

FDP control in multivariate linear models using the bootstrap

Samuel Davenport, Bertrand Thirion, Pierre Neuvial

University of California, San Diego

August 31, 2022

Notation and general framework

Definition

Given $D, L \in \mathbb{N}$ and a finite set $\mathcal{V} \subset \mathbb{R}^D$, we define a **random field** on \mathcal{V} to be a random function $f : \mathcal{V} \rightarrow \mathbb{R}^L$. We will say that f has **dimension** L .

Definition

Given functions $\mu : \mathcal{V} \rightarrow \mathbb{R}^L$ and $\mathfrak{c} : \mathcal{V} \times \mathcal{V} \rightarrow \mathbb{R}$ we write $f \sim \mathcal{G}(\mu, \mathfrak{c})$ if f is a random field on \mathcal{V} with mean μ and covariance $\mathfrak{c}(u, v) = \text{cov}(f(u), f(v))$ and such that $\text{vec}(f)$ has a multivariate Gaussian distribution.

Suppose that we observe random fields $y_i : \mathcal{V} \rightarrow \mathbb{R}$, for $1 \leq i \leq n$ and some number of subjects n . At each voxel we assume that

$$Y_n(v) = X_n\beta(v) + E_n(v)$$

- $Y_n(v) = [y_1(v), \dots, y_n(v)]^T$: the response at each $v \in \mathcal{V}$
- $\beta : \mathcal{V} \rightarrow \mathbb{R}^p$: vector of parameters
- X_n : design matrix (which is itself random)
- $E_n = [\epsilon_1, \dots, \epsilon_n]^T$ - the noise - is an n -dimensional random field. We will assume that $(\epsilon_m)_{m \in \mathbb{N}}$ is an i.i.d sequence.

Testing contrasts

Then given contrasts, $c_1, \dots, c_L \in \mathbb{R}^p$ for some number of contrasts $L \in \mathbb{N}$, we are interested in testing the null hypotheses:

$$H_{0,l}(v) : c_l^T \beta(v) = 0$$

for $1 \leq l \leq L$ and each $v \in \mathcal{V}$.

We can test these using the t -statistic:

$$T_{n,l}(v) = \frac{c_l^T \hat{\beta}_n(v)}{\sqrt{\hat{\sigma}_n(v)^2 c_l^T (X_n^T X_n)^{-1} c_l}}. \quad (1)$$

For $n \in \mathbb{N}$, $1 \leq l \leq L$ and $v \in \mathcal{V}$ we can define p -values,

$$p_{n,l}(v) = 2(1 - \Phi_{n-r_n}(|T_{n,l}(v)|)) \quad (2)$$

where Φ_{n-r_n} is the CDF of a t -statistic with $n - r_n$ degrees of freedom.

- These are asymptotically valid
- Under an additional assumption of Gaussianity they are valid in the finite sample

Simultaneous coverage

- Let $\mathcal{H} = \{(l, v) : 1 \leq l \leq L \text{ and } v \in \mathcal{V}\}$ and $m = |\mathcal{H}|$.
- For $H \subseteq \mathcal{H}$, let $|H|$ denote the number of elements within H .
- let $\mathcal{N} \subset \mathcal{H}$ index the null hypotheses.

Given $0 < \alpha < 1$ we want,

$$V : \{H : H \subset \mathcal{H}\} \rightarrow \mathbb{N}$$

such that

$$\mathbb{P}(|S \cap \mathcal{N}| \leq V(S), \forall S \subset \mathcal{H}) \geq 1 - \alpha. \quad (3)$$

If (3) holds then, with probability $1 - \alpha$, simultaneously over all $S \subset \mathcal{H}$, $V(S)$ provides an upper bound on the number of false positives within S .

Joint Error Rate (JER)

Define the **joint error rate (JER)** of the collection $(R_k)_{1 \leq k \leq K} \subset \mathcal{H}$

$$\text{JER}((R_k(\lambda))_{1 \leq k \leq K}) := \mathbb{P}(|R_k \cap \mathcal{N}| > k - 1, \text{ some } 1 \leq k \leq K) \quad (4)$$

(Blanchard, Neuvial, Roquain, et al., 2020) showed that if

$$\text{JER}((R_k)_{1 \leq k \leq K}) \leq \alpha$$

then the bound $\bar{V}_\alpha : \{H : H \subset \mathcal{H}\} \rightarrow \mathbb{R}$, sending $S \subset \mathcal{H}$ to

$$\bar{V}_\alpha(S) = \min_{1 \leq k \leq K} (|S \setminus R_k| + k - 1) \wedge |S|, \quad (5)$$

satisfies (3) and thus provides an α -level bound over the number of false positives within each chosen rejection set.

