Notip: Non-parametric True Discovery Proportion estimation for brain imaging

Alexandre Blain, Bertrand Thirion, Pierre Neuvial

Notip: Non-parametric True Discovery Proportion estimation for brain imaging

Alexandre Blain[a,b], Bertrand Thirion[a], Pierre Neuvial[b]

[a]*Inria, CEA, Université Paris-Saclay, Paris, France*
[b]*Institut de Mathématiques de Toulouse, UMR 5219, Université de Toulouse, CNRS, UPS, Toulouse, France*

Abstract

Cluster-level inference procedures are widely used for brain mapping. These methods compare the size of clusters obtained by thresholding brain maps to an upper bound under the global null hypothesis, computed using Random Field Theory or permutations. However, the guarantees obtained by this type of inference - i.e. at least one voxel is truly activated in the cluster - are not informative with regards to the strength of the signal therein. There is thus a need for methods to assess the amount of signal within clusters; yet such methods have to take into account that clusters are defined based on the data, which creates circularity in the inference scheme. This has motivated the use of *post hoc* estimates that allow statistically valid estimation of the proportion of activated voxels in clusters. In the context of fMRI data, the All-Resolutions Inference framework introduced in [24] provides post hoc estimates of the proportion of activated voxels. However, this method relies on parametric threshold families, which results in conservative inference. In this paper, we leverage randomization methods to adapt to data characteristics and obtain tighter false discovery control. We obtain Notip: a powerful, non-parametric method that yields statistically valid estimation of the proportion of activated voxels in data-derived clusters. Numerical experiments demonstrate substantial power gains compared with state-of-the-art methods on 36 fMRI datasets. The conditions under which the proposed method brings benefits are also discussed.

1. Introduction

The mapping of the human brain consists in associating regions of the brain with cognitive functions or disorders. This is important both for basic neuroscience, e.g. the understanding of brain function, and medical applications, as it allows to identify regions that carry disease-related signal. The most popular modality to map brain function is functional Magnetic Resonance Imaging (fMRI), as it is non-invasive and offers decent spatial resolution (about $2mm$ isotropic) and full brain coverage.

FMRI data are sampled on a discrete 3D lattice and subject to various preprocessing steps [10], resulting in a set of *voxels* that contain a signal that reflects brain activity. After suitable statistical analysis, relevant brain territories can be reported. More precisely, practitioners define a *contrast*, that is, a linear combination of a set of images, typically corresponding to the comparison between two or more conditions or groups of participants, and seek to test hypotheses $\mathbf{H_{0,i}}$: "Voxel i is *inactive* for this contrast", meaning that it does not show any effect for the selected contrast, versus $\mathbf{H_{1,i}}$: "Voxel i is *active* for this contrast". This statistical problem entails a dire multiple testing issue as described in [11], as standard fMRI images comprise between $50k$ and $400k$ voxels (growing to millions with the development of high-resolution imaging).

In this context, if multiplicity is not accounted for, the number of false discoveries is unacceptably high. In other words, mere voxel-wise type 1 error control is not appropriate in the context of multiplicity. Family-Wise Error Rate (FWER) control can be used in this setting [11] but it is conservative, resulting in false negatives, which hurts reproducibility (see e.g. [27, 7]). A more powerful and commonly used approach is to control the False Discovery Rate (FDR) [13], which is systematically done using Benjamini-Hochberg procedure [3]. A caveat to this approach is that the FDR actually corresponds to the *expected* False Discovery Proportion (FDP). As noted by several authors [12, 16, 20], FDR control does not guarantee FDP control.

An alternative type of inference to increase statistical power is to perform inference at cluster-level, rather than voxel-level [22], because brain activation is organised in compact regions *(clusters)* in the brain volume. This type of inference tests whether regions above a given threshold are larger than expected under the null hypothesis, or whether the total amount of signal in these regions [26] exceeds its expected value under

a null distribution. However, this approach suffers from several problems [8], such as the arbitrary choice of cluster-forming threshold [30], or the difficulty to establish a null distribution for cluster size and aggregated signal. To address this last issue, reliable non-parametric solutions have been proposed [29, 8]. However, the arbitrariness regarding cluster-forming threshold is hard to deal with. To overcome it, one may define such clusters or regions, and *then* assess the proportion of active voxels in each region, i.e the True Discovery Proportion, TDP = 1 - FDP. Such a region of interest could be defined a priori, using an anatomical atlas, or a posteriori, based on the fMRI data. For instance, one might wonder what is the proportion of active regions in a *blob*, i.e. a contiguous set of statistical values that are higher than the image background. Yet, such a definition of the clusters after seeing the data raises a double-dipping issue, which can lead to massive false positive inflation [17].

To illustrate this statistical bias, let us consider a classical example of invalid post-selection inference. Users often perform a first round of tests to identify potentially interesting regions (i.e., regions comprising significant signal). If inference is performed only on smallest p-values obtained at this first round, then the FDP is not controlled, as shown in [6]. To bypass this double-dipping issue, one can use *post hoc* estimates that control the FDP. The first method of that kind is a parametric method called All-resolutions inference (ARI) [24].

In this paper, we introduce the Notip procedure, that adapts non-parametrically to data correlation structure. We study whether such a procedure can yield superior power while offering the same statistical guarantees. We perform extensive experiments on dozens of fMRI datasets to compare the number of detections obtained by this approach with that of existing methods.

The paper is organized as follows. In Section 2, existing methods for the post hoc control of FDP are introduced via the notion of *Joint Error Rate* (JER) proposed by [6]. Our main contribution is the Notip method presented in Section 3: a non-parametric data-driven approach that relies on the JER framework to obtain sharper post hoc FDP control. Numerical experiments and results on fMRI data reported in Sections 4 and 5 show that substantial power gains can be obtained from the proposed method, while controlling the FDP of the detected regions at a fixed level. Finally, we discuss the benefits of our proposed methodology, and outline some possible limitations.

2. False Discovery Proportion control by Joint Error Rate control

The point of this article is to build an inference method that takes into account multiplicity and circularity by achieving post hoc FDP control, while maintaining satisfactory statistical power.

2.1. Notation

We denote by m the number of hypotheses, i.e. the number of voxels under consideration (typically spanning a given brain template). In the context of fMRI, m generally ranges from $50,000$ to $400,000$. We denote the set of true null hypotheses (voxels with no effect) by H_0, and by $m_0 = |H_0|$ its cardinal. Given a set of m p-values associated to each hypothesis, we denote by $p_{(k:m)}$ the k^{th} one in ascending order. For a set S of hypotheses of interest (i.e. the set of voxels in a region of interest), the aim is to control the number of false positives in S, that is $|S \cap H_0|$, or equivalently, the corresponding proportion of false positives: $\mathrm{FDP}(S) = |S \cap H_0| \,/\, |S|$.

2.2. Post hoc FDP control

The most common approach to address large-scale multiplicity problems is to control the False Discovery Rate (FDR) [3]. This is generally done by the Benjamini-Hochberg (BH) procedure [3], which uses different significance thresholds depending on the ranks of the p-values: the k^{th} p-value is compared to $t_k^{Simes} = \alpha k/m$. The BH procedure controls the FDR under the PRDS (Positive Regression Dependency on a Subset) assumption [4]. However, since the FDR is the **expected** FDP, FDR control is a weak statistical guarantee on the **actual** FDP [16]. This can be problematic when the FDP distribution has heavy tails, which can happen when the tested hypotheses are dependent. In such cases, the FDR might be controlled while FDP quantiles diverge (see Figure 2.1 in [21]). We thus choose to focus on the control of the actual number (or proportion) of false positives.

A post hoc upper bound V on the number of false positives is an integer-valued function of subsets S of hypotheses that satisfies:

$$\mathbb{P}\left(\forall S,\ |S \cap H_0| \leq V(S)\right) \geq 1 - \alpha\,. \tag{1}$$

Since $\mathrm{FDP}(S) = |S \cap H_0| \,/\, |S|$, obtaining a bound V satisfying (1) is strictly equivalent to obtaining a post hoc upper bound on the FDP. This equivalence will be used implicitly throughout the paper.

