Comparing fMRI inter-subject correlations between groups

Jussi Tohka[a,f], Frank E. Pollick[b], Juha Pajula[c], Jukka-Pekka Kauppi[d,e]

[a] *A.I. Virtanen Institute for Molecular Sciences,University of Eastern Finland, Kuopio, Finland*
[b] *School of Psychology, University of Glasgow,UK*
[c] *VTT Technical Research Centre of Finland, Tampere, Finland*
[d] *Faculty of Information Technology, University of Jyväskylä, Finland*
[e] *Department of Computer Science and HIIT, University of Helsinki, Finland*
[f] *Corresponding author, jussi.tohka@uef.fi*

Abstract

Inter-subject correlation (ISC) based analysis is a conceptually simple approach to analyze functional magnetic resonance imaging (fMRI) data acquired under naturalistic stimuli such as a movie. We describe and validate the statistical approaches for comparing ISCs between two groups of subjects implemented in the ISC toolbox, which is an open source software package for ISC-based analysis of fMRI data. The approaches are based on permutation tests. We validated the approaches using five different data sets from the ICBM functional reference battery tasks. In these experiments, we created two matched groups of subjects and assumed that no group difference exists. Based on the experiments, we recommend the usage of subject-wise permutations, instead of element-wise permutations following Chen et al. (2016). However, we observed that the null-distributions should be voxel-specific and not based on pooling all voxels across the brain as is typical in fMRI. This was the case even if studentized permutation tests were used. Additionally, we experimented with an fMRI dataset acquired using a dance movie stimulus for comparison of the group of adult males in autism spectrum to the matched control group. The experiment confirmed the differences between voxel-based permutation tests and global model based permutation tests.

Keywords: functional magnetic resonance, fMRI, brain, permutation test, naturalistic stimuli

1. Introduction

Inter-subject correlation (ISC) based analysis, originally introduced by Hasson et al. (2004), is a conceptually simple approach to analyze functional magnetic resonance imaging (fMRI) data acquired under naturalistic stimuli such as a movie. In the ISC based analysis of the functional magnetic resonance

imaging (fMRI) data, the extent of shared processing across subjects during the experiment is determined by calculating correlation coefficients between the fMRI time series of the subjects in the corresponding brain locations and then averaging the correlation coefficients. ISC maps have been shown to align well with the activation maps of the traditional block design stimuli when the stimuli can be modeled by a reference time-course (Pajula et al., 2012), and it has been applied in a number of studies (Hasson et al., 2008; Abrams et al., 2013; Jääskeläinen et al., 2008; Englander et al., 2012; Kauppi et al., 2010; Nummenmaa et al., 2012; Wilson et al., 2007). In addition to localizing shared processing within a group, the ISC methodology has been extended to compare ISCs between two similar, but not identical, stimuli within a single group of subjects (Herbec et al., 2015; Reason et al., 2016), as well as to compare between two groups of subjects experiencing the same stimuli (Hasson et al., 2009; Salmi et al., 2013). In this paper, we are interested in the latter scenario. An example application, highlighted also in this paper, is a comparison of the ISCs of subjects diagnosed with the autism spectrum disorder to the ISCs of matched controls.

In the previous study involving a comparison of ISCs between different groups of subjects, Hasson et al. (2009) compared the extent and strength of ISCs of autism and typical groups during free-viewing of a movie. They computed ISC maps separately for both groups as well as between the two groups. To construct the ISC maps, they averaged Fisher z transformed subject-pairwise ISCs within each voxel/region-of-interest. The maps were thresholded based on a maximum value obtained from the identical procedure but using correlations of forward and reversed time courses. In another fMRI study involving ISC based group comparison, Salmi et al. (2013) used t-statistic to assess difference between the average ISCs of the autistic and normal groups. A permutation test was used to assess statistical significance of the t-statistic. In the test, subjects were randomly exchanged between the groups before re-calculating the t-statistic.

Chen et al. (2016) proposed permutation and bootstrap based methods for statistical hypothesis testing for comparing ISCs between two groups, and Chen et al. (2017) introduced a linear mixed effects model as an alternative to non-parametric ISC hypothesis tests. In particular, Chen et al. (2016) demonstrated that a permutation test exchanging the components of the correlation matrix leads to excessively liberal hypothesis tests while exchanging subjects between the two groups leads to tests with approximately correct alpha levels. In this paper, we will extend their analysis in several ways: 1) we will re-confirm the above mentioned conclusion by Chen et al. (2016) by new experiments; 2) we will introduce studentized statistics for testing the differences of ISCs between group; 3) we will show that

the voxel-level null-distributions are more appropriate than the image level null-distributions; 4) we will discuss multiple comparisons corrections, which were not discussed in Chen et al. (2016, 2017).

The methods, which we describe, are implemented in the ISC toolbox (Kauppi et al., 2014), which is an open source software package for ISC analysis available at https://www.nitrc.org/projects/isc-toolbox/. The toolbox consists of three parts: (1) a startup GUI for setting-up parameters for the analysis, (2) a main program that computes ISC maps based on selected parameters, and (3) a GUI-based visualization tool for the exploration of the findings. The GUIs are designed to make the analysis easier but a whole analysis pipeline can also be carried out from Matlab's command line.

