# Exploring new methods for statistical inference in fMRI data analysis: The spatial specificity paradox

Weeterings, Bas

Universiteit
Leiden
The Netherlands

# Exploring New Methods for Statistical Inference in fMRI data-analysis

The Spatial Specificity Paradox.

B.A.A. Weeterings

Thesis advisor: Dr. W. Weeda. Methodology and Statistics, Department of Psychology, Leiden University

Defended on June 14, 2023

**MASTER THESIS**
**STATISTICS AND DATA SCIENCE**
**UNIVERSITEIT LEIDEN**

**Abstract**

Brain activity in fMRI studies is represented by voxels; units of graphic information defining a small location in the brain. In a typical case, the brain is visualized using somewhere around 200.000 voxels. To measure activity every location or voxel is tested individually, with every voxel using a separate hypothesis test; this leads to a massive multiple testing problem. One way this problem is solved is by Bonferroni-like corrections on single voxels, however Bonferroni is notorious for it's conservativeness (Samuel-Cahn, 1996). Instead of correcting for every test at the voxel level, one can also test groups (called clusters) of voxels. Hypothesis-testing on clusters reduces the multiple testing problem by accept- or rejecting entire clusters, but leads to a new problem known as the 'spatial specificity paradox': inference on the voxel level accurately locates activation at the cost of having low power for each test, whereas inference on the cluster level has more power but cannot localize activation any more accurate than "there is at least one voxel active in this cluster". Recently a solution called All-Resolutions Inference (ARI) was developed based on closed-testing to tackle this problem (Rosenblatt, Finos, Weeda, Solari, & Goeman, 2018). This method offers one way to quantify activation within clusters, without losing too much power. This project aims to assess and compare the quality of these new methods using simulation studies and real data applications.

**Keywords:** True Discovery Proportion, Permutation Test, Multiple Testing, Selective Inference, fMRI Cluster Analysis.

# Contents

# 1 Introduction

## Classic Inference for fMRI-data; Voxels, $Z$-scores and the BOLD Signal

An fMRI experiment typically involves multiple subjects that are measured for a prolonged time period. A fMRI-scan measures what's called the blood-oxygen-level-dependent (BOLD) signal at each location (also called a voxel) across time. In fMRI-analysis the BOLD signal is a reflection of changes in deoxyhemoglobin, driven by changes in blood flow and oxygenation which are coupled to underlying neuronal activity via a process termed neurovascular coupling (Hillman, 2014). That's in short the reward from a fMRI-scan; a large collection of BOLD signals spread-out over the brain across time. If the researcher is interested only in some subsection of the brain we can limit inference to only voxels inside the region of interest, but in any case the end-result is always a large data set with a BOLD observation per subject per voxel per time point. Via multiple processing steps these data are aggregated into single $z$-scores per voxel. Being outside the scope of this article it's good awareness that BOLD estimation and in particular aggregation of the BOLD signals is a complex process in and of itself all happening prior to cluster inference. Voxels and the BOLD signal and serve as pre-knowledge before we can start working with aggregated data and look at inference; how can the activation values be used to retrieve brain areas (clusters) that make sense and be helpful to practitioners. Typically we compare the BOLD signals between two conditions or stimuli (within subject), or two conditions between groups (between subject). Either way, we apply a voxel-wise statistical test resulting in a three-dimensional map of voxel-wise Z-values.

## The Problem of Spatial Specificity in Neuroimaging

The brain map is made by storing voxel-wise Z-values stored a 3D-array. The three dimensions represent the MNI-coordinates, which indicate the location of the voxel or cluster. Classically we define a voxel-wise null hypothesis stating that the voxel is not active i.e. the BOLD signal is not related to the experimental stimulus while ignoring the fact that image data are really correlated. Instead of treating the image as a bag of voxels we can threshold the data and hypothesis-test at the cluster-level. By accept- or rejecting entire clusters, a much smaller number of hypothesis-tests is required but leads to the question of what threshold will show us signal, otherwise known as the spatial specificity paradox; a high threshold leads to good spatial specificity, but poor power (high risk of false negatives), while a low threshold will lead to poor specificity (high risk of false positives), but good power. In summary, inference on the voxel-level accurately locates activation at the cost of having low power for each test, whereas inference on the cluster-level has more power but cannot localize activation any more accurate than "there is at least one voxel active in this cluster". Recently new methods for cluster-inference were developed; a parametric method called All-Resolutions Inference (ARI) by Rosenblatt et al. (2018) as well as a non-parametric method called permutation-based All-Resolutions Inference (pARI) Andreella, Hemerik, Finos, Weeda, and Goeman (2023) forming the leading cause for this article; how do these methods for fMRI-analysis differ in the amount of *truly* active voxels they report?

## Study Design

The projects' aim is to analyse and compare spatial specificity in fMRI data-analysis for new cluster inference methods, specifically All-Resolutions Inference (ARI), permutation-based All-Resolutions Inference (pARI) as well as a more simple procedure counting active voxels using a more classical multiple test correction. Readers interested in more detail on the inference framework are recommended the following methodology section. Ultimately the goal is to compare

methods on their ability to detect active voxels i.e. true discovery proportions as measure for spatial specificity. The research can be divided into two main parts. For the first part methods are used in real data applications and for the second part methods are used on a simulation. The useful thing about simulation is it allows the activation in the clusters to be exactly known. The design can be realistic, for example by pre-selecting a desired area of the brain or unrealistic, for example by activating a square in the brain. The latter clearly will never occur in nature, but might show where different methods perform better or break down given circumstance. For example we may have the expectation for the permutation-based method to be more reliant on the amount of available data i.e. participants in the study, a parameter we can control in our simulation environment. For the reader interested to try fMRI-data analysis or simply in further details the full project; code with a manual for the pipeline is available on GitHub and the Open Science Framework (OSF).

## 2 Methodology

### Theory and Algorithm for the All-Resolutions Inference Framework (ARI)

Let the brain $B$ be a collection of $m$ voxels. We assume that a test statistic for activation can be calculated for each voxel which can be converted into a voxel-wise p-value orderable such that $p_1 \leq p_2 \leq \cdots \leq p_m$. Let voxel set $A \subseteq B$ be the unknown set of all truly active voxels. Finally, denote $\mathcal{S} = 2^B$ as the collection of all $|\mathcal{S}| = 2^m$ possible voxel sets. For every subset of voxels $S \subseteq \mathcal{S}$ the number of true discoveries in $S$ is $\pi(S) = |B \backslash A|$ (Chen, Goeman, Krebs, Meijer, & Weeda, 2022) and if $S$ is non-empty, the corresponding activation of all truly active voxels (TDP) is denoted by Rosenblatt et al. (2018):

$$\pi(S) = a(S)/|S| \tag{1}$$

The TDP informs about the extent of spatial activation within $S$. Following Rosenblatt et al. (2018) we say that there is good spatial localisation of the signal in S if the TDP is high enough. ARI uses the methods of Goeman and Solari (2011) and Goeman, Meijer, Krebs, and Solari (2019) to construct lower confidence bounds $\overline{\pi}(S)$ for the set-wise proportion of active voxels. The TDP lower confidence bound $\overline{\pi}(S)$ was derived by Goeman, Meijer, Krebs, and Solari (2019) using the closed testing procedure Marcus, Eric, and Gabriel (1976) with Simes (1986) local tests. It is given by (1) for non-empty subsets $S \subseteq B$ where $\overline{\pi}(S)$ is the lower confidence bound given by:

$$\mathbb{P}(\forall S \in \mathcal{S} : \overline{\pi}(S) \leq \pi(S)) \geq 1 - \alpha \tag{2}$$

It's worthwhile to mark that ARI is more powerful for larger sets in comparison to small ones. ARI may give large values for $\overline{\pi}(S)$ even if no voxel in $S$ is significant when using Hommel as correction for multiple testing error (Hommel, 1988).

