Post hoc inference via JER control

Etienne Roquain[1]
Joint work with Gilles Blanchard[2] and Pierre Neuvial[2]

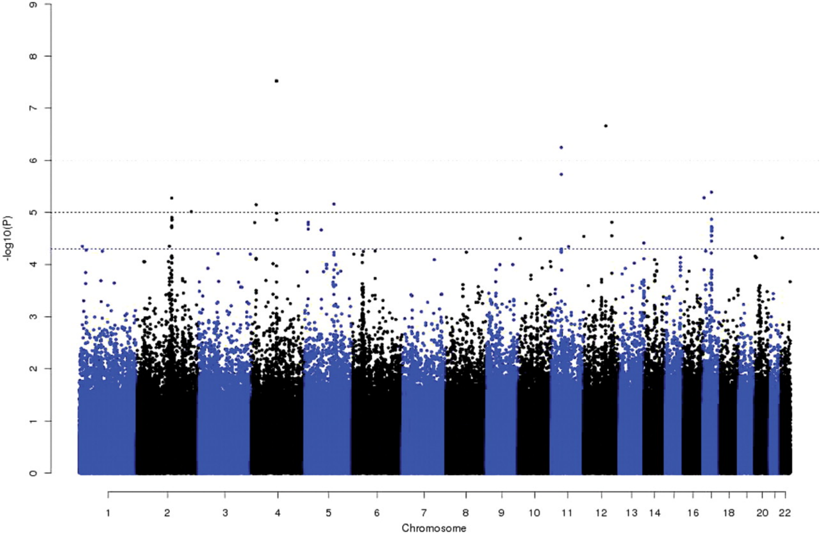
[1]LPMA, Université Pierre et Marie Curie, France
[2]Institut für Mathematik, Universität Potsdam, Germany
[3]IMT, Université Paul Sabatier, France

Mathematical Method of Modern Statistics, 12/07/2017

Arxiv 1703.02307

Find signal in massive datasets

- GWAS interesting SNPs? [Saad et al., 2011]



Multiple inferences

- Multiple testing:
  - derive the rejection set R
  - such that from FDR(R) $\leq \alpha$

  [Benjamini and Hochberg (1995)] . . . [Bogdan et al. (2014)], [Barber and Candès (2015)]

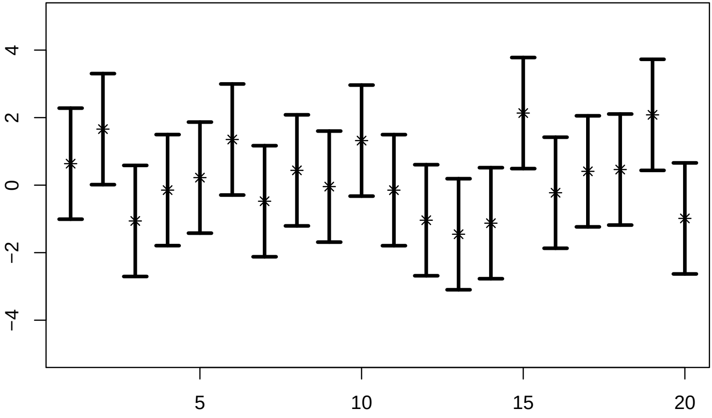
- Post-selective inference
  - Inference after specific selection [Lockhart et al. (2014) and Fithian et al. (2014)]
  - Inference after arbitrary selection
    - ⋆ confidence intervals on selected parameters [Benjamini and Yekutieli (2005)], [Berk et al. (2013)]
    - ⋆ estimator/bound on signal quantity after selection [Goeman and Solari (2011)]

CI no selection

Let

$$X \sim \mathcal{N}(\theta, I_m) \in \mathbb{R}^m, \quad \theta \in \mathbb{R}^m,$$

90% CI for each θ_i

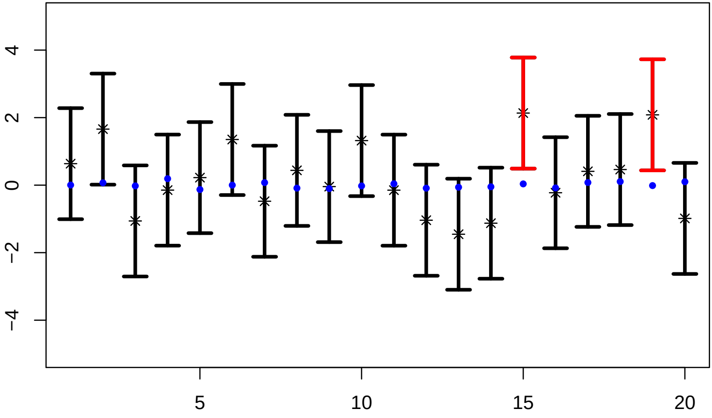


CI no selection

Let

$$X \sim \mathcal{N}(\theta, I_m) \in \mathbb{R}^m, \quad \theta \in \mathbb{R}^m,$$

