mathcal{I}}^{1-\alpha},$$

where $R_{\mathcal{I}}^{1-\alpha}$ is the $(1-\alpha)$-quantile of the values $\#\mathcal{I} \cap \mathcal{R}(g_j X)$, $1 \leq j \leq w$.

we consider the c.t.p. based on these local tests.

Upper bound based on c.t.p.

By Goeman and Solari,

$$\overline{V}_{\text{ct}} := \overline{V}_{\text{ct}}(\mathcal{R}) = \max\{\#\mathcal{I} : \mathcal{I} \subseteq \mathcal{R}, H_{\mathcal{I}} \text{ is not rejected by c.t.p.}\}$$

is a $1-\alpha$-upper bound for $\#\mathcal{R} \cap \mathcal{N} = V$.

In theory this bound is ideal.

Problem: naively calculating $\overline{V}_{\text{ct}}$ is often infeasible. Indeed, to check if $H_{\mathcal{I}}$ is rejected by c.t.p., requires to check if all $H_{\mathcal{J}}$ with $\mathcal{I} \subseteq \mathcal{J}$ are rejected...

Shortcut

The bound $\overline{V}_{\text{ct}}$ equals

$$\min\Bigg[R,\ \min\left\{1 \leq M \leq R :\ \text{for all } \mathcal{I} \subseteq \mathcal{R} \text{ with } \#\mathcal{I} = M,\ M > R^{1-\alpha}_{\mathcal{I} \cup R^c}\right\} - 1\Bigg]$$

Using this shortcut, we can often calculate $\overline{V}_{\text{ct}}$ when there are many hypotheses.

When $\binom{R}{\overline{V}_{\text{ct}}}$ is large, calculating $\overline{V}_{\text{ct}}$ is infeasible.
$\rightarrow$ Conservative shortcut

Simulations

$m = 100,\ D = (0, 0.01)$ and $\alpha = 0.1$

π_0	Correlation	$\mathbb{E}(R)$	$\mathbb{E}(\overline{V}/R)$	$\mathbb{E}(\overline{V}_{\mathrm{CT}}/R)$
0.9	no	8.8 ± 0.1	0.35 ± 0.01	0.33 ± 0.01
0.9	yes	7.6 ± 0.2	0.46 ± 0.01	0.45 ± 0.01
0.7	no	24.6 ± 0.5	0.18 ± 0.01	0.12 ± 0.01
0.7	yes	20.8 ± 1.1	0.23 ± 0.01	0.18 ± 0.02
0.5	no	40.0 ± 0.6	0.16 ± 0.01	0.07 ± 0.00
0.5	yes	34.1 ± 1.8	0.18 ± 0.01	0.11 ± 0.01

Conclusion

- Until now SAM was only heuristic
- We have extended SAM with a CI for the FDP
- Using closed testing, we have decreased the estimate and upper bound