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

Alexandre Blain, Bertrand Thirion, Pierre Neuvial

Notip: Non-parametric True Discovery Proportion control 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 [25] 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*, for Non-parametric True Discovery Proportion control: a powerful, non-parametric method that yields statistically valid guarantees on the proportion of activated voxels in data-derived clusters. Numerical experiments demonstrate substantial gains in number of detections compared with state-of-the-art methods on 36 fMRI datasets. The conditions under which the proposed method brings benefits are also discussed.

Keywords: fMRI, Brain mapping, False Discovery Proportion control, Selective Inference

1. Introduction

The mapping of the human brain consists of 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 [11], 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 [12], 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 [12] but it is conservative, resulting in false negatives, which hurts reproducibility (see e.g. [28, 8]). A more powerful and commonly used approach is to control the False Discovery Rate (FDR) [14], 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). The FDP is the proportion of false discoveries among all discoveries. As noted by several authors [13, 17, 21], 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 [23], 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 [27] exceeds its expected value under a null distribution. However, this approach suffers from several problems [9], such as the arbitrary choice of cluster-forming threshold [31], 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 [30, 9]. 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 [18].

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 [7]. To bypass this double-dipping issue, one can use *post hoc* estimates that control the FDP. Note that an upper bound on the FDP is equivalent to a lower bound on the TDP. The first method of that kind is a parametric method called All-resolutions inference (ARI) [25].

In this paper, we introduce the Notip procedure, that adapts non-parametrically to data correlation . The use of non-parametric procedures also renders the inference robust to mis-specification of the statistics distribution. We study whether such a procedure can yield less conservative inference 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 [7]. Our main contribution is the Notip method presented in Section 3: a nonparametric 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 gains in the number of detections are 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 kth 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: $\text{FDP}(S) = |S \cap H_0| / |S|$.

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 kth 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 [17]. 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 [22]). 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:

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

Since $\text{FDP}(S) = |S \cap H_0| / |S|$, 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 [15], 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 [19] combined with the following inequality:

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

Equation (2) is an immediate consequence of the Simes inequality [26], and also holds under the PRDS assumption. The All-resolutions inference (ARI) method [25] provides a tighter post hoc bound that uses the thresholds $\alpha k / h(\alpha)$ instead of $t_k^{Simes} = \alpha k / m$ in (2), where $h(\alpha) \leq m$ is the so-called Hommel value [16]. $h(\alpha)$ 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 [7]. Letting $R_k^{Simes} = \left\{i : p_i \leq t_k^{Simes}\right\}$, Equation (2) can be written as:

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

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:

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

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 [7]. Therefore, as illustrated in the bottom panel of Figure 1, the final post hoc FDP upper bound is $V^{Simes}(S)/|S|$, where

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

As noted by [7], the bound (4) coincides with the bound originally proposed by [15]. 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 $R_k = \{i : p_i \leq t_k\}$. While setting $k_{\max} = m$ is natural in (4), setting $k_{max} < m$ can be useful when t_k is calibrated from the data, as discussed in the next sections. The Joint Error Rate (JER) of the threshold family t is defined by [7] as:

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

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

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

provides a valid FDP upper bound for any threshold family t such that $\text{JER}(t) \leq \alpha$ [7]. This bound can be calculated in $O(|S|)$ for a given set S using Algorithm 1 in [10].
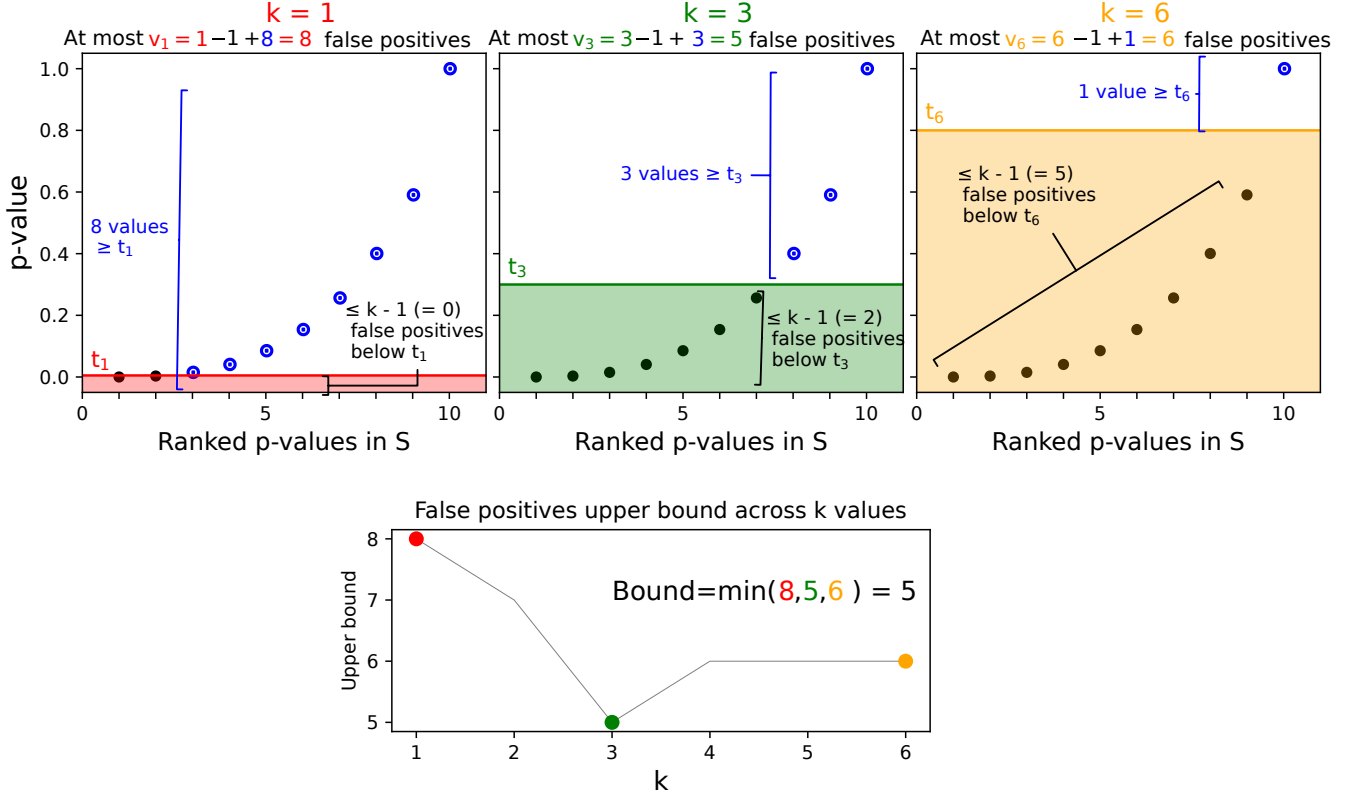
Figure 1: **Computation of the post hoc bound** (6) **on the number of false positives**, given a set S of 10 p-values, using a JER controlling threshold family. Top panels: computation of k-th bound $V_k(S) = \sum_{i \in S} \mathbb{1}\{p_i(X) \geq t_k\} + k - 1$ for 3 values of k, with horizontal colored lines representing the associated thresholds t_k. Bottom panel: The post hoc upper bound(6) corresponds to the minimum of all $V_k(S)$. In this example, the bound guarantees that the number of false positives in the S is at most 5 with probability $> 90\%$.

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 can be conservative for positively dependent p-values [7], 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 (that is, JER control) holds:

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

In order to reach this goal more generally, we consider collections of threshold families called **templates** since their introduction in [7]. Formally, a template is set of functions $\lambda \mapsto (t_k(\lambda))_k$ 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 $k = 1 \ldots m$ and $\lambda > 0$.