2. ISC group comparison

Let us denote a number of time points in an fMRI time course by T, and a number of subjects of ith group by N_i $(i = 1, 2)$. The fMRI time courses of the voxel j from all the subjects in two groups are $X_j^{(1)}$ and $X_j^{(2)}$, where $X_j^{(i)}$ is a $N_i \times T$ matrix. Further, let $R_j^{(i)} = corr(X_j^{(i)})$ and $Z_j^{(i)}$ its z-transform [1]. We denote the element (n, m) of $Z_j^{(i)}$ by $z_j^{(i)}(n, m)$, which is the z-transformed correlation coefficient between the time courses of the subjects m and n. A test statistic for comparing the two ISCs between two groups is the difference between the means of z-transformed correlations:

$$S_j = \frac{1}{N_1(N_1-1)/2} \sum_{n=1}^{N_1-1} \sum_{m=2,m>n}^{N_1} z_j^{(1)}(n,m) - \frac{1}{N_2(N_2-1)/2} \sum_{n=1}^{N_2-1} \sum_{m=2,m>n}^{N_2} z_j^{(2)}(n,m). \tag{1}$$

To develop the hypothesis testing for S_j, we introduce a matrix $X_j = \begin{bmatrix} X_j^{(1)} \\ X_j^{(2)} \end{bmatrix}$ and define $R_j = corr(X_j)$ and Z_j as the (element-wise) z-transformation of R_j. In terms of Z_j, we can write S_j in an equivalent form

$$S_j = \frac{1}{N_1(N_1-1)/2} \sum_{n=1}^{N_1-1} \sum_{m=2,m>n}^{N_1} z_j(n,m) - \frac{1}{N_2(N_2-1)/2} \sum_{n=1}^{N_2-1} \sum_{m=2,m>n}^{N_2} z_j(n+N_1, m+N_1). \tag{2}$$

This form will be needed later when developing permutation based strategy to form inferences.

[1] The correlation matrix of X, $corr(X)$, is a matrix of pairwise correlation coefficients between the columns of X. $R_j^{(i)}$ denotes the correlation matrix of $X_j^{(i)}$. We apply Fisher's z-transform to the elements of this matrix to obtain the z-transformed matrix: $Z_j^{(i)} = \operatorname{arctanh}\left(R_j^{(i)}\right)$.

Often in the randomization based hypothesis testing, it is useful to consider studentized test statistics when the interest is in difference between particular parameters, such as means or medians, of two distributions (Chung and Romano, 2013). This is because a permutation test is sensitive to all differences between the two distributions, rather than to the difference in a particular parameter (Chung and Romano, 2013). This leads us to using the studentized test statistics

$$t_j^0 = \frac{S_j}{\sqrt{\hat{v}_j}}, \tag{3}$$

where $\hat{v}_j$ is an estimate of variance of S_j. We approximate the variance for each group by

$$\hat{v}_j^{(i)} = \frac{2}{N_i(N_i-1)} \sum_{n=1}^{N_1-1} \sum_{m=2,m>n}^{N_1} (\overline{z}_j - z_j^{(i)}(n,m))^2, \tag{4}$$

$\overline{z}_j = \frac{1}{N_i(N_i-1)} \sum_n \sum_{m>n} z_j^{(i)}(m,n)$. This is a biased variance estimate as it does not account for the dependencies between the elements of the correlation matrix $z_j^{(i)}(n,m)$. On the other hand, it is an approximation, up to a multiplication by a constant, of the leave-one-subject-out variance defined by Kauppi et al. (2017). We then define

$$\hat{v}_j = \frac{N_1}{N_1+N_2}\hat{v}_j^{(1)} + \frac{N_2}{N_1+N_2}\hat{v}_j^{(2)}. \tag{5}$$

It is important to note that the variance estimates can be also written in terms of the matrix X_j as the statistic S_j in Eq. (2) above as this is required for the subject-wise permutation strategy below. We then define the test statistic as the SAM statistic (Tusher et al., 2001; Xie et al., 2005)

$$t_j = \frac{S_j}{\sqrt{\hat{v}_j + v_0}}, \tag{6}$$

where v_0 is a small positive constant. We will study two versions of this statistic: 1) t_j^0, where $v_0 = 0$, and 2) t_j^1, where v_0 is set as the 0.25% percentile of $\hat{v}_j$ across the brain voxels. Finally, we emphasize that just normalizing the test statistic values (S_j) is not enough for a studentized permutation test, but the variance estimates must be re-computed for each permutation.

We study two types of the permutation strategies to form the null distributions as Chen et al. (2016):

- **Subject-wise (SW) permutation**. The rows and columns of the matrices Z_j are permuted by a permutation π acting on the set $1, 2, \ldots, N_1 + N_2$. The row and column indexes are permuted with the same permutation. Eq. (2) or Eq. (3) or Eq. (6) is used to compute the test statistic value

after the permutation. This type of permutation strategy necessitates computing also between-group ISCs (i.e., the ISCs of time series of subjects from different groups). Note that this procedure corresponds to random swapping of subjects between the two groups before computing subject-pairwise ISCs. However, it is faster to pre-compute the (z-transformed) correlation matrices and work with them than to compute the test statistics based on time series.

- **Element-wise (EW) permutation**. The elements of the correlation matrices $Z_j^{(1)}$ and $Z_j^{(2)}$ are randomly swapped. In total, there are $N_1(N_1-1)/2 + N_2(N_2-1)/2$ elements in these matrices. Eq. (1) can be used to compute the test statistic value after the permutation. Note that after the permutation, the matrices $\pi(Z_j^{(1)})$ and $\pi(Z_j^{(2)})$ do not have to be proper correlation matrices and the permutation satisfies the exchangeability under extremely limited assumptions that do not hold in practice. Hence, the α-level is expected to be overly liberal.

Repeating the permutation procedure B times leads to the hypothesis test using one of the above permutation processes. However, an important question remains: should one generate the null-model for each voxel independently or can one assume the same null model for all voxels? The latter approach is usually taken in fMRI and, especially, with studentized t_j statistics one might assume that this is reasonable. However, as we will show, also t_j statistic values correlate with the average ISC $S_j^{(1)} + S_j^{(2)}$, and thus the global-null-model suffers from problems. This is why it is advisable to proceed by computing a null-model separately for every voxel to obtain a null-distribution (and a p value) for that voxel. This p-value is then Gaussianized using the p-to-Z transform. The disadvantage of this procedure is that it limits the options for multiple comparisons correction (MCC) by essentially ruling out the permutation-based cluster-extent or peak corrections. The Gaussian Random Field (GRF) based corrections are still available by turning voxel p-values into a Z-field by the p-to-Z transform. However, one should be aware of the caveat that the GRF-based cluster correction has been deemed to be overly liberal in certain settings (Eklund et al., 2016). Similarly, corrections based on the false disovery rate (FDR) are still available (Ge et al., 2003). In particular, we implement FDR corrections based on Storey's procedure (Storey, 2002; Storey and Tibshirani, 2003), which we have found to be better powered than the typical Benjamini-Hochberg procedure (Benjamini and Hochberg, 1995).

