Imaging Neuroscience
an open access journal

Comment

Selective inference for fMRI cluster-wise analysis, issues, and recommendations for critical vector selection: A comment on Blain et al.

Angela Andreella[a], Anna Vesely[b], Wouter Weeda[c], Jelle Goeman[d]

[a]Department of Economics, Ca' Foscari University of Venice, Venice, Italy
[b]Department of Statistical Sciences, University of Bologna, Bologna, Italy
[c]Department of Psychology, Leiden University, Leiden, The Netherlands
[d]Department of Biomedical Data Sciences, Leiden University Medical Center, Leiden, The Netherlands

Corresponding Author: Angela Andreella (angela.andreella@unive.it)

ABSTRACT

Two permutation-based methods for simultaneous inference on the proportion of active voxels in cluster-wise brain imaging analysis have recently been published: Notip and pARI. Both rely on the definition of a critical vector of ordered *p*-values, chosen from a family of candidate vectors, but differ in how the family is defined: computed from randomization of external data for Notip and determined a priori for pARI. These procedures were compared to other proposals in the literature, but an extensive comparison between the two methods is missing due to their parallel publication. We provide such a comparison and find that pARI outperforms Notip if both methods are applied under their recommended settings. However, each method carries different advantages and drawbacks.

Keywords: fMRI cluster analysis, brain mapping, multiple testing, permutation test, selective inference, true discovery proportion

1. INTRODUCTION

Cluster-extent-based thresholding is a common approach in functional Magnetic Resonance Imaging (fMRI) analysis to explore which parts of the human brain are activated under some stimuli of interest. This approach permits controlling the Type I error at the level of clusters of adjacent voxels, gaining power with respect to voxel-wise inference approaches by exploiting the intrinsic spatial structure of fMRI data (Nichols & Hayasaka, 2003).

However, the method is affected by the so-called spatial specificity paradox. This paradox arises because the larger the identified cluster, the less information we obtain from classic cluster inference about the signal within it. Indeed, the method tests the null hypothesis that none of the voxels in the cluster are active. Rejecting this null hypothesis only allows to claim the presence of at least one active voxel within the cluster. Consequently, larger clusters provide less information about the number and spatial location of active voxels (Woo et al., 2014). Moreover, conducting follow-up inference within the cluster, or "drilling down," introduces a "double-dipping" problem and leads to an inflated Type I error rate (Kriegeskorte et al., 2009).

The spatial specificity paradox can be resolved by making post-hoc inference on the True Discovery Proportion (TDP), that is, the proportion of false null hypotheses

Received: 23 June 2023 (originally at *Neuroimage*); 2 September 2023 (transfer to *Imaging Neuroscience*) Revision: 21 May 2024 Accepted: 23 May 2024
Available Online: 7 June 2024

Imaging Neuroscience, Volume 2, 2024
https://doi.org/10.1162/imag_a_00198

within a subset. In neuroimaging, post-hoc TDP inference procedures provide lower confidence bounds on the proportion of active voxels within clusters, simultaneously over all possible clusters of interest. The simultaneity characteristic of the confidence bounds makes them valid even under post-hoc selection, allowing for follow-up inference within the cluster, unlike the cluster-extent-based thresholding approach (Goeman et al., 2023; Rosenblatt et al., 2018).

The first approach that proposed simultaneous inference on TDP in the fMRI context is the "All-Resolution Inference" (ARI) method developed by Rosenblatt et al. (2018). However, ARI is parametric and can have low power in some scenarios, especially if correlated data such as fMRI are analyzed. It is well known that statistical analyses based on the permutation theory are superior in terms of power and underlying assumptions in fMRI data analysis since they adapt to the correlation structure of the p-values (Helwig, 2019; Winkler et al., 2014). Permutation-based approaches to compute lower bounds for the TDP were first proposed by Meinshausen (2006) and Hemerik et al. (2019). However, these methods analyze only clusters consisting of the smallest k p-values. The SansSouci method of Blanchard et al. (2020) extended this type of permutation-based simultaneous confidence bounds for the TDP to have the same flexibility as ARI, that is, for clusters defined in different ways, even post-hoc, as many times as the researcher wants. An alternative permutation-based TDP method was proposed by Vesely et al. (2023).