The **calibration** procedure introduced in [6, 7] 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 [24, 2]. Randomized p-values can be obtained similarly for two-sample tests, using class label permutations instead of sign-flipping.
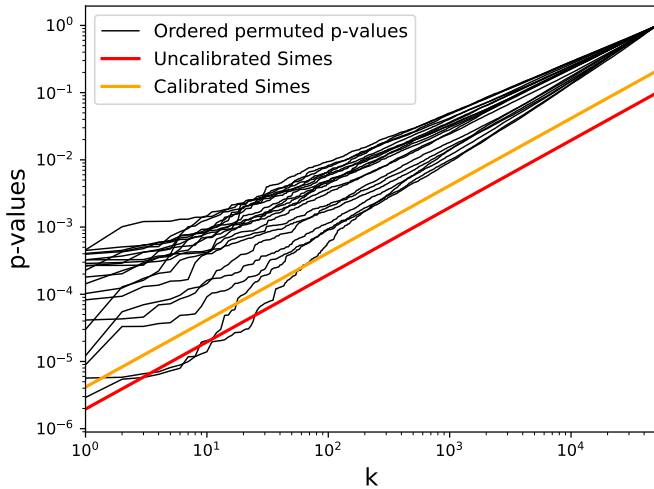
4

Figure 2: **Addressing the conservativeness of the Simes inequality by calibration**. A set of 20 randomized p-value curves are computed on real data (black curves). Two JER controlling families at level 10% are shown as colored lines. Both of them 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 higher threshold families that cross the same number of black curves. The calibrated Simes family (in orange) is the least possible conservative threshold family that crosses at most 2 curves.

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 $X_{flipped}$.

1: **function** GET_RANDOMIZED_P_VALUES(X, B)
2:     $n, p \leftarrow$ shape(X)
3:                             $\triangleright$ n subjects, p voxels
4:     pval0 $\leftarrow$ zeros(B, p)
5:     **for** $b \in [1, B]$ **do**
6:         flip $\leftarrow$ diag(draw_random_vector($\{-1, 1\}^n$))
7:                            $\triangleright$ matrix of shape (n, n)
8:         $X_{flipped} =$ flip.X
9:         pval0[b] $\leftarrow$ one_sample_t_test($X_{flipped}$, 0)
10:                        $\triangleright$ 0 = null hypothesis
11:     **end for**
12:
13:     pval0 $\leftarrow$ sort_lines(pval0)
14:         $\triangleright$ Sort each vector of randomized p-values
15: **return** pval0
16: **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 dependent the data, the more the parametric Simes bound is expected to be conservative, see e.g. [7]. 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 [6, 7]. The bound obtained from this calibration procedure is equivalent to the bound considered in [1].

3. Main contribution: data-driven templates and Notip procedure

(see [6, 1, 10])

The calibrated Simes family can lead to tighter post hoc bounds, yet it still relies on the Simes template, which is linear in k, as illustrated in Figure 2. Instead of only optimising λ for a given template shape (e.g. a linear shape for the Simes template), the second degree of freedom that can be exploited to achieve better statistical power while still controlling the JER is **to learn the template function, or, equivalently, its shape when displayed as a graph** Figure 2 illustrates 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 [20]. However, since the method proposed in that paper does not distinguish between the learning and calibration steps, it suffers from circularity biases, as noted by [7, 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 dataset that is independent from the datasets on which inference is performed. First, we compute B randomized p-value curves on training data using Algorithm 1 and extract quantile curves $t^b = (t_k^b)_k$ for $b = 1 \ldots B$, 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 [7], that has been discretized over a set of B values.

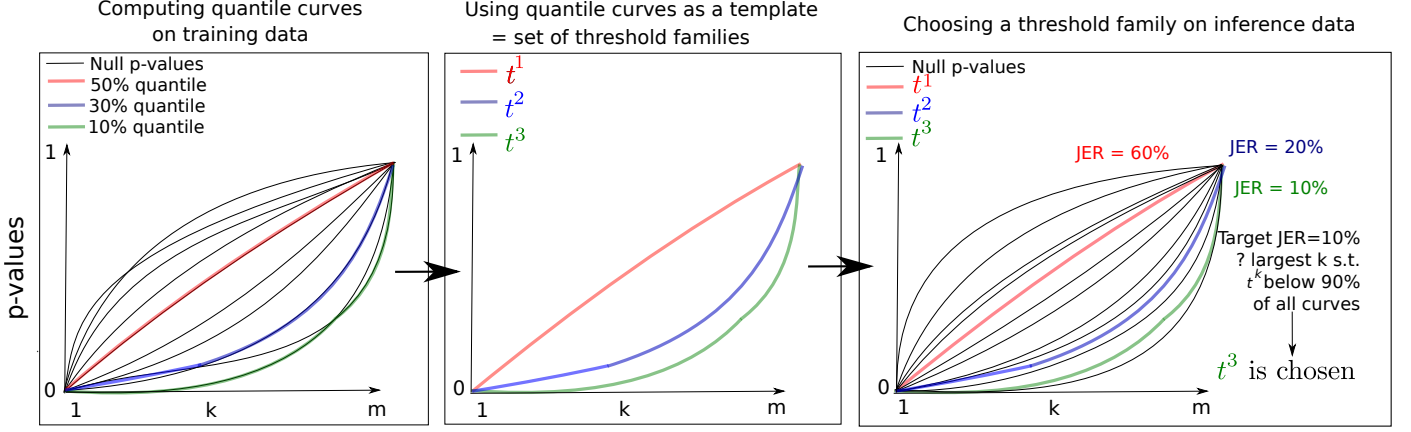[1]Rigorously, the ARI bound corresponds to using Simes inequality with the Hommel value h instead of m.

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 do not have a parametric shape (contrary e.g. to the Simes template), but follow the shape of sorted null *p*-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. To perform calibration, we evaluate the empirical JER of all threshold families of the learned template. Then, we select the largest $b \in \{1, \dots B\}$ such that JER control holds on inference data for the threshold family t^b. To avoid evaluating the JER of all threshold families, this search is done by dichotomy in practice. The resulting method is called "Notip" for Non-parametric true discovery proportion. As described above, Notip requires a training dataset in order to learn the template. An example of a template learned from a training data set is displayed in Figure 14. Note that learned templates do not have a parametric shape, but follow the shape of randomized *p*-value curves.

The calibration process depends on the parameter k_{max}, whose choice induces the following trade-off. On the one hand, since $JER((t_k^b)_{1 \leq k \leq K}) \leq JER((t_k^b)_{1 \leq k \leq K'})$ for all $K \leq K'$, choosing a smaller k_{max} allows calibration to choose a largest value of b in the dichotomy, leading to a less conservative family. On the other hand, a larger value for k_{max} leads to more thresholds considered in the min in the bound written in Equation 6, and hence to a possibly tighter bound. Guidelines to choose k_{max} as well as an informed default choice for fMRI data are given in Section 8.2.

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.

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 inference data by choosing the least conservative family of the learned template that empirically controls the JER.

