FDP control in multivariate linear models using the bootstrap

Samuel Davenport[1*], Bertrand Thirion[2], Pierre Neuvial[3]

[1] Division of Biostatistics, University of California San Diego, United States

[2] Inria, Université Paris-Saclay

[3] Institut de Mathématiques de Toulouse, Université de Toulouse

[*] Corresponding author email: sdavenport@health.ucsd.edu

Corresponding author address: Division of Biostatistics, 9500 Gilman Dr, La Jolla, CA 92093, USA

September 24, 2023

Abstract

In this article we develop a method for performing post hoc inference of the False Discovery Proportion (FDP) over multiple contrasts of interest in the multivariate linear model. To do so we use the bootstrap to simulate from the distribution of the null contrasts. We combine the bootstrap with the post hoc inference bounds of Blanchard et al. (2020) and prove that doing so provides simultaneous asymptotic control of the FDP over all subsets of hypotheses. We demonstrate, via simulations, that our approach provides simultaneous control of the FDP over all subsets and is typically more powerful than existing, state of the art, parametric methods. We illustrate our approach on functional Magnetic Resonance Imaging data from the Human Connectome project and on a transcriptomic dataset of chronic obstructive pulmonary disease.

Keywords: FDP control, bootstrap, simultaneous inference, post hoc inference

1 Introduction

Statistical analysis of functional Magnetic Resonance Imaging data grounds the inference of associations between external conditions (such as disease status and experimental factors) and the signals recorded in brain regions, that are assumed to reflect brain activity. In particular, practitioners typically aim to uncover associations between local signals and conditions that are specific to a given area; such specificity is essential for interpretation purposes. The most standard framework is that of mass-univariate inference, in which models are fit separately at each brain location, in order to detect significant associations. This framework is simple and computationally efficient but, given mm-scale resolution reached by current imaging setups, results in a dire multiple comparison issue.

Statistical analysis of genomic data encounters a similar multiple comparison problem. In particular, this is the case in Genome-Wide Association Studies that aim to identify Single Nucleotide Polymorphisms that are associated with one or more phenotypes of interest, and in gene expression studies, the goal of which is to identify genes where activity is associated with one or more variables of biological or clinical interest.

In this field, the state-of-the-art framework is also based on univariate tests that are performed for each genomic marker. While imaging data typically consist of a smooth volume-domain voxel grid, the dependence structure of genomic data is dictated by the interdependence between genomic markers, which is mediated by haplotypic blocks encountered in Genome-Wide Association Studies and by gene networks or pathways in expression studies.

In both of these scientific fields (and in others), control of the false discovery rate (FDR) has quickly become a de facto standard, as it yields image or genome-level error control together with acceptable power (Genovese et al., 2002; Storey and Tibshirani, 2003). In practice, most researchers control the FDR using the Benjamini-Hochberg procedure (Benjamini and Hochberg, 1995), under the assumption of positive regression dependence (Benjamini and Yekutieli, 2001). This assumption is generally considered reasonable given the positive correlation that typically exists between voxels or genomic markers. However, users often interpret FDR-control as a control of false discovery proportion (FDP), which is incorrect, as the FDR is only the expected value of the FDP. Overall, this approach can result in unreliable error control, especially when there is dependence within the data, see Korn et al. (2004) and Figure 2.1 in Neuvial (2020). It is instead desirable to provide probabilistic control on the *proportion* or *number* of false discoveries.

Genomic and brain imaging datasets frequently involve the simultaneous test of several contrasts (Smyth, 2004; Alberton et al., 2020). Such simultaneous tests are important because they can ground double dissociation (Henson, 2006), ensuring the specificity of discoveries and leading to unambiguous interpretations of the results. A difficulty arises here as the tests of the different contrasts that are considered at each feature (voxel/gene) are typically dependent and it may no longer be reasonable to assume positive regression dependence. It is thus of interest to consider controlling the FDP under the null hypothesis for each contrast, without making unwanted assumptions.

The notion of *post hoc inference* was introduced by Goeman and Solari (2011), following earlier works by Genovese and Wasserman (2006); Meinshausen (2006) on the probabilistic control of the FDP. The idea of post hoc inference is to provide confidence bounds on the number or proportion of true/false discoveries among arbitrary and possibly data-driven subsets of variables of interest. By construction, such guarantees address the issue of circular inference (Rosenblatt et al., 2018).

Post hoc bounds can be obtained as a by-product of the control of a multiple testing risk called the joint error rate (JER) by a simple interpolation argument (Blanchard et al., 2020). Using this construction, state-of-the-art post hoc bounds (Goeman and Solari, 2011; Rosenblatt et al., 2018) can be recovered from the Simes inequality, a classical result from the multiple testing literature (Simes, 1986). The resulting bounds are valid under positive regression dependence. They have also been shown to be conservative in genomics and neuroimaging applications (Blanchard et al., 2021).

Since the joint error rate only depends on the joint distribution of the test statistics under the null hypothesis, joint error rate control can alternatively be obtained by randomization techniques (Blanchard et al., 2020, 2021; Hemerik et al., 2019; Andreella et al., 2023). In particular, sharp data-driven joint error rate control and associated post hoc bounds have been obtained for one-sample tests using sign-flipping, and for two-sample tests using permutations (Blanchard et al., 2020, 2021). When inferring on contrasts in the linear model, on which we will focus, exchangeability does not generally hold. As such it is not possible to provide valid finite sample inference using permutation. In order to perform post hoc inference over contrasts we will need to be able to

obtain the joint null distribution of the test-statistics of multiple contrasts within the framework of the linear model. To do so we use the bootstrap, adjusting the approach of Westfall (2011) to the multivariate setting. Justification for bootstrapping the residuals in a one-dimensional linear model was first provided in Freedman (1981) based on theory proved in Bickel and Freedman (1981). These results and their proofs were extended to multivariate linear models in Eck (2018).

The primary contribution of our work is to theoretically extend the results of Blanchard et al. (2020) to the non-exchangeable setting of the linear model, allowing us to provide asymptotically valid simultaneous FDP control. To do so we prove that the bootstrap can be combined with the interpolation approach of Blanchard et al. (2020) in order to provide asymptotically valid post-hoc bounds. This requires a different proof strategy to that of Blanchard et al. (2020), in order to establish asymptotic control of the joint error rate, because their proofs strongly rely on exchangeability of the resampled pivotal statistics which does not hold in our setting. As part of this contribution we illustrate how these methods can be applied in practice on both brain imaging and genetics datasets and provide software tools to implement them which are available in a python package at: `http://github.com/sjdavenport/pyperm`. As a further contribution we include an alternative proof of the consistency of the bootstrap in the multivariate linear model which relies on the Lindeberg CLT.

Proofs and further theoretical and simulation results are available in the supplementary material - sections of which will be denoted using the suffix S. Code to reproduce the analyses and figures of this paper is available at at: `http://github.com/sjdavenport/lmfdp`.

2 Notation and general framework

2.1 Random Fields on a Lattice

Throughout we will take $(\Omega, \mathcal{F}, \mathbb{P})$ to be a probability space, write $\mathbb{E}$ to denote expectation and will define random variables with respect to this space. We will also take $\mathbb{N}$ to be the set of positive integers. We will primarily be working with random fields, observed at a finite number of points, as our data. These are defined as follows.

Definition 2.1. 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 measurable mapping $g : \Omega \to \left\{h : \mathcal{V} \to \mathbb{R}^L\right\}$. We say that g has **dimension** L.

Given $\omega \in \Omega$ and $v \in \mathcal{V}$ we will write $g(\omega, v) = g(\omega)(v)$ and will typically drop dependence on ω and simply refer to the random variable $g(v) : \Omega \to \mathbb{R}^L$ when indexing g and say that g is a random field on $\mathcal{V}$. We define the mean of g to be the function $\mu : \mathcal{V} \to \mathbb{R}^L$ sending $v \in \mathcal{V}$ to $\mathbb{E}[g(v)]$. To each g we associate a covariance $\mathfrak{c}$ and a correlation function ρ which map $\mathcal{V} \times \mathcal{V}$ to $\mathbb{R}^{L \times L}$ and are defined as

$$\mathfrak{c}(u, v) = \text{cov}(g(u), g(v)) = \mathbb{E}\Big[(g(u) - \mu(u))(g(v) - \mu(v))^T\Big]$$

and $\rho(u, v) = \mathfrak{c}(u, v)(\mathfrak{c}(u, u)\mathfrak{c}(v, v))^{-1/2}$ for all $u, v \in \mathcal{V}$.

For $1 \leq j \leq L$, we define the random fields $g_j : \Omega \to \{g : \mathcal{V} \to \mathbb{R}\}$ which send $\omega \in \Omega$ to $g_j(\omega)(\cdot) = g(\omega)(\cdot)_j = g(\cdot)_j$. We will call $g_1, \ldots, g_L$ the **components** of g and will write the combination as $g = [g_1, \ldots, g_L]^T$. Convergence in distribution and probability (which we will denote by $\xrightarrow{d}$ and $\xrightarrow{\mathbb{P}}$) of random fields is well defined via vectorization,

see Section S-1 for a formalization of this and for how to define operations on random fields. We also define Gaussian random fields as follows.

Definition 2.2. Given functions $\mu : \mathcal{V} \to \mathbb{R}^L$ and $\mathfrak{c} : \mathcal{V} \times \mathcal{V}$ we write $g \sim \mathcal{G}(\mu, \mathfrak{c})$ if g is a random field with mean μ and covariance $\mathfrak{c}$ and the vector $(g_j(v) : v \in \mathcal{V}, 1 \leq j \leq L)$ has a multivariate Gaussian distribution.

