False discovery proportion control by permutations

Proving properties of SAM

Jesse Hemerik, Jelle Goeman

Radboud university medical center, The Netherlands

80 Years After Bonferroni

Radboudumc

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 exact conf. statements about FDP

FDP

We test hypotheses $H_1, ..., H_m$

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

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

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

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

Part 2: our results

Results on $\widehat{FDP}$

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

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

$\widehat{FDP}' = \widehat{FDP} \cdot \widehat{\pi}_0$ is not

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 $\#R(g_j X),\ 1 \le j \le w$

Recall permutation test:

- Consider:
  - data X with any distribution
  - a group G of transformations from and to the range of X
  - a test statistic $T(X)$
- $\boxed{H_0\colon\ X \overset{d}{=} gX \text{ for all } g \in G.}$
- Let

$$T^{(1)} \leq \ldots \leq T^{(\#G)}$$

  be the sorted values $T(gX)$, $g \in G$.
- Then $P(T(X) > T^{(\lceil (1-\alpha)\cdot \# G \rceil)}) \leq \alpha$.

Proof upper bound

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

$$\#\mathcal{N} \cap R(g_j X), \quad 1 \leq j \leq w.$$

By permutation principle:

$$P(\#\mathcal{N} \cap R(X) > V^{1-\alpha}) \leq \alpha.$$

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

Conservativeness (1)

- By permutation principle the $(1-\alpha)$-quantile of the values
$$\#\mathcal{N} \cap R(g_j X), \quad 1 \leq j \leq w,$$
is a $(1-\alpha)$-upper bound for V.

- But we don't know $\mathcal{N}$, so use the $(1-\alpha)$-quantile of the values
$$\#R(g_j X), \quad 1 \leq j \leq w.$$

- So real error rate can be much smaller than α.

Conservativeness (2)

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

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

Unknown properties. It's not median-unbiased

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

Better upper bound

Let E be the event that $V \leq V^{1-\alpha}$.
Thus $P(E) \geq 1 - \alpha$.

Suppose E holds. Thus $V \leq V^{1-\alpha} \leq \overline{V}$. So among R there are no more than $\overline{V}$ true hypotheses

Use this information to find better bound $\overline{V}^1$
Continue like this, finding $\overline{V}^1 \geq \overline{V}^2 \geq \overline{V}^3 \ldots$

Improved upper bound $= \min_i \overline{V}^i$

Part 3: Relation to closed testing

SAM bound $\geq$

Bound of iterative method $\min_i \overline{V}^i \geq$

Bound derived from closed testing procedure

General definition closed testing

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

For each H_I, define a test of level α. (So $2^m - 1$ *local tests*)

C.t.procedure rejects all H_I with property that all H_J with $J \supseteq I$ are rejected by their local tests

Deriving upper bounds using c.t.p.

Write $\mathcal{X} = \{I \subseteq \{1, ..., m\} : \quad H_I \text{ rejected by c.t.p.}\}$

Let $K \subseteq \{1, ..., m\}$ be any set.

By Goeman and Solari (2011):

An upper bound to $\#\mathcal{N} \cap K$ is

$$\max\{\#I : I \subseteq K, I \notin \mathcal{X}\} \vee 0.$$

With probability $\geq 1 - \alpha$ these bounds are valid uniformly over all $K \subseteq \{1, ..., m\}$.

Our c.t.p.

In the SAM context, recall

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

For each H_I consider local test that rejects iff

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

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

Connection to our iterative method

Consider the c.t.p. based on these local tests.
Write $R := R(X)$

Upper bound for $V = R \cap \mathcal{N}$ is

$$\max\{\#I : I \subseteq R \text{ and } I \notin \mathcal{X}\}$$
$$= \ldots = \ldots \leq \ldots = \overline{V}^1.$$

Using $\overline{V}^1$, by analogous argument $\overline{V}^2$ follows, etc.

Uniform bounds

For every $K \subseteq \{1, ..., m\}$ a (uniform) bound for $\#K \cap \mathcal{N}$ is

$$\max\{\#I : I \subseteq K \text{ and } I \notin \mathcal{X}\} = ... = ... \leq ... = ... =$$
$$\min\{\#K, \#K \cap R^c + R^{(1-\alpha)}_{K \cup R^c}\} =: \overline{V}(K)$$

Relation to iterative method

- An upper bound to $R \cap \mathcal{N}$ is

$$\max\{\overline{V}(K) : K \subseteq R, \#K = \overline{V}(R)\} =$$
$$\max\{\min\{\#K, R^{(1-\alpha)}_{K \cup R^c}\} : K \subseteq R, \#K = \overline{V}(R)\}.$$

  But this is exactly $\overline{V}^1$. Analogously $\overline{V}^2, \overline{V}^3, \ldots$ follow

- Likewise, for every $I \subseteq \{1, \ldots, m\}$ we can improve $\overline{V}(I)$

Computational feasibility

- SAM bound $\geq$
  Bound of iterative method $\min_i \overline{V}^i \quad \geq$
  Bound from c.t.p.
- Iterative method faster than using c.t.p.
- But still computationally intensive
- $\rightarrow$ Shortcut

Use of random permutations

Suppose we want to use only w permutations from G

Drawing with replacement: Take $g_1 := id$. Draw $g_2, ..., g_w$ with replacement from G

Drawing without replacement: Take $g_1 := id$. Draw $g_2, ..., g_w$ without replacement from $G \setminus \{id\}$

Conclusion

- Until now SAM was only heuristic
- We have proven properties of SAM and extended it to give confidence statements about the FDP
- We have improved SAM without losing coverage

References

First SAM paper:

Tusher, V.G., Tibshirani, R. and Chu, G. (2001). Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences* **98** 5116-5121.

Rationale behind $\widehat{\pi}_0$:

Storey, J.D. et al. (2004). Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach. *JRSS: Series B (Statistical Methodology)* **66** 187-205.

Details about ($\widehat{\pi}_0$ as used in) SAM *R* package *samr*:

Chu, G. et al. Significance Analysis of Microarrays: users guide and technical document.

Deriving FDP upper bounds using closed testing:

Goeman, J. J. and Solari, A. (2011). Multiple testing for exploratory research. *Statistical Science* **26** 584-597