Conditions for validity

of re-sampling based multiple testing

with applications in genomics

Violeta Calian

Science Institute,University of Iceland

Dunhaga 3, 107 Reykjavik, Iceland

and

Jason C. Hsu

Department of Statistics, The Ohio State University

Columbus OH 43210, USA

Outline

1. Problem, motivation and our solution
2. Error rates and error control
3. The MCP principle
4. Models. Test statistics. Critical values
5. Theoretical results on re-sampling distributions. Examples
6. Conclusions

1 Problem, motivation and our solution

1.1 The problem

Inference for high-dimensional (d) multivariate distributions:

Huge number of hypotheses to be tested.

Data : $\mathbf{X} \sim F_X(k_1, k_2, ...)$

where:

k_a = cumulants of the *unknown* , multivariate distribution F_X

Model : function of $\theta =$ (vector-) parameter

Sample $\mathbf{X^1}, \mathbf{X^2}, ..., \mathbf{X^m} \sim F_X(k_1, k_2, ...)$ (i.i.d)

$m << d$

Sample estimate: $\hat{\theta} \sim P_*(k_{*1}, k_{*2}, ...)$ = functional(F_X)

Re-sample: $\hat{\theta}_\mathbf{r} \sim P_r(k_{r1}, k_{r2}, ...)$ = functional(F_X)

Strategies for hypotheses testing:

Assume MVNormal distributions and do exact/approximate MCP adjustments

Create empirical distributions by re-sampling methods and apply MCP adjustments

1.2 Motivation: gene expression levels microarray data

Clinical uses:

- *Disease prognosis*: Gene expression levels measured by microarrays may be used to predict an individual's prognosis in a certain disease. For example, MammaPrint by Agendia provides prognostics for breast cancer patients based on expression levels in tumor samples measured by microarrays.

- *Pharmacogenomics*: Pharmacogenomics is the co-development of a *drug* that targets a subpopulation of the patient population, as well as a *device* that can be used to predicts whether a patient is in this subpopulation of responders to the drug. An example of such a drug is Herceptin for HER-2/neu positive breast cancer patients.

Developing a clinical device for prognostic or pharmacogenomics is a 2-stage process:

The first stage: find marker genes to train a prognostic algorithm, based on data with known disease outcomes. $---->$ uses MCP to select genes.

The second stage: a separate clinical trial validating the prediction algorithm, in terms of sensitivity and specificity.

1.3 Our solution

Hypotheses $<---->$ Parameters $<---->$

Test statistics ($\mathbf{T}$) $<---> * maxT <--->$

critical values c_I defined by $P(maxT > c_I) = \alpha$ so that:

P(*more than* m *mistakes*) $< \alpha$

The exact, theoretical distribution of the tests is compared with the analytically derived re-sampling distributions, all being expressed as:

functionals of the data underlying multivariate (unknown) distributions.

2 Error rates and error control

Testing the multiple null hypotheses $H_{01}, \ldots, H_{0d}$

Type I errors: rejecting a true null hypothesis.

Type II errors: failing to reject a false null hypothesis.

Impossible to be simultaneously minimized $- - - >$ solution:

Specify an acceptable level α for the Type I error rate. Then select the procedure (in this class) which minimizes Type II (maximizes power).

Error rate control, at α - level should satisfy:

$$P(\text{an incorrect decision}) \leq \alpha. \qquad (1)$$

1. the *comparison-wise error rate*:

   each H_{0i} is tested so that $P(\text{reject } H_{0i}) \leq \alpha$

   when H_{0i} is true.

2. the *experiment-wise error rate*

   (*weak control of the family-wise error rate*):

   when all H_{0i} are true,

$$P(\text{reject at least one } H_{0i}) \leq \alpha.$$

3. the *family-wise error rate* (FWER), *strongly* controlled:

   (*control of the maximum Type I error rate*)

   regardless of which $\{H_{0i},\ i \in I\}, I \subseteq \{1, \ldots, k\}$, are true,

$$P(\text{reject at least one } H_{0i},\ i \in I) \leq \alpha.$$

4. the *False Discovery Rate* (FDR):

   the expected *proportion* of incorrectly rejected (true) null hypotheses is no more than α :

$$E\left(\frac{\text{no. of true } H_{0i} \text{ rejected}}{\text{total no. of } H_{0i} \text{ rejected}}\right) \leq \alpha.$$

   Consequence: good performance when majority of H_{0i} are true.