Two recent approaches have appeared in the literature to compute a lower bound for the TDP: Notip by Blain et al. (2022) and pARI by Andreella et al. (2023). Both methods build upon the work of Blanchard et al. (2020), each proposing a different specific permutation-based TDP approach tailored to neuroimaging applications. In the work by Blain et al. (2022), the authors compare their methods with ARI and SansSouci; the gain in power and reliability of permutation-based approaches over parametric methods is apparent. However, due to the parallel publication process, Notip and pARI have not yet been compared to each other. Blain et al. (2022) have made a comparison with pARI, but the settings of the method used in the study were not those recommended by Andreella et al. (2023). Therefore, a proper comparative analysis is still lacking. In this manuscript, we provide such an analysis.

The paper is organized as follows. Section 2 briefly revisits inference on the TDP. Subsection 2.1 gives a general formulation of the permutation methods cited above (i.e., SansSouci, pARI, and Notip) before describing in detail the similarities and dissimilarities between Notip and pARI in Subsection 2.2. Finally, Section 3 revisits the analyses presented in Blain et al. (2022), comparing them to pARI as defined in Andreella et al. (2023). In this comparison, we follow Blain et al. (2022) exactly in terms of the choice of the datasets and evaluation criteria. We show that we replicate the results shown in Blain et al. (2022) regarding Notip, then add the pARI method under the specifications recommended by Andreella et al. (2023). By following exactly the analysis choices made in the Notip paper, we make sure not to favor the pARI method, with which we are more familiar.

2. CONTROLLING TRUE DISCOVERY PROPORTIONS

Consider the brain $B = \{1,\dots,m\} \subset \mathbb{N}$ composed of m voxels and, for each voxel $i \in B$, a p-value p_i corresponding to the null hypothesis that it is not active under the condition of interest. We define by $A \subseteq B$ the unknown set of truly active voxels and by $S \subseteq B$ a generic non-empty subset of hypotheses of interest (i.e., a cluster of voxels). For any choice of S, interest lies in the number of true discoveries $a(S) = |A \cap S|$ or, equivalently, the TDP $|A \cap S| / |S|$, where $|S|$ stands for the cardinality of the set S. For a chosen error rate $\alpha \in (0,1)$, TDP procedures aim to construct lower $(1-\alpha)$-confidence bounds for these quantities, simultaneously over all possible choices of S. The confidence bounds for the number of true discoveries, denoted by $\bar{a}(S)$, are such that

$$\Pr\left(\bar{a}(S) \le a(S)\right) \ge 1 - \alpha \qquad (1)$$

for all $S \subseteq B$. An analogous formulation holds for the confidence bounds for the TDP, which can be immediately derived from $\bar{a}(S)$ (Goeman & Solari, 2011).

The simultaneity of the confidence bounds makes them valid even under post-hoc selection and so allows the user to decide which sets of hypotheses S to analyze in a flexible and post-hoc manner. Therefore, methods with this property give information on the amount of true signal inside any set of voxels. The collection of voxels can be defined in various ways, allowing researchers to choose the method that suits their needs. Examples include clusters based on a searchlight, anatomical regions of interest (ROIs), functional ROIs, and data-driven regions (e.g., cluster-extent-based thresholding). Users can drill down into a region multiple times to more precisely identify the location of true active voxels by applying any region selection rule, whether data-driven or not.

2.1. TDP based on critical vectors and permutations

To bound the TDP, pARI and Notip, like ARI and SansSouci, use a strategy based on critical vectors for ordered p-values. They compute the simultaneous lower $(1-\alpha)$-confidence bound for the number of true discoveries in a cluster S as

$$\bar{a}(S) = \max_{1 \le u \le |S|} 1 - u + \left|\left\{ i \in S : p_i \le \ell_u \right\}\right| \tag{2}$$

where $\ell = (\ell_1, \ldots, \ell_m) \in [0,1]^m$ is a suitable non-decreasing vector called critical vector, or in some cases template (Blain et al., 2022; Blanchard et al., 2020). Different critical vectors have been proposed, but in order to obtain valid simultaneous confidence bounds as in Equation (1), it must satisfy the following condition:

$$\Pr\left(\bigcap_{i=1}^{|N|} \left\{ q_{(i)} \ge \ell_i \right\} \right) \ge 1 - \alpha, \tag{3}$$

where $N = B \setminus A$ is the unknown set of inactive voxels, and $q_{(1)} \le \ldots \le q_{(|N|)}$ are their sorted p-values. This means that the curve of the sorted p-values corresponding to inactive voxels should lie completely above the critical vector with probability at least $1-\alpha$.

