Controlling False Positive Rates in Neuroimaging

Samuel Davenport

University of California, San Diego

December 18, 2023

1. Robust FWER control in neuroimaging using Random Field Theory (work with Armin Schwartzman, Thomas E. Nichols and Fabian Telschow)
   - False Positives in Clustersize inference
   - Voxelwise Random Field theory
   - Resting State Validations
   - Gaussianization Transformation

2. FDP control in multivariate linear models using the residual bootstrap (work with Bertrand Thirion, Pierre Neuvial)
   - Simultaneous FDP control
   - Simulation results
   - Applications to fMRI and transcriptomics

Robust FWER control in neuroimaging using Random Field Theory (work with Armin Schwartzman, Thomas E. Nichols and Fabian Telschow)

False Positives in Clustersize inference

In 2016 (Eklund, Nichols, & Knutsson, 2016) showed that clusterwise RFT (a commonly used multiple testing method) had massively inflated false positive rates. However they actually showed that the opposite held true for voxelwise RFT.

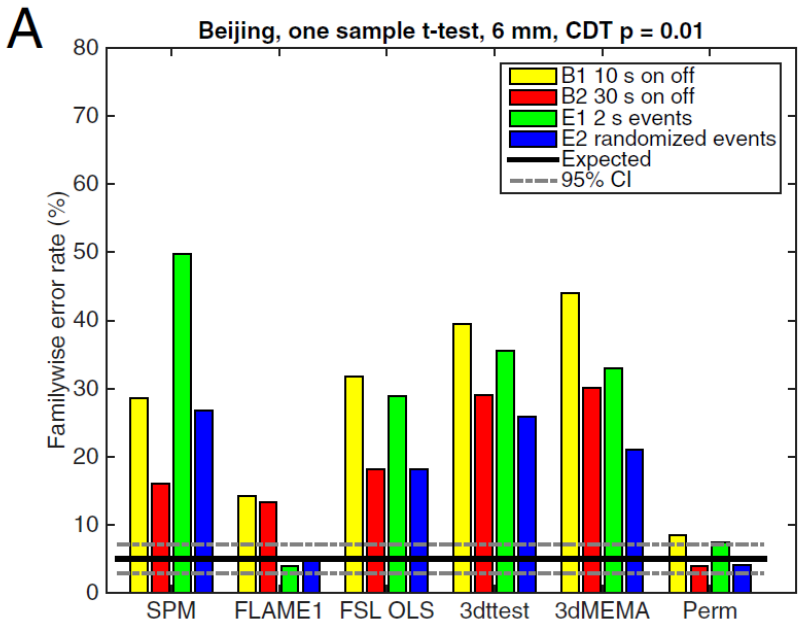
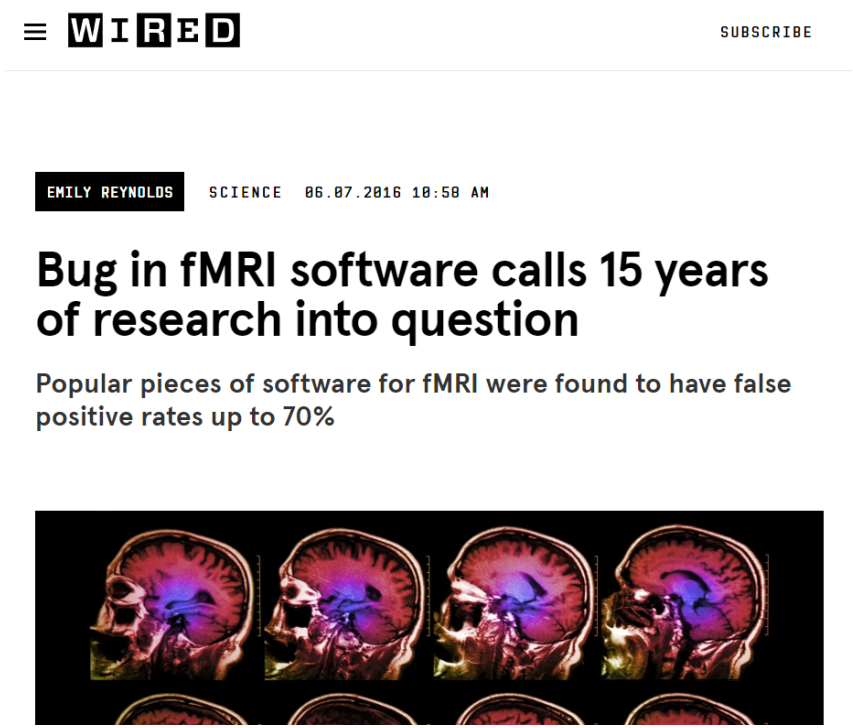


Figure 1: (Eklund et al., 2016) Figure 1: clusterwise RFT is inflated

Popular media

(Eklund et al., 2016) originally claimed that "These results question the validity of some 40,000 fMRI studies and may have a large impact on the interpretation of neuroimaging results"



Actually the results only affected 3500 papers (still a lot!).

fMRI data is non-stationary

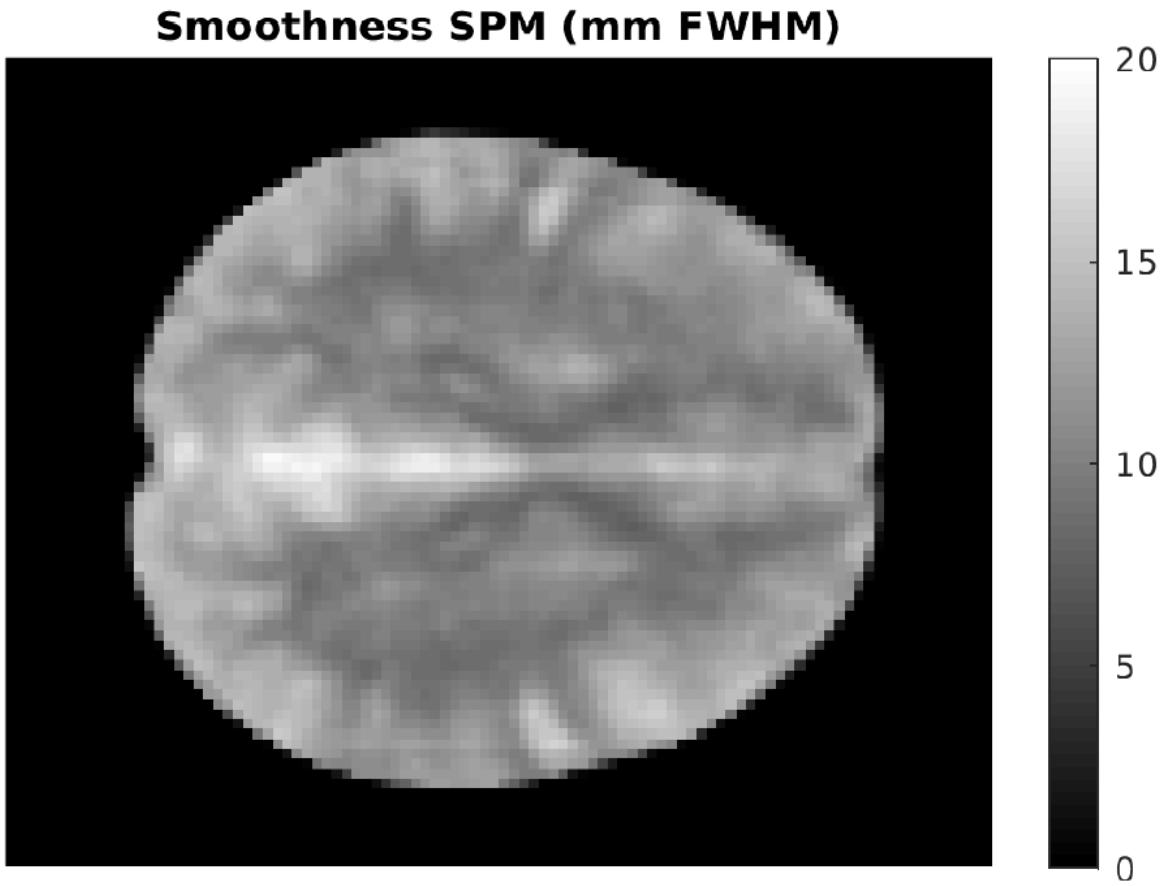
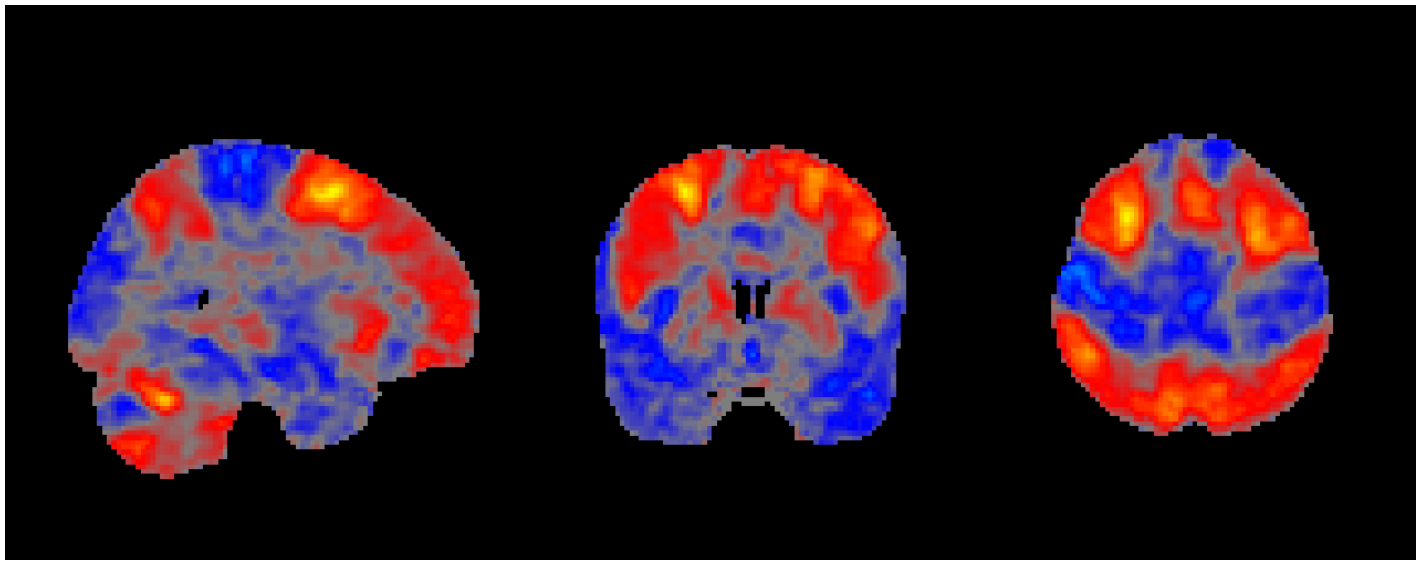