Require: $X_{train}, X_{infer}, B_{train}, B_{infer}, \alpha, k_{max}$
1: $\text{pvals}_{train} \leftarrow \text{get\_randomized\_p\_values}(X_{train}, B_{train})$
2:                          ▷ array of shape (B_{train}, n_{voxels})
3:                          ▷ lines of pvals_{train} are sorted
4: **for** $b \in [1, B_{train}]$ **do**:
5:      $\text{learned\_templates}[b] \leftarrow \text{quantiles}(\text{pvals}_{train}, b/B_{train})$
6: **end for**
7: $\text{pvals}_{infer} \leftarrow \text{get\_randomized\_p\_values}(X_{infer}, B_{infer})$
8:                          ▷ vector of shape (B_{infer}, n_{voxels})
9: **for** $b \in [1, B_{train}]$ **do** :
10:     $\widehat{JER}_b \leftarrow \text{estimate\_jer}(\text{pvals}_{infer},$
                    $\text{learned\_templates}[b], k_{max})$
11: **end for**
12: $b_{calibrated} \leftarrow \max\{b \in [1, B_{train}] \text{ s.t. } \widehat{JER}_b \leq \alpha\}$
13:                 ▷ Choose largest b such that JER control holds
14: **if** $b_{calibrated} = 0$ **then**
15:     **return** Calibrated_Simes
16:                          ▷ No suitable learned template found
17: **end if**
18: $\text{chosen\_template} \leftarrow \text{learned\_templates}[b_{calibrated}]$
19: **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.

```
 1: function ESTIMATE_JER(pvals, thr, k_max)
 2:     (B_infer, p) ← shape(pvals)
 3:     JER ← 0
 4:     for b' ∈ [1, B_infer] do:
 5:         for i ∈ [1, ..., k_max] do:
 6:             diff[i] ← pvals[b'][i] - thr[i]
 7:                         ▷ Check JER control at rank i
 8:         end for
 9:         if min(diff) < 0 then:
10:             JER ← JER + 1/B_infer
11:             ▷ Increment risk if JER control event is violated
12:         end if
13:     end for
14: return JER
15: end function
```

Once Algorithm 2 has been run, according to [7, 6], the bound defined in Equation 6 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 [10].

4. Experiments

4.1. Data

4.1.1. FMRI data

To investigate the potential gain in number of detections yielded by using data-driven templates, we performed experiments on an fMRI dataset, collection 1952 [29] 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 [29]. 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 in group analyses, 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.

Collection 1952 only contains elementary *'versus baseline'* contrasts, with a massive amount of non-specific signal. In order to obtain meaningful inference examples, we paired 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 significantly higher or lower 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 resulting list of 36 contrast pairs is given in Table 6.

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. The variability of the Notip method with regards to the choice of the training set is studied in Section 4.3. For the rest of the experiments, we use a single training set. Although learning 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 choose for training data a pair of contrasts with 113 subjects and 51199 voxels smoothed using $FWHM$ (full width at half-maximum) $= 4mm$ and at least 2% of active voxels (with probability $\geq 95\%$ according to ARI). This is the pair of contrasts with the lowest proportion of active voxels that we could find among contrast pairs with at least 100 subjects. This choice is referred to as the optimal template in the rest of the paper. It is explained in the Discussion. This template is learnt using $B_{train} = 10,000$ permutations and we choose $k_{max} = 1,000 \simeq \lfloor m/50 \rfloor$ for reasons detailed in Section 8.2. 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.

4.1.2. Synthetic data

For some of the experiments described below, we have generated simulated data using the pyrft package: https://github.com/sjdavenport/pyrft. This package allows generates smooth noisy random fields that resemble fMRI data. In this controlled setup, the ground truth is known. An example of such simulated data can be found in Section 8.7. The simulation setting is the following, with π_0 the proportion of null voxels: $\alpha = 0.05$, $\pi_0 = 0.9$, $FWHM = 4mm$, $n_{train} = 100$, $n_{infer} = 50$, $q = 0.1$, $B_{train} = B_{infer} = 1000$.

4.1.3. Code to reproduce the experiments

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: `https://github.com/alexblnn/Notip`. This repository contains a script per experiment.

The analysis we performed on this data can be divided into 6 main experiments that are detailed in the rest of this section.

4.2. Variation of the number of detections for all three methods

To compare different choices of templates and investigate whether data-driven templates yield a gain in number of detections 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 aim for when applying FDR controlling procedures such as the Benjamini-Hochberg procedure. We denote by S_t the largest region (i.e. subset of voxels) such that its FDP upper bound is smaller than some user-defined value $q \in [0, 1]$, called the FDP budget. It corresponds to the maximum FDP that one is willing to tolerate in a given region. Formally, we solve the following optimisation problem for any template t:

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

where $V_\alpha^t(S)/|S|$ is the upper bound on the FDP at risk level α computed on S using the template t. By construction of the bound (6), the solution of (8) is a p-value level set, of the form $\{i/p_i \le \tau\}$ for some τ [7, Section 7.4]. As such, $|S_t|$ can be obtained in linear time in m using Algorithm 1 in [10].

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

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

The calibration procedure on any a priori fixed template

controls the JER [7, 6]. Therefore, it makes sense to compare the number of detections obtained by different template choices (i.e. ARI, calibrated Simes and learned template) for a given error control $1 - \alpha$. We compare the number of detections for several values of q, the FDP budget, for a given risk $\alpha = 5\%$.

We also perform the same experiment on the simulated data described in Section 4.1.2. In this case, since the ground truth is known, we can compare the empirical True Positive Rate (TPR) of all three methods. This quantity represents the proportion of true signal recovered by the template t for the region S_t defined in (8). Formally, we defined the TPR in S_t as the ratio of the lower bound on the true positives in S_t to the number of truly activated voxels in S_t:

$$\text{TPR}(S_t) = \frac{|S_t| - V_\alpha^t(S_t)}{|H_1|}.$$

Where $|H_1|$ corresponds to the number of truly activated voxels. As such, $\text{TPR}(S_t)$ is an empirical measure of power for the template t.

4.3. Comparison with FDR control

The above experiment on the number of detections leads to a natural comparison based on the "BH region", that is the region obtained using the BH procedure that controls the FDR (= expected FDP). More precisely, we compare the size of the BH region to the size of FDP controlling regions. Conversely, we also compute FDP upper bounds on the BH region. This illustrates the difference between FDR control and FDP control with a concrete example.

4.4. Variation of the number of detections 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 dual experiments. First, we compute the number of detections for the three possible methods as in Section 4.2, with the difference that the template is learned 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_{infer}, varying from $n_{infer} = 8$ to $n_{infer} = 200$.

4.5. Sensitivity to the choice of training data

Since Notip requires learning a template on training data before performing inference, the choice of such data and its impact on the performance of the method is an important question. To assess this sensitivity quantitatively, we fix an fMRI contrast pair for inference. Then, we compare the number of detections for each template choice -as described in Section 4.2- using the 36 different fMRI contrast pairs as 36 different training sets for the Notip method. It should be noted that ARI and calibrated Simes do not depend on the chosen training set; their number of detections is computed once and for all. These 36 fMRI contrast pairs differ in several ways such as the number of subjects, the nature of the contrasts, the fMRI study or quantity of signal. This allows us to evaluate the robustness of Notip to poorly matched training and inference data. In this experiment, we also include the optimal template choice we used for all other experiments (i.e. least amount of signal and maximum number of subjects).

4.6. Influence of data smoothness

Another potential source of mismatch is the smoothing done in preprocessing of fMRI data. To assess the consequences on performance of a potential smoothing mismatch between training and inference data, we consider the case where the smoothing parameter FWHM is different in the training and inference data, using FHWM = 4mm for the training data and FWHM = 8mm for the inference data.

4.7. Using Notip on a single dataset

When learning a template on separate data is inconvenient, or to avoid the computational cost of learning the template, a natural idea is to use Notip on a single dataset. In such a setting, circularity biases may appear as in [20]. The workaround that we propose to retain valid FDP control is to perform two independent rounds of randomization - one for training and one for inference. While this approach is formally not covered by the theoretical framework of [7], we have performed experiments to assess its FDP control and power on the simulated data described in Section 4.1.2.