90% CI for each θ_i

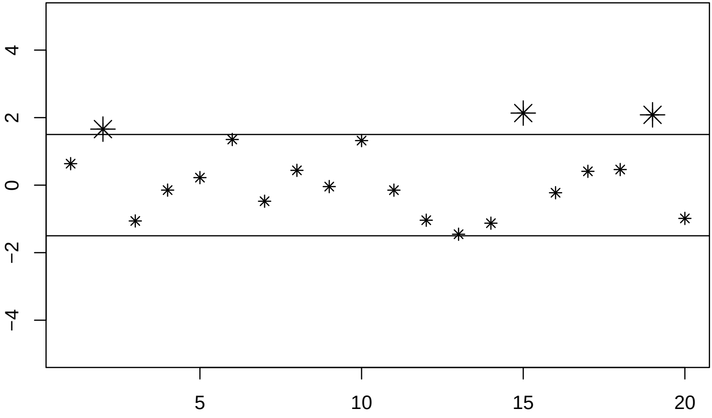


CI after selection

Let

$$X \sim \mathcal{N}(\theta, I_m) \in \mathbb{R}^m, \quad \theta \in \mathbb{R}^m,$$

90% CI for each θ_i after selection

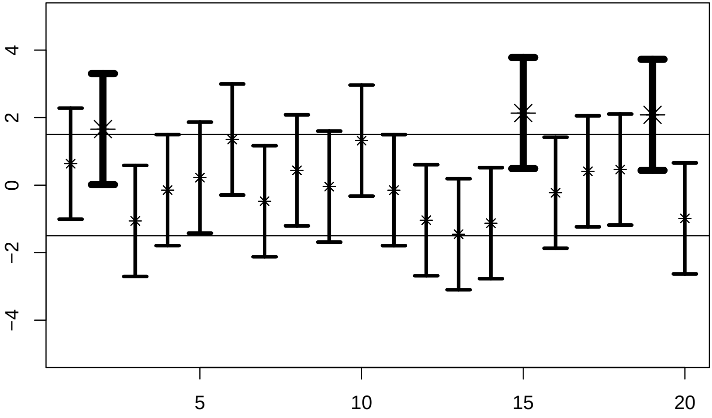


CI after selection

Let

$$X \sim \mathcal{N}(\theta, I_m) \in \mathbb{R}^m, \quad \theta \in \mathbb{R}^m,$$

90% CI for each θ_i after selection

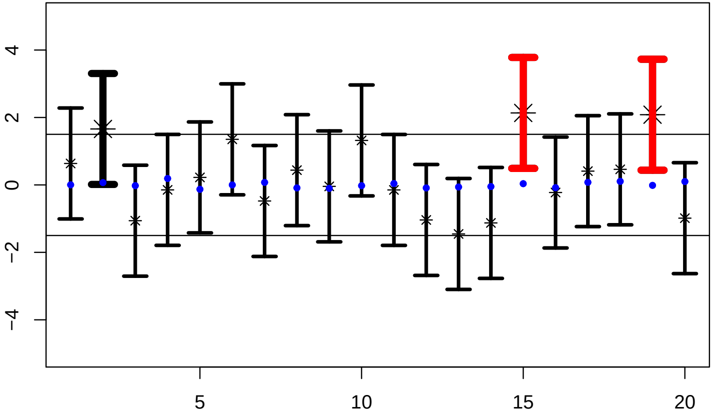


CI after selection

Let

$$X \sim \mathcal{N}(\theta, I_m) \in \mathbb{R}^m, \quad \theta \in \mathbb{R}^m,$$

90% CI for each θ_i after selection

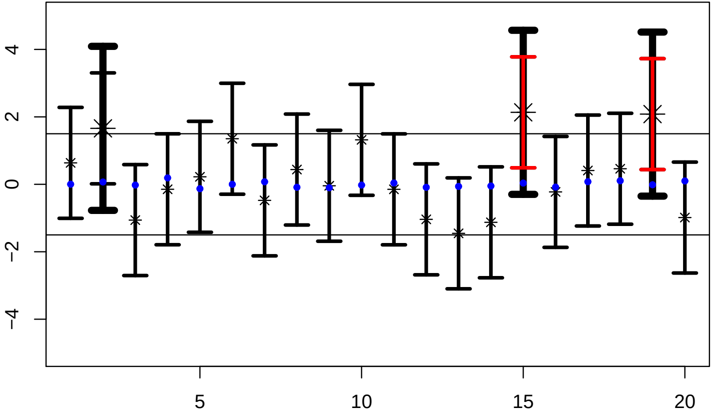


CI after selection

Let

$$X \sim \mathcal{N}(\theta, I_m) \in \mathbb{R}^m, \quad \theta \in \mathbb{R}^m,$$