### Local Simes Test and Inequality for Cluster Error Control

ARI's error control is guaranteed under the assumption of the Simes inequality (Rosenblatt et al., 2018). To derive (1) we start by defining for every voxel set $S$ the null hypothesis:

$$H_S : a(S) = 0 \tag{3}$$

$H_S$ is the Random Field Theory (RFT) null hypothesis for cluster-wise inference: rejecting $H_S$ indicates that there is at least one active voxel in $S$. We test every $H_S$ with the Simes test

(Simes, 1986), rejecting $H_S$ at level $\alpha$ if and only if $p_s \leq \alpha$, where

$$p_S = \min_{1 \leq i \leq |S|} \frac{|S|}{i} p_{(i:S)} \tag{4}$$

and $p_{(i:S)}$ is the $i$th smallest p-value among voxels in $S$. The Simes test is valid if $P(p_S \leq \alpha) \leq \alpha$ for all $S$ for which $H_S$ is true. For the validity of the ARI procedure as a whole, however, we only need this to hold for the set $S = B \backslash A$ of all non-active voxels, the largest set for which $H_S$ is true. We assume that:

$$\mathbb{P}(p_{B \backslash A} \leq \alpha) \leq \alpha \tag{5}$$

The Simes inequality is the most important assumption for ARI and is frequently used in multiple testing literature; Hommel (1988), Hochberg (1988) and Benjamini and Hochberg (1995) procedures make the same assumption.

### Controlling the FWER using Closed Testing

The tests for $2^m$ hypotheses $H_S$ must be corrected for multiple testing. A powerful method for this is closed testing (Marcus et al., 1976). In closed testing a hypothesis $H_S$ is rejected if and only if $H_S$ is rejected for all $I \supseteq S$. (Goeman, Meijer, Krebs, & Solari, 2019) have proven that closed testing with Simes tests rejects a hypothesis $H_S$ if and only if:

$$\min_{1 \leq i \leq |S|} \left\{ \frac{h}{i} p_{(i:S)} \right\} \leq \alpha \tag{6}$$

We can calculate a FWER-adjusted p-value, $p_S$, for any region hypothesis $H_S$. Such adjusted p-values are defined as the smallest $\alpha$-level that allows rejection of $H_S$ within the closed testing procedure. The useful duality holds that $p_S \leq \alpha$ if and only if $\overline{\pi}(S) > 0$.

### Calculating the Proportion of Truly Active Voxels

Lower confidence bounds for the percentage of truly active voxels (TDP) follow from the result of the closed testing procedure given by Goeman and Solari (2011): if for some $K \geq 0$, $H_I$ is false for all subsets $I \subseteq S$ with $|I| = |S| - k$, then there is at least one active voxel in each such $I$, and therefore there are at least $k - 1$ active voxels in $S$. By setting $\overline{\pi}(S) = \overline{\alpha}(S)/|S|$ we get via the FWER-control of the closed testing procedure and (2) it follows:

$$\mathbb{P}(\forall S \in \mathcal{S} : \overline{a}(S) \leq a(S)) \geq 1 - \alpha \tag{7}$$

Which if we translate to the case of the Simes test returns:

$$\overline{\alpha}(S) = \min \left\{ 0 \geq k \geq |S| : \min_{1 \leq i \leq |S| - k} \frac{h}{i} p_{(i+k:S)} > \alpha \right\} \tag{8}$$

It has been shown by Goeman, Meijer, Krebs, and Solari (2019) that $\overline{\alpha}(S)$ is always at least as large as the naive bound that simply counts the number of FWER-significant voxels in $S$ and often much larger especially when the number of voxels is large.

### Permutation-Based All-Resolutions Inference

Permutation-based All-Resolutions Inference (pARI) introduces the concept of closed testing based on a critical vector. The Simes-based critical vector for ARI is $l_i = i\alpha/h$, where $h$ is a

random variable that can be calculated using the short-cut defined by Goeman, Meijer, Krebs, and Solari (2019). It is the largest set size of a subset of the brain not rejected by the Simes test. However the Simes-based critical values $(l_1, \ldots, l_m)$ can be overly strict when p-values are positively correlated. The critical vector in pARI leads to a less conservative test while controlling the FWER i.e. such that the number of false positives are below 5%. The permutation method takes into account the dependence structure and marginal distributions of the p-values. It is not required for the null p-values to be uniformly distributed. Instead we require that the null p-values are exchangeable with the corresponding post-permutation p-values. Consider a group of permutations or sign-flipping transformations or any other data transformation that preserves the distribution of the test statistics under the null hypothesis. The method is based on $w$ random permutations or sign-flipping transformations. Let $p_1^1, \ldots, p_m^1$ be the p-values for the real data, and for every $2 \leq j \leq w$ let $p_1^j, \ldots, p_m^j$ be the p-values obtained for the $j$-th random permutation of the data. The permutation-based critical vector $l(\lambda_a)$ by Andreella et al. (2023) is defined as:

$$\lambda_a = \sup\{w^{-1} | 1 \leq j \leq w : p_i^j \geq l_i(\lambda) \ \forall i \in B| \geq 1 - \alpha\} \tag{9}$$

satisfying (8) such that we may keep FWER-control. Permutation-based All-Resolutions Inference uses a iterative approximation to approach $l(\lambda_a)$ defined by Hemerik, Solari, and Goeman (2019) in order to maintain a balance between power and a realistic computation time. Because of the increase in power, the expectation would be for pARI to result in higher activity (TDP) within each cluster at the cost of longer computation time. We should also notice that pARI is dependent on the amount of permutations possible to be sampled whereas ARI is not. As such pARI requires at least some sufficient amount of participants in a given study before it provides an accurate approximation of within-cluster activity.

### Voxel-Count Procedure

Another way to correct for the FWER, namely the most conservative way, would be to simply count the number of active voxels above some significant activation level alpha and correct for the number of tests. Essentially this is like a bonferroni test where the number of tests is equal to the non-zero, non-masked p-values. This is almost but not entirely the length of the brain. We can store the values in one large array and transform them back to the original 3D space (i.e. Pmap) as well as order the p-values and check if each pass activation level alpha. Because of the test's conservativeness, it's probably best looked at as a baseline or lower bound for the TDP, rather then a viable method for inference. The mathematics of the procedure are briefly explained in Algorithm 1. For simplicity the formatting of data is omitted, readers interested in handling data-structures may be interested in the code included with the pipeline for more detail.

**Algorithm 1** Count procedure

**Require:** $p_1 \leq p_2 \cdots \leq p_m$
  $m \leftarrow || < p_1, \ldots, p_m > ||$ ▷ size of the multiple testing problem ≈ length of the brain
  $\alpha_{adj} \leftarrow \frac{\alpha}{m}$ ▷ using a bonferroni-like correction
  **function** CountBrainCluster(*Pmap*, *Statmap*, $\alpha$, *clusters*)
    **for** $\forall$ clusters $C$ **do**
      $|C_i| \leftarrow || < p_1, \ldots, p_j > ||$
      $Z_i \leftarrow \max(\forall Z \in C_i)$
      $\#discoveries_i \leftarrow 0$
      **for** $\forall\ p \in$ cluster $C_i$ **do**
        **if** $p_j < \alpha_{adj}$ **then** ▷ FDP
          $discoveries_i \leftarrow discoveries_i + 1$
        **end if**
      **end for**
      $active\ proportion_C \leftarrow \frac{\#discoveries_i}{|C_i|}$ ▷ TDP
      $true\ null_C \leftarrow |C_i| - \#discoveries_i$
      $false\ null_C \leftarrow |C_i| - true\ null_C$
    **end for**
    **return** *clusters*
  **end function**

The algorithm is build such that it can be used to accept different sets of p-values (reconstructed into the appropriate Pmap) and activation level alpha. As a extra note the false discovery rate (FDR) and Benjamin-Hoch method (Benjamini & Hochberg, 1995) were briefly investigated during the project and were not included for further analysis. The FDR resulted in a upper-bound of sorts for the TDP, which for all but one actual dataset reached full activation ($TDP \mathrel{\widehat{=}} 100\%$) too quickly even for clusters realistically too small to be meaningful. For the Benjamin-Hoch method the behaviour looks similar to All-Resolution Inference, while taking much longer to compute, a problem meant to be solved by ARI in the first place. Finally outside of adjusting activation level alpha for the bonferroni-correction, the conventional $\alpha = 0.05$ is used for all other analysis.