Our current implementation in the ISC toolbox produces both global-null and voxel-null models with the voxel-null model strongly recommended. For the MCC, only Storey's FDR corrections are available for voxel-null models. For the global null, we have implemented three MCC strategies 1) FDR-

based correction, permutation-based 2) voxel-wise or 3) cluster-extent-based extreme statistics over the brain (Nichols and Holmes, 2002). This leads to strong control of the (voxel or cluster-wise) family-wise error-rate (FWER). The cluster extent correction replaces the original voxel-wise statistics by the extents of clusters exceeding a pre-defined cluster defining threshold. As the parametric tests for test-statistics are not available, we set the cluster defining threshold based on raw (i.e., uncorrected) p-values approximated by the permutation process. We sample M voxels from the brain or gray matter mask, repeat permutations B_0 times, and pool the resulting B_0M test statistic values to form the null distribution.

For the voxel-null models, which require many more permutation iterations (since the MCC cannot be incorporated in the permutation framework), we use a following strategy to speed up the computations: first, we compute B_1 permutation iterations and then fine tune the p-values with more iterations for only those voxels that have sufficiently low the initial p-values (from the first B_1 iterations). This cycle is repeated several times with the p-value threshold for fine tuning lowered within every iteration. This procedure saves a substantial amount of computation expense. We recommend to set B to 25000 for the global-null models and B_1 to 5000 for the voxel-null models, and all the experiments reported in this work are based on these settings.

3. Materials and methods

3.1. Null experiments

We generated five data sets where we do not expect to see any between group effect in ISC as the groups were based on the matched groups of healthy young subjects. For this, we used the fMRI data from 36 healthy young adults (18 men and 18 women; the average age was 28.2 years from the range from 20 to 36 years) during the ICBM functional reference battery (FRB) tasks `https://ida.loni.usc.edu/login.jsp?project=ICBM`. The ICBM project (Principal Investigator John Mazziotta, M.D., University of California, Los Angeles) is supported by the National Institute of Biomedical Imaging and BioEngineering. ICBM is the result of efforts of coinvestigators from UCLA, Montreal Neurological Institute, University of Texas at San Antonio, and the Institute of Medicine, Juelich/Heinrich Heine University, Germany. We have used the same data earlier in several ISC evaluation experiments (Pajula et al., 2012; Pajula and Tohka, 2014, 2016).

All five FRB tasks were block-design tasks (12 blocks per run (6 off-on blocks) and 3 volumes at the beginning of the run to wait for magnetisation stabilisation), which were highly standardized. In the auditory naming task (AN), subjects were instructed to listen to the description of an object from a sound file and then think their answer silently to the description. In the external order (EO) task, the subjects were presented with four abstract design stimuli followed by a fifth stimulus and required to recall whether the final abstract design was among the four presented previously. In the HA task subjects were instructed to imitate the presented hand configuration with their right hand. In the VG task, the images of certain objects were shown to the subjects on the screen and subjects were instructed to generate a verb associated to the object silently in their mind without saying it aloud. In the OM task, subjects watched an image including a central cross in the middle surrounded by 10 black boxes. Subjects were instructed to concentrate on the central cross and saccade to the surrounding box if it changed white for a moment. After this, they were instructed to return their gaze immediately to the central cross. For a more detailed description of the five tasks, see (Pajula et al., 2012).

The functional data was collected with a 3 Tesla Siemens Allegra fMRI scanner and the anatomical T1 weighted MRI data with an 1.5 Tesla Siemens Sonata scanner. The TR/TE times for the functional data were 4 s/32 ms, flip angle 90 degree, pixel spacing 2 mm and slice thickness 2 mm. The parameters for the anatomical T1 data were 1.1 s/4.38 ms, 15 degree, 1 mm and 1 mm, correspondingly. The fMRI data was preprocessed (including motion correction, stereotactic registration, temporal high-pass filtering with a cutoff period of 60s, spatial filtering with 5mm isotropic kernel) using FSL as described in (Pajula et al., 2012).

We divided 36 subjects randomly into two groups of 18 subjects so that the groups were age and sex matched. We expected to see no difference between the two groups in this setting. We checked that there was no group differences by using the standard general linear model based hypothesis test as implemented in FSL's (FMRIB's Software Library, www.fmrib.ox.ac.uk/fsl) FEAT (FMRI Expert Analysis Tool) Version 6.00. Higher-level analysis was carried out using FLAME (FMRIB's Local Analysis of Mixed Effects) stage 1 and stage 2 (Beckmann et al., 2003; Woolrich et al., 2004; Woolrich, 2008). Note that the standard analysis is different from the ISC-based analysis in that it requires reference time courses whereas the ISC-based analysis does not. The false positive rates for this analysis are presented in the Appendix.

3.2. Autism spectrum experiment

Additionally, we have applied the method to compare inter-subject correlation of 10 male adults on the autism spectrum (ASD) to 10 age and IQ matched typically developed (TD) male adults while they viewed a 90 second clip of a solo ballet dance. The TD group was comprised of 10 individuals with an average age of 26.5 6.9, and the ASD group included 10 individuals with an average age of 28.4 9.0. The TD group had an average Autism Quotient (AQ) score of 12.7 4.8 (N=9) and an average Intelligence Quotient (IQ) score of 118.6 6.8. The ASD group had an average AQ of 37.11 7.3 (N=9) and an average IQ score of 118.3 5.9. All were right handed as assessed by the Edinburgh handedness inventory. Participants were recruited from the participant database at the School of Psychology, University of Glasgow. None of the participants had experience in practicing ballet dance and none regularly watched dance performances. Ethical permission for the study was obtained from the Greater Glasgow and Clyde National Health Service ethics board.