5. Results

5.1. Variation of the number of detections for different template types

A comparison of the number of detections obtained for the three possible methods at hand, i.e. ARI, calibrated Simes

and Notip is displayed in Figure 4. To obtain this figure, we used 36 pairs of fMRI contrasts. The number n_{infer} of subjects ranged from 25 to 120 across inference contrast pairs.
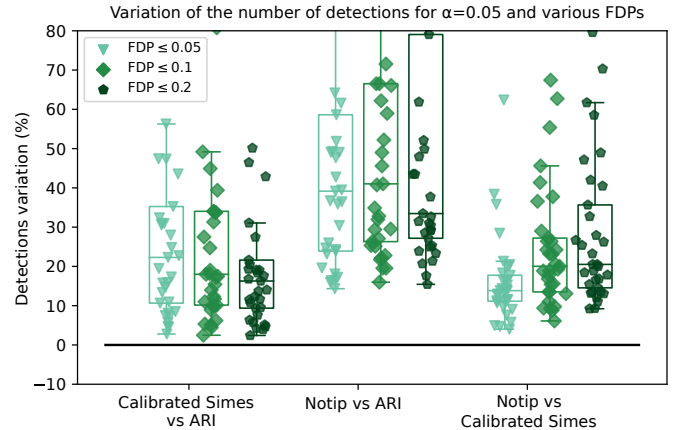

Figure 4: **Comparison of the number of detections between ARI, calibrated Simes and learned templates across 36 pairs of fMRI contrasts from Neurovault collection 1952.** After learning the template on a single contrast pair (see section 4), we perform inference on all 36 pairs. For each contrast pair,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

In Figure 4, we notice that learned templates yield a substantial gain in detections compared to both other template choices for all target FDPs. On average, learned templates offer a $\sim$ **40% increase** in detections compared to the ARI method and a $\sim$ **20% increase** compared to calibrated Simes. Gains in number of detections can vary largely across contrast pairs. This is essentially due to variance contained in the data, as all three methods exhibit similar TPR variability on simulated data (see Section 8.4). A concrete example of inference on fMRI data is shown in Figure 5.
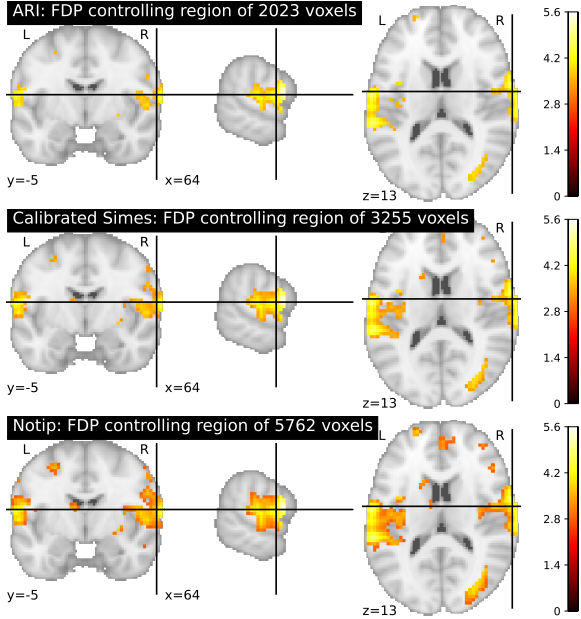
Figure 5: **Comparison of the number of detections between ARI, calibrated Simes and learned template on fMRI data.** For a pair of fMRI contrasts "look negative cue" vs "look negative rating" we compute the largest possible region such that FDP ≤ 0.1 with risk level $\alpha = 0.05$ for the three possible templates: ARI, calibrated Simes template and learned template. Notice that the number of detections is markedly higher (+ 77 %) using the learned template compared to the calibrated Simes template.

We have also performed the same experiment on simulated data. In this setting, we can report the actual TPR of the methods instead of region sizes. The empirical FDP for these simulations are reported in Figure 16.
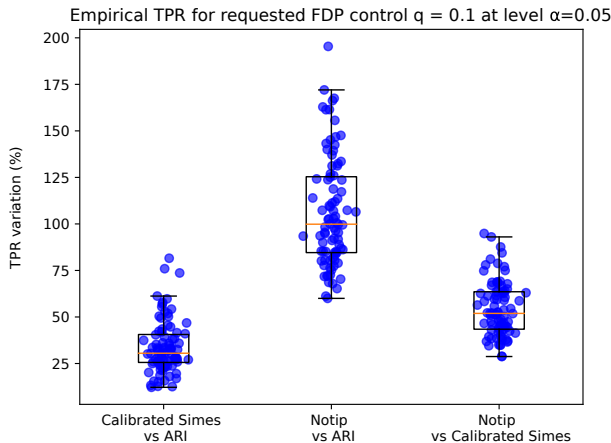

Figure 6: **TPR comparison for a FDP budget** $q = 0.1$ **at risk level** $\alpha = 0.05$. We run 100 simulations and report the TPR. Notice that Notip offers substantial gains in TPR compared to both ARI (100 % on average) and calibrated Simes (50 % on average).

Figure 6 illustrates the TPR gains achieved using Notip on simulated data compared to both ARI to both ARI (100 % on average) and calibrated Simes (50 % on average). Overall,

simulations support the fact that Notip offers substantial performance gains compared to both ARI and calibrated Simes.
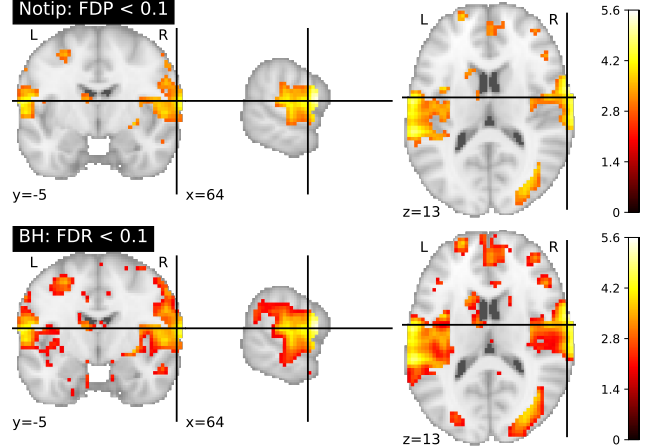
5.2. Comparison with FDR control

Figure 7: **Comparison of the number of detections 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 such that FDP ≤ 0.1 at risk level $\alpha = 0.05$ for the learned template and the largest possible region such that FDR ≤ 0.1 using the BH procedure. BH region size: 13814 voxels. Learned template region size: 5762 voxels.

Since FDR control is a much weaker guarantee than FDP control, it is expected that the BH procedure yields substantially more detections compared to FDP controlling procedures, as seen in Figure 7. However, FDP being the targeted guarantee, it is interesting to compute FDP upper bounds on the FDR controlling region yielded by BH. Concretely, we are trying to obtain a bound on the FDP of a region that only has a guarantee on its FDR. Table 1 shows the FDP upper bounds computed on the FDR controlling region using all three possible methods.

	ARI	Calibrated Simes	Notip
FDP Upper bound	61%	45%	25%

Table 1: **FDP upper bounds on the FDR controlling region obtained using the BH procedure (at level** $q = 10\%$**).** Notice that Notip yields smaller FDP bounds than ARI and calibrated Simes . This upper bound remains higher than the FDR guarantee (10%), which is more permissive by design.

Notip leads to a less conservative FDP upper bound than ARI and calibrated Simes. However, at risk level $\alpha = 5\%$, Notip is only able to guarantee that the FDP is less than 25% while the FDR is controlled at level 10%. This illustrates the difference between FDR control and FDP control, the latter being less permissive by design. While the BH procedure

guarantees that the **expected** FDP is below 10%, Notip guarantees explicitly that the **actual** FDP is below 25% with high probability ($\geq$ 95%). It should be noted that on a single inference run, a guarantee on the **expected** FDP has no clear interpretation, whereas the guarantee on the **actual** FDP is directly interpretable.