As described in [14], the comparison between ordered p-values and $\left(t_k^{Simes}\right)_{k=1..m}$ can also provide post hoc FDP control. This can be done using closed testing [18] combined with the following inequality:

$$\mathbb{P}\left(\exists k \in \{1,\dots,m_0\} : p_{(k:m_0)} < t_k^{Simes}\right) \leq \alpha\,. \tag{2}$$

Equation (2) is an immediate consequence of the Simes inequality [25], and also holds under the PRDS assumption. The All-resolutions inference (ARI) method [24] provides a tighter post hoc bound that uses the thresholds $\alpha k/h(\alpha)$ instead of $\alpha k/m = t_k^{Simes}$ in (2), where $h(\alpha) \leq m$ is the so-called Hommel value [15]. $h(\alpha)$ represents an $1-\alpha$-level upper confidence bound on the number m_0 of true null hypotheses.

2.3. Joint Error Rate

An alternative construction of post hoc bounds has been introduced by [6]. Letting $R_k^{Simes} = \left\{i : p_i \leq t_k^{Simes}\right\}$, Equation (2) can be written as:

$$\mathbb{P}\left(\forall k, \left|R_k^{Simes} \cap H_0\right| \leq k-1\right) \geq 1-\alpha\,. \tag{3}$$

Equation (3) can be interpreted as the simultaneous control of all $k-$Family-Wise Error Rate (FWER), where the $k-$FWER is the probability of obtaining at least k false positives. Each set R_k^{Simes} yields a valid FDP upper bound over any subset S:

$$\begin{aligned}
|S \cap H_0| &= \left|S \cap \overline{R_k^{Simes}} \cap H_0\right| + \left|S \cap R_k^{Simes} \cap H_0\right| \\
&\leq \left|S \cap \overline{R_k^{Simes}}\right| + \left|R_k^{Simes} \cap H_0\right| \\
&= \sum_{i \in S} 1\left\{p_i(X) \geq t_k^{Simes}\right\} + \left|R_k^{Simes} \cap H_0\right| \\
&\leq \sum_{i \in S} 1\left\{p_i(X) \geq t_k^{Simes}\right\} + k - 1 \\
&=: V_k^{Simes}(S)\,,
\end{aligned}$$

where the last inequality holds with probability at least $1-\alpha$ by (3).

The computation of $V_k^{Simes}(S)$ is illustrated in the top panels of Figure 1 for $k \in \{1,3,6\}$. Since (3) holds simultaneously for all k, the minimum over k of all $V_k^{Simes}(S)$ is a valid upper bound on the false positives in S [6]. Therefore, as illustrated in the bottom panel of Figure1, the final post hoc FDP upper bound is $V^{Simes}(S)/|S|$, where

$$V^{Simes}(S) = \min_{1 \leq k \leq |S|} \left\{ \sum_{i \in S} 1\left\{p_i(X) \geq t_k^{Simes}\right\} + k - 1 \right\}. \tag{4}$$

The bound $V^{Simes}(S)$ is called an *interpolation* bound, as it generalizes statistical control from a given family $(R_k)_k$ to any subset S of hypotheses.

As noted by [6], the bound (4) coincides with the bound originally proposed by [14]. This can be generalized as follows by replacing $t^{Simes} := (t_k^{Simes})_{1 \leq k \leq m}$ with any threshold family $t := (t_k)_{1 \leq k \leq k_{max}}$ corresponding to $R_k = \{i : p_i \leq t_k\}$. Here, the k_{max} parameter controls the length of threshold families. This can be exploited when the signal is a priori parsimonious, as discussed in Section 8.1. The Joint Error Rate (JER) of the threshold family t is defined by [6] as:

$$JER(t) = \mathbb{P}\left(\exists k \in \{1,\dots,k_{max} \wedge m_0\} : p_{(k:m_0)} < t_k\right). \tag{5}$$

With this notation, both Equations 2 and 3 are equivalent to $\mathrm{JER}(t^{Simes}) \leq \alpha$. By the interpolation argument outlined above, the bound

$$V^t(S) = \min_{1 \leq k \leq |S| \wedge k_{max}} \left\{ \sum_{i \in S} 1\left\{p_i(X) \geq t_k\right\} + k - 1 \right\} \tag{6}$$

provides a valid FDP upper bound for any threshold family t such that $\mathrm{JER}(t) \leq \alpha$ [6]. This bound can be calculated in $O(|S|)$ for a given set S using Algorithm 1 in [9].

2.4. Tighter FDP upper bounds via randomization

The Simes inequality (2) ensures JER control at level at most α for the threshold family $(\alpha k/m)_k$. While this control is sharp for independent p-values, it is typically conservative for positively dependent p-values [6], leading to conservative FDP bounds. The first degree of freedom that can be leveraged to obtain tighter bounds for a given α is to choose the least conservative threshold family among a pre-defined set of families. In the case of the Simes family, this is done by choosing the threshold family $(\lambda k/m)_k$ associated to the largest λ such that the following inequality holds:

$$\mathbb{P}\left(\exists k \in \{1,\dots,m_0\} : p_{(k:m_0)} < \frac{\lambda k}{m}\right) \leq \alpha\,. \tag{7}$$

In order to reach this goal more generally, we consider collections of threshold families called **templates** as introduced in [6]. Formally, a template is set of functions $\lambda \mapsto (t_k(\lambda))_k$

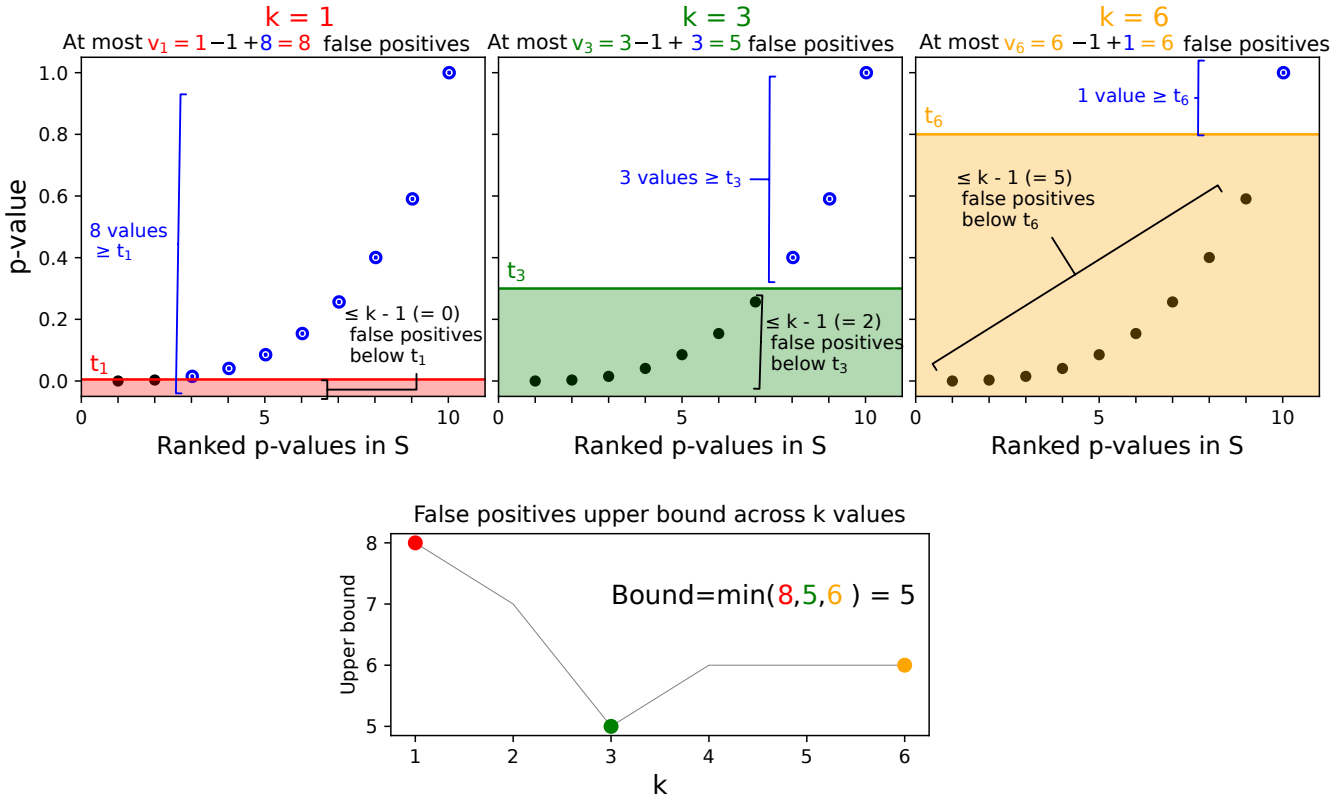

Figure 1: **Computation of the post hoc bound** (6) **on the number of false positives**, given a set of 10 p-values, using a JER controlling threshold family. Top panels: computation of k-th bound for 3 template indices k, with horizontal colored lines representing the associated thresholds t_k. Bottom panel: post hoc bound computation, which corresponds to the minimum of all k-th bounds. In that case, we find that the number of false positives in the set of 10 p-values is no more than 5.

such that any fixed value of λ corresponds to a threshold family. For example, the Simes template corresponds to the choice: $t_k(\lambda) = \lambda k/m$ for all $k = 1 \dots m$ and $\lambda > 0$.

