

Multiple testing for high dimensional biological data

Katie Pollard & Mark van der Laan

Division of Biostatistics, U.C. Berkeley

www.stat.berkeley.edu/~laan

www.bepress.com/ucbbiostat/paper121

Data and Null Hypotheses

Data: X_1, \dots, X_n i.i.d. observations of a p -dimensional vector $X \sim P$

- gene expression measurements
- gene expression plus outcomes or covariates (e.g.: survival)
- SNPs and an outcome (e.g.: response to treatment)
- occurrence of sequence motifs

Parameters: μ_1, \dots, μ_p s.t. $\mu_j(P) \in \mathcal{R}$.

- location parameters (means, medians, differences in means)
- regression parameters (association between gene j 's expression and outcome)

Null Hypotheses:

$$H_{0,j} : \mu_j(P) = \mu_j^0, j = 1, \dots, p,$$

where μ_j^0 are hypothesized null values, frequently zero.

Test Statistics

Test $H_{0,j}, j = 1, \dots, p$, with

$$D_{jn} = \mu_{jn} - \mu_j^0$$

or $T_{jn} = (\mu_{jn} - \mu_j^0) / sd(\mu_{jn})$.

If μ_{jn} is an **asymptotically linear** estimator of μ_j , that is,

$$\sqrt{n}(\mu_{jn} - \mu_j) = \frac{1}{n} \sum_{i=1}^n IC_j(X_i|P) + op(1), \quad (1)$$

for $j = 1, \dots, p$ then we know that as $n \rightarrow \infty$

$$Z_n \equiv \sqrt{n}(\mu_n - \mu(P)) \stackrel{D}{\Rightarrow} N(0, \Sigma(P)), \quad (2)$$

where $\Sigma(P) = E(IC(X)IC(X)^\top)$ is the covariance of the vector **influence curve** $IC(X) = \{IC(X)_j : j = 1, \dots, p\}$ of μ_n . Let

$$Z \sim Q_0(P) \equiv N(0, \Sigma(P)) \quad (3)$$

denote the limit (in distribution) of Z_n .

Error Rates

Given a vector $c \in \mathbb{R}^p$, consider a corresponding **multiple testing procedure** $MT(c)$ defined by:

$$\text{Reject } H_{0,j}, \text{ if } |T_{jn}| > c_j, j = 1, \dots, p. \quad (4)$$

Let:

- $V_n = V_n(c, T_n) = \sum_{j=1}^p I(|T_{jn}| > c_j, \mu_j(P) = \mu_j^0)$ be the number of false positives of $MT(c)$,
- $\theta(F) \in (0, 1) \in \mathfrak{R}$ be a particular type-I-error rate, and
- k be a user supplied constant.

Error rates $\theta(F_{V_n})$ which are functions of the distribution of V_n :

- $\int x dF_{V_n}(x)/p = E(V_n)/p$: per-comparison error rate (PCER),
- $\int x dF_{V_n}(x) = E(V_n)$: per-family error rate (PFER),
- $1 - F_{V_n}(k - 1) = Pr(V_n \geq k)$: family-wise error rate (FWER),

Cut-off Rule and Error Control

Let α be a target error rate and let $c = c(Q, \alpha) \in \mathbb{R}^p$ denote a vector function cut-off rule such that if $T_n \sim Q$, then $MT(c)$ has the property that $\theta(F_{V_n}) = \alpha$. Let $Q_n(P)$ denote the distribution of T_n based on X_1, \dots, X_n when $X \sim P$.

Weak Control: $\theta(F_{V_n}) = \alpha$ under a *particular* null distribution Q_{0n} .

Strong Control: $\theta(F_{V_n}) = \alpha$ under $Q_n(P)$ (i.e.: the truth).

Asymptotic Strong Control (ASC): $\theta(F_{V_n}) = \alpha_n$ under $Q_n(P)$ and

$$\alpha_n \rightarrow \alpha \text{ as } n \rightarrow \infty.$$

So, $MT(c)$ depends critically on the choice of distribution under which the error rate is controlled.

Null Distributions

Seek to control the error rate under a **test statistic distribution** that satisfies the null hypotheses and is as close as possible to the true test statistic distribution $Q_n(P)$. The correct null distribution is the **projection** of $Q_n(P)$ onto the space of mean zero distributions.

Results:

1. $Q_0 \equiv N(0, \Sigma(P))$ is the asymptotically correct null distribution for the vector of test statistics $\sqrt{n}(\mu_n - \mu^0)$.
2. If $c_0 \equiv c(Q_0, \alpha)$ then $MT(c_0)$ has ASC.

Estimate Q_0 by a choice of Q_{0n} , e.g.:

- $\tilde{Q}_{0n} = N(0, \Sigma_n)$
- Bootstrap $Q_{0n}^\#$
- $Q_n(P_{0n})$, where P_{0n} is an estimated data null distribution

Bootstrap Estimated Null Distribution

Let

- \tilde{P}_n be an estimator of P
- $\tilde{\mu}_n = \mu(\tilde{P}_n)$ be the parameter estimate under \tilde{P}_n
- $\mu_n^\#$ be μ_n applied to n i.i.d. copies $X_1^\#, \dots, X_n^\#$ of $X^\# \sim \tilde{P}_n$
- $Q_{0n}^\#$ be the distribution of $Z_n^\# = \sqrt{n}(\mu_n^\# - \tilde{\mu}_n)$

Estimate Q_0 with $Q_{0n}^\#$. Under weak regularity conditions, it is known that $Z_n^\# \xrightarrow{D} Z \sim Q_0$ conditional on \tilde{P}_n , and hence $Q_{0n}^\#$ converges to Q_0 conditional on the data.

