TAMPEREEN TEKNILLINEN YLIOPISTO
TAMPERE UNIVERSITY OF TECHNOLOGY

Juha Pajula

# Inter-Subject Correlation Analysis for Functional Magnetic Resonance Imaging

Properties and Validation

Julkaisu 1378 • Publication 1378

Tampere 2016

Tampereen teknillinen yliopisto. Julkaisu 1378
Tampere University of Technology. Publication 1378

Juha Pajula

# Inter-Subject Correlation Analysis for Functional Magnetic Resonance Imaging

Properties and Validation

Thesis for the degree of Doctor of Science in Technology to be presented with due permission for public examination and criticism in Tietotalo Building, Auditorium TB109, at Tampere University of Technology, on the 22th of April 2016, at 12 noon.

Tampereen teknillinen yliopisto - Tampere University of Technology
Tampere 2016

**Supervisors:**

Jussi Tohka

Department of Bioengineering and Aerospace Engineering
Universidad Carlos III de Madrid, Spain

Department of Signal Processing,
Tampere University of Technology, Finland

Ulla Ruotsalainen

Department of Signal Processing
Tampere University of Techonology, Finland

**Pre-examiners:**

Miika Koskinen

Faculty of Medicine
University of Helsinki, Finland

Michael Hanke

Institute of Psychology
Otto-von-Guricke-University, Germany

**Opponent:**

Christian Windischberger

Center for Medical Physics and Biomedical Engineering
Medical University of Vienna, Austria

Faculty of Computing and Electrical Engineering
Tampere University of Technology, Finland

ISBN 978-952-15-3721-9 (printed)
ISBN 978-952-15-3723-3 (PDF)
ISSN 1459-2045

# Abstract

Inter-subject correlation (ISC) analysis for functional magnetic resonance imaging (fMRI) is a data driven approach to detect the brain activity during complex stimuli. The ISC measures are computed as a correlation between the fMRI time courses of the studied group of subjects. The ISC analysis is developed especially for fMRI studies with naturalistic stimuli, like movies, music, video games or annotated stories. The naturalistic stimuli are typically used to study the higher cognitive functions of a human brain such as emotions or humor.

This thesis investigates the properties of ISC analysis for fMRI. Three major aspects of the ISC analysis were studied: the accuracy of the analysis when compared with the general linear model (GLM) based analysis, the effects of spatial smoothing and the effect of sample size on the results. In addition, the openly available implementation of ISC analysis that was used in this study, ISCtoolbox for Matlab, was improved by developing a built-in support for cluster computing environments. The improvements were published with the new version of the ISCtoolbox. These improvements were mandatory due to the high computational costs of ISC analysis and the high number of ISC analyses required for the studies of this thesis.

Four international journal publications are included in the thesis. The first one describes the properties of the ISC analysis and ISCtoolbox for Matlab implementation. The second one investigates the accuracy of the ISC analysis with a block design fMRI data when compared with the GLM analysis. The third study investigates the effects of spatial smoothing on the ISC analysis and uses the GLM analysis as a reference for the testing. The fourth study tests how the sample size affects the ISC analysis results.

The ISC analysis was verified to be an efficient non-parametric data analysis method for fMRI data especially in studies with naturalistic stimuli. In addition to this the studies indicated that ISC can successfully be applied also to the traditional block design data, where it is able to detect activations with similar accuracy as the GLM analysis. The spatial smoothing was found to be a mandatory pre-processing step for the ISC analysis. When the thresholds for ISC results were corrected with false discovery rate multiple comparisons correction, the ISC analysis was able to tolerate slightly larger Gaussian smoothing kernels than the GLM analysis. The sample size investigation verified that the ISC analysis can produce fairly stable results when the number of subjects included in the study was more than 20. At least 30 subjects were found to guarantee the high similarity of the results. The implementation of the ISC analysis as ISCtoolbox was found to be highly efficient in cluster computing environments.

# Preface

The work of this thesis was carried out at the Department of Signal Processing (SGN) at the Tampere University of Technology (TUT) between fall 2010 and fall 2015. Most of the computations were conducted in Outolintu and Merope computing clusters.

I would like to express my sincere gratitude to docent Jussi Tohka and Vice Rector Ulla Ruotsalainen; you made this possible. I also want to thank the current and past members of $M^2$oBSI (Methods and Models for Biological Signals and Images) group; Antonietta Pepe, Elaheh Moradi, Defne Us, Jussi Forma, Uygar Tuna, Erman Acar and everyone else who I have met within the activities of the group. Special thanks for Jukka-Pekka Kauppi for all the discussions we had; let the work with ISCtoolbox continue also in the future.

Half of the thesis was funded by the assistantship of the SGN department from where I would also thank Sari Peltonen, Alpo Värri, Heikki Huttunen and everyone else who I have worked with. Thank you for Noora for a good company in OHBM conferences. Special gratitude I want to express for my pre-examiners, Michael Hanke and especially for docent Miika Koskinen: The comments during the examining process raised this work for the level it is now.

I also want to thank my parents Satu and Antti, Brother Miikka as well as Aunt Raija: At least you tried to understand my work. Lastly I want to express my deepest gratitude for Tiina: For the whole five and a half years you listened to my self-talk and especially I want to thank you for the support you gave me during the summer 2015 and January 2016. All Tuska festivals and concerts we had during these years, were the necessary breaks to keep my work going on.

Doing science is not a profession, doing science is a way of life.

Tampere, February 2016 Juha Pajula

# Contents

# Acronyms

| | |
|---|---|
| **3D** | Three dimensional |
| **AFNI** | Analysis of Functional Neuroimages |
| **AN** | Auditory naming |
| **BOLD** | Blood-oxygen-level dependent |
| **CPU** | Central processing unit |
| **EO** | External order |
| **FDR** | False discovery rate |
| **FEAT** | FMRI Expert Analysis Tool |
| **FWER** | Family-wise error rate |
| **fMRI** | Functional magnetic resonance imaging |
| **FPR** | False positive rate |
| **FRB** | Functional reference battery |
| **FSL** | FMRIB Software Library |
| **FWHM** | Full width at half maximum |
| **GLM** | General linear model |
| **GPGPU** | General-purpose computing on graphics processing units |
| **GPU** | Graphics processing units |
| **GRF** | Gaussian random field |
| **GUI** | Graphical user interface |
| **HA** | Hand imitation |
| **HRF** | Haemodynamic response function |
| **ICA** | Independent component analysis |
| **IDA** | Image data archive |
| **ICBM** | International Consortium for Brain Mapping |
| **ISC** | Inter-subject correlation |
| **LONI** | Laboratory of Neuro Imaging |
| **LPCA** | Local Principal Component Analysis |
| **MAE** | Mean absolute error |
| **MRI** | Magnetic resonance imaging |
| **MR** | Magnetic resonance |
| **NIDM** | Neuroimaging Data Model |
| **NiPyPe** | Neuroinformatics Python pipeline |
| **OM** | Oculomotor |
| **RAM** | Random-access memory |
| **RF** | Radio frequency |
| **ROI** | Region of interest |
| **SGE** | Sun Grid Engine |
| **Slurm** | Simple Linux Utility for Resource Management |
| **SNR** | Signal-to-noise ratio |

**SPM** Statistical Parametric Mapping
**W3C** World-Wide Web Consortium
**VG** Verbal generation
**ZPF** Fisher's Z transform

# List of Publications

I Kauppi J-P., Pajula J., Tohka J., "A versatile software package for inter-subject correlation based analysis of fMRI", *Frontiers In Neuroinformatics*, vol 8, no. 2, Jan. 2014.

II Pajula J., Kauppi J-P., Tohka J., "Inter-Subject Correlation in fMRI: Method Validation against Stimulus-Model Based Analysis", *PlosONE*, vol 7, no. 8, p. e41196 Aug. 2012.

III Pajula J., Tohka J., "Effects of spatial smoothing on inter-subject correlation based analysis of FMRI", *Magnetic Resonance Imaging*, vol 32, no. 9, pp. 1114-1124 Nov. 2014.

IV Pajula J., Tohka J. "How Many is Enough? Effect of Sample Size in Inter-Subject Correlation Analysis of fMRI", *Computational Intelligence and Neuroscience*, vol 2016, ID 2094601, pp. 1-10 Jan. 2016.