2.2 Linear Model Framework

Let $\mathcal{V} \subset \mathbb{R}^D$ be a finite set of points corresponding to the domain of interest (this could for instance be the voxels of the brain or points representing genes). Suppose that we observe random fields $y_i : \mathcal{V} \to \mathbb{R}$, for $1 \leq i \leq n$ and some number of subjects $n \in \mathbb{N}$. At each point $v \in \mathcal{V}$, we assume that

$$Y_n(v) = X_n\beta(v) + E_n(v) \tag{1}$$

where for each $v \in \mathcal{V}$, $Y_n(v) = [y_1(v), \ldots, y_n(v)]^T$ is a vector giving the observed data, $\beta(v) \in \mathbb{R}^p$ is the vector of parameters (for some $p \in \mathbb{N}$), $X_n \in \mathbb{R}^{n \times p}$ is the design matrix of the covariates (note that this may include nuisance variables) and $E_n = [\epsilon_1, \ldots, \epsilon_n]^T$ is an n-dimensional random field on $\mathcal{V}$ which represents the unobserved noise, where $(\epsilon_n)_{n \in \mathbb{N}}$ is an i.i.d sequence of 1-dimensional random fields on $\mathcal{V}$. Note that we give the design matrix X_n a subscript n as we will allow it to grow with n. Let $\mathfrak{c}_\epsilon$ and ρ_ϵ be the covariance and correlation functions of ϵ_1 respectively.

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 $v \in \mathcal{V}$. For each $v \in \mathcal{V}$ we can test the intersection hypothesis

$$H_0(v) : c_l^T\beta(v) = 0 \text{ for } 1 \leq l \leq L$$

using an F-test at each $v \in \mathcal{V}$ given by

$$F_n(v) = \frac{(C\hat{\beta}_n(v))^T(C(X_n^TX_n)^{-1}C^T)^{-1}(C\hat{\beta}_n(v))/\text{rank}(C)}{\hat{\sigma}_n(v)^2}. \tag{2}$$

Here $\hat{\beta}_n(v) = (X_n^TX_n)^{-1}X_n^TY_n(v)$ and $C = (c_1, \ldots, c_L)^T \in \mathbb{R}^{L \times p}$ is the matrix of contrasts. $\hat{\sigma}_n^2 : \mathcal{V} \to \mathbb{R}$ is the estimate of the variance based on the residuals which sends $v \in \mathcal{V}$ to

$$\hat{\sigma}_n^2(v) = \frac{1}{n - r_n}\left\|Y_n(v) - X_n\hat{\beta}_n(v)\right\|^2.$$

where r_n is the rank of X_n. The individual null hypotheses can be tested using test statistics:

$$T_{n,l}(v) = \frac{c_l^T\hat{\beta}_n(v)}{\sqrt{\hat{\sigma}_n(v)^2c_l^T(X_n^TX_n)^{-1}c_l}}. \tag{3}$$

Under $H_{0,l}(v)$ and assuming that the noise is Gaussian, conditional on X_n, $T_{n,l}(v)$ is distributed as a t-statistic with $n - r_n$ degrees of freedom. This allows a p-value to be calculated to test $H_{0,l}(v)$ for each contrast l at each point v, namely, $p_{n,l}(v) = 2(1-\Phi_{n-r_n}(|T_{n,l}(v)|))$ where Φ_d is the CDF of a t-statistic with $d \in \mathbb{N}$ degrees of freedom. Dropping the Gaussianity assumption, the p-values are still asymptotically valid under reasonable assumptions (see e.g. Theorem S-3.4). Moreover, for each $1 \leq l \leq L$, $T_{n,l}$ is a 1-dimensional random field and we define $T_n = [T_{n,1}, \ldots, T_{n,L}]^T$.

2.3 Bounds on the False Discovery Proportion

The above framework gives us $m = L|\mathcal{V}|$ different hypothesis tests, and results in a multiple testing problem, which can be quite severe e.g. if the size of $\mathcal{V}$ is large. Let $\mathcal{H} = \{(l, v) : 1 \leq l \leq L \text{ and } v \in \mathcal{V}\}$ index the hypotheses. For $H \subseteq \mathcal{H}$, let $|H|$ denote the number of elements within H. Finally let $\mathcal{N} \subseteq \mathcal{H}$ index the true null hypotheses. Given $0 < \alpha < 1$ we will seek to provide a function $V : \{H : H \subseteq \mathcal{H}\} \to \mathbb{N}$ such that

$$\mathbb{P}(|H \cap \mathcal{N}| \leq V(H), \text{ for all } H \subseteq \mathcal{H}) \geq 1 - \alpha. \tag{4}$$

If (4) holds then simultaneously over all $H \subseteq \mathcal{H}$, with probability $1 - \alpha$, $V(H)$ provides an upper bound on the number of false positives within H. Suppose that for some $K \in \mathbb{N}$ we have sets $R_1, \ldots, R_K \subseteq \mathcal{H}$ (which depend on the data) and constants $\zeta_1, \ldots, \zeta_K \in \mathbb{N}$ and define

$$\text{JER}((R_k, \zeta_k)_{1 \leq k \leq K}) := \mathbb{P}(|R_k \cap \mathcal{N}| > \zeta_k, \text{ some } 1 \leq k \leq K) \tag{5}$$

to be the joint error rate of the collection $(R_k, \zeta_k)_{1 \leq k \leq K}$. Blanchard et al. (2020) showed that if $\text{JER}((R_k, \zeta_k)_{1 \leq k \leq K}) \leq \alpha$, then the bound $\overline{V} : \{H : H \subseteq \mathcal{H}\} \to \mathbb{R}$, sending $H \subseteq \mathcal{H}$ to

$$\overline{V}(H) = \min_{1 \leq k \leq K} (|H \setminus R_k| + \zeta_k) \wedge |H|, \tag{6}$$

satisfies (4) and thus provides an α-level bound over the number of false positives within each chosen rejection set. If the sets $R_1, \ldots, R_K$ are nested then $\overline{V}$ is in fact the optimal interpolation bound[1] among the post hoc bounds that can be derived from JER control. We will follow the approach of Blanchard et al. (2020) and define the collections $(R_k, \zeta_k)_{1 \leq k \leq K}$, that we will use, using template families. An important practical feature of the bound $\overline{V}(H)$ is that it can be computed in linear time in $|H|$, see e.g. Algorithm 2 in Enjalbert-Courrech and Neuvial (2022). For simplicity we did not consider the closed testing-based post hoc bounds introduced by Goeman and Solari (2011). This type of post hoc bounds are briefly discussed in Section 6.

Definition 2.3. Given $K \in \mathbb{N}$, we say that a family of functions $(t_k)_{1 \leq k \leq K}$ is a **template family** if for each $1 \leq k \leq K$, $t_k : [0, 1] \to \mathbb{R}$, $t_k(0) = 0$ and t_k is strictly increasing and continuous. The parameter K is called the **size** of the template.

The simplest and most commonly used template family is the linear template which, for $K \in \mathbb{N}$, is given by $t_k(x) = \frac{xk}{m}$ for $1 \leq k \leq K$ and $x \in [0, 1]$. Existing post hoc bounds associated with this template are described in Section 3.3. However other choices are available and the optimal choice of template may depend on the dataset under consideration: we refer to Section 6 for further details and a discussion of the choice of template as well as to Hemerik et al. (2019); Blanchard et al. (2020); Blain et al. (2022); Andreella et al. (2023). Given a template family and $\lambda \in [0, 1]$, for each $1 \leq k \leq K$ and $n \in \mathbb{N}$, we will take $R_k(\lambda) = \{(l, v) \in \mathcal{H} : p_{n,l}(v) \leq t_k(\lambda)\}$, set $\zeta_k = k - 1$, and let $p^n_{(k:\mathcal{N})}$ be the kth smallest p-value in the set $\{p_{n,l}(v) : (l, v) \in \mathcal{N}\}$ (setting $p^n_{(k:\mathcal{N})} = 1$ if $k > |\mathcal{N}|$). We will refer to the collection $(R_k(\lambda), k - 1)_{1 \leq k \leq K}$ as the canonical reference family.

[1]Here we define optimal in the sense that given another function $V' : \{H : H \subseteq \mathcal{H}\}$, such that for all $A \subseteq \mathcal{H}, |R_k \cap A| \leq \zeta_k$ for $1 \leq k \leq K$ implies that $|R \cap A| \leq V'(R)$ for all $R \subseteq \mathcal{H}$, it follows that $V'(R) \leq \overline{V}(R)$. Optimality in this sense follows by Blanchard et al. (2020)'s Proposition 2.5.

Lemma 2.4. *For each* $\lambda \in [0,1]$,

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

Thus for a given template family, in order to obtain an upper bound on the number of false positives we can choose a threshold $\lambda \in [0,1]$ such that

$$\mathbb{P}\left(\min_{1\leq k\leq K\wedge|\mathcal{H}|} t_k^{-1}(p^n_{(k:\mathcal{N})}) \leq \lambda\right) \leq \alpha. \tag{7}$$

Then the joint error rate of the family $(R_k(\lambda), k-1)_{1\leq k\leq K}$ is controlled to a level α and so the corresponding bound: $\overline{V}$, provides a $(1-\alpha)$-level simultaneous upper bound on the number of false positives.

Blanchard et al. (2020) chose λ via permutation testing, using the fact that under an exchangeability assumption permutation allows the probability in (7) to be controlled exactly. In the linear model, permutation of the response does not satisfy the exchangeability assumption when there are multiple potentially non-zero covariates in the model (see Appendix S-4.3 for a discussion of this). In what follows we take a different approach that proceeds via bootstrapping the data and results in asymptotic control of the error rate.

For $\alpha \in (0,1)$ $\overline{V}(H)$ provides an $(1-\alpha)$-level simultaneous upper bound on the number of false positives within H. From (4) we have

$$\mathbb{P}\left(\frac{|H\cap\mathcal{N}|}{|H|} \leq \frac{\overline{V}(H)}{|H|},\ \forall H\subseteq\mathcal{H}\right) \geq 1-\alpha. \tag{8}$$

It thus follows that for each $H \in \mathcal{H}$, $\frac{\overline{V}(H)}{|H|}$ provides an upper bound on the proportion of false positives within H also known as the **false discovery proportion** or **FDP**. Similarly $\frac{|H|-\overline{V}(H)}{|H|}$ provides a $(1-\alpha)$-level simultaneous lower bound on the **true discovery proportion** or **TDP**.