90% CI for each θ_i after selection



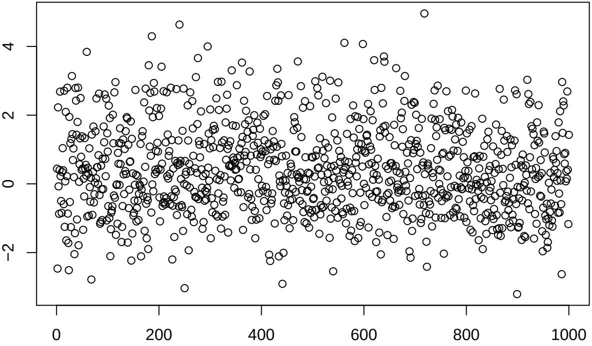
A solution : [Benjamini and Yekutieli (2005)] take $1 - 0.1|R|/m$ (or so)

Estimating true null quantity

Let

$$X \sim \mathcal{N}(\theta, I_m) \in \mathbb{R}^m, \quad \theta \in \mathbb{R}_+^m,$$

Parameter $m_0(\theta) = \#$zeros in θ (true null number)

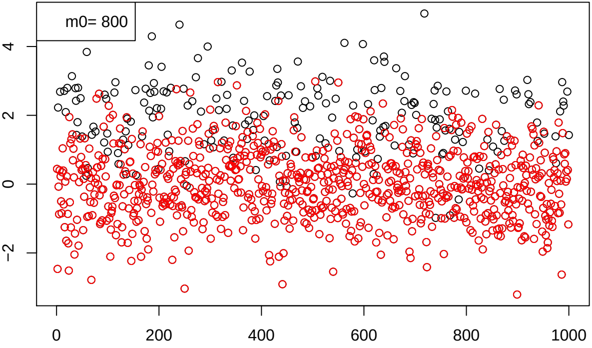


Estimating true null quantity

Let

$$X \sim \mathcal{N}(\theta, I_m) \in \mathbb{R}^m, \quad \theta \in \mathbb{R}_+^m,$$

Parameter $m_0(\theta) = \#\text{zeros in } \theta$ (true null number)

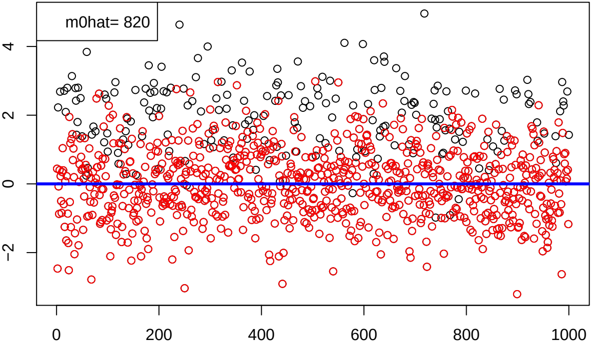


Estimating true null quantity

Let

$$X \sim \mathcal{N}(\theta, I_m) \in \mathbb{R}^m, \quad \theta \in \mathbb{R}_+^m,$$

Parameter $m_0(\theta) = \#$zeros in θ (true null number)



$$\hat{m}_0 = 2\#\{i : X_i \leq 0\} \geq 2\#\{i : \theta_i = 0, X_i \leq 0\} \approx m_0.$$

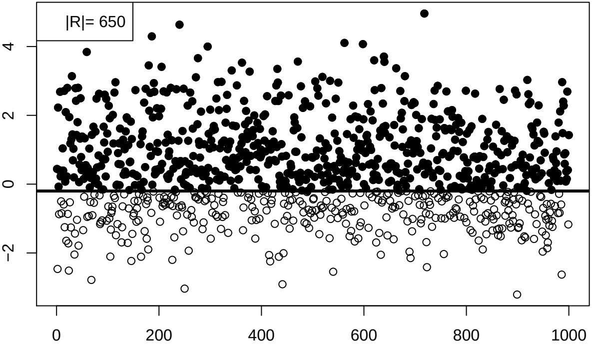
[Storey (2002)]

Estimating m_0 after selection

Let

$$X \sim \mathcal{N}(\theta, I_m) \in \mathbb{R}^m, \quad \theta \in \mathbb{R}_+^m,$$

Parameter $V(R) = \#$zeros in θ in selected R (false positives in R)