# 1 Introduction

## 1.1 Basics of Functional Magnetic Resonance Imaging

The principles and details of magnetic resonance imaging (MRI) and functional magnetic resonance imaging (fMRI) are described in various books. For this reason only simplified descriptions of these methods are repeated here. The descriptions are based on Hornak (1996-2014) and Poldrack et al. (2011).

Magnetic resonance (MR) measurements are based on orienting the molecules or atoms in a strong magnetic field, exciting them out of the orientation with a radio frequency (RF) pulse and then measuring the induced RF signal when the excited molecules or atoms are returning to the orientation. The imaging of different tissue properties requires different RF pulse sequences, for which reason the needed RF pulse sequence depends on the aimed application. The RF signals induced from exited molecules have different characteristics depending on the tissue properties of the measured location which makes possible to visualize different properties from the living tissue.

A typical task-based fMRI is measured when the subjects are under an active stimulus. The fMRI studies measured without the active stimuli are commonly referred to a resting state fMRI (Biswal et al., 1995). The most common fMRI measure type is blood-oxygen-level dependent (BOLD) fMRI (Ogawa et al., 1990b,a), which is based on the state of hemoglobin in the blood. The BOLD contrast is based on the paramagnetic properties of hemoglobin depending on whether it carries oxygen or not: Deoxygenated hemoglobin molecule is paramagnetic and oxygenated hemoglobin is not, which causes measurable difference in RF signal. This thesis is focusing only on the task-based BOLD fMRI studies.

From the physiological point of view, the basic principle of BOLD fMRI is that active brain areas consume more oxygen, which increases the blood flow to the active locations. The oxygen in blood is carried by hemoglobin molecules (Ogawa et al., 1998). When hemoglobin releases the oxygen, it becomes paramagnetic which decreases the measured MRI signal. On the other hand, the increased oxygen consumption causes an increase in the blood flow in the activated area. This produces the local surplus of oxygenated blood and increase in the measured MRI signal from this area. These phases are visible in detected BOLD response signals: First there is small undershoot compared with the baseline due to the increased use of oxygen, then a clear peak over the baseline is detected due to increased oxygenated blood flow and finally, the peak is decreased to the baseline again when the surplus of oxygenated blood ends.

By examining the BOLD response signals, it is possible to detect the areas where the oxygen consumption is increased due to neuronal activity. This makes the BOLD fMRI an indirect measure as it does not measure the actual brain activation, but a phenomenon

which is triggered by the activity of the brain cells in a certain location. The correlation between the increased oxygen consumption and the neuronal firing was proven in the studies of Mukamel et al. (2005) and Nir et al. (2007). Due to the indirect nature of the BOLD measurement, there is a delay of 2-6 seconds between the BOLD response compared with the actual neuronal activity. With model based analyses it is critical to take this delay into account among with other physiological properties of BOLD signal (Logothetis, 2003; Poldrack et al., 2011).

## 1.2 Functional Magnetic Resonance Imaging Data Analysis

The starting point for the task-based fMRI analysis is the data acquisition with an external stimulus, that can be used also to trigger the subject to perform various tasks. The subject is concentrating on the stimulus when he is measured with a MRI scanner. The stimuli can practically include instructions to perform tasks of any kind for the subject, for example watching pictures, hearing music or tapping a finger according to the stimulus. In practice, the tasks can be anything which is possible to perform actively in the scanner without introducing too much distortion to the measurements.

After measurement, the fMRI data requires various processing steps before it is suitable for statistical data analysis. Together these steps are commonly referred to a data processing pipeline, which can be divided roughly into pre-processing steps and analysis procedures. Typical pre-processing steps are: image reconstruction and correction, registration and normalization as well as spatial and temporal smoothing. (Poldrack et al., 2011). Common aims for data pre-processing are to remove the measurement artifacts and improve the signal-to-noise ratio (SNR) of the data.

In task-based fMRI studies the analysis should be designed according to the used stimuli. And further, the design of the stimuli should take into account the analysis method, which will be used to analyze the data. Traditional stimuli to fMRI studies typically have a strictly controlled structure from where the parametric model is easy to derive. For example, a model is easy to derive from a block design study where subjects have to stay in rest otherwise and when they detect stimulus, such as a sound or an image, they have to push a button. From stimulus of this kind, a parameterized signal model is easy to define: The signal is "ON" when an image is shown and "OFF" when no image is present and the subject is in rest. A typical way to analyze these studies is to use a general linear model (GLM) based analysis. These kind of studies with a block design can reveal efficiently which brain areas are active when the subject reacts to the activating stimuli, but they are challenging to apply for data acquired with more complex stimuli.

Studying higher brain functions, such as emotions or fear, requires more complex study setup than just tapping with finger on the required moments. Typical stimuli, which people are facing in everyday life is extremely rich over multiple senses. It is challenging to mimic these kind of stimuli with strictly controlled block design type of study design and even more challenging to model with parametric models. Compared with block design studies a more complex approach is used in event related design (D'Esposito et al., 1999; Foerde et al., 2006), where the length of the events and timing is more random. The most complex stimulus is mimicking the real life situations as closely as possible and those are commonly called naturalistic. These kind of stimuli can be, for example, video games (Kätsyri et al., 2013), music (Glerean et al., 2012), annotated stories (Hanke et al., 2014) or movies (Bartels et al., 2008; Hasson et al., 2004; Jääskeläinen et al., 2008; Kauppi et al., 2010). The fMRI studies with movie stimulus are typically referred to

neurocinematics (Hasson et al., 2008). Compared with a traditional strictly controlled stimuli, it is challenging to define a comprehensive model for a long movie due to the richness and complexity of it as stimuli. For this reason data-driven methods without any parametric model for stimuli are preferred.

The inter-subject correlation (ISC) analysis (Hasson et al., 2004; Kauppi et al., 2010, 2014) is a potential non-parametric method for naturalistic studies. The ISC analysis is purely based on the data itself as it simply computes the similarity between the corresponding time series of the subjects under common stimuli. The basic assumption in ISC is that subjects are engaged in the common stimuli in a similar manner within the subject group. This synchronization is then possible to detect using a correlation measure between the corresponding voxel time courses of the subjects' fMRI images.

## 1.3 Objectives

This thesis investigates the properties of the ISC analysis and validates the accuracy of the ISC analysis by comparing it with GLM analysis within the same data. The properties of ISC analysis were investigated in two studies. In the first study, the effects of spatial smoothing were studied and in the second study, the effect of sample size on the analysis results was investigated.

In addition, the openly available ISC implementation called ISCtoolbox for Matlab (Kauppi et al., 2014) was updated during the thesis work and with the first publication of this thesis the version 2.0 of ISCtoolbox was released. In the scope of this thesis the new version of ISCtoolbox had improvements in the computing speed and usability of the graphical user interface (GUI), as well as a new built-in support for parallel computing environments. The support for parallel environments was required due to the large number of ISC analyses applied for the tests of this thesis work.

## 1.4 Motivation

The current trends in cognitive brain studies are commonly aiming to study the higher brain functions associated for example to emotions or humour. These trends have lead the naturalistic stimuli to become increasingly popular. As the parametric model for complex naturalistic stimuli is challenging to derive, the data-driven methods such as ISC analysis or independent component analysis (ICA) (Mckeown et al., 1998) are more suitable for these studies. Where ISC searches the similarity of the responses within the group of subjects, ICA is typically used to search for the functional patterns in the brain. The validation of the ISC analysis makes the method more feasible for scientific community and the verification of the critical properties of it helps the users make more informed choices regarding the ISC analysis.

ISCtoolbox was originally developed to help researchers to apply the ISC analysis with a proper statistical methodology. Within this thesis it was further developed to make possible to apply larger, more data intense studies, with it. Large data intense studies require more computing power which commonly leads to an increased computing time and requirement of larger computing resources. To face these challenges many universities and research centers have their own computing cluster or have access to some shared computing resources. A typical challenge with highly efficient computing environments is that they are often complex to utilize. For this reason, the easy and efficient built-in

support for distributed computing was developed for ISCtoolbox. It was also aimed to make ISCtoolbox more appealing for non-technical scientists or clinical staff.

## 1.5 Structure of The Thesis

The thesis is based on four journal publications which all concern the ISC analysis. The first publication describes the ISCtoolbox implementation of the ISC analysis and the last three publications are investigating the ISC analysis properties.

