Large Scale Multiple Hypothesis Testing
Correlation and Permutations

Jelle Goeman and Aldo Solari

36th Annual Conference of the
International Society for Clinical Biostatistics
August 23, 2015 - Utrecht

Outline

1 **Dependence structures**

2 **Permutation null distribution**

3 **Westfall & Young permutation method**

4 **Meinshausen's permutation method**

Dependence structures

Null p-values

$q_1, \ldots, q_{m_0}$ p-values of true hypotheses

Marginal distribution

(super)uniform

Joint distribution

Unknown dependence structure.
Independence structure not realistic for genomics or fMRI data.

Assumptions and FWER control

No assumptions
Conservative except for 'worst case' dependence

Simes' inequality
Conservative except for independence

Permutation methods
Exact because they adapt to the dependence structure

Variability of the number of false discoveries

Dependence structure

Affects the variance of $V = \sum_{i=1}^{m_0} \mathbb{1}\{q_i \leq \alpha\}$, but not its mean

Mean-Variance structure of V

$$
\begin{aligned}
\mathrm{E}(V) &= \sum_{i=1}^{m_0} \mathrm{E}(\mathbb{1}\{q_i \leq \alpha\}) = m_0 \cdot \alpha \\
\mathrm{Var}(V) &= \sum_{i=1}^{m_0} \sum_{j=1}^{m_0} \mathrm{Cov}(\mathbb{1}\{q_i \leq \alpha\}, \mathbb{1}\{q_j \leq \alpha\}) \\
&= \underbrace{m_0 \cdot \alpha \cdot (1 - \alpha)}_{\text{independence structure}} + \\
&\quad + \underbrace{2 \cdot \sum_{i < j} \left\{ \mathrm{P}(q_i \leq \alpha, q_j \leq \alpha) - \mathrm{P}(q_i \leq \alpha) \cdot \mathrm{P}(q_j \leq \alpha) \right\}}_{\text{overdispersion term}}
\end{aligned}
$$

Example: exchangeable correlation structure

Test statistics and p-values

$$(Z_1, \ldots, Z_{m_0}) \sim \mathcal{N}\Big(\underset{m_0 \times 1}{0}, (1 - \rho) \underset{m_0 \times m_0}{I} + \rho \underset{m_0 \times m_0}{1} \Big)$$

Gaussian joint null distribution with exchangeable correlation structure $\rho \geq -1/(m_0 - 1)$

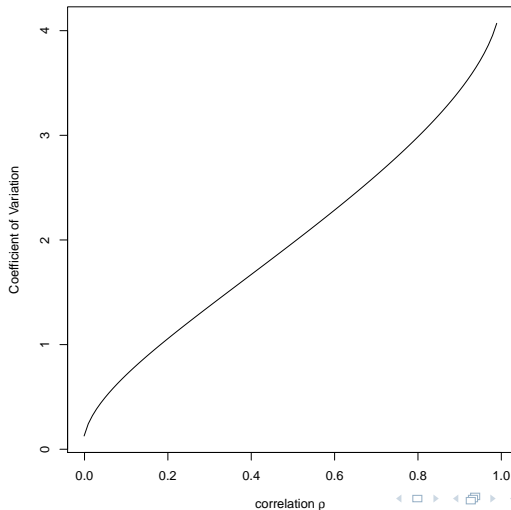$$(q_1, \ldots, q_{m_0}) = (1 - \Phi(Z_1), \ldots, 1 - \Phi(Z_{m_0}))$$

