

Cluster Size Matters: A Comparative Study of Notip and pARI for Post Hoc Inference in fMRI

Nils Peyrouset,¹ Pierre Neuvial^{1*}, Bertrand Thirion²

¹Institut de Mathématiques de Toulouse; Université de Toulouse; CNRS;
UPS, F-31062 Toulouse Cedex 9, France

²INRIA, Université Paris-Saclay, CEA

*Correspondence: pierre.neuvial@math.univ-toulouse.fr

Abstract

All Resolutions Inference (ARI) is a post hoc inference method for functional Magnetic Resonance Imaging (fMRI) data analysis that provides valid lower bounds on the proportion of truly active voxels within any, possibly data-driven, cluster. As such, it addresses the paradox of spatial specificity encountered with more classical cluster-extent thresholding methods. It allows the cluster-forming threshold to be increased in order to locate the signal with greater spatial precision without overfitting, also known as the drill-down approach. Notip and pARI are two recent permutation-based extensions of ARI designed to increase statistical power by accounting for the strong dependence structure typical of fMRI data.

A recent comparison between these papers based on large voxel clusters concluded that pARI outperforms Notip. We revisit this conclusion by conducting a systematic comparison of the two. Our reanalysis of the same fMRI data sets from the Neurovault database demonstrates the existence of complementary performance regimes: while pARI indeed achieves higher sensitivity for large clusters, Notip provides more informative and robust results for smaller clusters. In particular, while Notip supports informative “drill-down” exploration into subregions of activation, pARI often yields non-informative bounds in such cases, and can even underperform the baseline ARI method.

1 Introduction

A classical approach to statistical inference for functional Magnetic Resonance Imaging (fMRI) data is cluster-extent-based thresholding. This method aims to identify clusters of adjacent voxels containing at least one active voxel [Nichols and Hayasaka, 2003]. This approach suffers from two known limitations. First, larger clusters provide less information than smaller ones, a phenomenon known as the spatial specificity paradox [Woo et al., 2014]. Second, when clusters are

zoomed-in (or “drilled-down”) by choosing a more stringent threshold, a form of double dipping occurs, resulting in the loss of statistical control [Kriegeskorte et al., 2009].

Post hoc inference aims to address these limitations by providing statistical guarantees on the number or proportion of active voxels in arbitrary, possibly user-defined clusters [Goeman and Solari, 2011]. The first application of post hoc inference to fMRI data is the All Resolutions Inference (ARI) [Rosenblatt et al., 2018]. ARI relies on the Simes inequality [Simes, 1986], which can be conservative due to the strong positive dependence typically found in fMRI data [Andreella et al., 2023]. Recently, several improvements to ARI have been proposed specifically for fMRI data analysis, leading to the Notip [Blain et al., 2022] and pARI [Andreella et al., 2023] approaches.

All of these methods are based on the idea of calibration [Blanchard et al., 2020], which uses permutation or sign-flipping to learn and adapt to the dependency structure of the data set at hand. From a theoretical perspective, these methods differ only in the choice of the so-called *template* (set of thresholds), which implicitly determines the relative weight given to smaller or larger sets of voxels. A recent comparison between Notip and pARI based on the data sets originally analyzed in [Blain et al., 2022] has shown that pARI outperforms Notip for large clusters [Andreella et al., 2024].

In this paper, based on the same data sets, we demonstrate the existence of regimes in which one method outperforms the other one, a conclusion in line with the idea of “No Free Lunch”. The increased sensitivity of pARI for larger clusters (already noted by Andreella et al. [2024]) comes at the price of decreased sensitivity for smaller clusters. In practice, for clusters of several hundreds of voxels, pARI is generally outperformed by Notip, but also by the baseline method ARI. This implies that contrary to Notip and ARI, pARI does not provide informative results when drilling down into the clusters with the highest signal values.

In the remainder of the paper, we provide a self-contained description of the compared methods (Section 2) and report extensive results on 37 fMRI data sets (Section 3), and provide a short discussion of the consequences of these results (Section 4).

2 Methods

2.1 Post hoc inference for true discovery proportions

For each of m voxels, we test the null hypothesis that voxel i is not active under the condition of interest. The set $\mathcal{H}$ of all tested hypotheses is then identified to the set of all voxels, i.e. $\mathcal{H} = [m]$, where $[n] = \{1, \dots, n\}$ for any integer n . We denote by $\mathcal{H}_0 \subset \mathcal{H}$ the (unknown) subset of true null hypotheses. Let $m_0 = |\mathcal{H}_0|$ be the (unknown) number of true null hypotheses and $\pi_0 = m_0/m$ be the corresponding proportion. For an arbitrary selection of voxels $S \subset \mathcal{H}$, $|S \cap \mathcal{H}_0|$ is the number of false positives within S , that is, the number of voxels that

are selected whereas their corresponding null hypothesis is true (i.e., inactive voxels). The corresponding True Discovery Proportion (TDP) is then defined as $\text{TDP}(S) = 1 - |S \cap \mathcal{H}_0|/|S|$.