The first publication describes the details of ISCtoolbox (Kauppi et al., 2014) as well as multiple smaller usability aspects of the toolbox. The improvements were needed for ISCtoolbox as the testing procedures required an extensive number of analyses and for this reason a built-in support for Simple Linux Utility for Resource Management (Slurm) (Yoo et al., 2003) and Sun Grid Engine (SGE) based computing grids (Love, 2013; Scalable Logic, 2013; Univa Corporation, 2013) were added to the toolbox. The second publication (Pajula et al., 2012) validates the accuracy of ISC analysis by comparing it with the traditional GLM analysis. The GLM analysis was used as a golden standard with the block design data. The validation demonstrated the detection accuracy of the ISC but left open how the ISC method behaves in different situations. The third publication (Pajula and Tohka, 2014) studies a typical question in fMRI analysis: How the spatial smoothing affects the accuracy of the analysis. The main conclusion was that the implementation of ISCtoolbox can handle spatial smoothing with a slightly wider Gaussian kernel than GLM analysis with the same data. The fourth publication (Pajula and Tohka, 2016) answers to another critical question for the study design: How many subjects should be included in ISC analysis to ensure the good quality of the results? The inspection of the results suggested that minimum of 20 subjects should be included in the ISC analysis, but more than 30 subjects should be used for reproducible results.

The structure of the thesis is the following. First, the common data processing pipeline for fMRI is introduced in Chapter 2. Chapter 3 presents the traditional study setup for block design analysis and the structure of a typical GLM analysis. In Chapter 4, the focus is moved to the studies with naturalistic stimuli and the ISC analysis. The main concepts and implementation considerations of ISCtoolbox are described in Chapter 5. Chapter 6 introduces the methods and concepts which were used to study the properties of ISC analysis. The journal publications included in the thesis are listed in Chapter 7. Finally, in Chapter 8, the outcome of the thesis work is discussed with the future aspects of the ISC analysis and Chapter 9 draws the final conclusions from the work.

# 2 Standard fMRI Analysis Pipeline

Data processing in fMRI starts from data acquisition and ends to the final statistical analysis. Between these, the fMRI data is reconstructed from the measured signals and pre-processed for the statistical analyses. Here the whole process from data acquisition to final analysis is referred to fMRI data analysis pipeline. The fMRI data has multiple artifact sources where the subject movement and the magnet field inhomogeneities in the scanner device are the most critical ones. In principle, the environment, subject and used devices can all cause different artefacts to the data and all of them should be corrected or taken into account in the analysis design. The required pre-processing steps depend also on the designed statistical analysis. For example, if the analysis is done for a group of subjects, the image data has to be converted into a common spatial domain.

This chapter describes the typical parts of the fMRI data processing pipeline and the main principles how they can affect the final analysis results. The main focus is given to the pre-processing as it is common for all analysis methods. Pre-processing has a significant role also when investigating the properties of a group analysis and so it has a critical role again in comparison of GLM and ISC methods. The GLM and ISC analyses are described in detail in Chapter 3 and 4. The details of the data processing pipeline here are following the description of Poldrack et al. (2011) and a typical pipeline used by the FMRIB Software Library (FSL) software package (Jenkinson et al., 2012).

## 2.1 General Structure of the Pipeline

Figure 2.1 illustrates three different pipelines used by four different software packages: ISCtoolbox, FSL and Analysis of Functional Neuroimages (AFNI) and Statistical Parametric Mapping (SPM). The pipelines of ISCtoolbox and FSL software packages were illustrated because they were used in this thesis. SPM was included as it demonstrates a slightly different approach in pipeline structure with normalization, spatial smoothing and statistical analysis. The first four steps illustrates the pre-processing steps which are handling the errors and the artifacts of the data. They are typical for all fMRI pipelines. The order of the last three steps depends on the selected software and the used analysis pipeline implementation. For example, a clear difference is between FSL and SPM software: FSL normalizes the data after the statistical analysis compared with SPM, which applies the normalization before the spatial smoothing.

According to Poldrack et al. (2011) the motion correction and slice timing correction can be applied in any order and the order should be determined separately for each study. In practice, the slice timing correction can improve the fMRI analysis results (Sladky et al., 2011), but in certain statistical analyses it can also be left out from the pipeline as proposed in Poldrack et al. (2011). As mentioned earlier, the FSL (Jenkinson et al., 2012;

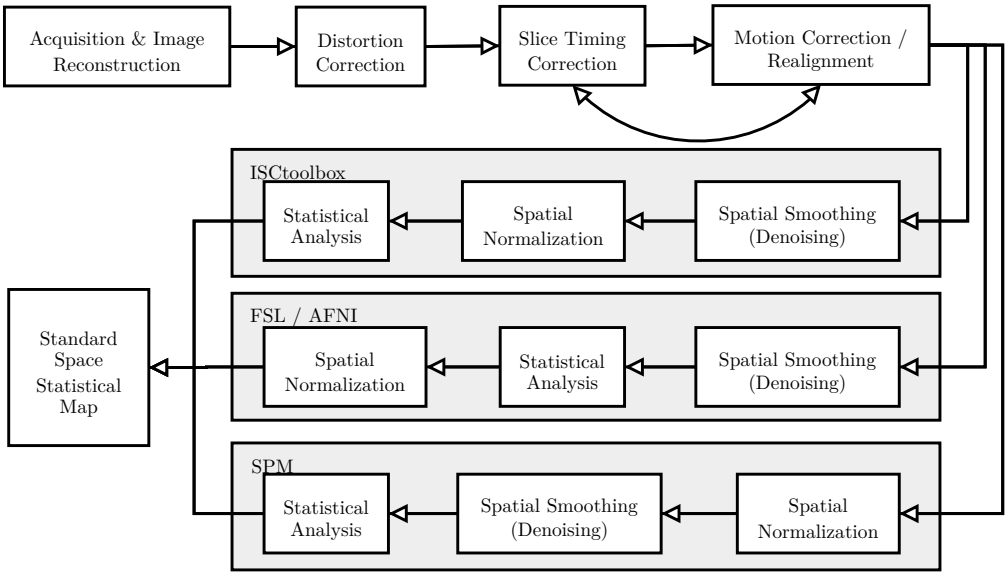



**Figure 2.1:** Data analysis pipeline for fMRI data. The acquisition & image reconstruction, distortion correction, slice timing correction and motion correction are typical for most of the fMRI pipelines compared with spatial smoothing, spatial normalization and statistical analysis which order depends on the used analysis software and study design. Traditionally acquisition & image reconstruction, distortion correction, slice timing correction, motion correction and spatial smoothing are considered as pre-processing steps, but also the spatial normalization is often included in pre-processing procedures.

Smith et al., 2004) and SPM (Friston, 2008) software have different naming and different order for normalization, smoothing and statistical analysis. The main reason why FSL and AFNI apply the analysis before normalization is that less spatial data points have to be analyzed in the original space. The normalization also requires typically less computing resources with a single statistical three dimensional (3D) volume compared with the multiple volumes of the fMRI data (Poldrack et al., 2011; Smith et al., 2004). In this thesis, the selected software were FEAT GLM analysis software (Beckmann et al., 2003; Woolrich et al., 2004) from FSL software package and ISCtoolbox for Matlab (Kauppi et al., 2014). The used pre-processing pipelines were the ISCtoolbox and FSL pipelines described in Figure 2.1.

## 2.2 Common Processing Steps

The image acquisition is the step where the fMRI data is measured. The subject is placed in the scanner and the data is measured at the same time as the subject concentrates on the given stimuli. Commonly, the image reconstruction from RF signals is done straight after the measurement by the scanner software.

The first pre-processing step after the fMRI data is reconstructed from the RF signals is the distortion correction. The scanner, the subject and the environment can cause distortions and variation to the homogeneity of the magnetic fields, which can produce artifacts to the measured data. With the distortion correction, the errors caused by the inhomogeneities of the magnetic fields are corrected.

As the 3D volumes of the fMRI data are measured slice-by-slice, all slices are not measured exactly at the same moment. The slice timing correction aims to correct the timing of the slices inside each volume at a synchronous time point (Sladky et al., 2011).

Critical to all fMRI analyses is the motion correction. As the image acquisition can last from a couple of minutes to dozens of minutes (see for example (Hanke et al., 2014)) the subjects typically move slightly inside the scanner. To minimize the movement, head frames have been developed to keep the head still during the session. Even if the subject externally stays still, the brain can still move slightly due to blood flow and breathing. As fMRI analyses are based on the time series of spatial voxels inside the brain the corresponding brain locations should be aligned in all time points. For this reason all spatial movement between acquisition volumes should be corrected carefully. In SPM software this procedure is referred to realignment instead of motion correction.