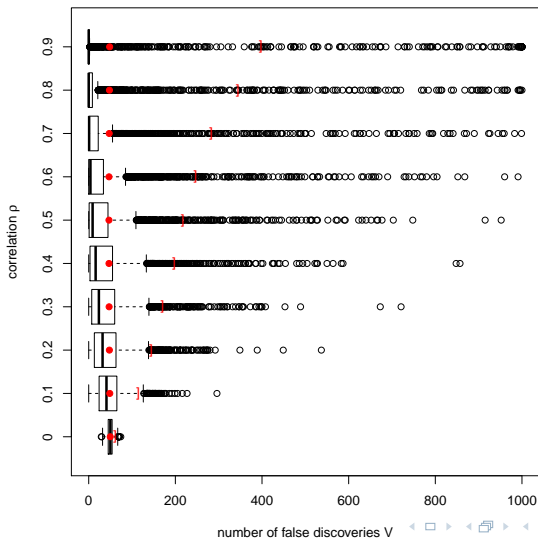Gaussian Copula with parameter ρ

Variance of V

$$\mathrm{Var}(V) = m_0 \cdot \alpha \cdot (1 - \alpha) + m_0 \cdot (m_0 - 1) \cdot \{ C_2(\alpha, \alpha; \rho) - \alpha^2 \}$$

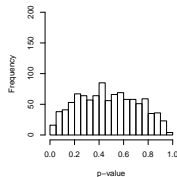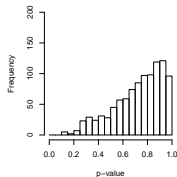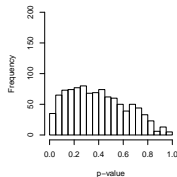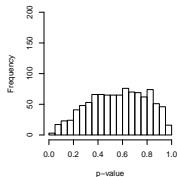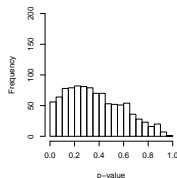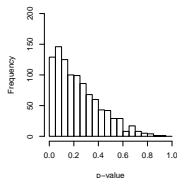where $C_2(u, v; \rho) = \Phi_2(\Phi^{-1}(u), \Phi^{-1}(v); \rho)$

Coefficient of variation $\dfrac{\sqrt{\mathrm{Var}(V)}}{|\mathrm{E}(V)|}$ as a function of ρ

for $m_0 = 1000$ and $\alpha = 0.05$

Boxplot of V (2000 realizations) as a function of ρ
for $m_0 = 1000$ and $\alpha = 0.05$

Examples of p-values histograms for $m_0 = 1000$ and $\rho = 1/2$
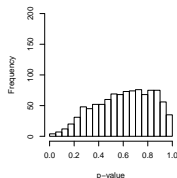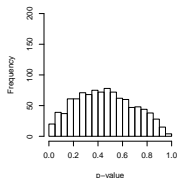
One particular example with $m_0 = 1000$ and $\rho = 1/2$
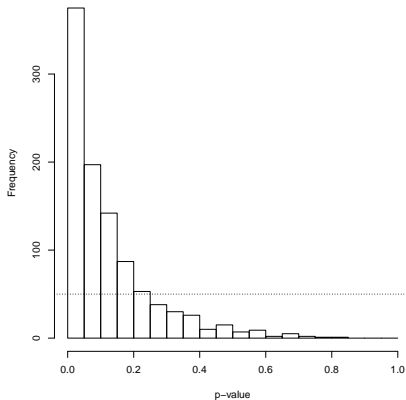
Number of rejections by testing at $\alpha = 0.05$: 375
Expected number of false discoveries: $m_0 \cdot \alpha = 50$
Estimated number of false discoveries (Storey$-1/2$): 2.8
Bonferroni and Simes inequalities satisfied

Conservativeness of Bonferroni

Example

$(q_1, \ldots, q_{m_0}) \sim$
Gaussian Copula with
parameter ρ

$$\mathrm{FWER} = \mathrm{P}\left(\bigcup_{i=1}^{m_0} \left\{ q_i \leq \frac{\alpha}{m_0} \right\} \right)$$