Figure 2: Smoothness varies across the brain (Eklund et al., 2016)

Taking non-stationarity into account allows us to improve these methods!

The problem

Given random images $Y_1, ..., Y_n : \mathcal{V} \to \mathbb{R}$ we want to test if $\mathbb{E}(Y_1(v)) = 0$, for $v \in \mathcal{V}$. To do so we first smooth and then combine these images to obtain a test-statistic T, and can control the FWER using $\mathbb{P}(\sup_{v \in \mathcal{V}} T(v) > u)$.



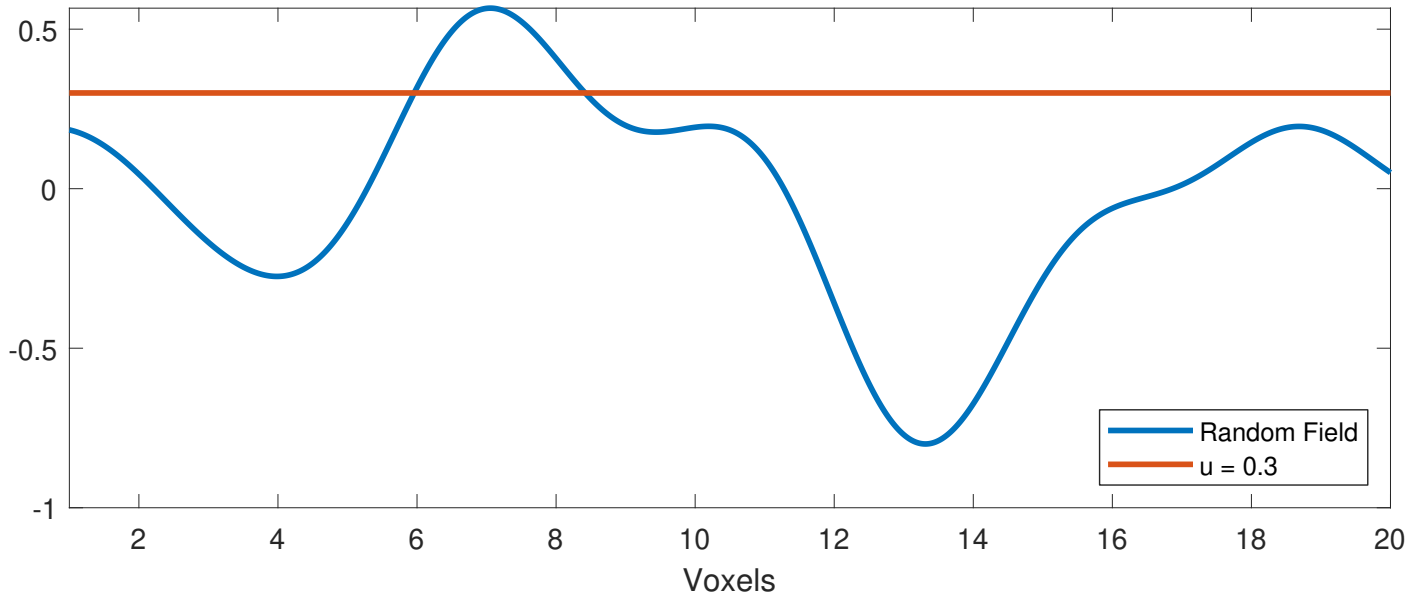
We will consider voxelwise inference.

Voxelwise RFT

Let $M_u(T)$ be the number of local maxima of T above a threshold u then assuming that T is twice differentiable,

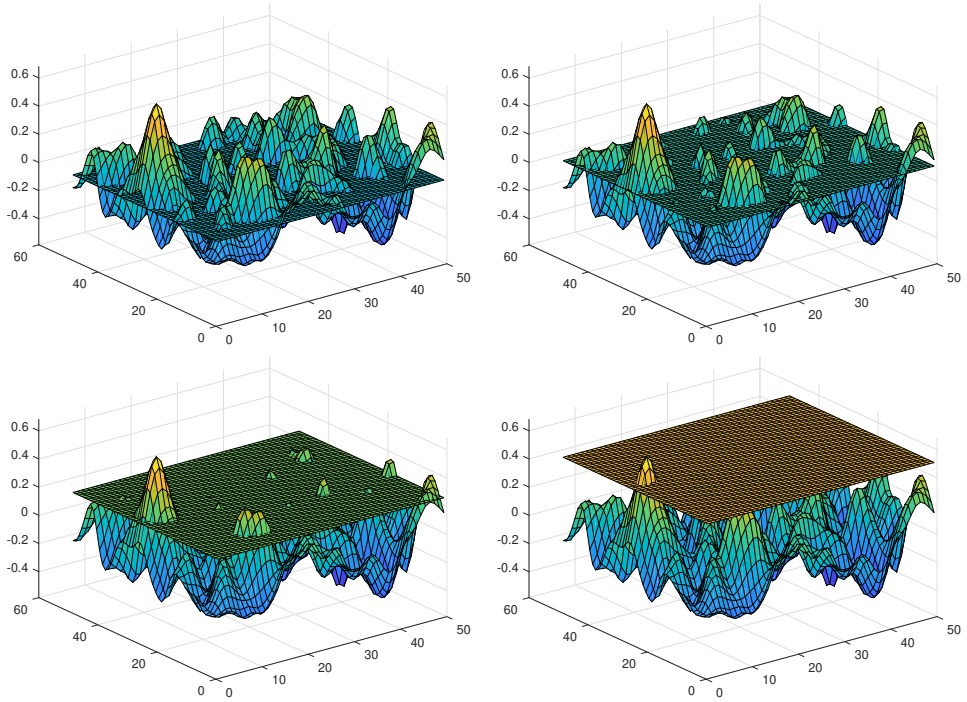
$$\mathbb{P}\left(\sup_{v\in\mathcal{V}} T(v) > u\right) = \mathbb{P}(M_u(T) \geq 1) \leq \mathbb{E}[M_u(T)]$$

because T exceeds u if and only if there is at least one local maxima above u.



The Euler Characteristic approximation

When there are no holes the Euler Char is the number of connected components i.e. clusters. At high thresholds it equals the number of local maxima.



This is useful because

$$\mathbb{E}[M_u(T)] \approx \mathbb{E}[\chi(\mathcal{A}_u(T))].$$

Performance of Traditional RFT

(Eklund et al., 2016) also showed that the opposite held true for voxelwise RFT (which we will introduce and aim to fix in this talk)

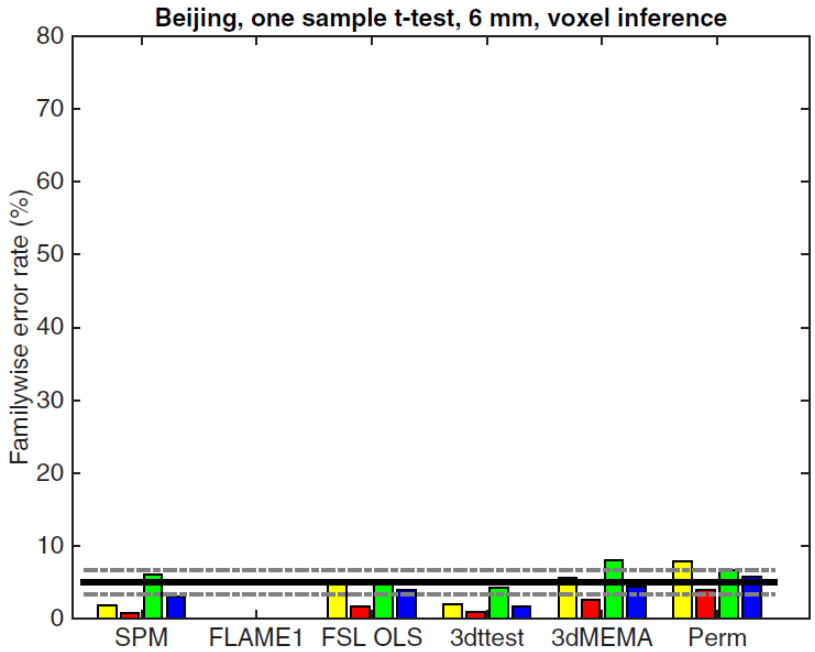


Figure 3: (Eklund et al., 2016) Figure 1c: voxelwise RFT is conservative.