5. generalised family-wise error rates (*gFWER*):

$$P(\text{reject more than } m \text{ of true } H_{0i}, i \in I) \leq \alpha$$

with m - a “reasonable” number.

gFWER at $m = 0$ is FWER.

3 The MCP principle

(*Stefansson et all (1988); FinnerStrassburger(2002)*)

Simple example: hypotheses to be tested: $\{\theta_0^i = 0\}, i = 1, 2$

Partitioning principle creates the **disjoint** hypotheses:

$\theta_0^1 = 0$ AND $\theta_0^2 \neq 0$

$\theta_0^1 \neq 0$ AND $\theta_0^2 = 0$

$\theta_0^1 = 0$ AND $\theta_0^2 = 0$

(number of hypotheses: $2^d - 1$, with $d = 2$)

Consequences:

(i) at most one of the disjoint hypotheses can be true.

(ii) we do not need any multiplicity adjustment

(iii) the number of disjoint hypotheses can be very large (if d is large)

More general form of Partitioning principle

Hypotheses to be tested $H_{0i} : \theta \in \Theta_i,\ i \in I,$

Scientific hypotheses to be proven: the complements Θ_i^{c}, of $\Theta_i,\ i \in I,$ and their intersections.

1. Partition $\bigcup_{i \in I} \Theta_i$ into disjoint $\Theta^\star_J,\ J \subseteq I$, as follows: For each $J \subseteq I$, let $\Theta^\star_J = \bigcap_{i \in J} \Theta_i \bigcap (\bigcap_{j \notin J} \Theta_j^{\mathsf{c}})$. Then $\{\Theta^\star_J,\ J \subseteq I\}$, including $\Theta^\star_\emptyset = \bigcap_{j \in I} \Theta_j^{\mathsf{c}}$, partition the parameter space. Note that $\Theta^\star_J$ can be interpreted as the part of the parameter space in which exactly $H_{0i},\ i \in J$, are true, and $H_{0j},\ j \notin J$ are false.

2. Test each $H_{0J}^{P} : \theta \in \Theta_{J}^{\star}$ at level α. Since the null hypotheses are disjoint, at most one null hypothesis is true. Therefore, even though no multiplicity adjustment to the levels of the tests is made, the probability of rejecting at least one true null hypothesis is at most α.

3. For all $J \subseteq I$, infer $\theta \notin \Theta_J$ if all $H^\star_{0J'}$ such that $J \subseteq J'$ are rejected. That is, infer the intersection null hypothesis H_{0J} is false if all null hypotheses $H^\star_{0J'}$ implying it are rejected. In terms of useful scientific inference of rejecting the original null hypotheses $H_{0i},\ i \in I$, since $\Theta_i = \bigcup_{J \ni i} \Theta^\star_J$, *Partitioning rejects* H_{0i} if *all* H^{P}_{0J} such that $J \ni i$ are rejected.

Notations:

Test statistic: $\mathbf{T} = (T_1, T_2, ..., T_g)$

$[1], \ldots, [g]$ random indices such that $T_{[1]} < \cdots < T_{[g]}$

Step-down Algorithm

Step 1: If $T_{[g]} > c_{\{[1],\ldots,[g]\}}$, then infer $\theta_{[g]} \neq 0$ and go to step 2; else stop.

Step 2: If $T_{[g-1]} > c_{\{[1],\ldots,[g-1]\}}$, then infer $\theta_{[g-1]} \neq 0$ and go to step 3; else stop.

. . .

Step g : If $T_{[1]} > c_{\{[1]\}}$, then infer $\theta_{\{[1]\}} \neq 0$ and stop; else stop.

Conditions of validity

S0: A level-α test for $H_I^\star$ should satisfy

$$sup_{\theta \in \Theta_I} P_\theta \{ max_{i \in I} \, T_i > c_I \} \leq \alpha$$

where $\Theta_I = \{\theta : \theta_i = 0 \text{ for } i \in I \text{ and } j \neq 0 \text{ for } j \notin I\}$ (! the supremum of this rejection probability may or may not occur at $\theta_1 = \cdots = \theta_g = 0$.)