### Data Preliminaries and Pre-processing

All fMRI-data used can be found in the fMRIdata package accessed in R or from GitHub. It provides a collection of pre-processed fMRI data-sets hosted from OpenNEURO; an online network where many fMRI-data-sets are publicly available. Specifically the data-sets used for the purposes of this article are: the Auditory data (ds000158); research regarding the human voice areas specifically spatial organisation, the Arrow data (ds000102); performance during a flanker task, the Food data (ds000157); a picture viewing task, and finally the Rhyme data (ds000003); a rhyme judgement experiment. Implementation of ARI has been provided in the R environment with the R package hommel (Goeman, Meijer, & Krebs, 2019), and specifically for fMRI data analysis, the R package ARIbrain which can be installed directly within R or found on GitHub. Likewise the package pARI can also be installed directly or from GitHub.

### Analysis Pipeline

In order to easily apply the analysis for multiple data-sets each dataset (real or simulated) passes equally (has the same parameter tuning unless specified) through the analysis pipeline

in R. The pipeline provides a document-standard to perform All Resolutions Inference (ARI) and permutation-based inference (pARI) and can be use any set of fMRI-data as input when data from participants are provided in NIfTI-format (a term for an image or data stored in a 3D format). Such a file representing a participant is also known as a cope or copes in plural. For parameter selection there are two main parameters that can be adjusted; the alpha and cluster threshold-value. By default these will be the typical values of 0.05 and 3.2 respectively, these being the most commonly used values for standard analysis. Both values are held constant within any particular analysis, but can be changed to perform the same analysis with different parameter tuning. The pipeline contains a data folder where the user should present his or her copes and a mask as NIfTI files, and a stats folder where output (stat- and cluster-maps) are stored by the pipeline automatically. A statmap is a description for the data-file containing the voxel-wise Z-values in 3D space. A clustermap holds the 3D data after thresholding. The full process from masking, mapping and threshold is also illustrated by Figure 2.1. The pipeline reads any listed copes and stores them in R. If no copes are given, a simple simulation can be ran to create some copes as well. Using the listed copes and a mask as input a function is called to return the map of Z (and optionally p)-values to the user in NIfTI format. The maps are stored in the stats folder. From the Z-values clusters are computed by thresholding the data, by default with $Z > 3.2$, the exception is during the simulation study where the threshold-value varies as parameter across simulations. Clusters are used as input for pARI, ARI, as well as the count procedure. Each method returns the respective true null, false null and active proportions per cluster. Notice that for each method only the active proportion of voxels within the clusters changes, while the clusters (size and location i.e. statmap) used as input are equal. The output-table also presents the size, highest Z-value, the location or MNI-coordinates and amount of clusters. The MNI-coordinates refer to a system of coordinates for indexing voxels in a volume applicable to anatomical atlases. Finally the pARI and ARI-objects created in R are used to store the true discovery proportion (TDP) brain maps in the stats folder.

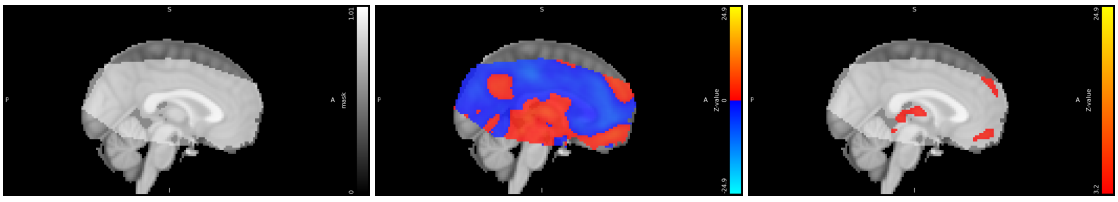

Figure 2.1: Process of masking and thresholding a brain map. On the left the mask; a binary operator based on aggregation of the original BOLD signals. In the middle raw Z-values per voxel contained in the mask. On the right data after thresholding ($Z > 3.2$).

### All-Resolution Inference using *pARI* and *ARIbrain*

pARI is the package developed to compute the permutation-based All-Resolution Inference (ARI) method. Permutation-based ARI does not assume any distribution of the null distribution of the p-values. It needs to satisfy the exchangeability assumption as all permutation-based methods (Pesarin & Salmaso, 2010). Like parametric ARI this method computes simultaneous lower confidence bounds for the number of true discoveries. pARIbrain returns the lower bounds of true discoveries for each cluster and allows for circular analysis controlling for the multiplicity of inferences. pARI can be found on zenodo. ARIbrain is the package for All-Resolution Inference (ARI) in neuroscience. ARI is used to compute lower bounds for proportions of active voxels (or any other spatially located units) within given clusters.

### Neuroimaging using *FSLeyes*

FSLeyes is a wxPython application for visualising neuroimaging data. The application is hosted on the FSLeyes Gitlab. In the pipeline we make use of FSLeyes twice: once to inspect the statmap, and once more to visualise the clusters. The statmap gives an overview of how Z-values are distributed across the brain during the experiment. The clustermap shows active brain-regions given above a specified threshold. Note the clustermap does not provide any information regarding truly active voxels, only the cluster size and location, precisely without showing how many voxels were active inside a particular cluster. However the clustermap is insightful as comparison material, for example are higher TDP's found more frequently for a particular method and cluster size when given particular circumstances.

### Set-up for fMRI-data Simulation

Inside the analysis pipeline a function is included to create simulations. Input for the analysis is the amount of simulations (or when unspecified ten copes are created by default) and a signal and a noise parameter (with the default signal-to-noise ratio being 1:1). A list is created where every cope is represented by one large array. This large array is sorted back into the 3D space of the brain, which for the sake of simplicity receives the same dimensions as the data-sets from the fMRIdata package. The package neuRosim is used to create spatial noise and fill the array with correlated data. For this purpose a Gaussian random field is used, with a full-width half-maximum (FWHM) of 3; using as rule of thumb for an appropriate FWHM three times the size of, in our case simulated "1mm" voxel brain. After spatial noise is created the signal is added to the selected rows inside the brain. For the purposes of this article we do not care much for the shape itself, instead we care for the (detected) proportion of active voxels for different methods while varying threshold-values and signal-to-noise ratios. Each cope is again written into NIfTI format, after which the analysis pipeline is followed in equal fashion to running a regular dataset, except this time no real data (copes) were needed to be input.

## 3 Data Application

### Auditory Data

Pernet et al. (2015) used auditory data to perform cluster analysis for localization of the voice-sensitive 'temporal voice areas' (TVA) of the human auditory cortex. The contrasts for the experiment were made between forty eight-second long blocks of vocal (20 blocks) versus non-vocal (20 blocks) sounds from Belin, Zatorre, Lafaille, Ahad, and Pike (2000). As the data have been pre-processed (much like the other analysed data-sets) we can quite easily use the auditory data as input for the pipeline to apply ARI, pARI, and the voxel count procedure. Figure 3.1 shows the distribution of Z-values across the brain. The slices are roughly through the middle of the brain and in respective order from left to right show the sagittal, coronal and horizontal plane in radiological orientation.

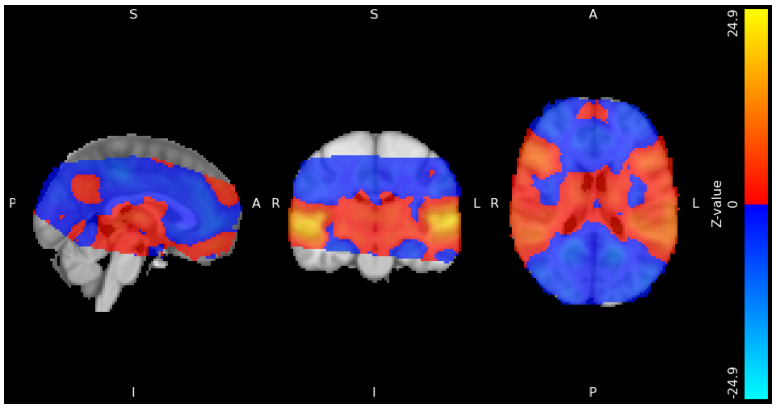


Figure 3.1: Activation map of the vocal vs. non-vocal contrast for the Auditory data. Colors indicate Z-values per individual voxel.