2.4 Bootstrapping

In what follows we shall resample our data using the residual bootstrap (Freedman, 1981) which is robust to potential asymmetries in the data as it does not rely on sign flipping. Given $n \in \mathbb{N}$, in our notation, this proceeds by calculating the residuals

$$\hat{E}_n = Y_n - X_n\hat{\beta}_n = (I_n - X_n(X_n^TX_n)^{-1}X_n^T)E_n, \tag{9}$$

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

$$\hat{\beta}_n = (X_n^TX_n)^{-1}X_n^TY_n = \beta + (X_n^TX_n)^{-1}X_n^TE_n. \tag{10}$$

Given a number of bootstraps to perform: $B \in \mathbb{N}$ for each $1 \leq b \leq B$, conditional on the data, a selection: $\hat{\epsilon}_1^b, \ldots, \hat{\epsilon}_n^b$ is chosen independently with replacement from $\left\{\hat{E}_{n,1}, \ldots, \hat{E}_{n,n}\right\}$ resulting in a combined random field $E_n^b = [\hat{\epsilon}_1^b, \ldots, \hat{\epsilon}_n^b]^T$. Given this let $Y_n^b = X_n\hat{\beta}_n + E_n^b$ and define bootstrapped parameter estimates

$$\hat{\beta}_n^b = (X_n^TX_n)^{-1}X_n^TY_n^b. \tag{11}$$

Define the estimate of variance using the bootstrap residuals to be

$$(\hat{\sigma}_n^b)^2 = \frac{1}{n}\sum_{i=1}^{n}(E_{n,i}^b)^2 - \left(\frac{1}{n}\sum_{i=1}^{n} E_{n,i}^b\right)^2. \tag{12}$$

We can use bootstrap to infer on our desired null distribution. To do so we will require the following assumption.

Assumption 1.

a) For $n \in \mathbb{N}$, $X_n = [x_1, \ldots, x_n]^T$ for a sequence of i.i.d vectors $(x_n)_{n\in\mathbb{N}}$ in $\mathbb{R}^p$ with bounded density and such that $\mathbb{E}\left[\|x_1\|^{2+\delta}\right] < \infty$ for some $\delta > 0$. Let $\Sigma_X = \mathbb{E}[x_1 x_1^T]$.

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 $\min_{v\in\mathcal{V}} \operatorname{var}(\epsilon_1(v)) > 0$.

Suppose $(X_n)_{n\in\mathbb{N}}$ and $(\epsilon_n)_{n\in\mathbb{N}}$ satisfy Assumption 1. Then consistency of the multivariate bootstrap (Freedman, 1981; Eck, 2018) implies that, conditional on $(X_m, Y_m)_{m\in\mathbb{N}}$ for almost all sequences $(X_m, Y_m)_{m\in\mathbb{N}}$, for each $1 \le b \le B$, as $n \to \infty$,

$$\sqrt{n}(\hat{\beta}_n^b - \hat{\beta}_n) \xrightarrow{d} \mathcal{G}(0, \mathfrak{c}_\epsilon \Sigma_X^{-1}) \tag{13}$$

$$\text{and} \quad \hat{\sigma}_n^b \xrightarrow{\mathbb{P}} \sigma.$$

In particular, let $T_n^b : \mathcal{V} \to \mathbb{R}$ be the L-dimensional random field on $\mathcal{V}$ such that, for $1 \le l \le L$,

$$T_{n,l}^b = \frac{c_l^T(\hat{\beta}_n^b - \hat{\beta}_n)}{\hat{\sigma}_n^b \sqrt{c_l^T (X_n^T X_n)^{-1} c_l}}. \tag{14}$$

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 $1 \le b \le B$,

$$T_n^b \xrightarrow{d} \mathcal{G}(0, \mathfrak{c}') \tag{15}$$

as $n \to \infty$. Here $\mathfrak{c}' : \mathcal{V} \times \mathcal{V} \to \mathbb{R}$ takes $u, v \in \mathcal{V}$ to $\mathfrak{c}'(u,v) = \rho_\epsilon(u,v) A C \Sigma_X^{-1} C^T A^T$ where $A \in \mathbb{R}^{L\times L}$ is a diagonal matrix with $A_{ll} = (c_l^T \Sigma_X^{-1} c_l)^{-1/2}$ for $1 \le l \le L$. Crucially the limiting distribution in this result is the same as the limiting distribution of the test-statistics (3) under the global null that $\beta = 0$, see Lemma S-3.4. It follows that the bootstrap provides consistent estimates of the quantiles of functionals of the data, see Section S-2.10.

We provide two proofs of these bootstrap consistency results. The first demonstrates how to translate the results of Eck (2018) into our notation, see Section S-2.4. The second instead uses the Lindeberg CLT to establish convergence and is provided in Section S-2.3, see Theorems S-2.5 and S-2.7 and their proofs for further details.

Remark 2.5. *In the univariate setting consistency of the bootstrap was originally proved in Freedman (1981). This result has been used widely in the multivariate setting however it was only recently that Eck (2018) formally showed that that the proof of Freedman (1981) extends to multiple dimensions. The approach of Freedman (1981) and Eck (2018), takes advantage of the fact that convergence in distribution is equivalent to convergence in the Mallows metric (Bickel and Freedman (1981)). We provide an alternative proof of the consistency of the bootstrap in Section S-2 which instead relies on elementary probability tools such as the Lindeberg CLT and the triangular law of large numbers.*

3 Joint error rate control in the linear model

In this section we will state and prove our main results. We will show that given $0 < \alpha < 1$, choosing λ to be the α-quantile of the bootstrapped distribution results in asymptotic $1-\alpha$ level control of the joint error rate and thus results in simultaneous control of the FDP.

3.1 Joint error rate control

To set this up, given a test-statistic $T : \mathcal{V} \to \mathbb{R}^L$, a subset $H \subseteq \mathcal{H}$, and $n \in \mathbb{N}$, for $1 \leq k \leq |H|$, let $p^n_{(k:H)}(T)$ be the kth minimum value in the set

$$\{2 - 2\Phi_{n-r_n}(|T_l(v)|) : (l, v) \in H\}.$$

Using the results we have proved so far we can obtain the following theorem.

Theorem 3.1. *For $H \subseteq \mathcal{H}$, let $f_{n,H} : \{g : \mathcal{V} \to \mathbb{R}^L\} \to \mathbb{R}$ send*

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

and for $n, B \in \mathbb{N}$ and $\alpha \in (0, 1)$, let

$$\lambda^*_{\alpha,n,B}(H) = \inf\left\{\lambda : \frac{1}{B}\sum_{b=1}^{B} 1\big[f_{n,H}(T^b_n) \leq \lambda\big] \geq \alpha\right\}$$

be the α-quantile of the bootstrap distribution based on $B \in \mathbb{N}$ bootstraps, of $f_{n,H}(T_n)$ conditional on the observed data. Assume that Assumption 1 holds and that $n - r_n \to \infty$ almost surely. Then for all $H \subseteq \mathcal{H}$ such that $\mathcal{N} \subseteq H$,

$$\lim_{n\to\infty} \lim_{B\to\infty} \mathbb{P}\big(f_{n,\mathcal{N}}(T_n) \leq \lambda^*_{\alpha,n,B}(H)\big) \leq \alpha.$$

This limit holds with equality if $H = \mathcal{N}$. Furthermore, taking $H = \mathcal{H}$, it follows that

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

Applying this result and using Claim 2.4 we are thus able to obtain asymptotic control of the joint error rate of the canonical reference family. Following the discussion in Section 2.3 this means that we obtain asymptotic post hoc FDP control. In particular we having the following corollary.

Corollary 3.2. *Under the assumptions of Theorem 3.1, for $0 < \alpha < 1$, and $H \subseteq \mathcal{H}$, let*

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

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

Thus in order to provide FDP control, given a number of bootstraps $B \in \mathbb{N}$, we can calculate $\lambda^*_{\alpha,n,B}(H)$, the α-quantile of the bootstrap distribution of $f_{n,H}(T_n)$ conditional on the observed data. Then $\overline{V}_{\alpha,n,B}(H)$ provides a $(1-\alpha)$ level simultaneous upper bound on the number of false positives in $H \subseteq \mathcal{H}$.

3.2 Bootstrap step-down procedure

It is possible to improve on the power of the above procedure by taking a step-down approach in the spirit of (Romano and Wolf, 2005). This is based on the idea that joint error rate control implies familywise error rate control, see Section S-4.2. As such it is possible to obtain an estimate of the set of null hypotheses and thereby obtain a tighter bound. The procedure, which adapts the step-down procedure of Blanchard et al. (2020) to our setting, can be iterated as follows.

Algorithm 1 step-down bootstrap

1: Set $j \leftarrow 0$ and $H_n^{(0)} \leftarrow \mathcal{H}$
2: **repeat**
3: Set $j \leftarrow j+1$, $\lambda_{n,j} \leftarrow \lambda^*_{\alpha,n,B}(H_n^{(j-1)})$ and $H_n^{(j)} \leftarrow \{(l,v) : p_{n,l}(v) \geq t_1(\lambda_{n,j})\}$
4: **until** $H_n^{(j)} = H_n^{(j-1)}$
5: Set $\hat{H}_n \leftarrow H_n^{(j)}$ and **return** $\hat{H}_n$

As the following theorem demonstrates, the step-down approach controls the joint error rate and therefore provides simultaneous FDP control.

Theorem 3.3. *Under the assumptions of Theorem 3.1, for* $0 < \alpha < 1$, *let* $\hat{H}_n$ *be the set generated by applying Algorithm 1. Then*

$$\lim_{n\to\infty} \lim_{B\to\infty} \mathbb{P}\Big(f_{n,\mathcal{N}}(T_n) < \lambda^*_{\alpha,n,B}(\hat{H}_n) \Big) \leq \alpha.$$

Thus, for $H \subseteq \mathcal{H}$, *letting* $\overline{V}_{\alpha,n,B}(H) = \min_{1\leq k\leq K}(|H \setminus R_k(\lambda^*_{\alpha,n,B}(\hat{H}_n))| + k - 1) \wedge |H|$, *it follows that*

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

In the definition of $\lambda^*_{\alpha,n,B}$ we require the computation of $|\mathcal{H}|$ statistics for each bootstrap each of which is based on a sample of size n. As such the complexity of these algorithms is $O(nB|\mathcal{H}|)$.