S1: Tests for all hypotheses are based on statistics $T_i, \; i = 1, \ldots, g$, whose values do not vary with $H_{0I}^\star$;

S2: The level-α test for $H_{0I}^\star$ is of the form of rejecting $H_{0I}^\star$ if $max_{i \in I} \, T_i > c_I$;

S3: Critical values c_I have the property that if $J \subset I$ then $c_J \leq c_I$.

Comments

- Hochberg's (1988) method is a step-up version of the Bonferroni test.
- Holm's (1979) method is a step-down version of the Bonferroni test.
- they both are special cases of Partitioning testing and control FWER.
- Partitioning is related to (although more fundamental than) Closed Testing *(MarcusPeritzGabriel(1976))*.

4 Models. Test statistics. Critical values

Model 1: linear model

$\mathbf{X} = \mu_{\mathbf{X}} + \epsilon$ and $\mathbf{Y} = \mu_{\mathbf{Y}} + \epsilon$

Hypotheses of interest: is there a difference between the mean - vector - components of X and Y ?

$\theta^i = \mu_X^i - \mu_Y^i$

$H_{0i} : \theta^i = 0$

Data: observations $\mathbf{X}^\beta$, $\mathbf{Y}^\beta$, with $\beta = 1, 2, \ldots,$ sample size.

Model 2: linear mixed effects model

$$\mathbf{X}_\alpha = \mu_\mathbf{X} + \mathbf{Z}_\alpha + \epsilon$$

and

$$\mathbf{Y}_\alpha = \mu_\mathbf{Y} + \mathbf{Z}_\alpha + \epsilon$$

with:

$\mathbf{Z}_\alpha$ - random (subject) effects and $\mu_\mathbf{X}$, $\mu_\mathbf{Y}$ - fixed (group) effects, random effects and ϵ are independently distributed.

Hypotheses of interest: is there a difference between the mean - vector - components of X and Y ?

$$\theta^i = \mu_X^i - \mu_Y^i$$

Data: observations ${\mathbf{X}_\alpha}^\beta$, ${\mathbf{Y}_\alpha}^\beta$, with $\beta = 1, 2, \ldots$, sample size and $\alpha = 1, 2, \ldots$ within-subject-sample-size.

No normality assumptions.

Test statistics

(i) $T_i = \bar{X}_i - \bar{Y}_i$, with $i = 1, 2, ...k$, so $\mathbf{T} = \bar{\mathbf{X}} - \bar{\mathbf{Y}}$

(ii) standard multivariate t-test:

$$\mathbf{t}_{stand} = (\hat{k}_2(\mathbf{T}))^{-1/2}\mathbf{T} \qquad (2)$$

where $\hat{k}_2(\mathbf{T})$ = the estimated variance-covariance matrix Σ of the random variable $\mathbf{T}$

In practice, an other version of this test *is used*:

$$\mathbf{t}_{pract} = (\hat{k}_{2,diag}(\mathbf{T}))^{-1/2}\mathbf{T} \tag{3}$$

where $(\hat{k}_{2,diag}(\mathbf{T}))^{-1/2} = \Sigma_{diag}^{-1/2}$ and Σ_{diag} is a diagonal matrix with same diagonal elements as Σ.

Relation between the 2 versions:

$$\mathbf{t}_{pract} = \Sigma_{diag}^{-1/2} \Sigma^{1/2} \mathbf{t}_{stand} = \Omega \mathbf{t}_{stand} \qquad (4)$$

$$\left(k_a(\mathbf{t}_{pract})\right)^{i_1 i_2 \ldots i_a} =$$

$$\sum_{j_1 j_2 \cdots j_a} \Omega_{i_1 j_1} \Omega_{i_2 j_2} \ldots \Omega_{i_p j_a} \left(k_a(\mathbf{t}_{stand})\right)^{j_1 j_2 \cdots j_a}$$

5 Theoretical results on $maxT$ re-sampling distributions

Types of re-sampling:

(i) with replacement (bootstrap) / without replacement (permutations)

(ii) from raw data (observations) / from model-residuals

Comparing the test statistic distributions obtained by various re-sampling methods with the true distribution: in terms of cumulants.