While in the scanner, all participants viewed three dance videos and in the present analysis we examine just one. This stimulus was a video (60 fps, 1280 by 720 resolution) of a Romantic ballet dance (Giselles solo dance in Act II of Giselle), 90 seconds in duration. The video was also converted to black and white, the ballerinas face was blurred out and there was no associated audio track. Stimulus presentation was controlled by Presentation software (Neurobehavioural systems, Inc). Before beginning the experiment, participants were instructed to simply relax and enjoy watching the dances while being scanned.

Data were acquired from a single functional T2*-weighted acquisition (EPI, TR 2000 ms; TE 30 ms; 32 Slices; 3mm^3 voxels; FOV of 210, imaging matrix of 70 x70) using a 3T Tim Trio Siemens scanner. The run took 270 seconds with a total of 90 seconds for each dance presentation. There were 8 seconds of blank at the beginning and 36 seconds at the end of the run and 16 seconds of blank between the first and second as well as the second and third dance presentation. The Romantic style dance chosen for analysis occurred randomly in either the first or second position. An anatomical scan was performed at the end of the scanning session that comprised a high-resolution T1-weighted anatomical scan using a 3D magnetization prepared rapid acquisition gradient recalled echo (ADNI-MPRAGE) T1-weighted sequence (192 slices; 1mm cube isovoxel; Sagittal Slice; TR = 1900 ms; TE = 2.52; 256 x 256 image resolution). The fMRI data were preprocessed in Brain Voyager QX (Vers.2.6, Brain Innovation B.V., Maastricht, Netherlands). This included: 3D Motion Correction with Trilinear/sinc interpolation, slice scan-time correction, linear removal, and high-pass filtering with cutoff set to 1 cycle. Spatial smoothing with a Gaussian kernel of 6 mm FWHM was also applied. This was followed by normalization of functional

scans into common Talairach space, and co-registration of functional and anatomical data. Finally, the functional data were trimmed using Matlab to obtain the 45 volumes (90 seconds) for each dance, used later for ISC analysis.

4. Results

4.1. Null experiment

The results of the null experiment are listed in Tables 1 and 2. In the Tables, the fractions of significant voxels in the experiment across the gray matter mask are displayed when the α-level is varied. We call this fraction as (observed) false positive rate and it should be as close as possible to the nominal α-level (0.05, 0.01, or 0.001 in the tables). As the tables show, the subject-wise permutations led to the approximately correct observed false positive rates while the element-wise permutations produced too liberal p-values (i.e. the observed false positive rate was much higher than the nominal α-level). This result agrees with Chen et al. (2016). The tables also list the correlation between the test statistic for the group difference and the average ISC of the two groups. This correlation should be as close to zero as possible (as the average ISC between two groups should be independent of the ISC difference). As the tables show, only the subject-wise voxel-null models displayed desirable behavior whereas subject-wise global-null models displayed correlations from 0.33 (SWstu0 with AN task, see the table caption for the explanation of the abbreviation) to 0.57 (HA task with SW). Note that the studentization reduced the correlation but did not eliminate it. The undesirable performance of the global-null models is visualized in Figure 1. As can be seen in the Figure, the global-null model (thresholded at $p < 0.05$, SW-permutations with studentization) produced large clusters of significant voxels. This performance was not expected for a null experiment. Instead, the voxel-null model produced a comparison map which appeared more realistic with small clusters of significant voxels across the brain.

4.2. Autism spectrum experiment

The results of the comparison maps (both voxel null-model and global-null model, studentisized with v_0 is set as the 0.25% percentile of $\hat{v}_j$ across the brain voxels) thresholded at p = 0.01 (uncorrected) showed several brain regions that differed between the groups. The voxel-null results are shown in Fig. 2. Ignoring small clusters less than 108 mm^3 revealed 4 regions where the ISC map was greater for the ASD group, and 1 region where the ISC map was greater for the TD group. The ASD group had greater

		Voxel-null				Global-null			
task	method	$p < 0.05$	$p < 0.01$	$p < 0.001$	ρ	$p < 0.05$	$p < 0.01$	$p < 0.001$	ρ
AN	EW	0.1102	0.0454	0.0177	0.3503	0.1118	0.0489	0.0214	0.4402
	EWstu0	0.1102	0.0454	0.0177	0.3503	0.1102	0.0456	0.0181	0.3664
	EWstu1	0.1102	0.0455	0.0179	0.3532	0.1115	0.0479	0.0202	0.4041
AN	SW	0.0452	0.0086	0.0008	0.0162	0.0450	0.0085	0.0008	0.3944
	SWstu0	0.0450	0.0086	0.0009	0.0152	0.0449	0.0084	0.0005	0.3334
	SWstu1	0.0452	0.0086	0.0009	0.0156	0.0450	0.0084	0.0006	0.3613
EO	EW	0.1882	0.1043	0.0531	0.3780	0.1809	0.0996	0.0468	0.5345
	EWstu0	0.1883	0.1041	0.0530	0.3772	0.1887	0.1049	0.0567	0.4155
	EWstu1	0.1880	0.1040	0.0531	0.3776	0.1889	0.1083	0.0613	0.4632
EO	SW	0.0436	0.0074	0.0006	0.0128	0.0484	0.0085	0.0009	0.5054
	SWstu0	0.0433	0.0074	0.0006	0.0124	0.0463	0.0080	0.0006	0.3867
	SWstu1	0.0436	0.0074	0.0006	0.0121	0.0474	0.0081	0.0007	0.4393

Table 1: The results of the null experiment with the AN and EO stimuli of the ICBM-FRB. The columns $p < 0.05$, $p < 0.01$, and $p < 0.001$ list the fraction of the significant voxels at each α level (false positive rate). The closer the value is to the nominal α-level, the better the result is. The columns ρ list the correlation between the test statistic for group difference and the average ISC of the two groups. Lower the absolute correlation, the better the result. EW and SW are the permutation strategies, stu0 is the studentized statistic with $v_0 = 0$, stu1 is the studentized statistic with v_0 as 0.25 % quantile of voxel-wise variances across the mask (see Section 2). Only SW permutations with voxel-wise null models produced satisfactory results, independent of the studentization strategy.