In Figure 1, we give a graphical intuition of the computation of $\bar{a}(S)$, as defined in Equation (2). The solid black line is the curve of the sorted p-values in the cluster S of interest; the dashed red and dotted blue lines are two critical vectors (of pARI and Notip, respectively). If there were no signal in S, the black curve would be completely to the left of (i.e., above) each critical vector with probability $1-\alpha$. As it happens, the curve is way to the right of (i.e., below) the critical vector, indicating the presence of much signal. The lower bound $\bar{a}(S)$ to the number of active voxels, according to (2), is given as the maximal horizontal distance between the curve and the critical vector. It is clear from the figure that the shape of the critical vector is crucial and that different critical vectors may give very different TDP values.

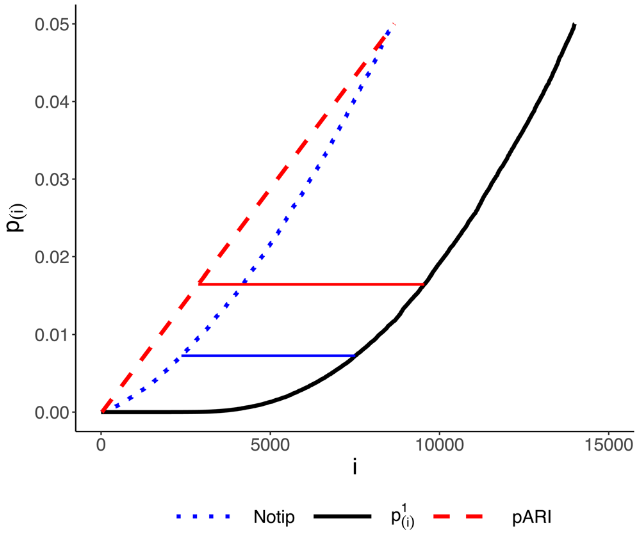

Fig. 1. Graphical intuition of Equation (2). The black solid line represents the vector of sorted observed p-values $p_{(1)} \le \ldots \le p_{(m)}$. For each method (red for pARI, blue for Notip), the broken line represents the resulting critical vector; then, $\bar{a}(S)$ is computed as the length of the solid segment, which is the largest distance between the curve of the observed p-values and the critical vector.

To construct a critical vector that satisfies Equation (3), both Notip and pARI rely on a high number w of transformations of the data, $w-1$ of which can be random permutations or sign-flipping transformations or any other random data transformations that preserve the distribution of the test statistics under the null hypothesis (Winkler et al., 2014), while the remaining one must be the original, untransformed data (Hemerik & Goeman, 2018). The p-value curves arising from $w = 40$ such data transformations are illustrated in Figure 2, with each thin grey curve a p-value curve for a permutation. To find the critical vector, a pre-specified set of candidate critical vectors $\ell(\lambda) = (\ell_1(\lambda), \ldots, \ell_m(\lambda))$, $\lambda \in \Lambda$, is chosen, such that each ℓ_i is non-decreasing in λ. These candidate critical vectors are illustrated as the dashed red lines in Figure 2. In order to satisfy Equation (3), the final critical vector is chosen as the highest curve such that $(1-\alpha)100\%$ of the sorted p-value curves lie above it. That is, if $p^j_{(1)} \le \ldots \le p^j_{(m)}$ are the sorted p-values obtained for the j-th random permutation, then λ is chosen as the largest value such that

$$\left|\left\{ j : p^j_{(1)} > \ell_1(\lambda), \ldots, p^j_{(m)} > \ell_1(\lambda) \right\}\right| \ge (1-\alpha)w. \tag{4}$$

The resulting critical curve is given as the thick red line in Figure 2.