Also observed in (Worsley, 2005) in simulations.

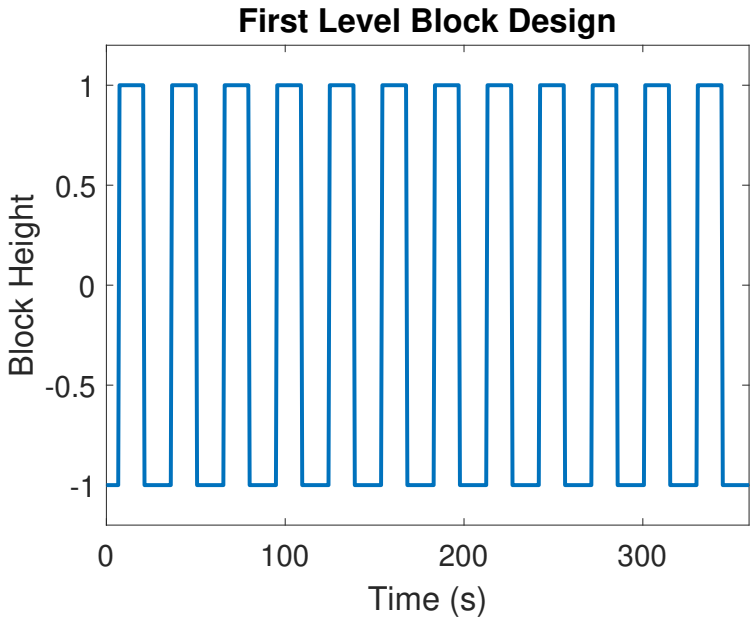
Assumptions of Random Field Theory in Neuroimaging

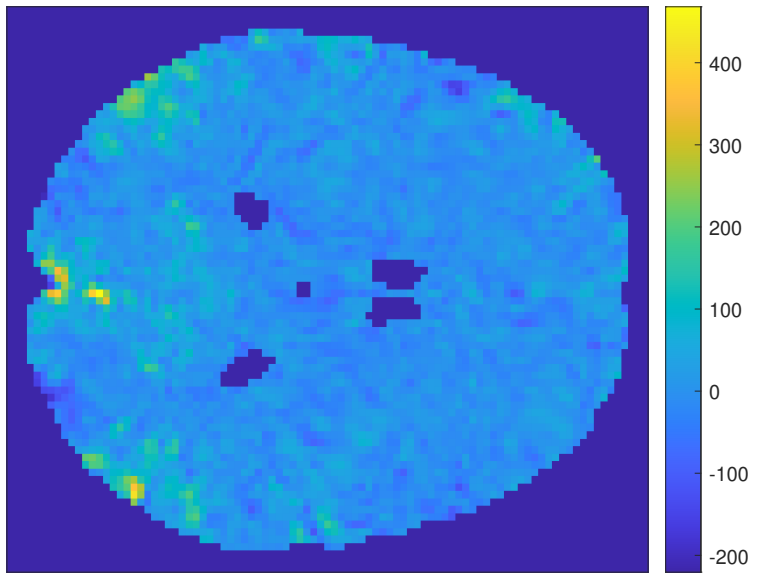
- Smoothness assumption (we showed this wasn't needed in (Telschow & Davenport, 2023))
- Stationarity (used for LKC calculation in fMRI software, but not needed: (Taylor et al., 2006))
- Gaussianity (questionable validity in fMRI)

These assumptions do not hold and so it is important to validate using realistic simulations.

Resting State Validation

We processed data from 7000 subjects from the UK biobank. Each subject has a time series of 490 images. Combine these into one contrast image using a block design at each voxel.





The results is 7000 contrast images (one for each subject). These have mean zero by construction as we randomized the blocks.

Resting state LKCs

We can estimate the true EEC distribution using the resting state data. For $j = 1, \ldots, 5000$ we draw $N = 20$ subjects with replacement and compute

$$u \mapsto \frac{1}{5000} \sum_{j=1}^{5000} \chi(\mathcal{A}_u(T_{j,N})).$$

were $T_{j,N}$ is the jth test-statistic based on the N random subsamples. We compare this to 3 EEC estimation approaches:

- Using non-stationary LKCs adapted to convolution fields - based on (Telschow & Davenport, 2023) and (Taylor 2006)).
- Two stationary approaches: Kiebel and Forman.

For other comparisons see (Telschow & Davenport, 2023).

Expected Euler characteristic curve - original data

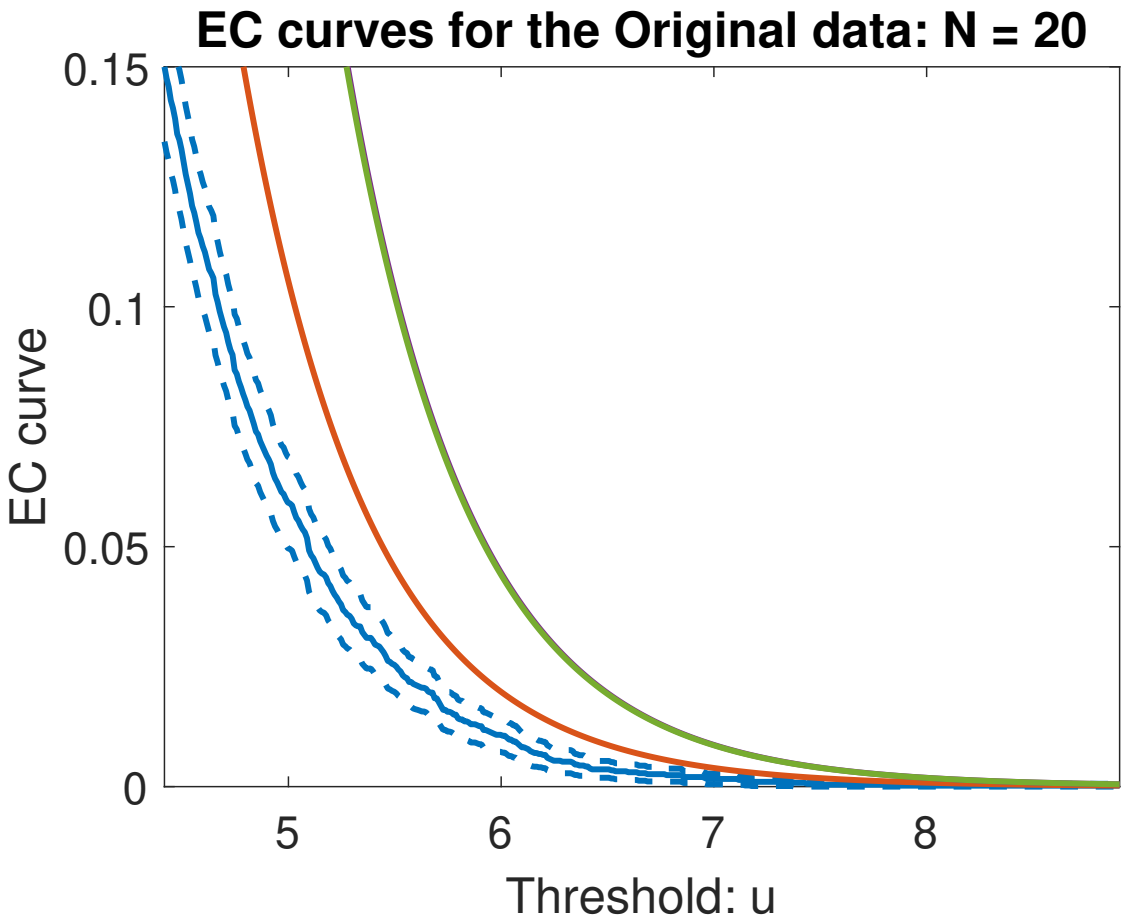
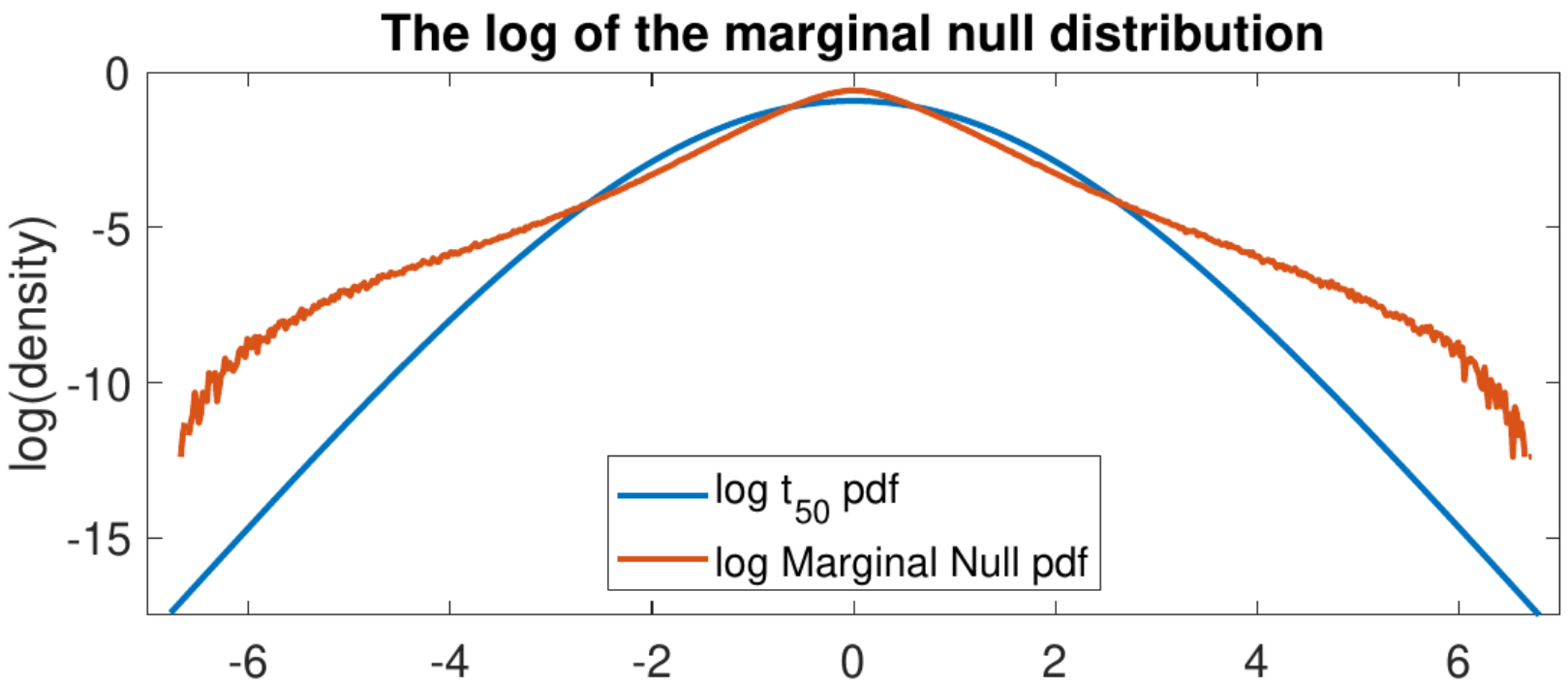