Remark 3.4. *The results in this subsection and the one previous have been stated for two-sided p-values however they also hold for one-sided p-values,* $1 - \Phi_{n-r_n}(|T_{n,l}(v)|)$ *without change. All that is required to show this is to re-define* $p^n_{(k:H)}(T)$ *as the* k*th minimum value in the set*

$$\{1 - \Phi_{n-r_n}(T_l(v)) : (l,v) \in H\}.$$

In this scenario we would use the one-sided p-values to test the null hypotheses that $c_l^T \beta(v) \leq 0$ *at each* $v \in \mathcal{V}$ *and* $1 \leq l \leq L$.

3.3 Parametric Approaches

In this section we will discuss two existing parametric[2] approaches to simultaneous FDP inference which are based on the Simes inequality (17). The first one is the original Simes post hoc bound introduced in Goeman and Solari (2011). The second one is the method

[2]Here we use the term parametric to indicate that dependency assumptions on the data are required in order for the methods to be valid.

of Rosenblatt et al. (2018) and Goeman et al. (2019). It corresponds to an improvement on the basic Simes bound that is adaptive to the proportion of true null hypotheses - i.e. it is a step-down version of the Simes bound. This method has been applied to brain imaging data in Rosenblatt et al. (2018), and is called **ARI** which stands for "**All Resolutions Inference**". Both methods can be conveniently formulated in terms of the bound $\overline{V}$ defined in (6), associated to the linear template family $(t_k)_{1\leq k\leq m}$, where $t_k(\lambda) = \lambda k/m$, i.e.

$$\overline{V}_\lambda(S) = \min_{1\leq k\leq m}\left\{\sum_{i\in S} 1\left[p_i \geq \frac{\lambda k}{m}\right] + k - 1\right\}. \tag{16}$$

As noted by Blanchard et al. (2020), the Simes post hoc bound of Goeman and Solari (2011) is simply $\overline{V}_\alpha$. Moreover, letting $\overline{\alpha} = \alpha m/h(\alpha)$, where

$$h(\alpha) = \max\left\{i \in \{1,\ldots,m\}, \forall j \in \{1,\ldots,i\}, p_{(m-i+j)} > \frac{\alpha j}{i}\right\},$$

the ARI bound of Goeman et al. (2019) is $\overline{V}_{\overline{\alpha}}$. The quantity $h(\alpha)$ is called the Hommel factor (Hommel, 1988) and can be interpreted as a $(1-\alpha)$-level upper confidence bound on $|\mathcal{N}|$, the number of true null hypotheses.

If the null p-values satisfy positive regression dependence then both of these methods result in simultaneous $(1-\alpha)$-level FDP control. This is shown formally in Goeman and Solari (2011) and Goeman et al. (2019) via closed testing and can also be shown to hold by combining the Simes inequality with the joint error rate framework of Section 2.3. To see this note that if the null p-values are positive regression dependent (Sarkar et al., 2008), then the Simes inequality is satisfied, that is:

$$\mathbb{P}\left(\exists 1 \leq k \leq |\mathcal{N}| : p^n_{(k:\mathcal{N})} \leq \frac{\alpha k}{|\mathcal{N}|}\right) \leq \alpha, \tag{17}$$

with equality if the null p-values are independent.

In particular taking $\lambda = \alpha$, and noting that $|\mathcal{N}| \leq m$, the Simes inequality implies that (7) holds (taking $K = m$ and $(t_k)_{1\leq k\leq m}$ to be the linear reference family). Moreover, Goeman et al. (2019)'s Lemma 2 implies that if the null p-values satisfy positive regression dependence, then

$$\mathbb{P}\left(\exists 1 \leq k \leq |\mathcal{N}| : p^n_{(k:\mathcal{N})} \leq \frac{\alpha k}{h(\alpha)}\right) \leq \alpha. \tag{18}$$

Thus taking $\lambda = \overline{\alpha} = \alpha m/h(\alpha)$, it follows that (7) holds with respect to the linear reference family. In particular the Simes procedure, which uses $\overline{V}_\alpha$ as a bound, and ARI, which uses $\overline{V}_{\overline{\alpha}}$, provide simultaneous $(1-\alpha)$-level control of the FDP.

In our results, presented in the following sections, we compare the performance of the non-parametric bootstrap approach to these parametric alternatives.

4 Simulation Results

4.1 Simulation Setup

In order to assess empirically that our method correctly controls the joint error rate we run numerical simulations. We create the noise in these simulations by generating 2-dimensional stationary Gaussian random fields on domains which are 25 by 25, 50 by 50

and 100 by 100 pixels. To do so we smooth Gaussian white noise with a Gaussian kernel with full width at half maximum (FWHM) in $\{0, 4, 8\}$ (in pixel units), accounting for edge effects to ensure stationarity (see e.g. Davenport and Nichols (2022)), and scaled so that the variance is 1 everywhere.

We let the total number of subjects n range from 20 to 100. For each n, smoothness level, image size and $\pi_0 \in \{0.5, 0.8, 0.9, 1\}$, we run 5000 simulations - each with 100 bootstraps - to test the joint error rate. For each simulation we do the following. First we generate n Gaussian random fields $\epsilon_1, \ldots, \epsilon_n$ as described above and add signal to them (as detailed in the next paragraph). We then randomly divide these images into 3 disjoint groups: $G_1, G_2, G_3 \subset \{1, \ldots, n\}$ - performing assignment to each group with equal probability (we eliminate assignments where a given group has no entries). We test for the difference between the first and the second group and between the second and the third group - giving us $L = 2$ contrasts to differentiate between. We thus, in total, test 5000 hypotheses for the 50 by 50 scenario and 20000 for the 100 by 100 case.

We vary the amount of signal in the datasets as follows. Given a proportion π_0 we randomly choose a subset $\mathcal{N}$ of size $\pi_0|\mathcal{H}|$ of $\mathcal{H} = \{(l, v) : 1 \leq l \leq 2, v \in \mathcal{V}\}$ to be null (which is thus different in each simulation) and add signal to ensure that the remainder are non-null. To do so, for $1 \leq i \leq n$, and each $v \in \mathcal{V}$, we set

$$Y_i(v) = 1[i \in G_2, (1, v) \notin \mathcal{N}] + 1[i \in G_3, (1, v) \notin \mathcal{N}] + 1[i \in G_3, (2, v) \notin \mathcal{N}] + \epsilon_i(v).$$

This ensures that the power of the test to detect a difference at h is equal for any $h \in \mathcal{N}^C$. If $\pi_0 = 1$ then all hypotheses are null. An example realisation is shown in Figure S-4.

In the next subsections we compare our bootstrap procedures, in terms of false positive control and power, to two parametric alternatives: the Simes procedure (Goeman and Solari, 2011) and its step-down: all resolutions inference (ARI, Rosenblatt et al. (2018)) which are described in Section 3.3.

4.2 False Positive control

In each simulation setting, for $1 \leq j \leq 5000$, we calculate a test statistic random field $T_n^{(j)}$ and obtain λ thresholds for the single-step bootstrap, step-down bootstrap, Simes and ARI methods, where we have used 1000 bootstraps for the non-parametric procedures. For each method we obtain λ-thresholds $\lambda_1, \ldots, \lambda_{5000}$ allowing us to estimate the joint error rate via the statistic

$$\frac{1}{5000} \sum_{j=1}^{5000} 1[f_{n,\mathcal{N}}(T_n^{(j)}) \leq \lambda_j]$$

which we refer to as the **empirical joint error rate**. Here $1[\cdot]$ denotes the indicator function.

The results for the 50 by 50 simulations are displayed in Figure 1 and those for the other domain sizes are shown in Figures S-5 and S-6. The results for the bootstrap methods are shown in blue whilst those for the parametric methods are shown in red. The solid lines indicate the step-down methods (i.e. ARI and the step-down bootstrap). These plots demonstrate that, given a reasonable number of subjects ($N \geq 80$) and smoothness of the underlying data (FWHM $= 4, 8$), the joint error rate of the bootstrap procedures converge to the nominal level, in this case 0.1.

Empirically the parametric procedures are valid in all settings considered. However, their control of the joint error rate is substantially below the nominal level when

the applied smoothing is non-zero, while the bootstrap approaches demonstrate tighter control. The step-down procedures provide an improvement on their single-step counterparts. This difference increases as π_0 decreases. See Section 4.3 for further details on the effect of π_0.

