Journal of the Royal Statistical Society Series B: Statistical Methodology, 2023, **85**, 1128–1153
https://doi.org/10.1093/jrsssb/qkad067
Advance access publication 4 July 2023
Original Article

ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS
Journal of the Statistics Society
Series B
Statistical Methodology
B

Cluster extent inference revisited: quantification and localisation of brain activity

Jelle J. Goeman[1], Paweł Górecki[2], Ramin Monajemi[1], Xu Chen[1], Thomas E. Nichols[3,4] and Wouter Weeda[5]

[1]Biomedical Data Sciences, Leiden University Medical Center, Leiden, The Netherlands
[2]Faculty of Mathematics, Informatics and Mechanics, Institute of Informatics, University of Warsaw, Warsaw, Poland
[3]Nuffield Department of Population Health, Big Data Institute, Li Ka Shing Centre for Health Information and Discovery, University of Oxford, Oxford, UK
[4]Nuffield Department of Clinical Neurosciences, Wellcome Centre for Integrative Neuroimaging, FMRIB, University of Oxford, Oxford, UK
[5]Methodology and Statistics, Psychology, Leiden University, Leiden, The Netherlands

Address for correspondence: Jelle J. Goeman, Biomedical Data Sciences, Leiden University Medical Center, Postbus 9600, 2300 RC Leiden, The Netherlands. Email: j.j.goeman@lumc.nl

Abstract

Cluster inference based on spatial extent thresholding is a popular analysis method multiple testing in spatial data, and is frequently used for finding activated brain areas in neuroimaging. However, the method has several well-known issues. While powerful for finding regions with some activation, the method as currently defined does not allow any further quantification or localisation of signal. In this paper, we repair this gap. We show that cluster-extent inference can be used (1) to infer the presence of signal in any region of interest and (2) to quantify the percentage of activation in such regions. These additional inferences come for free, i.e. they do not require any further adjustment of the alpha-level of tests, while retaining full family-wise error control. We achieve this extension of the possibilities of cluster inference by embedding the method into a closed testing procedure, and solving the graph-theoretic *k*-separator problem that results from this embedding. We demonstrate the usefulness of the improved method in a large-scale application to neuroimaging data from the Neurovault database.

1 Introduction

We consider large-scale multiple testing problems in which the hypotheses are structured in a *d*-dimensional rectangular grid. In such problems, the spatial organisation of the hypotheses can be important, as the signal can often be assumed to be spatially clustered.

Our primary motivation comes from functional magnetic resonance imaging (fMRI) studies, which study brain activation in response to a mental task (Ogawa et al., 1992). In such studies, brain activity is measured for around 200,000 of voxels, three-dimensional equivalents of pixels. This results in per-voxel null hypotheses of no activation that form a regular 3-*d* grid. Signal is expected to cluster, and researchers therefore aim to find regions (‘clusters’) of activation rather than individual active voxels. Regular grids of hypotheses arise in other contexts as well, for example 1-*d* structures in DNA methylation (Jaffe et al., 2012) and 2-*d* structures in climate science (Sommerfeld et al., 2018).

The state-of-the-art solution for inference on regions in fMRI is cluster inference by cluster extent thresholding (Forman et al., 1995; Friston et al., 1994; Nichols, 2012). In brief, starting from

Received: September 30, 2021. **Revised:** February 17, 2023. **Accepted:** June 8, 2023

a test statistic per voxel, this method finds all clusters of contiguous voxels for which the test statistic exceeds a predefined threshold. Clusters are declared significant if their extent (number of voxels) exceeds the extent threshold, which is defined as the $(1-\alpha)$-quantile of the distribution of the maximal extent of such clusters under the global null hypothesis. This extent threshold can be determined either analytically, using the assumption that the z-scores come from a Gaussian random field (Eklund et al., 2016; Friston et al., 1994; Worsley et al., 1996), or more robustly by permutations (Hayasaka & Nichols, 2003).

Cluster inference has strong control of family-wise error rate (FWER) at the level of clusters (Worsley et al., 1992). This means that, regardless of the amount of signal present in the data, with probability at least $1-\alpha$ no cluster null hypothesis is falsely rejected. The cluster null hypothesis is the hypothesis that none of the voxels in the cluster is truly active. The inferential statement that can be made from cluster inference is, therefore, that, with $1-\alpha$ simultaneous confidence, every significant cluster contains at least one active voxel.

However, cluster-level FWER control has been criticised as insufficient to support the conclusions researchers would typically like to draw from neuroimaging experiments. For example, Woo et al. (2014) argued that, especially at low z thresholds, clusters can become too large and span multiple anatomical brain areas, challenging the interpretation of the results. The following three inferential conclusions are often (implicitly or explicitly) drawn from cluster inference result, though they are not supported by the theory (Woo et al., 2014).

1. *'A large significant cluster contains a substantial number of active voxels'*. Cluster-level FWER control only supports the statement that at least one voxel in the cluster is confidently active, not that many, or let alone, all voxels are active.
2. *'A large significant cluster is a more substantial scientific finding than a small significant cluster'*. In fact, the assertion that at least one voxel in a large cluster is active, is a less precise, and therefore weaker finding than the same assertion in a small cluster. This counter-intuitive property is known as the Spatial Specificity Paradox.
3. *'Substantial overlap between a significant cluster and an anatomical brain area indicates evidence for the presence of activity in that anatomical brain area'*. A significant cluster confidently contains at least one active voxel, but unless that cluster is completely contained in the anatomical area, such activity may lie outside the anatomical brain area.

Still, the three unsupported conclusions from cluster inference, sketched above, seem intuitively quite reasonable. If a cluster exceeds the minimal size k for a significant cluster by a large margin, it is natural to suppose that there is a substantial amount of signal in the cluster, and at least more than in another cluster with an extent just over k. If the large cluster largely overlaps with an anatomical region, it is reasonable to suppose that some of the signal in the cluster must be in the anatomical region.

This paper strengthens cluster inference by presenting an improvement of the method that allows much stronger and more informative conclusions to be drawn, avoiding the problems sketched above. We construct this improvement of cluster inference by remarking that cluster inference is a special case of a true discovery guarantee method, as defined by Goeman et al. (2021). This suggests a possibility for uniform improvement of the method, by embedding it into a closed testing procedure, which we will construct.

Rather than returning a p-value for each supra-threshold cluster, the new method returns a *true discovery proportion* (TDP) for every region, a simultaneous lower confidence bound for the proportion of truly active voxels in the region (Genovese & Wasserman, 2006; Goeman & Solari, 2011). By quantifying how widely spread a signal is within a brain region, TDP-based inference avoids the spatial specificity paradox (Rosenblatt et al., 2018). Moreover, TDP can be calculated for any brain region, not just for supra-threshold clusters; this way also the amount of signal in anatomical regions may be assessed. Being a uniform improvement of classic cluster inference, there is no power loss when switching from classic cluster-based inference to the method proposed in this paper; the new method will always yield TDP >0 for any cluster that is significant according to classic cluster-based inference. The new method retains strict FWER control over all reported findings: with probability at least $1-\alpha$ no reported TDP is greater than the proportion of truly active voxels in the corresponding region.

There is a substantial literature on methods providing simultaneous lower bounds for TDP, or equivalently upper bounds for false discovery proportions (Blanchard et al., 2020; Goeman et al., 2019; Katsevich & Ramdas, 2020), which have all been shown to be special cases of closed testing procedures (Goeman et al., 2021). TDP methods have also been applied to neuroimaging data before (Andreella et al., 2023; Blain et al., 2022; Rosenblatt et al., 2018; Vesely et al., in press). However, all methods proposed thus far treat the indices of hypotheses as exchangeable, neglecting their spatial arrangement. In contrast, we build a closed testing procedure using local tests that take the spatial structure into account. This results in a non-exchangeable closed testing procedure with very different mathematical properties. In particular, an unexpected and interesting connection arises between graph theory and multiple testing procedures, as we shall see below.

A major challenge of constructing closed testing procedures is computational. We will show that calculating TDP for a brain region amounts to solving an instance of a graph-theoretic k-separator problem (Ben-Ameur et al., 2015). We propose two novel and fast algorithms to solve the k-separator problem in the lattice graph induced by brain connectivity, in order to find shortcuts for the closed testing procedure.

To illustrate the performance of the method, we will apply the novel lower bound on 818 data sets from the Neurovault database (Gorgolewski et al., 2015). We will first illustrate the intended workflow of the new method using an n-back working memory data set (Barch et al., 2013), which we will introduce in the next section as a motivating example.

Throughout the paper, we use the notational convention that, except for the probability distribution P, all capitals are sets and all lower case variables are scalars or vectors. Random variables are in boldface. The proofs of all lemmas and theorems are in the online supplementary material, Section A.

2 Motivating example

We will first illustrate and preview the new method with a concrete motivating example, which is typical for fMRI cluster inference. Although the details are specific for the neuroimaging context, the example serves to illustrate the paradoxes researchers run into when inferring on spatial data in terms of clusters.

The Human Connectome Project (HCP; Van Essen et al., 2013) consists of neuroimaging data of over 5,000 subjects performing multiple cognitive tasks, of which brain activity is measured by the proxy of changes in blood oxygenation levels (BOLD). In our example, we will use fMRI data obtained from 80 unrelated individuals, each performing an n-back working memory task (Barch et al., 2013). During this task participants are sequentially shown a series of letters (e.g. ‘E’, ‘D’, ‘Z’, ‘X’, ‘M’). After the sequence is shown, participants are asked to recall letters from a specific position in the sequence. For example, in the 0-back condition this is the last letter shown (‘M’), in the 2-back condition this is the letter in second-to-last position (‘Z’). In n-back tasks, higher values of n are theoretically associated with larger memory load for the participants. We focussed on the 2-back vs. 0-back contrast, for which the null hypothesis of interest per voxel was that the BOLD signal was identically distributed between the 2-back and 0-back conditions. For the calculation of per-voxel test statistics, we followed a standard processing pipeline (Glasser et al., 2013) using FSL, a popular software package for cluster-extent inference (Woolrich et al., 2001). This is a two-stage analysis, in which the 2-back vs. 0-back contrast is first analysed for each subject separately and the results are subsequently aggregated across subjects into a group-level z-statistic for each of the 257,659 voxels in the brain, using standard methods described by Beckmann et al. (2003). Each of these z-statistics is standard normal under their respective per-voxel null hypothesis.

Before seeing the data, a cluster-forming threshold of $z = 3.1$ was chosen. Clusters were formed by all connected neighbouring supra-threshold voxels. Using standard theory, which we will revisit in Section 3, a permutation-based extent threshold of 72 was found, indicating that all clusters consisting of more than 72 voxels are significant. This led to six significant clusters and several nonsignificant clusters. The details of the significant clusters are shown in Figure 1 and Table 1.

With classic cluster inference, the analysis ends here. The researchers may claim that some signal is present in each significant cluster, but the amount of signal is undetermined. This is especially tantalising for the biggest cluster A, that visually consists of several sub-regions. No statement can be made about the presence of signal in these sub-clusters. Cluster C overlaps for a large

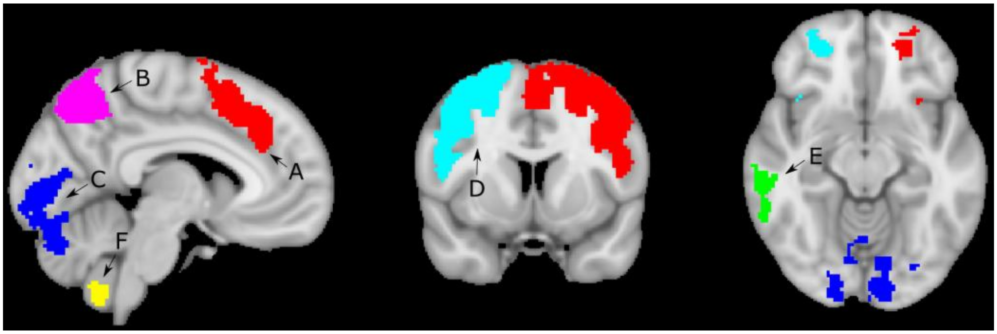

Figure 1. Task-related brain activation for the 2-back vs. 0-back contrast across all subjects. Six significant clusters A, B, C, D, E, F are indicated.

part with the cerebellum, but since it is not fully contained in the cerebellum, the researcher may not confidently claim the presence of signal here from the overlap with Cluster C. In contrast, Cluster F, which is relatively small and would not attract the most attention in the publication, does substantiate a claim about the presence of signal in the cerebellum since it is completely contained in it. Paradoxically, Cluster F is the most precise finding, since it localises the presence of signal to a precision of no more than 100 voxels.

The theory developed in this paper will allow much more informative statements to be made about clusters A to F.