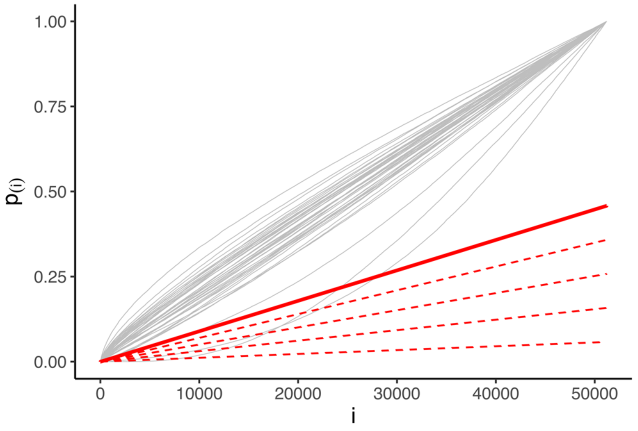

Fig. 2. λ-calibration step: the grey lines represent the vector of sorted p-values given by a random permutation of the data randomly sampling 40 permutations. The red dashed lines are the candidate critical vectors for pARI having different λ values. The solid red line is the optimal pARI critical vector having the largest λ across the ones that cross the null distribution of the p-values represented by the grey lines at most α% of the times.

This permutation-based process allows the method to incorporate the unknown spatial correlation structure of voxels in the calibration of the critical vector, and so to gain power compared to parametric methods.

2.2. Differences between pARI and Notip

The construction just described is common to pARI and Notip. However, pARI and Notip differ in their definition of the set of candidate vectors from which the optimal critical vector is selected, which we call a family of critical vectors (also called, in some cases, a set of learned templates as in Blain et al. (2022) and Blanchard et al. (2020)).

For neuroimaging data, Andreella et al. (2023) recommend the shifted Simes family, given by

$$\ell_i(\lambda) = \frac{(i-\delta)\lambda}{m-\delta} \tag{5}$$

where $\delta \in \{0,1,\ldots,m-1\}$, a shift parameter, is a fixed value that must be chosen independently of the data. The SansSouci approach used the same Simes-based family defined in Equation (5) with $\delta = 0$. Choosing δ larger has the result of losing all power for clusters S of size δ or less, but in a trade-off, this results in substantially higher power for larger clusters. Andreella et al. (2023), therefore, recommended $\delta > 0$ in general, following Hemerik et al. (2019), and substantially larger than 1 if interest is in large clusters. However, δ is not allowed to depend on the sizes of clusters found, so a sensible default must be fixed. They recommended $\delta = 3^3 = 27$ when interest is on clusters of large size, as is common in neuroimaging, so we take this as pARI's default value.

Blain et al. (2022), in contrast, define the family using $\tilde{w}$ permutations on external data with $\tilde{m} \approx m$ voxels. Let $\tilde{p}_{(1)}^{j} \le \ldots \le \tilde{p}_{(\tilde{m})}^{j}$ be the sorted vector of p-values for the j-th permutation of the external data. In the family of candidate critical vectors proposed by Blain et al. (2022), $\ell_i(\lambda)$ is the λ-quantile of the vector $\left(\tilde{p}_{(i)}^{1},\ldots,\tilde{p}_{(i)}^{\tilde{w}}\right)$ if $i \le k_{max}$, and $\ell_i(\lambda) = 1$ otherwise, where $k_{\max} \in \{1,\ldots,m\}$ is some fixed bound chosen a priori. Formally,

$$\ell_i(\lambda) = \begin{cases} \tilde{p}_{(i)}^{(\lfloor \lambda \tilde{w} \rfloor)} & i \le k_{max} \\ 1 & \text{otherwise,} \end{cases} \tag{6}$$

where $\tilde{p}_{(i)}^{(j)}$ denotes the j-th smallest value among $\tilde{p}_{(i)}^{1},\ldots,\tilde{p}_{(i)}^{\tilde{w}}$.

Though seemingly similar in their use of permuted data, Equation (6) is markedly different from (4) above since (6) uses only the marginal distribution of the ordered p-values, whereas (4) uses their joint distribution. The relationship between the external data and the data under analysis should, therefore, not be seen as the usual relationship between a training and a validation set. In fact, Meinshausen (2006) proposed using the same data in (4) and (6), and though Hemerik et al. (2019) and Blanchard et al. (2020) pointed out that doing so destroys the formal validity of the method, the choice of Meinshausen (2006) is generally fine in practice.