as function of m_0 and ρ
with $\alpha = 0.05$
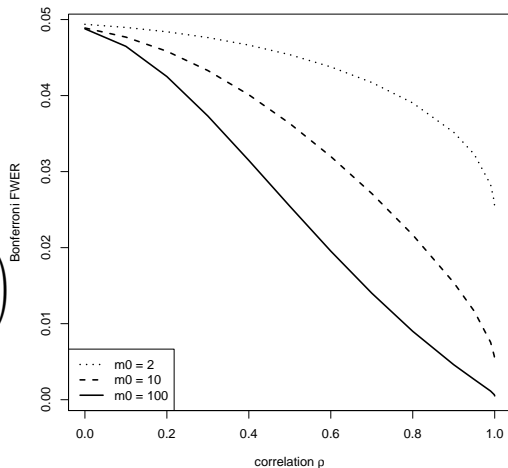
Exact FWER control

$$\mathrm{FWER} = \mathrm{P}\left(\bigcup_{i=1}^{m_0}\left\{q_i \leq \tilde{\alpha}\right\}\right) = \mathrm{P}\left(q_{(1)} \leq \tilde{\alpha}\right) = \alpha$$

where $\tilde{\alpha}$ is the α-quantile of $\min Q = q_{(1)} = \min(q_1, \ldots, q_{m_0})$

But we need to know the distribution of $\min Q$!

Example: exchangeable correlation structure

Test statistics

$$(Z_1, \ldots, Z_{m_0}) \sim \mathcal{N}\left(\underset{m_0 \times 1}{0}, (1-\rho)\underset{m_0 \times m_0}{I} + \rho \underset{m_0 \times m_0}{1}\right)$$

maxZ

$\max Z = Z_{(m_0)} = \max(Z_1, \ldots, Z_{m_0})$ with density function

$$f_{\max Z}(z) = m_0\ \phi(z)\ \Phi_{m_0-1}\left(z\sqrt{1-\rho}\underset{m_0-1 \times 1}{1}; \underset{m_0-1 \times 1}{0}, (1-\rho)\underset{m_0-1 \times m_0-1}{I} + \rho\underset{m_0-1 \times m_0-1}{1}\right)$$

Exact FWER control

$$\mathrm{FWER} = \mathrm{P}\left(Z_{(m_0)} > \tilde{c}\right) = \alpha$$

where $\tilde{c}$ is the $(1-\alpha)$-quantile of $\max Z$

Density function and $(1 - \alpha)$-quantile of $\max Z$ as a function of ρ with $m_0 = 10$
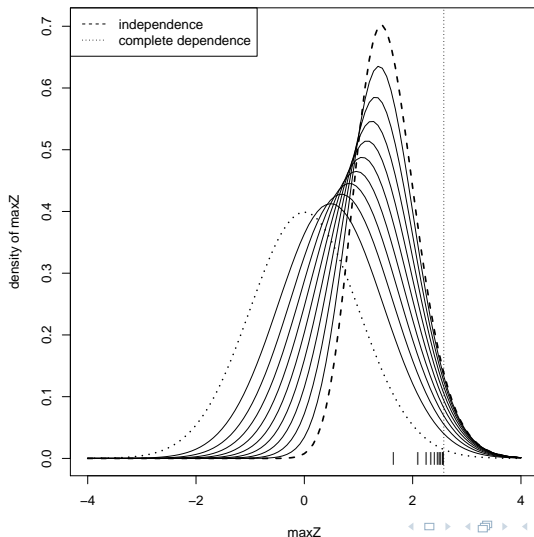
Outline

1 Dependence structures

2 Permutation null distribution

3 Westfall & Young permutation method

4 Meinshausen's permutation method

Monte Carlo null distribution

- Assume that data $\underset{n \times m_0}{Y_0}$ is generated from $t_0 \in \bigcap_{i \in \mathcal{T}} H_i$

- $(q_1, \ldots, q_{m_0}) \sim q_0$
- Draw $(b - 1)$ IID $(q_1^*, \ldots, q_{m_0}^*)$ from q_0
- Store the results in a $b \times m_0$ matrix
  Get the minimal p-values along each row

$$
\underset{b \times m_0}{Q} = \left[\begin{array}{ccccc} q_1 & \cdots & q_i & \cdots & q_{m_0} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ q_1^* & \cdots & q_i^* & \cdots & q_{m_0}^* \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots \end{array} \right] \qquad \underset{b \times 1}{q_{(1)}} = \left[\begin{array}{c} q_{(1)} \\ \cdots \\ q_{(1)}^* \\ \cdots \\ \cdots \end{array} \right]
$$

- Monte Carlo $\tilde{\alpha}$ is the α-quantile of $\underset{b \times 1}{q_{(1)}} \overset{b \to \infty}{\to}$ to the true $\tilde{\alpha}$
- Knowledge of q_0 required!