The **calibration** procedure introduced in [5, 6] uses randomization (see [2]) to obtain samples from the joint distribution of p-values under the null hypothesis. As the JER (5) is a function of this distribution, these so-called randomized p-values allow us to select the largest possible λ such that the JER is controlled. Algorithm 1 describes how to compute such randomized p-values in the case of one-sample tests, using sign-flipping [23, 2]. Randomized p-values can also be obtained for two-sample tests using class label permutations instead of sign-flipping.

Algorithm 1 Computing randomized p-values using sign-flipping. For a number B of sign-flips, compute p-values using a one-sample t-test on the flipped data $X_{flipped}$.

function GET_RANDOMIZED_P_VALUES(X, B)
    $n, p \leftarrow$ shape(X)
                ▷ n subjects, p voxels
    pval0 $\leftarrow$ zeros(B, p)
    for $b \in [1, B]$ **do**
        flip $\leftarrow$ diag(draw_random_vector($\{-1, 1\}^n$))
                ▷ matrix of shape (n, n)
        $X_{flipped} =$ flip.X
        pval0[b] $\leftarrow$ one_sample_t_test($X_{flipped}$, 0)
                ▷ 0 = null hypothesis
    end for

    pval0 $\leftarrow$ sort_lines(pval0)
            ▷ Sort each vector of randomized p-values
return pval0
end function

Figure 2 illustrates the conservativeness of the parametric Simes template on real data and the benefit yielded by calibration using randomized p-values curves. Choosing $\lambda > \alpha$ in (4) leads to a less conservative bound. Note that, the more depen-

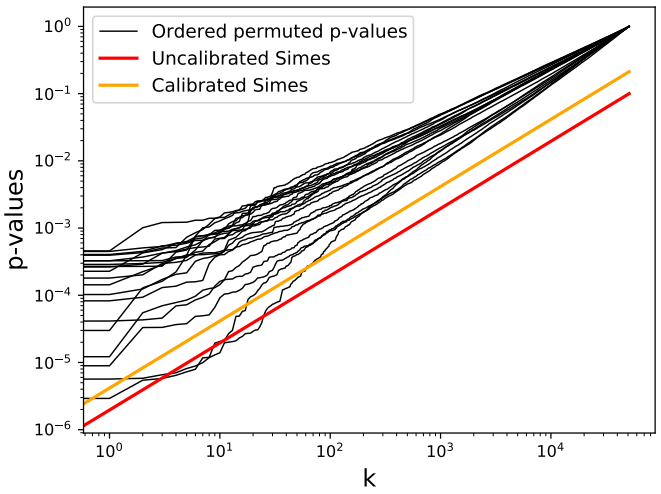

Figure 2: **Conservativeness of the Simes inequality and calibration**, illustrated on a set of 20 randomized p-values curves computed on real data and two JER controlling families at level 10%. Notice that both threshold families cross 2 curves (= 10% of all curves) which indeed corresponds to controlling the JER at level 10%. The uncalibrated Simes family (in red) is conservative since it is possible to choose larger threshold families that cross the same number of black curves. The calibrated Simes family is the least possible conservative threshold family that crosses at most 2 curves.

dent the data, the more we expect the original Simes bound to be conservative, see e.g. [6]. Thus, calibration should be particularly useful for smooth data.

While the ARI procedure corresponds to using Simes inequality without calibration[1] for JER control, calibration using the Simes template can be considered the state-of-the-art method for this problem [5, 6]. The bound obtained from this calibration procedure is equivalent to the bound considered in [1].

The second degree of freedom that can be exploited to achieve better statistical power while still controlling the JER is **to change the shape of the template**, instead of only optimising λ for a given template shape. In the next section, we introduce a data-driven approach to define a candidate template.

3. Main contribution: data-driven templates

Using the above-described calibration procedure to select a threshold family based on the inference data typically yields a substantial power gain (see [5, 1, 9]) even if the template shape is still linear as in the parametric ARI method. Yet, we notice in Figure 2 that for small k, permuted p-value curves are not exactly linear. This suggests that using a non-linear template shape could be relevant for fMRI data. Several other parametric templates are considered in [1], but the authors report that none of these attempts outperformed the Simes template. An ideal template should approximately reproduce the shape of randomized p-values curves computed from real data. Therefore, we propose to learn a template directly from the data.

A related idea has been explored in [19]. However, since the same data set was used for both the learning step and the calibration step, the method proposed in that paper suffers from circularity biases, as noted by [6, Remark 5.3]. Indeed, in the JER framework, the template has to be fixed a priori.
In order to address this issue, we propose to *learn* a template from an fMRI contrast that is independent from the contrasts on which inference is performed. We thus assume that training data are available to us to learn the template. Such data can easily be obtained from public data repositories.

First, we compute B randomized p-value curves using Algorithm 1 and extract quantile curves $t^b = (t^b_k)_k$ for $b = 1 \dots B$, as shown in the left panel of Fig. 3. These quantile curves are then viewed as a set of B sorted threshold families (middle panel), which is called a **learned template**. Note that it is indeed a template in the sense of [6], that has been discretized over a set of B values.

After obtaining a learned template, calibration is performed on the inference data (i.e. any inference contrast) as would be done with a parametric template. This is shown in the right panel of Figure 3 and in Section 2. In other words, we select the largest $b \in \{1, \dots B\}$ such that JER control holds on inference data for the threshold family t^b. In practice, this is done by dichotomy.

The complete procedure is summarized in Algorithm 2, with lines 1-7 corresponding to the training step and lines 8-20 corresponding to the inference step. The latter step requires the computation of the empirical JER for a given family, which is described in Algorithm 3.

[1] Rigorously, the ARI bound corresponds to using Simes inequality with the Hommel value h instead of m

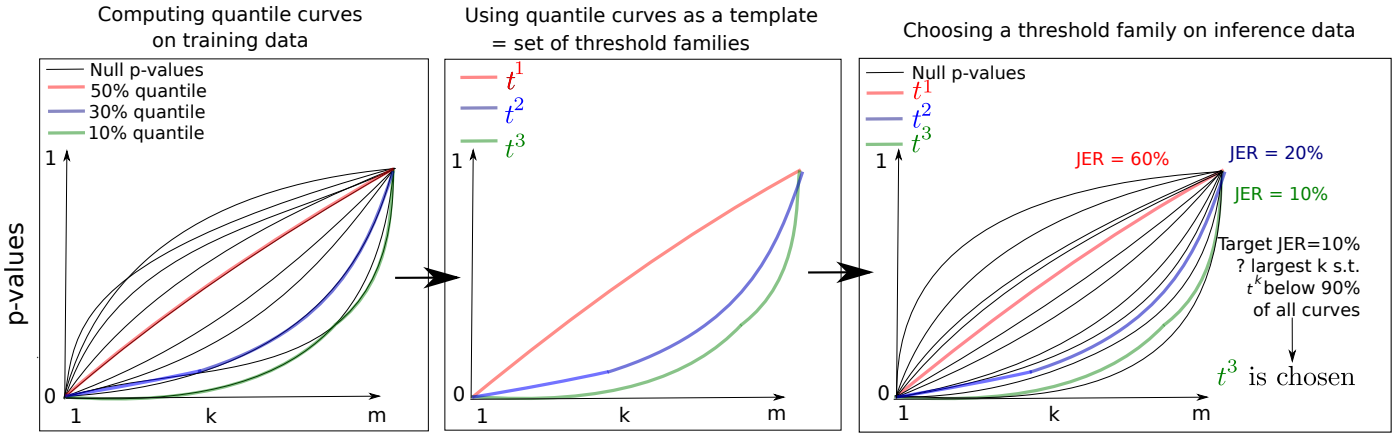

Figure 3: **Learning a template from training data and using this template for calibration on inference data.** Left panel: quantiles of randomized p-value curves are computed on training data. Middle panel: the resulting quantile curves are used as a template (the so-called learned template). Right panel: calibration is performed on inference data using the learned template. Notice that learned templates have varying shapes, contrary to parametric families such as Simes template.