5.3. Variation of the number of detections for low sample sizes

The above results demonstrate that data-driven templates yield consistent gains in number of detections 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 8.
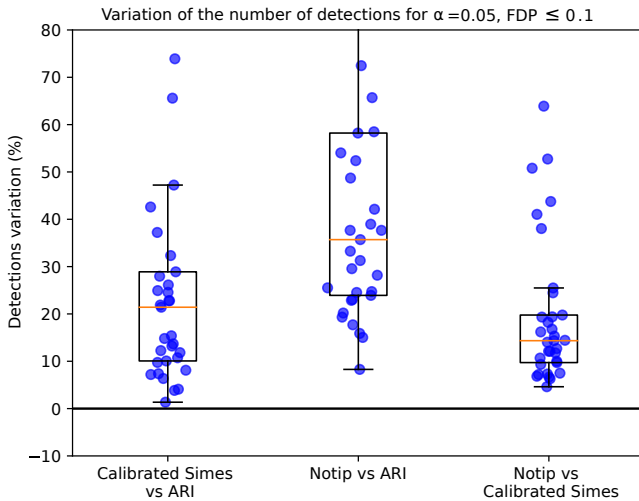

Figure 8: **Comparison of the number of detections between ARI, calibrated Simes and a learned template using a subsampled training set.** Here, the template is learned using $n_{train} = 10$ subjects instead of $n_{train} = 113$ subjects. Learned templates still perform better than the calibrated Simes template on average, but subsampling the training set leads to a sub-optimal number of detections , compared with Figure 4.

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 9: gains in number of detections remain consistent- yet more variable for smaller sample sizes - across datasets with different number of subjects. As noted in [8], high variance is unavoidable when inferring on small datasets (e.g. $n_{infer} \leq$ 25). For a single dataset comprising 17 subjects, the learned template performs substantially worse than calibrated Simes.
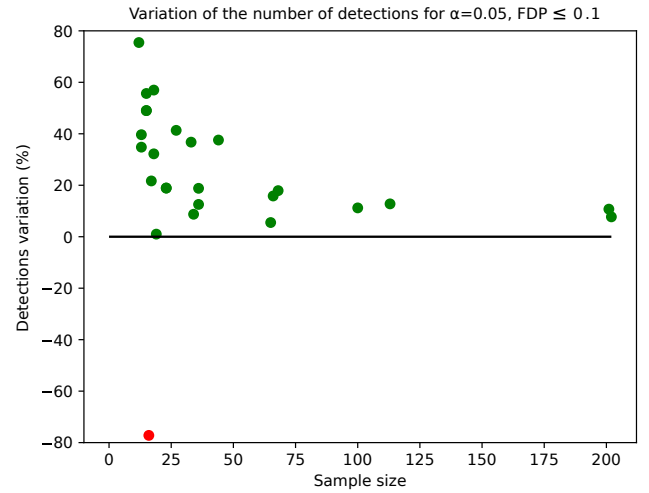

Figure 9: **Comparison of the number of detections between learned template and calibrated Simes for many contrast pairs with a different numbers of subjects.** The gains in number of detections remain consistent across datasets with different number of subjects. However, for a single dataset comprising 17 subjects, the learned template performs substantially worse than calibrated Simes.

5.4. Sensitivity to the choice of training data

Figure 10 displays the variation of the number of detections made by Notip compared to ARI and calibrated Simes using 36 different training sets. All training contrast pairs except one yield more detections than calibrated Simes, with gains ranging from 10% to 80%. This shows that the Notip procedure is robust to poorly matched training and inference data, since contrast pairs considered for training vary along many dimensions: number of subjects, nature of contrasts, fMRI study, quantity of signal... In the worst possible case, Notip performs marginally worse than calibrated Simes. Also note that the optimal template used in all other experiments (corresponding to the template learned from the training data with minimal signal and maximum number of subjects as described in Section 4.1.1) outperforms all other choices.
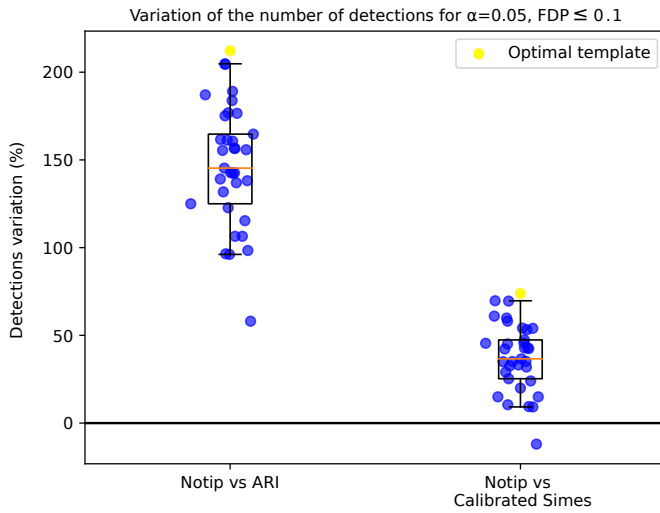
Figure 10: **Variation of the number of detections using many training sets**. For a fixed contrast pair "look negative cue" vs "look negative rating" and 36 different training contrast pairs, we compute the largest possible regions that ensure FDP ≤ 0.1 at risk level $\alpha = 0.05$. Note that for all training contrast pairs except one, Notip performs better than calibrated Simes, with gains ranging from 10 % to 80 % for the optimal template choice described in Section 4.1. In the worst case, Notip performs slightly worse than calibrated Simes.
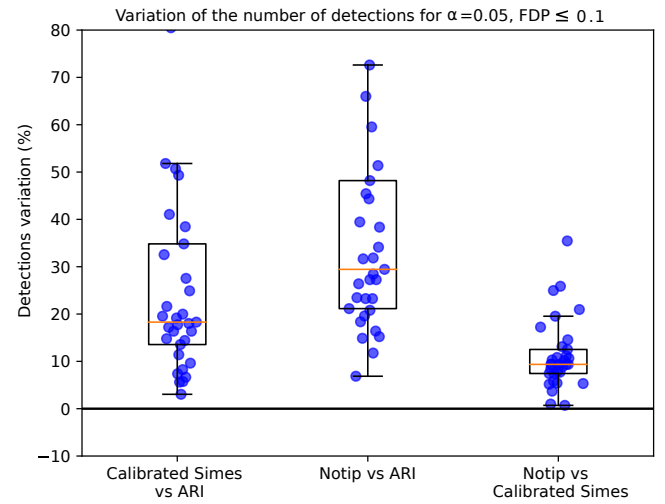

Figure 11: **An example of mismatch between the smoothing factors of training and inference data.** After learning the template on a single contrast pair (see Section 4) with smoothing full width at half maximum (FWHM) 4mm, we perform inference on all 36 pairs smoothed with FWHM 8mm. For each contrast 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.

5.5. Influence of data smoothness

We have seen in Figure 10 that Notip is robust to mismatches of training and inference data across different dimensions (number of subjects, quantity of signal...). We now examine the robustness of Notip with regards to a mismatch of the smoothing parameter between training and inference data.

Figure 11 shows that the smoothing parameter of the training data and the inference data should be matched for optimal performance. Otherwise performance gains relative to the calibrated Simes method are reduced, albeit still positive.