Let $K \in \mathbb{N}$ and suppose we have a set of, strictly increasing and continuous template functions

$$t_k : [0, 1] \rightarrow \mathbb{R} \tag{6}$$

for each $1 \leq k \leq K$. Given $n \in \mathbb{N}$, define

$$R_k(\lambda) = \{(l, v) \in \mathcal{H} : p_{n,l}(v) \leq t_k(\lambda)\},$$

for each $\lambda \in [0, 1]$. We will refer to the collection $(R_k(\lambda))_{1 \leq k \leq K}$ as the canonical reference family. The simplest example is the linear template family i.e. $t_k(\lambda) = \frac{\lambda k}{m}$.

Controlling the JER

Let $p_{(k:\mathcal{N})}^n$ be the k th smallest p -value in the set $\{p_{n,l}(v) : (l, v) \in \mathcal{N}\}$ (and set $p_{(k:\mathcal{N})}^n = 1$ if $k > |\mathcal{N}|$).

Claim

For each $\lambda \in [0, 1]$,

$$JER((R_k(\lambda))_{1 \leq k \leq K}) = \mathbb{P}\left(\min_{1 \leq k \leq K \wedge |\mathcal{H}|} t_k^{-1}(p_{(k:\mathcal{N})}^n) \leq \lambda\right).$$

Bootstrapping in the Linear Model

Bootstrapping

Let

$$\hat{E}_n = Y_n - X_n \hat{\beta}_n = (I_n - X_n (X_n^T X_n)^{-1} X_n^T) E_n.$$

where I_n is the $n \times n$ identity matrix and

$$\hat{\beta}_n = (X_n^T X_n)^{-1} X_n^T Y_n = \beta + (X_n^T X_n)^{-1} X_n^T E_n.$$

Given $B \in \mathbb{N}$ for each $1 \leq b \leq B$, conditional on the data, $\hat{\epsilon}_1^b, \dots, \hat{\epsilon}_n^b$ are chosen independently with replacement from $\{\hat{E}_{n,1}, \dots, \hat{E}_{n,n}\}$ resulting in a combined random field $E_n^b = [\hat{\epsilon}_1^b, \dots, \hat{\epsilon}_n^b]^T$. Let

$$Y_n^b = X_n \hat{\beta}_n + E_n^b$$

and let

$$\hat{\beta}_n^b = (X_n^T X_n)^{-1} X_n^T Y_n^b$$

be the bootstrapped parameter estimates.

Assumption

a) For $n \in \mathbb{N}$, $X_n = [x_1, \dots, x_n]^T$ for a sequence of i.i.d vectors $(x_n)_{n \in \mathbb{N}}$ in \mathbb{R}^p such that $\mathbb{E}[\|x_1\|^{5/2}] < \infty$ and whose multivariate density is bounded above.

b) $(\epsilon_n)_{n \in \mathbb{N}}$ is an i.i.d sequence of 1-dimensional random fields on \mathcal{V} which is independent of $(x_n)_{n \in \mathbb{N}}$ and such that $\max_{v \in \mathcal{V}} \mathbb{E}[\epsilon_1(v)^4] < \infty$ and $\min_{v \in \mathcal{V}} \text{var}(\epsilon_1(v)) > 0$.

Theorem

(Bootstrap convergence.) Suppose that $(X_m)_{m \in \mathbb{N}}$ and $(\epsilon_m)_{m \in \mathbb{N}}$ satisfy Assumption 1. Then conditional on $(X_m, Y_m)_{m \in \mathbb{N}}$, for almost every sequence $(X_m, Y_m)_{m \in \mathbb{N}}$, for each $1 \leq b \leq B$,

$$\sqrt{n}(\hat{\beta}_n^b - \hat{\beta}_n) \xrightarrow{d} \mathcal{G}(0, \mathbf{c}_\epsilon \Sigma_X^{-1}).$$

Consistent as

$$\sqrt{n}(\hat{\beta}_n - \beta) \xrightarrow{d} \mathcal{G}(0, \mathbf{c}_\epsilon \Sigma_X^{-1}). \quad (7)$$

- (Freedman, 1981) proved a version of this in 1D based on convergence in the Mallows metric.
- (Eck, 2018) extended this proof to the multivariate case.
- We have a (substantially simpler) proof based on the Lindeberg CLT.