Post hoc inference [Goeman and Solari, 2011] aims at building an $(1 - \alpha)$ a **TDP lower bound**, that is, a function $\overline{\text{TDP}}_\alpha$ such that

$$\mathbb{P}\left(\forall S \subset \mathcal{H}, \text{TDP}(S) \geq \overline{\text{TDP}}_\alpha(S)\right) \geq 1 - \alpha. \quad (1)$$

That is, with probability greater than $1 - \alpha$, the proportion of true discoveries of any subset S is at least $\overline{\text{TDP}}_\alpha$. We emphasize that the "for all S " in (1) is *inside* the probability: this implies that a TDP lower bound is valid for any number of possibly data-driven sets S . In the context of fMRI studies, such a TDP lower bound is applicable to all voxel clusters obtained by thresholding a statistical map. Moreover, multiple cluster-forming thresholds may be chosen, possibly based on the results of the data analysis, without compromising the statistical validity of the TDP lower bound. Therefore, as argued by Rosenblatt et al. [2018], post hoc methods address the problem of double dipping in fMRI data analysis, allowing users to "drill down" from the cluster level to sub-regions, and even to individual voxels, in order to pinpoint the origin of the activation".

Joint Error Rate Control. Blanchard et al. [2020] have shown that post hoc bounds may be systematically derived from the control of a statistical risk called the Joint Error Rate, by a simple interpolation principle. We consider a vector of p -values associated with each voxel: $\mathbf{p} = (p_1, \dots, p_m)$. For a positive integer K , let $\mathbf{t} = (t_k)_{k \in [K]}$ be a non-decreasing vector of thresholds in $(0, 1)$ aka template. The Joint Error Rate of the family $\mathbf{t}$ is defined by

$$\text{JER}(\mathbf{t}) = \mathbb{P}(\exists k \in \mathcal{H}_0 \cap [K], p_{(k:\mathcal{H}_0)} < t_k), \quad (2)$$

where for $A \subset \mathcal{H}$ we denote by $p_{(k:A)}$ the k -th smallest p -value among $(p_i)_{i \in A}$. Blanchard et al. [2020] have shown that if $\text{JER}(\mathbf{t}) \leq \alpha$, then a TDP lower bound (1) is given by the function $\overline{\text{TDP}}^{\mathbf{t}}$, defined for $S \subset \mathcal{H}$ by

$$\overline{\text{TDP}}^{\mathbf{t}}(S) = |S|^{-1} \left(\max_{k \in [K]} 1 - k + \sum_{i \in S} \mathbf{1}\{p_i < t_k\} \right). \quad (3)$$

After an initial sorting of the p -values, the bound $\overline{\text{TDP}}^{\mathbf{t}}(S)$ can be computed in linear time ($O(|S|)$) from (3), as shown by Enjalbert-Courrech and Neuviat [2022]¹. This framework for deriving post hoc bounds is particularly convenient in practice. Indeed, since the *computational* problem of the efficient evaluation of the post hoc bound is solved once and for all, the only remaining challenge to obtain a post hoc bound is the *statistical* problem of finding a JER controlling family $\mathbf{t}$.

¹A generic implementation applicable to any JER controlling family $\mathbf{t}$ is provided in the R package `sanssouci` and in the Python package `sanssouci.python`.

2.2 Building JER controlling families

The Simes family. The simplest example of JER controlling family is the Simes family, defined by $t_k = \alpha k/m$ for $k \in [m]$. The Simes [1986] inequality states that for independent or positively associated test statistics [Sarkar, 2008], in the sense of the PRDS property introduced by Benjamini and Yekutieli [2001], we have:

$$\mathbb{P}(\exists k \in \mathcal{H}_0, p_{(k:\mathcal{H}_0)} < \alpha k/m_0) \leq \alpha. \quad (4)$$

As the left hand side of (4) is exactly the JER of the Simes family, the Simes inequality trivially implies that the Simes family controls JER at level $m_0\alpha/m = \pi_0\alpha$, and a fortiori at level α . It has been shown in Blanchard et al. [2020] that the post hoc bound obtained by interpolation recovers the Simes bound obtained by Goeman and Solari [2011] by combining closed testing [Marcus et al., 1976] with a dedicated computational shortcut. The All Resolutions Inference (ARI) method [Rosenblatt et al., 2018] is an improved version of the above Simes-based method, where the thresholds $t_k = \alpha k/m$ are replaced with $t_k = \alpha k/h(\alpha)$, where $h(\alpha) \leq m$ is the Hommel value introduced in Hommel [1988], which satisfies $m_0 \leq h(\alpha)$ with probability $1 - \alpha$ [Goeman et al., 2019].