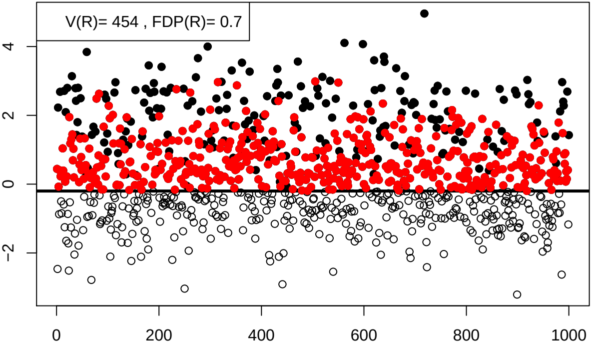


Estimating m_0 after selection

Let

$$X \sim \mathcal{N}(\theta, I_m) \in \mathbb{R}^m, \quad \theta \in \mathbb{R}_+^m,$$

Parameter $V(R) = \#$zeros in θ in selected R (false positives in R)

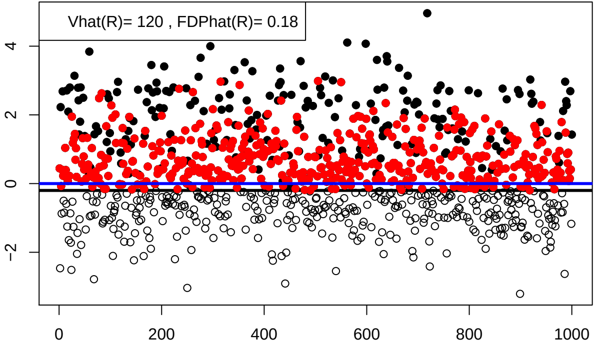


Estimating m_0 after selection

Let

$$X \sim \mathcal{N}(\theta, I_m) \in \mathbb{R}^m, \quad \theta \in \mathbb{R}_+^m,$$

Parameter $V(R) = \#$zeros in θ in selected R (false positives in R)

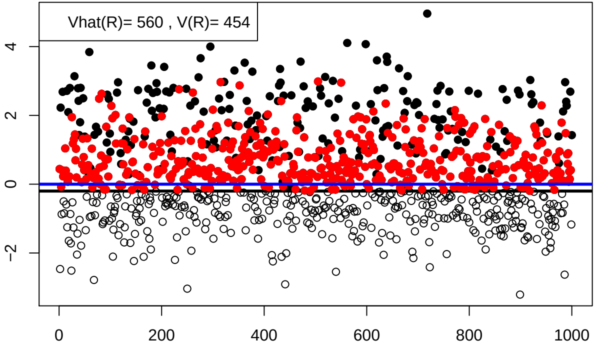


$\hat{V}(R) = 2\#\{i \in R : X_i \leq 0\}$ fails

A basic idea

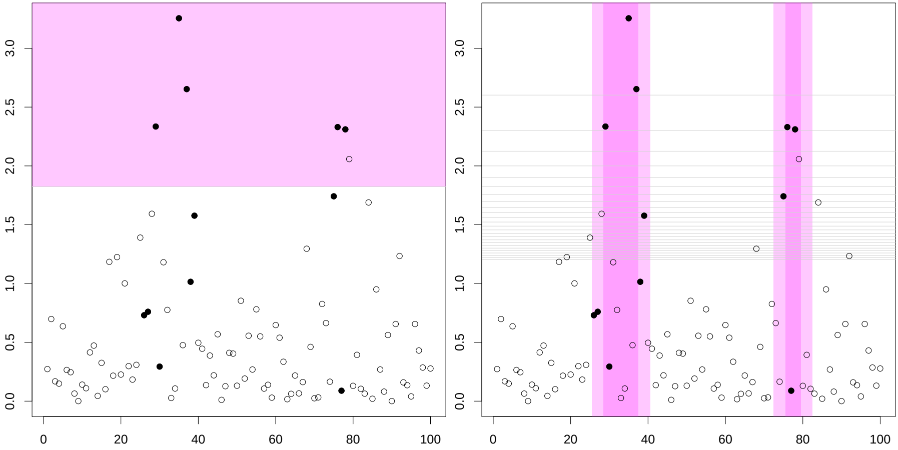
$$\begin{aligned} V(R) &= \#\{i \in R : \theta_i = 0\} \\ &= \#\{i \in R : \theta_i = 0, X_i \leq 0\} + \#\{i \in R : \theta_i = 0, X_i > 0\} \\ &\leq \#\{i \in R : X_i \leq 0\} + \#\{i \in R : \theta_i = 0, X_i > 0\} \\ &\leq \#\{i \in R : X_i \leq 0\} + \#\{i : \theta_i = 0, X_i > 0\} \\ &\approx \#\{i \in R : X_i \leq 0\} + m/2 =: \overline{V}(R) \end{aligned}$$

A basic idea



$$\overline{V}(R) = \#\{i \in R : X_i \leq 0\} + m/2$$
$$= \#\{i \in R : X_i \leq 0\} + |R|/2 \, \frac{m}{|R|}$$

What is *R*?