Test statistic re-sampling distributions

Proposition 1 (*true distribution*)

Let $\mathbf{X} \sim F_X$, $\mathbf{Y} \sim F_Y$, and cumulants of general multivariate (dimension K) distributions F_X, F_Y exist and are finite (at least up to some order q). Let $\mathbf{T} = \bar{\mathbf{X}} - \bar{\mathbf{Y}}$. Then the cumulants of the test statistic *true* distribution P_{theor} are:

$$k_a(\mathbf{T}^{theor}) = m^{1-a} k_a(F_X) + (-1)^a n^{1-a} k_a(F_Y) \quad (5)$$

Proposition 2 (*re-sampling distributions*)

Let $\mathbf{X} \sim F_X$, $\mathbf{Y} \sim F_Y$, and assume that cumulants of general multivariate (dimension K) distributions F_X, F_Y exist and are finite (at least up to some order q). Let $\mathbf{T} = \bar{\mathbf{X}} - \bar{\mathbf{Y}}$ and assume large sample sizes.

Then: $k_a(P_{Bboot}) \approx k_a(P_{resBoot}) \approx k_a(P_{theor})$ and $k_a(P_{permut}) = k_a(P_{pboot}) \approx k_a(P_{respool})$, for any $a > 1$ and k_1 is the zero - vector for all methods except bootstrap raw data. The cumulants for all methods are explicitly given as functionals of the original distributions as follows:

a1) *re-sampling with replacement*, from the raw data samples (P_{Bboot}):

$$k_a(\mathbf{T}^{Bboot}) = m^{1-a}k_a(F_X) + (-1)^a n^{1-a}k_a(F_Y) \quad (6)$$

(large sample size n approximation, order $O(n^{-1})$)

a2) re-sampling with replacement, separately, on each group of *residuals*:

$$k_a(\mathbf{T}^{resBoot}) = \frac{k_a(F_X)}{m^{a-1}}\left((1-1/m)^a + (-1)^a\frac{m-1}{m^a}\right) + \quad (7)$$
$$(-1)^a\frac{k_a(F_Y)}{n^{a-1}}\left((1-1/n)^a + (-1)^a\frac{n-1}{n^a}\right)$$

b1, b2) *permutations* (P_{permut}) =

re-sampling with replacement (P_{pboot}) of/from the *pooled* sample:

$$k_a(\mathbf{T}^{permut}) = \frac{k_a(F_X)}{m^{a-1}} + \frac{(-1)^a k_a(F_Y)}{n^{a-1}} - \tag{8}$$

$$(k_a(F_X) - k_a(F_Y))\frac{1/m^a - (-1)^a/n^a}{1/m + 1/n} =$$

$$\left(\frac{1}{m^{a-1}} + (-1)^a \frac{1}{n^{a-1}}\right) \frac{mk_a(F_X) + nk_a(F_Y)}{m+n} =$$

$$k_a(\mathbf{T}^{pboot})$$

b3) re-sampling with replacement from pooled residuals ($P_{respool}$):

$$k_a(\mathbf{T}^{respool}) = \frac{1/m^{a-1}+(-1)^a/n^{a-1}}{m+n} \cdot \left[\frac{k_a(F_X)}{m^{a-1}}(m-1)\left((m-1)^{a-1}+(-1)^a\right) + \frac{k_a(F_Y)}{n^{a-1}}(n-1)\left((n-1)^{a-1}+(-1)^a\right)\right] \quad (9)$$

Simple examples ($a = 2$):

re-sampling *raw data*:

$$k_2(\mathbf{T}^{Bboot}) \approx k_2(\mathbf{T}^{theor}) = \frac{1}{m}k_2(F_X) + \frac{1}{n}k_2(F_Y) \quad (10)$$

$$k_2(\mathbf{T}^{pboot}) = k_2(\mathbf{T}^{permut}) = \frac{1}{n}k_2(F_X) + \frac{1}{m}k_2(F_Y) \quad (11)$$

and re-sampling *residuals*:

$$k_2(\mathbf{T}^{respool}) = k_2(\mathbf{T}^{permut}) - \frac{k_2(F_X) + k_2(F_Y)}{mn} \quad (12)$$

$$k_2(\mathbf{T}^{resBoot}) = k_2(\mathbf{T}^{Bboot}) - \frac{k_2(F_X)}{m^2} - \frac{k_2(F_Y)}{n^2} \quad (13)$$

$maxT$ distributions

$$P(maxT = a) = \sum_{i=1}^{K} P(\mathbf{T}'_{(\mathbf{i})}|T^i = a) = \qquad (14)$$

$$\sum_{i=1}^{K} \prod_{l \neq i} \int_{-\infty}^{a} dT'_l P(\mathbf{T}'_{(\mathbf{i})}|T_i = a)$$

where $P(maxT) = P(max_{i \in I} T_i)$ and $\mathbf{T}'_{(\mathbf{i})} = (T^1, ..., T_{i-1}, T_{i+1}, ..., T_K)$ is the vector obtained from $\mathbf{T}$ by removing the component i.