Algorithm 2 Learning template on training data and calibrating on inference data. A template is learnt by computing permuted p-values and extracting quantile curves. Then, this template is used to perform calibration on testing data by choosing the least conservative family of the learned template that empirically controls the JER.

Require: $X_{train}, X_{test}, B_{train}, B_{test}, \alpha, k_{max}$

$pvals_{train} \leftarrow$ get_randomized_p_values(X_{train}, B_{train})
▷ vector of shape (B_{train}, n_{voxels})
for $b \in [1, B_{train}]$ **do**:
    learned_templates[b] $\leftarrow$ quantiles($pvals_{train}$, $\frac{b}{B_{train}}$)
end for
learned_templates $\leftarrow$ learned_templates[:, :k_{max}]
▷ retain first k_{max} columns
$pvals_{test} \leftarrow$ get_randomized_p_values(X_{test}, B_{test})
▷ vector of shape (B_{test}, n_{voxels})
for $b \in [1, B_{train}]$ **do** :
    $\widehat{JER}_b \leftarrow estimate\_jer(pvals_{test}$, learned_templates[$b$])
end for
$b_{calibrated} \leftarrow$ card$\{b \in [1, B_{train}]$ s.t. $\widehat{JER}_b \leq \alpha\}$
▷ Choose largest b such that JER control holds
if $b_{calibrated} = 0$ **then**
    return Calibrated_Simes
▷ No suitable learned template found
end if
chosen_template $\leftarrow$ learned_templates[$b_{calibrated}$]
return chosen_template

Algorithm 3 JER estimation on randomized p-values. The empirical JER is computed for a given template and a matrix of permuted p-values. This computation is directly based on equation 5.

function ESTIMATE_JER(pvals, thr, k_{max})
    (B_{test}, p) $\leftarrow$ shape(pvals)
    $\widehat{JER} \leftarrow 0$
    for $b' \in [1, B_{test}]$ **do**:
        for $i \in [1, ..., k_{max}]$ **do**:
            diff[i] $\leftarrow$ pvals[b'][i] - thr[i]
                ▷ Check JER control at rank i
        end for
        if min(diff) < 0 **then**:
            $\widehat{JER} \leftarrow \widehat{JER} + 1/B_{test}$
            ▷ Increment risk if JER control event is violated
        end if
    end for
return $\widehat{JER}$
end function

Once Algorithm 2 has been run, according to [6, 5], Equation 6 yields is a valid FDP upper bound. This bound can be computed on any subset of interest S in linear time in $|S|$ using Algorithm 1 in [9]. The complete procedure leading to this bound is called **Notip** for Non-parametric True Discovery Proportion estimation.

4. Experiments

4.1. Data

To investigate the potential power gain yielded by using data-driven templates, we performed experiments on an fMRI dataset, collection 1952 [28] of the Neurovault database (`http://neurovault.org/collections/1952`). This dataset is an

aggregation of 20 different fMRI studies, consisting of statistical maps obtained at the individual level for a large set of contrasts. These images have been preprocessed using the procedure described in [28]. In particular, they have been spatially normalized to MNI space using SPM12 software, and resampled to 3mm isotropic resolution. In the present case, the inference question concerns one-sample tests, i.e. identifying what brain regions show a significant increase of activity for the contrast of interest, as opposed to the baseline, across participants. The group-level statistic and associated p-value are obtained through a one-sample t-test on the individual z-maps.

Since collection 1952 only contains elementary *'versus baseline'* contrasts, we had to find relevant pairs of contrasts to obtain meaningful inference examples. Such 'versus baseline' contrasts contain a massive amount of non-specific signal, hence we pair them with control contrasts. A typical interesting contrast pair is "words vs baseline" vs "face vs baseline"; by subtracting these two contrasts, we obtain the more relevant "words vs face" contrast, which aims at uncovering brain regions with higher signal for word images than for face images stimuli.

To obtain consistent results, we excluded contrasts with too few subjects and/or trivial signal. The full list of 36 contrast pairs is given in Table 3.

In order to use data-driven templates on fMRI data, we have to choose a training set beforehand, on which we learn a template once and for all.

Although choosing a different template for each contrast pair would produce statistically valid inference, the computational cost would be high and this would lead to a loss in generality (i.e. the user would have to learn a template per inference contrast pair, instead of doing it once). For these experiments, we chose a training pair of contrasts to learn the data-driven template with 113 subjects and 51199 voxels smoothed using $FWHM = 4mm$ and 2% of active voxels (as estimated using ARI). This is the pair of contrasts with the lowest proportion of active voxels we could find among contrast pairs with at least 100 subjects. This choice is explained in the Discussion. We learn this template using $B = 10,000$ permutations and $k_{max} = 1,000 \simeq \lfloor m/50 \rfloor$ for reasons detailed in Section 8.1. Note that we also apply the same choice of k_{max} when using the Simes template, so that both templates are compared on a fair basis.

Data manipulation is mostly performed through Nilearn v0.9.0, nibabel v.3.1.1. The proposed statistical methods are implemented in the sanssouci package `https://github.com/pneuvial/sanssouci.python`. The experiments presented in this section can be reproduced using the code at the following address: `https://github.com/alexblnn/Notip`. This repository contains a script per experiment.

The analysis work we performed on this data can be divided into 4 main experiments that are detailed in the rest of this section.

4.2. Detection rate variation for different template types

To compare different choices of templates and investigate whether data-driven templates yield a detection rate gain over existing methods, we compute the size of the largest possible region that satisfies a target error control for each choice of template on the 36 chosen contrast pairs. This is typically the type of inference that users perform with FDR controlling procedure such as the Benjamini-Hochberg procedure. Formally, we solve the following optimisation problem for any template t:

$$|S_t| = \max_S |S| \quad \text{s.t.} \quad \frac{V_\alpha^t(S)}{|S|} \leq q\,, \tag{8}$$

where $q \in [0,1]$ is the FDP budget, and $V_\alpha^t(S)/|S|$ the upper bound on the FDP at risk level α computed on S using template t. Note that $|S_t|$ can be obtained in linear time in m using Algorithm 1 in [9].

Then, we compute the relative size difference of S_t for all possible pairs of templates. Formally, the **detection rate variation** between the learned template (i.e., the Notip procedure) and the calibrated Simes template is defined as:

$$\frac{|S_{Learned}| - |S_{Simes}|}{|S_{Simes}|}$$

[6, 5] show that the calibration procedure on any a priori fixed template indeed controls the JER.

Therefore, it makes sense to compare the detection rate associated with different template choices (i.e. ARI, calibrated Simes and learned template) by comparing the number of detections for a given error control. We compare the number of detections for several values of q, the target FDP budget, for a given risk $\alpha = 5\%$.

4.3. Comparison with FDR control

The above experiment on detection rate variation leads to a natural comparison with the regions obtained using the BH procedure that controls the FDR (= expected FDP). More precisely, we compare the size of the BH region and the size of FDP controlling regions. Conversely, we also estimate the FDP on the BH region to evaluate how accurately the FDP is controlled using BH procedure.

4.4. Detection rate variation for low sample sizes

Because of the high cost of acquisition, many fMRI datasets comprise few subjects. This may lead to unstable behavior and limited statistical power. To study the impact of sample size on the inference procedure both at training and inference step, we perform two experiments. First, we compute the detection rate for the three possible templates as in the first experiment, with the difference that this time we learn the template using $n_{train} = 10$ subjects instead of $n_{train} = 113$. Second, we use the standard template with 113 subjects but this time infer on 25 pairs of fMRI contrasts with any number of subjects n_{test}, varying from $n_{test} = 8$ to $n_{test} = 130$.