Figure 4: Blue: resting estimate EEC + 95% uncertainty, Red: SuRF LKC approximation. Green: Stationary LKC estimates (Kiebel + Forman).

Why doesn't it work? - fMRI data is highly non-Gaussian



Gaussianization transformation

More formally, at each voxel v we standardize and demean the fields Y_n:

$$Y_n^{S,D} = \frac{Y_n - \hat{\mu}}{\hat{\sigma}}. \tag{1}$$

Going back to the original data we standardize it (without demeaning) to yield:

$$Y_n^S = \frac{Y_n}{\hat{\sigma}}$$

and for each voxel v and subject n we compare $Y_n^S(v)$ to the null distribution to obtain a quantile

$$q_n(v) = \frac{1}{N|\mathcal{V}|} \sum_{n=1}^{N} \sum_{v' \in \mathcal{V}} 1[Y_n^S(v) \leq Y_n^{S,D}(v')].$$

The Gaussianized fields are then given by

$$Y_n^G(v) = \Phi^{-1}(q_n(v)) \quad v \in \mathcal{V},\ n = 1, \ldots, N.$$

Expected Euler characteristic curve - Gaussianized data

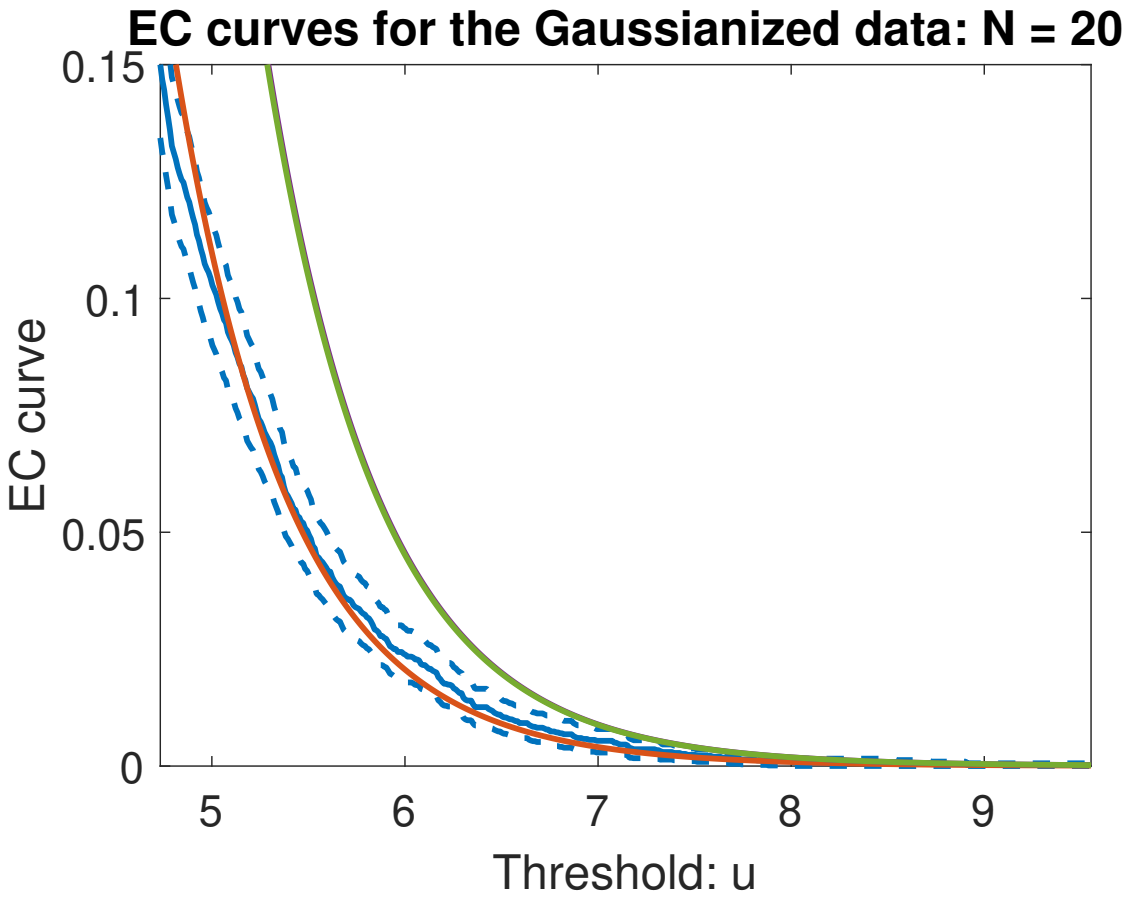


Figure 5: Blue: resting estimate EEC + 95% uncertainty, Red: SuRF LKC approximation. Green: Stationary LKC estimates (Kiebel + Forman).

Conclusions

- We need to be careful about imposing parametric assumptions (like stationarity)
- fMRI data is (highly) non-Gaussian and we should be careful about making this assumption in our models.
- Using a transformation can accelerate convergence of the CLT allowing for improved LKC estimation and control of the FWER.
- We use 7000 images instead of the between 100-200 samples used in (Eklund et al., 2016) meaning that we don't suffer from the same level of bias due to dependence between the draws.

FDP control in multivariate linear models using the residual bootstrap (work with Bertrand Thirion, Pierre Neuvial)

Linear Model

Suppose that we observe random images $Y_n : \mathcal{V} \to \mathbb{R}$, for $1 \leq n \leq N$ and some number of subjects n. At each voxel we assume that

$$\mathbf{Y}_N(v) = X_N \beta(v) + E_N(v)$$

- $\mathbf{Y}_N(v) = [Y_1(v), \ldots, Y_N(v)]^T$: the response at each $v \in \mathcal{V}$
- $\beta : \mathcal{V} \to \mathbb{R}^p$: vector of parameters
- X_N: design matrix (which is itself random)
- $E_N = [\epsilon_1, \ldots, \epsilon_N]^T$ - the noise - is an n-dimensional random image. We will assume that $(\epsilon_m)_{m \in \mathbb{N}}$ is an i.i.d sequence with finite variance.