1. We calculate TDP per cluster, a lower bound to the number of truly active voxels. Clusters A, B, C, D, E, F get TDPs of 37%, 40%, 33%, 37%, 19%, and 10%, respectively. This indicates that clusters A to D are the main findings of the experiment, but shows that the localisation of the signal is only moderately precise.
2. We also find TDPs for any other (anatomical) brain regions of interest. We find, for example, significant evidence of signal in cerebellum, mostly from the overlap with cluster C, though with a small TDP of 5.8%.

The TDP values we find are guaranteed to be consistent with the cluster p-values in the sense that $p \leq 0.05$ if and only if TDP is positive. Compared to the p-values, the TDP is more informative since it quantifies the pervasiveness of the signal within the cluster. The full analysis results are given in Section 9.

3 Classic cluster inference

We give a short recap of classic cluster inference to set the scene and introduce notation.

The brain is partitioned into hundreds of thousands of voxels, forming a rectangular grid in $\mathbb{Z}^d$. We will usually think of $d = 3$, but we will write our theory for general $d \geq 1$. The brain $B \subset \mathbb{Z}^d$ is an irregularly shaped, finite collection of voxels. It is not always the entire brain that is of interest to the researcher, and a mask $M \subseteq B$ is chosen, before seeing the data, limiting all inference to voxels in M.

We define a neighbour relationship between voxels, saying that voxels $v, w \in \mathbb{Z}^d$ are neighbours if $v - w \in \{-1, 0, 1\}^d$. This neighbourhood definition is known as 26-connectivity in neuroimaging since it gives each voxel 27 neighbours (26 plus itself) if $d = 3$. The voxels and the neighbour relation together induce an undirected graph when the voxels are seen as nodes and the neighbour relationships as edges. We call a voxel set $V \subseteq \mathbb{Z}^d$ a *cluster* if its induced graph is connected. We call voxel sets V and W *disconnected* if no voxel of V is a neighbour of a voxel of W.

Let Ω be our statistical model and $\mathrm{P} \in \Omega$ the unknown probability distribution of the data. For each voxel $v \in B$ we define a voxel-wise null hypothesis $H_v \subseteq \Omega$ stating that the voxel v is not active. Note that in general a hypothesis H is true if and only if $\mathrm{P} \in H$. Researchers are usually not particularly interested in individual voxels, since these are considered too small to represent relevant brain processes. Instead, researchers look at clusters of neighbouring voxels. For every voxel

Table 1. Task-related brain activation for the 2-back vs. 0-back contrast across all subjects

Cluster	Size	p-Value	max (z)	X	Y	Z
A	8,870	<0.001	8.87	44	72	60
B	8,526	<0.001	9.51	19	42	61
C	7,956	<0.001	9.20	63	33	20
D	6,652	<0.001	9.73	31	67	64
E	350	0.004	5.18	15	46	28
F	100	0.027	6.56	49	35	10

Note. Columns show the size, p-value, maximum z-statistic, and coordinates of the maximum for all clusters.

set $V \subseteq M$, we define the voxel set null hypothesis as $H_V = \bigcap_{v \in V} H_v$. This hypothesis states that all of the voxel-wise null hypotheses for voxels in V are true, i.e. that none of the voxels in V are active. The hypothesis $H_\emptyset = \Omega$ is always true.

We assume that we have z-score test statistics $(\mathbf{z}_v)_{v \in B}$ for every voxel null hypothesis. The z-score $\mathbf{z}_v$ is expected to be small in absolute value if H_v is true and large if H_v is false. We make no assumptions about the marginal or joint distribution of the z-scores at this point. Cluster inference uses these voxel z-scores to make inference at the cluster level. First, before seeing the data the researcher selects a z-score cut-off z. Next, the researcher finds the set of all supra-threshold voxels in the mask, $M \cap \mathbf{Z}$, where

$$\mathbf{Z} = \{v \in B : \mathbf{z}_v > z\} \tag{1}$$

is the collection of all supra-threshold voxels. Equation (1) uses one-sided tests. Two-sided tests can be done either using $|\mathbf{z}_v| > z$ in equation (1) or by repeating the analysis twice: once with $\mathbf{z}_v$ and once with $-\mathbf{z}_v$, using half the α-level.

The supra-threshold voxel set $\mathbf{Z} \cap M$ is not in general a cluster, but it is always a union of clusters. We can uniquely write $\mathbf{Z} \cap M = \mathbf{C}_1 \cup \cdots \cup \mathbf{C_n}$, where $\mathbf{C}_1, \ldots, \mathbf{C_n}$ are disconnected clusters. Cluster inference now claims the presence of signal in every $\mathbf{C}_i$ for which $|\mathbf{C}_i| > k_M$, where $|\cdot|$ is the cardinality of a set, and k_M is the cluster-extent threshold for mask M. The cluster-extent threshold is defined as the $(1-\alpha)$-quantile of the maximum size of a supra-threshold cluster under the global null. Formally, the size of the largest supra-threshold voxel is $\chi_{M \cap \mathbf{Z}}$, where

$$\chi_V = \max\{|C| : C \subseteq V \text{ is a cluster}\}.$$

This maximum is always defined since the empty set is a cluster. The cluster-extent threshold k_M therefore has the property that, for every $\mathrm{P} \in H_M$,

$$\mathrm{P}(\chi_{M \cap \mathbf{Z}} > k_M) \leq \alpha. \tag{2}$$

We remark that k_M is allowed to be random, as it would be, e.g. in permutation approaches. We also remark that we deviate slightly from the usual definition of k_M, which uses $\geq$ in the first inequality in equation (2).

To achieve equation (2), it is important to take both the marginal and the joint distribution of the z-scores into account. Various parametric and semi-parametric methods have been proposed. Friston et al. (1994) assume that $(z_v)_{v \in M}$ follows a stationary Gaussian random field on M, and that each H_v, $v \in B$, is the hypothesis that z_v has zero mean. In this case, k_M can be approximated using the expected Euler characteristic of the field, and equation (2) holds as long as z is large enough and the field is sufficiently smooth (Eklund et al., 2016; Worsley et al., 1996). Alternatively, a k_M achieving equation (2) may be calculated from other assumptions, e.g. t-fields, χ^2-fields, or F-fields (Worsley et al., 1996). Permutation-based approaches (Hayasaka & Nichols, 2003) can efficiently accommodate the joint distribution of the z-scores. In the rest of the paper, we will not use any specific set of distributional assumptions. We will simply assume k_M can be calculated for every $M \subseteq B$ such that equation (2) holds.

Larger masks allow larger supra-threshold clusters, and therefore larger cluster-extent thresholds. We will assume that if $M \subseteq N$, then,

$$k_M \leq k_N. \tag{3}$$

This relationship is natural since $\chi_{M \cap Z} \leq \chi_{N \cap Z}$, surely. We may take $k_\emptyset = 0$ without loss of generality.

4 Closed testing for cluster inference

Having described classic cluster inference, we can now construct its embedding into a closed testing procedure. We will exploit theory of Goeman et al. (2021), who provide a general method to construct a closed testing procedure from an existing multiple testing procedure.

4.1 Local test and effective local test

A closed testing procedure is built from local tests, which are hypothesis tests for a voxel set null hypothesis H_V. We will define such a local test for every voxel set $V \subseteq M$ as the test that rejects when cluster inference with mask $M = V$ rejects at least one voxel set null hypothesis. The resulting test rejects when $\phi_V = 1$, where

$$\phi_V = \mathbb{1}\{\chi_{V \cap Z} > k_V\}. \tag{4}$$

This is a valid local test due to the assumption that equation (2) holds for every $M \subseteq B$, and therefore for $M = V$: we have for every $\mathrm{P} \in H_V$ that $\mathrm{P}(\phi_V = 1) \leq \alpha$. If $V = \emptyset$, then $\phi_V = 0$, so the test never rejects. We will use the local test (4) for every $V \subseteq M$ as the building block for the new closed testing procedure.

The local test ϕ_V is a valid hypothesis test for the presence of signal in V if the researcher restricted attention to V before seeing the data. If the researcher chooses $V \subseteq M$ after seeing the data, a multiple testing correction needs to be performed over all $2^{|M|}$ hypothesis choices $(H_V)_{V \subseteq M}$. This is what closed testing does.

Marcus et al. (1976) proved that such correction for multiple testing can be achieved by the effective local test, defined for any local test as

$$\psi_V = \min\{\phi_W : V \subseteq W \subseteq M\}.$$

The effective local test controls voxel set-level FWER over all $(H_V)_{V \subseteq M}$, having the property that for every $\mathrm{P} \in \Omega$,

$$\mathrm{P}(\psi_V = 0 \text{ for all } V \subseteq M \text{ with } \mathrm{P} \in H_V) \geq 1 - \alpha. \tag{5}$$

Remembering that $\mathrm{P} \in H_V$ if and only if H_V is true, we see that with probability at least $1 - \alpha$ no true voxel set null hypothesis is rejected even when ψ_V is applied on all $V \subseteq M$.

4.2 Shortcut

However, ψ_V is difficult to calculate, since it involves calculating ϕ_W, and therefore k_W, for exponentially many $V \subseteq W \subseteq M$. We propose to approximate ψ_V for every $V \subseteq M$ by an alternative test that is easier to compute:

$$\underline{\psi}_V = \mathbb{1}\{\chi_{V \cap Z} > k_M\}.$$

For every $V \subseteq M$, the test $\underline{\psi}_V$ rejects at most as often as ψ_V, as Lemma 1 states.

Lemma 1 For every $V \subseteq M$, we have $\underline{\psi}_V \leq \psi_V$.

The alternative test $\underline{\boldsymbol{\psi}}_V$ is a shortcut for the effective local test $\boldsymbol{\psi}_V$: it sacrifices some power for ease of computation. By Lemma 1, $\underline{\boldsymbol{\psi}}_V$ retains the error guarentees of $\boldsymbol{\psi}_V$. Combining the lemma with equation (5), we obtain voxel set-level FWER for $\underline{\boldsymbol{\psi}}_V$. For every $\mathrm{P} \in \Omega$,

$$\mathrm{P}(\underline{\boldsymbol{\psi}}_V = 0 \text{ for all } V \subseteq M \text{ with } \mathrm{P} \in H_V) \geq 1 - \alpha.$$

We can check that the test $\underline{\boldsymbol{\psi}}_V$ reproduces the FWER guarantee of classic cluster inference. Classic cluster inference rejects all clusters $\mathbf{C} \subseteq M \cap \mathbf{Z}$ with $|\mathbf{C}| > k_M$. For such $\mathbf{C}$, we have $\chi_{\mathbf{C}\cap\mathbf{Z}} = \chi_{\mathbf{C}} = |\mathbf{C}| > k_M$, so that $\underline{\boldsymbol{\psi}}_{\mathbf{C}} = 1$.

However, $\underline{\boldsymbol{\psi}}_V$ allows useful additional conclusions that are not endorsed by classic cluster inference. If $A \subseteq B$ is an anatomical region of interest, we may reject H_A and claim the presence of activity in A if $\chi_{A\cap\mathbf{Z}} > k_M$, that is when there are at least k_M connected supra-threshold voxels within A. This provides a partial solution to the desired inference problem 3 in the *Introduction* to this paper, since it defines precisely how large a 'substantial overlap' between a significant cluster and an anatomical region must be to allow a claim of activity in the region: the overlap must contain a connected area of size at least k_M. Note that the region of interest A does not have to be chosen before seeing the data for such inference to be valid, since FWER control is over all $V \subseteq M$.

4.3 TDPs from closed testing

The major gain of the closed testing formulation is not in voxel set-level FWER control, but in simultaneous TDP lower bounds for every cluster (Genovese & Wasserman, 2006; Goeman & Solari, 2011). When voxel-wise null hypotheses are point null hypotheses, upper bounds to TDP are impossible (Goeman & Solari, 2011). We will now construct such TDP bounds.

Let $A_\mathrm{P} = \{v \in B : \mathrm{P} \notin H_v\}$ be the set of all truly active voxels in the brain. For voxel set $V \subseteq B$ the number of truly active voxels in V is

$$a_\mathrm{P}(V) = |V \cap A_\mathrm{P}|.$$

If the researcher would claim that voxel set V is active, the researcher would be right about $a_\mathrm{P}(V)$ voxels, and wrong about $|V| - a_\mathrm{P}(V)$ of them. We call

$$\pi_\mathrm{P}(V) = \frac{a_\mathrm{P}(V)}{|V|},$$

or 0 if $V = \emptyset$, the TDP of set V. This is our target of inference. We will infer on $\pi_\mathrm{P}(V)$ through $a_\mathrm{P}(V)$, which is easier to work with.

