Normal Theory Multiple Test Procedures

Jinxi Liu

October 25, 2017

Preliminaries

Consider testing the set of hypotheses $H_i : \theta_i \leq 0$ versus $A_i : \theta_i > 0$ for $i = 1, 2, ..., k$. Assume that unbiased estimators $\hat{\theta}_1, ..., \hat{\theta}_k$ are available, each based on a sample of size N, with a multivariate normal distribution, $var(\hat{\theta}_i) = \tau^2\sigma^2$ and $corr(\hat{\theta}_i, \hat{\theta}_j) = \rho$, where τ^2 and ρ are known constants and σ^2 is an unknown error variance. Let S^2 be an unbiased estimator of σ^2 having v degrees of freedom such that vS^2/σ^2 has a χ^2_v distribution independent of the $\hat{\theta}_i$.

Preliminaries

Let $t_i = \hat{\theta}_i/(s\tau)$, where s is the observed value of S. Then under H_i for $i = 1, ..., k$, $t_1, ..., t_k$ are observations from k-variate central t statistics, $T_1, ..., T_k$, with v df and common correlation τ. Without loss of generality, assume that $m(0 \leq m \leq k)$ null hypotheses are true. Further suppose that the hypotheses have been relabeled if necessary so that $T_1, ..., T_m$ correspond to the true null hypotheses. If the statistics $T_1, ..., T_k$ are consistent for the tests of $H_1, ..., H_k$

T_i consistent test statistic for the test of H_i versus A_i; $i = 1, ..., k$; then asymptotically$(N \to \infty)$ we will have the following ordering:

Preliminaries

$T_{(1)} \leq T_{(2)} \leq ... T_{(m)} \leq T_{(m+1)} \leq ... T_{(k)}$ corresponding to $H_{(1)} \leq H_{(2)} \leq ... H_{(m)} \leq H_{(m+1)} \leq ... H_{(k)}$, where $H_{(1)}, ..., H_{(m)}$ are true null hypotheses and $H_{(m+1)}, ..., H_{(k)}$ are false null hypotheses.

Denote $P_{r:s}(\cdot)$, $r \leq s$, to be the probability under any parameter configuration with $H_{(1)}, ..., H_{(r)}$ are true and $H_{(r+1)}, ..., H_{(s)}$ are false.

Step-down procedure

The normal theory step-down FDR controlling procedure (NSD) will be determined upon specification of the critical constants $c_1, ..., c_k$. This section describes how the critical constants are computed for one-sided tests.

Step-down procedure

$$FDR = E\left(\frac{V}{R}\right) = \sum_{r=1}^{k} \frac{1}{r} E(v|R=r) P_{m:k}(R=r)$$

$$= \sum_{r=1}^{k} \frac{1}{r} E(V|R=r) \times \tag{1}$$

$$P_{m:k}(T_{(k)} \geq c_k, T_{(k-1)} \geq c_{k-1}, \ldots,$$

$$T_{(k-r+1)} \geq c_{k-r+1}, T_{(k-r)} \leq c_{k-r})$$

where the last condition $T_{(k-r)} \leq c_{k-r}$ is imposed only when $r < k$.

Step-down procedure

The above expression can be seen to be asymptotically equivalent (as $N \to \infty$) to

$$\sum_{r=k-m+1}^{k} \frac{(r-k+m)}{r} P_{m:m}(T_{(m)} \geq c_m, ..., T_{(k-r+1)} \geq c_{k-r+1}, T_{(k-r)} < c_{k-r}) \quad (2)$$

If $c_1, ..., c_k$ are chosen so that expression (2) is contained below α for any $m \in \{1, 2, ..., k\}$ then the FDR will asymptotically be held below α.

Step-down procedure

The statistics $T_1, ..., T_m$ come from the central multivariate t-distribution by assumption. Therefore, we can represent the variables as

$$T_i = \frac{\sqrt{1-\rho}Z_i - \sqrt{\rho}Z_0}{U}, i = 1, ..., k,$$

where$Z_i, i = 1, ..., k$, are independent N(0,1) variables and U is an independent $\sqrt{\chi^2/v}$ variable. Expression (2) becomes

$$\sum_{r=k-m+1}^{k} \frac{r-k+m}{r} \int_0^\infty \int_{-\infty}^\infty p(Z_{(m)}) \geq d_m, p(Z_{(m-1)}) \geq d_{m-1}, ...,$$

$$p(Z_{(k-r+1)}) \geq d_{k-r+1}, , p(Z_{(k-r)}) \leq d_{k-r})\phi(z_0)dz_0 f_v(u)du, \quad (3)$$

Step-down procedure

where

$$d_i = \frac{c_i u + \sqrt{\rho} z_0}{\sqrt{1-\rho}}, 1 \leq i \leq k,$$

$\phi(\cdot)$ is the standard normal density function, and $f_v(\cdot)$ is the density function of U.