16 / 41

Permutation null distribution

$$
\underset{b \times m_0}{Q} =
\left[
\begin{array}{ccc}
q_1 & \cdots & q_{m_0} \\
\cdots & \cdots & \cdots \\
q_1^* & \cdots & q_{m_0}^* \\
\cdots & \cdots & \cdots \\
\cdots & \cdots & \cdots
\end{array}
\right] =
\left[
\begin{array}{ccc}
q_1(Y_0) & \cdots & q_{m_0}(Y_0) \\
\cdots & \cdots & \cdots \\
q_1(\pi Y_0) & \cdots & q_{m_0}(\pi Y_0) \\
\cdots & \cdots & \cdots \\
\cdots & \cdots & \cdots
\end{array}
\right]
$$

where

- $\pi \in \boldsymbol{\pi}$ is a transformation of Y_0 (*null invariant*)
- $\boldsymbol{\pi} = \{id, \pi_2, \ldots, \pi_b\}$ is an algebraic group

Example: permutations

$$
\underset{n \times m_0}{Y_0} = \left[\begin{array}{ccccc}
y_{1,1} & \cdots & y_{1,i} & \cdots & y_{1,m_0} \\
\cdots & \cdots & \cdots & \cdots & \cdots \\
y_{j,1} & \cdots & y_{j,i} & \cdots & y_{j,m_0} \\
\cdots & \cdots & \cdots & \cdots & \cdots \\
y_{n,1} & \cdots & y_{n,i} & \cdots & y_{n,m_0}
\end{array}\right]
$$

$$
\underset{n \times m_0}{\pi Y_0} = \left[\begin{array}{ccccc}
y_{\pi(1),1} & \cdots & y_{\pi(1),i} & \cdots & y_{\pi(1),m_0} \\
\cdots & \cdots & \cdots & \cdots & \cdots \\
y_{\pi(j),1} & \cdots & y_{\pi(j),i} & \cdots & y_{\pi(j),m_0} \\
\cdots & \cdots & \cdots & \cdots & \cdots \\
y_{\pi(n),1} & \cdots & y_{\pi(n),i} & \cdots & y_{\pi(n),m_0}
\end{array}\right]
$$

where

- $(\pi(1), \ldots, \pi(n))$ is a permutation of $(1, \ldots, n)$
- $\boldsymbol{\pi}$ is the permutation group with $b = n!$
- πY_0 shuffles the rows of Y_0

Permutation null distribution

analogous to Monte Carlo, but:

Not required

Knowledge of q_0!

Required: null invariance condition

Under $\bigcap\limits_{i \in \mathcal{T}} H_i$

$$\underset{b \times m_0}{Q} (Y_0) \stackrel{d}{=} \underset{b \times m_0}{Q} (\pi Y_0)$$

for every $\pi \in \boldsymbol{\pi}$.

The joint distribution of Q is invariant under any transformation $\pi \in \boldsymbol{\pi}$ of the data Y_0

📄 Goeman and Solari (2010). The sequential rejection principle of familywise error control. *The Annals of Statistics*, 38:3782–3810

Randomization hypothesis

Under $\bigcap_{i \in \mathcal{T}} H_i$

$$Y_0 \stackrel{d}{=} \pi Y_0$$

for every $\pi \in \boldsymbol{\pi}$.

The distribution of Y_0 is not affected by the transformations in $\boldsymbol{\pi}$

📕 Lehmann and Romano (2005), *Testing Statistical Hypotheses*, Springer, New York. Chapter 15, p. 633.