Goeman and Solari (2011) proved that, for any closed testing procedure with effective local tests $(\boldsymbol{\psi}_V)_{V\subseteq M}$, random variables defined, for all $V \subseteq M$, as

$$\mathbf{a}(V) = \min\{|V \setminus W| : W \subseteq V, \boldsymbol{\psi}_W = 0\}, \tag{6}$$

have the property that, for all $\mathrm{P} \in \Omega$,

$$\mathrm{P}(\mathbf{a}(V) \leq a_\mathrm{P}(V) \text{ for all } V \subseteq M) \geq 1 - \alpha. \tag{7}$$

A lower bound for the TDP follows immediately: $\boldsymbol{\pi}(V) = \mathbf{a}(V)/|V|$, or 0 if $V = \emptyset$, is a simultaneous lower bound for the TDP all $V \subseteq M$. By equation (7), for all $\mathrm{P} \in \Omega$, we have

$$\mathrm{P}(\boldsymbol{\pi}(V) \leq \pi_\mathrm{P}(V) \text{ for all } V \subseteq M) \geq 1 - \alpha.$$

The lower bound $\mathbf{a}(V)$, and its companion $\boldsymbol{\pi}(V)$ provide much stronger statements than the effective local test. Where $\boldsymbol{\psi}_V$ only gives confidence whether or not there is signal present in V, $\mathbf{a}(V)$ gives confidence for the amount of signal. There is no information lost in reporting $\mathbf{a}(V)$ rather than rejection or non-rejection $\boldsymbol{\psi}_V$, since $\mathbf{a}(V) \geq \boldsymbol{\psi}_V$, as follows immediately from the definition. The simultaneity of equation (7) implies family-wise error control over all $V \subseteq M$ considered or reported: with probability at least $1 - \alpha$ no reported $\mathbf{a}(V)$, $V \subseteq M$, overestimates the number of truly active voxels $a_{\mathrm{P}}(V)$ in V, even if V was chosen after seeing the data.

Since $\mathbf{a}(V)$ involves the expression $\boldsymbol{\psi}_V$, which is difficult to calculate, we use the shortcut $\underline{\boldsymbol{\psi}}_V$ to get a partial shortcut for $\mathbf{a}(V)$. We write

$$\breve{\mathbf{a}}(V) = \min\{|V \setminus W| : W \subseteq V,\ \underline{\boldsymbol{\psi}}_W = 0\}.$$

By Lemma 1, $\breve{\mathbf{a}}(V) \leq \mathbf{a}(V)$, so $\breve{\mathbf{a}}(V)$ inherits the property (7). Moreover, $\breve{\mathbf{a}}(V)$ can be rewritten in a relatively simple form. The formulation of $\breve{\mathbf{a}}(V)$ and its property are our first main result. We formulate it as a theorem.

Theorem 1 Let

$$\breve{\mathbf{a}}(V) = s_{k_M}(V \cap \mathbf{Z}), \tag{8}$$

where $s_k(V) = \min\{|R| : \chi_{V \setminus R} \leq k\}$. Then, for all $\mathrm{P} \in \Omega$,

$$\mathrm{P}(\breve{\mathbf{a}}(V) \leq a_{\mathrm{P}}(V) \text{ for all } V \subseteq M) \geq 1 - \alpha. \tag{9}$$

Although $\breve{\mathbf{a}}(V)$ may yield smaller TDP than $\mathbf{a}(V)$, the resulting TDP lower bounds are still at least as powerful as the statements of classic cluster inference, as the next theorem asserts: all clusters found by classic cluster inference have a strictly positive TDP bound.

Theorem 2 If $\mathrm{C} \subseteq (\mathbf{Z} \cap M)$, with $|\mathrm{C}| > k_M$, is a cluster, then $\breve{\mathbf{a}}(\mathrm{C}) > 0$.

5 Calculating TDPs

The shortcut (8) reduces a computation time of $\mathbf{a}(V)$ that is exponential in $|M|$ to a computation time for $\breve{\mathbf{a}}(V)$ that is exponential in $|V|$. This is still prohibitive for most regions V. In this section, we discuss algorithms for $\breve{\mathbf{a}}(V)$. We show that this calculation is equivalent to solving a problem known as the k-separator problem in graph theory. For the specific case of that problem in the voxel graph with 26-connectivity, we obtain a lower bound to $\breve{\mathbf{a}}(V)$ that has computation time $O(|V|^{1+1/d})$, and a fast heuristic algorithm, coupled with simulated annealing, that approaches $\breve{\mathbf{a}}(V)$ from above. Both the lower bound and the simulated annealing algorithm rely on a duality between our k-separator problem and tiling problem on a slightly larger object, which we will derive and explain.

5.1 The *k*-separator problem

From Theorem 1, we see that we have efficient computation of $\breve{\mathbf{a}}(V)$ whenever we can efficiently compute $s_k(V)$, for $V \subseteq \mathbf{Z}$. The value of $s_k(V)$ is the minimum number of voxels that must be removed from V in order that the remainder falls apart into disconnected components of size k. The quantity $s_k(V)$ can be defined for any graph, and is known in graph theory literature as the k-separator problem (Ben-Ameur et al., 2015). The k-separator problem is NP-hard, even for small fixed values of k. For example, with $k = 1$ we have a classic vertex cover problem (NP-hard), while for $k = 2$ the problem is equivalent to the computation of dissociation number which is NP-complete for a class of bipartite graphs (Yannakakis, 1981). Ben-Ameur et al. (2015) proposed polynomial time solutions to several constrained variants of the k-separator problem; however, none of them is applicable in our case. In the next few sections, we present novel solutions tailored to the specific type of graph induced by the neuroimaging context.

5.2 Preliminaries

Any voxel set V can always be written as a union of disconnected clusters. The next lemma says that it is sufficient to calculate s_k for these clusters.

Lemma 2 If $V = C_1 \cup \cdots \cup C_n$, where $C_1 \ldots, C_n$ are disconnected clusters, then

$$s_k(V) = \sum_{i=1}^{n} s_k(C_i).$$

Without loss of generality, therefore, we can focus on calculating $s_k(V)$ only for $V \subseteq B$ that are clusters. However, the results in the remainder of this section are for general voxel sets V.

5.3 Positive neighbours

For our solutions to the k-separator problem we will exploit a duality between k-separating V and tiling a somewhat larger object. To construct this duality, we first need to introduce to $\mathbf{Z}^d$ the directed relationship of being ‘positive neighbours’.

We say that $w \in \mathbb{Z}^d$ is a *positive neighbour* of $v \in \mathbb{Z}^d$ if $w - v \in \{0, 1\}^d$. We write

$$\{v\}^+ = \{v + e : e \in \{0, 1\}^d\}$$

for the voxel set of all positive neighbours of v. If $w \in \{v\}^+$ we call v a *negative neighbour* of w, since $v - w \in \{-1, 0\}^d$. Note that the positive and negative neighbours do not partition the neighbours. For example, if $d = 2$, $w = (-1, 1)$, though a neighbour of $v = (0, 0)$, is neither its positive or its negative neighbour. Moreover, every v is always both a positive and a negative neighbour of itself.

The concept of the positive neighbours allows the definition of three useful derived voxel sets from every finite voxel set $V \subset \mathbb{Z}^d$. We define the *cover* V^+ of V as

$$V^+ = \{v + e : v \in V, e \in \{0, 1\}^d\} = \bigcup_{v \in V} \{v\}^+$$

the set of all voxels in V and their positive neighbours. The *interior* V^- of V is

$$V^- = \{v \in V : \ v + e \in V \text{ for all } e \in \{0, 1\}^d\}.$$

the set of all $v \in V$ that only have positive neighbours in V. Finally, the *shave* of V is $V^0 = V \setminus V^-$. This is the ‘positive edge’ of V, the set of voxels in V that have at least one positive neighbour outside V. These three derived voxel sets will allow us to rewrite the k-separator problem into a tiling problem.

5.4 Tiling

To calculate $s_k(V)$, we are interested in k-separators, defined as voxel sets $R \subseteq V$ with the property that $\chi_{V \setminus R} \leq k$. The value of $s_k(V)$ is the minimum $|R|$ over all k-separators. In this section, we will show that minimising $|R|$ over all k-separators is equivalent to minimising a function $t_k(T_1, \ldots, T_n)$ over all tilings $T_1, \ldots, T_n$ of V^+. The latter will turn out to be an easier problem formulation to work with.

Define a *tiling* of V^+ as a collection of pair-wise disjoint voxel sets $T_1, \ldots, T_n$, called *tiles*, such that $\bigcup_{i=1}^{n} T_i = V^+$. Note that every two distinct tiles from a tilling are disjoint as sets but their voxels may induce a connected graph. Given a tiling $T_1, \ldots, T_n$ of V^+, we will be interested in the function

$$t_k(T_1, \ldots, T_n) = \sum_{i=1}^{n} |T_i^0 \cap V| + \sum_{i=1}^{n} (|T_i^- \cap V| - k)_+, \tag{10}$$

where $(\cdot)_+$ is the positive part function. This function is the link between tilings and k-separators, as the following two lemmas state.

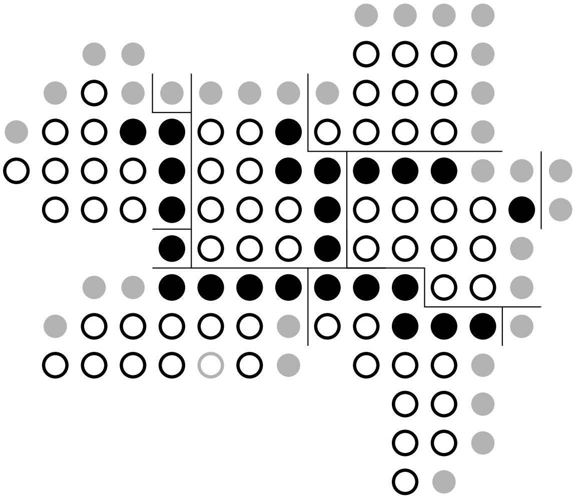

Figure 2. Illustration of a k-separator and a corresponding tiling, with $d = 2$ and $k = 10$. The voxel set V comprises of all black voxels (open and filled). The set V^+ comprises of V and all the grey voxels (open and closed). The k-separator R is the set of all filled black voxels. The corresponding tiling is indicated by the lines. All filled voxels are part of the shave T^0 for their respective tile T; open voxels are part of the interior T^-.

Lemma 3 For every tiling $T_1, \ldots, T_n$ of V^+ there exists a k-separator R of V such that

$$|R| = t_k(T_1, \ldots, T_n).$$

Lemma 4 For every k-separator R of V there exists a tiling $T_1, \ldots, T_n$ of V^+ such that $T_1, \ldots, T_n$ are clusters, and

$$|R| \geq t_k(T_1, \ldots, T_n).$$

To get some intuition why these lemmas are true, it is helpful to consider a property of neighbours and positive neighbours proven as Lemma 9 in the online supplementary material, Section A: two voxels are neighbours if and only if they have a common positive neighbour. It follows that voxel sets V and W are disconnected if and only if V^+ and W^+ are disjoint. It is this connection between disconnectedness of sets and simple disjointness of slightly larger sets that is exploited in Lemmas 3 and 4. Loosely, if R cuts V as $V \setminus R = C_1 \cup \ldots \cup C_n$, with $C_1, \ldots, C_n$ pair-wise disconnected, then $C_1^+, \ldots, C_n^+ \subseteq V^+$ are pair-wise disjoint tiles. Vice versa if $T_1, \ldots, T_n \subseteq V^+$ are pair-wise disjoint tiles, then their interiors $T_1^-, \ldots, T_n^- \subseteq V$ are pair-wise disconnected; if these interiors are of size a most k, then $R = (T_1^0 \cup \ldots \cup T_n^0) \cap V$ separates V. We illustrate the link between k-separator and tiling with an example in Figure 2.

Combining Lemmas 3 and 4, it follows that minimising $|R|$ over all k-separators is equivalent to minimising $t_k(T_1, \ldots, T_n)$ over all tilings. We formulate this result as a theorem.

Theorem 3 We have

$$s_k(V) = \min \{t_k(T_1, \ldots, T_n) : T_1, \ldots, T_n \text{ is a tiling of } V^+\}.$$

The minimum is attained for a tiling for which $T_1, \ldots, T_n$ are all clusters.

Theorem 3 rewrites the k-separator problem but does not simplify it. There is no obvious way to minimise $t_k(T_1, \ldots, T_n)$ in polynomial time. However, we will exploit this theorem in the next three sections to construct a lower bound to $s_k(V)$, and a heuristic approximation to it.