The critical values c_I are solutions of:

$$\alpha = \int_{c_I}^{\infty} P(maxT = a) da \qquad (15)$$

Distribution of the vector test statistic may be expressed in terms of cumulants:

$$P(\mathbf{T}) =$$

$$\int e^{i\rho T} e^{-ik_1\rho - \frac{1}{2}\rho k_2 \rho + \sum' \frac{1}{a!} k_a^{n_1 n_2 \dots n_k} \rho_1^{n_1} \rho_2^{n_2} \dots \rho_k^{n_k} + \dots} \frac{d^g \rho}{(2\pi)^g} = \tag{16}$$

$$\frac{1}{|\sqrt{k_2}|} \int e^{iz\rho} e^{-\rho^2/2} e^{\sum' \frac{1}{a!} \tilde{k}_a^{n_1 n_2 \dots n_{ig}} \rho_1^{n_1} \rho_2^{n_2} \dots \rho_k^{n_k} + \dots} \frac{d^g \rho}{(2\pi)^g} \tag{17}$$

where $\tilde{k}_a = k_a/(\sqrt{k_2})^a$, formally.

Note: $\tilde{k}_a$ will be of order $m^{-a/2+1}$ (since k_a is of order m^{-a+1} for any a).

$$P(\mathbf{T}) = \frac{1}{|\sqrt{k_2}|}\phi(z)(1 + \text{ higher order terms }) \qquad (18)$$

where higher order terms
$O(m^{-1/2}), O(m^{-1}), O(m^{-3/2}),$ ... can be expressed in terms of multidimensional Hermite polynomials and
$\mathbf{T} = k_1(T) + \mathbf{z}\sqrt{k_2(T)}$.

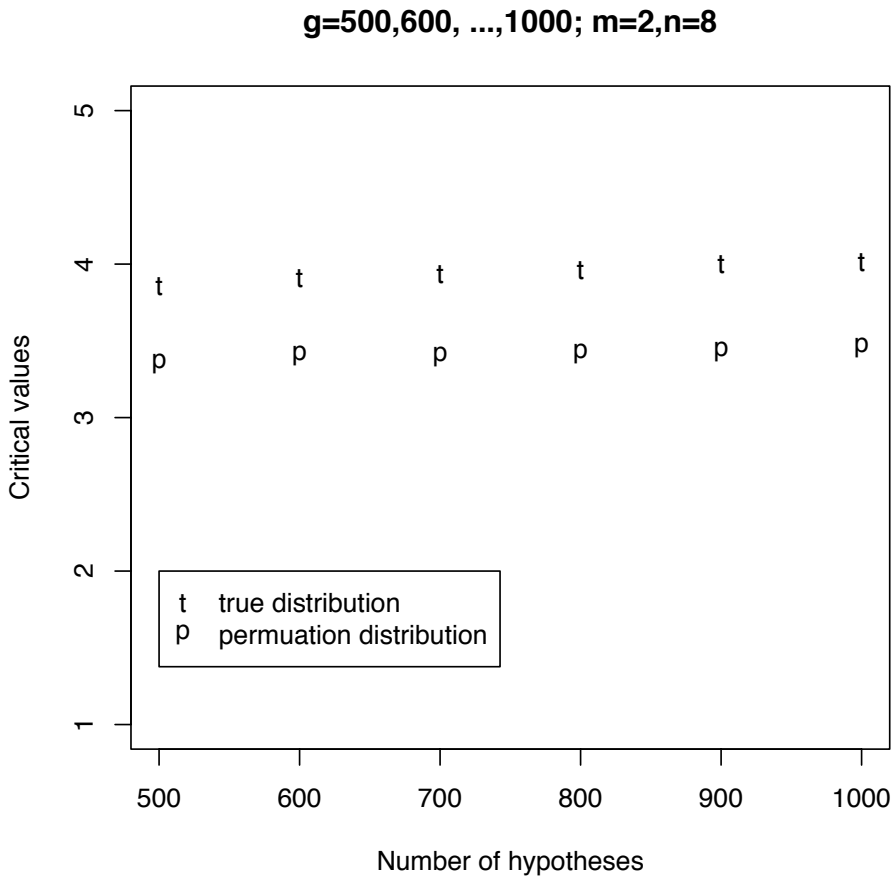
Example 1: same means, different correlations