Sufficient condition
Implies the null invariance condition

Exchangeability
When $\boldsymbol{\pi}$ are permutations,
randomization hypothesis $\equiv$ Y_0 is (row) exchangeable

Example: multivariate linear model

$$\underset{n\times m}{Y} = \underset{n\times 1}{X}\ \underset{1\times m}{\beta} + \underset{n\times q}{Z}\ \underset{q\times m}{\gamma} + \underset{n\times m}{\varepsilon}$$

where

- Y contains the response (e.g. gene expression of m genes)
- X contains the covariate of interest (e.g. treatment vs control)
- Z contains the nuisance covariates (e.g. intercept, age)
- Rows of ε are IID from any distribution with zero mean and covariance matrix Σ (e.g. dependence structure of the genes)
- $\bigcap_{i=1}^{m} H_i = \bigcap_{i=1}^{m} \left\{ \beta_i = 0 \right\}$

Example: multivariate linear model

Nuisance covariates: intercept only

Under $\bigcap_{i=1}^{m} H_i$

$$\underset{n \times m}{Y} = \underset{n \times 1}{1} \underset{1 \times m}{\gamma} + \underset{n \times m}{\varepsilon}$$

and Y is (row) exchangeable:
rows of Y are IID with mean vector γ and covariance matrix Σ

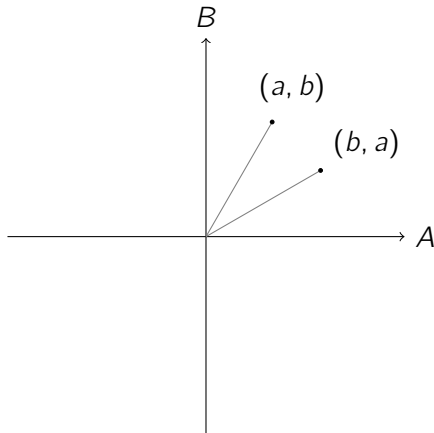### Nuisance covariates: intercept and age

Under $\bigcap_{i=1}^{m} H_i$

$$\underset{n \times m}{Y} = \underset{n \times 2}{Z} \underset{2 \times m}{\gamma} + \underset{n \times m}{\varepsilon}$$

but (row) exchangeability fails:
rows of Y are independent but not identically distributed

📄 Solari, Finos, Goeman (2014). Rotation-based multiple testing in the
multivariate linear model. *Biometrics*, 70:954–961
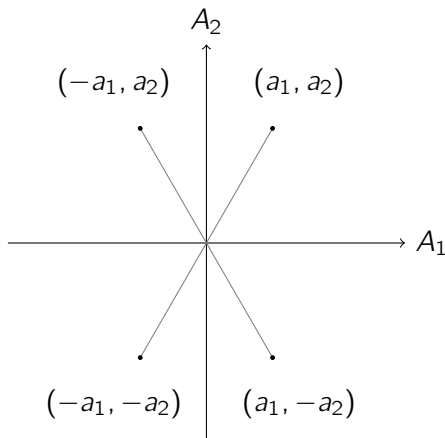
Two-sample situation:
1 IID sample of size n_A from A, 1 IID sample of size n_B from B
Assume location shift model: $A \stackrel{d}{=} B + \theta$
Under $H_0 : \theta = 0$, (a, b) and its permutation (b, a) are equally likely

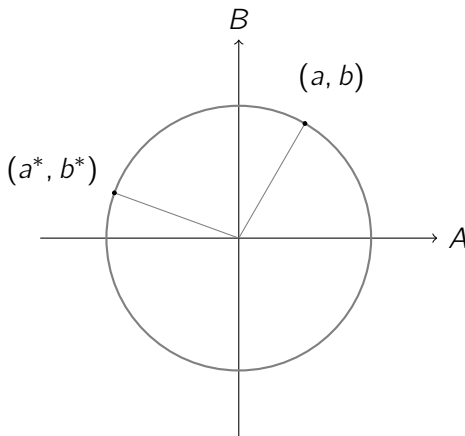*One-sample situation:*

one i.i.d. sample of size n from A

Assume symmetry around θ model: $A - \theta \overset{d}{=} \theta - A$

Under $H_0 : \theta = 0$, (a_1, a_2), $(-a_1, a_2)$, $(a_1, -a_2)$ and $(-a_1, -a_2)$ are equally likely