5.6. Using Notip on a single dataset

To assess whether using Notip with the same dataset for training and inference controls the FDP, and whether it yields performance gains compared to ARI and calibrated Simes, we performed 1000 simulations. For each of these runs, we report the empirical FDP and TPR of all three methods.
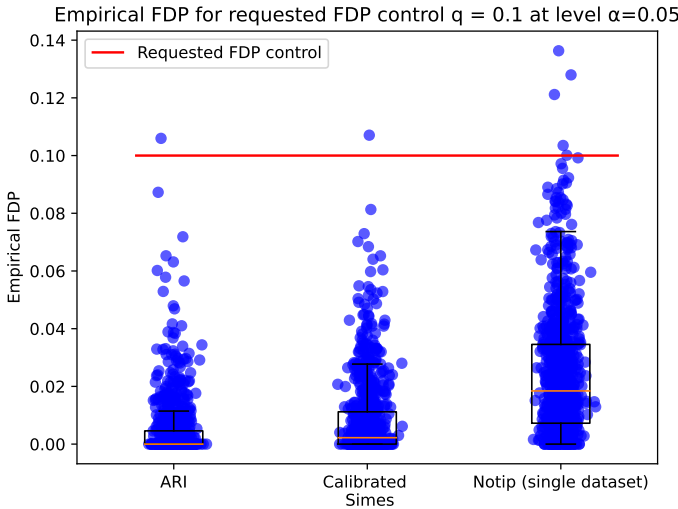
Figure 12: **False Discovery Proportion achieved for a FDP budget** $q = 0.1$ **with risk level** $\alpha = 0.05$ **using Notip on a single dataset**. We run 1000 simulations and report the empirical FDP for each one. Notice that Notip (single dataset) controls the FDP at level $\alpha = 0.05$ since FDP control is violated for 5 runs, i.e. $0.5\% < 5\%$ of all simulations. As expected, ARI and calibrated Simes also control the FDP.

Notice that as seen in Figure 12, Notip (single dataset) indeed controls the FDP, as only 5 points are above the red line - i.e. the FDP was above the budget $q = 0.1$ in 0.5% of experiments ($< \alpha = 5\%$). As expected, ARI and calibrated Simes control the FDP more conservatively.
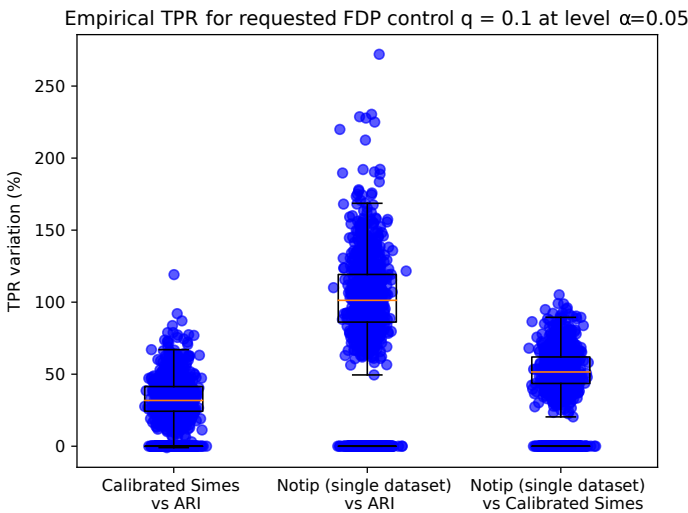

Figure 13: **TPR comparison for an FDP budget** $q = 0.1$ **with risk level** $\alpha = 0.05$ **using Notip on a single dataset**. We run 1000 simulations and report the empirical TPR for each one. Notice that Notip (single dataset) offers substantial performance gains compared to both ARI (100 % on average) and calibrated Simes (50 % on average).

As seen in Figure 13, Notip (single dataset) yields sub-

stantial performance gains compared to ARI and calibrated Simes: 50% on average compared to calibrated Simes, and 100% on average compared to ARI. These gains are comparable to those obtained using the classical Notip method on simulated data (see Figure 6).

6. Discussion

In this paper, we have proposed the Notip procedure, that allows users to obtain statistical guarantees on 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 obtain an upper bound on the FDP, or, equivalently, a lower bound on the TDP in any cluster of interest (see an example in Section 8.5).

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

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 informative guarantee than controlling the FDR, its expected value. We show that our procedure yields a higher number of detections 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 [7], but the gains are expected to be marginal [10].

The gains in detections are maintained across practically all possible training sets, even in cases of poor matching between the training and inference datasets, as seen in Figure 10. Figure 11 also illustrates the robustness of Notip, this time in the case of a poor match of smoothing parameters between the training and inference data. In this case, the gain in detections obtained by using the learned template is reduced, albeit still non-negligible (30% compared to ARI and 9% compared to calibrated Simes). We found that choosing training contrast pairs that contain a large number of subjects and low signal is optimal for performance.

13

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.

Notip comes with an additional computational cost compared to classical calibration using the Simes template, since we have to learn the template before inference. Generally, this additional cost is acceptable in practice since learning a template on a contrast and inferring on a contrast have the same time complexity. If learning a template ex ante is inconvenient or simply impossible, for instance when users only have a single dataset at hand, we have shown numerically that it is possible (though not formally supported by the theory) to use Notip on a single dataset.

We have used $10,000$ 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 $B_{infer} = 1,000$ takes around 45 seconds). This can be trivially parallelized, as natively done in the implementation we propose.

A current 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 [10].

We have achieved the goal of obtaining valid post hoc FDP control - rather than FDR control, or even weaker guarantees on clusters - while maintaining a satisfactory number of detections. 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`.

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. Supplement to "post hoc confidence bounds on false positives using reference families". *Submitted to AoS*, 2018.

[6] 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. Chapman and Hall/CRC, 1st edition edition, November 2021.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

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

[14] 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.

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

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

[17] 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.

[18] 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.

[19] 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.

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

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

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

[23] 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.

[24] 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.

[25] 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.

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

[27] 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.

[28] 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.

[29] 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.

[30] 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.

[31] 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. Visualising learned templates

Figure 14 shows an example of a learned template computed on the contrast pair used as a training set in all the experiments of the paper, using $B = 1000$. We retain a set of 20 quantile curves for clarity - as $1,000$ or $10,000$ curves would not be suitable for visualization. Notice that all threshold families of the learned template are non-linear for small values of k, i.e. $k \leq 1000$. By definition, these curves are similar to the randomized p-values curves displayed in Figure 2.

Figure 14: **Visualising a learned template in log-log scale**. This learned template was computed using $B = 1000$ on the same contrast pair used as training data for the experiments described in Sections 4 and 5. A set of 20 curves are retained for clarity. The curves are increasingly lighter for higher order quantiles.

8.2. 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 [7]. While some guidelines are given in the Discussion of [7], the choice of k_{max} remains an open question. Equation 6 may be written as follows:

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

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 [7] 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

$$k_{max} = q_{max}|S_{max}|, \qquad (10)$$

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 $q \geq 0.5$. In the case of fMRI, one is generally interested in sparse activation extent, as widespread effects 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 $k_{max} = 0.02m = 1,000$. This is the default value of k_{max} in the implementation we propose. To illustrate the effect of the choice of k_{max} we display the variation of the number of detections of all three methods on 36 fMRI datasets across 9 different inference settings for varying k_{max} in Figure 15. Except for extremely small or large values of k_{max} Notip is at worst slightly sub-optimal and $k_{max} = 1,000$ is a safe default.

As noted in [7], 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.

8.3. FDP control on simulated data

In section 4.2 we report the empirical TPR for experiments on simulated data, for which the ground truth is known. We also compute the FDP for each simulation run to verify that, as expected, Notip indeed controls the FDP.

Figure 16: **False Discovery Proportion achieved for a FDP budget** $q = 0.1$ **with risk level** $\alpha = 0.05$. We run 100 simulations and report the empirical FDP for each one. All three methods control the FDP, but Notip is less conservative than ARI and Calibrated Simes.

8.4. Variability of Notip

We have observed relatively high variability in number of detections when comparing Notip to ARI and calibrated Simes in Figure 4. One may wonder whether this variability is inherent to the Notip procedure or stems from the data. To assess this, we report the empirical TPR of each method (rather than the 3 pairwise comparisons) on simulated data, in the same setup as in Figure 6.