It is also possible to perform the slice timing correction after the motion correction. Both of these corrections reorganizes the data spatially and can cause errors to each other: The slice timing correction can mix the movement based intensity variation between volumes acquired in different time points. On the other hand, the motion correction can mix the data points between the slices acquired with slightly different moments around the current time point. The order of the procedures should be selected so that the common error is minimized. Poldrack et al. (2011) suggest that slice timing correction should be avoided when possible. They also point out that with short repetition times (2 seconds or less) the slice timing correction could be left out completely as most event related analyses are somewhat robust for small misspecification in timing.

## 2.3 Analysis Dependent Processing Steps

While the first processing steps were needed in almost every analyses, the spatial smoothing and normalization procedures are highly dependent from the selected analysis method and analysis type itself. As visible in Figure 2.1 the spatial smoothing, normalization and statistical analysis are done in different order depending on the selected software.

Spatial smoothing (Khullar et al., 2011a,b; Lindquist and Wager, 2008; Pajula and Tohka, 2014; Wink and Roerdink, 2004) is a critical procedure for most of the traditional fMRI analyses. Three common reasons to apply the spatial smoothing are: to increase the SNR, ensure the certain level of smoothness in the data, and to improve the spatial registration. When considering the motion correction, the spatial smoothing acts as a guarantee that at least some part of the corresponding time series from different subjects are overlapping spatially. At group-level studies this property of spatial smoothing is even more important due to the varying shapes and sizes of an individual brain. The improvement of SNR is the most critical reason for spatial smoothing and for this reason it is often called denoising. The traditional way to denoise data is to apply a 3D Gaussian smoothing kernel to the data. The size of the kernel varies, but a common rule of thumb is that the full width at half maximum (FWHM) of the smoothing kernel should be at least two times the univariate voxel size of the original data (Friston et al., 1996; Pajula and Tohka, 2014). For example, if the original fMRI data has the spatial voxel size of 2x2x2mm$^3$ then the selected FWHM of the smoothing kernel should be at least 4mm, but often the selected kernel is even larger than this. If there is no need for Gaussian smoothness in the data, also other kernels and methods such as wavelet denoising can be used (Khullar et al., 2011a,b; Pajula and Tohka, 2014). As the spatial smoothing averages the data

with the spatial neighborhood, it can also cause false positive voxels by spreading the data from activated brain area to the neighboring brain areas (Pajula and Tohka, 2014).

When spatial smoothing aims to improve the possibility to detect signals in voxels associated with the activated brain tissue, the spatial normalization makes it possible to compare the data from different subjects or studies. As a term, the spatial normalization refers to the procedure of registering the data from its original spatial space to some common template and thus into a common space of the template. Two commonly used templates are called Talairach (Talairach and Tournoux, 1988) and MNI-152 (Mazziotta et al., 2001) templates, from which the latter is more favorable as it is newer and defined from the average of 152 subjects instead of a single brain as the Talairach template.

The spatial accuracy of the fMRI data is typically lower than the accuracy of structural MRI data from the same subject. For this reason the high resolution structural MRI images are often used with the fMRI images to ensure the high quality image registration. In these cases, the fMRI data is registered to the corresponding MRI image and the MRI image is registered to the used common template. Then the affine transformation received from the registration of the MRI image to the common template is applied for the fMRI data which was registered to the MRI image. This is a common procedure for the data processing pipeline of FSL software.

The spatial registration methods can be divided in two main categories Poldrack et al. (2011): linear and non-linear methods. Linear methods allow only linear transformation (translations, scalings, and rotations) compared with non-linear methods which apply high dimensional morphing to the images. Linear methods cannot morph the image to exactly like the template, but they guarantee that the structure within the image stays the same. The non-linear methods can morph the image fit exactly to the template, but they cannot guarantee that the relative distances or sizes are no more the same (or that certain voxels are not combined). The selection between linear and non-linear registration should be done according to the data and the selected analysis type.

Image normalization is mandatory for all group studies, but it has multiple challenges also due to large variation in human brain anatomy. In general, brains from different individuals have the same structure, but when the brains are investigated in detail the sizes of the different structures and even the whole volume of the brain can vary significantly Poldrack et al. (2011). Various studies have investigated the differences and reasons for them (Zatorre et al., 2012). Learning of new skills can change the size of the brain areas which are associated to the activity involved in the skill. This was shown for example by Maguire et al. (2006) with taxi and bus drivers in London and Draganski et al. (2004) for a test group who learned to juggle. Within the studies with young people (Sanchez et al., 2012) as well as elderly people (Fjell and Walhovd, 2010) it has been noticed that aging changes the structure of the brain. All of these cause challenges for registering the image into a template which typically represents an average human brain in a certain scale of age. Further away from the average the subject is, more probable the registration errors are.

## 2.4 Data Analysis and Post-processing

The aim of pre-processing is to prepare the data for the statistical analysis and ensure the validity of the analysis. As noted earlier the type of analysis can affect the needed processing pipeline significantly and even require extra steps for data processing. In this thesis two analysis methods were used. The first was the ISC analysis which was validated

by the second method, the group GLM analysis. The GLM analysis was used as a golden standard in comparison with ISC.

With fMRI studies the data analysis methods can be divided roughly in two categories: parametric methods and data-driven methods. The traditional data analysis in fMRI is based on a parametric model which describes the used stimuli and takes into account the physical properties of BOLD fMRI signal. The data-driven methods normally do not have any parametric stimulus model as the analysis is mostly or completely based on the data itself. The ISC analysis is this kind of data-driven method. In the ISC analysis, the brain activity measure is based on the similarity of the corresponding time courses within the group of subjects. In practice, correlations are computed between time courses from one subject with the corresponding time courses from another subject and the procedure is repeated for all possible subject pairs in the study.

Another notable difference between the fMRI data analysis methods is that some methods concentrate only on a single-subject and other methods are applied in a group of subjects. The GLM analysis in its basic form is applied to a single-subject, compared with the ISC analysis which is always applied to a group of subjects. Also GLM can be applied to a group of subjects, but then the analysis is done at two levels. The analysis is first applied at a single-subject level and then the single-subject results are combined in the group-level analysis.

If the GLM analysis is applied at the group-level, the alignment within the group of subjects becomes crucial. The same applies with ISC: The common alignment is the most critical part of the pre-processing pipeline as the whole analysis is based on the assumption that subjects have some common behavior in brain activity when they get engaged to the common stimuli. If the corresponding time series between different subjects are not on the same spatial location, the whole analysis can fail. On the other hand, this is the reason why spatial smoothing is important for the ISC analysis. The correlations between the time series are larger if the SNR is higher, but the smoothing also ensures that at least some contents of the corresponding time series over different subjects are in the corresponding spatial locations. Compared with GLM, ISC has no requirements for the smoothness of the data (Pajula and Tohka, 2014; Woolrich et al., 2001) as there is no parametric model to fit into the time series.

After the statistical analysis, the results are typically saved for visualization or further investigation. The later use of the data requires not only saving the volumetric data itself but also the metadata, which describes how the digital volumetric data is physically mapped to the real world. Currently, most of the analysis software are supporting a medical image data format called Nifti[1] Cox et al. (2004). This data format includes the volumetric data as well as the metadata describing the properties of the volumetric data. Nifti files [*.nii*] can also be compressed with gzip[2] [*nii.gz*] to save hard drive space and some software such as FSL can use the compressed data straight without decompressing it first.