In Notip, $k_{\max}$ is a tuning parameter, compable to δ in pARI, and like $\delta > 0$, use of $k_{max} < m$ was recommended for a different family by Hemerik et al. (2019). Effectively, all p-values higher than the k_{max}-th one are ignored by Notip. Like δ, the choice of $k_{\max}$ induces a trade-off: small values can lead to a less conservative family of critical vectors but also to smaller lower bounds for the TDP. Blain et al. (2022) describe $k_{\max}$ as the largest size of the cluster for which a high proportion of active voxels is guaranteed. They suggested to fix $k_{\max} = 1{,}000$.

As a further improvement, Andreella et al. (2023) proposed a step-down version of pARI, which outperforms the SansSouci method in terms of power even if the same critical vector family is used. This improvement comes at the price, however, of high computational time. In this paper, we use the faster version of pARI without the step-down.

3. COMPARISON ON NEUROVAULT DATA

In this section, we compare the Notip and pARI approaches, following exactly the analysis performed originally by Blain et al. (2022). The comparison between pARI and Notip methods primarily emphasizes power, as error control has been previously established in the respective papers (Andreella et al., 2023; Blain et al., 2022). The Neurovault database (Varoquaux et al., 2018) contains data from many fMRI studies. Here, we analyzed collection 1952 (http://neurovault.org/collections/1952), consisting of statistical maps from 20 different studies. First, the images were preprocessed following the procedure outlined in Varoquaux et al. (2018) (i.e., spatial normalization to MNI space using SPM12 software, resampled to a 3 mm isotropic resolution). Then, the data were preprocessed using the Python code made available by Blain et al. (2022) at https://github.com/alexblnn/Notip, resulting in 36 contrast pairs. Specifically, we analyzed elementary "versus baseline," and control contrasts from collection 1952, containing data from a large number of different cognitive tasks (e.g., visual, auditory). For a complete overview of the contrasts analyzed, please refer to Table 6 in Blain et al. (2022).

The analysis was carried out using the pARI R package (https://CRAN.R-project.org/package=pARI) for applying pARI, and the Python code made available by Blain et al. (2022) for applying Notip. Figures 1 and 2, above, have been computed using the first dataset of this collection, that is, “shapes versus baseline” contrast versus “faces versus baseline” contrast from the HCP study. To make Figure 1 clearer, we considered the cluster composed of the smallest 15,000 voxels.

Here, we redo only those analyses from Blain et al. (2022) in which they compare performance between the Notip and competing methods. It is not straightforward to compare different TDP methods because each method gives 2^m TDP confidence bounds. A method that performs better for some TDP bounds may be worse for other bounds, even within the same data or simulation scenario. We follow Blain et al. (2022) in their choice of metric for comparing methods, which focuses on the size of the largest cluster found at a fixed TDP threshold. Other metrics are possible; for example, Andreella et al. (2023) used the TDP of clusters defined at a fixed cluster-defining threshold as their metric. In all the analysis, we fix the number of permutations used to compute the Notip critical vector $\tilde{w}$ to 10,000, and the number of permutations used to calculate the null distribution of the p-values to 1,000.

The left-hand side of Figure 3 reproduces the results of Blain et al. (2022; Figure 4, right-hand side), in which they compare Notip to pARI with $\delta = 0$, that is, to Sans-Souci. The relative number of detections between Notip and pARI, defined as

$$\frac{|S|_{\text{Notip}} - |S|_{\text{pARI}}}{|S|_{\text{pARI}}}, \tag{7}$$

where $|S|$ is the largest possible region that reaches a fixed TDP level, is analyzed. The boxplots presented in Figure 3 show the distribution of this metric over 36 contrasts maps from Neurovault collection 1952 data and TDP thresholds 0.8, 0.9, 0.95 with α fixed at 0.05. The results on the left-hand side of Figure 3 reproduce almost exactly the results presented in Blain et al. (2022). There are minor differences due to the use of random permutations. In addition, we noticed that the code provided by Blain et al. (2022) did not consider the mandatory inclusion of the identity transformation, which we included to get exact α control (Hemerik & Goeman, 2018), even though due to the high number of permutations (i.e., $w = 1{,}000$) this makes almost no difference. The right-hand side of Figure 3 makes the same comparison but with pARI’s recommended setting of $\delta = 27$.