The first critical constant is determined by setting (2) equal to α with $m = 1$,

$$\alpha = \frac{1}{k} P_{1:1}(T_{(1)} \geq c_1),$$

which has the solution $c_1 = t_v(k\alpha)$, the upper $k\alpha$ point of the t distribution with v df.

If $k\alpha$ exceeds $1/2$ then c_1 is taken to be zero by convention so as not to allow the possibility of rejecting based on a negative t value.

Step-down procedure

Given c_1, we obtain c_2 by setting (2) with m $= 2$ equal to α,

$$\alpha = \frac{1}{k-1}\int_0^\infty \int_{-\infty}^\infty P(Z_{(2)} \geq d_2, Z_{(1)} < d_1)\phi(z_0)f_v(u)dz_0du + \frac{2}{k}\int_0^\infty \int_{-\infty}^\infty P(Z_{(2)} \geq d_2, Z_{(1)} \geq d_1)\phi(z_0)f_v(u)dz_0du$$

and solving for $c_2 \in [c_1, \infty)$. The integrals are computed via numerical integration. A unique solution will exist as long as $k \leq 1/\alpha$. If$k > 1/\alpha$, then c_2 is taken to equal c_1, and the FDR is conservative

In general, c_j is obtained by considering $c_1, ..., c_{j1}$ fixed and setting (3) with $m = j$ equal to α. If a solution exists in $[c_{j-1}, 1)$,then it is c_j. Otherwise, c_j is set equal to c_{j-1}.

Step-down procedure

Simulation results:

The power of the NSD is higher than the power of the BH procedure, even in the case when correlation $= 0$.

The raise in power becomes higher as the number of false null hypotheses increases.

Extension

So far we have assumed that the estimators $\hat{\theta}_i$ share a common correlation coefficient. In practice, this is rarely the case even approximately. Here we consider calculation of the critical constants by simulation, without any assumption on the known correlation structure.

Extension

Eq. (2) is still a valid expression for the asymptotic FDR, where $T1, ..., Tk$ come from the central multivariate t distribution with correlation Λ.

Without loss of generality, suppose that the hypotheses have been ordered so that $t_1 \geq t_2 \geq ... \geq t_k$, where now the correlation is Λ'.

The first constant, c_1, may be found directly from the same equation used before

$$\alpha = \frac{1}{k} P_{1:1}(T_{(1)} \geq c_1),$$

The second constant, c_2, then is found by considering (2) with $m = 2$ equal to α.

Extension

$$\alpha = \frac{1}{k-1} P_{2:2}(Z_{(2)} \geq d_2, Z_{(1)} < d_1) + \frac{2}{k} P_{2:2}(Z_{(2)} \geq d_2, Z_{(1)} \geq d_1) \tag{4}$$

To solve (4), simulate a large number M of k-variate central t statistics $(T_1^*, ..., T_k^*)$ with correlation Λ'. For each simulation, order the t statistics, $T_1^* and T_2^*$, so that $T_{(1)}^* \leq T_{(2)}^*$.

Next assign each simulation a coefficient, either $1/(k-1)$ or $2/k$, as $T_{(1)}^* < c_1$ or $T_{(1)}^* \geq c_1$ respectively.

Then order the simulations based on the value of $T_{(2)}^*$.

Extension

Now consider the sequence of partial sums of the simulation coefficients, starting with the simulation that has the largest $T^*_{(2)}$. Find the simulation for which this partial sum is largest yet still less than or equal to $M\alpha$. The corresponding value of $T^*_{(2)}$ is then taken as c_2, assuming that this is greater than c_1.

The constants are determined recursively.

Extension

The constants are determined recursively, with the $j_t h$ constant being determined from the equation

$$\alpha = \sum_{r=k-j+1}^{k} \frac{(r-k+j)}{r} P_{j:j}(T_{(j)} \geq c_j, ..., T_{(k-r+1)} \geq c_{k-r+1}, T_{(k-r)} < c_{k-r}) \quad (5)$$

Eq. (5) is solved from the same set of M simulated k-variate t statistics that were used to determine $c_2, ..., c_{j-1}$. The statistics are ordered within each simulation, and the appropriate coefficients assigned to each simulation based on the relative size of $T^*_{(1)}, ..., T^*_{(j-1)}$ and $c_1, ..., c_{j-1}$.

Extension

Next the simulations are ordered by the value of $T^*_{(j)}$. The sequence of partial sums of the simulation coefficients are computed, starting with the simulation with the largest $T^*_{(j)}$.

Finally, c_j is taken as the value of $T^*_{(j)}$ from the simulation for which the partial sum is greatest but not more than $M\alpha$. Monotonicity is enforced so that $c_j \geq c_{j-1}$.