[1] http://nifti.nimh.nih.gov/
[2] http://www.gzip.org/

# 3 Typical fMRI Studies with General Linear Model

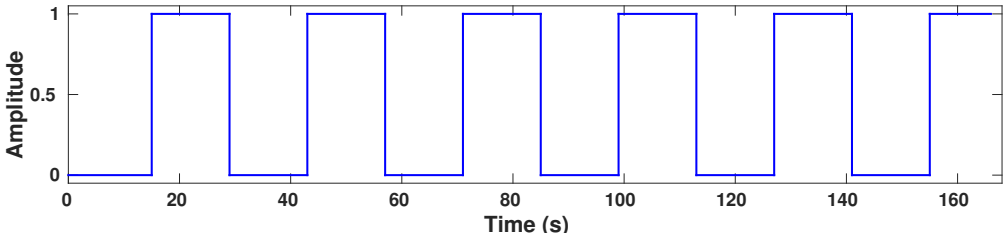
## 3.1 Typical Setup

Traditionally, fMRI studies have been focused on certain brain functions, for example, how watching pictures activates the visual cortex. The basic design for this kind of study has two aspects which are commonly defined together: The stimuli and the parametric model of the expected brain activation. For this reason the stimuli, such as watching the pictures, should have a temporal structure that is possible to describe with a parametric model. A simplified principle of the analysis is that the parametric model predicts the shape of the BOLD time course on the brain areas which have been activated from the given stimuli. In other words, the stimulus creates a BOLD response which shape can be predicted by the model. This typically limits the research to strictly controlled lab experiments that are fairly easy to predict.

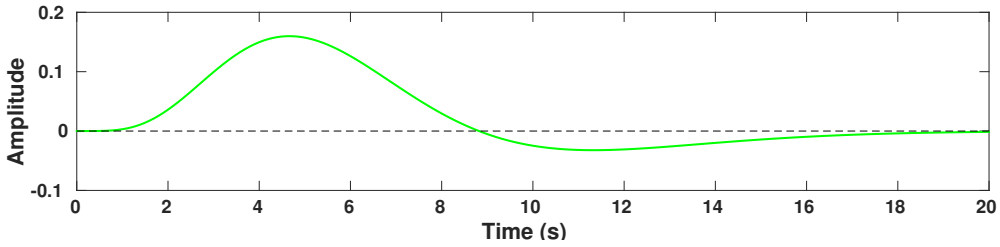
A typical parametric model for a BOLD fMRI analysis has a boxcar signal convolved with a canonical haemodynamic response function (HRF) (Friston et al., 1994). An example of a boxcar based model with real life fMRI time courses are presented in Figure 3.1. In principle, the boxcar model defines the active and rest phases of the stimuli. When the subject is resting the model is 'OFF', or zero, when the subject is active the model is 'ON', or one. The HRFs can be defined with various functions like single gamma function, double gamma function or finite impulse response basis set (Lindquist et al., 2009). Here the discussion is concentrated on double gamma HRF as it was used in the studies of this thesis. Double gamma HRF is defined with two gamma functions, where the first gamma function describes the main peak of the BOLD data and the second gamma function accounts the post stimulus undershoot. An example of double gamma HRF is presented in the panel (b) of Figure 3.1. The panel (c) presents the model of GLM analysis. The presented model is computed with FSL FEAT with a double gamma convolution and the boxcar model. The panel (d) presents 37 BOLD time courses acquired from the highest peak location of ISC analysis of the VG task in Pajula et al. (2012). In comparison of the mean time course (red line) and the parametric model of the panel (c) the similarity of the shape is clear.

### 3.1.1 Typical fMRI Stimuli

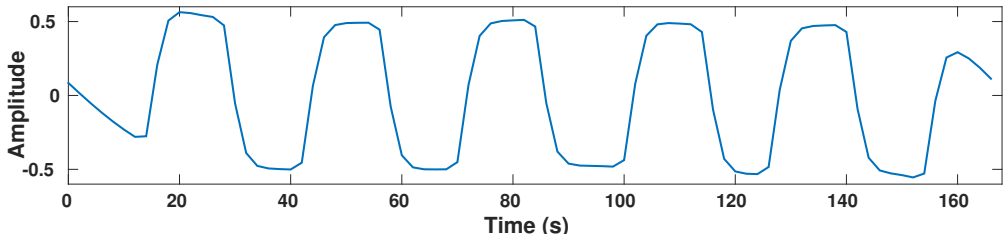
In task-based fMRI studies, the brain should be active during the measurement and in the case of a group study the activation should have a common source with every subject in the study group. In general, the stimuli of fMRI study can be anything that is possible to perform inside the scanner without causing too large movements for the head. The type



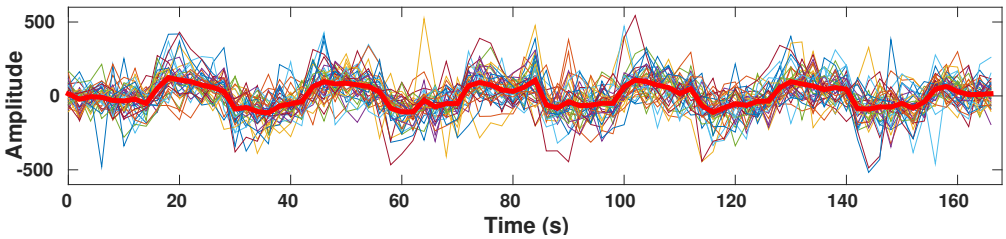

**(a)** Binary model for GLM analysis




**(b)** Double gamma HRF




**(c)** Binary model convolved with double gamma HRF




**(d)** Corresponding fMRI time courses from 37 subjects under the same stimuli.

**Figure 3.1:** Examples of signals from GLM analysis. The panel (a) presents the boxcar model for stimuli which has six 'OFF-ON' repeats on the block design stimuli. The panel (b) shows the double gamma HRF. The panel (c) presents the boxcar model convolved with double gamma HRF. The model was created with FSL FEAT for a GLM analysis. The panel (d) presents the time courses of 37 subjects from VG task included in the analysis of Pajula et al. (2012). The time courses were selected from the voxel with the highest peak value from the ISC analysis of the VG task. The thick red line is the mean over the time series from 37 subjects. The mean shape of measured BOLD signals clearly follows the model which corresponds to the results of Pajula et al. (2012)

of the stimuli depends purely on the aim of the study. For example, to measure which areas of the brain are associated to hand activity requires an activity where subjects have to perform a specific task with their hands according to the external stimuli. This kind of activity could be for example a button push in certain moments triggered by sound or shown pictures. The design of moments and button pushes are then an applicable starting point for the parametric model of the activation time course.

## 3.2 General Linear Model Analysis

One of the most commonly used fMRI data analysis method is the GLM analysis (Beckmann et al., 2003; Nelder and Wedderburn, 1972; Woolrich et al., 2004). It has been implemented in various analysis software like FSL (Jenkinson et al., 2012; Smith et al., 2004), SPM (Friston, 2008), NIAK[1] (Bellec et al., 2011), Freesurfer[2] (Dale et al., 1999) and AFNI[3] (Cox, 1996; Cox and Hyde, 1997). Even though the principle of the analysis is the same in all software there are also certain differences. As the studies included in this thesis are all applied using the FSL FEAT implementation of the GLM analysis, the descriptions of the analysis procedures here are mainly based on the FSL FEAT user guide[4], Poldrack et al. (2011) and Worsley and Friston (1995).

### 3.2.1 Single Subject Analysis

The basic principle in a single-subject analysis with the GLM is to predict the measured fMRI time series of the subject using the model which is commonly defined according to the used stimulus. As presented in Figure 3.1 the stimulus model is typically formed as a convolution between the selected HRF and the boxcar model describing the stimulus. What follows is a simplified presentation of the process, for more detailed description of the process see Poldrack et al. (2011).For a single-subject, the similarity between the model and the time course of each voxel is computed with GLM by fitting the model for the measured time courses. The basic GLM equation for a single time course is defined in Equation 3.1:

$$y(t) = a + b \cdot x(t) + e(t), \tag{3.1}$$

where $y(t)$ is the measured time course and $x(t)$ is the model which is defined according to the stimuli. The parameter $a$ is the baseline constant and $b$ is the model parameter. The $e(t)$ is the random error in model fitting. The standard assumption for the error is that it is normally distributed with zero mean. The estimation is applied by adjusting the parameters $a$ and $b$ so that the squared difference in the left and the right side of the equation is minimized. This leads to the estimate $\hat{y}(t)$ as defined in Equation 3.2:

$$\hat{y}(t) = \hat{a} + \hat{b}x(t), \tag{3.2}$$

where $\hat{a}$ and $\hat{b}$ are the found parameter estimates according to Equations 3.3:

$$\begin{aligned} \hat{b} &= r_{xy}\frac{s_y}{s_x}, \\ \hat{a} &= \bar{y} - \hat{b}\bar{x}, \end{aligned} \tag{3.3}$$

[1] http://www.nitrc.org/projects/niak
[2] http://freesurfer.net/
[3] http://afni.nimh.nih.gov/
[4] http://fsl.fmrib.ox.ac.uk/fsl/fslwiki/FEAT/UserGuide

where $r_{xy}$ is the correlation coefficient between, $x$ and $y$, $s_x$ and $s_y$ are the standard deviation of $x$ and $y$. The $\bar{x}$ and $\bar{y}$ are the average of $x$ and $y$.