Where Notip almost always outperformed pARI without the shift, we note that the reverse is true for the recommended shifted version of pARI. To investigate further, Figure 4 plots the largest cluster sizes found by pARI ($\delta = 27$) against those found by Notip. Also, from this plot, we see that the size of the largest cluster found is almost always greater with pARI than with Notip, and this effect is especially pronounced when the largest cluster contains many voxels (i.e., top right part of Figure 4).

Finally, Table 1 reproduces results from Table 2 in Blain et al. (2022), to which we added results for pARI with $\delta = 27$. The contrast pair “look negative cue vs look negative rating” of the Neurovault database is analyzed. The clusters are computed by thresholding the statistical map

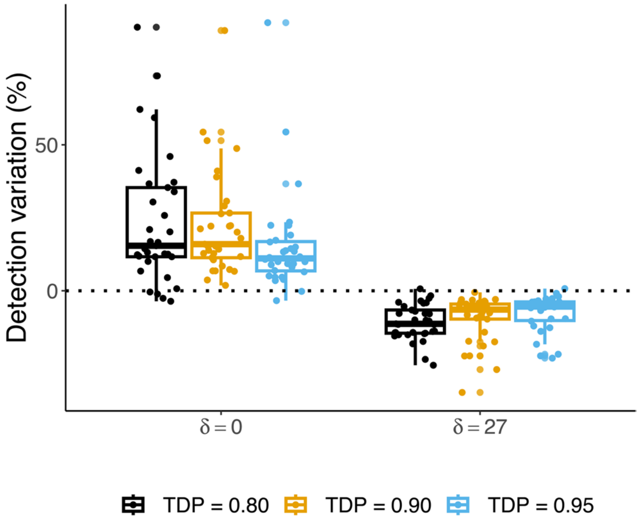

Fig. 3. Percentage variations detected defined as $\frac{|S|_{\text{Notip}} - |S|_{\text{pARI}}}{|S|_{\text{pARI}}}$. The left side is the non-recommended setting for pARI (i.e., fixing $\delta = 0$), which we show only to reproduce the results of Blain et al. (2022). Instead, the right side represents the results using the recommended setting for pARI as shown by Andreella et al. (2023) when $\delta = 27$. Since the comparison is given in terms of variation as defined above, values below 0 indicate better performance in pARI than in Notip.

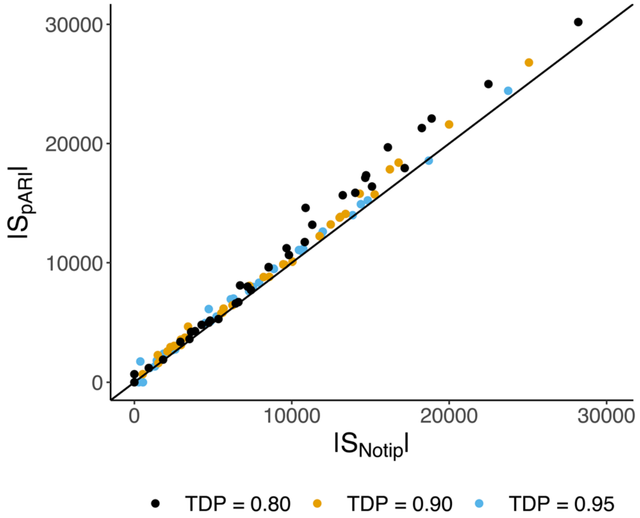

Fig. 4. Size of the largest clusters found by pARI with $\delta = 27$ ($|S_{\text{pARI}}|$) and Notip ($|S_{\text{Notip}}|$) with TDP $\geq t \in \{0.8, 0.9, 0.95\}$.

Table 1. Clusters identified with threshold $|z| > 3$: clusters size and TDP lower bound at risk level $\alpha = 0.05$ using two possible critical vectors (Notip, and Simes-based pARI with $\delta = 27$) on contrast pair "look negative cue vs look negative rating."