4.5. Influence of data smoothness

As in numerous statistical learning problems, the statistical properties of training and testing data ought to be well matched for the method to perform as expected. To assess the consequences on performance of a potential mismatch between training and inference data, we consider the case where smoothing parameter FWHM (full width at half maximum) is different in the training and inference data, using FHWM = 4mm for the training data and FWHM = 8mm for the inference data.

5. Results

5.1. Detection rate variation for different template types

A comparison the detection rate obtained for the three possible methods at hand, i.e. ARI, calibrated Simes and the learned template is displayed in Figure 4. To obtain this figure, we used 36 pairs of fMRI contrasts. The number n_{test} of subjects in each inference contrast pair ranged from 25 to 120.

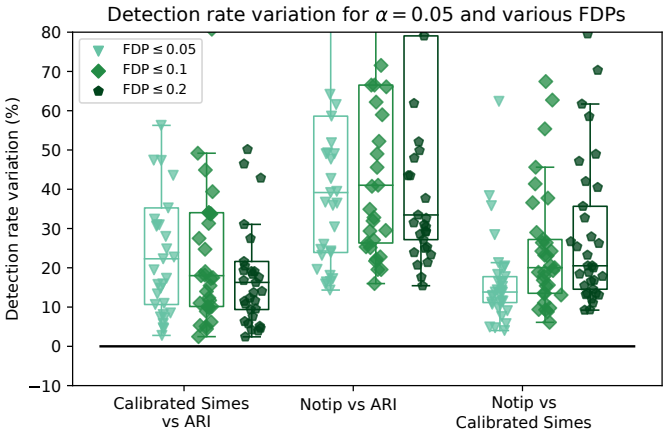

Figure 4: **Detection rate comparison between ARI, calibrated Simes and learned templates across 36 pairs of fMRI contrasts from Neurovault collection 1952.** After learning the template on a single contrast pair (see section 4), we perform inference on all 36 pairs. For each contrast pair, we compute the largest possible region that satisfies FDP control at level $q \in \{0.05, 0.1, 0.2\}$ with risk level $\alpha = 0.05$.

In Figure 4, we notice that learned templates yield a substantial gain in detection rate compared to both other template choices for all requested controls. On average, learned templates offer a $\sim$ **40% increase** in detection rate compared to the ARI method and a $\sim$ **20% increase** compared to calibrated Simes. A concrete example of inference on fMRI data is shown in Figure 5.

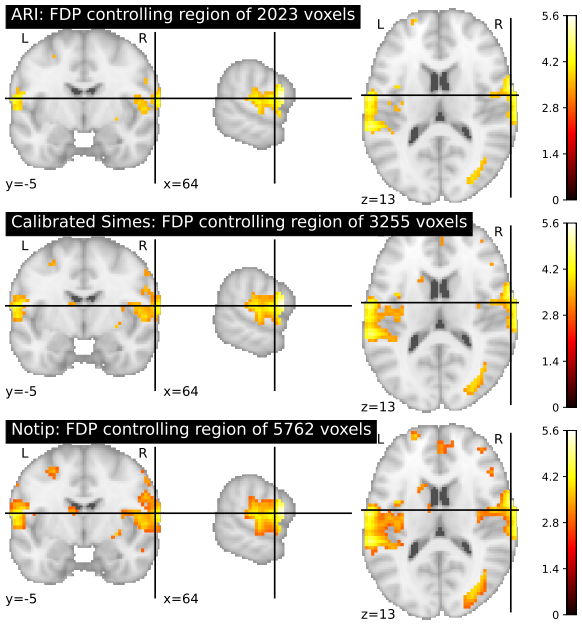

Figure 5: **Detection rate comparison between ARI, calibrated Simes and learned template on fMRI data.** For a pair of fMRI contrasts "look negative cue" vs "look negative rating" we compute the largest possible region that controls the FDP at level $q = 0.1$ with risk level $\alpha = 0.05$ for the three possible templates: ARI, calibrated Simes template and learned template. Notice that the detection rate is markedly higher (+ 77 %) using the learned template compared to the calibrated Simes template.

5.2. *Comparison with FDR control*

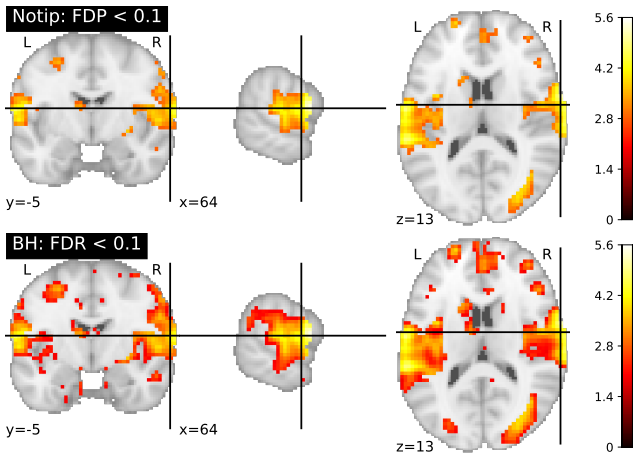

Figure 6: **Detection rate comparison between learned template and the BH procedure on fMRI data.** For a pair of fMRI contrasts "look negative cue" vs "look negative rating" we compute the largest possible region that controls the FDP at level $q = 0.1$ with risk level $\alpha = 0.05$ for the learned template and the largest possible FDR controlling region at level $q = 0.1$ using the BH procedure. BH region size: 13814 voxels. Learned template region size: 5762 voxels.

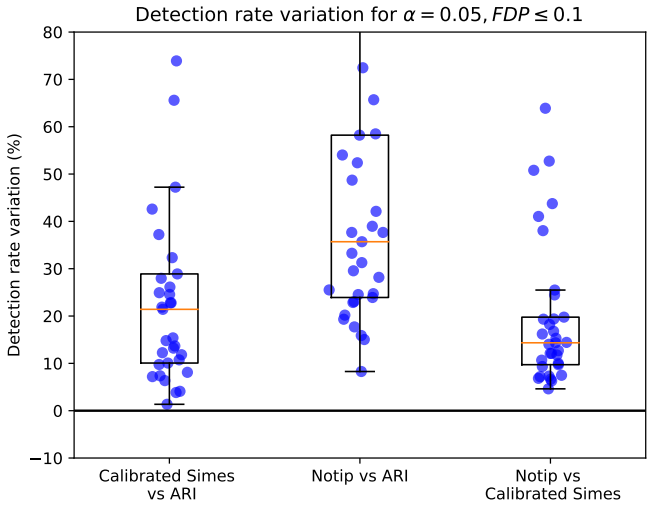

Figure 7: **Power comparison between ARI, calibrated Simes and a learned template using a subsampled training set.** Here, the template is learned using $n_{train} = 10$ subjects instead of $n_{train} = 113$ subjects. Learned templates still perform better than the calibrated Simes template on average, but subsampling the training set leads to a sub-optimal detection rate, compared with Figure 4.

Since FDR control is a much weaker guarantee than FDP control, it is expected that the BH procedure yields a substantially higher detection rate compared to FDP controlling procedures, as seen in Figure 6. However, FDP being the true quantity of interest, it is interesting to estimate the FDP on the FDR controlling region yielded by BH. Table 1 shows the estimated FDP on the FDR controlling region estimated with both the calibrated Simes template and the learned template.

	ARI	Calibrated Simes	Learned template
Estimated FDP	61%	45%	25%

Table 1: **Estimated FDP on the FDR controlling region obtained using the BH procedure (at level $q = 10\%$).** Notice that the learned template method yields more detections than the calibrated Simes template, but the estimated FDP remains above the FDR guarantee (10%). In other words, in this region the FDR is controlled but likely not the FDP at level $\alpha = 0.05$ (if it were the case, we would have an estimated FDP below 10%).

5.3. *Detection rate variation for low sample sizes*

The above results demonstrate that data-driven templates yield consistent power gains over existing methods that offer the same guarantees. In this section we investigate whether these gains subsist in sub-optimal conditions. Namely, when the template is learned on very few subjects or if inference is done on experiments with few subjects. The first point is illustrated in Figure 7.