Figure 3.2 shows clusters after thresholding for activation map 3.1. Detailed results are presented in Table 3.1. Note that only clusters with a size of hundred voxels or more are shown; considering how easily small clusters are found in a simulation where there is no signal (and therefore no clusters) we conclude small clusters ($|C| < 100$) are too likely to be a fluctuation that gets picked up by ARI and pARI, and offer no relevance and neither can be interpreted in a meaningful manner.

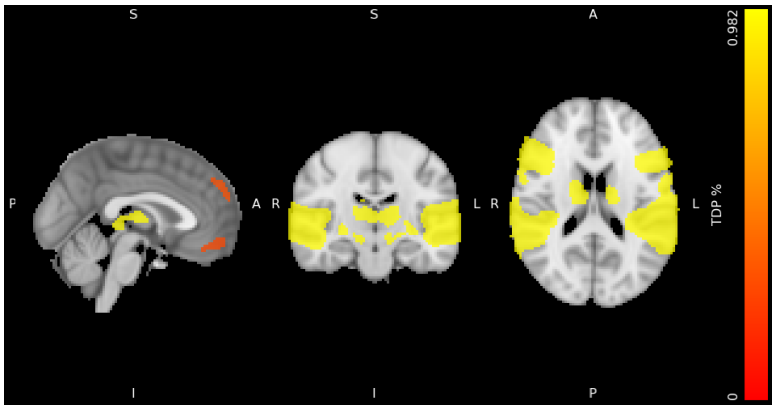


Figure 3.2: Activation map after cluster-forming threshold ($Z > 3.2$) for the Auditory data using permutation-based All-Resolutions Inference. Colors indicate the TDP for each of the clusters.

Table 3.1: Cluster inference for Auditory data: comparison of true discovery proportions by method for clusters $|C| > 100$ identified with threshold $Z > 3.2$. MNI-coordinates indicate location of peak activity $Z_{max}$ within-cluster.

| *Cluster* | *Size* | *% active* | | | *MNI* | | | *Statistic* |
|---|---|---|---|---|---|---|---|---|
| $C$ | $\|C\|$ | $\bar{\pi}(C)_{Count}$ | $\bar{\pi}(C)_{ARI}$ | $\bar{\pi}(C)_{pARI}$ | $x$ | $y$ | $z$ | $Z_{max}$ |
| 9 | 31427 | 0.6288 | 0.9181 | 0.9502 | 76 | 58 | 38 | 24.508 |
| 8 | 442 | 0.5430 | 0.6493 | 0.6742 | 71 | 61 | 62 | 10.100 |
| 7 | 269 | 0.0706 | 0.2082 | 0.2565 | 43 | 90 | 56 | 5.687 |
| 6 | 251 | 0.0040 | 0.1514 | 0.2231 | 44 | 88 | 30 | 5.332 |

From the auditory data nine clusters were discovered in total. When discussing clusters of significant size ($|C| \geq 100$) the data show one large cluster ($|C| = 31427, Z_{max} = 24.51$) with relatively high activity ($\bar{\pi}(C) = 0.950, 0.918, 0.628$ for pARI, ARI and the count procedure respectively) and three smaller clusters varying in size ($251 \leq |C| \leq 442$) and activity ($0.004 \leq \bar{\pi}(C) \leq 0.674$). Based on the results from Table 3.1 pARI outputs higher activity i.e. TDP per cluster across the board. Clearly the count procedure is less effective in detecting activity compared to ARI and pARI by a substantial margin.

### Arrow Data

For the arrow data functional imaging data were acquired from 26 healthy adults while they performed a slow event-related Eriksen Flanker task (Kelly, Uddin, Biswal, Castellanos, & Milham, 2008). In the flanker task arrows are shown and the irrelevant information (the flanker) has to be ignored (Willemssen, Hoormann, Hohnsbein, & Falkenstein, 2004). Participants had to indicate the direction of a central arrow in an array of 5 arrows. In congruent trials the flanking arrows pointed in the same direction as the central arrow (e.g., >>>>>), while in more demanding incongruent trials the flanking arrows pointed in the opposite direction (e.g., >><>>). Figure 3.3 shows activation values across the brain for the particular task. Figure 3.4 shows the clusters from the activation-values. Next table 3.2 shows basic inference per cluster and active proportions for each method individually.

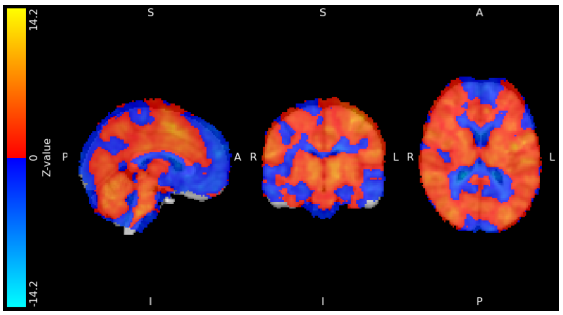


Figure 3.3: Activation map of the congruent vs. incongruent contrast for the Arrow data. Colors indicate Z-values per voxel.

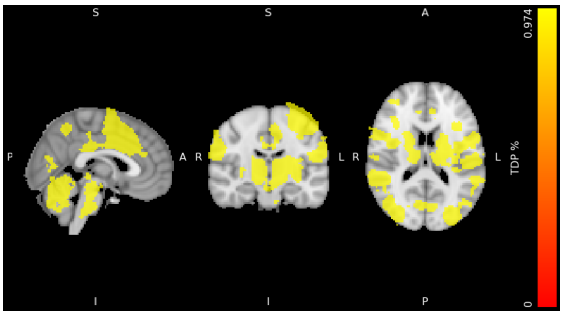


Figure 3.4: Activation map after cluster-forming threshold ($Z > 3.2$) for the Arrow data. Colors indicate the TDP for each of the clusters.

Table 3.2: Cluster inference for Arrow data: comparison of true discovery proportions by method for clusters $|C| > 100$ identified with threshold $Z > 3.2$. MNI-coordinates indicate location of peak activity $Z_{max}$ within-cluster.

| *Cluster* | *Size* | *% active* | | | *MNI* | | | *Statistic* |
|---|---|---|---|---|---|---|---|---|
| $C$ | $\|C\|$ | $\bar{\pi}(C)_{Count}$ | $\bar{\pi}(C)_{ARI}$ | $\bar{\pi}(C)_{pARI}$ | $x$ | $y$ | $z$ | $Z_{max}$ |
| 24 | 69947 | 0.0933 | 0.8044 | 0.9334 | 36 | 38 | 26 | 13.977 |
| 23 | 424 | 0.0000 | 0.0000 | 0.0000 | 64 | 91 | 46 | 4.755 |
| 22 | 140 | 0.0000 | 0.0000 | 0.0143 | 23 | 48 | 36 | 6.761 |

From the arrow data twenty-four clusters were discovered. After small clusters are omitted, the data show one large cluster ($|C| = 699947, Z_{max} = 13.98$) with high activity ($\bar{\pi}(C)_{pARI} = 0.934, \bar{\pi}(C)_{ARI} = 0.804, \bar{\pi}(C)_{Count} = 0.093$). Notice even for the second and third largest cluster the highest statistical value ($Z_{max} = 4.76$) turns out to be too low to really find much activity

within these clusters. Also notice it would be possible to break down the large (and perhaps too dominant) cluster into smaller clusters by increasing the threshold from 3.2 to for example 4.2 if doing so happens to be more meaningful. Ultimately changing the cluster threshold would depend on the researcher or healthcare provider's goals and falls out of scope for this project. Finally the percentage active for the count procedure stays low ($\bar{\pi}(C) = 0.093$), even for the large highly active cluster ($\bar{\pi}(C)_{ARI,pARI} \geq 0.804$).

## Food Data

Smeets, Kroese, Evers, and de Ridder (2013) had thirty female subjects perform a passive viewing task with blocks of food and nonfood images. The experiment investigated neuronal response based on the participant's dietary concerns. Subjects alternately viewed 24 s blocks of palatable food images (8 blocks) and non-food images (i.e., office utensils; 8 blocks) and rated attractiveness on a one through seven Likert-scale. Figure 3.5 shows overall brain activity applying the pipeline from section three. Figure 3.5 shows activation values across the brain. Table 3.3 presents detailed inference for each method based on the clustermap.