Assume $A \sim N(0, \sigma^2)$ and, independently, $B \sim N(\theta, \sigma^2)$
Under $H_0 : \theta = 0$, (a, b) and any rotation (a^*, b^*) are equally likely,
$(a^*, b^*) = (a \cdot \cos \tau - b \cdot \sin \tau, a \cdot \sin \tau + b \cdot \cos \tau)$ with $\tau \in [0, 360]$

Outline

1 Dependence structures

2 Permutation null distribution

3 Westfall & Young permutation method

4 Meinshausen's permutation method

Westfall & Young permutation method

Powerful
Attains the exact level α by adapting to the unknown dependence structure of the null p-values

Two variants
$\max T$ and $\min P$

Required
Null invariance condition (sufficient: randomization hypothesis)

Not required

- assumptions on the dependence structure of the null p-values
- (super)uniformity of null p-values: W&Y provides FWER control even with invalid and possibly anti-conservative null p-values

Single-step maxT

1 Construct the permutation null distribution under $\bigcap\limits_{i=1}^{m} H_i$

Calculate the minimal p-values along each row

$$
\underset{b \times m}{P} = \left[
\begin{array}{cccccc}
p_1 & \cdots & p_i & \cdots & \cdots & p_m \\
\cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\
p_1^* & \cdots & p_i^* & \cdots & \cdots & p_m^* \\
\cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\
\cdots & \cdots & \cdots & \cdots & \cdots & \cdots
\end{array}
\right]
\qquad
\underset{b \times 1}{p_{(1)}} = \left[
\begin{array}{c}
p_{(1)} \\
\cdots \\
p_{(1)}^* \\
\cdots \\
\cdots
\end{array}
\right]
$$

Calculate $\tilde{\alpha}$, the α-quantile of $\underset{b \times 1}{p_{(1)}}$

2 Reject all hypotheses with p-values $< \tilde{\alpha}$

Sequential (step-down) maxT

1. Remove all columns of $\underset{b \times m}{P}$ corresponding to the r hypotheses rejected

   Re-calculate the minimal p-values along each row

$$
\underset{b \times m - r}{P} = \left[\begin{array}{ccccccc} p_1 & \cdots & \cancel{p_r} & \cdots & \cancel{\times} & p_m \\ \cdots & \cdots & \cancel{\times} & \cdots & \cancel{\times} & \cdots \\ p_1^* & \cdots & \cancel{p_r^*} & \cdots & \cancel{\times} & p_m^* \\ \cdots & \cdots & \cancel{\times} & \cdots & \cancel{\times} & \cdots \\ \cdots & \cdots & \cancel{\times} & \cdots & \cancel{\times} & \cdots \end{array} \right] \qquad \underset{b \times 1}{p_{(1)}} = \left[\begin{array}{c} p_{(1)} \\ \cdots \\ p_{(1)}^* \\ \cdots \\ \cdots \end{array} \right]
$$

   Re-calculate $\tilde{\alpha}$, the α-quantile of $\underset{b \times 1}{p_{(1)}}$

2. Reject all hypotheses with p-values $< \tilde{\alpha}$

Repeatedly apply ① and ② until no new rejections occur

minP

- Same as $\max T$, except that it replaces $\underset{b \times m}{P}$ by $\underset{b \times m}{\tilde{P}}$ with per-hypothesis permutation p-values

$$\tilde{p}_{ji} = \frac{\#\{l : p_{li} \leq p_{ji}\}}{b}$$

where p_{ji} is the (j, i)th element of $\underset{b \times m}{P}$

- Computationally intensive
- Guarantees an uniform distribution of the input p-values
- Permutation p-values are discrete: it can happen that $\tilde{p}_{(1)} = \tilde{p}^*_{(1)}$ for most permutations, resulting in zero power