Testing contrasts

Then given contrasts, $c_1, \ldots, c_L \in \mathbb{R}^p$ 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:

$$T_{N,l}(v) = \frac{c_l^T \hat{\beta}_N(v)}{\sqrt{\hat{\sigma}_N(v)^2 c_l^T (X_N^T X_N)^{-1} c_l}}. \tag{2}$$

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

$$p_{N,l}(v) = 2(1 - \Phi_{N-r_N}(|T_{N,l}(v)|)) \tag{3}$$

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

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,

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

such that

$$\mathbb{P}(|S \cap \mathcal{N}| \leq V(S),\ \forall S \subset \mathcal{H}) \geq 1 - \alpha. \tag{4}$$

If (4) holds then, with probability $1 - \alpha$, simultaneously over all $S \subset \mathcal{H}$, $V(S)$ provides a upper bound on the number of false positives within S.

Providing a bound

Let $p^N_{(k:\mathcal{N})}$ be the kth smallest p-value in the set $\{p_{N,l}(v) : (l,v) \in \mathcal{N}\}$ (and set $p^N_{(k:\mathcal{N})} = 1$ if $k > |\mathcal{N}|$). Let $K \in \mathbb{N}$ and suppose we have a set of, strictly increasing and continuous template functions

$$t_k : [0,1] \to \mathbb{R} \tag{5}$$

for each $1 \leq k \leq K$. (Blanchard et al 2020) showed that given λ s.t.

$$\mathbb{P}\left(\min_{1 \leq k \leq K \wedge |\mathcal{H}|} t_k^{-1}(p^N_{(k:\mathcal{N})}) \leq \lambda \right) \leq \alpha$$

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

$$\overline{V}_\alpha(S) = \min_{1 \leq k \leq K} (|S \setminus R_k(\lambda) + k - 1) \wedge |S|, \tag{6}$$

where $R_k = \{(l,v) \in \mathcal{H} : p_{N,l}(v) \leq t_k(\lambda)\}$ satisfies (4) and thus provides an α-level bound over the number of false positives within each chosen rejection set.

Controlling the probability

Theorem

Let $f : \left\{g : \mathcal{V} \to \mathbb{R}^L\right\} \to \mathbb{R}$ *send*

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

and for $N, B \in \mathbb{N}$ *and* $\alpha \in (0, 1)$, *let* $\lambda^*_{\alpha,N,B}$ *be the* α *quantile of* $f(T^b)$. *Assume that* $N - r_N \to \infty$. *Then*

$$\lim_{n\to\infty} \lim_{B\to\infty} \mathbb{P}\left(\min_{1 \leq k \leq K \wedge |\mathcal{H}|} t_k^{-1}(p^N_{(k:\mathcal{N})}) \leq \lambda^*_{\alpha,N,B}\right) \leq \alpha.$$

Main Result

Theorem

Under the assumptions of Theorem 2.1, for $0 < \alpha < 1$, *and* $H \subseteq \mathcal{H}$, *let*

$$\overline{V}_{\alpha,N,B}(H) = \min_{1 \leq k \leq K} (|H \setminus R_k(\lambda^*_{\alpha,N,B})| + k - 1) \wedge |H|.$$

Then $$\lim_{n \to \infty} \lim_{B \to \infty} \mathbb{P}\big(|H \cap \mathcal{N}| \leq \overline{V}_{\alpha,N,B}(H),\ \forall H \subseteq \mathcal{H}\big) \geq 1 - \alpha.$$

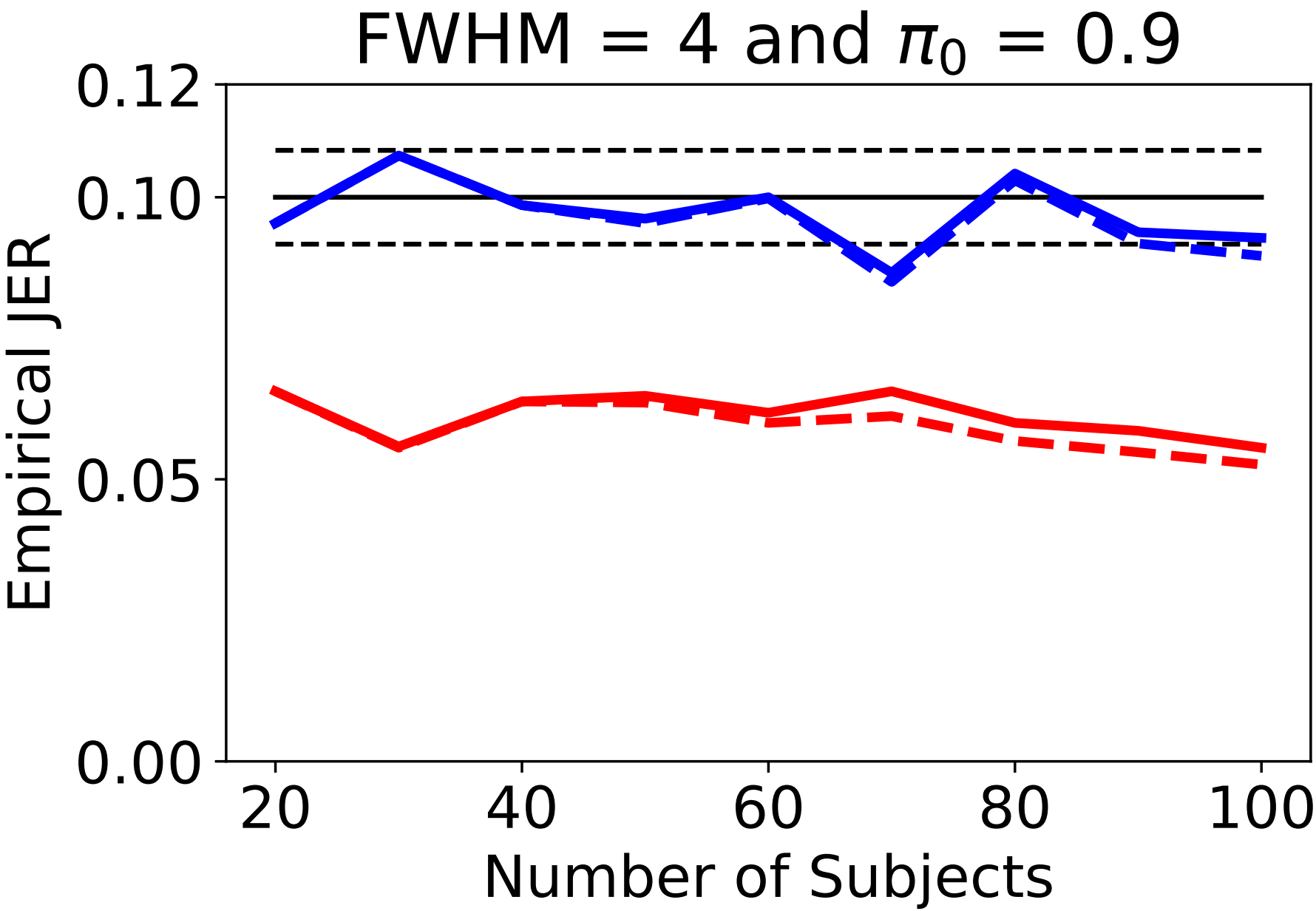
Iterating a step down version of this procedure is available.

Simulation description

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

- 50×50 GRFs smoothed with FWHM $= 4$
- $N = \{20, 30, \ldots, 100\}$ 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 $\pi_0 \in \{0.8, 0.9\}$ of the hypotheses to have non-zero mean 1.
- Compared the parametric and bootstrap methods.
- Bootstrap uses 1000 bootstraps
- Further simulations in (Davenport, Thirion, & Neuvial, 2022).