		True discovery proportion		
			Simes-based pARI	
Cluster ID	Cluster size	Notip	$\delta = 0$	$\delta = 27$
1	7,695	0.26	0.23	**0.34**
2	14,877	0.45	0.32	**0.58**
3	14,445	0.50	0.37	**0.60**
4	5,238	0.29	0.24	**0.34**
5	4,563	**0.30**	**0.30**	0.29
6	12,555	0.35	0.16	**0.52**
7	6,075	0.17	0.09	**0.24**
8	25,812	0.66	0.46	**0.76**
9	6,507	0.17	0.15	**0.20**

For each cluster, the values in bold indicate the best result, that is, TDP (lower limit) higher.

at absolute values greater than 3 and keeping only clusters composed of at least 150 voxels (Woo et al., 2014). Again, we can note how imposing $\delta = 27$ significantly increases the method's power; pARI is, in fact, more powerful than Notip in all clusters, except the smallest one, that is, it returns greater lower bounds for the TDP.

We can conclude that the shifted version of Simes-based pARI performs remarkably well and, in most cases, surpasses the Notip approach, emphasizing the importance of choosing an appropriate critical vector (and shift value) for gaining power.

Please refer to the online Supplementary Materials for further analysis.

4. DISCUSSION

We have seen that pARI outperformed Notip in almost all settings considered by Blain et al. (2022) when the shift parameter δ of pARI was appropriately set at $\delta = 27$. This finding may seem counterintuitive since Notip uses additional information in the form of external data. It should be realized, however, that in this external data, Notip looks only marginally at the ordered p-values. The added value of this information may be limited in practice, as also illustrated by the experience (Blain et al., 2022; Meinshausen, 2006) that double dipping by reusing the data under analysis as if they were external does not break the validity of the method in practice.

Both Notip and pARI have a tuning parameter (k_{max} and δ, respectively). The presence of an additional parameter can be considered a drawback, especially since it has to be chosen before seeing the data. Both methods, therefore, recommend a default value ($k_{max} = 1{,}000$ and $\delta = 27$) for applications in neuroimaging. It is interesting to note that k_{max} and δ have complementary effects: $k_{max} < m$ focuses power of Notip away from very large clusters, while $\delta > 0$ focuses power of pARI away from small ones. It could be an interesting avenue of further research to formulate an alternative method that has both a k_{max} and a δ parameter (e.g., as considered in a different context by Hemerik et al. (2019)).

It can be argued that Notip has a second tuning parameter in the choice of the external data. This can be avoided by re-use of the data under analysis, but the resulting method has no formal proof of error control. Whether data are reused or not, this additional analysis step makes the procedure more computationally expensive. For the analyses presented here (i.e., considering standard Notip and the single-step version of pARI), Notip takes approximately 42 minutes, while pARI takes only 1 minute. pARI, on the other hand, becomes computationally expensive if the step-down version is used.

Various trade-offs characterize both methods and can be seen as two out of many possible analysis choices. The comparison that we have given here shows that the choice of the family matters, but further analyses are needed to study each method's power properties in more detail and to determine which method should be preferred in which settings. This could also help in finding even better families than those considered by Notip and pARI.

DATA AND CODE AVAILABILITY

The data underlying this study are those used in Blain et al. (2022), available in the NeuroVault database at http://neurovault.org/collections/1952. The code to preprocess the data and apply the Notip method is available at the GitHub repository https://github.com/alexblnn/Notip. The code for the pARI method is developed in the R package pARI, at https://CRAN.R-project.org/package=pARI.

AUTHOR CONTRIBUTIONS

Angela Andreella: conceptualization, methodology, software, data curation, formal analysis, investigation, writing—original draft, and writing—review and editing. Anna Vesely: conceptualization, methodology, formal analysis, and writing—review and editing. Wouter Weeda and Jelle Goeman: conceptualization, methodology, writing—review and editing, and supervision.

ETHICS

This research relies on existing data sources, and no primary data collection was undertaken.

DECLARATION OF COMPETING INTEREST

The authors declare no competing interests.

ACKNOWLEDGMENTS

Angela Andreella gratefully acknowledges financial support from Ca' Foscari University of Venice via Grant No. PON 2014-2020/DM 1062. Anna Vesely acknowledges financial support from the Ministry of University and Research within the Complementary National Plan PNC-I.1 "Research initiatives for innovative technologies and pathways in the health and welfare sector." The authors express gratitude to Jesse Hemerik for valuable discussions related to this work.