R from the data in any possible way

Aim

Observe $X \sim P$ with parameter $\theta = \theta(P) \in \mathbb{R}^m$.

Number of false positives in $R \subset \{1, \ldots, m\}$:

$$V(R) = |R \cap \mathcal{H}_0|, \quad \mathcal{H}_0 = \{i \,:\, \theta_i = 0\}.$$

Post hoc bound

$\overline{V}(\cdot) \in \mathbb{N}$, such that for all P,

$$\mathbf{P}\left(\forall R \subset \{1, \ldots, m\} \,:\, V(R) \leq \overline{V}(R)\right) \geq 1 - \alpha$$

- agnostic method on R
- desirable to have sharp $\overline{V}(R)$ for R containing large X_i's
- take reference sets $(R_k)_k$ making only few false discoveries

Aim

Observe $X \sim P$ with parameter $\theta = \theta(P) \in \mathbb{R}^m$.

Number of false positives in $R \subset \{1, \ldots, m\}$:

$$V(R) = |R \cap \mathcal{H}_0|, \quad \mathcal{H}_0 = \{i \,:\, \theta_i = 0\}.$$

Post hoc bound

$\overline{V}(\cdot) \in \mathbb{N}$, such that for all P,

$$\mathbf{P}\left(\forall R \subset \{1, \ldots, m\} \,:\, V(R) \leq \overline{V}(R)\right) \geq 1 - \alpha$$

- agnostic method on R
- desirable to have sharp $\overline{V}(R)$ for R containing large X_i's
- take reference sets $(R_k)_k$ making only few false discoveries

Method

JER control

$\mathfrak{R} = \{R_k\}_k$ reference family such that

$$JER(\mathfrak{R}) = \mathbf{P}(\exists k \; : \; V(R_k) \geq k) \leq \alpha$$

That is, $\mathcal{E} = \{\forall k \; : \; |R_k \cap \mathcal{H}_0| \leq k - 1\}$ is of proba $\geq 1 - \alpha$.

Lemma (interpolation)

On the event $\mathcal{E}$, $\forall R$,

$$V(R) \leq \overline{V}(R) = \min_k \{|R_k^c \cap R| + k - 1\}$$

- JER control offers post hoc bound

Method

JER control

$\mathfrak{R} = \{R_k\}_k$ reference family such that

$$JER(\mathfrak{R}) = \mathbf{P}(\exists k \ : \ V(R_k) \geq k) \leq \alpha$$

That is, $\mathcal{E} = \{\forall k \ : \ |R_k \cap \mathcal{H}_0| \leq k - 1\}$ is of proba $\geq 1 - \alpha$.

Lemma (interpolation)

On the event $\mathcal{E}$, $\forall R$,

$$V(R) \leq \overline{V}(R) = \min_k \{|R_k^c \cap R| + k - 1\}$$

- JER control offers post hoc bound

Simes inequality

Proposition [Simes (1986)]

If $(p_i, 1 \leq i \leq m)$ available with $(p_i, i \in \mathcal{H}_0)$ i.i.d. $U(0,1)$,

$$\mathbf{P}(\exists k \ : \ p_{(k:\mathcal{H}_0)} \leq \alpha k/m) \leq \alpha.$$

we have $\leq$ if positive dependence [Benjamini and Yekutieli (2001)]

Corollary

Simes reference family $\mathfrak{R}$ with $R_k = \{i \ : \ p_i \leq \alpha k/m\}$ satisfies

$$JER(\mathfrak{R}) = \mathbf{P}(\exists k \ : \ V(R_k) \geq k) \leq \alpha$$

and thus provides a post hoc bound ([Goeman and Solari (2011)]).

- Calibrated for independence only
- Why threshold $t_k \propto k$?

Simes inequality

Proposition [Simes (1986)]

If $(p_i, 1 \leq i \leq m)$ available with $(p_i, i \in \mathcal{H}_0)$ i.i.d. $U(0,1)$,

$$\mathbf{P}(\exists k \;:\; p_{(k:\mathcal{H}_0)} \leq \alpha k/m) \leq \alpha.$$

we have $\leq$ if positive dependence [Benjamini and Yekutieli (2001)]

Corollary

Simes reference family $\mathfrak{R}$ with $R_k = \{i \;:\; p_i \leq \alpha k/m\}$ satisfies

$$JER(\mathfrak{R}) = \mathbf{P}(\exists k \;:\; V(R_k) \geq k) \leq \alpha$$

and thus provides a post hoc bound ([Goeman and Solari (2011)]).