Empirical Error Rate



Power - definition

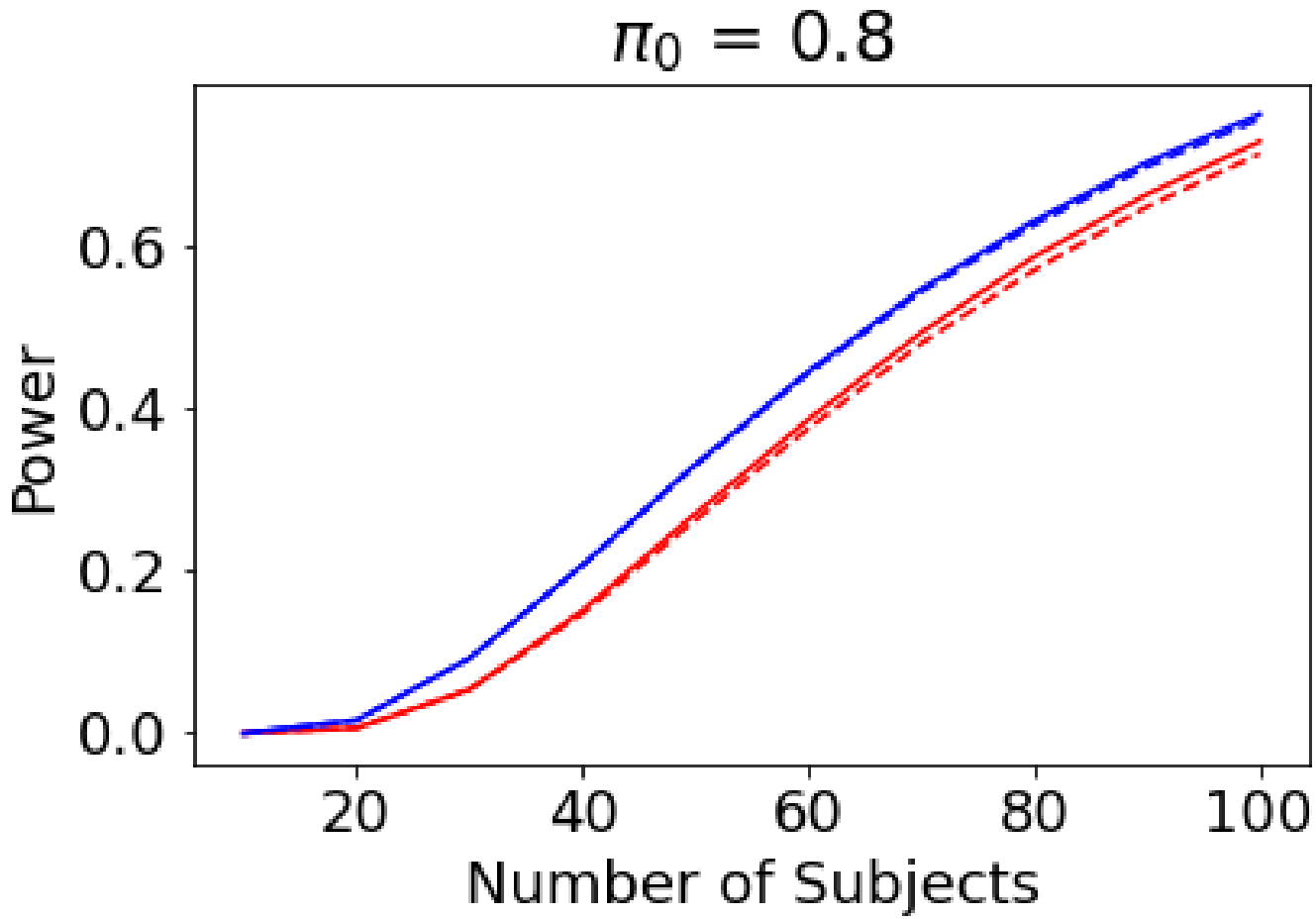
Given a set $R \subset \mathcal{H}$, define

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

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)

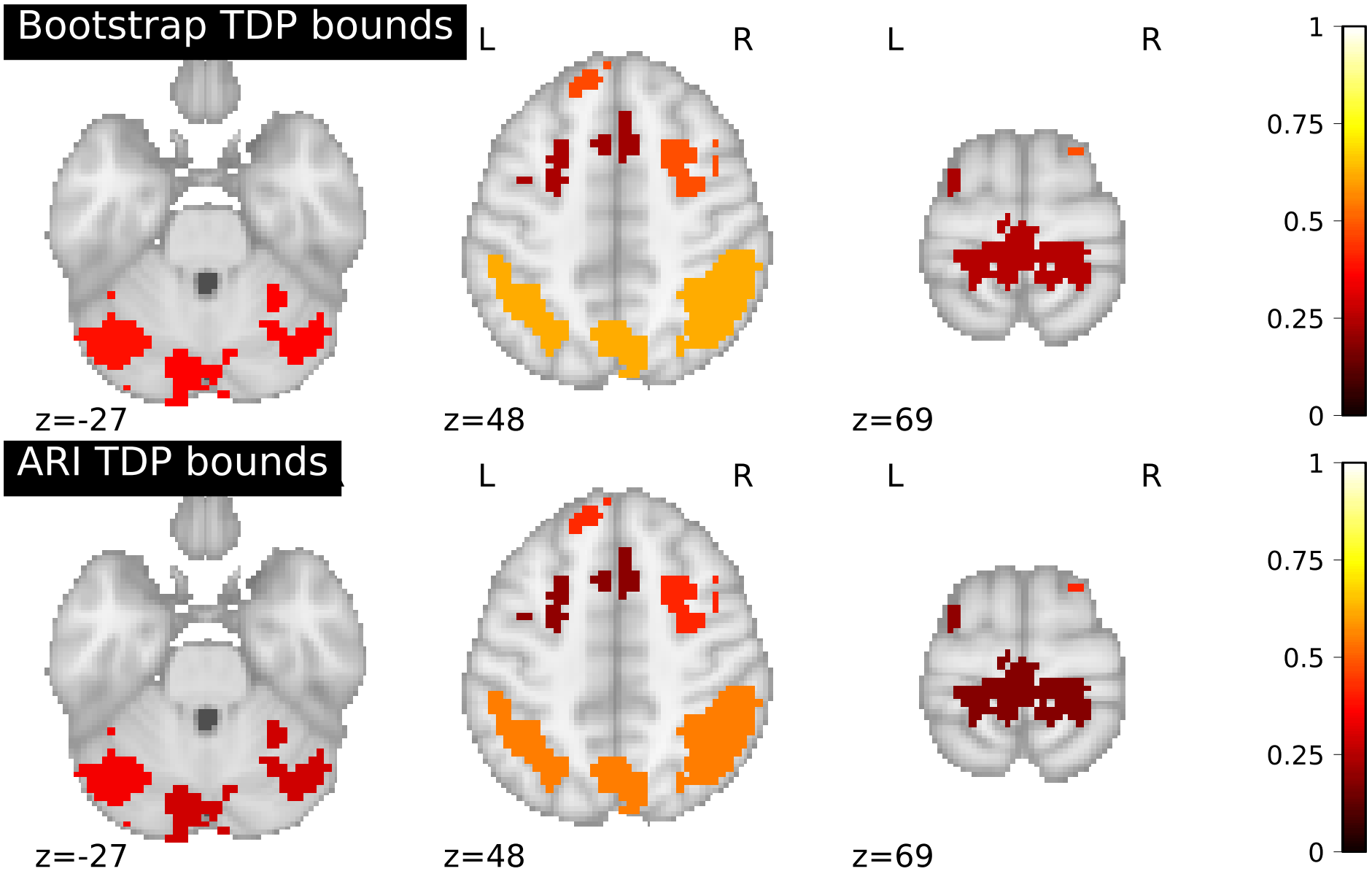
Power - Results (In the FWHM = 4 setting)



fMRI data model

- 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 intelligence
- Used 1000 bootstraps