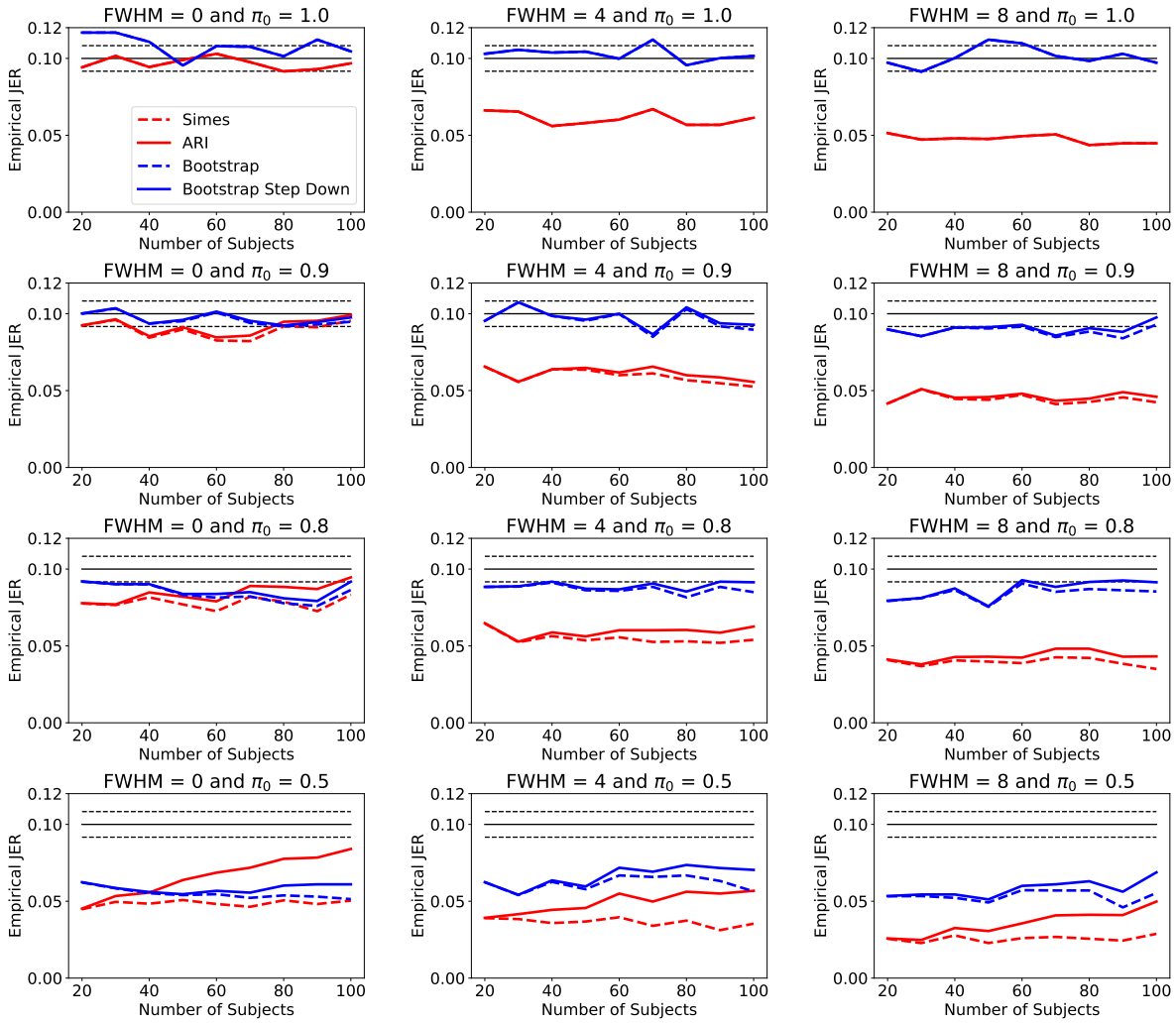

Figure 1: Comparing the empirical joint error rate across methods for the simulation setting described in Section 4.1 for $\alpha = 0.1$ on the domain of size 50 by 50 pixels. The bootstrap procedures typically provide tighter control of the joint error rate than the parametric ones, except under independence. The bootstrap methods are shown in blue whilst the parametric methods are shown in red. The solid lines indicate the step-down methods. The thin flat black dashed lines provide 95% marginal confidence bands based on the normal approximation to the binomial distribution.

4.3 Power

In this section we compare the power of the various methods in the simulation setting described in Section 4.1 in the case where the applied FWHM is 4 pixels. We have chosen to focus on this level of smoothness because it represents a realistic level of applied smoothness and illustrates the benefits that can be achieved when using the bootstrap under dependence.

Here we shall use a notion of power originally proposed in Blanchard et al. (2020) to compare the ability of joint error rate controlling procedures to detect signal. Given a set $R \subseteq \mathcal{H}$, define

$$\text{Pow}(R) := \mathbb{E}\left[\frac{|R| - \overline{V}(R)}{|R \cap (\mathcal{H} \setminus \mathcal{N})|}\middle| |R \cap (\mathcal{H} \setminus \mathcal{N})| > 0\right] \tag{19}$$

where for each method $\overline{V}$ is the corresponding post-hoc bound. Here we consider the following choices of R with which we compare the power (as in Blanchard et al. (2020)). 1) $R = \mathcal{H}$ and 2) taking R to be the hypotheses of $\mathcal{H}$ which are rejected by the Benjamini Hochberg procedure, applied to the p-values $\{p_{n,l}(v) : (l, v) \in \mathcal{H}\}$, at a level 0.05. Note that, unlike in Blanchard et al. (2020), no additional level of randomness in the choice of the sets in 2) is prescribed. We also consider taking $R = \{(l, v) : p_{n,l}(v) \leq 0.05\}$, see Section S-6.5, the results for which are similar in nature to scenario 1 from above. The results for cases 1) and 2) are illustrated graphically in Figure 2. These are for simulations on the 50 by 50 domain.

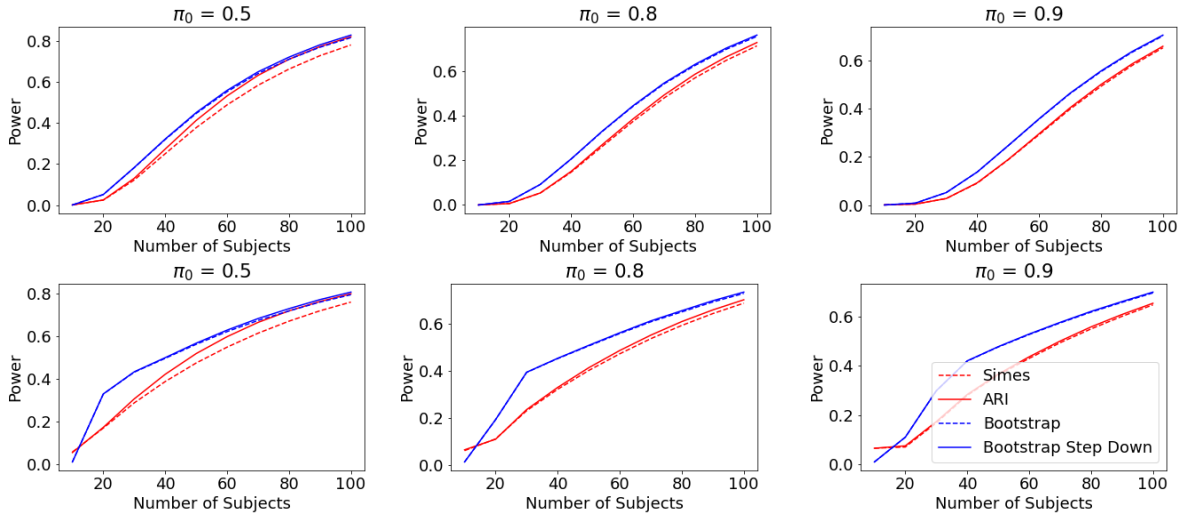

Figure 2: Plotting the power of the different methods against the number of subjects. The power for setting 1 (i.e. $R = \mathcal{H}$) is shown in the top row and the power for setting 2 (i.e. taking R to be the Benjamini-Hochberg rejection set) is shown in the bottom row.

From these plots we can see that overall the bootstrap based approaches have a higher power than the parametric ones. The power of ARI only becomes comparable (or higher) to that of the bootstrap in the extreme scenario ($\pi_0 = 0.5$) given a large enough sample size. Additionally the bootstrap is not robust at the smallest sample size considered (i.e. $n = 10$) where it is slightly conservative. However it is important to note that in typical high-dimensional applications (neuroimaging, genetics) $\pi_0 > 0.9$ and n is often substantially greater than 10.

The lower the value of π_0, the greater the increase in power that is obtained by using the step-down algorithms. ARI is always more powerful than Simes by construction. In the relatively sparse scenarios (i.e. $\pi_0 \geq 0.8$) they have a very similar power however for $\pi_0 = 0.5$, ARI provides a marked improvement over Simes. The bootstrap step-down always improves on the standard bootstrap approach though the difference is not particularly large: even when $\pi = 0.5$ this increase is relatively small. The similarity of the standard and step-down procedures, for both the parametric and bootstrap methods,

is consistent with the results obtained on real data which are described in the next subsections.

5 Real Data results

5.1 Neuroimaging data application

We have 3D functional Magnetic Resonance Imaging data from $n = 386$ unrelated subjects, who performed an m-back working memory task, from the Human Connectome Project. After pre-processing (described in Section S-5) we obtain a 3-dimensional contrast image for each subject. We fit a linear model to these images including sex, height, weight, body mass index, two different measures of blood pressure, handedness and IQ (measured using the PMAT24_A_CR test score). We consider sex and IQ as a variables of interest. We obtain test-statistic contrasts for sex and IQ and a p-value at each voxel for each contrast. We form clusters using a cluster defining threshold on the p-values of $p = 0.001$, with each cluster being a contiguous set of voxels above the threshold (clusters are defined separately for each contrast of interest).

We use our bootstrap framework, performing the resampling using 1000 bootstraps, to provide a lower bound on the proportion of active voxels within each cluster, taking $\alpha = 0.1$. This illustrates that multiple clusters, in different regions of the brain, have a relatively large proportion of active voxels for the contrast of IQ. For the contrast of sex only a single cluster has a non-zero lower bound on the number of true positives. The bounds provided using the step-down bootstrap procedure are the same as the single-step version in this example.

We compare to the results that are obtained using Simes and ARI bounds (taking $\alpha = 0.1$) and see that our bootstrap approach results in higher lower bounds on the number of active voxels. In this setting the bounds obtained by the parametric procedures are very similar to each other, which is not surprising given the sparsity of the signal. For the IQ contrast the lower bounds provided by the bootstrap and ARI for the number of true positives and on the TDP within each cluster are shown graphically in the upper panel of Figure 3. The corresponding plot for the sex contrast is shown in Section S-6.2. Direct comparison of the lower bounds is shown in Figure 4.

5.2 Transcriptomic data application

In this section, we illustrate the application of our methods to a specific gene expression data set. Gene expression studies use microarray or sequencing biotechologies in order to measure the activity (or "expression level") of a large number of genes simultaneously. We focus on a study of chronic obstructive pulmonary disease (COPD), see Bahr et al. (2013), whose main goal was to identify genes whose expression level is significantly associated with lung function. In order to do this, the authors fit a linear model for this association for each gene, while controlling for the following covariates: age, sex, body mass index, parental history of COPD, and two smoking variables (smoking status and pack-years). The number of subjects is $n = 135$ while the number of genes is $V = 12,531$, leading to a large-scale multiple testing problem. Using the Benjamini-Hochberg method to control the FDR at the 5% level, 1,745 genes were found to be significantly associated.