5.5 A lower bound

First, we construct a lower bound to $s_k(V)$. Replacing $s_k(V)$ by its lower bound in Theorem 1 retains the TDP guarantee implied by that theorem. As a consequence, the lower bound will be a shortcut to the closed testing procedure: it retains the guarantee on the TDP, but sacrifices some inferential power for computational reasons. We will derive this shortcut in two stages. First, in this section, we will calculate a shortcut with $O(|V|)$ time complexity. Next, in Section 5.6, we will construct a more powerful shortcut in $O(|V|^{1+1/d})$ time.

The rationale behind the shortcut is that to minimise the expression (10) we should favour tiles T with $|T^- \cap V| \le k$, since for such tiles the second term of equation (10) disappears. For such tiles, minimising t amounts to finding tiles T with as small as possible edge ratio $|T^0|/|T|$. However, if $|T^-| \le k$, the edge ratio is bounded from below by the most efficient such ratio possible. This optimal edge ratio r_k can be used to bound $s_k(V)$. We formulate this result as Theorem 4.

Theorem 4

$$s_k(V) \ge r_k \cdot |V^+| - |V^+ \setminus V|,$$

where

$$r_k = \min\{|V^0| / |V| : \emptyset \ne V \subset \mathbb{Z}^d,\ \ |V^-| \le k\}. \tag{11}$$

Define $\underline{s}_k(V) = r_k \cdot |V^+| - |V^+ \setminus V|$. How can we interpret this lower bound? We see that $\underline{s}_k(V)$ is large if its size $|V|$ is large relative to the size $|V^+|$ of its cover. It takes large values therefore for large and compact V, and small values for smaller or irregular sets V. The calculation of r_k is given in Lemma 5. We plot r_k for $k = 1, \ldots, 100$ and $d = 2, 3, 4$ in Figure 3.

Lemma 5 If $k = 0$, we have $r_k = 1$. If $k > 0$, we have

$$r_k = \min_{1 \le j \le k} \frac{f_{d,j} - j}{f_{d,j}},$$

where $f_{d,k} = 0$ if $d = 0$ or $k = 0$, and, for $d > 1$, we have recursively

$$f_{d,k} = b^+_{d,k} + f_{d-1,k-b_{d,k}}.$$

Here,

$$b_{d,k} = \left(\lfloor k^{1/d} \rfloor\right)^{d-l_{d,k}} \left(\lfloor k^{1/d} \rfloor + 1\right)^{l_{d,k}},$$

and

$$b^+_{d,k} = \left(\lfloor k^{1/d} \rfloor + 1\right)^{d-l_{d,k}} \left(\lfloor k^{1/d} \rfloor + 2\right)^{l_{d,k}},$$

where

$$l_{d,k} = \left\lfloor \frac{\log(k) - d\log(\lfloor k^{1/d} \rfloor)}{\log(\lfloor k^{1/d} \rfloor + 1) - \log(\lfloor k^{1/d} \rfloor)} \right\rfloor.$$

In the example object of Figure 2, we find from Lemma 5 that with $d = 2$ and $k = 10$ we have $r_k = 7/16$. With $|V| = 84$ and $|V^+| = 118$, we get $\underline{s}_k(V) = 17.6$, so $s_k(V) \ge 18$.

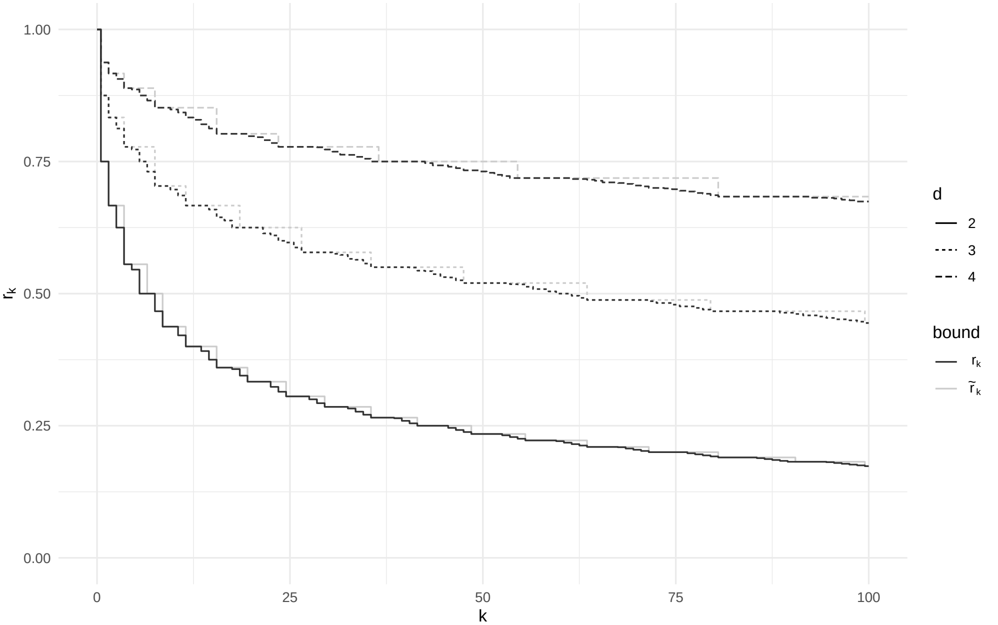

Figure 3. The thresholds r_k and $\tilde{r}_k$, defined in Theorem 4 and Lemma 8, respectively, as a function of the extent threshold k for dimensions $d = 2, 3, 4$.

5.6 Pruning

Irregularly shaped objects V have low $\underline{s}_k(V)$. It can therefore pay to prune V to $V' \subseteq V$ in order to bound $s_k(V)$ from below by $\underline{s}_k(V') \le s_k(V') \le s_k(V)$. We will use this to get an improved bound on $s_k(V)$.

Suitable choices are $V' = (V^-)^+$, $V'' = (((V^-)^-)^+)^+$, etc., which prune away increasingly broad extremities of V. We illustrate V' in Figure 4. In this example, we have $|V'| = 78$, $|(V')^+| = 106$, and we find $\underline{s}_k(V') = 18.4$, so $s_k(V) \ge 19$. Further pruning to V'' leads to $|V''| = 69$ and $|(V'')^+| = 92$ for $\underline{s}_k(V'') = 17.3$ (see figures in the online supplementary material, Section C). Further pruning does not lead to better bounds. In any case, as Lemma 6 states, pruning more than $|V|^{1/d}$ times is never necessary.

Lemma 6 If $i \ge \lfloor |V|^{1/d} \rfloor$, then $V^{(i)} = \emptyset$.

Taking pruning into account, and using that $s_k(V) > 0$ if $|V| > k$, we define the improved bound

$$\check{s}_k(V) = \mathbb{1}\{\chi_V > k\} \vee \max\left\{\left\lceil \underline{s}_k(V^{(i)}) \right\rceil : i = 0, 1, \ldots, |V|^{1/d}\right\},$$

where $V^{(i)}$ is obtained from V by performing the $(\cdot)^-$ operation i times, followed by the $(\cdot)^+$ operation i times.

Taking everything together, the proposed procedure and its TDP guarantee property are summarised in the following theorem, which proves that the lower bound is a shortcut to the closed testing procedure.

Theorem 5 For every $V \subseteq M$, let

$$\underline{\mathbf{a}}(V) = \sum_{i=1}^{\mathbf{n}} \check{s}_{k_M}(\mathbf{C}_i),$$

where $\mathbf{C}_1, \ldots, \mathbf{C}_\mathbf{n}$ are disconnected clusters such that $\mathbf{C}_1 \cup \cdots \cup \mathbf{C}_\mathbf{n} = V \cap \mathbf{Z}$. Then, for all $\mathrm{P} \in \Omega$,

$$\mathrm{P}(\underline{\mathbf{a}}(V) \le a_\mathrm{P}(V) \text{ for all } V \subseteq M) \ge 1 - \alpha.$$

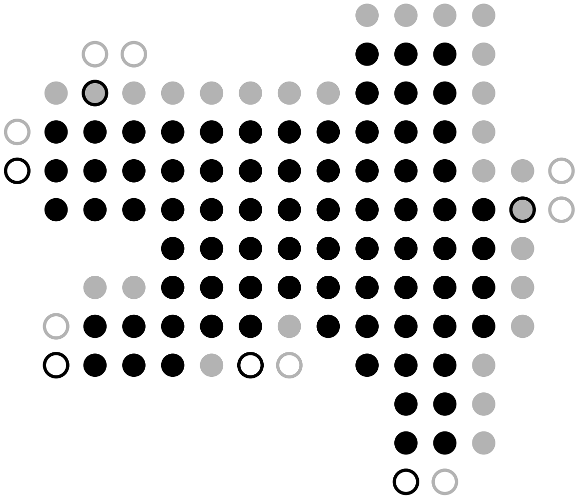

Figure 4. Illustration of the pruning V' of the voxel set V from Figure 2. The voxel set V consists of all black voxels (open and filled); the set V^+ additionally comprises of the grey voxels (open and filled). The pruned set $V' = (V^-)^+$ consists of the filled black voxels, and its cover $(V')^+$ of all filled grey voxels. We see that each voxel removed to obtain V' nets a reduction in size of two voxels for $(V')^+$, resulting in a net gain in $\underline{s}_k(V')$ relative to $\underline{s}_k(V)$, since $r_k \leq 1/2$.

Computational complexity for $\underline{s}_k(V)$, ignoring constants in d, is $O(|V|)$, and for $\check{s}(V)$ is $O(|V|^{1+1/d})$, so that is also the computational complexity of $\underline{\mathbf{a}}(V)$ if V is a supra-threshold cluster. For general V, complexity is the sum of the complexity of its comprising clusters, which is $O(|V|^{1+1/d})$ in the worst case that V is a supra-threshold cluster.

It is easy to verify that the shortcut of Theorem 5 also retains the property of Theorem 2 that it uniformly improves classic cluster inference. Still, it sacrifices some power, since the lower bound $\check{s}_k(V)$ may be (much) smaller than $s_k(V)$. The difference between $s_k(V)$ and $\check{s}(V)$ can be expected to be relatively large especially if $|V|/k$ is small and if V is irregularly shaped.

5.7 Heuristic algorithms to minimise *k*-separators

The strength of the shortcut of the previous paragraph is its guaranteed TDP control, as expressed in Theorem 5. To obtain this control the shortcut sacrifices power in exchange for computational efficiency. In this section we present an alternative computational approach that aims to approximate $s_k(V)$ heuristically as closely as possible, instead of bounding it from below. The algorithm has two parts. First, a heuristic algorithm finds a good separator. Next, an attempt is made to find a local improvement of the solution using simulated annealing. The second phase of the algorithm uses Theorem 3.

The first heuristic algorithm finds clusterings with acceptable sizes of separator sets. The algorithm consists of two phases: inferring an initial clustering, and improving regions consisting of a small number of neighbouring clusters. In the first phase, the algorithm starts from an empty clustering. It generates a small number of candidate clusters, where the number is a small integer, usually between 1 and 10. Each candidate cluster is created starting from a randomly chosen available voxel by a sequence of insertions of adjacent voxels such that the induced size of its separator is kept small. Then, the best candidate cluster, i.e. the cluster with the separator's minimal size, is inserted into the current clustering. The procedure is repeated until there is no space to insert a new cluster. The second phase consists of repetitions of local improvements. The algorithm randomly takes a small number of neighbouring clusters, removes them from the current clustering, and applies a procedure similar to the first phase to find a better setting of clusters.

We follow up on the optimal heuristic separator using a simulated annealing algorithm, as follows. The separator of V is translated to a tiling of V^+ according to Lemma 4. In each step, the algorithm chooses a random voxel $v \in V^+$ and a random neighbour $w \in V^+$ of v. If v and w are part of the same tile T with interior size $|T^- \cap V| > k$, the algorithm proposes to start a new tile

$\{v\}$; otherwise it proposes to reassign v from its old tile to the tile of w. If the target function t' of the proposed tiling is lower than or equal to the target function t of the previous step, the proposal is always accepted. Otherwise, the proposal is accepted with a probability that is a decreasing function of $t' - t$ and of the current iteration number. After a maximum number of iterations is reached, the algorithm returns the best solution it found during its travels through the search space.

The first algorithm was implemented in C and the simulated annealing in Python. The algorithms are usually invoked with a time limit setting. Pseudo-code for both heuristic algorithms are given in the online supplementary material, Section B.

The heuristic algorithms are not guaranteed to find the global minimum with a finite running time. If the algorithm did not find the correct solution, the value found is larger than the actual minimum $s_k(V)$, so there is no formal guarantee of TDP control comparable to Theorem 5. Still, the overstatement of $s_k(V)$ may often be less than the understatement of $s_k(V)$ due to the lower bound (4). The heuristic approach may therefore be the preferred solution in practice if computation time is not an issue and a small overstatement of TDP is acceptable.