Convergence of the bootstrapped t -statistics

Theorem

(*Bootstrap test-statistic convergence.*) Suppose that $(X_m)_{m \in \mathbb{N}}$ and $(\epsilon_m)_{m \in \mathbb{N}}$ satisfy Assumption 1 and, for each $1 \leq b \leq B$, let $T_n^b : \mathcal{V} \rightarrow \mathbb{R}$ be the L -dimensional random field on \mathcal{V} such that, for $1 \leq l \leq L$,

$$T_{n,l}^b = \frac{c_l^T (\hat{\beta}_n^b - \hat{\beta}_n)}{\hat{\sigma}_n^b \sqrt{c_l^T (X_n^T X_n)^{-1} c_l}}.$$

Then conditional on $(X_m, Y_m)_{m \in \mathbb{N}}$, for almost every sequence $(X_m, Y_m)_{m \in \mathbb{N}}$, for each $1 \leq b \leq B$,

$$T_n^b \xrightarrow{d} \mathcal{G}(0, \mathbf{c}')$$

as $n \rightarrow \infty$. In particular it follows that

$$T_n^b | \mathcal{N} \xrightarrow{d} \mathcal{G}(0, \mathbf{c}') | \mathcal{N}.$$

JER Control in the Linear Model

Bootstrapped quantile

Let $f_n : \{g : \mathcal{V} \rightarrow \mathbb{R}^L\} \rightarrow \mathbb{R}$ send

$$T \mapsto \min_{1 \leq k \leq K \wedge |\mathcal{H}|} t_k^{-1}(p_{(k:\mathcal{H})}^n(T))$$

For each $n, B \in \mathbb{N}$ and $0 < \alpha < 1$, define the α -quantile of the bootstrapped distribution of $f_n(T_n)$ as

$$\lambda_{\alpha,n,B}^* = \inf \left\{ \lambda : \frac{1}{B} \sum_{b=1}^B 1[f_n(T_n^b) \leq \lambda] \geq \alpha \right\}.$$

Theorem

Assume Assumption 1 holds and that $r_n = o(n)$.

$$\begin{aligned} \text{Then, } & \lim_{n \rightarrow \infty} \lim_{B \rightarrow \infty} \text{JER}((R_k(\lambda_{\alpha, n, B}^*))_{1 \leq k \leq K}) \\ &= \lim_{n \rightarrow \infty} \lim_{B \rightarrow \infty} \mathbb{P} \left(\min_{1 \leq k \leq K \wedge |\mathcal{H}|} t_k^{-1}(p_{(k:\mathcal{N})}^n) \leq \lambda_{\alpha, n, B}^* \right) \leq \alpha \end{aligned}$$

I.e. the joint error rate is asymptotically bounded at a level α .

Iterating a step down version of this procedure is available.

Under PRDS, for $0 < \alpha < 1$, the Simes inequality implies that

$$\mathbb{P}\left(\exists k \in \{1, \dots, m\} : p_{(k:\mathcal{N})}^n < \frac{\alpha k}{m}\right) \leq \frac{\alpha |\mathcal{N}|}{m}.$$

Thus defining the linear template family as $t_k(x) = \frac{xk}{m}$, it follows that

$$\text{JER} = \mathbb{P}\left(\min_{1 \leq k \leq K \wedge |\mathcal{H}|} t_k^{-1}(p_{(k:\mathcal{N})}^n) \leq \alpha\right) \leq \alpha.$$

Thus \bar{V}_α (constructed using the sets $R_k(\alpha)$) is a valid post-hoc bound.

- This works best under independence as then the inequality becomes exact.
- PRDS may not hold (especially in the contrast cases);

(Rosenblatt, Finos, Weeda, Solari, & Goeman, 2018) introduced a version of this that estimates $|\mathcal{N}|$ using the hommel value h . It can be shown that under PRDS,

$$\text{JER} = \mathbb{P} \left(\min_{1 \leq k \leq K \wedge |\mathcal{H}|} t_k^{-1}(p_{(k:\mathcal{N})}^n) \leq \frac{\alpha m}{h} \right) \leq \alpha.$$