Figure 15: **Comparison of the number of detections 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.
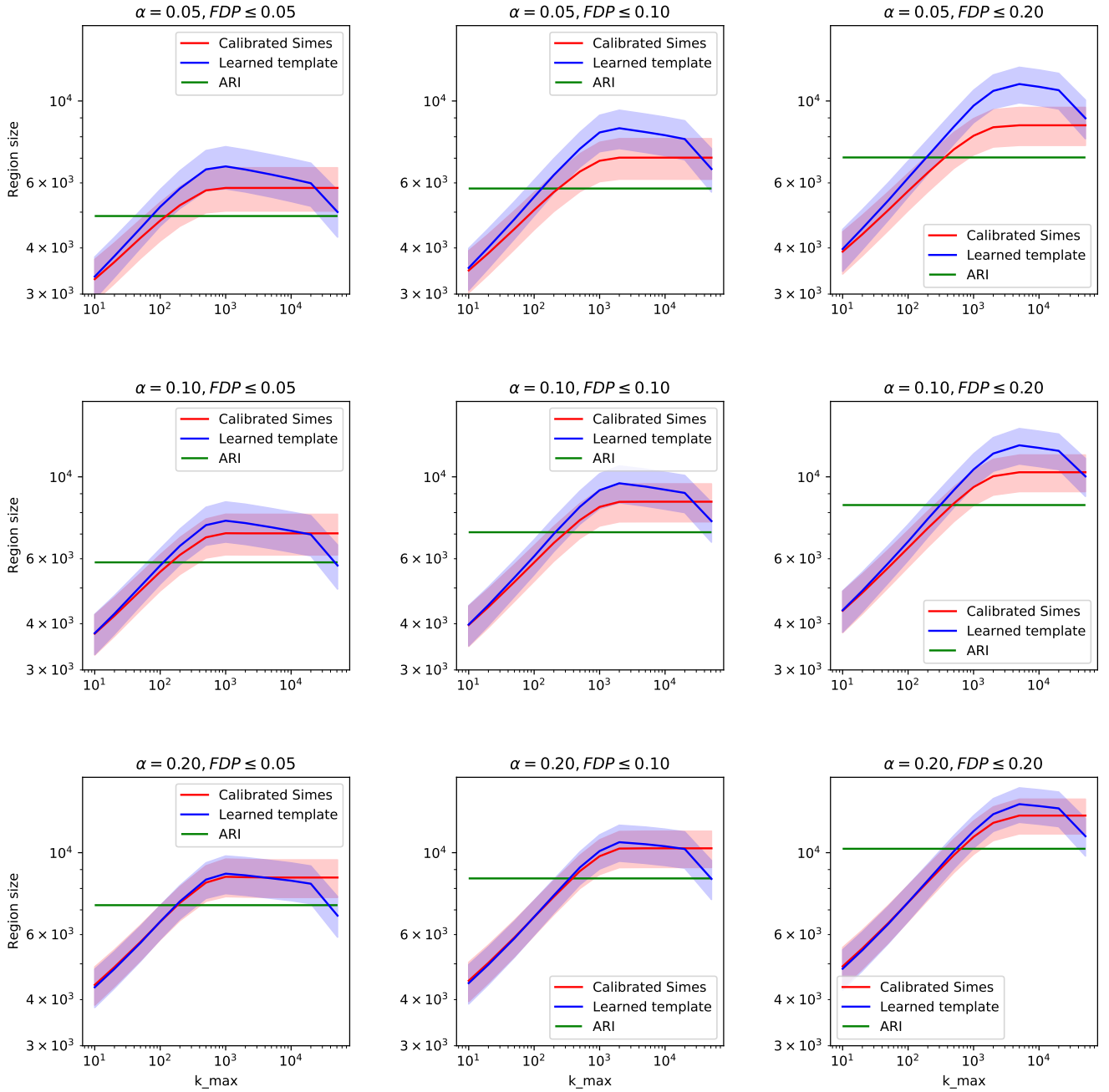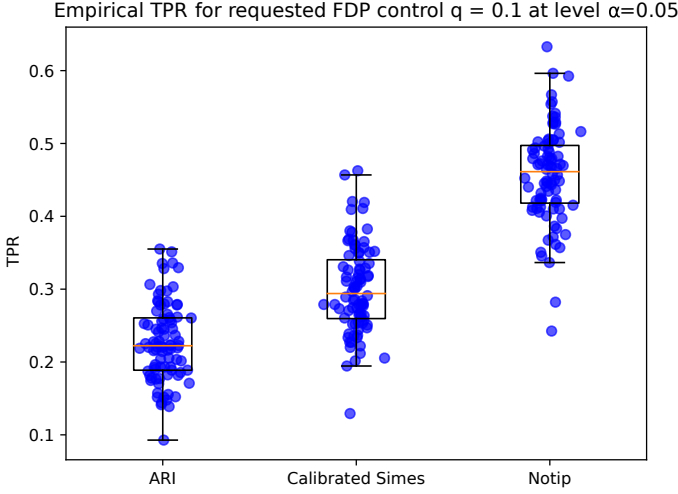
Figure 17: **TPR comparison for an FDP budget** $q = 0.1$ **with risk level** $\alpha = 0.05$. We run 100 simulations and report the empirical TPR for each one. Notice that the variability of performance is similar for all three methods.

Figure 17 indicates that all three methods exhibit similar performance variability on simulated data. This suggests that the variability observed in Figure 4 is due to the data itself rather than to the Notip method.

8.5. TDP lower bounds on clusters

Throughout the paper, we chose to focus on FDP upper bounds - and thus on FDP controlling regions - to make Notip comparable with other methods that control the FDR or the FWER. Since Notip is a post hoc method, it can also be used for inference on data-driven clusters. In this setting, it is natural to formulate the results in terms of TDP lower bounds (obtained as 1 - FDP upper bounds), since users generally want a positive guarantee when inferring on clusters. This is illustrated in Table 2. Notice that Notip is able to offer less conservative guarantees on the TDP in all clusters than both ARI and calibrated Simes. In Table 3 we retain 3 clusters among the 9 found in Table 2 for further study, i.e. changing the cluster-forming threshold to assess its impact on performances of all three methods. In Tables 4 ($z > 2.5$) and 5 ($z > 3.5$), notice that the same clusters are detected with varying sizes. The TDP guarantees remain less conservative using Notip than both ARI and calibrated Simes when the cluster-forming threshold is either lowered to 2.5 or upped to 3.5.

8.6. Additional details on [7]

For self-containedness, this subsection contains additional details on FDP control via the control of the Joint Error Rate,

as described in [7], as well as precise references to the paper's Theorems.

JER control and FDP upper bound

JER control as defined in 5 corresponds to Equation 2 of [7], and the post-hoc bound defined in 6 corresponds to Equation 3 of [7]:

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

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

Randomization

Since the distribution $p_{(k:m_0)}$ is unknown in practice, we use randomization to sample from it, as described in Section 2.4. This is detailed in Section 3 of [7]. In one-sample designs, which is the setting considered throughout experiments, we perform random sign-flipping of samples. Let us denote $\mathscr{G} = \{-1, 1\}^n$ the group of all possible sign-flippings s of size n. This group acts on X of shape (n, m) in the following way:

$$(s.X)_{i,j} = s_i X_{i,j}, i \in \mathbb{N}_n, j \in \mathbb{N}_m.$$

In words, for each sample index $i \in [\![1, n]\!]$, we either sign-flip the sample X_i or leave it untouched. Notice that under the null hypothesis, the distribution of $p_i(X)$ and of $p_i(s.X)$ are equal. Therefore, we are able to approximate the joint distribution of $(p_i(X))_{i \in \mathscr{H}_0(P)}$ conditionally on X by $(p_i(s.X))_{i \in \mathscr{H}_0(P)}$ (see [2]).