5.8 Heuristic algorithm performance

A heuristic algorithm for a computationally hard problem cannot guarantee to find the optimal solution. Also estimating the error of such approaches is usually a difficult task. One way to proceed is to use exact solution approaches such as exhaustive enumeration, dynamic programming, or integer linear programming formulations. However, in the case of intractable problems, they can only be applied to small instances. Here, we propose a different approach. First, we show that some instances of the k-separator problem are tractable by showing their exact solution. Next, to estimate an error of the heuristic algorithm given the input consisting of multiple data sets, we generate a number of tractable instances matching properties of the input and jointly apply the heuristic algorithm under the same parameter setting. Finally, knowing the exact solution of tractable instances, we can estimate the solution error of the input data sets. The main result is formulated below in Lemma 7

> **Lemma 7** Let $k = n^d$ and c be a vector of d positive integers. If the dimensions of a hyperrectangle R are $(n+1)c_i - 1$ for $i = 1, \ldots, d$, then the bound of Theorem 4 is exact, so that the optimal k-separator of R has $|R| - n^d \Pi c_i$ voxels.

Since our algorithm is not utilising the information on the shape of the input voxel sets, nor the clusters are formed as cubes in the sampling, we believe that the benchmark of correctness based on hyperrectangles is a good indicator of how scores from the heuristic differ from the optimal ones.

To estimate the error of the heuristic algorithm, we inferred a collection of hyperrectangle tests based on the three-dimensional data sets from the Neurovault repository (see Section 10). Our goal was to cover the whole range of k values and data sets sizes from the input repository. Therefore, we set k bounded above 1,000, and the hyperrectangle, i.e. cuboid, sizes to maximum 18,000 voxels. Being consistent with the notation from Lemma 7, each test is uniquely determined by four integer parameters n, c_1, c_2, $c_3 \leq 10$, where k is n^3, and the corresponding cuboid has dimensions $(n+1)c_i - 1$, for each i. After rejecting too large cuboids, we obtained 1,064 tests, which enlarged the input repository by nearly 9%.

The experiment indicated that nearly 50% of tests were completed with no error, and the worst errors of 5%–6% has only $\sim$ 5% of tests. A more detailed summary is depicted in Figure 5 with boxplots of errors for each value of k, where the 0% represents no error. The results obtained on cuboid tests indicate that the sizes of separators inferred by our heuristic algorithm are optimal in nearly half of the cases. For the rest of the cases, the error is usually below 4% with high confidence and the median error is below 2%.

6 Choosing thresholds

6.1 Voxel-wise inference

An alternative to cluster-extent inference is classic voxel-wise inference. In voxel-wise inference, FWER is controlled over all voxel-wise null hypotheses. This is achieved by finding the

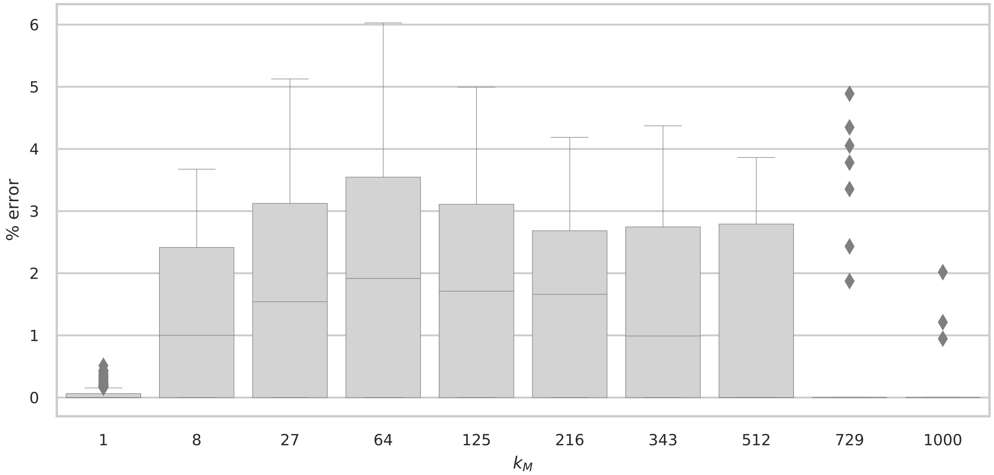

Figure 5. Upper bound heuristic performance: a boxplots of errors as a percentage of the true *k*-separator. Note that all errors are overestimates by construction.

$(1-\alpha)$-quantile of the distribution of the maximal *z*-score under the global null hypothesis H_M, and rejecting the null hypothesis whenever a voxel's *z*-score exceeds this threshold. Although cluster-extent inference is often contrasted sharply with voxel-wise inference, suggesting that these are two very different modes of operation. It was noted by Poline et al. (1997) and Friston et al. (1994) that classic voxel-wise inference is simply a special case of cluster-extent inference, obtained by choosing $k_M = 0$. It follows that we can get a TDP per cluster from voxel-wise inference.

In classic voxel-wise inference, we reject H_v for all voxels $v \in \mathbf{Z} = \{v \in M : \mathbf{z}_v \geq z\}$, where z is chosen as the smallest value such that

$$\mathrm{P}(|M \cap \mathbf{Z}| > 0) \leq \alpha \tag{12}$$

holds for all $\mathrm{P} \in H_M$. It has been shown (Friston et al., 1991; Worsley et al., 1992) that voxel-wise inference controls voxel-wise FWER, i.e. for all $\mathrm{P} \in \Omega$,

$$\mathrm{P}(\mathbf{Z} \nsubseteq A_{\mathrm{P}}) \leq \alpha.$$

We can embed voxel-wise inference into the closed testing procedure we have constructed by remarking that $|M \cap \mathbf{Z}| > 0$ if and only if $\chi_{M \cap \mathbf{Z}} > 0$. Therefore, equation (12) is equivalent to

$$\mathrm{P}(\chi_{M \cap \mathbf{Z}} > 0) \leq \alpha,$$

which is simply equation (2) with $k_M = 0$, and the latter is a valid choice for k_M. The closed testing procedure resulting from this choice is a relatively simple one, as the following theorem states.

Theorem 6 If $k_M = 0$, then for all $V \subseteq M$ we have

$$\underline{\mathbf{a}}(V) = \check{\mathbf{a}}(V) = \mathbf{a}(V) = |V \cap \mathbf{Z}|.$$

The theorem says how to calculate TDP for clusters when doing voxel-wise inference: the TDP lower bound for a set V is simply the fraction of voxel-wise significant voxels among the voxels in V. Supra-threshold clusters obtained with $k_M = 0$ always have a TDP of 100%.

6.2 Choosing k_M

Cluster-extent inference assumes that z and k_M are chosen in such a way that equation (2) holds. It is common in cluster-extent inference to fix the *z*-score threshold z, and to calculate k_M as the smallest value such that equation (2) is satisfied (Friston et al., 1994). However, we saw in the previous section that the order is reversed in voxel-wise inference: there $k_M = 0$ is fixed, and z is chosen as the smallest value of z satisfying equation (2). In this section, we argue that the order of fixing k_M calculating z should be generally preferred, both from the perspective of power and obtaining a good TDP bound.

It is perfectly valid to choose k_M first, and to find a value of z that corresponds to this k_M, as previously proposed by Bullmore et al. (1999). The relationship between z and k_M depends only on the null model H_M, and not on the observed z-scores. For cluster inference based on random field theory, the relationship between z and k_M depends on the smoothness of the field, which is estimated from the independent residuals. For cluster inference based on permutations, k_M is calculated from the matrix of all permutation z-scores, and can be calculated without knowing which permutation corresponds to the real data. We present a fast algorithm for finding z based on k_M using permutations in the online supplementary material, Section E.

It is generally (slightly) more powerful to choose k_M rather than z. The reason for this is that k_M is discrete, while z is continuous. When fixing z and calculating k_M there is almost always a smaller value of z that would result in the same value of k_M. Using this value instead of the previously chosen z would result in a uniformly more powerful method but still controls TDP, since equation (2) still holds. We may therefore, after choosing z and finding k_M, always re-calibrate our z.

Alternatively, we may simply choose k_M and find z as the smallest value such that equation (2) holds, as is done in voxel-wise inference. This has the important advantage that the achievable TDP can be better controlled.

7 Upper bounds

In this section, we present two upper bound results that impose hard limits on the TDP that can be achieved with closed testing based on cluster-extent inference. The first bound, in Section 6.2, limits what can be achieved using the lower bound; this result helps to choose the settings of that method. The second bound, in Section 7, limits what can be achieved in terms of TDP by the full closed procedure (6). Since closed testing procedures can only be uniformly improved by improving their local tests (Goeman et al., 2021), and that the room for such improvements is limited if equation (2) is tight, this sets a limit on the potential of any method that is consistent with classic cluster extent inference.

The maximal achievable TDP from the shortcut can be calculated as a function of k_M and cluster size $|\mathbf{C}|$ by the following theorem.

Theorem 7 For every cluster $\mathbf{C} \subseteq \mathbf{Z}$, we have

$$\underline{\mathbf{a}}(\mathbf{C}) \leq \left\lceil \frac{r_{k_M} - r_{|\mathbf{C}|}}{1 - r_{|\mathbf{C}|}} \cdot |\mathbf{C}| \right\rceil \vee \mathbb{1}\{|\mathbf{C}| > k_M\}.$$

By Theorem 7, to achieve a TDP of γ, for some $\gamma > 1/k_M$, we need a cluster $\mathbf{C}$ with

$$r_{|\mathbf{C}|} \leq \frac{r_{k_M} - \gamma}{1 - \gamma}.$$

The maximal TDP according to $\underline{a}$ for different values of k_M and different cluster size $|\mathbf{C}|$ is given in Figure 6. Since $r_{|\mathbf{C}|} \to 0$ as $|\mathbf{C}| \to \infty$, the TDP lower bound $\underline{\mathbf{a}}(\mathbf{C})/|\mathbf{C}|$ achieved by the shortcut of Theorem 5 is at most r_{k_M} for very large clusters, and much smaller than that for small and irregular clusters. The maximal TDP values converge to r_k as the cluster size increases. Clusters may achieve the maximal TDP if they are highly compact. Irregular clusters tend to have (much) smaller TDP.

We see from Figure 6 that, with large values of k_M, it is difficult or even impossible to achieve good TDP even for large clusters, so a small value of k_M is recommended if large TDP is desired. Assuming that we are interested in finding clusters with TDP ≥ 1/2, a sweet spot with $d = 3$ seems to be $k_M = 14$, for which $r_{k_M} = 2/3$. To achieve TDP ≥ 1/2, clusters need to have $r_{|V|} \leq 1/3$, which implies $|V| \geq 339$.

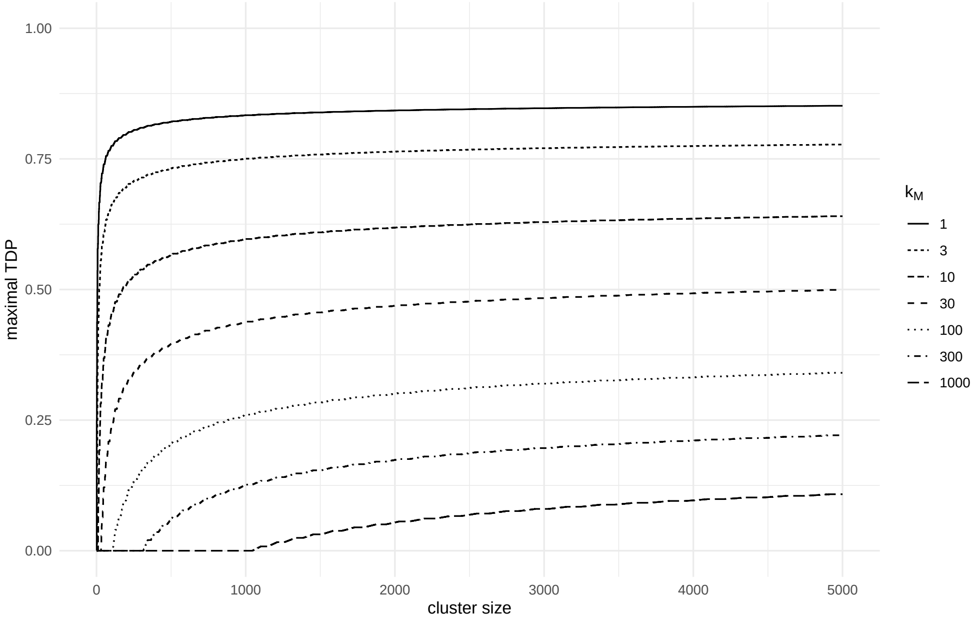

Figure 6. The maximal true discovery proportion (TDP) according to the shortcut $\underline{a}(V)$, defined in Theorem 5, as a function of the extent threshold k_M and cluster size, for dimensions $d = 3$.