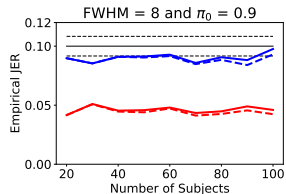
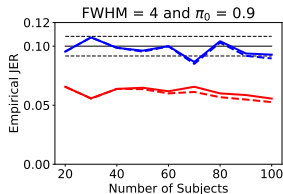
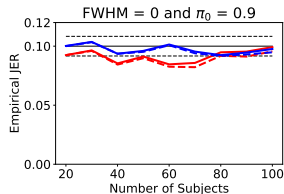
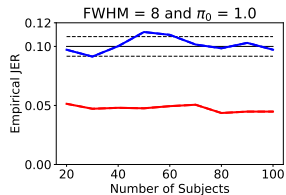
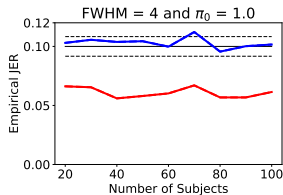
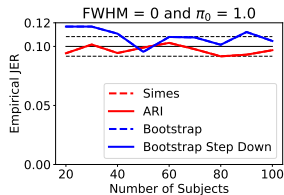
- The $\bar{V}_{\frac{\alpha m}{h}}$ (constructed using the sets $R_k(\frac{\alpha m}{h})$) is thus a valid post-hoc bound.
- Known as All Resolutions Inference or (ARI)
- It's the step down version of the Simes bound

Results

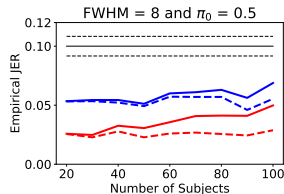
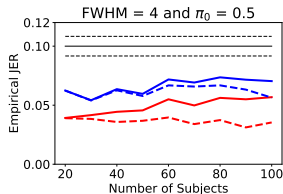
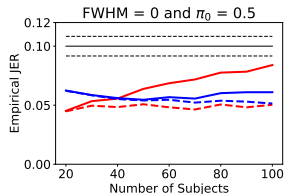
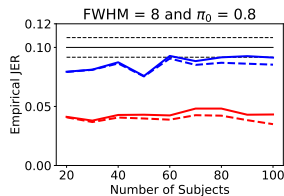
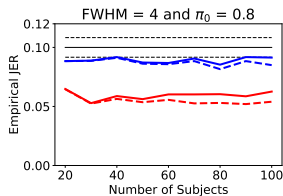
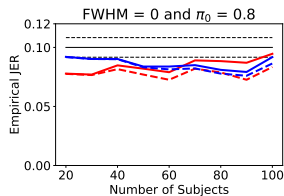
We ran 2D simulations to test the performance of the methods.

- 50×50 GRFs smoothed with $\text{FWHM} = 0, 4, 8$
- $N = \{20, 30, \dots, 100\}$ subjects
- randomly divided the subjects into 3 groups
- tested the difference between the first and the second and between the second and the third group at each pixel
- Randomly assigned a proportion $\pi_0 \in \{0.5, 0.8, 0.9, 1\}$ of the hypotheses to have non-zero mean 1.
- Compared the parametric and bootstrap methods.
- Bootstrap uses 100 bootstraps

Empirical JER



Empirical JER - continued



Given a set $R \subset \mathcal{H}$, define

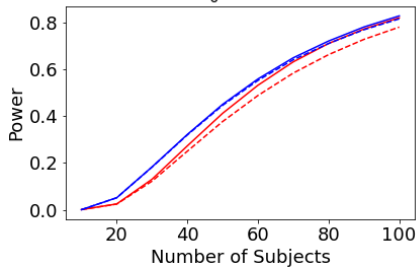
$$\text{Pow}(R) := \mathbb{E} \left[\frac{|R| - \bar{V}(R)}{|R \cap (\mathcal{H} \setminus \mathcal{N})|} \mathbb{1}_{|R \cap (\mathcal{H} \setminus \mathcal{N})| > 0} \right]$$

we take $R = \mathcal{H}$ (in this talk).

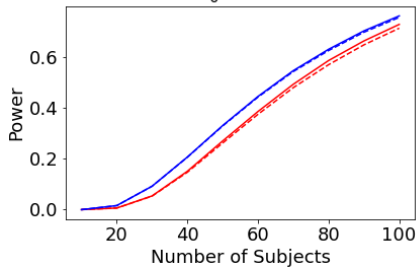
- This is a measure of the bounds on the true discovery proportion and so serves as a measure of power.
- Same notion of power as that of (Blanchard et al., 2020).
- Consider the same simulation setting where the FWHM = 4

Power - Results (In the FWHM = 4 setting)