We fit this linear model to the data, regressing the gene level data against the controlled covariates and lung function and considering a single contrast for lung function.

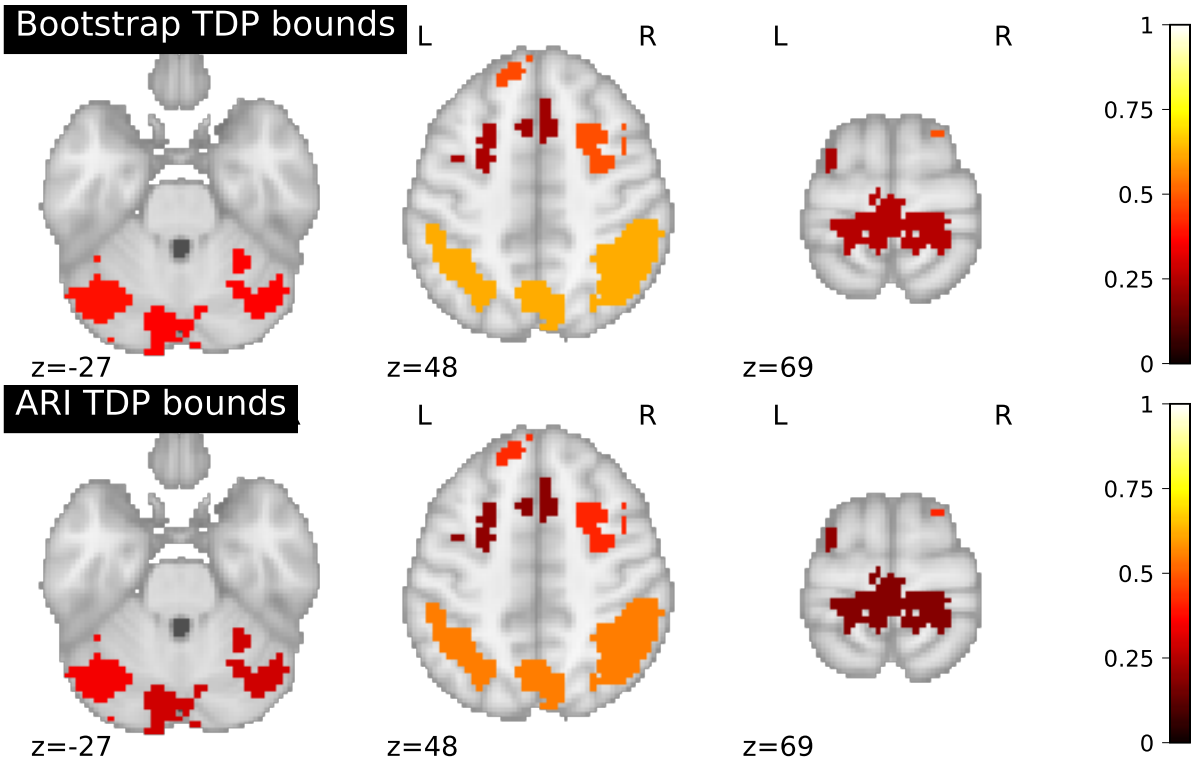

Figure 3: TDP bounds within clusters for the contrast for IQ in the linear regression model fit to the HCP data. Each cluster is shaded a single colour which is the lower bound on the TDP. The upper panel gives the TDP bounds within each cluster provided by the bootstrap procedure. The lower panel gives the bounds provided by using ARI. The bounds given by the bootstrap are larger (as indicated by the lighter colours) indicating that the method is more powerful. Note that these images are 2D slices through the 3D brain and so voxels that are part of the same cluster are not necessarily connected.

We performed 1000 bootstraps and used these to obtain $\lambda^*_{\alpha,135,1000} = 0.22$, where we took $\alpha = 0.1$. This allows us to provide a $(1-\alpha)$-level simultaneous lower bounds on the number of true positives within any specified set of genes. In particular it allows us to conclude (with 90% confidence) that at least 1,354 of the 1,745 genes within the Benjamini-Hochberg significance set are active. The stepdown bootstrap provides the same bound as the single-step version in this case. Simes and ARI provide lower bounds on the number of true positives in this set of 917 and 966 respectively, which are substantially less informative than the bootstrap bounds.

In the absence of prior information on genes, a natural idea is to rank them by decreasing statistical significance. Our post hoc methods provide upper confidence curves on the proportion of true positives among the most significant genes. Such curves are displayed in Figure 5, where the blue lines correspond to our proposed single-step and step-down bootstrap-based methods, and thered lines correspond to the parametric approaches of Goeman and Solari (2011) and Rosenblatt et al. (2018). These results are consistent with the numerical experiments of Section 4. First, the bootstrap method yields post hoc bounds that are substantially more informative that their parametric counterpart. Second, the difference between single-step methods and their step-down counterpart is very small, which is consistent with the fact that the signal is expected to be sparse in such genomic data sets, corresponding to π_0 close to 1. For the bootstrap

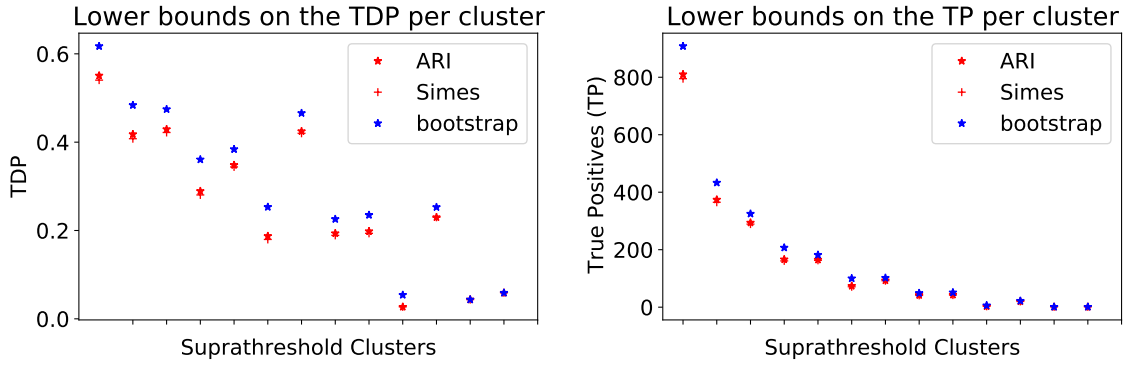

Figure 4: Comparing the TDP and true positive lower bounds across clusters for the different methods. The bootstrap lower bounds are consistently higher than the parametric methods. Clusters are organized from left to right in terms of their size. Only one cluster for the sex contrast is found: this is the 2nd smallest cluster overall with a TP lower bound of 1 voxel. The sizes and bounds of the clusters in the IQ contrast are larger. For the largest cluster we are able to conclude that it contains 908 true positives using the bootstrap approach.

there is in fact no difference between the single-step and step-down approach in this example.

A widely used approach in differential expression studies is to select genes based on the conjunction of a threshold on the p-values and a threshold on its effect size (Cui and Churchill, 2003). Ebrahimpoor et al. (2020) recently noted that this type of double selection can lead to inflated numbers of false discoveries when used in conjunction with FDR-based multiple testing corrections, whereas post hoc inference is by construction robust against this issue. The use of our proposed post hoc bounds in this context is illustrated in the volcano plot in Figure 6 (Cui and Churchill, 2003). In this plot, each gene is represented in two dimensions by estimates of its effect size (x axis, also kown as "fold change" in genomics) and p-value (y axis), in a logarithmic scale. Figure 6 illustrates a particular selection, corresponding to the genes whose p-value is below 0.001 and whose effect size is above 0.5. Our bootstrap-based bound ensures that with probability $1 - \alpha = 90\%$, among these 546 genes, at least 490 are true positives, corresponding to a FDP below 0.1. Importantly, the p-value and effect size thresholds can be chosen post hoc, and multiple such choices can be made without compromising the statistical coverage of the associated bound. For example, the bounds associated to the gene subsets with positive and negative effect size are also displayed in Figure 6.

6 Discussion

In this paper we have introduced a bootstrap method which provides simultaneous control of the FDP over subsets of hypotheses of multiple contrasts in the linear model. We have proved the asymptotic validity of this approach and shown, via simulation, that the error rate is controlled to the correct level given a reasonable number of subjects.

From our simulations and real data examples, we can see that the bootstrap approach typically provides better bounds than existing, state of the art parametric methods (i.e. Simes and ARI). This occurs because we are able to model the dependence within the

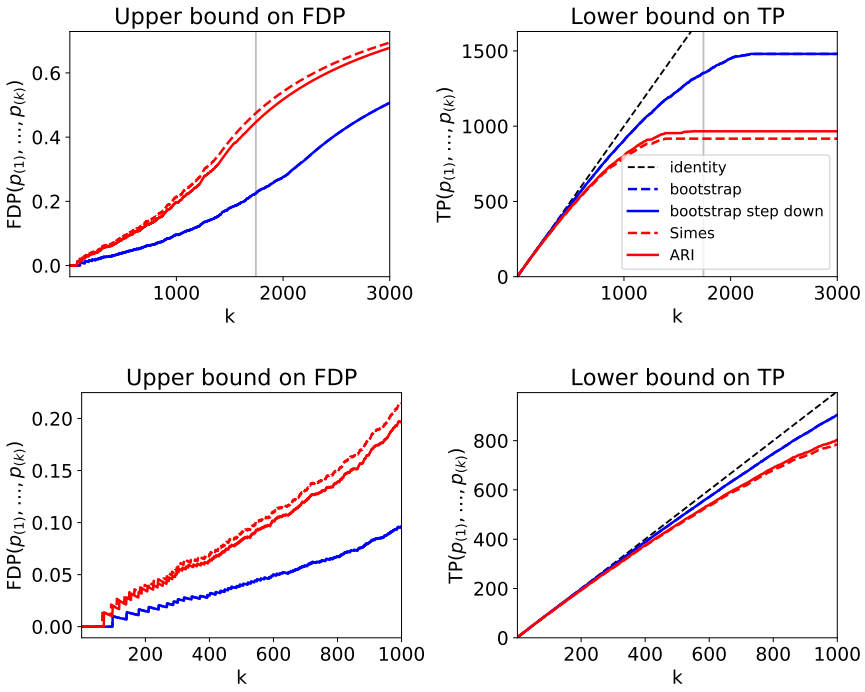

Figure 5: False discovery proportion and true positive plots for the transcriptomic dataset. In the upper panels, for $k = 1, \ldots, 3000$, upper bounds on the FDP and lower bounds on the number of true positives are provided by each of the methods for the sets comprised of the hypotheses with the k smallest p-values. The silver vertical line corresponds to the location of the Benjamini-Hochberg rejection set. The lower panels provide a zoomed in version of the same plot for for the 1000 smallest p-values. The bootstrap methods provide substantially better bounds than the parametric ones. ARI slightly improves on Simes while the step-down bootstrap is indistinguishable from the single-step bootstrap approach in this setting.