The Simes inequality and the ARI method, which is based on it, are usually conservative in high-dimensional cases with dependent test statistics [Blanchard et al., 2020, Enjalbert-Courrech and Neuvial, 2022]. Such situations are a common use case in neuroimaging or genomic data (see e.g. Hayasaka and Nichols [2003]). This translates into the conservativeness of the associated post hoc bound: in such scenarios, the coverage of the post hoc bound (1) can be substantially larger than $1 - \alpha$.

JER Calibration. To address this conservativeness, a natural idea is to seek for other JER controlling families, whose JER is closer to the *risk budget* α . Given a threshold family $\mathbf{t}$ the JER (2) only depends on the *joint* distribution of the null p -values, which is generally unknown. To address this issue, Blanchard et al. [2020] have introduced a generic approach to approximating the JER. This approach involves sampling from the joint distribution using randomization-based methods. In particular, the work of Blanchard et al. [2020] covers the classical cases of group label permutations for two-group testing and sign flipping for one-group testing. More general linear models are covered in Davenport et al. [2025]. Starting from a set of candidate families $\mathbf{t}$ called a template, JER calibration methods select the family $\mathbf{t}^*$ whose JER is the largest among those below the target risk/budget α . For a graphical illustration of the JER calibration principle, see Blanchard et al. [2021], Blain et al. [2022], Andreella et al. [2024].

2.3 Existing JER calibration methods

Following Andreella et al. [2024], we focus on the two most recent post hoc inference methods for the mass-univariate analysis of neuroimaging data [Blain

et al., 2022, Andreella et al., 2023]. Both of them are based on JER calibration, and they only differ by the choice of candidate families (a.k.a. template).

Andreella et al. [2023] have introduced the permutation ARI (pARI) method. As its name suggests, it is inspired by the ARI method: the recommended candidate threshold families for neuroimaging data are of the form $t_k^\delta(\lambda) = \frac{\lambda(k-\delta)}{(m-\delta)} \mathbb{1}_{\{k > \delta\}}$, where δ is an integer hyperparameter which has to be specified before data analysis to avoid circularity issues. This hyperparameter indirectly controls where the method concentrates its power, via the minimal size of a region where non trivial inference can be made. The choice $\delta = 0$ recovers the Calibrated Simes method introduced by Blanchard et al. [2020], whose numerical performance had already been studied by Enjalbert-Courrech and Neuvial [2022] for genomic applications. Andreella et al. [2023] recommend “fixing $\delta = 1$ if the practitioner is interested in computing the lower bound for the TDP in small clusters, while $\delta > 1$ if the attention is focused on large cluster”. In their application to fMRI data, Andreella et al. [2023] chose $\delta = 1$ for their analysis of Auditory Data, and $\delta = 27$ for their analysis of Rhyme Data. According to the follow-up paper Andreella et al. [2024], the choice $\delta = 27$ is recommended for the analysis of fMRI data.

Blain et al. [2022] have introduced the Notip method, where the main innovation is that the candidate threshold families are *data-driven* instead of considering a pre-specified parametric part. In practice, Notip performs a first round of permutation on the data set at hand, and uses the successive empirical quantiles of the obtained null statistics as threshold families. The size K of the threshold families is set to 2% of the total number of voxels, that is, $K = 1000$ when $m = 50000$ voxels².

3 Results

Following Andreella et al. [2024], we consider both methods using the parameter values recommended by their authors, i.e. $\delta = 27$ for pARI and $k_{\max} = 1000$ for Notip. We start by studying one particular contrast, and then give general results on a set of 37 contrasts.

3.1 Focus on one contrast

Here, we focus on the “Look negative cue vs Look negative rating” contrast, taken from the Neurovault collection³. This dataset was already studied in [Blain et al., 2022, Figures 5 and 7] to illustrate the face validity of the Notip method. It was also used in [Blain et al., 2022, Tables 2 to 5] to compare Notip to the baseline method ARI and to pARI with $\delta = 0$ (referred to as

²Assuming that the proportion of active voxels in a typical fMRI data set is typically small (say, below 5%) and considering that TDP bounds below 1/2 are not informative, Blain et al. [2022] have shown that one can focus on the 2.5% percent of largest p -values, rounded to 2% for simplicity.

³The corresponding data are available from <http://neurovault.org/collections/1952>.

“calibrated Simes“ in [Blain et al. \[2022\]](#). The FDP bounds obtained by these methods are compared for each cluster obtained by a cluster-defining threshold of $z \in \{2.5, 3, 3.5\}$. This comparison shows that Notip outperforms the other methods available at that time (therefore not including pARI with $\delta = 27$), for all values of z .