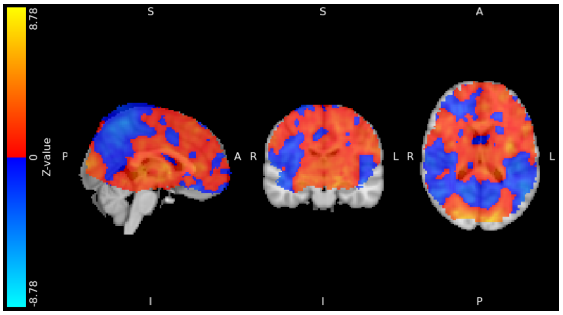


Figure 3.5: Activation map of the food vs. non-food (utensils) contrast for the Food data. Colors indicate Z-values per voxel.

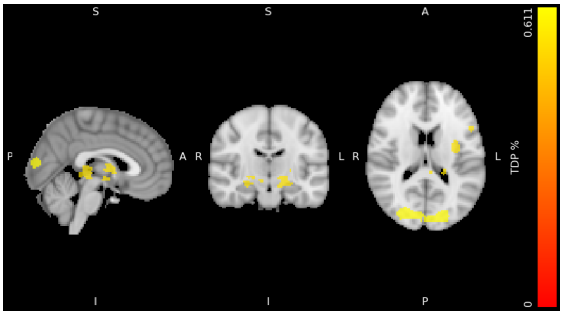


Figure 3.6: Activation map after cluster-forming threshold ($Z > 3.2$) for the Food data. Colors indicate the TDP for each of the clusters.

Table 3.3: Cluster inference for Food data: comparison of true discovery proportions by method for clusters $|C| > 100$ identified with threshold $Z > 3.2$. MNI-coordinates indicate location of peak activity $Z_{max}$ within-cluster.

| *Cluster* | *Size* | *% active* | | | *MNI* | | | *Statistic* |
|---|---|---|---|---|---|---|---|---|
| $C$ | $\|C\|$ | $\bar{\pi}(C)_{Count}$ | $\bar{\pi}(C)_{ARI}$ | $\bar{\pi}(C)_{pARI}$ | $x$ | $y$ | $z$ | $Z_{max}$ |
| 51 | 3331 | 0.0096 | 0.0852 | 0.2894 | 62 | 65 | 34 | 8.478 |
| 50 | 3023 | 0.0437 | 0.3348 | 0.4750 | 54 | 20 | 49 | 8.643 |
| 49 | 223 | 0.0000 | 0.0000 | 0.0000 | 50 | 79 | 65 | 5.081 |
| 48 | 134 | 0.0000 | 0.0000 | 0.0000 | 31 | 76 | 33 | 5.051 |

Summarising the food data, two moderately and roughly equally sized clusters were found ($3032 \leq |C| \leq 3331$) with corresponding activation values between $8.478 \leq Z_{max} \leq 8.643$. Following clusters had too low of a activation value for all three methods to detect activation ($Z_{max} \leq 5.081$). Notice pARI finds a higher active proportion, particularly in the first cluster, contrasting ARI which detects a low active proportion; pARI finds active proportions from $[0.289 \leq \bar{\pi}(C)_{pARI} \leq 0.475]$ whereas ARI's ranged from $[0.085 \leq \bar{\pi}(C)_{ARI} \leq 0.335]$. Given the small clusters pARI seems to return higher active proportions, but these data alone are not sufficient to generalise this particular behaviour into a conclusion.

### Rhyme Data

Figure 3.7 shows the statmap for an experiment where thirteen participants performed rhyming judgments on words (Xue & Poldrack, 2007). The basic paradigm throughout was a same–different judgment task using a Korean characters as stimuli. Two characters then flashed subsequently and subjects were asked to judge whether the two characters were identical or not. Applying the pipeline to the statmap results in the clusters shown in Figure 3.8. Cluster and method specific inference are presented in Table 3.4.

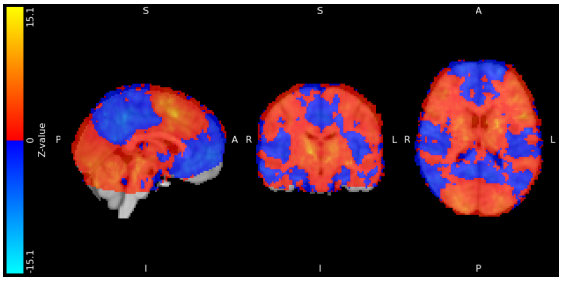


Figure 3.7: Activation map of the identical vs. non-identical contrast for the Rhyme data. Colors indicate Z-values per voxel.

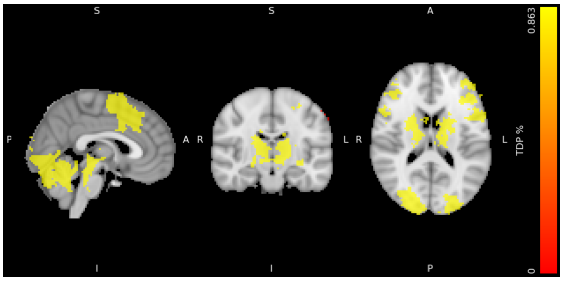


Figure 3.8: Activation map after cluster-forming threshold ($Z > 3.2$) for the Rhyme data. Colors indicate the TDP for each of the clusters.

Table 3.4: Cluster inference for Rhyme data: comparison of true discovery proportions by method for clusters $|C| > 100$ identified with threshold $Z > 3.2$. MNI-coordinates indicate location of peak activity $Z_{max}$ within-cluster.

| *Cluster* | *Size* | *% active* | | | *MNI* | | | *Statistic* |
|---|---|---|---|---|---|---|---|---|
| $C$ | $\|C\|$ | $\bar{\pi}(C)_{Count}$ | $\bar{\pi}(C)_{ARI}$ | $\bar{\pi}(C)_{pARI}$ | $x$ | $y$ | $z$ | $Z_{max}$ |
| 139 | 34115 | 0.0013 | 0.3817 | 0.6748 | 44 | 70 | 61 | 14.904 |
| 138 | 1546 | 0.0000 | 0.0000 | 0.0000 | 58 | 33 | 59 | 8.470 |
| 137 | 606 | 0.0000 | 0.0000 | 0.0000 | 31 | 61 | 63 | 7.935 |
| 136 | 158 | 0.0000 | 0.0000 | 0.0000 | 21 | 50 | 62 | 5.941 |

In summary the rhyme data show one large cluster ($|C| = 34115, Z_{max} = 14.90$) in which by estimation about 38% to 68% of the voxels are active [$0.382 \leq \bar{\pi}(C) \leq 0.675$]. Any smaller cluster ($|C| \leq 1546$) showed no activity ($Z_{max} \leq 8.470$). In this case it turns out the highest Z-statistic is not sufficiently high to find any more activity in the smaller clusters.

### Comparing ARI and pARI's True Discovery Proportions

Because the results of voxel-counting are better viewed as a conservative baseline for the true TDP, our focus here will be dedicated to ARI versus pARI. Figure 3.9 shows the active proportions for every cluster (discovered previously in section three) for ARI and pARI covering the four data-sets: auditory, arrow, food and rhyme data. Proportions for ARI are shown on the x-axis against pARI's proportions on the y-axis. The size of each dot reflects the size of the particular cluster.

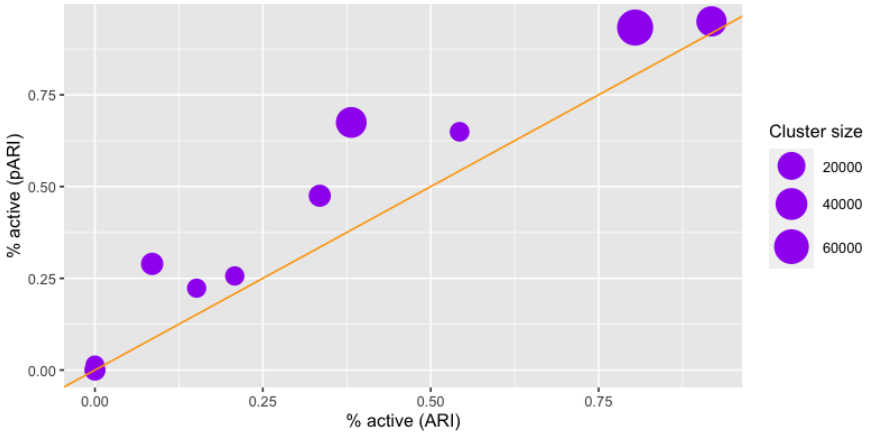


Figure 3.9: Meta-analysis of clusters comparing true discovery proportions of ARI (x-axis) versus pARI (y-axis). Clusters are based on the data applications from Sections 3.1 through 3.4. Size of the dot indicates size of the cluster.