Calibration

The final step of the procedure is to perform calibration using the randomized p-values that we previously computed. Once calibration has been performed, a valid post-hoc FDP upper bound is obtained via 6. Calibration is described in Section 3 and written explicitly in Algorithms 2 and 3. In [7], calibration is described more formally, in order to obtain a proof that this procedure indeed yields JER control. For a given template $t_k(\cdot)$ with $1 \le k \le k_{max}$ and a risk level α, the goal is to find the largest λ such that $JER(t_k(\lambda)) \le \alpha$. Formally, we want to compute:

$$\lambda(\alpha) = \max\left\{\lambda \ge 0 : \widehat{JER}(t) \le \alpha\right\}$$

With:

$$\widehat{JER}(t) = \frac{1}{B} \sum_{b=1}^{B} 1\left\{\exists k \in \{1, \ldots, k_{\max}\} : p^b_{(k:m_0)} < t_k(\lambda)\right\}$$

This corresponds to Equation 20 of [7]. Note that the empirical JER is computed using Algorithm 3 in practice. In the case of Notip, λ takes discrete values; the maximum is computed using dichotomy as described in Section 2. Theorem 4.8 of [7] shows that the calibration procedure indeed controls the JER, thus leading to a valid post-hoc FDP upper bound. The proof of Theorem 4.8 can be found in Section 7 of the supplementary material of [5].

8.7. An example of simulated data

Here is an example of simulated data computed in 2D for clarity. We use 3D images in the experiments to mimick fMRI data. Here, we use a 10×10 2D grid and generate the ground truth, a binary mask that defines the signal. Then, we generate n_{infer} null images and n_{infer} images that comprise signal. Substracting these two sets of images results in a list of n_{infer} one-sample images, as in fMRI experiments. In Figure 18 an example of simulated ground truth is displayed, while Figure 19 shows an example of simulated one-sample image. Figure 19 is a noisy version of the ground truth shown in Figure 18.

Figure 19: **A simulation draw.** This 2D 10×10 grid represents a draw of one-sample image comprising signal at locations determined by the binary mask shown in Figure 18. This is a typical example of input data in experiments on simulated data; the goal is then to recover the binary mask using inference methods such as Notip, ARI or calibrated Simes.
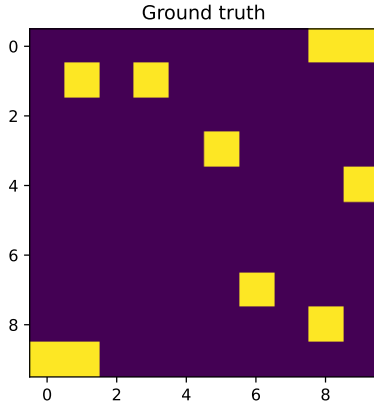

Figure 18: **Simulated ground truth.** This binary mask locates the simulated signal on a 2D 10×10 grid. Signal locations have been drawn randomly and account for $(1 - \pi_0)\%$ of the image, the rest of the image being null data.
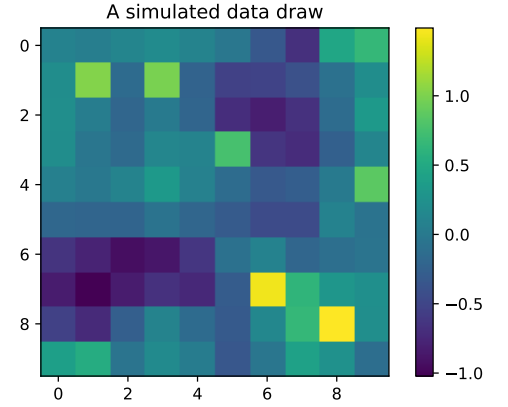
						True Discovery Proportion		
Cluster ID	X	Y	Z	Peak Stat	Cluster Size (mm3)	ARI	Calibrated Simes	Notip
1	-33.0	-94.0	-17.0	5.63	7695	0.17	0.24	0.26
1a	-45.0	-79.0	-26.0	4.56				
1b	-48.0	-61.0	-26.0	4.13				
1c	-51.0	-64.0	-35.0	4.08				
2	66.0	2.0	16.0	5.47	14877	0.20	0.33	0.45
2a	69.0	-22.0	10.0	4.67				
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 TDP lower bound at risk level** $\alpha = 5\%$ using the three possible templates (ARI, Calibrated Simes and Notip) 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`.

						True Discovery Proportion		
Cluster ID	X	Y	Z	Peak Stat	Cluster Size (mm3)	ARI	Calibrated Simes	Notip
1	66.0	2.0	16.0	5.47	14877	0.20	0.33	0.45
1a	69.0	-22.0	10.0	4.67				
1b	69.0	-10.0	13.0	4.59				
1c	69.0	-28.0	13.0	4.43				
2	-12.0	-82.0	-8.0	5.40	14445	0.27	0.38	0.50
2a	30.0	-73.0	-8.0	4.96				
2b	-24.0	-61.0	-11.0	4.91				
2c	30.0	-46.0	-11.0	4.64				
3	-63.0	-34.0	16.0	4.95	25812	0.30	0.48	0.66
3a	-63.0	-10.0	13.0	4.90				
3b	-27.0	-19.0	4.0	4.85				
3c	-57.0	-19.0	7.0	4.68				

Table 3: **Cluster localization (z > 3), size, peak statistic and TDP lower bound at risk level** $\alpha = 5\%$ using the three possible templates (ARI, Calibrated Simes and Notip) on contrast pair 'look negative cue vs look negative rating'. Cluster subpeaks are also reported when relevant. Notice that we retained 3 clusters (originally of indices 2, 3 and 8 of Table 8.5).

Cluster ID	X	Y	Z	Peak Stat	Cluster Size (mm3)	True Discovery Proportion		
						ARI	Calibrated Simes	Notip
1	66.0	2.0	16.0	5.47	28593	0.13	0.18	0.29
1a	69.0	-22.0	10.0	4.67				
1b	69.0	-10.0	13.0	4.59				
1c	69.0	-28.0	13.0	4.43				
2	-12.0	-82.0	-8.0	5.40	23355	0.19	0.23	0.35
2a	30.0	-73.0	-8.0	4.96				
2b	-24.0	-61.0	-11.0	4.91				
2c	30.0	-46.0	-11.0	4.64				
3	-63.0	-34.0	16.0	4.95	43092	0.19	0.25	0.42
3a	-63.0	-10.0	13.0	4.90				
3b	-27.0	-19.0	4.0	4.85				
3c	-57.0	-19.0	7.0	4.68				

Table 4: **Cluster localization (z > 2.5), size, peak statistic and TDP lower bound at risk level** $\alpha = 5\%$ using the three possible templates (ARI, Calibrated Simes and Notip) on contrast pair 'look negative cue vs look negative rating'. Cluster subpeaks are also reported when relevant.

Cluster ID	X	Y	Z	Peak Stat	Cluster Size (mm3)	True Discovery Proportion		
						ARI	Calibrated Simes	Notip
1	66.0	2.0	16.0	5.47	7425	0.38	0.48	0.69
1a	69.0	-22.0	10.0	4.67				
1b	69.0	-10.0	13.0	4.59				
1c	69.0	-28.0	13.0	4.43				
2	-12.0	-82.0	-8.0	5.40	8397	0.46	0.53	0.73
2a	30.0	-73.0	-8.0	4.96				
2b	-24.0	-61.0	-11.0	4.91				
2c	30.0	-46.0	-11.0	4.64				
3	-63.0	-34.0	16.0	4.95	9585	0.46	0.55	0.76
3a	-63.0	-10.0	13.0	4.90				
3b	-57.0	-19.0	7.0	4.68				
3c	-60.0	-49.0	25.0	4.59				

Table 5: **Cluster localization (z > 3.5), size, peak statistic and TDP lower bound at risk level** $\alpha = 5\%$ using the three possible templates (ARI, Calibrated Simes and Notip) on contrast pair 'look negative cue vs look negative rating'. Cluster subpeaks are also reported when relevant.

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 6: **36 pairs of fMRI contrasts used for experiments.** These contrasts images have been downloaded from Neurovault 1952 collection.