This comparison was complemented by [\[Andreella et al., 2024, Figure 4 and Table 1\]](#), where the pARI method with $\delta = 27$ was added for $z = 3$. They observed that pARI (with $\delta = 27$) outperformed Notip in this case. We were able to reproduce this observation (see [Table 1](#)). In order to complement this study, we have considered other choices for z . The results for $z = 3.5$ are reported in [Figure 1](#), where the clusters are represented on a glass brain plot. The comparison results are more contrasted than those reported in [Andreella et al. \[2024\]](#) for $z = 3$: while Notip and pARI yield comparable TDP bounds for the largest clusters (clusters 2, 3 and 8), Notip performs better than pARI for smaller clusters. This behavior is somewhat expected: as noted by [Andreella et al. \[2024\]](#), “ $k_{\max}$ focuses power of Notip away from very large clusters, while $\delta > 0$ focuses power of pARI away from small ones”.

A more complete picture is brought by [Figure 2](#), where the TDP lower bounds associated with all possible choices of z (or equivalently, all possible p -value level sets) are displayed for each method. For each value of k , we plot for each method the TDP lower bound $\text{TDP}(S_k)$ obtained for the set $S_k = \{i \in \mathcal{H}, |Z_i| \geq Z_{(k)}\}$ of voxels corresponding to the k largest Z scores. In particular, the values of k corresponding to $Z_{(k)} \in \{3, 3.5, 4, 4.5\}$ are highlighted by dotted vertical lines.

Since all of the compared method are valid post hoc bounds, the performance of a method can be quantified by the TDP bound, with higher values corresponding to a better bound. First, the Notip and pARI curves cross each other: this reflects the fact that no method is uniformly more powerful than the other one. This point illustrates the absence of “free lunch” predicted by the theory outlined in [Section 2](#): both methods optimize the same objective function, targeting a JER of α by estimating the joint null p -value distribution using permutations. However, they have different constraints, which are encoded by the choice of a template.

As illustrated in [Figure 1](#) for a specific value of z , the performance of Notip and pARI is comparable in the region $3 \leq z \leq 4$, with Notip better for larger values of z (i.e., smaller values of k) and pARI better than Notip for smaller values of z (i.e., larger values of k). For smaller sets, the performance of pARI drops massively. This is expected for very small sets: by construction, pARI cannot detect any signal in sets of size less than $\delta = 27$, corresponding here to 729 mm^3 . However, pARI performs worse than the baseline ARI method for sets smaller than 800 voxels. This is alarming since the ARI method is known to be conservative for fMRI data [\[Blain et al., 2022\]](#). In fact, this conservativeness was the main motivation of the pARI method [\[Andreella et al., 2023\]](#).

A major feature of post hoc methods is their ability to “further ‘drill down’ from the cluster level to sub-regions, and even to individual voxels, in order to pinpoint the origin of the activation” [\[Rosenblatt et al., 2018\]](#). In theory, all