For $\check{\mathbf{a}}(V)$ we have a weaker bound $\check{\mathbf{a}}(V) \leq \tilde{r}_{k_M} \cdot |V|$, from Lemma 8, below, that bounds TDP by $\tilde{r}_k \approx r_k$. The value of $\tilde{r}_k$ is illustrated in Figure 3 in Section 5.5. This bound suggests that also when using the heuristic approximation to the k-separator problem, a researcher would want to use a value of k_M that yields $\tilde{r}_{k_M}$ substantially above the target TDP, e.g. getting a TDP over 0.5 is impossible if $k_M > 64$, and remains unlikely unless k_M is substantially smaller than 64, since the bound of Lemma 8 is not very tight.

Lemma 8 We have $s_k(V) \leq \tilde{r}_k \cdot |V|$, where $\tilde{r}_k = (b^+_{d,k} - b_{d,k})/b^+_{d,k}$.

Note that $b_{d,k}$ and $b^+_{d,k}$ are defined in Lemma 5.

Theorem 7 and Lemma 8 give upper bounds for the shortcuts to the closed testing procedure. Such bounds are useful for researchers intending to use these shortcuts. We can also consider an upper bound to the full closed testing procedure (7) itself, given below in Theorem 8. This bound is of fundamental and practical interest, as we will explain.

Theorem 8 Let $\bar{\mathbf{a}}(V) = s_{k_{M \setminus \mathbf{Z}}}(V \cap \mathbf{Z})$, then, for every $V \subseteq M$,

$$\mathbf{a}(V) \leq \bar{\mathbf{a}}(V).$$

In the proof of this theorem in the online supplementary material (Section A), we will prove a slightly tighter bound. Note the similarity of $\bar{\mathbf{a}}(V)$ with $\check{\mathbf{a}}(V)$, the only difference being that k_M is replaced by $k_{M \setminus \mathbf{Z}}$. This difference will be small unless $|\mathbf{Z}|$ is large relative to $|M|$.

Practically, Theorem 8 can be used to bound the loss $\mathbf{a}(V) - \underline{\mathbf{a}}(V)$ of the shortcut $\underline{\mathbf{a}}(V)$ relative to the full closed testing procedure $\mathbf{a}(V)$. It limits the potential for further computational improvements. In practice, unless $|\mathbf{Z}|$ is large relative to $|M|$ we will have $k_M \approx k_{M \setminus \mathbf{Z}}$, so that $\check{\mathbf{a}}(V) \approx \bar{\mathbf{a}}(V)$, and $\check{\mathbf{a}}(V) \approx \mathbf{a}(V)$.

More fundamentally, we can combine Theorem 8 with the insights from Goeman et al. (2021). We have constructed $\mathbf{a}(V)$ as the unique closed testing procedure induced by cluster-extent inference. By Goeman et al. (2021) closed testing procedures are optimal, so there is no room for improvement of the method outside the closed testing framework. Moreover, improvement within the closed testing framework is limited to improvement of the local test, and there is hardly

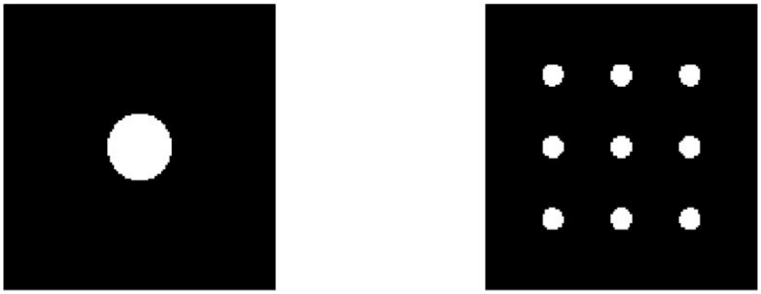

Figure 7. Two-dimensional simulated signal illustration. Focal signal (left) with one large circle in the middle; distributed signal (right) with nine identical circular regions.

room for that if z and k_M are optimised for equation (2). It follows that Theorem 8 gives a clear upper bound to the TDP arising from any method that is based on cluster-extent thresholding. Any method that achieves the result of Theorem 2 would also be constrained by the result of Theorem 8.

8 Simulation

In this section, Monte Carlo simulation is conducted to demonstrate the validity of our proposed methods, to investigate the tightness the TDP, and to see the gap between the upper and lower TDP bounds.

8.1 Set-up

Two-dimensional images, each with 128×128 pixels, were simulated. Two spatial signal configurations were considered, shown in Figure 7: (1) a focal configuration with a single large circle of signal in the middle and (2) a distributed configuration with nine small circular regions of signal spread out. The number of pixels with signal was 716 for both configurations. The simulated images were created by filling each pixel with spatially correlated noise, starting from i.i.d. standard Gaussian noise and smoothing with a spatial Gaussian smoothing kernel with full width at half maximum (FWHM) of four pixels, i.e. with $\sigma = 1.7$ pixels. Signal was added according to the chosen configuration at a fixed signal amplitude of $d = 0.1$ and $d = 0.05$, respectively. We considered 20 sample sizes n between 10 and 200 with an increment of 10, and a total of 1,000 images were generated for each simulation setting. We calculated z-scores for each voxel using a one-sample t-test. Clusters of interest were defined as all connected components of $\mathbf{Z}$ as defined in Section 3, using z-score thresholds $z = 0.348 \times \sqrt{n}$ for each sample size n.

To calculate the k_M threshold at $\alpha = 0.05$ fulfilling equation (2), we simulated a second independent null field without signal for each combination of each sample size and threshold, smoothed in the same way. We calculated k_M as the 95% quantile of the empirical distribution of the maximum cluster size in this null field. Clusters of size k_M or smaller were discarded in accordance with standard practice. Subsequently, the TDP bound was calculated using both the heuristic algorithm of Section 5.7 and the lower bound of Theorem 5.

8.2 Results

Figure 8 shows the average size of the clusters found as a function of sample size. To be precise, we calculate the total volume $\sum |\mathbf{C}_i|$ for all significant clusters, i.e. clusters $\mathbf{C}_i \subseteq \mathbf{Z}$ for which $\underline{\mathbf{a}}(\mathbf{C}_i) > 0$. The figure displays this 100 times this volume divided by true signal volume $|A_{\mathrm{P}}|$.

The figure shows a qualitative difference between the two signal amplitudes. At the high amplitude ($d = 0.1$) the clusters are consistent for the signal, with clusters converging to the true signal as the sample size increases. In contrast, at the low amplitude the clusters capture a vanishing fraction of the true signal.

Figure 9 shows the error rate of the method, which is well controlled at $\alpha = 0.05$ for all settings. The lower bound is conservative for large and for small sample sizes, while the heuristic algorithm is only conservative for large sample size. We explain this for small sample size by the compactness

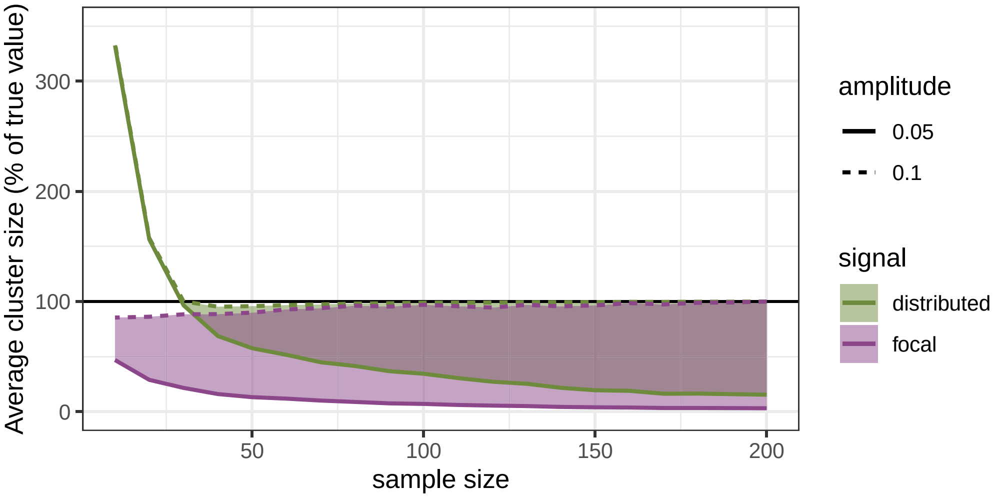

Figure 8. Average cluster sizes (expressed in percent of the true value) for focal (purple) and distributed (green) signals with the amplitudes of $d = 0.1$ (dashed line) and $d = 0.05$ (solid line).

of the chosen signal regions, for which the lower bound method tends to underestimate TDP. For large sample size, conservativeness is due to discreteness of k_M, so that the α-level in equation (2) is not exhausted. The heuristic algorithm also controls its error rate quite well in this simulation, despite the lack of a theoretical guarantee.

Figure 10 shows the TDP bounds found by the method. Displayed is the average value of the TDP over all significant clusters, i.e. over all clusters with TDP > 0. Note that the number of such clusters is much smaller for the low signal amplitude setting than for the high amplitude setting, and much larger for the distributed configuration of signal than for the focal one. We see that in all settings the TDP of significant clusters goes to 1 as sample size increases. This is because the value of k_M decreases with the sample size, eventually reaching $k_M = 0$. The difference in TDP between the lower bound and the heuristic algorithm is appreciable but not overly large, almost never exceeding 10%.

9 Application: HCP *n*-back task revisited

We illustrate the use of the new method using a more extensive analysis of the data set introduced in Section 2.

A z-score threshold z and cluster-extent threshold k_M can be defined in any way that satisfies equation (2); that is, fixing one threshold, the smallest value of the other still satisfying equation (2) can be calculated. We present the permutation-based thresholds in this section, using the fast algorithm for finding z as a function of k_M using permutations given in the online supplementary material, Section E. For comparison, the analysis with thresholds based on random field theory is given in the online supplementary material, Section F.

We present two alternative permutation-based analyses. First, we fixed $z = 3.1$, which corresponds to $k_M = 72$ in these data (Table 2). Next, we fixed $k_M = 14$ and calculated the corresponding z-threshold $z = 3.7$ (Table 3). Nonsignificant supra-threshold clusters were not displayed. The TDP bounds for relevant overlapping anatomical regions are also displayed. The clusters A to F in Table 2 are the same as in Table 1 and are visualised in Figure 1. A visualisation of the clusters in Table 3 are given in the online supplementary material.

TDP was calculated both using heuristic algorithms and using the lower bound of Theorem 5. Our heuristic algorithms were run for several hours on a cluster to produce these results, and we believe that these results are sufficiently close to the true minimum. Shorter running times of 20–60 s would give TDP results up to only 5% higher than the reported values. Comparing the heuristic results and the lower bound, the lower bound was closest to the heuristic solution for large clusters and small k_M, as expected from the theory.

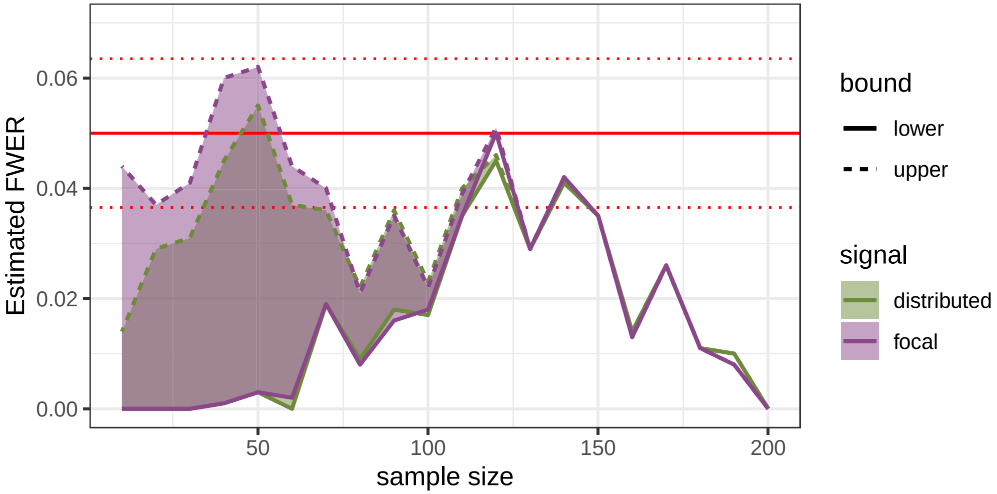

Figure 9. Estimated family-wise error rates (FWERs) for focal (purple) and distributed (green) signals with the amplitudes of $d = 0.1$. The red dotted horizontal lines represent the binomial confidence intervals for FWER at $\alpha = 0.05$ (solid horizontal line). Shown are the results for lower bound (solid line) and upper bound (dashed line) based on the heuristic algorithm. The results for $d = 0.05$ (not shown) are almost identical, since the same realisation of the noise field was used for both simulations.