Unstable performance may occur when inferring on data with few subjects, even if the template is learned on a large number of subjects ($n_{train} = 113$ here). This is illustrated in Figure 8: detection rate gains remain consistent across datasets with different number of subjects. However, for a single dataset comprising 17 subjects, the learned template performs substantially worse than calibrated Simes.

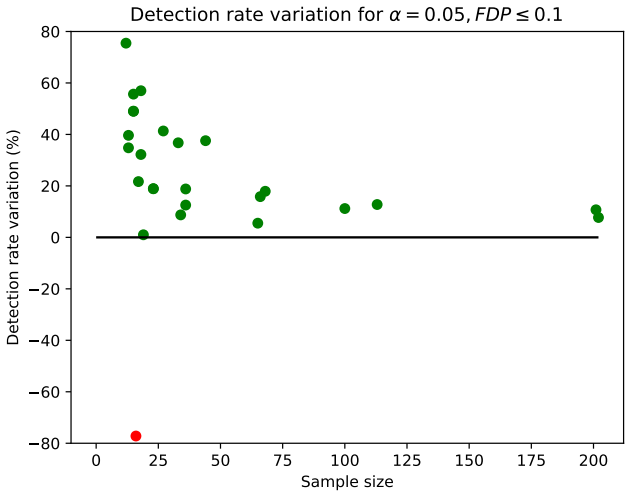

Figure 8: **Power comparison between learned template and calibrated Simes for many contrast pairs with a different numbers of subjects.** The detection rate gains remain consistent across datasets with different number of subjects. However, for a single dataset comprising 17 subjects, the learned template performs substantially worse than calibrated Simes.

5.4. *Influence of data smoothness*

Figure 9 shows that the smoothing parameter of the training data and the inference data have to be matched. Otherwise performance gains relative to the calibrated Simes method are reduced, albeit still positive.

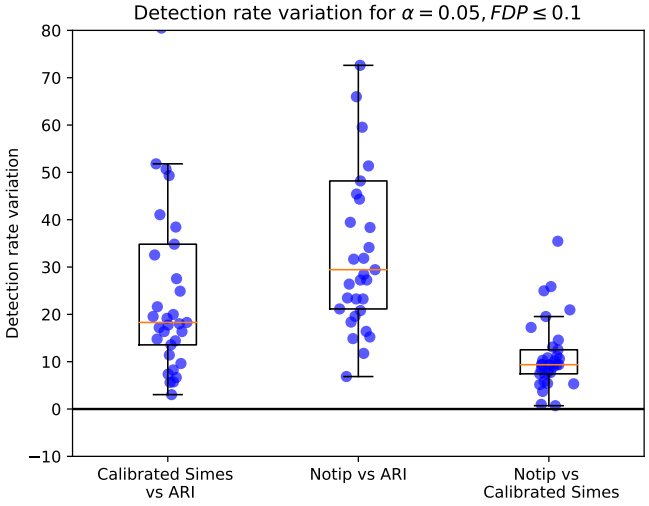

Figure 9: **An example of mismatch between the smoothing factors of training and inference data.** After learning the template on a single contrast pair (see Section 4) with smoothing full width at half maximum (FWHM) 4mm, we perform inference on all 36 pairs smoothed with FWHM 8mm. For each contrast pair, we compute the largest possible region that satisfies FDP control at level 0.1 with risk level $\alpha = 0.05$. The learned template still performs marginally better than calibrated Simes in this case, but gains are substantially lower in this regime.

6. Discussion

In this paper, we have proposed the Notip procedure, that allows users to estimate the proportion of truly activated voxels in any given cluster. There are at least two ways to perform inference on fMRI data using this procedure. First, one can threshold a statistical map to obtain the largest possible region that satisfies a requested FDP control. Second, users can also estimate the FDP on any cluster, as is usually done in the literature (see an example in Section 8.2).

This type of analysis is meant to mitigate the arbitrariness of cluster-forming thresholds in cluster-level inference, which remains a popular framework. The crucial observations is that estimates computed on these clusters may be plagued by circularity biases.

We have introduced a data-driven approach to obtain valid post hoc FDP control, thus achieving this goal. Moreover, controlling the FDP is a substantially more precise guarantee than controlling the FDR, its expected value. While FDP control comes at an unavoidable power cost compared to FDR control, we show that our procedure yields a higher detection rate than existing methods that offer the same statistical guarantees, namely ARI and calibrated Simes. We could go further by applying a step-down procedure as described in [6], but the gains are expected to be marginal [9].

However, this gain in detection rate is not systematic. First, it depends on the choice of the training set for learning the data-driven template. Interestingly, we found that certain learned templates outperformed the others in terms of detection rate. These templates correspond to the training contrast pairs that contain a large number of subjects and low signal. This is coherent with intuition since a large number of subjects and minimal signal allow a more stable and accurate estimation of the distribution of p-values under the null. Therefore, when selecting a template, it is useful to rely on a large-sample dataset with small signal magnitude.

One should also be careful when using data-driven templates on small datasets, as their performance is sub-optimal in this setting. In general, users should thus pay attention to the matching of training and testing data. For instance, if the smoothing parameter is poorly matched between the training and testing data, the detection rate gain obtained by using the learned template is reduced. It still remains non-negligible (9% compared to calibrated Simes).

Overall, even in deteriorated inference settings, the learned template offers substantial gains; this attests of the robustness of the Notip method.

This method also comes with an additional computational cost compared to classical calibration using the Simes template, since we have to learn the template before inference. However this additional cost is acceptable in practice since learning a template on a contrast and inferring on a contrast have the same time complexity.

We use $10,000$ permutations for better resolution when learning the template instead of the typical $1,000$ permutations used at the inference step. Learning a template using $B_{train} = 10,000$ permutations with a standard laptop (on a single thread) takes around 7 minutes, while inferring on a contrast pair (using $B_{test} = 1,000$ takes around 45 seconds). This can be trivially parallelized, as it is natively in the implementation we propose.

Another limitation of the proposed method is that it only handles one-sample or two-sample designs. This method could be extended to multivariate linear models in future work.

The idea of learning templates is not specific to fMRI data and could also be used on other types of data on which the calibration procedure is useful such as genomics [9].

We have achieved the goal of obtaining valid post hoc FDP control - rather than FDR control, or even weaker guarantees on clusters - while maintaining satisfactory power. This allows users in the brain imaging community to use more reliable inference methods that provide robust guarantees, avoiding circularity biases. The efforts to build such methods appear to us as important goal for the brain imaging field. The Python code used in this paper is available at `https://github.com/alexblnn/Notip`. This code relies on the sanssouci package available at `https://github.com/pneuvial/sanssouci.python`.

7. Acknowledgments

This project was funded by a UDOPIA PhD grant from Université Paris-Saclay and also supported by the FastBig ANR project (ANR-17-CE23-0011), the KARAIB AI chair (ANR-20-CHIA-0025-01) and the SansSouci ANR project (ANR-16-CE40-0019). The authors thank Laurent Risser and Nicolas Enjalbert-Courrech for their precious help on writing and improving the sanssouci Python code, and Samuel Davenport for useful discussions about this work.

References

[1] Angela Andreella, Jesse Hemerik, Wouter Weeda, Livio Finos, and Jelle Goeman. Permutation-based true discovery proportions for fmri cluster analysis. *arXiv preprint arXiv:2012.00368*, 2020.

[2] Sylvain Arlot, Gilles Blanchard, and Etienne Roquain. Some nonasymptotic results on resampling in high dimension, i: Confidence regions, ii: Multiple tests. *arXiv preprint arXiv:0712.0775*, 2007.

[3] 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.

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

[5] Gilles Blanchard, Pierre Neuvial, and Etienne Roquain. On agnostic post hoc approaches to false positive control. In Xinping Cui, Thorsten Dickhaus, Ying Ding, and Jason C. Hsu, editors, *Handbook of Multiple Comparisons*, Chapman & Hall/CRC Handbooks of Modern Statistical Methods. 1st edition edition, November 2021.

[6] 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.

[7] Katherine S. Button, John P. A. Ioannidis, Claire Mokrysz, Brian A. Nosek, Jonathan Flint, Emma S. J. Robinson, and Marcus R. Munafò. Power failure: why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, 14(5):365–376, May 2013.