Random permutations

All possible permutations computationally not feasible

Usually, the number of permutations $\#\boldsymbol{\pi} = n!$ is too large

Random permutations

A set of randomly drawn permutations is not a group, but valid if

- draw $(b-1)$ times with replacement from $\boldsymbol{\pi}$ and add $\{id\}$
- draw $(b-1)$ without replacement from $\boldsymbol{\pi} \setminus \{id\}$ and add $\{id\}$

📄 Hemerik and Goeman (2015) Exact testing with random permutations,
arXiv:1411.7565v1

How many?

- $\max T$ requires $\geq 1/\alpha$, but 1000 is recommended at $\alpha = 5\%$
- $\min P$ requires $\geq m/\alpha$, but preferably many more

Application: Golub data

Leukemia ALL/AML study

- Y: gene expression of $m = 3051$ genes for $n = 38$ patients
- X: AML vs ALL
- Z: intercept

Single-step maxT

With $b = 1000$, the single-step $\max T$ method gives
$\tilde{\alpha} = 2.01 \times 10^{-5}$ (Bonferroni $= \alpha/m = 1.63 \times 10^{-5}$), which leads
to 104 rejections (Bonferroni $= 98$)

Sequential maxT

Stepping down by removing the rejected hypotheses leads to 3
additional rejections (Holm $= 98$)

R lab: permutations

Outline

1 Dependence structures

2 Permutation null distribution

3 Westfall & Young permutation method

4 Meinshausen's permutation method

Meinshausen's permutation method

FDP confidence bounds

$$P(\mathcal{Q}(\mathcal{R}) \leq \bar{\mathcal{Q}}(\mathcal{R})) \geq 1 - \alpha$$

holds for rejected sets of the form $\mathcal{R} = \{H_i : p_i \leq t\}$
simultaneously over $t \in [0, 1]$

Variant of Simes' inequality

Finds permutation-based critical values $\tilde{c}_1, \ldots, \tilde{c}_{m_0}$ such that

$$q_{(i)} > \tilde{c}_i \quad \text{for all } i = 1, \ldots, m_0$$

with probability at least $1 - \alpha$

Assumptions & Power

Same as in Westfall & Young permutation method

Permutation-based critical values

1. Construct the permutation null distribution under $\bigcap_{i=1}^{m} H_i$

   Sort the p-values along each row

   $$\begin{pmatrix} P \\ b \times m \end{pmatrix} = \left[\begin{array}{ccccc} p_{(1)} & \cdots & p_{(i)} & \cdots & p_{(m)} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ p_{(1)}^{*} & \cdots & p_{(i)}^{*} & \cdots & p_{(m)}^{*} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots \end{array} \right]$$

2. The values of $\tilde{c}_1, \ldots, \tilde{c}_m$ are chosen in such a way that a proportion $1 - \alpha$ of the permutation sorted p-values lie everywhere above $\tilde{c}_1, \ldots, \tilde{c}_m$:

   $$\frac{\#\{* : p_{(1)}^{*} \geq \tilde{c}_1, \ldots, p_{(m)}^{*} \geq \tilde{c}_m\}}{b} \geq 1 - \alpha$$
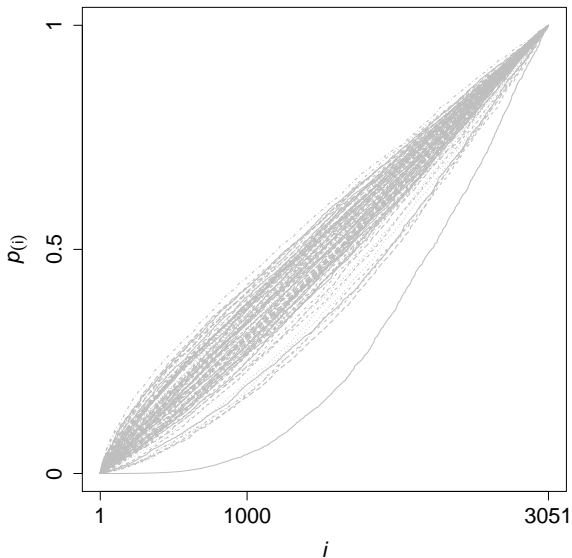
Permutation curves