Comparing the $z = 3.1$ and $k_M = 14$ settings, the results clearly show a trade-off between detection and TDP. The lower cluster-extent threshold k_M, that corresponds to a higher z-threshold, returns smaller clusters with larger TDP, while the high k_M results in larger clusters with smaller TDP. For anatomical regions it is not a priori clear whether larger TDP would be found with high or low values of k_M. In this data set, increased TDP bounds were perceived when k_M was small, i.e. when the z-threshold was large. Corresponding anatomical regions of the clusters were identified using the Harvard–Oxford cortical structural atlas and MNI structural atlas as available in FSL (Jenkinson et al., 2012).

10 Application: Neurovault

Next, we applied the new algorithm to a selection of 818 data sets from the Neurovault database (neurovault.org; Gorgolewski et al., 2015). The Neurovault database consists of unthresholded maps from neuroimaging studies. We selected 818 representative functional MRI data sets containing group-level statistics maps. For the calculation of clusters we used two settings: a standard z-threshold of $z = 3.1$, and a k-threshold of $k_M = 14$. The corresponding k_M- and z-thresholds, respectively, were estimated using Gaussian Random Field Theory (Forman et al., 1995). As residual data were unavailable, we estimated smoothness of the random field on the z-statistics image. The $z = 3.1$ setting produced values of k_M ranging from 71 to 507 (1st and 9th decile). Details of the selected images and estimation procedures can be found in the online supplementary material, Section D.

For each data set we estimated the TDP of each supra-threshold cluster obtained using $z = 3.1$ and $k_M = 14$. We then calculated for each TDP value how many supra-threshold voxels with at least that TDP were significant on average across all data sets. This allows us to visualise the relationship between the size of the clusters detected and the TDP of those clusters for different methods. We plot the theoretical lower bound of both methods (according to Theorem 5), and the solution as estimated using the heuristic methods. For reference we also calculated the number of voxels above the Gaussian random field voxel-wise threshold (equivalent to a $k_M = 0$ setting).

Figure 11 shows the results of the analysis across all data sets. It displays, for each mimimal TDP γ, the total volume $\sum |\mathrm{C}_i|$ for all clusters $\mathrm{C}_i \subseteq \mathrm{Z}$ for which $\underline{\mathbf{a}}(\mathrm{C}_i) \geq \gamma$ ('lower'), and $\overline{\mathbf{a}}(\mathrm{C}_i) \geq \gamma$ ('upper'). As can be seen the $z = 3.1$ setting (purple) leads to larger cluster sizes but with low TDP's. For $k_M = 14$ (green), the size of the clusters with low TDP's is smaller, but there are

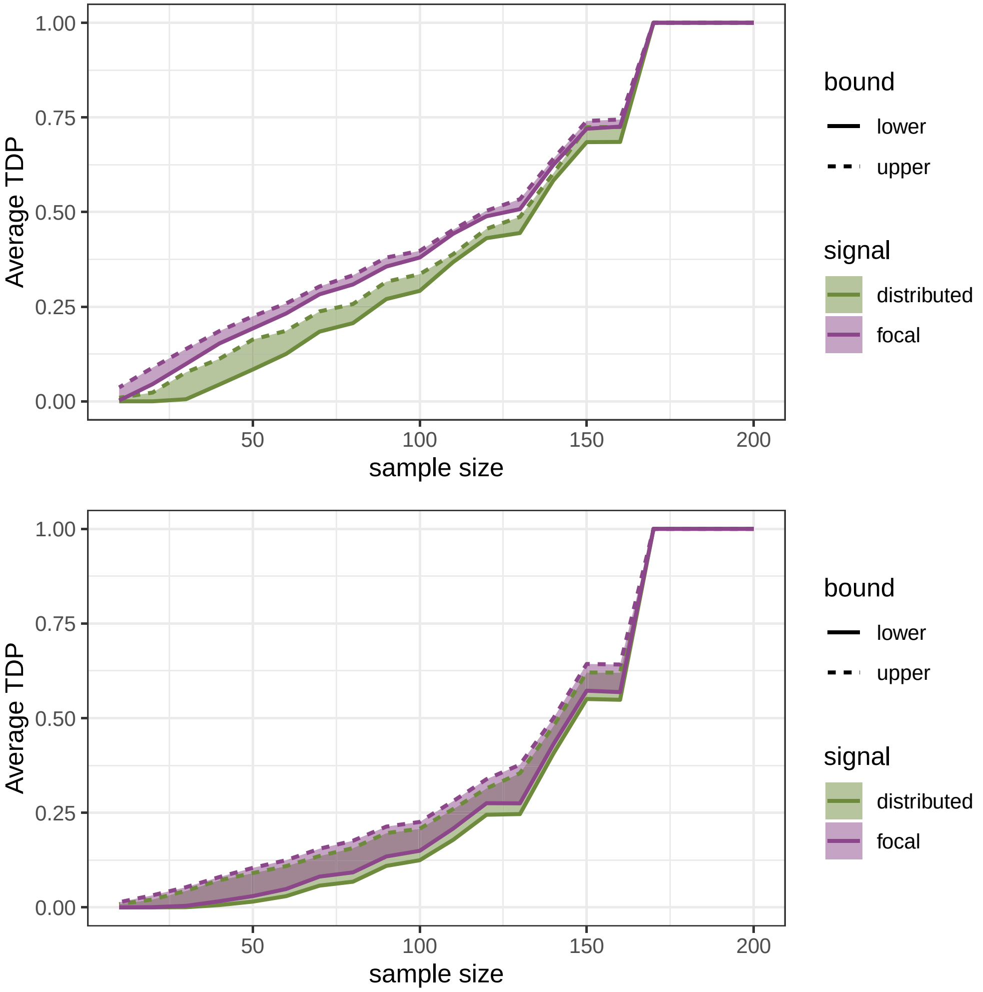

Figure 10. Average true discovery proportion (TDP) bounds for all significant clusters for focal (purple) and distributed (green) signals with the amplitudes of $d = 0.1$ (top) and $d = 0.05$ (bottom). Shown are the results for lower bound (solid line) and upper bound (dashed line) based on the heuristic algorithm.

more clusters with a more reasonable (albeit still relatively small) TDP. Both methods detect larger regions than voxel-wise inference ($k_M = 0$, black line) at low TDP thresholds, but smaller regions at high TDP. The figure shows a clear trade-off between detection and TDP: at low k_M settings, small regions are detected with large TDP; with high k_M, larger regions are detected, but TDP is (much) lower.

We note that the estimation of the smoothness using the z-statistics rather than the residuals tends to overestimate the smoothness if there is much signal. As a result, it is likely that we have overestimated values of k_M when $z = 3.1$ and overestimated z when $k_M = 14$. The TDP results in Figure 11 are therefore likely an underestimate of what would be found if the full data sets would have been available.

11 Discussion

We have presented a uniform improvement of classic cluster inference that allows much more meaningful and informative inference to be obtained from that method. In the first place, the new method allows inference on anatomical regions of interest and data-driven supra-threshold

Table 2. Results for supra-threshold clusters, defined by the cluster-forming z-threshold of $Z > 3.1$ and the resulting minimal cluster-extent threshold $k_M = 72$ based on permutation

	Cluster			Anatomical region					Location			
ID	Size	TDP	LB	Region	Size	Overlap	TDP	LB	x	y	z	$Z_{\max}$
A	8,870	0.368	0.265	MFG	18,250	4,049	0.082	0.061	44	72	60	8.87
				FP	33,571	2,021	0.020	0.013				
				IC	6,591	564	0.025	0.016				
B	8,526	0.402	0.307	sLOC	27121	5,142	0.069	0.049	19	42	61	9.51
				AG	13,689	4,260	0.117	0.089				
				pSMG	14,829	3,804	0.097	0.074				
				Precuneous	18,119	2,491	0.051	0.037				
C	7,956	0.332	0.201	Cerebellum	39,724	6,551	0.057	0.037	63	33	20	9.20
D	6,652	0.372	0.265	MFG	18,250	4,035	0.083	0.061	31	67	64	9.73
				FP	33,571	2,587	0.026	0.018				
				IC	6,591	589	0.026	0.017				
E	350	0.191	0.037	pMTG	11,420	310	0.006	0.001	15	46	28	5.18
				tMTG	9,735	271	0.005	0.000				
F	100	0.140	0.010	Cerebellum	39,724	100	0.000	0.000	49	35	10	6.56
Total	32,454	0.367	0.257	MFG	18,250	8,084	0.165	0.122				
				Cerebellum	39,724	6,651	0.058	0.037				
				sLOC	27,121	5,142	0.069	0.049				
				FP	33,571	4,608	0.046	0.031				
				AG	13,689	4,260	0.117	0.089				
				pSMG	14,829	3,804	0.097	0.074				
				Precuneous	18,119	2,491	0.051	0.037				
				IC	6,591	1,153	0.051	0.033				
				pMTG	11,420	310	0.006	0.001				
				tMTG	9,735	271	0.005	0.000				

Note. The results from the heuristic algorithms are indicated by true discovery proportion (TDP), the lower bound of Theorem 5 by LB.

clusters within the same analysis. Moreover, regions of interest do not have to be specified before seeing the data. Secondly, rather than (only) a p-value, the new method provides a true discovery proportion for every brain region. By quantifying the spatial extent of activation within the brain region, the TDP is much more informative than the p-value, which only quantifies the evidence for the presence of any signal at all. TDP is also less prone to overinterpretation than the p-value. In the Neurovault analysis we have found many examples of brain regions with a seemingly impressive $p < 0.001$ that had unremarkable TDPs of 20% or less. We recommend that TDP is always reported with (or even instead of) the p-value in fMRI cluster inference.

Despite making these additional inferences, error control remains as strict as with classic cluster inference: with probability at least $1 - \alpha$ no regions get an estimated TDP that is larger than the true value. To guarantee this error control, the method does not require any additional model assumptions. It can assume either that the z-scores of inactive voxels follow a Gaussian random field or that they are invariant under permutations.

Inference on brain regions in terms of TDP can be said to solve the Spatial Specificity Paradox (Woo et al., 2014), but by doing so it makes the same paradox painfully visible. At the usual setting with a cluster-forming threshold of $z = 3.1$ most significant brain regions have a TDP less than 20%–30%. Our analyses have made it clear that there is a trade-off involved in choosing the

Table 3. Results for supra-threshold clusters, defined by cluster-extent threshold $k_M = 14$ and the resulting cluster-forming z-threshold of $Z > 3.7$, based on permutation

	Cluster			Anatomical region					Location			
ID	Size	TDP	LB	Region	Size	Overlap	TDP	LB	x	y	z	$Z_{\max}$
1	7,231	0.606	0.532	sLOC	27,121	4,293	0.091	0.078	19	42	61	9.51
				Precuneous	18,119	2,123	0.067	0.058				
2	6,899	0.577	0.487	MFG	18,250	3,224	0.102	0.087	44	72	60	8.87
				SFG	18,946	2,880	0.085	0.073				
				PCG	9,245	1,558	0.096	0.084				
				IC	6,591	494	0.040	0.034				
3	5,345	0.546	0.438	Cerebellum	39,724	4,840	0.067	0.054	63	33	20	9.20
4	5,143	0.575	0.487	MFG	18,250	3,285	0.104	0.089	31	67	64	9.73
				FP	33,571	1,893	0.031	0.026				
				SFG	18,946	1,745	0.052	0.043				
5	202	0.391	0.158	OP	15,486	156	0.004	0.001	39	22	36	5.72
				ICC	7,134	110	0.006	0.003				
6	128	0.375	0.148	pMTG	11,420	128	0.004	0.002	15	46	28	5.18
7	66	0.379	0.182	Cerebellum	39,724	66	0.001	0.000	49	35	10	6.56
8	61	0.361	0.115	FP	33,571	61	0.001	0.000	31	86	29	5.77
9	56	0.321	0.143	FP	33,571	56	0.001	0.000	57	88	29	5.16
10	39	0.308	0.103	OP	15,486	39	0.001	0.000	51	15	42	5.35
11	22	0.182	0.045	Thalamus	4,602	17	0.000	0.000	43	53	43	4.55
12	21	0.095	0.048	Cerebellum	39,724	21	0.000	0.000	42	36	10	4.85
Total	25,213	0.573	0.482	MFG	18,250	6,509	0.206	0.176				
				Cerebellum	39,724	4,927	0.068	0.054				
				SFG	18,946	4,625	0.137	0.116				
				sLOC	27,121	4,293	0.091	0.078				
				Precuneous	18,119	2,123	0.067	0.058				
				FP	3,3571	2,010	0.032	0.026				
				PCG	9,245	1,558	0.096	0.084				
				IC	6,591	494	0.040	0.034				
				OP	15,486	195	0.004	0.001				
				pMTG	11,420	128	0.004	0.002				
				ICC	7,134	110	0.006	0.003				
				Thalamus	4,602	17	0.000	0.000				

Note. The results from the heuristic algorithms are indicated by true discovery proportion (TDP), the lower bound of Theorem 5 by LB.

cluster-forming threshold. Low thresholds result in many large clusters but with low TDP; higher thresholds have less detection power but much higher TDP. In the extreme, voxel-wise inference was shown to be a special case of cluster-extent inference that always returns a TDP of 100%. In order to obtain TDP substantially higher than a reasonably minimal threshold of 50%, we recommend cluster thresholding with $k_M = 14$ or less, resulting in much larger z-thresholds than usually recommended in the field (Eklund et al., 2016).