10,000 sample data sets

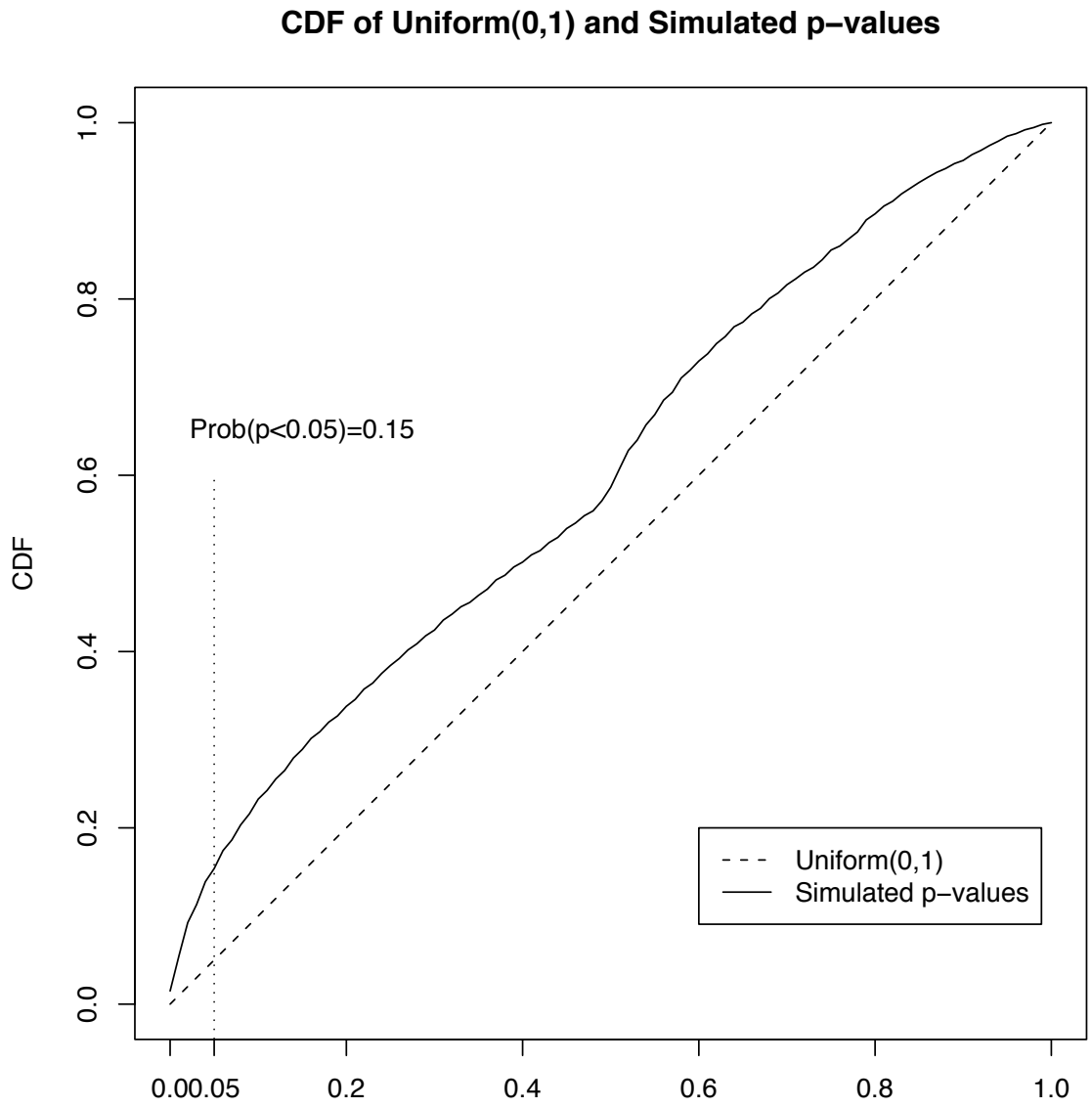
$g = 500, 600, ..., 1000.$ dimensional distributions:

$MVN_g(\mu_X, \mathbf{\Sigma}_X)$ and $MVN_g(\mu_Y, \mathbf{\Sigma}_Y)$ where $\mu_X = \mu_Y = 0$, $\mathbf{\Sigma}_X$ has all the diagonal elements equal to 1 and all the off-diagonal elements equal to zero, while $\mathbf{\Sigma}_Y$ has all the diagonal elements equal to 1 and all the off-diagonal elements equal to 0.9.