Define

$$R_n(c, Z_n^\#) \equiv \sum_{j=1}^p I(|Z_{jn}^\#| > c_j).$$

Let c_n be a solution of $\theta \left(F_{R_n(c, Z_n^\#)} \right) = \alpha$. Then, $MT(c_n)$ is a **bootstrap based multiple testing procedure** controlling θ at level α .

Asymptotic Strong Control (ASC)

Theorem: Let $c_{0n} \equiv c(Q_{0n}, \alpha)$ and suppose $c_{0n} \xrightarrow{P} c_0$ for $n \rightarrow \infty$.
Then

$$\limsup_{n \rightarrow \infty} \theta \left(F_{V_n(c_{0n})} \right) \leq \alpha. \quad (5)$$

If the mapping $Q \rightarrow c(Q, \alpha)$ is continuous, then $c(Q_{0n}, \alpha) \rightarrow c(Q_0, \alpha)$ whenever $Q_{0n} \Rightarrow Q_0$. Hence,

- $\tilde{Q}_{0n} = N(0, \Sigma_n)$ provides ASC if $\Sigma_n \rightarrow \Sigma(P)$ as $n \rightarrow \infty$,
- the bootstrap $Q_{0n}^\#$ method provides ASC,
- $Q_n(P_0)$ does **not** provide ASC unless

$$\Sigma(P_0) = \Sigma(P). \quad (6)$$

Condition (6) is the formal analogue of the **subset pivotality condition** (Westfall and Young, 1993, p.42-43).

Equivalence of Multiple Testing and Confidence Regions

Let F_n denote the distribution of $R_n(c, Z_n) = \sum_{j=1}^p I(|Z_{jn}| > c_j)$, where $Z_n = \sqrt{n}(\mu_n - \mu(P))$ and c is s.t. $\theta(F_n) = \alpha$. Then,

$$\{\mu : \sqrt{n}(\mu_n - \mu) < c\} \quad (7)$$

is a θ -specific $(1 - \alpha)\%$ confidence region for $\mu(P)$.

- If $\theta(\cdot)$ is the FWER, then the region defined by (7) is a $(1 - \alpha)\%$ *simultaneous* confidence region for $\mu(P)$.
- We can estimate F_n with the distribution $F_n^\#$ of $R_n(c, Z_n^\#)$, where $Z_n^\#$ is the bootstrap random variable $\sqrt{n}(\mu_n^\# - \mu_n)$.
- Let \tilde{c} be the solution of $\theta(F_n^\#) = \alpha$. Then,
 1. $\{\mu : \sqrt{n}(\mu_n - \mu) < \tilde{c}\}$ is an *asymptotically* correct θ -specific $(1 - \alpha)\%$ confidence region for $\mu(P)$.
 2. The equivalent multiple testing procedure is $MT(\tilde{c})$:

$$\text{Reject } H_{0,j} \text{ if } |\sqrt{n}(\mu_{jn} - \mu_j^0)| > \tilde{c}_j.$$

Example: Two sample problem

Suppose we have n_1 observations from Population 1 with mean μ_1 and n_2 observations from Population 2 with mean μ_2 . For example,

- tumor classes
- treatment regimes
- time points

For each gene, test the **null hypothesis** of no differential expression:

$$H_{0,j} : \mu_j = \mu_{2,j} - \mu_{1,j} = 0, j = 1, \dots, p.$$

Use **test statistics** such as:

$$D_{jn} = \bar{X}_{2,j} - \bar{X}_{1,j}, j = 1, \dots, p$$

or $T_{jn} = \frac{\bar{X}_{2,j} - \bar{X}_{1,j}}{\sqrt{\hat{\sigma}_{1,j}^2/n_1 + \hat{\sigma}_{2,j}^2/n_2}}, j = 1, \dots, p.$

Comparison of Null Distributions: Variance and Covariance

Let $COV(X_j, X_{j'})$ be ϕ_1 in population 1 and ϕ_2 in population 2.

Distribution	$Var(D_{jn})$	$Cov(D_{jn}, D_{j'n})$
Permutations	$\frac{\sigma_{1,j}^2}{n_2} + \frac{\sigma_{2,j}^2}{n_1}$	$\frac{\phi_1}{n_2} + \frac{\phi_2}{n_1}$
Bootstrap	$\frac{\sigma_{1,j}^2}{n_1} + \frac{\sigma_{2,j}^2}{n_2}$	$\frac{\phi_1}{n_1} + \frac{\phi_2}{n_2}$

Note:

- $VAR(T_{jn}) = 1$ for both distributions.
- But $COV(T_{jn}, T_{j'n})$ is not equivalent unless $n_1 = n_2$.

Conclusions

1. $Q_0 = N(0, \Sigma(P))$ is the asymptotically correct null distribution for the test statistics $\sqrt{n}(\mu_n - \mu^0)$ and provides ASC of type I error rates that are functions of the distribution of the number of false positives (e.g.: FWER).
2. For a finite sample, Q_0 can be estimated with a simple bootstrap method that provides ASC under weak conditions for any P .
3. Common practice of estimating Q_0 via a data null distribution P_0 only provides ASC when $\Sigma(P_0) = \Sigma(P)$.
4. We have provided a multivariate generalization of the equivalence between hypothesis testing and CIs.
5. Two Sample Problem: Permutation data null distribution P_{0n}
 - has the wrong covariance unless $n_1 = n_2$ or $\Sigma_1 = \Sigma_2$
 - hence, often fails to control the error rate
 - may be appropriate in some cases, e.g.: balanced designs