Permutation-based critical values
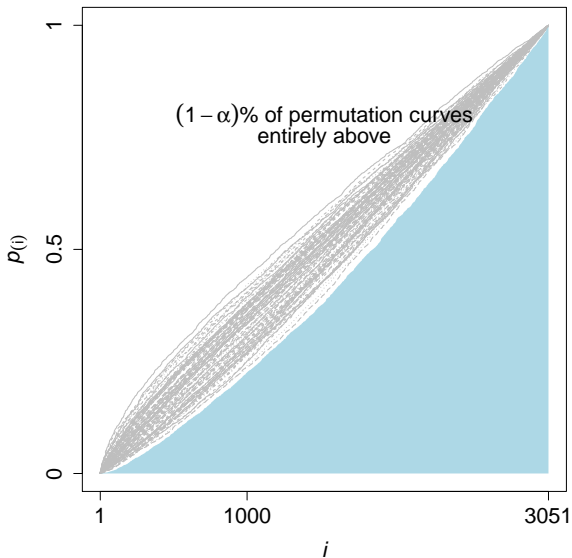
Simes' inequality

Application: Golub data

number of rejections
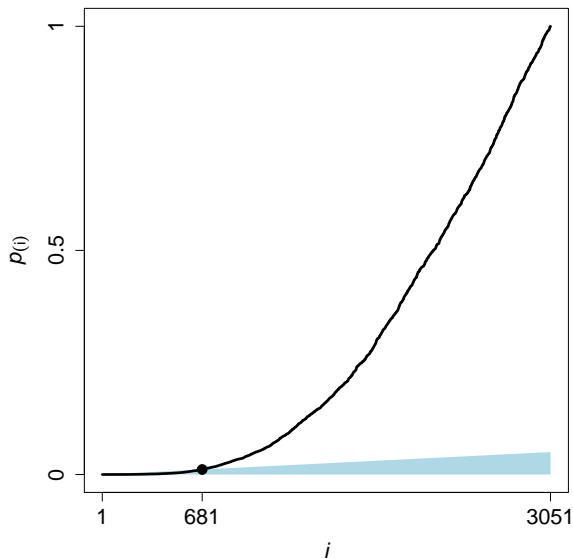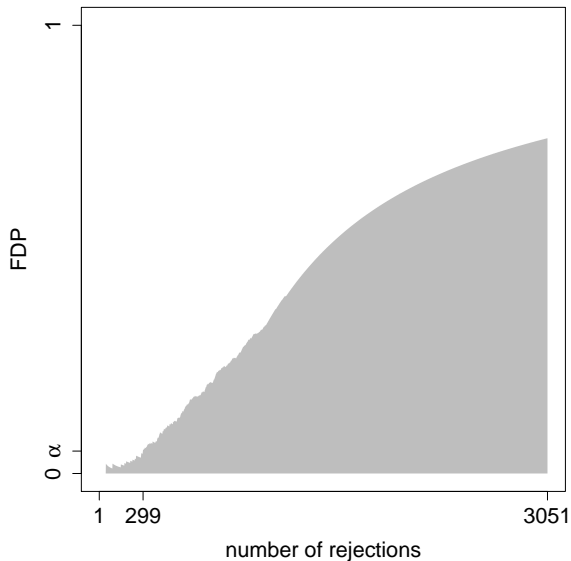
Application: Golub data

Meinshausen permutation method

With 95% confidence:

At least 769 hypotheses out of 3051 are false (FDP $\leq$ 74.8%)

The 44 hypotheses with best p-values are false

Goeman & Solari method (Hommel)

With 95% confidence:

At least 323 hypotheses out of 3051 are false (FDP $\leq$ 89.4%)

The 98 hypotheses with best p-values are false