data. The parametric methods, on the other hand, rely on the Simes inequality which is only exact under independence. Moreover the Simes inequality is only valid under positive regression dependence whereas the non-parametric bootstrap makes relatively few assumptions other than finite moments of the noise and the design. In real data situations there is often relatively strong dependence within the data in which case we would expect the bootstrap to give better bounds. This is illustrated in our brain imaging and transcriptomic examples where the bootstrap bounds provided substantial improvements over the ones derived using the parametric methods. Overall, our results are consistent with those obtained by Enjalbert-Courrech and Neuvial (2022) in the specific case of two-sample tests, where post hoc bounds based on non-parametric joint error rate control substantially outperformed their parametric counterpart.

The step-down bootstrap approach improves the power whilst maintaining control of the error rate. However in practice, as illustrated in the real datasets, the improvement is likely to be small as π_0 will be close to 1. Indeed in both of our real data examples there is no noticeable improvement. The improvement of ARI over Simes is typically non-zero but is rather small. These results demonstrate that the step-down methods, whether parametric or otherwise, appear to require a relatively small value of π_0 before they substantial improvement on their single step versions. It is worth noting that the improvement in the bounds provided by ARI relative to Simes is greater than that of

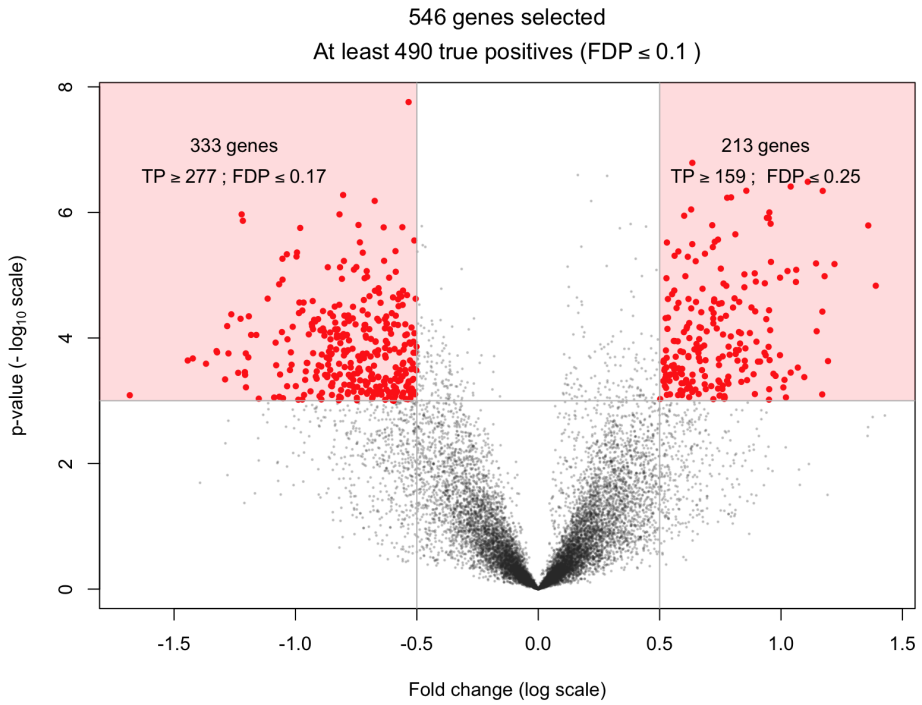

Figure 6: A volcano plot for the p-values for the transcriptomic data. For each gene this plots the estimated contrast effect size (labelled as fold change and corresponding to $c^T\hat{\beta}_{135}$ where c is the contrast vector for COPD) against the p-value, where both are measured in log scale. Two regions (shown via shading) are selected containing the genes whose p-values are less than 10^{-3} and for which the absolute fold change is greater than $10^{0.5}$. Bounds on the true positives (TP) and FDP, overall and for the shaded regions are provided on the plot.

the step-down bootstrap relative to the single-step version. One possible reason for this discrepancy is that in the step-down bootstrap (Algorithm 1), only the first threshold t_1 is used at each step. This implies that only part of the information on the true positives is exploited. One advantage of ARI over the bootstrap is computational time. However the bootstrap methods are relatively fast, for instance the neuroimaging application took under 10 minutes when run in serial with 1000 bootstraps, so the extra computational burden is a small price to pay for the increase in power. This burden can be decreased further by parallelizing as the bootstrap is trivially parallelizeable.

It is important to note that for the bootstrap approach, control of the FDP is asymptotic. In Section 4.2 we showed that, given reasonable sample sizes and smoothness levels (e.g. $n \geq 80$ for FWHM $= 4, 8$), the joint error rate was controlled at the correct rate. At low smoothness levels and sample sizes the error rate can be slightly inflated. In particular the convergence of the bootstrap when FWHM $= 0$ (in which the pixels are independent) is slower than when the smoothness is higher. Under independence we would thus recommend the use of ARI over resampling methods since, as discussed above, the Simes inequality is exact. At very small sample sizes, e.g. $n = 10$, the bootstrap can also be conservative (see Figure 2). However at the smoothness levels and sample sizes used in real data analyses, based on our theoretical and simulation results, we would expect the bootstrap to control the joint error rate to the desired level.

In order to obtain post hoc bounds on the FDP, we have followed the construction of

Blanchard et al. (2020), where the FDP control is obtained the joint error rate control by a straightforward interpolation argument. The resulting bounds are then computed in linear time. This construction differs from the initial method of Goeman and Solari (2011), in which post hoc bounds are obtained from closed testing. These authors have shown that closed testing-based post hoc bounds enjoy an admissibility property Goeman et al. (2021), which makes these bounds optimal in theory. However, such bounds are combinatorial in nature and dedicated computational shortcuts are required for them to be computed when the number of features exceeds a few dozens. In particular, the bounds implemented in Andreella et al. (2023) for permutation-based control of the FDP are in fact obtained by interpolation (see Theorem 1 in Andreella et al. (2023).

The template $(t_k)_{1\leq k\leq K}$ is a free parameter of the proposed method, and optimising this choice for a particular application could lead to tighter TDP bounds. For the numerical experiments of this paper, we have only considered the linear template, for which $t_k(\lambda) = \frac{\lambda k}{m}$, which is the most widely used in the post hoc inference literature (Goeman et al., 2019) and in particular for neuroimaging applications (Rosenblatt et al., 2018). Other parametric templates are considered in Blanchard et al. (2020); Andreella et al. (2023) and could readily be combined with the bootstrap-based approach introduced here. However, the experiments reported in these papers suggest that the linear template is a sensible choice when the feature subsets of interest are p-value level sets, as is often the case in the applications considered here. A natural idea to go beyond parametric templates is to learn from the data the shape of the template itself. This has been advocated by Meinshausen (2006), but the proof of the proposed method is invalid as it suffers from a circularity issue (Blanchard et al., 2020, Remark 5.3). Recently, Blain et al. (2022) used an independent data set to learn the optimal template in the case of one- and two-sample testing. Extending this idea to multivariate linear models as considered in the present paper is an interesting perspective for future research.

The idea of using the Lindeberg CLT has been used to establish bootstrap consistency for the standard non-parametric bootstrap (van der Vaart (1998)) and the multiplier bootstrap Kosorok (2003). It has also been used to prove certain results for high-dimensional linear models (Mammen, 1993) and for univariate robust regression (Shorack, 1982). However, as far as we are aware, it has not previously been used to establish bootstrap consistency in the multivariate linear model. Note that in this setting the quantities that need to be bootstrapped are the residuals which are dependent, as such existing consistency results for the non-parametric/multiplier bootstrap (van der Vaart (1998)/Kosorok (2003)) cannot be applied as they assume independence of the bootstrapped quantities. Moreover because we are bootstrapping the t-statistics we needed to additionally establish convergence in probability of the bootstrapped variance. This is a consequence of Theorem 1 of Eck (2018). However in the proof of Theorem S-2.5, where we have used an alternative proof to establish bootstrap consistency without relying on convergence in the Mallows metric, this aspect was tricky. To do so we first established a uniform bound on the expectation of the fourth moment of the bootstrapped residuals, see Lemma S-2.4 and used this in combination with the triangular weak law of large numbers to demonstrate convergence of the bootstrapped variance.

Our results can also be used to provide strong control of the familywise error rate over multiple contrasts (see Section S-4.2 for a proof, a formal discussion of this and the results of simulations). This comes about in two different ways. Firstly it arises as a direct consequence of joint error rate control when using the canonical reference family and taking $\zeta_k = k - 1$. Secondly the familywise error rate can be targeted directly, along

the lines of the approach of Westfall (2011) (i.e. not simultaneously with joint error rate control), this follows from Theorem 3.1 by taking $K = 1$, a result that is stated formally in Theorem S-4.1. There are a variety of methods to provide FWER control in the linear model but most of them are not suitable for the case where some of the variables of interest can take non-zero values. For instance Manly based permutation (Manly, 1986) provides weak, rather than strong control when there are multiple covariates in the model which may or may not be non-zero. This occurs because Manly permutation acts by permuting the Ys and thus does not generate resamples under the full null hypothesis - see Section S-4.3 for details. Freedman-Lane (Freedman and Lane, 1983) - another commonly used method - encounters similar issues. The bootstrap is able to avoid these issues, and thus provide strong control, because it centres the residuals before resampling. As discussed below, other forms of permutation testing could be used to provide the desired error rate control as an alternative to the bootstrap.