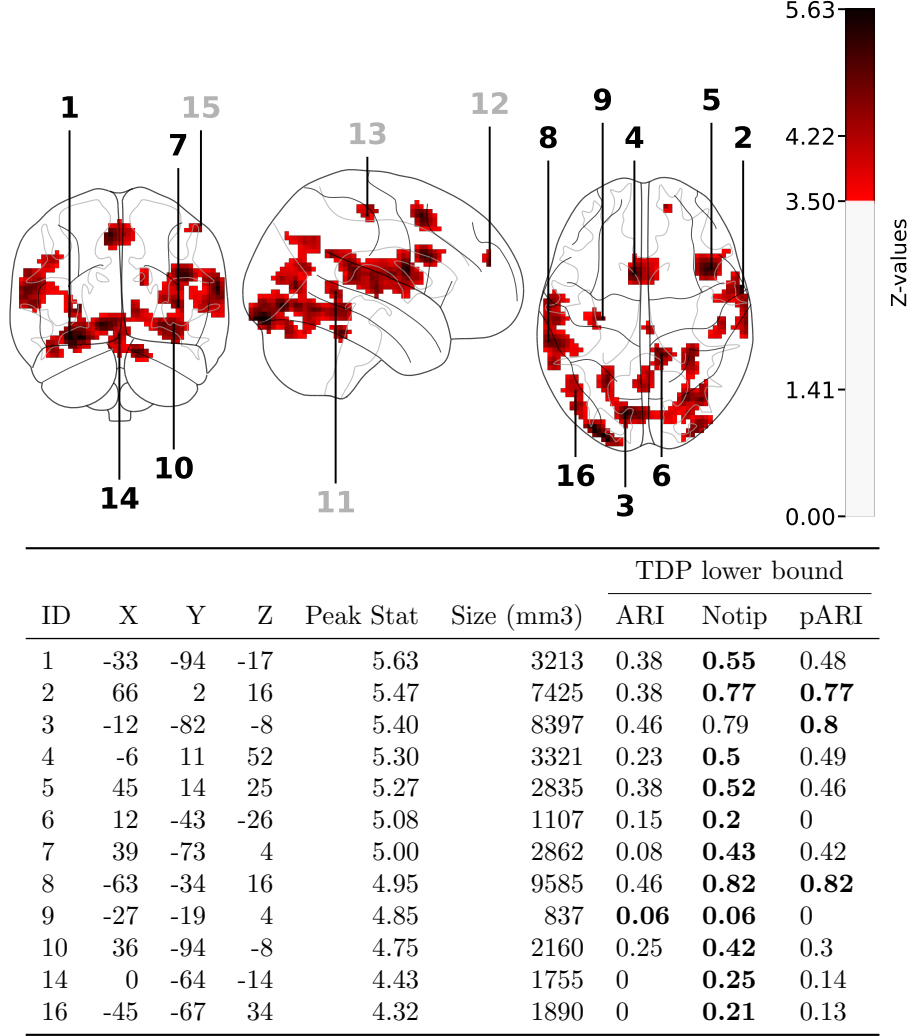


Figure 1: Clusters identified with threshold $z = 3.5$ for the “Look negative cue” vs “Look negative rating” data set: glass brain plot (top) and comparison between TDP lower bounds (bottom) For each cluster, the values in bold indicate the best result. Only clusters for which signal is detected by at least one method are reported.

of the methods discussed here have this capacity since they provide TDP lower bounds that are valid for all possible sets of voxels simultaneously. However, statistical validity (i.e., JER control) does not necessarily imply statistical power. In particular, *the TDP lower bounds obtained by pARI (with $\delta = 27$) for small sets of voxels are non-informative*. Table 2 illustrates this point numerically. It provides the TDP lower bounds for the same dataset when drilling down to

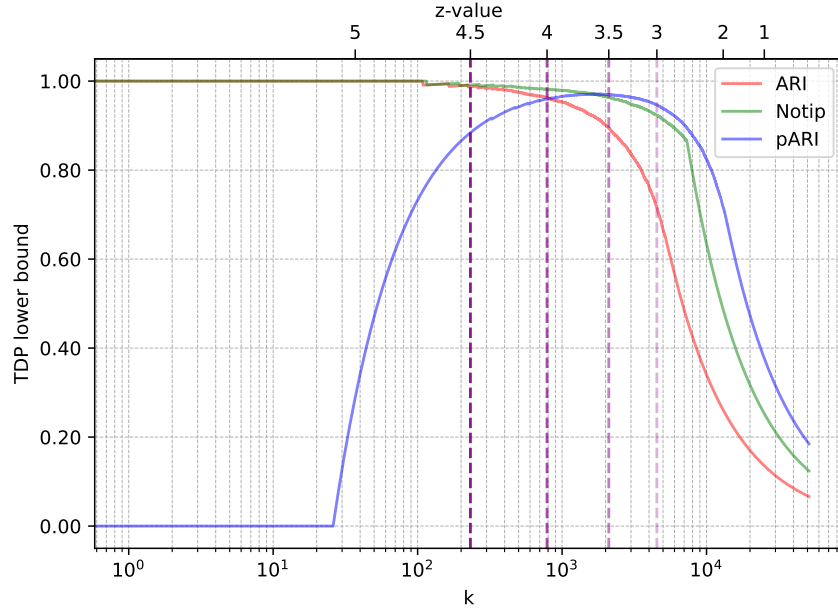


Figure 2: Confidence curve on the TDP for the “Look negative cue vs Look negative rating” contrast: for each $k \in [m]$, we plot the TDP lower bound $\overline{\text{TDP}}(S_k)$, where $S_k = \{i \in \mathcal{H}, |Z_i| \geq Z_{(k)}\}$ is the set of voxels with the k largest Z scores.

$z > 4$: pARI provides trivial (i.e. null) lower bounds for most clusters, and is outperformed by ARI (and a fortiori by Notip) even for the largest clusters of more than 3,000 mm³, corresponding to more than 100 voxels. In practice, the pARI method cannot drill down to $z > 4$ in this example. In contrast, the Notip method is uniformly more powerful than the baseline ARI method and enables informative drilling down. In the next section, we demonstrate that these observations are general, and not specific to this particular data set.

3.2 fMRI datasets from the Neurovault database

To consolidate the above results, we conducted experiments on a large fMRI data set: collection 1952 [Varoquaux et al., 2018] of the Neurovault database (<http://neurovault.org/collections/1952>). This dataset is an aggregation of 20 different fMRI studies and consists of statistical maps obtained at the individual level for a large set of contrasts. We focused on 37 fMRI contrasts: the “Look negative cue vs Look negative rating” contrast studied above, and the 36 contrasts introduced in Blain et al. [2022] and further studied in Andreella et al. [2024].

We perform the same analysis for each contrast as in Section 3.1. For each

threshold value of the Z statistic and each contrast, we obtain a list of clusters and compute a TDP bound for each compared method. The results corresponding to Table 1 and Figure 2 for each of these 37 contrasts are available at https://github.com/pilsneyrouset/comparison_Notip-pARI.