[8] Anders Eklund, Thomas E Nichols, and Hans Knutsson. Cluster failure: Why fmri inferences for spatial extent have inflated false-positive rates. *Proceedings of the national academy of sciences*, 113(28):7900–7905, 2016.

[9] Nicolas Enjalbert-Courrech and Pierre Neuvial. Powerful and interpretable control of false discoveries in differential expression studies. bioRxiv preprint: https://doi.org/10.1101/2022.03.08.483449, 2022.

[10] O. Esteban, C. J. Markiewicz, R. W. Blair, C. A. Moodie, A. I. Isik, A. Erramuzpe, J. D. Kent, M. Goncalves, E. DuPre, M. Snyder, H. Oya, S. S. Ghosh, J. Wright, J. Durnez, R. A. Poldrack, and K. J. Gorgolewski. fMRIPrep: a robust preprocessing pipeline for functional MRI. *Nat Methods*, 16(1):111–116, 01 2019.

[11] Karl J Friston, CD Frith, PF Liddle, and RSJ Frackowiak. Comparing functional (pet) images: the assessment of significant change. *Journal of Cerebral Blood Flow & Metabolism*, 11(4):690–699, 1991.

[12] Christopher Genovese and Larry Wasserman. A stochastic process approach to false discovery control. *The annals of statistics*, 32(3):1035–1061, 2004.

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

[14] Jelle J Goeman and Aldo Solari. Multiple testing for exploratory research. *Statistical Science*, 26(4):584–597, 2011.

[15] Gerhard Hommel. Multiple test procedures for arbitrary dependence structures. *Metrika*, 33(1):321–336, 1986.

[16] 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.

[17] Nikolaus Kriegeskorte, W Kyle Simmons, Patrick SF Bellgowan, and Chris I Baker. Circular analysis in systems neuroscience: the dangers of double dipping. *Nature neuroscience*, 12(5):535, 2009.

[18] Ruth Marcus, Peritz Eric, and K Ruben Gabriel. On closed testing procedures with special reference to ordered analysis of variance. *Biometrika*, 63(3):655–660, 1976.

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

[20] Pierre Neuvial. Asymptotic properties of false discovery rate controlling procedures under independence. *Electronic journal of statistics*, 2:1065–1110, 2008.

[21] Pierre Neuvial. *Contributions to statistical inference from genomic data*. Habilitation thesis, Université Toulouse III Paul Sabatier, September 2020.

[22] Jean-Baptiste Poline and Bernard M Mazoyer. Analysis of individual positron emission tomography activation maps by detection of high signal-to-noise-ratio pixel clusters. *Journal of Cerebral Blood Flow & Metabolism*, 13(3):425–437, 1993.

[23] A. Roche, S. Mériaux, M. Keller, and B. Thirion. Mixed-effect statistics for group analysis in fMRI: a nonparametric maximum likelihood

approach. *Neuroimage*, 38(3):501–510, Nov 2007.

[24] 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.

[25] R John Simes. An improved bonferroni procedure for multiple tests of significance. *Biometrika*, 73(3):751–754, 1986.

[26] Stephen M Smith and Thomas E Nichols. Threshold-free cluster enhancement: addressing problems of smoothing, threshold dependence and localisation in cluster inference. *Neuroimage*, 44(1):83–98, 2009.

[27] Bertrand Thirion, Philippe Pinel, Sébastien Mériaux, Alexis Roche, Stanislas Dehaene, and Jean-Baptiste Poline. Analysis of a large fmri cohort: Statistical and methodological issues for group analyses. *NeuroImage*, 35(1):105–120, 2007.

[28] Gaël Varoquaux, Yannick Schwartz, Russell A Poldrack, Baptiste Gauthier, Danilo Bzdok, Jean-Baptiste Poline, and Bertrand Thirion. Atlases of cognition with large-scale human brain mapping. *PLoS computational biology*, 14(11):e1006565, 2018.

[29] A. M. Winkler, G. R. Ridgway, M. A. Webster, S. M. Smith, and T. E. Nichols. Permutation inference for the general linear model. *Neuroimage*, 92:381–397, May 2014.

[30] Choong-Wan Woo, Anjali Krishnan, and Tor D. Wager. Cluster-extent based thresholding in fmri analyses: pitfalls and recommendations. *NeuroImage*, 91:412–419, May 2014. 24412399[pmid].

8. Appendix

8.1. Choice of k_{max}

The post hoc bound (6) is valid for any value of the parameter k_{max}, provided that this parameter is chosen *a priori* and not after data analysis [6]. While some guidelines are given in the Discussion of [6], the choice of k_{max} remains an open question. Equation 6 may be written as follows:

$$V(S) = \min_{1 \leq k \leq |S| \wedge k_{max}} V_k(S), \tag{9}$$

where $V_k(S) = \sum_{i \in S} 1\{p_i(X) \geq t_k\} + k - 1$. Each $V_k(S)$ is itself an upper bound on the number of false positives in S. The choice of k_{max} implies a tradeoff. On the one hand, large values of k_{max} can seem advantageous because the minimum in (9) is taken on a larger set of values of k. On the other hand, when the thresholds t_k are obtained by calibration — as in [6] or in the present paper, a smaller k_{max} leads to larger values of (t_k) for a given k, and thus to a tighter bound V_k. Noting that $V_k(S) \geq k-1$, the values of k such that $k > q|S|$ will yield $V_k(S)/|S| \geq q$ for any S. Therefore, these values of k are useless for obtaining a FDP bound less than q. This motivates a choice of k_{max} of the form

$$k_{max} = q_{max}|S_{max}|, \tag{10}$$

where q_{max} is the maximum proportion of false positives that can be tolerated by users and $|S_{max}|$ is the size of the largest set of voxels of interest.

In practice, the regions of interest are those in which a **high proportion of activated voxels** can be guaranteed. To be conservative, we set $q_{max} = 0.5$, which simply means that we are not interested in guaranteeing that the FDP is less than q for $q \leq 0.5$. In the case of fMRI, one is generally interested in sparse activation extent, as widespread effect are by definition not informative on the specific involvement of brain regions in the contrast of interest. As a default choice, we observe that most fMRI contrasts studied in the literature lead to less of 5% of the image domain to be declared activate, which amounts to setting $|S_{max}| = 0.05m$.

Finally, a reasonable choice seems to be $k_{max} = 0.5 * 0.05m = 0.025m$. In the context of the experiments we described where $m \simeq 50,000$, we settle for simplicity on using $k_{max} = 0.02m = 1,000$.

To illustrate the effect of the choice of k_{max} we display detection rate variations of all three methods on 36 fMRI datasets across 9 different inference settings for varying k_{max} in Figure 10. Except for extremely small or large values of k_{max} the method is at worst slightly sub-optimal and $k_{max} = 1,000$ seems to be a reasonable choice.

As noted in [6], no choice of k_{max} uniformly outperforms others. For example, the above choice, which is motivated by the *prior*: "$|S_{max}| = 0.05m$", may be poorly adapted in situations where very large regions are considered.

8.2. TDP estimation on clusters

While we chose to compare this method's power in the standard inference setting of fMRI - i.e. find the largest possible region that satisfies a certain control - the method also yields valid inference on the TDP of data-driven clusters. This is illustrated in Table 2.