After the time course estimate is found, the variance $\sigma^2$ of error $e(t)$ is estimated from the residual between the time course $y(t)$ and the time course estimate $\hat{y}(t)$ according to Equation 3.4:

$$e(t) = y(t) - \hat{y}(t) \tag{3.4}$$

from where the estimate of error variance $\hat{\sigma}^2$ is found by Equation 3.5:

$$\hat{\sigma}^2 = \frac{\sum_{t=1}^{T} e(t)^2}{T-2}, \tag{3.5}$$

where T is the number of time points in the signal. The statistical map is then computed using hypothesis testing based on parameter estimates $\hat{a}$ and $\hat{b}$ and $\hat{\sigma}$. For example, the t-statistic map corresponding to the null hypothesis that $b = 0$ is computed according Equation 3.6:

$$z = \frac{\hat{b}}{\hat{\sigma}/(s_x\sqrt{T-1})} \tag{3.6}$$

where the $\sigma$ is the stadard deviation of the error. The basic assumption on the GLM analysis is that the variance is constant over the whole data and the data has no autocorrelations. Both of these problems occur in fMRI data and for this reason they must be removed before the GLM analysis. Two main approaches are used to correct the autocorrelation problem: precoloring and prewhitening. In precoloring the data is low-pass filtered to increase the amount of known autocorrelation which then can be taken into account in the GLM modeling. In prewhitening, the amount of autocorrelation in the data is estimated and then removed from the data. Because the data is altered, the whitening must be applied also to the model of Equation 3.1 before fitting the model for the whitened data. In addition to prewhitening also temporal high-pass filtering is often applied to the data to remove the low frequency physiological noise which commonly exists in BOLD signals (Smith et al., 1999). This also affects to the study design: If the stimulus design includes repetitions in a too low frequency, the induced activation signal response may be removed from the time courses by the high-pass filtering.

### 3.2.2 Group Analysis

The group GLM analysis is based on the single-subject GLM analyses, which requires that the single-subject GLM analysis must be first applied to all subjects of the group study and then the group analysis is applied to the results of the single-subject analyses. For this reason the group analysis in GLM is commonly named as a second-level analysis. The group-level GLM analysis is applied with the same principle as the first-level analysis, but the inputs to the model fitting are now the estimates for the fitted first-level models of each subject included in the study.

The basic model for the group-level GLM can just have a weight of one for each subject, which defines the outcome as an average over the group of subjects. In general, two methods are common when modeling the group-level interfaces of the basic model. The simpler method is the fixed effects modeling, which uses only within-subject variance and ignores the cross-subject variance. This interpretation can be applied only within the used subject group but it cannot be generalized to larger population. The second modeling method is called mixed effects modeling which assumes that subjects included

in the analysis are randomly sampled from a larger distribution. The main difference between fixed and mixed effects modeling is that mixed effects modeling includes also the separately modeled between-subject (or the between-session) variance to the adjustment of the model.

In practice, GLM analysis can be applied also at three levels in studies where each subject has multiple sessions and the final analysis is done over these multiple sessions of the subjects. In this kind of study the first-level analysis is applied for all measurements of every subject within every session. The second-level is commonly applied over the sessions for the first-level single-subject results using the fixed effects group interface modeling. The fixed effects modeling is now accepted, as the analysis is concentrated only on a single-subject with multiple measuring sessions and it is not meant to be a general interpretation. The second-level produces then as many results as there were subjects included in the study and each of these subjects has a contrast between the sessions. The third level analysis is then applied for the second-level contrasts with mixed effects modeling. The mixed effects modeling is required on the third level analysis as the results should describe the whole population, not only the group of subjects included in the applied analyses.

## 3.3 Multiple Comparisons Correction

In principle, multiple comparisons corrections should be used with all statistical analyses to ensure the validity of used thresholds and the results acquired with them. An example of fMRI study without proper multiple comparisons correction was presented by Bennett et al. (2010). Recently, Hupé (2015) has pointed out multiple aspects how important it is to use the multiple comparisons corrections in any MRI study. Multiple comparisons corrections described here are mainly based on the definition of Ashby (2011).

The final phase for a fMRI analysis is to define which voxels are significantly active according to their statistics. Within the statistical map of group GLM analysis each voxel can be treated as a separated statistical test over the corresponding voxels of subjects included in the study. A simple statistical hypothesis for testing the activity the voxels could be that a voxel was not significant with a false positive rate (FPR) of $\alpha = 0.05$. This assumption requires that every voxel should be considered separately from each other. The FPR is defined in Equation 3.7:

$$FPR = \frac{FP}{FP + TN}, \tag{3.7}$$

where $FP$ is the number of false positive voxels (Type 1 error: non-activated voxels that are incorrectly defined as activated) and $TN$ is the number of true negative voxels (non-activated voxels that are correctly defined as non-activated). In practice, the definition allows five voxels from every 100 voxels to be defined as active even they were not (false positive). As an example, a single whole brain GLM statistic can have $10^5$ voxels. Even if there are no voxels with significant correspondence to model the FPR with $\alpha = 0.05$ hypothesis could still allow 5000 false positive voxels within the brain volume. Results of this kind are not acceptable and for this reason the testing hypothesis is commonly applied by defining the base assumption that each voxel is significant and the threshold allows only a limited number of false positives. This leads to the classic statement of the multiple comparisons problem: How to decide the criterion on each decision so that it guarantees that the experiment-wise false positive rate is lower than defined error rate

$\alpha_E$ (Ashby, 2011)? The FPR over all voxels in the study is referred to the family-wise error rate (FWER): The rate of type 1 error after applying multiple statistical tests in the study (Nichols and Hayasaka, 2003).

Due to the multiple comparisons problem it is highly inadvisable to define statistical thresholds for fMRI data without any statistical correction (Bennett et al., 2009, 2010). The problem is typically handled by controlling the FWER with multiple comparisons correction (Nichols and Hayasaka, 2003) with a suitable method. A simple statistical method used for controlling FWER is the Bonferroni correction, which assumes that all tests are independent (Ashby, 2011). The assumption of independence does not apply to fMRI data due to the spatial smoothing and physiological nature of BOLD signal. For this reason Bonferroni corrected thresholds are typically overly conservative with fMRI data (Poldrack et al., 2011).

A standard method for GLM analysis to control FWER is to use the multiple comparisons with a Gaussian random field (GRF) based multiple comparisons correction. With GRF correction the smoothness of the data is first estimated and the found smoothness estimates are used to evaluate the corrected thresholds with excursion set called Euler characteristic (Ashby, 2011; Worsley et al., 1992). The limitation of GRF based multiple comparisons correction is that it requires a certain level of smoothness for the data where it is applied. The rule of thumb, that at least two voxel FWHM kernel size for Gaussian smoothing should be used, is based on the smoothness requirements of GRF. Also GRF has shown to produce highly conservative thresholds. This may happen if the FWHM smoothness of the data is less than three or four times the voxel size. Another limitation is that the GRF correction tends to create overly conservative thresholds also if the number of subjects in the sample is less than 20 (Hayasaka and Nichols, 2003; Nichols and Hayasaka, 2003).

The false discovery rate (FDR) based multiple comparisons correction (Benjamini and Hochberg, 1995), popularized in neuroimaging by Genovese et al. (2002), has a different approach for handling the multiple comparisons problem. The basic idea behind the FDR method is that a few false positive values should be accepted. The FDR defines how many voxels are allowed to reject incorrectly based on the null hypothesis. In FDR, the rate $q$ defines that when considering all significance tests over every voxel it is guaranteed that $FDR \leq q$. In practice, the $q$ has a similar role as $\alpha_E$ with Bonferroni or GRF correction methods. According to the definition, FDR does not assume any spatial smoothness or any other spatial properties from the data. Compared with GRF this is a strength but it is also considered as a limitation of the FDR correction when it is used over the whole statistic (Chumbley and Friston, 2009). With FDR correction, it is not possible to define exactly which of the significant voxles were involved in the studied phenomenon as there is always certain amount of false positive voxels surviving from the threshold.