These results are summarized in Figure 3, where each panel corresponds to a value for the cluster-forming threshold z . For each method, the TDP bound of each cluster is plotted against its size. For a given value of z , larger clusters

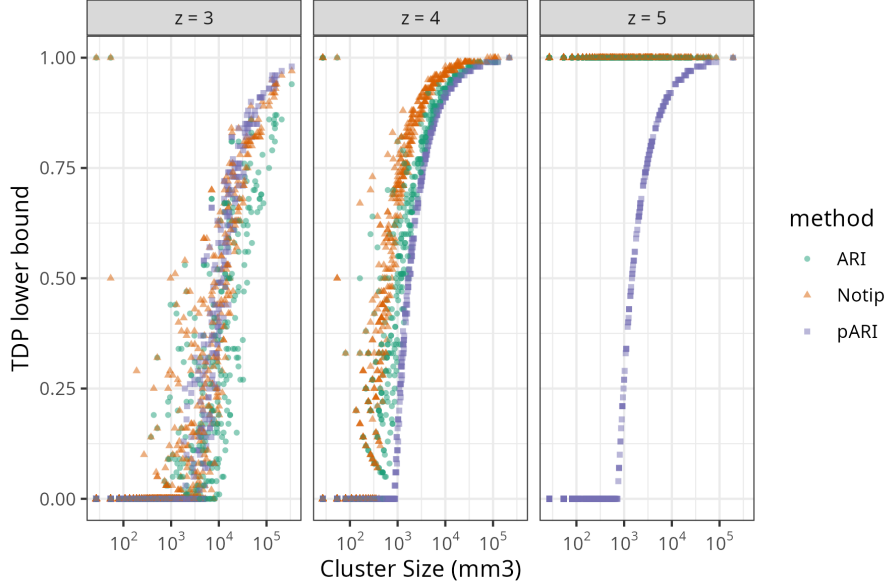


Figure 3: Lower bound on the True Discovery Proportion $\overline{\text{TDP}}(S)$ as a function of cluster size $|S|$, for each cluster S identified at the cluster forming threshold $z = 3$ (left panel), $z = 4$ (center panel), and $z = 5$ (right panel).

correspond to regions where the signal is stronger. Accordingly, the TDP bounds tend to be larger for larger clusters for a given method.

Comparing between methods based on the TDP bounds leads to the following conclusions. First, Notip consistently outperforms ARI. In particular, as noted above in the case of the dataset studied in Section 3.1, it retains and improves upon the drill-down ability of the ARI method.

When the cluster-forming threshold is set very high ($z = 5$), the signal is so strong that both methods report that the TDP is equal to 1. This corresponds to pure signal in all 481 clusters obtained across the 37 datasets: those are subsets of the Family-Wise Error Rate (FWER)-controlling set provided by the Bonferroni-Holm method [Holm, 1979]. In contrast, the pARI method detects pure signal in only 1 of these 481 clusters.

For a small value of the cluster-forming threshold ($z = 3$), pARI outperforms ARI for larger clusters, and its performance is globally similar to Notip's.

Consistent with the results reported by Andreella et al. [2024], pARI slightly outperforms Notip for larger clusters. However, the results differ markedly for smaller clusters. Here, pARI almost always yields null TDP bounds, while both Notip and ARI provide informative (i.e. non-null) TDP bounds.

Naturally, increasing the value of the cluster-forming threshold value to $z = 4$ leads to larger TDP bounds for all methods. However this improvement is not uniform across methods. pARI systematically underperforms compared to Notip and the baseline ARI method. pARI’s performance drops even more dramatically at $z = 5$, where it provides uninformative TDP bounds for clusters of size below 1,000 mm³ and non-trivial but massively underestimated TDP bounds for larger clusters.

4 Discussion

We have performed an extensive comparison between two recently proposed methods for post hoc inference for fMRI data: Notip [Blain et al., 2022] and pARI [Andreella et al., 2023]. As expected from the theory, since both are based on the same calibration principle [Blanchard et al., 2020], our numerical experiments confirm that neither of Notip nor pARI is uniformly more powerful than the other (no free lunch).

This study illustrates the importance and difficulty of objectively assessing the performance of methods. The Notip paper [Blain et al., 2022] focused on the size of the largest detected regions because the relative behavior of the different compared methods did not depend on the the region size. Notip remains consistently better than the baseline method, ARI. In contrast, the pARI method introduces a parameter δ , which indirectly controls the size of the smallest cluster for which informative TDP bounds can be obtained [Andreella et al., 2023]. Therefore, performance comparisons involving pARI must also consider smaller clusters.

Our experiments have shown that pARI can perform dramatically worse than Notip and even the baseline method ARI, especially in regions with a large amount of signal. Unfortunately, this precludes the drill-down approach advocated by Rosenblatt et al. [2018], where the cluster-forming threshold is increased in order to locate the signal with greater spatial precision. This limitation can be problematic since low cluster forming thresholds have been shown to lead to unreliable inference [Woo et al., 2014]. Specifically, [Woo et al., 2014] “recommend setting $p < .001$ as a lower limit default, and using more stringent primary thresholds or voxel-wise correction methods for highly powered studies”.