		Voxel-null				Global-null			
task	method	$p < 0.05$	$p < 0.01$	$p < 0.001$	ρ	$p < 0.05$	$p < 0.01$	$p < 0.001$	ρ
	EW	0.2060	0.1193	0.0642	0.4509	0.1977	0.1162	0.0638	0.5952
HA	EWstu1	0.2064	0.1200	0.0641	0.4515	0.2068	0.1245	0.0737	0.5348
	EWstu0	0.2058	0.1193	0.0634	0.4500	0.2060	0.1201	0.0677	0.4784
	SWstu1	0.0349	0.0060	0.0005	0.0224	0.0414	0.0078	0.0004	0.5102
HA	SWstu0	0.0340	0.0058	0.0004	0.0222	0.0387	0.0065	0.0003	0.4603
	SW	0.0357	0.0062	0.0006	0.0161	0.0439	0.0091	0.0004	0.5721
	EW	0.1892	0.0996	0.0458	0.3902	0.1841	0.0961	0.0437	0.5400
OM	EWstu1	0.1887	0.0990	0.0453	0.3890	0.1890	0.1023	0.0528	0.4992
	EWstu0	0.1891	0.0993	0.0455	0.3891	0.1894	0.1001	0.0491	0.4486
	SWstu1	0.0473	0.0091	0.0008	0.0445	0.0457	0.0113	0.0020	0.4628
OM	SWstu0	0.0476	0.0090	0.0008	0.0424	0.0462	0.0111	0.0016	0.4165
	SW	0.0474	0.0089	0.0009	0.0445	0.0463	0.0114	0.0023	0.5153
	EW	0.1431	0.0703	0.0308	0.4151	0.1445	0.0770	0.0375	0.5480
VG	EWstu1	0.1429	0.0701	0.0307	0.4142	0.1456	0.0762	0.0391	0.4906
	EWstu0	0.1428	0.0700	0.0308	0.4139	0.1432	0.0707	0.0333	0.4349
	SWstu1	0.0428	0.0084	0.0007	-0.0169	0.0414	0.0053	0.0001	0.4460
VG	SWstu0	0.0424	0.0083	0.0007	-0.0164	0.0401	0.0050	0.0001	0.3995
	SW	0.0424	0.0084	0.0008	-0.0204	0.0424	0.0057	0.0003	0.5019

Table 2: The results of the null experiment with the HA, OM, and VG stimuli of the ICBM-FRB. The columns $p < 0.05$, $p < 0.01$, and $p < 0.001$ list the fraction of the significant voxels (false positive rate) at each α level. The closer the value is to the α-level, the better the result is. The columns ρ list the correlation between the test statistic for group difference and the average ISC of the two groups. Lower the absolute correlation, the better the result. EW and SW are the permutation strategies, stu0 is the studentized statistic with $v_0 = 0$, stu1 is the studentized statistic with v_0 as the 0.25% quantile of across the mask variances (see Section 2.1). Only SW permutations with voxel-wise null models produced satisfactory results, independent of the studentization strategy.