SUPPLEMENTARY MATERIALS

Supplementary material for this article is available with the online version here: https://doi.org/10.1162/imag_a_00198.

REFERENCES

Andreella, A., Hemerik, J., Finos, L., Weeda, W., & Goeman, J. (2023). Permutation-based true discovery proportions for functional magnetic resonance imaging cluster analysis. *Statistics in Medicine*, *42*(14), 2311–2340. https://doi.org/10.1002/sim.9725

Blain, A., Thirion, B., & Neuvial, P. (2022). Notip: Non-parametric true discovery proportion control for brain imaging. *NeuroImage*, *260*, 119492. https://doi.org/10.1016/j.neuroimage.2022.119492

Blanchard, G., Neuvial, P., & Roquain, E. (2020). Post hoc confidence bounds on false positives using reference families. *The Annals of Statistics*, *48*, 1281–1303. https://doi.org/10.1214/19-aos1847

Goeman, J. J., Górecki, P., Monajemi, R., Chen, X., Nichols, T. E., & Weeda, W. (2023). Cluster extent inference revisited: Quantification and localisation of brain activity. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *85*(4), 1128–1153. https://doi.org/10.1093/jrsssb/qkad067

Goeman, J. J., & Solari, A. (2011). Multiple testing for exploratory research. *Statistical Science*, *26*(4), 584–597. https://doi.org/10.1214/11-sts356

Helwig, N. E. (2019). Statistical nonparametric mapping: Multivariate permutation tests for location, correlation, and regression problems in neuroimaging. *Wiley Interdisciplinary Reviews: Computational Statistics*, *11*(2), e1457. https://doi.org/10.1002/wics.1457

Hemerik, J., & Goeman, J. (2018). Exact testing with random permutations. *Test*, *27*(4), 811–825. https://doi.org/10.1007/s11749-017-0571-1

Hemerik, J., Solari, A., & Goeman, J. J. (2019). Permutation-based simultaneous confidence bounds for the false discovery proportion. *Biometrika*, *106*(3), 635–649. https://doi.org/10.1093/biomet/asz021

Kriegeskorte, N., Simmons, W. K., Bellgowan, P. S., & Baker, C. I. (2009). Circular analysis in systems neuroscience: The dangers of double dipping. *Nature Neuroscience*, *12*(5), 535–540. https://doi.org/10.1038/nn.2303

Meinshausen, N. (2006). False discovery control for multiple tests of association under general dependence. *Scandinavian Journal of Statistics*, *33*(2), 227–237. https://doi.org/10.1111/j.1467-9469.2005.00488.x

Nichols, T., & Hayasaka, S. (2003). Controlling the familywise error rate in functional neuroimaging: A comparative review. *Statistical Methods in Medical Research*, *12*(5), 419–446. https://doi.org/10.1191/0962280203sm341ra

Rosenblatt, J. D., Finos, L., Weeda, W. D., Solari, A., & Goeman, J. J. (2018). All-resolutions inference for brain imaging. *NeuroImage*, *181*, 786–796. https://doi.org/10.1016/j.neuroimage.2018.07.060

Varoquaux, G., Schwartz, Y., Poldrack, R. A., Gauthier, B., Bzdok, D., Poline, J.-B., & Thirion, B. (2018). Atlases of cognition with large-scale human brain mapping. *PLoS Computational Biology*, *14*(11), e1006565. https://doi.org/10.1371/journal.pcbi.1006565

Vesely, A., Finos, L., & Goeman, J. J. (2023). Permutation-based true discovery guarantee by sum tests. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *85*(3), 664–683. https://doi.org/10.1093/jrsssb/qkad019

Winkler, A. M., Ridgway, G. R., Webster, M. A., Smith, S. M., & Nichols, T. E. (2014). Permutation inference for the general linear model. *NeuroImage*, *92*, 381–397. https://doi.org/10.1016/j.neuroimage.2014.01.060

Woo, C.-W., Krishnan, A., & Wager, T. D. (2014). Cluster-extent based thresholding in fMRI analyses: Pitfalls and recommendations. *NeuroImage*, *91*, 412–419. https://doi.org/10.1016/j.neuroimage.2013.12.058