An alternative approach to controlling the joint error rate (and the FWER) over multiple contrasts could be developed by considering permutations of the residuals rather than bootstrapping. Importantly, like the bootstrap, methods based on permuting the residuals via these methods are typically only valid asymptotically. This is because, when dealing with multiple contrasts in the linear model, exchangeability does not hold and so permutation is not exact: see Section S-4.3 for further details. Permutation testing based methods in the linear model are very widespread (Winkler et al., 2014). As such establishing consistency results for these methods in the multivariate setting and using these to prove results on asymptotic joint error rate control is an interesting avenue for future research.

The choice of the error rate to control in a scenario where many hypotheses are being tested depends strongly on the goals of the researcher. The bounds that we have provided on the FDP provide more informative inference than simply controlling the FDR. As discussed in Neuvial (2020), under dependence controlling the FDR can lead to non-nonsensical results. Instead bounds on the FDP allow statements about the number of active voxels with a given set to be made. Moreover this inference is valid simultaneously over all sets and so guards against circular inference.

Acknowledgments

We are grateful to Alexandre Blain at Inria for his help with demonstrating how to use the sanssouci code to get ARI to work and for checking that the Python implementation was consistent with the implementation in the ARIbrain R package. We are grateful to François Bachoc at the University of Toulouse for his help with the proof of Lemma S-4.4. SD is grateful to Fabian Telschow at Humboldt University for useful discussions on bootstrapping.

SD was supported by NIH grant R01EB026859. SD and PN were supported by the SansSouci ANR project (ANR-16-CE40-0019). BT was supported by the KARAIB AI chair (ANR-20-CHIA-0025-01) and the FastBig ANR project (ANR-17-CE23-0011).

Data were provided in part by the Human Connectome Project, WU-Minn Consortium (Principal Investigators: David Van Essen and Kamil Ugurbil; 1U54MH091657) funded by the 16 NIH Institutes and Centers that support the NIH Blueprint for Neuroscience Research; and by the McDonnell Center for Systems Neuroscience at Washington University.

References

Bianca AV Alberton, Thomas E Nichols, Humberto R Gamba, and Anderson M Winkler. Multiple testing correction over contrasts for brain imaging. *NeuroImage*, 216:116760, 2020.

Angela Andreella, Jesse Hemerik, Livio Finos, Wouter Weeda, and Jelle Goeman. Permutation-based true discovery proportions for functional magnetic resonance imaging cluster analysis. *Statistics in Medicine*, 2023.

Timothy M Bahr et al. Peripheral blood mononuclear cell gene expression in chronic obstructive pulmonary disease. *American journal of respiratory cell and molecular biology*, 49(2):316–323, 2013.

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300, 1995.

Yoav Benjamini and Daniel Yekutieli. The control of the false discovery rate in multiple testing under dependency. *Annals of statistics*, pages 1165–1188, 2001.

Peter J Bickel and David A Freedman. Some Asymptotic Theory for the Bootstrap. *Annals of Statistics*, 9(6):1196–1217, 1981. ISSN 00905364. doi: 10.1214/aos/1176342871.

Alexandre Blain, Bertrand Thirion, and Pierre Neuvial. Notip: Non-parametric true discovery proportion control for brain imaging. *NeuroImage*, page 119492, 2022. URL `https://hal.archives-ouvertes.fr/hal-03649114`.

Gilles Blanchard, Pierre Neuvial, Etienne Roquain, et al. Post hoc confidence bounds on false positives using reference families. *Annals of Statistics*, 48(3):1281–1303, 2020.

Gilles Blanchard, Pierre Neuvial, and Etienne Roquain. On agnostic post hoc approaches to false positive control. In *Handbook of Multiple Comparisons*, Handbooks of Modern Statistical Methods. Chapman & Hall/CRC, November 2021. URL `https://hal.archives-ouvertes.fr/hal-02320543`.

Xiangqin Cui and Gary A Churchill. Statistical tests for differential expression in cDNA microarray experiments. *Genome Biol*, 4(4):210, 2003.

Samuel Davenport and Thomas E. Nichols. The expected behaviour of random fields in high dimensions: contradictions in the results of [1]. *Magnetic Resonance Imaging*, 2022.

Mitra Ebrahimpoor, Pietro Spitali, Kristina Hettne, Roula Tsonaka, and Jelle Goeman. Simultaneous enrichment analysis of all possible gene-sets: unifying self-contained and competitive methods. *Briefings in bioinformatics*, 21(4):1302–1312, 2020.

Daniel J Eck. Bootstrapping for multivariate linear regression models. *Statistics & Probability Letters*, 134:141–149, 2018.

Nicolas Enjalbert-Courrech and Pierre Neuvial. Powerful and interpretable control of false discoveries in two-group differential expression studies. *Bioinformatics*, 38(23): 5214–5221, 2022. doi: 10.1093/bioinformatics/btac693.

David Freedman and David Lane. A nonstochastic interpretation of reported significance levels. *Journal of Business & Economic Statistics*, 1(4):292–298, 1983.

David A Freedman. Bootstrapping regression models. *The Annals of Statistics*, 9(6): 1218–1228, 1981.

Christopher R Genovese and Larry Wasserman. Exceedance control of the false discovery proportion. *Journal of the American Statistical Association*, 101(476):1408–1417, 2006.

Christopher R Genovese, Nicole A Lazar, and Thomas E Nichols. Thresholding of statistical maps in functional neuroimaging using the false discovery rate. *Neuroimage*, 15(4):870–878, 2002.

Jelle J. Goeman and Aldo Solari. Multiple Testing for Exploratory Research. *Statistical Science*, 26(4):584–597, 2011. ISSN 0883-4237. doi: 10.1214/11-STS356.

Jelle J Goeman, Rosa J Meijer, Thijmen JP Krebs, and Aldo Solari. Simultaneous control of all false discovery proportions in large-scale multiple hypothesis testing. *Biometrika*, 106(4):841–856, 2019.

Jelle J Goeman, Jesse Hemerik, and Aldo Solari. Only closed testing procedures are admissible for controlling false discovery proportions. *The Annals of Statistics*, 49(2): 1218–1238, 2021.

Jesse Hemerik, Aldo Solari, and Jelle J Goeman. Permutation-based simultaneous confidence bounds for the false discovery proportion. *Biometrika*, 106(3):635–649, 2019.

R. Henson. Forward inference using functional neuroimaging: Dissociations versus associations. *Trends in cognitive sciences*, 10:64–69, 2006.

Gerhard Hommel. A stagewise rejective multiple test procedure based on a modified bonferroni test. *Biometrika*, 75(2):383–386, 1988.

Edward L Korn, James F Troendle, Lisa M McShane, and Richard Simon. Controlling the number of false discoveries: application to high-dimensional genomic data. *Journal of Statistical Planning and Inference*, 124(2):379–398, 2004.

Michael R Kosorok. Bootstraps of sums of independent but not identically distributed stochastic processes. *Journal of Multivariate Analysis*, 84(2):299–318, 2003.

Enno Mammen. Bootstrap and wild bootstrap for high dimensional linear models. *The annals of statistics*, 21(1):255–285, 1993.

Bryan FJ Manly. Randomization and regression methods for testing for associations with geographical, environmental and biological distances between populations. *Researches on Population Ecology*, 28(2):201–218, 1986.

Nicolai Meinshausen. False discovery control for multiple tests of association under general dependence. *Scandinavian Journal of Statistics*, 33(2):227–237, 2006.

Pierre Neuvial. *Contributions to statistical inference from genomic data*. Habilitation thesis, Université Toulouse III (France), 2020. Available from `https://tel.archives-ouvertes.fr/tel-02969229`.

Joseph P Romano and Michael Wolf. Exact and Approximate Stepdown Methods for Multiple Hypothesis Testing. *Journal of the American Statistical Association*, 100 (469):94–108, 2005. ISSN 0162-1459. doi: 10.1198/016214504000000539.

Jonathan D Rosenblatt, Livio Finos, Wouter D Weeda, Aldo Solari, and Jelle J Goeman. All-resolutions inference for brain imaging. *Neuroimage*, 181:786–796, 2018.

Sanat K Sarkar et al. On the simes inequality and its generalization. In *Beyond parametrics in interdisciplinary research: Festschrift in honor of Professor Pranab K. Sen*, pages 231–242. Institute of Mathematical Statistics, 2008.

Galen R Shorack. Bootstrapping robust regression. *Communications in statistics-theory and methods*, 11(9):961–972, 1982.

R J Simes. An improved Bonferroni procedure for multiple tests of significance. *Biometrika*, 73(3):751–754, 1986.

Gordon K. Smyth. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Statistical Methods in Genetics and Molecular Biology*, 3(3), 2004. doi: 10.2202/1544-6115.1027.

John D Storey and Robert Tibshirani. Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences*, 100(16):9440–9445, 2003.

A.W. van der Vaart. *Asymptotic Statistics*. 1998.

Peter H. Westfall. On Using the Bootstrap for Multiple Comparisons. *Journal of Biopharmaceutical Statistics*, 21(6):1187–1205, 2011. ISSN 1054-3406. doi: 10.1080/10543406.2011.607751. URL `http://www.tandfonline.com/doi/abs/10.1080/10543406.2011.607751`.

Anderson M. Winkler, Gerard R. Ridgway, Matthew A. Webster, Stephen M. Smith, and Thomas E. Nichols. Permutation inference for the general linear model. *NeuroImage*, 92:381–397, 2014. ISSN 10959572. doi: 10.1016/j.neuroimage.2014.01.060.