Computationally, the calculation of the TDP involves solving a k-separator problem. We presented two solutions to this problem: the lower bound retains the error control guarantee but is conservative; the heuristic solution is more accurate, but at the cost of losing error control if the method does not fully converge. Together, the two algorithms can be used to bracket the TDP

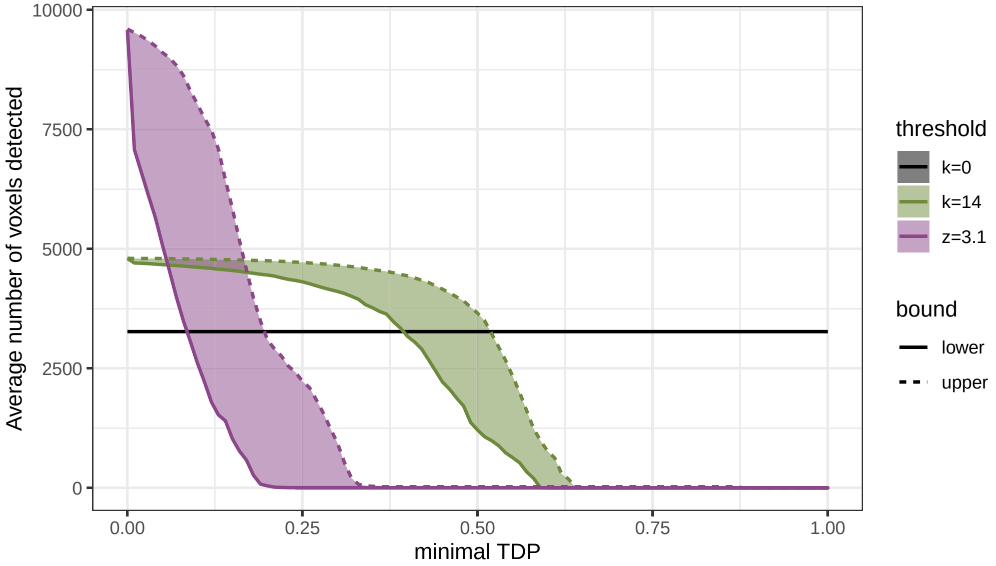

Figure 11. Relation between true discovery proportion (TDP) (x-axis) and the average number of voxels detected for three thresholds: $z = 3.1$ (purple), $k_M = 14$ (green), and $k_M = 0$ (black). The results for lower bound (solid line) and upper bound (dashed line) are shown based on the heuristic algorithm.

lower confidence bound. We recommend the heuristic solution in practice provided enough computing power is available.

Inference for neuroimaging in terms of TDP rather than p-values has been proposed by several authors before (Andreella et al., 2023; Blain et al., 2022; Rosenblatt et al., 2018; Vesely et al., in press). None of the proposed methods is expected to outperform any of the others uniformly (Goeman et al., 2021). A systematic and careful inventory should be performed to find out when to prefer which TDP methods with which tuning parameters. This large project is beyond the scope of this paper. In such a comparison, the method proposed in this paper will serve as an essential benchmark, representing the state-of-the-art of classic cluster analysis, which it is designed to be consistent with.

Although motivated by brain imaging data and methods, the novel method we have proposed can easily be applied in any setting in which hypotheses are structured on a d-dimensional rectangular grid, and interest is on inference in regions rather than individual hypotheses.

Acknowledgments

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

Conflict of interest: None declared.

Funding

This research was supported by Nederlandse Organisatie voor Wetenschappelijk Onderzoek, Grant Number: 639.072.412.

Supplementary material

Supplementary material is available online at *Journal of the Royal Statistical Society: Series B*.

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

Barch D. M., Burgess G. C., Harms M. P., Petersen S. E., Schlaggar B. L., Corbetta M., Glasser M. F., Curtiss S., Dixit S., Feldt C., Nolan D., Bryant E., Hartley T., Footer O., Bjork J. M., Poldrack R., Smith S., Johansen-Berg H., Snyder A. Z., …Consortium W.-M. H. (2013). Function in the human connectome: Task-fMRI and individual differences in behavior. *NeuroImage*, *80*, 169–189. https://doi.org/10.1016/j.neuroimage.2013.05.033

Beckmann C. F., Jenkinson M., & Smith S. M. (2003). General multilevel linear modeling for group analysis in FMRI. *NeuroImage*, *20*(2), 1052–1063. https://doi.org/10.1016/S1053-8119(03)00435-X

Ben-Ameur W., Mohamed-Sidi M.-A., & Neto J. (2015). The *k*-separator problem: Polyhedra, complexity and approximation results. *Journal of Combinatorial Optimization*, *29*(1), 276–307. https://doi.org/10.1007/s10878-014-9753-x

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. *Annals of Statistics*, *48*(3), 1281–1303. https://doi.org/10.1214/19-AOS1847

Bullmore E. T., Suckling J., Overmeyer S., Rabe-Hesketh S., Taylor E., & Brammer M. J. (1999). Global, voxel, and cluster tests, by theory and permutation, for a difference between two groups of structural MR images of the brain. *IEEE Transactions on Medical Imaging*, *18*(1), 32–42. https://doi.org/10.1109/42.750253

Eklund A., Nichols T. E., & Knutsson H. (2016). Cluster failure: Why fMRI inferences for spatial extent have inflated false-positive rates. *Proceedings of the National academy of Sciences*, *113*(28), 7900–7905. https://doi.org/10.1073/pnas.1602413113

Forman S. D., Cohen J. D., Fitzgerald M., Eddy W. F., Mintun M. A., & Noll D. C. (1995). Improved assessment of significant activation in functional magnetic resonance imaging (fMRI): Use of a cluster-size threshold. *Magnetic Resonance in Medicine*, *33*(5), 636–647. https://doi.org/10.1002/mrm.1910330508

Friston K. J., Frith C. D., Liddle P. F., & Frackowiak R. S. J. (1991). Comparing functional (PET) images: The assessment of significant change. *Journal of Cerebral Blood Flow and Metabolism*, *11*(4), 690–699. https://doi.org/10.1038/jcbfm.1991.122

Friston K. J., Worsley K. J., Frackowiak R. S. J., Mazziotta J. C., & Evans A. C. (1994). Assessing the significance of focal activations using their spatial extent. *Human Brain Mapping*, *1*(3), 210–220. https://doi.org/10.1002/hbm.460010306

Genovese C. R., & Wasserman L. (2006). Exceedance control of the false discovery proportion. *Journal of the American Statistical Association*, *101*(476), 1408–1417. https://doi.org/10.1198/016214506000000339

Glasser M. F., Sotiropoulos S. N., Wilson J. A., Coalson T. S., Fischl B., Andersson J. L., Xu J., Jbabdi S., Webster M., Polimeni J. R., Van Essen D. C., & Jenkinson M. (2013). The minimal preprocessing pipelines for the Human Connectome Project. *NeuroImage*, *80*, 105–124. https://doi.org/10.1016/j.neuroimage.2013.04.127

Goeman J. J., Hemerik J., & Solari A. (2021). Only closed testing procedures are admissible for controlling false discovery proportions. *The Annals of Statistics*, *49*(2), 1218–1238. https://doi.org/10.1214/20-AOS1999

Goeman J. J., Meijer R. J., Krebs T. J. P., & Solari A. (2019). Simultaneous control of all false discovery proportions in large-scale multiple hypothesis testing. *Biometrika*, *106*(4), 841–856. https://doi.org/10.1093/biomet/asz041

Goeman J. J., & Solari A. (2011). Multiple testing for exploratory research. *Statistical Science*, *26*(4), 584–597. https://doi.org/10.1214/11-STS356

Gorgolewski K. J., Varoquaux G., Rivera G., Schwarz Y., Ghosh S. S., Maumet C., Sochat V. V., Nichols T. E., Poldrack R. A., Poline J. -B., Yarkoni T., & Margulies D. S. (2015). NeuroVault.org: A web-based repository for collecting and sharing unthresholded statistical maps of the human brain. *Frontiers in Neuroinformatics*, *9*, 8. https://doi.org/10.3389/fninf.2015.00008

Hayasaka S., & Nichols T. E. (2003). Validating cluster size inference: Random field and permutation methods. *NeuroImage*, *20*(4), 2343–2356. https://doi.org/10.1016/j.neuroimage.2003.08.003

Jaffe A. E., Murakami P., Lee H., Leek J. T., Fallin M. D., Feinberg A. P., & Irizarry R. A. (2012). Bump hunting to identify differentially methylated regions in epigenetic epidemiology studies. *International Journal of Epidemiology*, *41*(1), 200–209. https://doi.org/10.1093/ije/dyr238

Jenkinson M., Beckmann C. F., Behrens T. E.J., Woolrich M. W., & Smith S. M. (2012). FSL. *NeuroImage*, *62*(2), 782–790. https://doi.org/10.1016/j.neuroimage.2011.09.015

Katsevich E., & Ramdas A. (2020). Simultaneous high-probability bounds on the false discovery proportion in structured, regression and online settings. *The Annals of Statistics*, *48*(6), 3465–3487. https://doi.org/10.1214/19-AOS1938

Marcus R., Eric P., & Gabriel K. R. (1976). On closed testing procedures with special reference to ordered analysis of variance. *Biometrika*, *63*(3), 655–660. https://doi.org/10.1093/biomet/63.3.655

Nichols T. E. (2012). Multiple testing corrections, nonparametric methods, and random field theory. *NeuroImage*, *62*(2), 811–815. https://doi.org/10.1016/j.neuroimage.2012.04.014

Ogawa S., Tank D. W., Menon R., Ellermann J. M., Kim S. G., Merkle H., & Ugurbil K. (1992). Intrinsic signal changes accompanying sensory stimulation: Functional brain mapping with magnetic resonance imaging. *Proceedings of the National Academy of Sciences of the United States of America*, *89*(13), 5951–5955. https://doi.org/10.1073/pnas.89.13.5951

Poline J. B., Worsley K. J., Evans A. C., & Friston K. J. (1997). Combining spatial extent and peak intensity to test for activations in functional imaging. *NeuroImage*, *5*(2), 83–96. https://doi.org/10.1006/nimg.1996.0248

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

Sommerfeld M., Sain S., & Schwartzman A. (2018). Confidence regions for spatial excursion sets from repeated random field observations, with an application to climate. *Journal of the American Statistical Association*, *113*(523), 1327–1340. https://doi.org/10.1080/01621459.2017.1341838

Van Essen D. C., Smith S. M., Barch D. M., Behrens T. E. J., Yacoub E., Ugurbil K., & Consortium W. -M. H. (2013). The WU-Minn Human Connectome Project: An overview. *NeuroImage*, *80*, 62–79. https://doi.org/10.1016/j.neuroimage.2013.05.041

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

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

Woolrich M. W., Ripley B. D., Brady M., & Smith S. M. (2001). Temporal autocorrelation in univariate linear modeling of FMRI data. *NeuroImage*, *14*(6), 1370–1386. https://doi.org/10.1006/nimg.2001.0931

Worsley K. J., Evans A. C., Marrett S., & Neelin P. (1992). A three-dimensional statistical analysis for CBF activation studies in human brain. *Journal of Cerebral Blood Flow & Metabolism*, *12*(6), 900–918. https://doi.org/10.1038/jcbfm.1992.127

Worsley K. J., Marrett S., Neelin P., Vandal A. C., Friston K. J., & Evans A. C. (1996). A unified statistical approach for determining significant signals in images of cerebral activation. *Human Brain Mapping*, *4*(1), 58–73. https://doi.org/10.1002/(SICI)1097-0193(1996)4:1¡58::AID-HBM4¿3.0.CO;2-O

Yannakakis M. (1981). Node-deletion problems on bipartite graphs. *SIAM Journal on Computing*, *10*(2), 310–327. https://doi.org/10.1137/0210022