Figure 3.9 clearly shows pARI returns higher active proportions compared to ARI. Notice in this particular research it was *always* the case that pARI returned higher discovery proportions, regardless of size; or at least the data do not show a specific relationship between size and method. Larger clusters in itself do correlate with relatively higher active proportions; for ARI the correlation between size and TDP is 76.5% ($\rho_{|C|,ARI} = 0.7654$) and slightly higher for pARI around 78.1% ($\rho_{|C|,pARI} = 0.7805$). Put differently this means we expect pARI to find slightly higher active discovery percentages compared to ARI.

# 4 fMRI-Data Simulation

## Creating the Signal

As end-product the simulation creates a square-shaped cluster inside the brain in the middle of the x-and y-axis, slightly higher up on the z-axes avoiding the amygdala or middle brain. Note this matters only visually as the cluster's location is irrelevant to ARI, pARI as well as the count procedure. The rough edge of the square does affect the random field estimate, however with the advantage that a square is much more easily created, moved, and in- or decreased in size compared to a more realistic blob-like shape one would normally find. The simulation represents ten copes, for which each spatially correlated noise is created across the brain using a Gaussian random field estimate. The signal is added to the brain (with noise) and this simulation is repeated for different signal-to-noise ratios starting at zero, five, and ten-to-one ultimately. The cluster threshold is varied in three levels using the most commonly used values of 2.3, 3.2 and 4.2. The amount of participants (i.e. copes) varies in levels of five, ten and twenty. Each scenario is the average of ten simulations or repetitions. The active proportions found in each scenario are shown in Table 4.1. Visually the activation levels are shown by Figures 4.1. 4.2 and 4.3. The left-hand side shows all Z-values, while the right-hand side is filtered to show all $|Z| \geq 2.3$ as to make the signal more distinguishable. It also roughly illustrates how many voxels (but really noise) pass our threshold-value by chance.

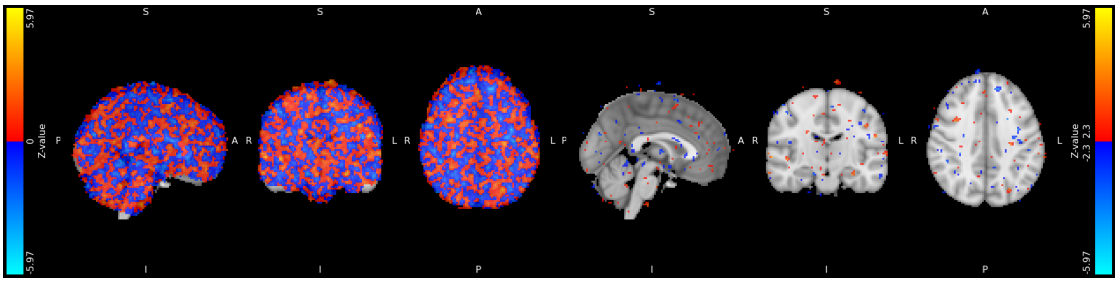

Figure 4.1: Activation map with only-noise or the brain at rest ($SNR = 0$). Spatial noise is created using a Gaussian random field with a FWHM of 3 simulating a 1mm voxel brain. The left-hand side shows original Z-values. The right-hand side is thresholded to $|Z| \geq 2.3$ to illustrate the (in this case absent) signal.

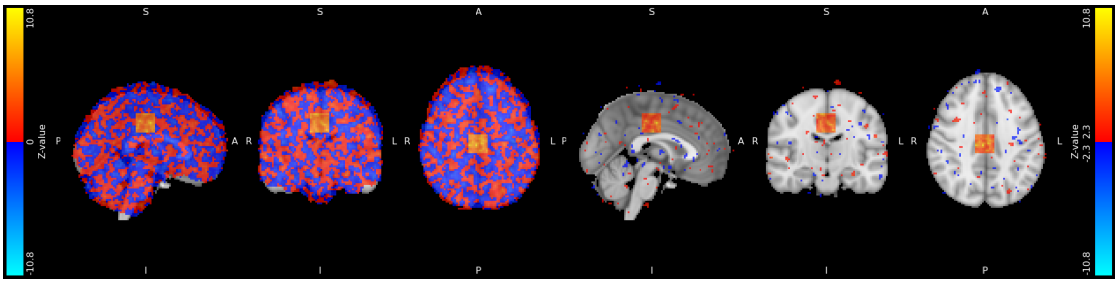

Figure 4.2: Activation map for a signal of medium strength ($SNR = 5$) with on both sides the cluster clearly distinguishable to the eye.

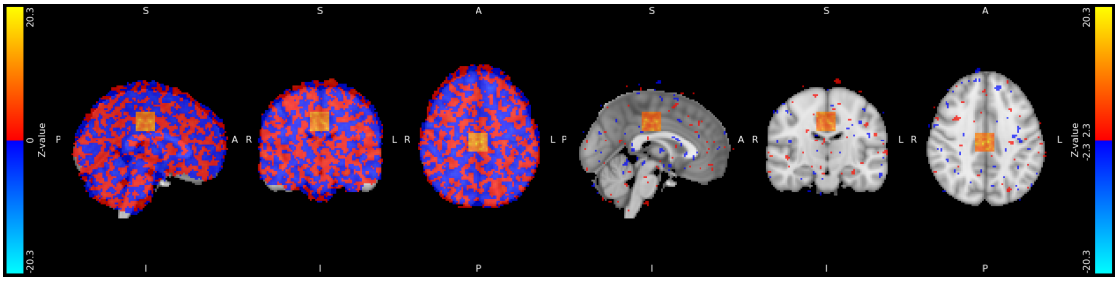

Figure 4.3: Activation map for the highest simulated activity ($SNR = 10$).

Starting from SNR zero (no signal at all or only noise) the cluster becomes increasingly more visible as the SNR increases to ten, where the squared-cluster is clearly visible and, at least expectedly, detectable for ARI and pARI. SNR 5 represents a half-way point between the only-noise and strongest signal scenarios and should still result in a detectable cluster. Note that the clustermaps for Figures 4.1, 4.2 and 4.3 have not been shown. By design it's corresponding clustermap is just the square in the middle of the brain and as such is not particularly helpful.

### Signal's Strength and Cluster-threshold's Effect on True Discovery Proportions

Figure 4.1 shows the active proportions per scenario. For the noise-only scenario the active proportions are (not unexpectedly) consistently zero regardless of threshold-level. For the strongest signal scenario ($SNR = 10$) ARI's active proportions range from $[0.781 \leq \bar{\pi}(C)_{ARI} \leq 0.790]$. Proportions for pARI are structurally higher ranging from $[0.816 \leq \bar{\pi}(C)_{pARI} \leq 0.825]$. The count procedure finds minimal activity, especially knowing the simulated cluster consists entirely of activity; $[0.064 \leq \bar{\pi}(C)_{Count} \leq 0.065]$. Notice these intervals are also an immediate reflection of the (activation) sensitivity to the threshold value, with the biggest range i.e. maximum change in discovery due to the threshold value being 0.090% for pARI (equal to the interval's bandwidth) and 0.085% for ARI. The count procedures value varies even less, but this is obfuscated by the fact that the count procedure's active proportions are so much lower to begin with.