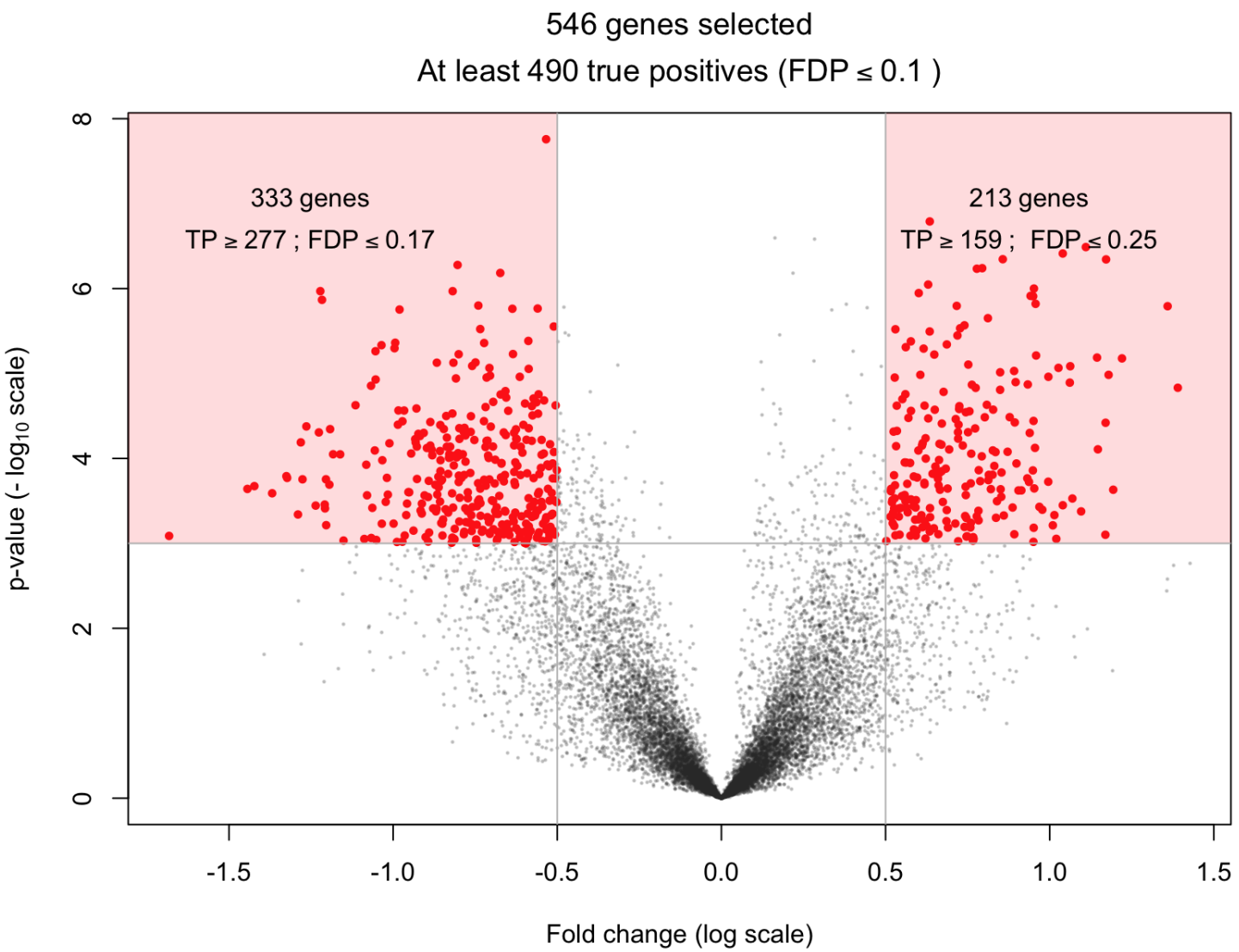
fMRI data analysis



Transcriptomic data analysis

- Have genetics data from 135 subjects from a COPD dataset
- 12531 genes
- run a regression against some controlled covariates and lung function and considered a single contrast for lung function.
- Used 1000 bootstraps.

Volcano plot



Conclusions Papers and Software

- This talk summarizes the work in three papers: (Telschow & Davenport, 2023), (Davenport, Schwarzman, Nichols, & Telschow, 2023).
- Software in MATLAB to perform RFT inference is available in the RFTtoolbox (Davenport & Telschow, 2023).
- Software in python to perform TDP inference available in the pyperm package.
- Slides available at `sjdavenport.github.io/talks`.

Bibliography

Davenport, S., Schwarzman, A., Nichols, T. E., & Telschow, F. (2023). Robust FWER control in neuroimaging using Random Field Theory: Riding the SuRF to continuous land Part 2.

Davenport, S., & Telschow, F. (2023). RFTtoolbox. Retrieved from `https://github.com/sjdavenport/RFTtoolbox`

Davenport, S., Thirion, B., & Neuvial, P. (2022). FDP control in multivariate linear models using the bootstrap. *arXiv preprint arXiv:2208.13724*.

Eklund, A., Nichols, T. E., & Knutsson, H. (2016). Cluster failure: Why fmri inferences for spatial extent have inflated false-positive rates. *Proceedings of the national academy of sciences*, *113*(28), 7900–7905.

Taylor, J. E., et al. (2006). A Gaussian Kinematic Formula. *The Annals of Probability*, *34*(1), 122–158.

Telschow, F., & Davenport, S. (2023). Riding the SuRF to Continuous Land - Part 1: precise FWER control for Gaussian Related Random Fields using Random Field Theory. *Preprint*.

FWER control on the Gaussianized data, $\alpha = 0.01$

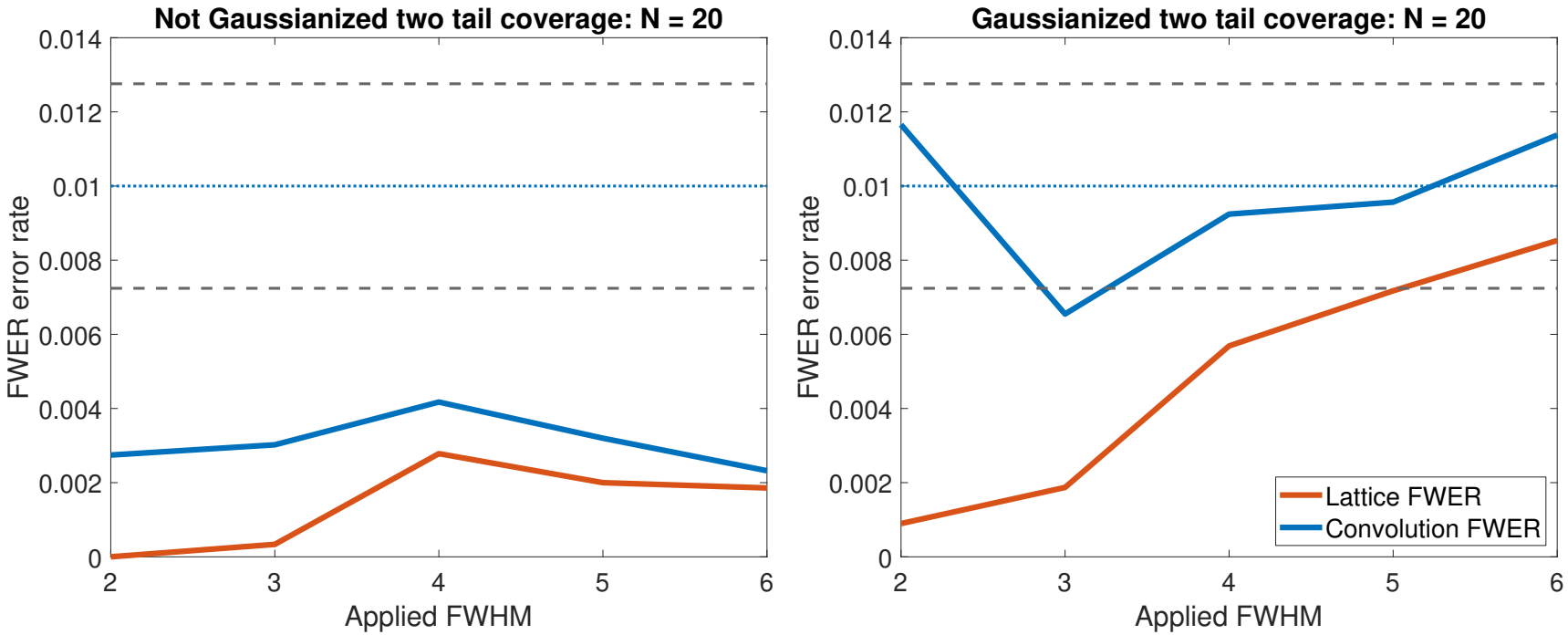
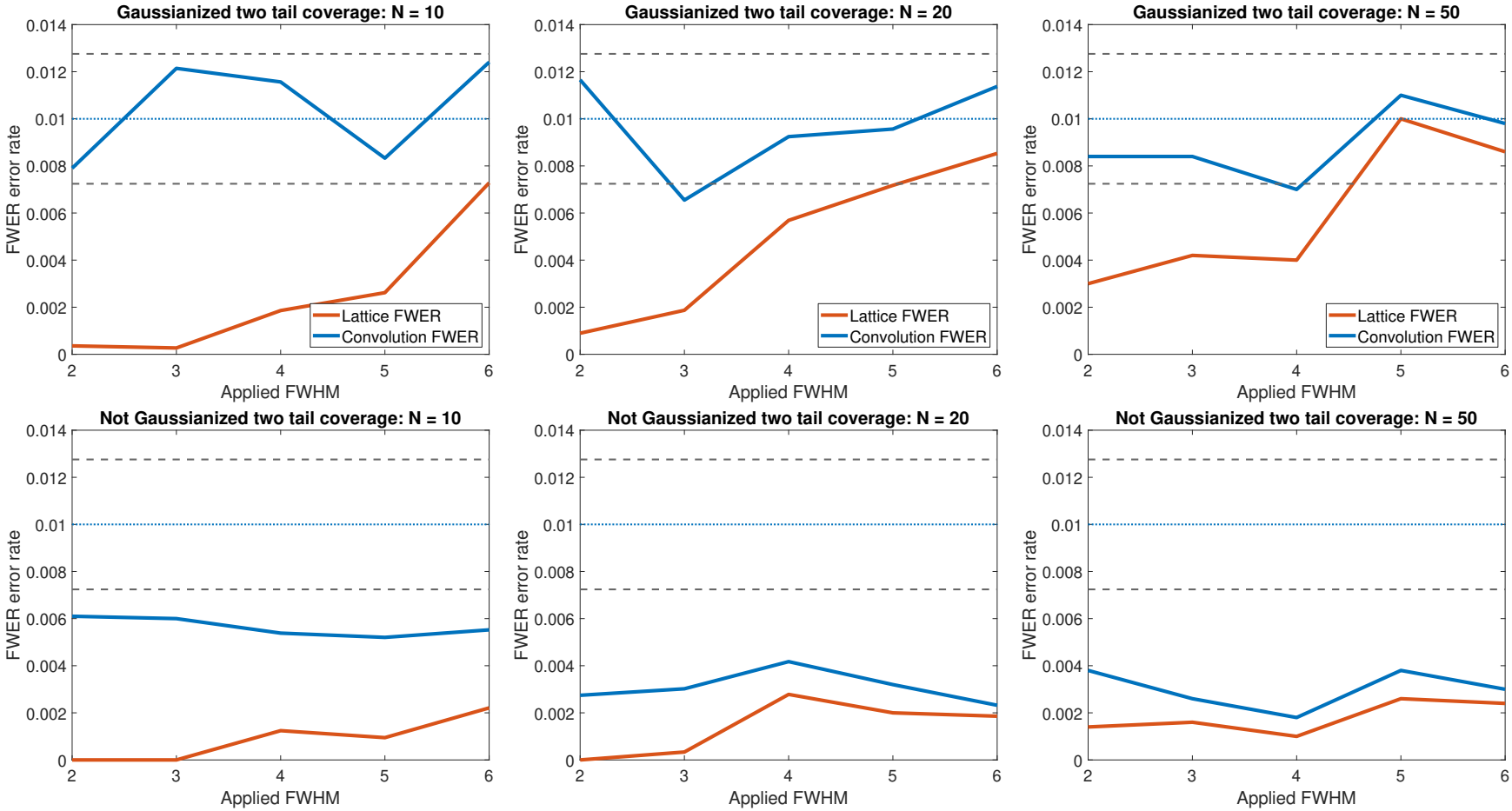
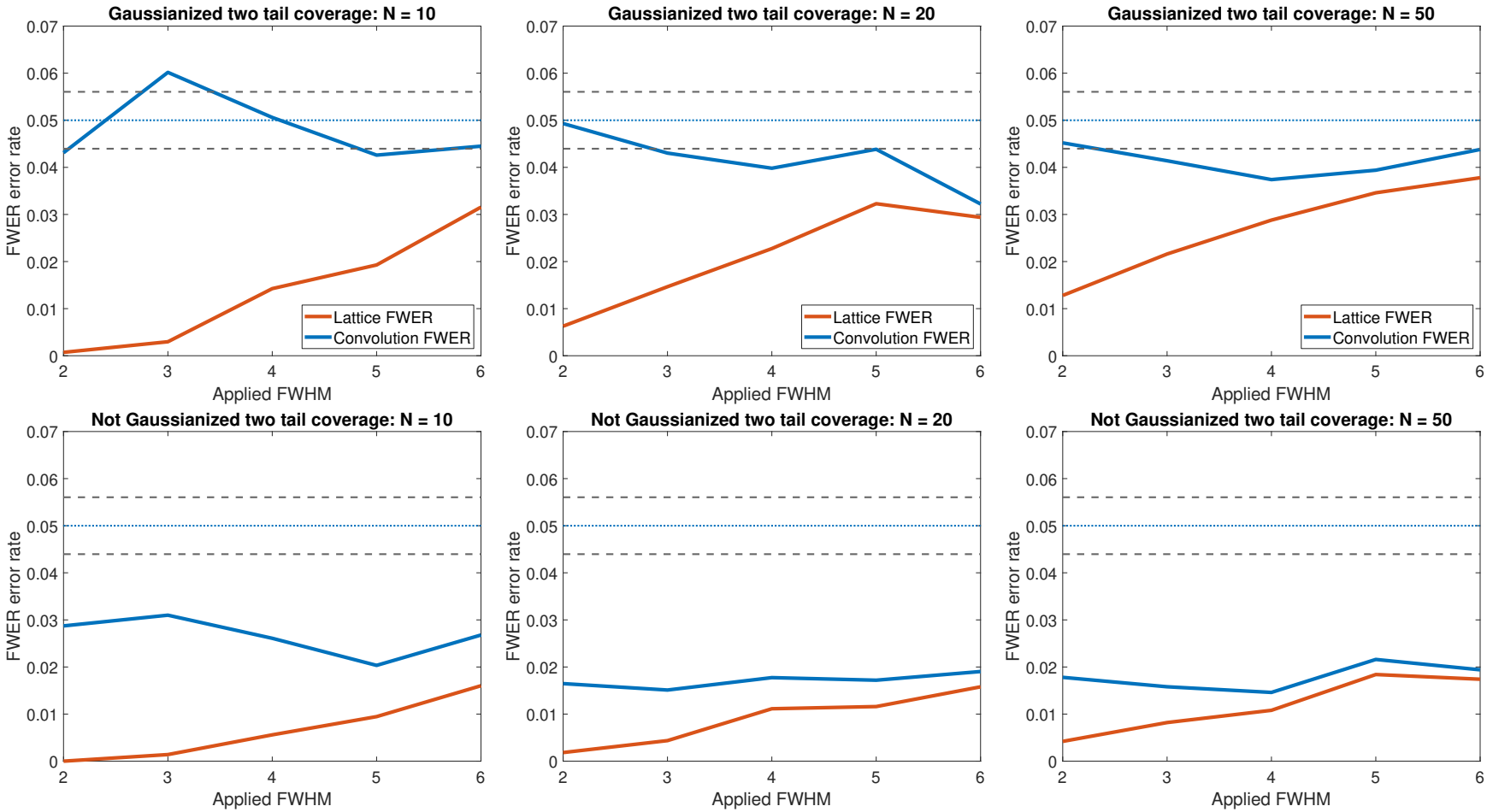