- Calibrated for independence only
- Why threshold $t_k \propto k$?

Simes inequality

Proposition [Simes (1986)]

If $(p_i, 1 \leq i \leq m)$ available with $(p_i, i \in \mathcal{H}_0)$ i.i.d. $U(0,1)$,

$$\mathbf{P}(\exists k \ :\ p_{(k:\mathcal{H}_0)} \leq \alpha k/m) \leq \alpha.$$

we have $\leq$ if positive dependence [Benjamini and Yekutieli (2001)]

Corollary

Simes reference family $\mathfrak{R}$ with $R_k = \{i \ :\ p_i \leq \alpha k/m\}$ satisfies

$$JER(\mathfrak{R}) = \mathbf{P}(\exists k \ :\ V(R_k) \geq k) \leq \alpha$$

and thus provides a post hoc bound ([Goeman and Solari (2011)]).

- Calibrated for independence only
- Why threshold $t_k \propto k$?

JER control with λ-adjustment

- $X \sim \mathcal{N}(\theta, \Gamma) \in \mathbb{R}^m$, $\theta \in \mathbb{R}^m$, Γ known
- p-values: $p_i = 2\overline{\Phi}(|X_i|)$, $1 \leq i \leq m$
- Reference family: $\mathfrak{R}$ with $R_k = \{i \,:\, p_i \leq t_k(\lambda)\}$, some kernel $t_k(\lambda)$

$$\begin{aligned} \mathrm{JER}(\mathfrak{R}) &= \mathbf{P}(\exists k \,:\, p_{(k:\mathcal{H}_0)} \leq t_k(\lambda)) \\ &\leq \mathbf{P}_{Z\sim\mathcal{N}(0,\Gamma)}\left(\min_k \left\{ t_k^{-1}(2\overline{\Phi}(|Z|_{(k)}))\right\} \leq \lambda\right) \text{ known !} \end{aligned}$$

Method

Compute $\lambda(\alpha, \Gamma)$ with bound $\leq \alpha$ and use $t_k(\lambda(\alpha, \Gamma))$

- Linear kernel: $t_k(\lambda) = \lambda k/m$ (Simes under independence)
- Balanced kernel: such that the $t_k^{-1}(2\overline{\Phi}(|Z|_{(k)}))$'s are all $U(0,1)$

JER control with λ-adjustment

- $X \sim \mathcal{N}(\theta, \Gamma) \in \mathbb{R}^m$, $\theta \in \mathbb{R}^m$, Γ known
- p-values: $p_i = 2\overline{\Phi}(|X_i|)$, $1 \leq i \leq m$

- Reference family: $\mathfrak{R}$ with $R_k = \{i \,:\, p_i \leq t_k(\lambda)\}$, some kernel $t_k(\lambda)$

$$\mathrm{JER}(\mathfrak{R}) = \mathbf{P}(\exists k \,:\, p_{(k:\mathcal{H}_0)} \leq t_k(\lambda))$$
$$\leq \mathbf{P}_{Z\sim\mathcal{N}(0,\Gamma)}\left(\min_k \left\{t_k^{-1}(2\overline{\Phi}(|Z|_{(k)}))\right\} \leq \lambda\right) \text{ known !}$$

Method

Compute $\lambda(\alpha, \Gamma)$ with bound $\leq \alpha$ and use $t_k(\lambda(\alpha, \Gamma))$

- Linear kernel: $t_k(\lambda) = \lambda k/m$ (Simes under independence)
- Balanced kernel: such that the $t_k^{-1}(2\overline{\Phi}(|Z|_{(k)}))$'s are all $U(0,1)$

JER control with λ-adjustment

- $X \sim \mathcal{N}(\theta, \Gamma) \in \mathbb{R}^m$, $\theta \in \mathbb{R}^m$, Γ known
- p-values: $p_i = 2\overline{\Phi}(|X_i|)$, $1 \leq i \leq m$

- Reference family: $\mathfrak{R}$ with $R_k = \{i \, : \, p_i \leq t_k(\lambda)\}$, some kernel $t_k(\lambda)$

$$\mathrm{JER}(\mathfrak{R}) = \mathbf{P}(\exists k \, : \, p_{(k:\mathcal{H}_0)} \leq t_k(\lambda))$$
$$\leq \mathbf{P}_{Z\sim\mathcal{N}(0,\Gamma)}\left(\min_k \left\{t_k^{-1}(2\overline{\Phi}(|Z|_{(k)}))\right\} \leq \lambda\right) \text{ known !}$$