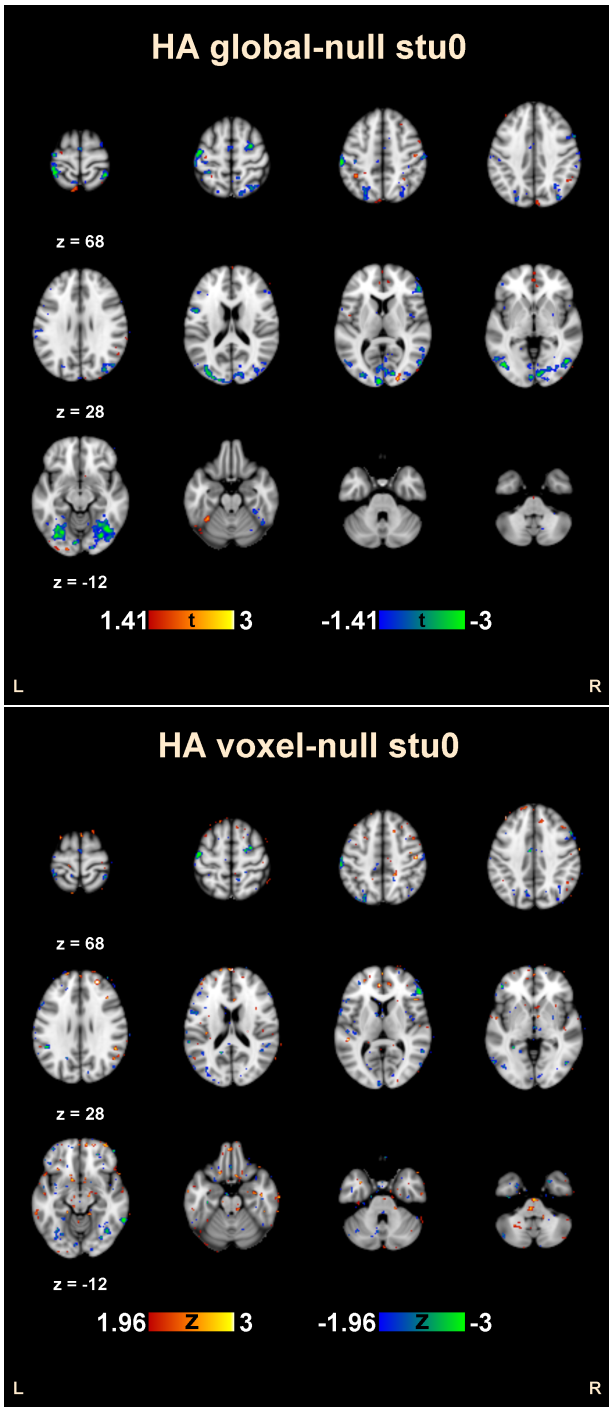

Figure 1: Uncorrected statistics maps of the HA task thresholded at $p < 0.05$ for global-null (upper panel) and voxel-null (lower panel) models. The subject-wise permutations with studentization was applied. The global-null model produced large clusters of significant voxels unlike the voxel-null model, while statistic map appears more realistic for a null-experiment where differences between groups are expected to be by chance.

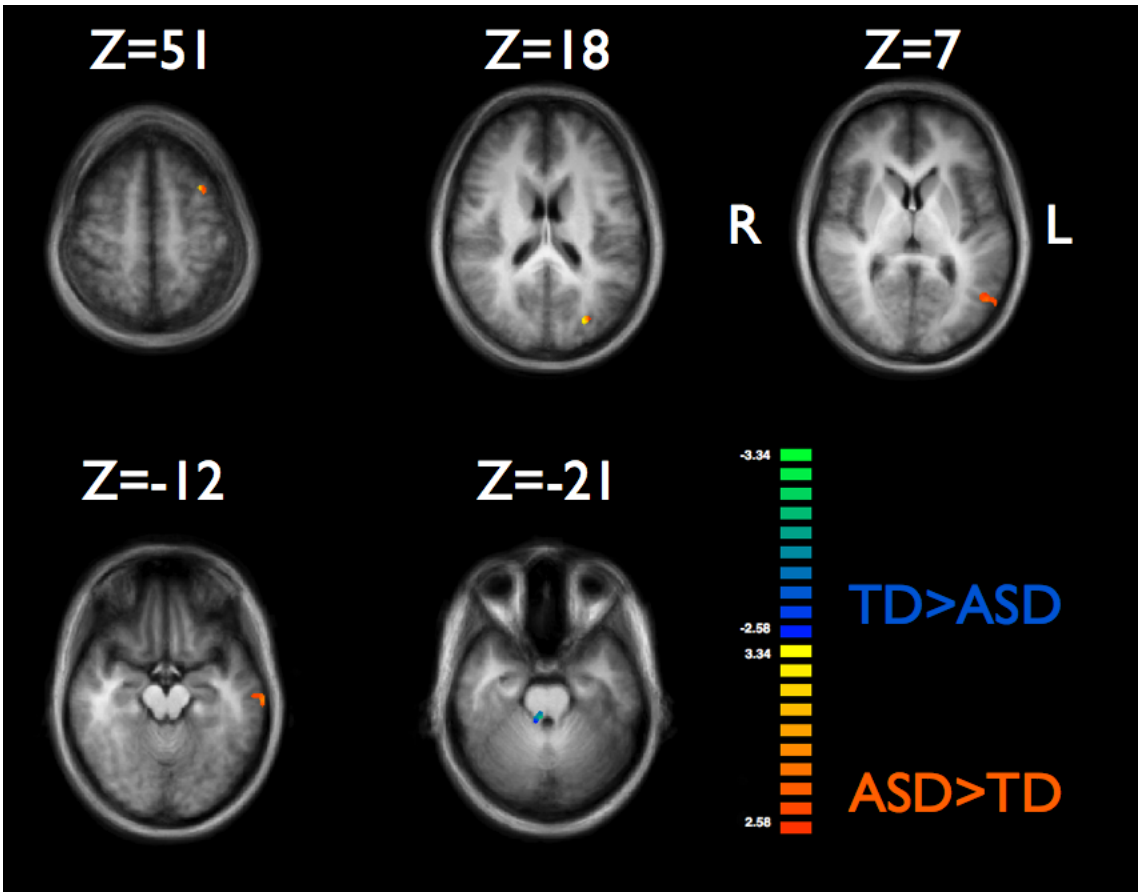

Figure 2: Uncorrected comparison map from the autism spectrum experiment, thresholded at $p = 0.01$. Voxel null model, studentisized with v_0 is set as the 0.25% percentile of $\hat{v}_j$ across the brain voxels.