Figure 8: Red: original lattice and Blue: Convolution field

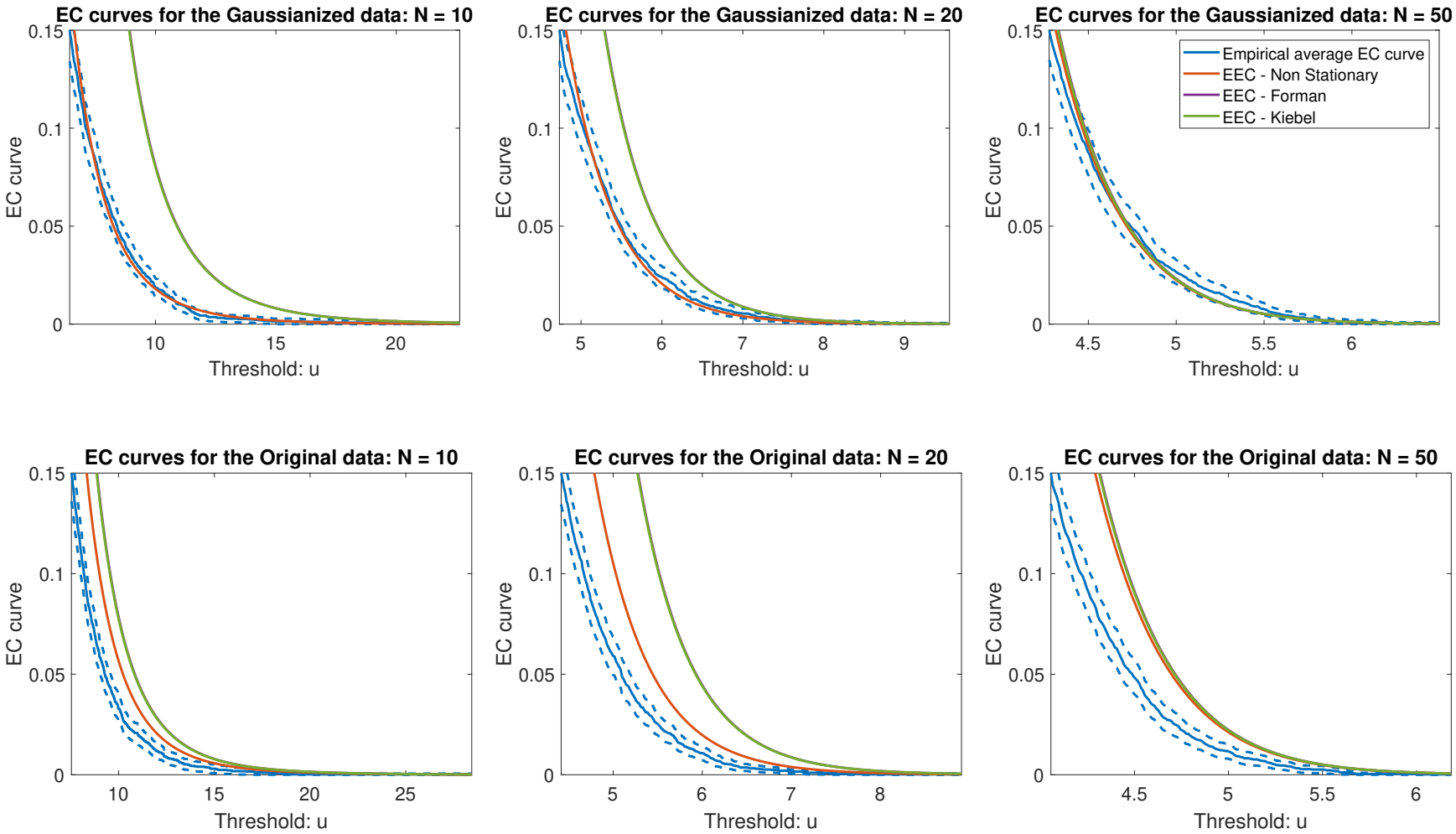
Controlling at $\alpha = 0.01$



Controlling at $\alpha = 0.05$



Expected Euler characteristic curve



LKC estimation results

We run 2D simulations, of white noise smoothed with a Gaussian Kernel. Kiebel and Forman are designed to estimate the LKCs under stationarity but they are biased. HPE and bHPE are unbiased but have a higher variance.