The methods discussed in this paper are not specific to fMRI studies and can be used in other contexts, such as genomics (see e.g. Enjalbert-Courrech and Neuviat [2022]), provided relevant hyperparameters are chosen.

Finally, we would like to remind users that the hyperparameters discussed in this work (δ for pARI and K for Notip) must be set prior to data analysis. Selecting them after the fact is another instance of double-dipping.

References

- A. Andreella, J. Hemerik, L. Finos, W. Weeda, and J. Goeman. Permutation-based true discovery proportions for functional magnetic resonance imaging cluster analysis. *Statistics in Medicine*, 42(14):2311–2340, 2023.
- A. Andreella, A. Vesely, W. Weeda, and J. Goeman. Selective inference for fmri cluster-wise analysis, issues, and recommendations for critical vector selection: A comment on blain et al. *Imaging Neuroscience*, 2:1–7, 2024.
- Y. Benjamini and D. Yekutieli. The control of the false discovery rate in multiple testing under dependency. *Annals of statistics*, pages 1165–1188, 2001.
- A. Blain, B. Thirion, and P. Neuvial. Notip: Non-parametric true discovery proportion control for brain imaging. *NeuroImage*, 260:119492, 2022.
- G. Blanchard, P. Neuvial, E. Roquain, et al. Post hoc confidence bounds on false positives using reference families. *Annals of Statistics*, 48(3):1281–1303, 2020.
- G. Blanchard, P. Neuvial, and E. Roquain. On agnostic post hoc approaches to false positive control. In X. Cui, T. Dickhaus, Y. Ding, and J. C. Hsu, editors, *Handbook of Multiple Comparisons*, Chapman & Hall/CRC Handbooks of Modern Statistical Methods. Chapman and Hall/CRC, 1st edition edition, Nov. 2021.
- S. Davenport, B. Thirion, and P. Neuvial. Fdp control in mass-univariate linear models using the residual bootstrap. *Electron. J. Statist.*, 19(1):1313–1336, 2025. ISSN 1935-7524. doi: 10.1214/25-EJS2354.
- N. Enjalbert-Courrech and P. Neuvial. Powerful and interpretable control of false discoveries in two-group differential expression studies. *Bioinformatics*, 38(23):5214–5221, 10 2022. ISSN 1367-4803.
- J. J. Goeman and A. Solari. Multiple testing for exploratory research. *Statistical Science*, 26(4):584–597, 2011.
- J. J. Goeman, R. J. Meijer, T. J. Krebs, and A. Solari. Simultaneous control of all false discovery proportions in large-scale multiple hypothesis testing. *Biometrika*, 106(4):841–856, 2019.
- S. Hayasaka and T. E. Nichols. Validating cluster size inference: random field and permutation methods. *Neuroimage*, 20(4):2343–2356, Dec. 2003.
- S. Holm. A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics*, pages 65–70, 1979.
- G. Hommel. A stagewise rejective multiple test procedure based on a modified bonferroni test. *Biometrika*, 75(2):383–386, 1988.

- N. Kriegeskorte, W. K. Simmons, P. S. Bellgowan, and C. I. Baker. Circular analysis in systems neuroscience: the dangers of double dipping. *Nature neuroscience*, 12(5):535, 2009.
- R. Marcus, P. Eric, and K. R. Gabriel. On closed testing procedures with special reference to ordered analysis of variance. *Biometrika*, 63(3):655–660, 1976.
- T. Nichols and S. Hayasaka. Controlling the familywise error rate in functional neuroimaging: a comparative review. *Statistical methods in medical research*, 12(5):419–446, 2003.
- J. D. Rosenblatt, L. Finos, W. D. Weeda, A. Solari, and J. J. Goeman. All-resolutions inference for brain imaging. *Neuroimage*, 181:786–796, 2018.
- S. K. Sarkar. On the simes inequality and its generalization. In *Beyond parameters in interdisciplinary research: Festschrift in honor of Professor Pranab K. Sen*, volume 1, pages 231–243. Institute of Mathematical Statistics, 2008.
- R. J. Simes. An improved bonferroni procedure for multiple tests of significance. *Biometrika*, 73(3):751–754, 1986.
- G. Varoquaux, Y. Schwartz, R. A. Poldrack, B. Gauthier, D. Bzdok, J.-B. Poline, and B. Thirion. Atlases of cognition with large-scale human brain mapping. *PLoS computational biology*, 14(11):e1006565, 2018.
- C.-W. Woo, A. Krishnan, and T. D. Wager. Cluster-extent based thresholding in fmri analyses: pitfalls and recommendations. *NeuroImage*, 91:412–419, May 2014. ISSN 1095-9572. doi: 10.1016/j.neuroimage.2013.12.058.