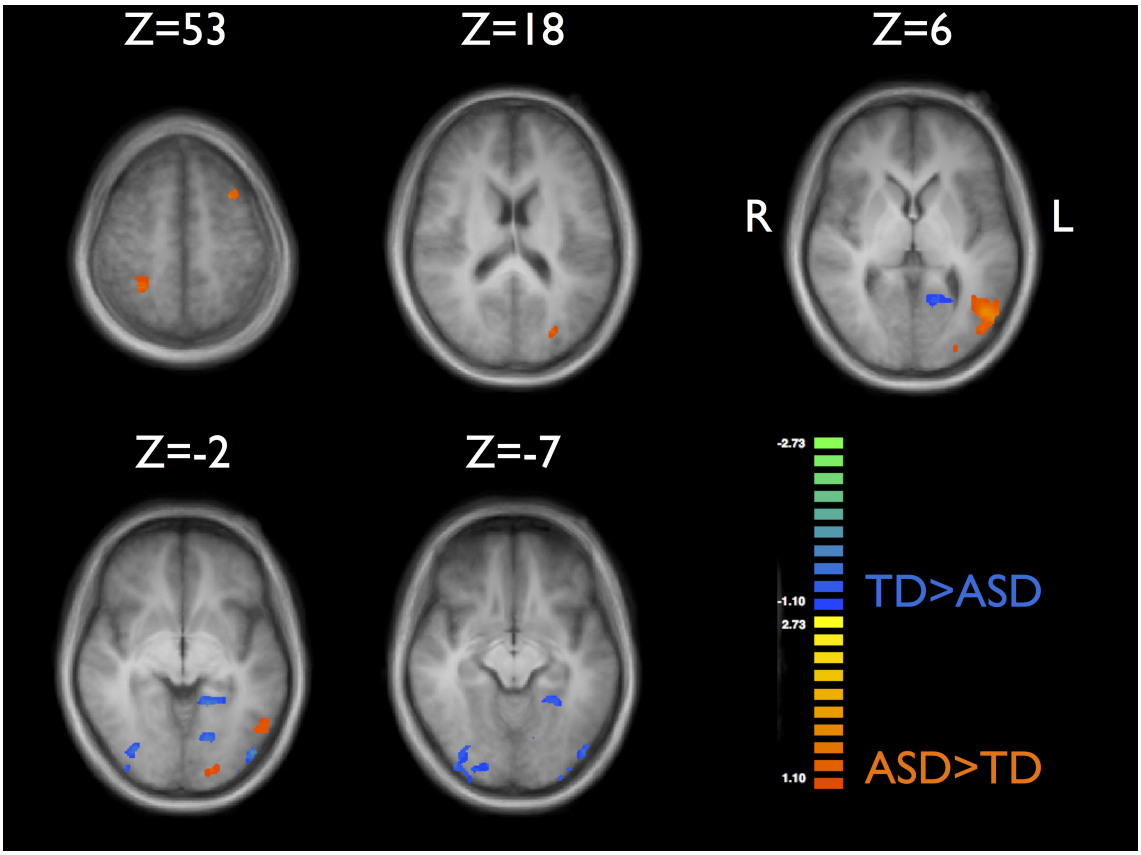

Figure 3: Uncorrected comparison map from the autism spectrum experiment, thresholded at $p = 0.01$. Global null model, studentisized with v_0 is set as the 0.25% percentile of $\hat{v}_j$ across the brain voxels.

ISC map values in frontal, parietal and temporal cortices as well as the left occipitotemporal cortex at Talairach coordinates (-49, -65 6), BA 37, which was the largest cluster at 271 mm^3. The TD group had greater ISC map values in a single cluster, located in the culmen. The global null results are shown in Figure 3. Ignoring small clusters less than 108 mm^3 revealed 6 clusters where there the ISC map was greater for the ASD group, and 5 regions where the ISC map was greater for the TD group. The ASD group had greater ISC map values in frontal and parietal cortex as well as the left occipitotemporal cortex at Talairach coordinates (-49, -65 6), BA 37, which was the largest cluster at 3036 mm^3. The TD group had greater ISC map values in several clusters located in the lingual gyrus and inferior occipital gyrus. The voxel and global null results qualitatively agreed for the contrast $ASD > TD$, but for the opposite contrast the results disagreed. As the null-experiment results (see Section 4.1) support the use of the voxel-null model instead of the global-null model, we recommend using the voxel-based null model if there is no other reason to prefer the global null instead.

5. Discussion

We have presented the ISC group comparison approach implemented in the ISCtoolbox software package https://www.nitrc.org/projects/isc-toolbox/. We have verified that the permutation-based ISC group comparison results in approximately correct type I error rates. Our results re-confirmed the finding by Chen et al. (2016) (who based their conclusion on synthetic data) that the element-wise permutations led to a high false positive rate. We thus strongly recommend subject-wise permutation strategy albeit element-wise permutations are available as an option in the ISCtoolbox. We have presented results that discourage the use of the global (across-brain) null-models, and instead we suggest to generate null-models independently for each voxel. The studentization of the test statistic was found to improve the performance of the global-null-models but not to cure their deficiencies entirely. The limitation of the current study is that it does not include an experiment where we would expect to see group difference in a certain, a-priori known location. Therefore, currently, we have no means to discuss about the statistical power of the methods studied.

Appendix: False positive rates for null experiment by general linear model

Since the null-experiment (section 3.1) was based on the block-design data, we can perform a group-difference test based on the standard General linear model (GLM) to help to verify that there are no

task	$p < 0.05$	$p < 0.01$	$p < 0.001$
EO	0.0469	0.0140	0.0030
AN	0.0474	0.0174	0.0036
HA	0.0172	0.0029	0.0004
OM	0.0700	0.0285	0.0063
VG	0.0402	0.0125	0.0023

Table 3: The results of the GLM-based analysis of the null-experiment of section 3.1. The columns $p < 0.05$, $p < 0.01$, and $p < 0.001$ list the fraction of the significant voxels at each α level (false positive rate). These results verify that also the standard GLM-based analysis produced false positive rates close to the nominal α-level, re-assuring that there was no group difference in fMRIs between the two matched groups of subjects.

significant group differences. The results of FSL's (FMRIB's Software Library, www.fmrib.ox.ac.uk/fsl) FEAT (FMRI Expert Analysis Tool) Version 6.00 with higher-level analysis was carried out using FLAME (FMRIB's Local Analysis of Mixed Effects) stage 1 and stage 2 (Beckmann et al., 2003; Woolrich et al., 2004; Woolrich, 2008) are in Table 3.

Acknowledgments

J.T's work is supported by the Academy of Finland. J-P.K was funded by the Academy of Finland Postdoctoral Researcher program (Research Council for Natural Sciences and Engineering; grant number 286019). Data collection and sharing for this project was, in part, provided by the International Consortium for Brain Mapping (ICBM; Principal Investigator: John Mazziotta, MD, PhD). ICBM funding was provided by the National Institute of Biomedical Imaging and BioEngineering. ICBM data are disseminated by the Laboratory of Neuro Imaging at the University of Southern California. We thank Naree Kim and Seonhee Jang for use of the dance video and Paula Regener for use of the fMRI data comparing typical and autism spectrum observers.