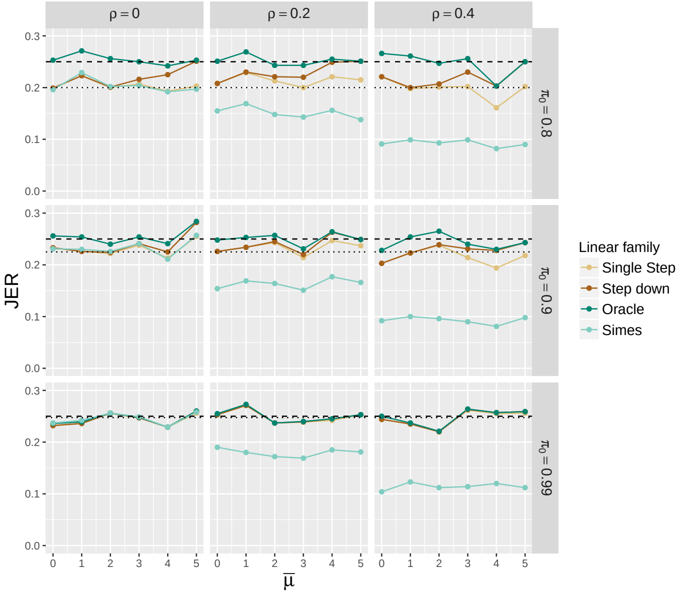
Method

Compute $\lambda(\alpha, \Gamma)$ with bound $\leq \alpha$ and use $t_k(\lambda(\alpha, \Gamma))$

- Linear kernel: $t_k(\lambda) = \lambda k/m$ (Simes under independence)
- Balanced kernel: such that the $t_k^{-1}(2\overline{\Phi}(|Z|_{(k)}))$'s are all $U(0,1)$

Illustration

- $\alpha = 0.25$
- $\Gamma = equi(\rho)$
- $m = 1000$
- $B = 1000$
- rep$= 1000$



Notions of power

Post hoc bound:

$$\mathbf{P}\left(\forall R \subset \{1,\ldots,m\} \;:\; |R\cap\mathcal{H}_0| \leq \overline{V}(R)\right) \geq 1-\alpha$$
$$\mathbf{P}\left(\forall R \subset \{1,\ldots,m\} \;:\; |R\cap\mathcal{H}_1| \geq \overline{S}(R)\right) \geq 1-\alpha,$$

for $\overline{S}(R) = |R| - \overline{V}(R)$ and $\mathcal{H}_1 = \mathcal{H}_0^c$.

Detection power: $R =$ all

For some procedure $\mathfrak{R}$, $\text{Pow}^*(\mathfrak{R}) = \mathbf{P}(\overline{S}(\{1,\ldots,m\}) > 0)$

Averaged power: R "random"

For some procedure $\mathfrak{R}$, $\text{Pow}(\mathfrak{R}) = \mathbf{E}\left(\frac{\overline{S}(R)}{|R\cap\mathcal{H}_1|} \;\Big|\; |R| > 0\right)$

Notions of power

Post hoc bound:

$$\mathbf{P}\left(\forall R \subset \{1,\dots,m\}\ :\ |R\cap\mathcal{H}_0| \leq \overline{V}(R)\right) \geq 1-\alpha$$
$$\mathbf{P}\left(\forall R \subset \{1,\dots,m\}\ :\ |R\cap\mathcal{H}_1| \geq \overline{S}(R)\right) \geq 1-\alpha,$$

for $\overline{S}(R) = |R| - \overline{V}(R)$ and $\mathcal{H}_1 = \mathcal{H}_0^c$.

Detection power: $R =$ all

For some procedure $\mathfrak{R}$, $\mathrm{Pow}^*(\mathfrak{R}) = \mathbf{P}(\overline{S}(\{1,\dots,m\}) > 0)$

Averaged power: R "random"

For some procedure $\mathfrak{R}$, $\mathrm{Pow}(\mathfrak{R}) = \mathbf{E}\left(\frac{\overline{S}(R)}{|R\cap\mathcal{H}_1|}\ \Big|\ |R| > 0\right)$

Notions of power

Post hoc bound:

$$\mathbf{P}\left(\forall R \subset \{1,\dots,m\} \,:\, |R \cap \mathcal{H}_0| \leq \overline{V}(R)\right) \geq 1-\alpha$$
$$\mathbf{P}\left(\forall R \subset \{1,\dots,m\} \,:\, |R \cap \mathcal{H}_1| \geq \overline{S}(R)\right) \geq 1-\alpha,$$

for $\overline{S}(R) = |R| - \overline{V}(R)$ and $\mathcal{H}_1 = \mathcal{H}_0^c$.