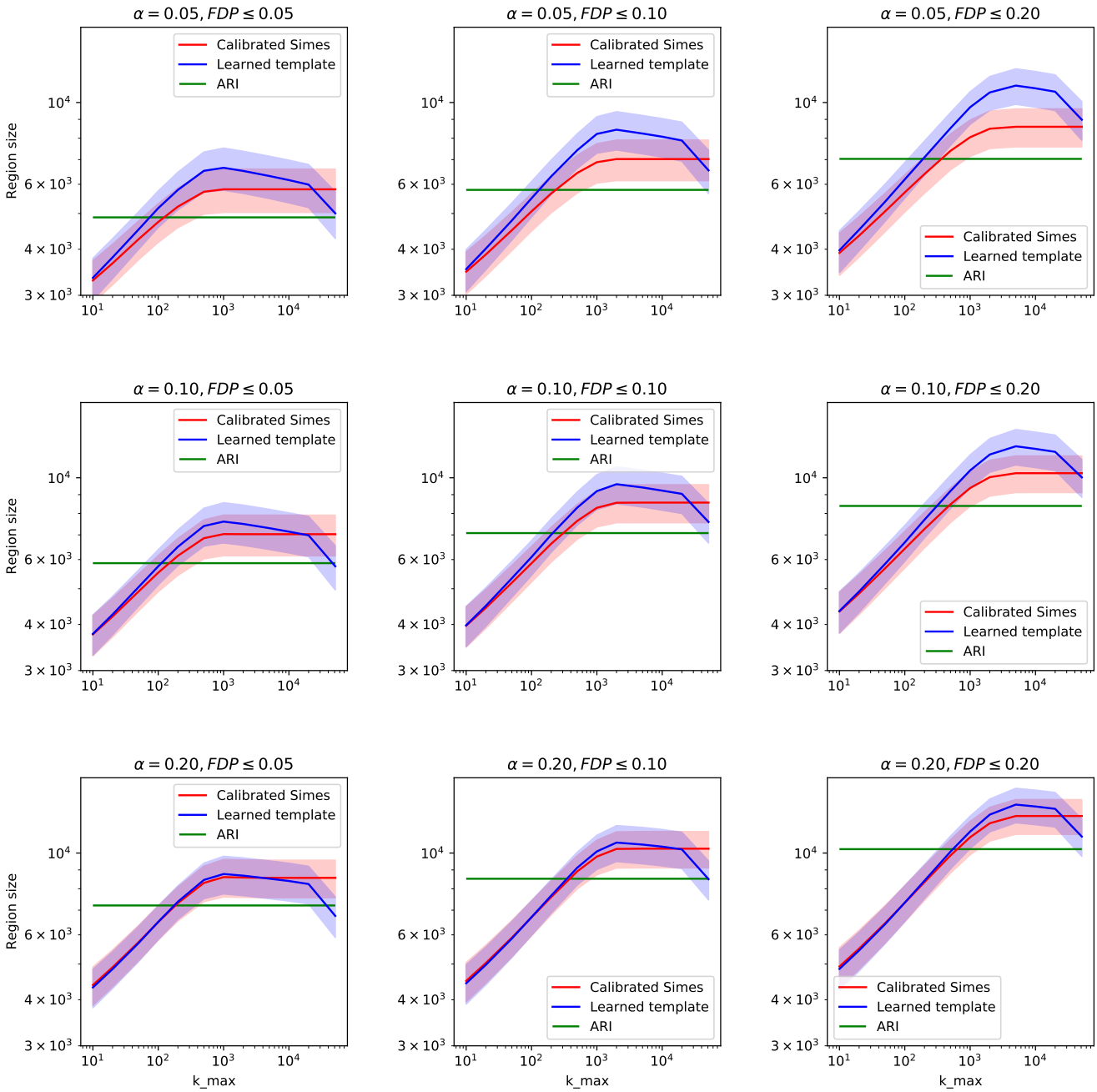

Figure 10: **Power comparison between learned template and calibrated Simes for various k_{max} values with 5% error bands in log-log scale.** Notice that the chosen k_{max} largely influences the maximum size of the FDP controlling region for the learned template.

						True Discovery Proportion		
Cluster ID	X	Y	Z	Peak Stat	Cluster Size (mm3)	ARI	Calibrated Simes	Learned
1	-33.0	-94.0	-17.0	5.63	7695	0.17	0.24	0.26
1a	-45.0	-79.0	-26.0	4.56				
1b	-48.0	-61.0	-26.0	4.13				
1c	-51.0	-64.0	-35.0	4.08				
2	66.0	2.0	16.0	5.47	14877	0.20	0.33	0.45
2a	69.0	-22.0	10.0	4.67				
2b	69.0	-10.0	13.0	4.59				
2c	69.0	-28.0	13.0	4.43				
3	-12.0	-82.0	-8.0	5.40	14445	0.27	0.38	0.50
3a	30.0	-73.0	-8.0	4.96				
3b	-24.0	-61.0	-11.0	4.91				
3c	30.0	-46.0	-11.0	4.64				
4	-6.0	11.0	52.0	5.30	5238	0.14	0.25	0.29
4a	6.0	8.0	55.0	4.19				
5	45.0	14.0	25.0	5.27	4563	0.24	0.30	0.30
5a	48.0	29.0	13.0	3.36				
6	12.0	-43.0	-26.0	5.08	12555	0.05	0.17	0.35
6a	0.0	-64.0	-14.0	4.43				
6b	3.0	-55.0	-11.0	4.26				
6c	3.0	-16.0	-32.0	4.23				
7	39.0	-73.0	4.0	5.00	6075	0.04	0.09	0.17
7a	39.0	-64.0	16.0	4.44				
7b	30.0	-82.0	10.0	4.42				
7c	27.0	-67.0	34.0	3.63				
8	-63.0	-34.0	16.0	4.95	25812	0.30	0.48	0.66
8a	-63.0	-10.0	13.0	4.90				
8b	-27.0	-19.0	4.0	4.85				
8c	-57.0	-19.0	7.0	4.68				
9	36.0	-94.0	-8.0	4.75	6507	0.08	0.15	0.17
9a	48.0	-70.0	-32.0	3.96				
9b	45.0	-70.0	-23.0	3.92				
9c	33.0	-82.0	-29.0	3.77				

Table 2: **Cluster localization (z > 3), size, peak statistic and estimated TDP** using the three possible templates (ARI, Calibrated Simes and Learned template) on contrast pair 'look negative cue vs look negative rating'. Cluster subpeaks are also reported when relevant. This table can be generated using script `https://github.com/alexblnn/Notip/blob/master/scripts/table_2.py`.

Study	Contrast 1	Contrast 2	$n_{subjects}$
HCP	shapes vs baseline	faces vs baseline	66
HCP	right hand vs baseline	right foot vs baseline	67
HCP	right foot vs baseline	left foot vs baseline	66
HCP	left hand vs baseline	right foot vs baseline	67
HCP	left hand vs baseline	left foot vs baseline	66
HCP	tool vs baseline	face vs baseline	68
HCP	face vs baseline	body vs baseline	68
HCP	tool vs baseline	body vs baseline	68
HCP	body vs baseline	place vs baseline	68
amalric2012mathematicians	equation vs baseline	number vs baseline	29
amalric2012mathematicians	house vs baseline	word vs baseline	37
amalric2012mathematicians	house vs baseline	body vs baseline	27
amalric2012mathematicians	equation vs baseline	word vs baseline	29
amalric2012mathematicians	visual calculation vs baseline	auditory sentences vs baseline	27
amalric2012mathematicians	auditory right motor vs baseline	visual calculation vs baseline	25
cauvet2009muslang	c16 music vs baseline	c02 music vs baseline	35
cauvet2009muslang	c16 language vs baseline	c01 language vs baseline	35
cauvet2009muslang	c02 language vs baseline	c16 language vs baseline	35
cauvet2009muslang	c04 language vs baseline	c16 language vs baseline	35
amalric2012mathematicians	face vs baseline	scramble vs baseline	85
ds107	scramble vs baseline	objects vs baseline	44
ds107	consonant vs baseline	scramble vs baseline	47
ds107	consonant vs baseline	objects vs baseline	44
ds108	reapp negative rating vs baseline	reapp negative cue vs baseline	32
ds108	look negative stim vs baseline	look negative rating vs baseline	34
ds108	reapp negative stim vs baseline	reapp negative rating vs baseline	34
ds109	false photo story vs baseline	false photo question vs baseline	36
ds109	false belief story vs baseline	false photo story vs baseline	36
ds109	false belief question vs baseline	false photo question vs baseline	36
ds109	false belief story vs baseline	false belief question vs baseline	36
ds109	false belief question vs baseline	false photo story vs baseline	36
pinel2007fast	visual right motor vs baseline	vertical checkerboard vs baseline	113
pinel2007fast	auditory right motor vs baseline	visual right motor vs baseline	121
ds107	scramble vs baseline	face vs baseline	85
amalric2012mathematicians	house vs baseline	scramble vs baseline	85
ds107	words vs baseline	face vs baseline	100

Table 3: **36 pairs of fMRI contrasts used for experiments.** These contrasts images have been downloaded from Neurovault 1952 collection.