References

Abrams DA, Ryali S, Chen T, Chordia P, Khouzam A, Levitin DJ, Menon V. Inter-subject synchronization of brain responses during natural music listening. European Journal of Neuroscience 2013;37(9):1458–69.

Beckmann CF, Jenkinson M, Smith SM. General multilevel linear modeling for group analysis in fmri. Neuroimage 2003;20(2):1052–63.

Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the royal statistical society Series B (Methodological) 1995;:289–300.

Chen G, Shin YW, Taylor PA, Glen DR, Reynolds RC, Israel RB, Cox RW. Untangling the relatedness among correlations, part i: nonparametric approaches to inter-subject correlation analysis at the group level. Neuroimage 2016;142:248–59.

Chen G, Taylor PA, Shin YW, Reynolds RC, Cox RW. Untangling the relatedness among correlations, part ii: Inter-subject correlation group analysis through linear mixed-effects modeling. NeuroImage 2017;147:825–40.

Chung E, Romano JP. Exact and asymptotically robust permutation tests. The Annals of Statistics 2013;:484–507.

Eklund A, Nichols TE, Knutsson H. Cluster failure: why fmri inferences for spatial extent have inflated false-positive rates. Proceedings of the National Academy of Sciences 2016;113(28):7900–5.

Englander ZA, Haidt J, Morris JP. Neural basis of moral elevation demonstrated through inter-subject synchronization of cortical activity during free-viewing. PLoS One 2012;7(6):e39384.

Ge Y, Dudoit S, Speed TP. Resampling-based multiple testing for microarray data analysis. Test 2003;12(1):1–77.

Hasson U, Avidan G, Gelbard H, Vallines I, Harel M, Minshew N, Behrmann M. Shared and idiosyncratic cortical activation patterns in autism revealed under continuous real-life viewing conditions. Autism Research 2009;2(4):220–31.

Hasson U, Furman O, Clark D, Dudai Y, Davachi L. Enhanced intersubject correlations during movie viewing correlate with successful episodic encoding. Neuron 2008;57(3):452–62.

Hasson U, Nir Y, Levy I, Fuhrmann G, Malach R. Intersubject synchronization of cortical activity during natural vision. science 2004;303(5664):1634–40.

Herbec A, Kauppi JP, Jola C, Tohka J, Pollick FE. Differences in fmri intersubject correlation while viewing unedited and edited videos of dance performance. Cortex 2015;71:341–8.

Jääskeläinen IP, Koskentalo K, Balk MH, Autti T, Kauramäki J, Pomren C, Sams M. Inter-subject synchronization of prefrontal cortex hemodynamic activity during natural viewing. The open neuroimaging journal 2008;2:14.

Kauppi JP, Jääskeläinen IP, Sams M, Tohka J. Inter-subject correlation of brain hemodynamic responses during watching a movie: localization in space and frequency. Frontiers in neuroinformatics 2010;4:5.

Kauppi JP, Pajula J, Niemi J, Hari R, Tohka J. Functional brain segmentation using inter-subject correlation in fmri. Human Brain Mapping 2017;38(5):2643–65.

Kauppi JP, Pajula J, Tohka J. A versatile software package for inter-subject correlation based analyses of fmri. Frontiers in neuroinformatics 2014;8.

Nichols TE, Holmes AP. Nonparametric permutation tests for functional neuroimaging: a primer with examples. Human brain mapping 2002;15(1):1–25.

Nummenmaa L, Glerean E, Viinikainen M, Jääskeläinen IP, Hari R, Sams M. Emotions promote social interaction by synchronizing brain activity across individuals. Proceedings of the National Academy of Sciences 2012;109(24):9599–604.

Pajula J, Kauppi JP, Tohka J. Inter-subject correlation in fmri: method validation against stimulus-model based analysis. PloS one 2012;7(8):e41196.

Pajula J, Tohka J. Effects of spatial smoothing on inter-subject correlation based analysis of fmri. Magnetic resonance imaging 2014;32(9):1114–24.

Pajula J, Tohka J. How many is enough? effect of sample size in inter-subject correlation analysis of fmri. Computational intelligence and neuroscience 2016;2016:2.

Reason M, Jola C, Kay R, Reynolds D, Kauppi JP, Grobras MH, Tohka J, Pollick FE. Spectators aesthetic experience of sound and movement in dance performance: A transdisciplinary investigation. Psychology of Aesthetics, Creativity, and the Arts 2016;10(1):42.

Salmi J, Roine U, Glerean E, Lahnakoski J, Nieminen-von Wendt T, Tani P, Leppämäki S, Nummenmaa L, Jääskeläinen IP, Carlson S, et al. The brains of high functioning autistic individuals do not synchronize with those of others. NeuroImage: Clinical 2013;3:489–97.

Storey JD. A direct approach to false discovery rates. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 2002;64(3):479–98.

Storey JD, Tibshirani R. Statistical significance for genomewide studies. Proceedings of the National Academy of Sciences 2003;100(16):9440–5.

Tusher VG, Tibshirani R, Chu G. Significance analysis of microarrays applied to the ionizing radiation response. Proceedings of the National Academy of Sciences 2001;98(9):5116–21.

Wilson SM, Molnar-Szakacs I, Iacoboni M. Beyond superior temporal cortex: intersubject correlations in narrative speech comprehension. Cerebral cortex 2007;18(1):230–42.

Woolrich M. Robust group analysis using outlier inference. Neuroimage 2008;41(2):286–301.

Woolrich MW, Behrens TE, Beckmann CF, Jenkinson M, Smith SM. Multilevel linear modelling for fmri group analysis using bayesian inference. Neuroimage 2004;21(4):1732–47.

Xie Y, Pan W, Khodursky AB. A note on using permutation-based false discovery rate estimates to compare different analysis methods for microarray data. Bioinformatics 2005;21(23):4280–8.