Table 4.1: Analysis of simulation: active proportions varying SNR, cluster-threshold and number participants $N$. Results represent averages of ten simulations or repeats.

| | $\bar{\pi}(C)_{Count}$ | | | $\bar{\pi}(C)_{ARI}$ | | | $\bar{\pi}(C)_{pARI}$ | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| $SNR$ / $u_{clus}$ | 0 | 5 | 10 | 0 | 5 | 10 | 0 | 5 | 10 | $N$ |
| 2.3 | 0 | 0 | 0.0007 | 0 | 0 | 0.0211 | 0 | 0 | 0.0007 | |
| 3.2 | 0 | 0 | 0.0007 | 0 | 0 | 0.0216 | 0 | 0 | 0.0007 | 5 |
| 4.2 | 0 | 0 | 0.0008 | 0 | 0 | 0.0218 | 0 | 0 | 0.0008 | |
| 2.3 | 0 | 0.0007 | 0.0639 | 0 | 0.0322 | 0.7815 | 0 | 0.0583 | 0.8157 | |
| 3.2 | 0 | 0.0008 | 0.0646 | 0 | 0.0346 | 0.7903 | 0 | 0.0628 | 0.8249 | 10 |
| 4.2 | 0 | 0.0010 | 0.0646 | 0 | 0.0417 | 0.7903 | 0 | 0.0756 | 0.8249 | |
| 2.3 | 0 | 0.0162 | 0.7876 | 0 | 0.2781 | 0.9538 | 0 | 0.2818 | 0.9537 | |
| 3.2 | 0 | 0.0165 | 0.8199 | 0 | 0.2821 | 0.9795 | 0 | 0.2871 | 0.9789 | 20 |
| 4.2 | 0 | 0.0228 | 0.8193 | 0 | 0.3627 | 0.9815 | 0 | 0.3696 | 0.9820 | |

Table 4.1 shows the importance of signal-to-noise ratio in the data, but even more so the amount of copes needed in order for the methods to detect signal at all. Even when the signal is strong compared to the noise ($SNR = 10$), with only five copes discovery proportions are near zero for pARI (not enough data to permute) and voxel-counting (maximally conservative as a test). If the signal is strong enough, with twenty copes ARI and pARI both go into the right direction varying between approximately 95% and 98% depending on selected threshold. If the signal is moderate but still very distinguishable ($SNR = 5$) true discovery proportions drop to between 27% and 36% however, while in reality we know all voxels to be truly active. Notice the threshold makes a bigger difference in this scenario (approximately 10%), whereas the change in TDP was much smaller for the 10:1 scenario (only 3%). The data shown in Figure 4.4 better illustrate the effects of threshold and signal-to-noise ratio on the true discovery proportions.

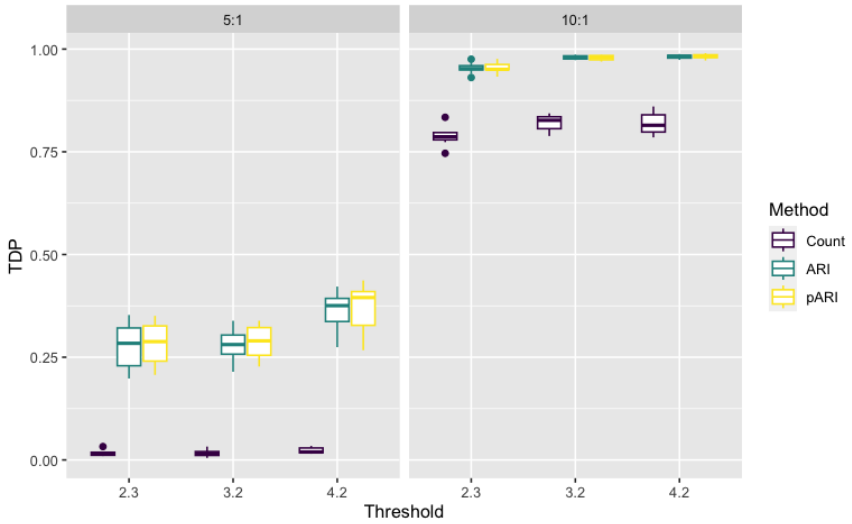


Figure 4.4: Effects of SNR and cluster-threshold on mean TDP by method.

To support the effects illustrated in Figure 4.4 Analysis of Variance was performed coupled with a linear model estimation. Since not all levels off interactions between threshold and method are significant the interaction was dropped to keep the selected model more simple. Full analysis of variance is included in appendix A.2 as well as the full linear model in appendix B.1. Details on the more parsimonious selected model are presented in Table 4.2.

Table 4.2: Model estimation ($R^2_{adj} = 0.9924$, $F = 2611$, $p < 0.001$).

| Coefficient | $\hat{\beta}$ | $\hat{se}$ | $t$ | $\mathbb{P}(>\|t\|)$ |
|---|---|---|---|---|
| Intercept | -0.002 | 0.008 | -0.309 | 0.7579 |
| Threshold 3.2 | 0.003 | 0.008 | 0.352 | 0.7250 |
| Threshold 4.2 | 0.060 | 0.008 | 7.120 | $< 0.001$ |
| ARI | 0.289 | 0.008 | 34.58 | $< 0.001$ |
| pARI | 0.294 | 0.008 | 35.21 | $< 0.001$ |
| SNR 10 | 0.792 | 0.011 | 74.40 | $< 0.001$ |
| Threshold 3.2 · SNR 10 | 0.025 | 0.012 | 2.100 | 0.0372 |
| Threshold 4.2 · SNR 10 | -0.030 | 0.012 | -2.560 | 0.0113 |
| ARI · SNR 10 | -0.127 | 0.012 | -10.70 | $< 0.001$ |
| pARI · SNR 10 | -0.132 | 0.012 | -11.14 | $< 0.001$ |

The model includes all main effects and four interaction effects. Let's for the moment discern between behaviour at the SNR 5 level versus at the 10 level. At the strong signal level (10:1) increasing the threshold from 2.3 to 3.2 increases the TDP at first, while increasing the threshold further to 4.2 makes no difference anymore. Compared to the 5 level where the difference between threshold 2.3 and 3.2 is small while increasing the threshold to 4.2 makes a proportionally bigger difference. Finally the count procedure gains much more from a high SNR then ARI and pARI; the increase in TDP is approximately 13% less for the latter two ($\hat{\beta}_{ARI \cdot SNR10:1} = -0.127$, $\hat{\beta}_{pARI \cdot SNR10:1} = -0.132$, $p < 0.001$). Note if we subtract the two coefficients the TDP difference between ARI and pARI itself is only 0.05%.

## 5 Conclusion and Discussion

This research studied methods for cluster-level inference on fMRI-data (particularly All-Resolutions Inference(ARI), permutation-based All-Resolutions Inference (pARI) and voxel-counting) on their ability to discover voxel activity as measure of spatial specificity; given equal clusters, what is the proportion of truly active voxels? Four data-sets and a simulation study learned the following regarding these methods. Most notably in data applications permutation-based All-Resolutions Inference always returned higher true discovery proportions when compared to All-Resolutions Inference, at least given the data in this article. If the amount of copes lie in the twenty to thirty range the TDP differences between ARI and pARI can be relatively substantial, however with more copes this difference is expected to shrink; in the auditory data TDP differences are much smaller compared to the arrow, food and rhyme data. Applying pARI however comes at the cost of heavy computational needs; especially if the data-set is large pARI easily requires multiple hours depending on hardware's computational capabilities. In contrast to ARI which typically requires a couple of minutes or less. If time is of the essence (or multiple analyses are needed) ARI could still be preferred. Without time-constraint pARI can be used instead. Counting voxels and applying a more classical multiple test correction was not a particularly viable for method for inference. The count procedure is too conservative; large, mainly active clusters returned proportions that are arguably too low to believe. Clearly the procedure is less effective in detecting activity by a substantial margin. Based on simulation-study the minimum recommended amount of copes to perform cluster-inference using any of the investigated methods is twenty. With ten copes ARI and pARI only detected strong signals corresponding to a signal-noise ratio of 10 and very low discovery proportions otherwise. With five copes pARI falls apart as there are not enough data to permute; to be safe one should always use ARI instead if the data has less than ten participants. Given twenty copes ARI and pARI pick up most activity (95%-98% depending on the threshold). Even the count procedure will detect substantial activity (78%-82%), but only for the strong signal. A middle-ground signal corresponding to SNR 5 is picked-up by ARI and pARI, but realistically only for at least twenty copes. With only ten copes discovery proportions are reduced drastically. If there's a medium signal increasing the threshold for from 3.2 to 4.2 also particularly improves TDP by around 9% for ARI and pARI or 3% for the count procedure, however again this works better with (more or at least) twenty copes as the difference quickly drops to between 0% and 3% for ten copes.