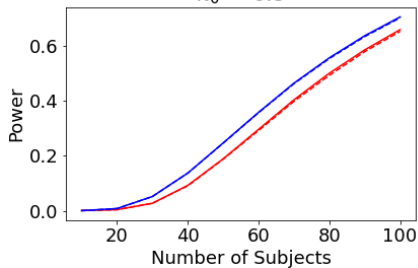
$\pi_0 = 0.5$



$\pi_0 = 0.8$



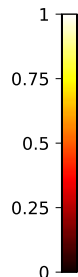
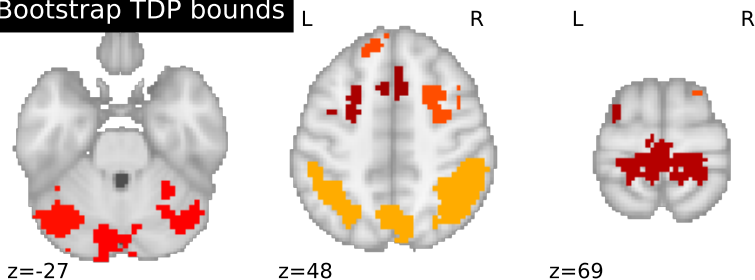
$\pi_0 = 0.9$



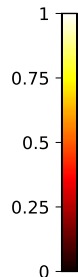
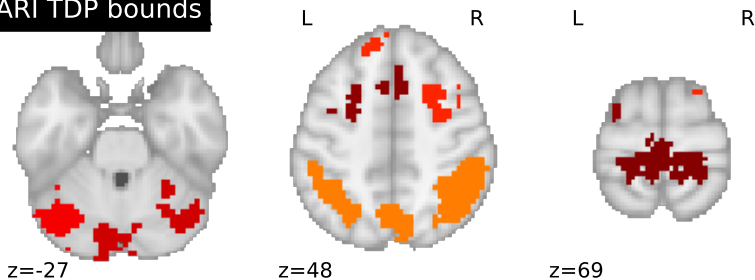
- fMRI data from 365 unrelated subjects from the HCP
- Subjects take a test the results of which are measured numerically.
- They also perform a working memory task
- At each voxel we fit a linear model of the fMRI data against: Age, Sex, Height, Weight, BMI, Blood pressure and the intelligence measure
- Test contrasts for Sex and intelligence
- Used 1000 bootstraps

fMRI data analysis

Bootstrap TDP bounds

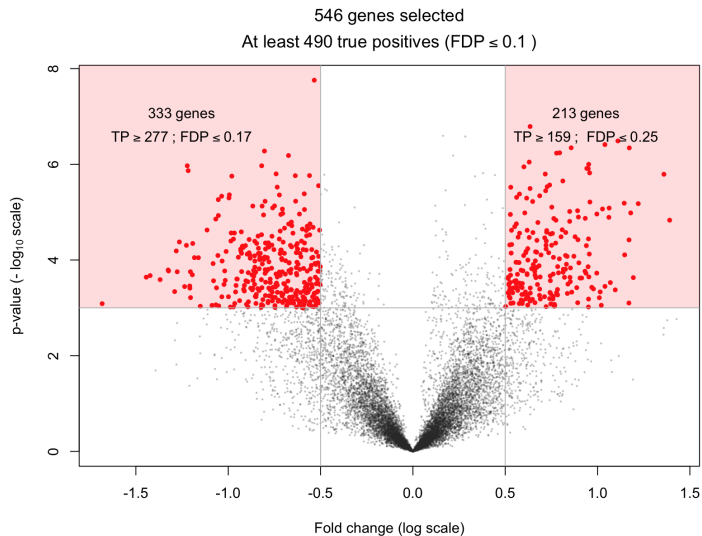


ARI TDP bounds



- Have genetics data from 135 subjects
- 12531 genes
- run a regression against some controlled covariates and lung function and considered a single contrast for lung function.

Volcano plot



- Using resampling approaches allows for large power gains when doing inference under dependence.
- Recommend non-parametric over parametric approaches in most cases
- The method is flexible and extends to other settings. I.e. other bootstrap settings.
- Code for implementation is available at github.com/sjdavenport/pyperm
- Pre-print available on arxiv (and from my website): (Davenport, Thirion, & Neuvial, 2022).

- Blanchard, G., Neuvial, P., Roquain, E., et al. (2020). Post hoc confidence bounds on false positives using reference families. *Annals of Statistics*, 48(3), 1281–1303.
- Davenport, S., Thirion, B., & Neuvial, P. (2022). Fdp control in multivariate linear models using the bootstrap. *arXiv preprint arXiv:2208.13724*.
- Eck, D. J. (2018). Bootstrapping for multivariate linear regression models. *Statistics & Probability Letters*, 134, 141–149.
- Freedman, D. A. (1981). Bootstrapping regression models. *The Annals of Statistics*, 9(6), 1218–1228.
- Rosenblatt, J. D., Finos, L., Weeda, W. D., Solari, A., & Goeman, J. J. (2018). All-resolutions inference for brain imaging. *Neuroimage*, 181, 786–796.