Appendix

A Additional numerical results

A.1 “Look negative cue” vs “Look negative rating” dataset

In Table 1, we reproduce the results obtained for $z = 3$ in [Blain et al., 2022, Table 2] and complemented by [Andreella et al., 2024, Table 1]. Note that the results of Table 1 are not exactly identical to those in Andreella et al. [2024] because of the numerical variability inherent to the use of random permutation in the analysis. We also provide additional results corresponding to $z = 4, 4.5$ and 5 in Tables 2, Tables 3 and 4.

ID	X	Y	Z	Peak Stat	Size (mm3)	TDP lower bound		
						ARI	Notip	pARI
1	-33	-94	-17	5.63	7695	0.17	0.3	0.33
2	66	2	16	5.47	14877	0.2	0.46	0.57
3	-12	-82	-8	5.40	14445	0.27	0.52	0.59
4	-6	11	52	5.30	5238	0.14	0.31	0.32
5	45	14	25	5.27	4563	0.24	0.33	0.28
6	12	-43	-26	5.08	12555	0.05	0.36	0.5
7	39	-73	4	5.00	6075	0.04	0.21	0.23
8	-63	-34	16	4.95	25812	0.3	0.66	0.75
9	36	-94	-8	4.75	6507	0.08	0.19	0.19

Table 1: “Look negative cue” vs “Look negative rating” dataset: comparison between lower bounds on the True Discovery Proportion for the cluster-defining threshold $z = 3$.

ID	X	Y	Z	Peak Stat	Size (mm3)	TDP lower bound		
						ARI	Notip	pARI
1	-33	-94	-17	5.63	1431	0.64	0.74	0.42
2	66	2	16	5.47	2997	0.74	0.87	0.72
3	-12	-82	-8	5.40	1431	0.58	0.75	0.42
4	-6	11	52	5.30	1485	0.51	0.76	0.44
5	45	14	25	5.27	1755	0.62	0.78	0.52
6	12	-43	-26	5.08	459	0.35	0.47	0
7	39	-73	4	5.00	405	0.2	0.4	0
8	30	-73	-8	4.96	567	0.29	0.43	0
9	-63	-34	16	4.95	3726	0.79	0.9	0.78
10	-24	-61	-11	4.91	594	0.32	0.45	0
11	-27	-19	4	4.85	216	0.25	0.25	0
12	36	-94	-8	4.75	1134	0.48	0.69	0.26
13	30	-46	-11	4.64	1188	0.5	0.68	0.3
14	-60	-49	25	4.59	324	0	0.08	0
15	-45	-79	-26	4.56	513	0	0.32	0
20	0	-64	-14	4.43	378	0	0.07	0
23	-45	-67	34	4.32	405	0	0.13	0

Table 2: “Look negative cue” vs “Look negative rating” dataset: comparison between lower bounds on the True Discovery Proportion for the cluster-defining threshold $z = 4$.

ID	X	Y	Z	Peak Stat	Size (mm3)	TDP lower bound		
						ARI	Notip	pARI
1	-33	-94	-17	5.63	918	0.91	0.94	0.21
2	66	2	16	5.47	513	0.89	0.89	0
3	-12	-82	-8	5.40	783	0.93	0.93	0.07
4	-6	11	52	5.30	594	0.86	0.91	0
5	45	14	25	5.27	783	0.93	0.93	0.07
6	12	-43	-26	5.08	189	0.86	0.86	0
7	39	-73	4	5.00	81	0.67	0.67	0
8	30	-73	-8	4.96	189	0.71	0.71	0
9	-63	-34	16	4.95	702	0.88	0.92	0
10	-24	-61	-11	4.91	162	0.67	0.67	0
11	-63	-10	13	4.90	108	0.5	0.5	0
12	-27	-19	4	4.85	81	0.67	0.67	0
13	36	-94	-8	4.75	432	0.81	0.88	0
14	-57	-19	7	4.68	108	0.5	0.5	0
15	69	-22	10	4.67	108	0.5	0.5	0
16	30	-46	-11	4.64	270	0.7	0.8	0

Table 3: “Look negative cue” vs “Look negative rating” dataset: comparison between lower bounds on the True Discovery Proportion for the cluster-defining threshold $z = 4.5$.

ID	X	Y	Z	Peak Stat	Size (mm3)	TDP lower bound		
						ARI	Notip	pARI
1	-33	-94	-17	5.63	378	1	1	0
2	66	2	16	5.47	135	1	1	0
3	-12	-82	-8	5.40	135	1	1	0
4	-6	11	52	5.30	81	1	1	0
5	45	14	25	5.27	216	1	1	0
6	12	-43	-26	5.08	27	1	1	0
7	39	-73	4	5.00	27	1	1	0

Table 4: “Look negative cue” vs “Look negative rating” dataset: comparison between lower bounds on the True Discovery Proportion for the cluster-defining threshold $z = 5$.