Concluding the research in this paper, comparing permutation-based All-Resolutions Inference against All-Resolutions Inference the permutation-based method leads to better spatial specificity (expressed in terms of true discovery proportions for equal clusters) in every typical case. The exception being when the amount of participants for a given study is smaller than ten, in which case there aren't enough permutations available to benefit from pARI's increased power. Compared to voxel-counting which serves mainly as a Bonferroni-like baseline, both ARI and pARI are a large improvement in spatial specificity; even more so when the signal is relatively weak or hard to distinguish. Compared to ARI the biggest downside of pARI is it's runtime, which is often hours compared to minutes. In situations where one would like to tune multiple parameters, perform multiple analysis or simply can't afford to leave hardware running for longer periods of time ARI will be more valuable. One recommendation for future research could be concerned with how ARI and pARI can be made more accurate for studies with few participants. In the current state twenty participants are recommended at the least, therefore the downside of these methods is any individual brain mapping is not possible yet. Permutations are of the table for individual scans, and we know ARI is conservative in this scenario. Potentially a less conservative test could be used, or we could pre-supply a critical vector for pARI estimated from a larger population with characteristics similar to the individual or small

group study. Using a prediction model one could compare critical vectors. Another goal could be to investigate how many permutations pARI requires and another could be to reduce pARI's runtime by exploring alternatives to the iterative estimation method currently used. Albeit more complicated, permutation-based All-Resolutions Inference has shown to be an improvement to spatial specificity compared to All-Resolutions Inference for most cases.

## References


Andreella, A., Hemerik, J., Finos, L., Weeda, W., & Goeman, J. J. (2023). Permutation-based true discovery proportions for functional magnetic resonance imaging cluster analysis. *Statistics in Medicine*.

Belin, P., Zatorre, R. J., Lafaille, P., Ahad, P., & Pike, B. (2000). Voice-selective areas in human auditory cortex. *Nature*, *403*(6767), 309–312.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, *57*(1), 289–300.

Chen, X., Goeman, J. J., Krebs, T. J., Meijer, R. J., & Weeda, W. D. (2022). Adaptive cluster thresholding with spatial activation guarantees using all-resolutions inference. *arXiv preprint arXiv:2206.13587*.

Goeman, J. J., Meijer, R., & Krebs, T. (2019). hommel: Methods for closed testing with simes inequality, in particular hommel's method. *R package version*, *1*.

Goeman, J. J., Meijer, R. J., Krebs, T. J., & Solari, A. (2019). Simultaneous control of all false discovery proportions in large-scale multiple hypothesis testing. *Biometrika*, *106*(4), 841–856.

Goeman, J. J., & Solari, A. (2011). Multiple testing for exploratory research.

Hemerik, J., Solari, A., & Goeman, J. J. (2019). Permutation-based simultaneous confidence bounds for the false discovery proportion. *Biometrika*, *106*(3), 635–649.

Hillman, E. M. (2014). Coupling mechanism and significance of the bold signal: a status report. *Annual review of neuroscience*, *37*, 161–181.

Hochberg, Y. (1988). A sharper bonferroni procedure for multiple tests of significance. *Biometrika*, *75*(4), 800–802.

Hommel, G. (1988). A stagewise rejective multiple test procedure based on a modified bonferroni test. *Biometrika*, *75*(2), 383–386.

Kelly, A. C., Uddin, L. Q., Biswal, B. B., Castellanos, F. X., & Milham, M. P. (2008). Competition between functional brain networks mediates behavioral variability. *Neuroimage*, *39*(1), 527–537.

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.

Pernet, C. R., McAleer, P., Latinus, M., Gorgolewski, K. J., Charest, I., Bestelmeyer, P. E., ... others (2015). The human voice areas: Spatial organization and inter-individual variability in temporal and extra-temporal cortices. *Neuroimage*, *119*, 164–174.

Pesarin, F., & Salmaso, L. (2010). The permutation testing approach: a review. *Statistica*, *70*(4), 481–509.

Rosenblatt, J. D., Finos, L., Weeda, W. D., Solari, A., & Goeman, J. J. (2018). All-resolutions inference for brain imaging. *Neuroimage*, *181*, 786–796.

Samuel-Cahn, E. (1996). Is the simes improved bonferroni procedure conservative? *Biometrika*, *83*(4), 928–933.

Simes, R. J. (1986). An improved bonferroni procedure for multiple tests of significance. *Biometrika*, *73*(3), 751–754.

Smeets, P. A., Kroese, F. M., Evers, C., & de Ridder, D. T. (2013). Allured or alarmed: counteractive control responses to food temptations in the brain. *Behavioural brain research*, *248*, 41–45.

Willemssen, R., Hoormann, J., Hohnsbein, J., & Falkenstein, M. (2004). Central and parietal event-related lateralizations in a flanker task. *Psychophysiology*, *41*(5), 762–771.

Xue, G., & Poldrack, R. A. (2007). The neural substrates of visual perceptual learning of words: implications for the visual word form area hypothesis. *Journal of cognitive neuroscience*, *19*(10), 1643–1655.

# A Analysis of Variance

Table A.1: Full ANOVA

| Effect | Df | Sum Sq | Mean Sq | $F$ | $\mathbb{P}(>F)$ |
|---|---|---|---|---|---|
| Threshold | 2 | 0.061 | 0.031 | 30.69 | $< 0.001$ |
| Method | 2 | 2.064 | 1.033 | 1033 | $< 0.001$ |
| SNR | 1 | 22.32 | 22.33 | 22465 | $< 0.001$ |
| Threshold:Method | 4 | 0.013 | 0.003 | 3.342 | $< 0.05$ |
| Threshold:SNR | 2 | 0.023 | 0.011 | 11.50 | $< 0.001$ |
| Method:SNR | 2 | 0.167 | 0.084 | 84.00 | $< 0.001$ |
| Residuals | 166 | 0.165 | 0.001 | | |

Table A.2: Reduced ANOVA

| Effect | Df | Sum Sq | Mean Sq | $F$ | $\mathbb{P}(>F)$ |
|---|---|---|---|---|---|
| Threshold | 2 | 0.061 | 0.031 | 29.09 | $< 0.001$ |
| Method | 2 | 2.065 | 1.033 | 984.7 | $< 0.001$ |
| SNR | 1 | 22.33 | 22.33 | 21293 | $< 0.001$ |
| Threshold:SNR | 2 | 0.023 | 0.011 | 10.87 | $< 0.001$ |
| Method:SNR | 2 | 0.167 | 0.084 | 79.62 | $< 0.001$ |
| Residuals | 170 | 0.178 | 0.001 | | |

# B Regression Analysis

Table B.1: Full linear model ($R^2_{adj} = 0.9928$, $F = 1908$, $p < 0.001$).

| Coefficient | $\hat{\beta}$ | $\hat{se}$ | $t$ | $\mathbb{P}(>\|t\|)$ |
|---|---|---|---|---|
| Intercept | 0.006 | 0.009 | 0.68 | 0.500 |
| Threshold 3.2 | 0.004 | 0.012 | 0.31 | 0.756 |
| Threshold 4.2 | 0.034 | 0.012 | 2.96 | $< 0.01$ |
| ARI | 0.277 | 0.012 | 24.017 | $< 0.001$ |
| pARI | 0.281 | 0.012 | 24.46 | $< 0.001$ |
| SNR 10:1 | 0.792 | 0.011 | 75.34 | $< 0.001$ |
| Threshold 3.2 · ARI | -0.001 | 0.014 | -0.08 | 0.935 |
| Threshold 4.2 · ARI | 0.037 | 0.014 | 2.64 | $< 0.01$ |
| Threshold 3.2 · pARI | -0.001 | 0.014 | -0.05 | 0.958 |
| Threshold 4.2 · pARI | 0.039 | 0.014 | 2.77 | $< 0.01$ |
| Threshold 3.2 · SNR 10:1 | 0.025 | 0.012 | 2.16 | $< 0.05$ |
| Threshold 4.2 · SNR 10:1 | -0.030 | 0.012 | -2.63 | $< 0.01$ |
| ARI · SNR 10:1 | -0.127 | 0.012 | -10.99 | $< 0.001$ |
| pARI · SNR 10:1 | -0.132 | 0.012 | -11.45 | $< 0.001$ |