Detection power: $R =$ all

For some procedure $\mathfrak{R}$, $\text{Pow}^*(\mathfrak{R}) = \mathbf{P}(\overline{S}(\{1,\dots,m\}) > 0)$

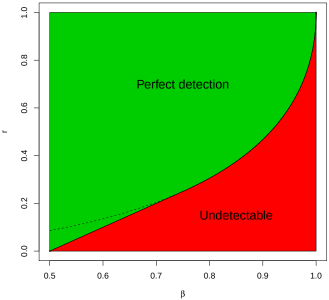
Averaged power: R "random"

For some procedure $\mathfrak{R}$, $\text{Pow}(\mathfrak{R}) = \mathbf{E}\left(\frac{\overline{S}(R)}{|R\cap\mathcal{H}_1|} \mid |R| > 0\right)$

Optimal detection

[Donoho and Jin (2004)]:

- Testing full null
- β sparsity parameter
- r effect size parameter
- Higher criticism attains the boundary



Theorem

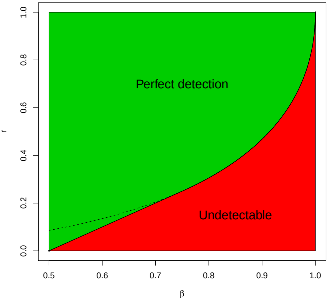
- for $r < \rho^\star(\beta)$, any JER controlling family has $\limsup_m \text{Pow}^*(\mathfrak{R}) \leq \alpha$;
- for $r > \rho^\star(\beta)$, balanced $\mathfrak{R}$ has $\text{Pow}^*(\mathfrak{R}) \to 1$.

Proof: balanced $\mathfrak{R}$ is a version of Higher criticism

Optimal detection

[Donoho and Jin (2004)]:

- Testing full null
- β sparsity parameter
- r effect size parameter
- Higher criticism attains the boundary



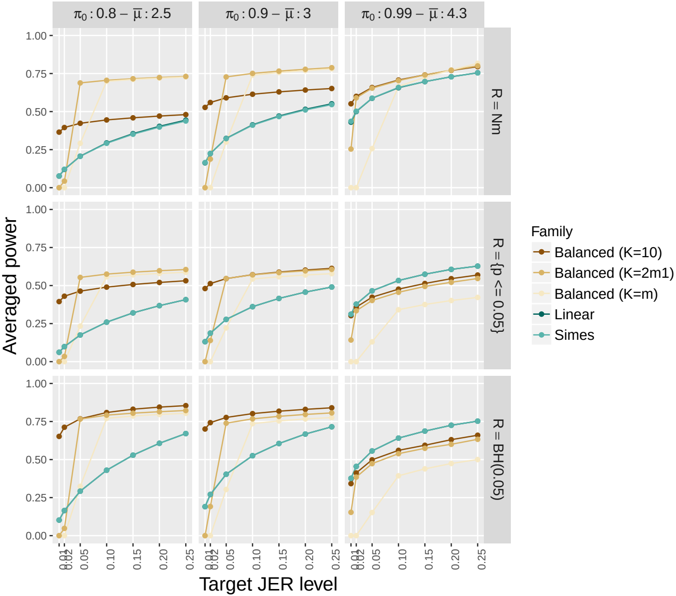
Theorem

- for $r < \rho^\star(\beta)$, any JER controlling family has $\limsup_m \text{Pow}^*(\mathfrak{R}) \leq \alpha$;
- for $r > \rho^\star(\beta)$, balanced $\mathfrak{R}$ has $\text{Pow}^*(\mathfrak{R}) \to 1$.

Proof: balanced $\mathfrak{R}$ is a version of Higher criticism

Illustration averaged power

- indep.
- $m = 1000$
- $B = 1000$
- rep$= 1000$



Outlook

Take home message

- Agnostic approach for false positive bound
- Price to pay: reference family (complexity K)

Todo

- Permutation (Γ unknown)
- Less conservative with structure constraints on R
- Multivariate test statistics

Advertising: ANR-16-CE40-0019 "Sanssouci"

- Postdoc position in Toulouse
- Worshop in Toulouse Feb 7-9, 2018

Outlook

Take home message

- Agnostic approach for false positive bound
- Price to pay: reference family (complexity K)

Todo

- Permutation (Γ unknown)
- Less conservative with structure constraints on R
- Multivariate test statistics

Advertising: ANR-16-CE40-0019 “Sanssouci"

- Postdoc position in Toulouse
- Worshop in Toulouse Feb 7-9, 2018

Outlook

Take home message

- Agnostic approach for false positive bound
- Price to pay: reference family (complexity K)

Todo

- Permutation (Γ unknown)
- Less conservative with structure constraints on R
- Multivariate test statistics

Advertising: ANR-16-CE40-0019 “Sanssouci"

- Postdoc position in Toulouse
- Worshop in Toulouse Feb 7-9, 2018