# 4 Naturalistic Stimulus and Inter-subject Correlation Analysis

## 4.1 Naturalistic Stimuli

The aim of naturalistic stimuli to fMRI is to simulate the everyday life encounters as closely as possible inside the MRI scanner. Naturalistic stimuli are needed as typical fMRI studies with the simple block design have limitations describing the complex environment people are facing in everyday life. It has been argued that studies with naturalistic stimuli can reveal the brain functions involved in daily encounters (Hasson et al., 2004, 2008; Jääskeläinen et al., 2008).

Studies with naturalistic stimuli typically aim to engage the subjects in the stimuli and by the engagement they aim to study the higher brain functions such as emotions (Botzung et al., 2010). Especially movies are found to be good stimuli for these studies (Golland et al., 2007; Jääskeläinen et al., 2008; Kauppi et al., 2010) and the studies using a movie stimulus are called neurocinematics studies (Hasson et al., 2008). Even though movies are the most common stimuli within naturalistic setups also other stimuli type such as computer games (Calhoun and Pearlson, 2012; Kätsyri et al., 2013) and listening to music (Abrams et al., 2013) or a narrated story (Hanke et al., 2014) are used.

With the naturalistic stimuli it is common that the stimuli have a complex non-parametric structure, which makes the parametric modeling of the stimuli challenging. With the movie stimuli, one possible approach is to annotate the movie and use the annotations as a starting point for the model. Annotations can be made with different criteria by professional or non-professional people. For example, researchers could rate a movie in short time windows according to the criteria if there is humor or not. This would then predict brain areas which are involved in the brain response connected to humor. The annotations could also be acquired from the studied subjects or a control group of subjects. In all of these approaches, the challenge is to determine how reliable and how comparable the individual ratings and model of the activity are compared with the actual activity triggered by the stimuli.

Non-parametric data driven methods such as ICA (Mckeown et al., 1998) or ISC (Hasson et al., 2008; Kauppi et al., 2014) do not require the parametric model of the data for the analysis. The ICA searches for similar independent signals from the data and it is typically used to detect the brain networks. A typical method in ICA is to search non-Gaussian signals with certain assumptions. A detailed overview of ICA is found in Hyvärinen and Oja (2000). The similarities between the GLM and ICA within the same data with simplified naturalistic stimulus were studied by Malinen et al. (2007). They found that ICA detected similar areas as GLM, but in addition many others which were

not detected by GLM. The main limitation with ICA is that it is complicated to apply the analysis at group level, as the detected networks vary between individuals and the detected networks may also vary between the analyses (Poldrack et al., 2011). Unlike ICA, the ISC analysis searches the similarity of the corresponding time series within the group of subjects. In the ISC analysis, the activity from the stimulus is assumed to be similar to all subjects in the study and in practice, the ISC analysis detects the areas where the brain functionality is the most similar within the group of subjects under the same stimuli.

## 4.2 Inter-subject Correlation Analysis

The origin of ISC analysis for fMRI is in the work of Hasson et al. (2004), where they presented clips from the movie "The Good the Bad and The Ugly" (Vincenzoni and Leone, 1966) and studied the similarities between the BOLD responses of the subjects. The similarity was computed with Pearson's correlation between the corresponding time series between all possible subject pairs.

### 4.2.1 Test statistic for detection of BOLD response similarity

The ISC analysis assumes that subjects under the same stimuli are engaged in the stimuli in a similar manner. This synchronization between the subjects is investigated by computing the voxelwise correlation between each possible subject pairs in the group of subjects. The principles of GLM and ISC analyses are compared in Figure 4.1. The ISC is computed between the corresponding time series of the subjects when GLM is computed between the model and the time series of the subject.

In principle, the correlation coefficient[1] is a natural measure of similarity between the fMRI time series of the studied subjects. Averaging the correlations of all possible subject pairs is a simple approach to study the triggered activity across the study group. In this thesis, the used implementation of ISC analysis was the ISCtoolbox, where the mean average of the voxel-wise correlation coefficients is defined as (Kauppi et al., 2010):

$$\overline{r} = \frac{1}{N(N-1)/2} \sum_{i=1}^{N} \sum_{j=2, j>i}^{N-1} r_{ij}, \tag{4.1}$$

where $\overline{r}$ denotes a group-level ISC in a given voxel, $N$ is the number of subjects in the study, and $r_{ij}$ is the correlation coefficient between the fMRI time series of subjects $i$ and $j$. The $r_{ij}$ is defined as:

$$r_{ij} = \frac{\sum_{n=1}^{N} [(s_i[n] - \bar{s_i})(s_j[n] - \bar{s_j})]}{\sqrt{\sum_{n=1}^{N} (s_i[n] - \bar{s_i})^2 \sum_{n=1}^{N} (s_j[n] - \bar{s_j})^2}}, \tag{4.2}$$

Here N is the total number of samples in time series, $s_i$ and $s_j$ are the time series obtained from the $i$th and $j$th subject, respectively, and $\bar{s_i}$ and $\bar{s_j}$ denote the means of $s_i$ and $s_j$. Because $r_{ii} = 1$ and $r_{ij} = r_{ji}$, the correlation coefficients are needed to compute only across $N(N-1)/2$ subject pairs (instead of $N^2$ subject pairs).

[1] Here a standard Pearson's correlation coefficient.

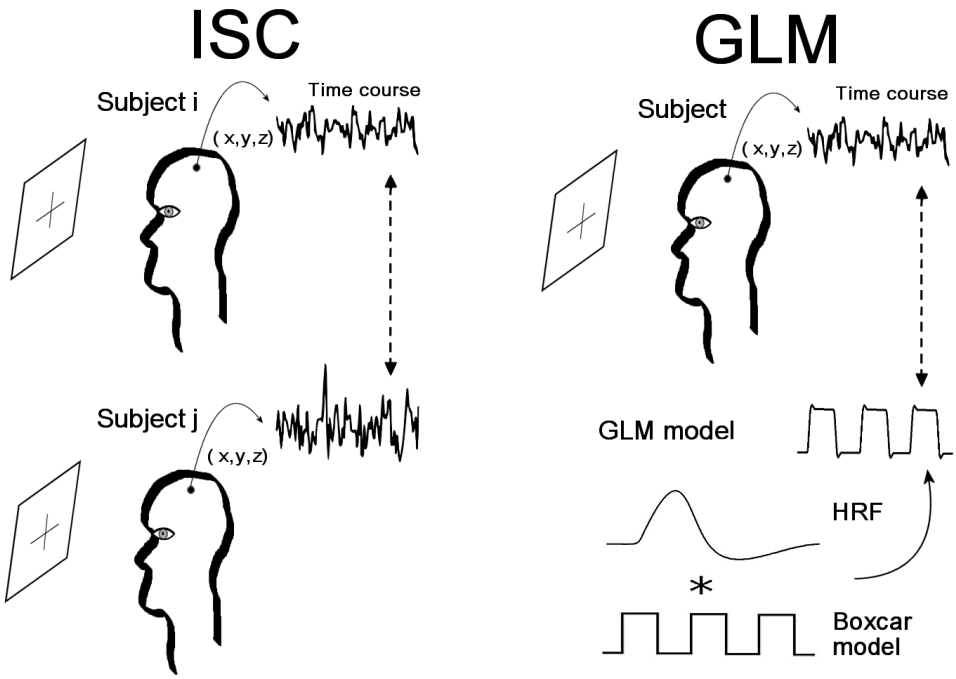



**Figure 4.1:** Comparison of principles behind ISC and GLM methods. The ISC analysis computes the similarity between the corresponding time courses of the subjects when the GLM computes the similarity between the time courses and the defined parametric model.

Even the principle of the analysis is simple, it has practical challenges with the implementation. The number of subject pairs increases roughly quadratically when $N$ increases and as Equation (4.1) is applied for every voxel within the brain, the number of correlation coefficients to be computed may be extremely high. This can slow down the computations and increase the memory and hard drive space requirements to challenging levels.