Sample sizes: $m = 2$ and $n = 8$.


g=500,600, ...,1000; m=2,n=8
Critical values
Number of hypotheses
1
2
3
4
5
500
600
700
800
900
1000
t
p
t true distribution
p permuation distribution

Example 2: same means, different skewness


CDF of Uniform(0,1) and Simulated p−values
CDF
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.05
0.2
0.4
0.6
0.8
1.0
Prob(p<0.05)=0.15
Uniform(0,1)
Simulated p−values

6 Conclusions

- An advantage of permutation testing is no knowledge of the distribution of the observations is required. Its control of error rate, however, only holds under the condition of identical distribution among groups to be compared. If the purpose of testing is to detect differences in means, then permutation testing may pick up unintended signals, rejecting an equality hypothesis for the wrong reason.

- The true and permutation distributions of the test statistic $\mathbf{T}$ will have the same even-order cumulants if $m = n$.
- The true and permutation distributions of the test statistic will not necessarily have the same a^{th} order cumulants for a odd, regardless of whether $m = n$, *unless* $k_a(F_X) = k_a(F_Y)$.
- If $\mathbf{X}$ and $\mathbf{Y}$ are not multivariate normal, then differences in cumulants of order higher than two can cause permutation test for equality of means to be liberal even if $m = n$.

- Best re-sampling solutions: re-sampling with replacement, residuals or raw data (with post-centering), *when* the sample sizes are reasonably *large*.

References

[HsuJC, 1996] Hsu, J. C. (1996). Multiple comparisons. Theory and methods. *Chapman & Hall*.

[Stefansson et all (1988)] Stefansson G., Kim W., Hsu J.C. (1988). On confidence sets in multiple comparisons In *Gupta S.S. and Berger J.O. (eds), Statitical decision theory and related topics IV*, 2: 89 - 104 *Springer Verlag, New-York.*

[Pollard and van der Laan (2005)] Pollard, K. S. and van der Laan, M. (2005). Re-Sampling-based multiple testing: Asymptotic control of type i error and applications to gene

expression data. *Journal of Statistical Planning and Inference*, 125:85–100.

[Pollard and van der Laan (2003)] Pollard, K. S. and van der Laan, M. (2003). Multiple testing for gene expression data: an investigation of null distributions with consequences for the permutation test. In *Proceedings of the 2003 International MultiConference in Computer Science and Engineering*, pages 3–9. METMBS'03 Conference.

[Huang et all (2006)] Huang Y., Xu H., Calian V., Hsu J. C. (2006) To permute or not permute. *Bioinformatics*, vol.22,18: 2244 - 2248.

[vantVeeretal(2002)] van ’t Veer L. J., Dai H., van de Vijver M. J., He Y. D., Hart A. M., Mao M., Peterse H. L., van der Kooy K., Marton M.J., Witteveen A., T., Schreiber G. J., Kerkhoven R., M., Roberts C., Linsley P.S., Bernards R., Friend S.H. Gene expression profiling predicts clinical outcome of breast cancer *Nature*, 415: 530 - 536.

[FinnerStrassburger(2002)] Finner H., Strassburger K., (2002) The partitioning principle: a powerful tool in multiple decision theory. *Annals of Statistics*, 30: 1194 - 1213

[MarcusPeritzGabriel(1976)] Peritz M., Peritz R., Peritz E., Gabriel K. R.(1976) On closed testing procedures with special

reference to ordered analysis of variance. *Biometrika*, 63: 655 - 660.

[XuHsu(2007)] Xu, H. and Hsu, J. C. (2007) Applying the generalized partitioning principle to control the generalized familywise error rate *Biometrical Journal*, 49, 1: 52 - 67.