### 4.2.2 Statistical Significance of ISC

A critical question for any analysis method is how to define what is significant in the results. If the threshold for significance is defined by hand, it is possible to select a threshold so that the resulting statistic just looks nice. The situation can be even more severe if the results do not contain any significant voxels, because then the surviving voxels are only false positives (See the study of Bennett et al. (2010) as an example). The correlation coefficients $r_{ij}$ in Eq. (4.1) are not independent because each subject is present in more than one subject pair (e.g. $r_{ij}$ and $r_{kj}$ are overlapping because they both depend on the same time series measured from subject $j$). Also BOLD fMRI signals are temporally correlated, for which reasons the standard tests for assessing the significance of $\overline{r}$ are not valid.

Various methods have been applied to define the statistical significance for ISC analysis. The original ISC analysis of Hasson et al. (2004) used cluster based Monte Carlo simulations to ensure that in any clusters the minimum corrected significance level was $p < 0.05$. In the study of Wilson et al. (2008) the ISC analysis was applied by first creating paired ISC statistical maps from all subject pairs and then performing the group-level

analysis over these maps. Under the null hypothesis that the expected value is zero, the group analysis was testing that which test-statistics of voxels were significantly greater than zero. They defined the statistical significance for group-level with uncorrected $p < 0.005$ voxel level thresholds, which were then corrected with minimum cluster-size based GRF correction. In Lerner et al. (2011) the statistical significance was defined with FDR corrected $q = 0.05$ thresholds, where the FDR correction was based on the null distribution which was computed using bootstrapping with a phase randomization procedure. They bootstrapped every empirical time course in every voxel 1000 times for creating the null distribution.

Implementation of ISCtoolbox uses a fully non-parametric resampling method to evaluate the significance of $\overline{r}$ (Kauppi et al., 2010). The used test against the null hypothesis is that $\overline{r}$ statistic is the same as for data with no specific time structure. In the toolbox, the "null" resampling distribution is computed by circularly shifting the time series of every subject by a random amount, which removes the time alignment across the subjects. The $\overline{r}$ statistic is then calculated with the randomized time series. The procedure is required to account the temporal autocorrelations of the fMRI data. In practice, it is nearly impossible to calculate all possible time shift combinations and for this reason the needed distribution is approximated with a large finite number of realizations. This is applied by randomizing the test across voxels and time points. By default, ISCtoolbox uses 100 million resampling iterations for creating the null distribution. The default number of iterations has been selected experimentally. The critical thresholds for significant ISCs are obtained in two phases. First the $p$-values of the true realizations are computed for each voxel from the null distribution and then the found $p$-values are corrected with FDR based multiple comparisons correction (Benjamini and Hochberg, 1995). The FDR correction was described in Chapter 3.

# 5 ISCtoolbox Software Implementation

ISCtoolbox for Matlab[1] (Kauppi et al., 2014) is an openly available software implementation of the ISC analysis. The original implementation (Kauppi et al., 2010) was further developed during this thesis by improving the efficiency of the implementation and adding a built-in support for the parallel computing environments. In addition to the basic ISC analysis with FDR corrected thresholds, ISCtoolbox includes also other ISC based analyses. The frequency band specific ISC analysis and the time window based ISC analysis are both based on the same analysis described in Equation 4.2. In addition to these two, ISCtoolbox includes also a phase synchronization analysis and a session comparison analysis between two separated ISC analysis sessions. In this thesis, these advanced methods were only applied in the testing of computing times in different computing environments. The subsection 5.1 shortly describes these features of ISCtoolbox in the scope of computational implementation. More detailed description of them is found in Kauppi et al. (2014).

## 5.1 Key Functionality of ISCtoolbox

**The frequency specific ISC analysis** separates the original full main band for equally divided and separated frequency bands. For each study with frequency specific ISC, the toolbox creates an equally divided filter bank, which is used to create the new frequency specific time series from the original time series. The ISC analysis is then applied for each frequency bands separately and separated thresholds are defined for each of them. In computational sense, the basic ISC analysis is applied $M + 1$ times, where $M$ is the number of frequency bands. The single run is the main band, which is the same as the basic ISC analysis without frequency specific ISC analysis.

In **the time window analysis** the time series are divided either in overlapping or nonoverlapping time windows. The basic ISC analysis is then applied separately: The voxelwise correlation is computed between corresponding time series within each time window. This creates a separated correlation value for each time window and each voxel gets a time activity curve according to the average correlation over the selected time windows. Computationally this means that ISC analysis is applied $N + 1$ times where $N$ is the total number of time windows and the single run is again the basic ISC analysis for the whole time series. For the significance testing of the time window analysis, the generation of the null distribution is randomized across all time windows which leads to a common threshold for all windows.

[1] https://www.nitrc.org/projects/isc-toolbox

**The inter-subject phase synchronization** is computed based on Equation 4.1, but replacing the Pearson correlation with the absolute angular distance measure (Vinck et al., 2010) and applying the Hilbert transform (Goswami and Hoefel, 2004) to the fMRI time-series to obtain their corresponding analytic signals. Time-window analysis as well as low frequency sub-band analysis have challenges to get a meaningful interpretation from the ISC measures due to modest temporal resolution. The temporal resolution can be increased by using the phase synchronization as a similarity measure. The inter-subject phase synchronization measure based on the instantaneous phase allows the analysis of the band-pass filtered signals on the basis of inherent temporal resolution of the time series. This measure can be combined with time-window analysis by applying it separately to each time window. This can be used to create temporal inter-subject synchronization curves over the selected region of interest (ROI). See the details in Kauppi et al. (2014)

**The session comparison analysis** evaluates the ISC differences between two ISC analysis sessions. The analysis requires that the same subjects have been in both sessions. First, a Z-transformation is applied to the correlations and then paired Fisher's Z transform (ZPF) (Raghunathan et al., 1996) statistics are computed from them. At the end, the paired ZPF statistics are combined to a sum ZPF statistic (see details from Kauppi et al. (2014)). The null hypothesis for the significance testing is that there is no difference between the sessions. The thresholds are acquired from approximated distribution, which is generated by flipping the sign of pairwise ZPFs randomly before computing the sum ZPF. In the toolbox, the default number of random permutations for session comparison testing is 25 000, which is determined experimentally. The multiple comparisons in session comparison analysis are controlled with FWER (Nichols and Holmes, 2002).

## 5.2 Computational Considerations

All of the methods described above are extremely demanding computationally. Even the basic analysis can take a day to compute with a reasonable desktop computer if the data set is large. ISCtoolbox is implemented with the advanced memory mapping technique of Matlab[2] to reduce the memory limitations. With this technique, the main band ISC analysis is typically able to run only with 8GB of random-access memory (RAM). When analyses become larger and the resolution of the fMRI images increases, the memory requirement increases and typically at least 16 GB or even more RAM is required to be able to apply the analysis. To meet these increased requirements the latest improvement for the ISCtoolbox implementation was the built-in support for Slurm (Yoo et al., 2003) and SGE N1 based (Love, 2013) parallel computing environments.

In the basic work flow of ISCtoolbox, the user defines and validates the data and parameters of the analysis with the GUI and, when the parameters are valid, starts the analysis. After this, the toolbox takes care of the rest of the work including the grid engine calls in parallel environment. The original main reason for implementing the automated parallel support in the toolbox was practical: The validation of the ISC analysis required repeated analyses and without parallel computing environment (or if the parallelization of the processes had to be done by hand) it would had been challenging or even impossible to apply all of the needed analyses in reasonable time.

[2] "Memory-mapping is a mechanism that maps a portion of a file, or an entire file, on disk to a range of addresses within an application's address space. The application can then access files on disk in the same way it accesses dynamic memory." Cited from: http://se.mathworks.com/help/matlab/import_export/overview-of-memory-mapping.html

## 5.3 Parallel Implementation of ISCtoolbox

ISCtoolbox has been developed so that it is possible to run the analysis procedure step-by-step and if needed, run some or all of the steps in a computing cluster with a grid engine. Even though the desktop computers become more efficient every year, the amount of data in a fMRI study increases simultaneously. As described earlier, all analyses implemented in ISCtoolbox are computationally demanding and resources to meet the demand are more typically available on computing clusters.

The parallelization of the analysis process can reduce the computing time in a cluster environment, but in addition it can also lower the minimum requirement of RAM memory for the whole process. Before the built-in support for computing cluster environments, applying the analysis with ISCtoolbox in these systems required deep knowledge on how the cluster environment works. In practice, a separated shell script was required for each step and for a single analysis flow this would had required more than 100 script files to be written and executed by the user.

From a parallelization point of view ISCtoolbox uses two key features: the step-by-step design and memory mapping mechanism of Matlab. The step-by-step design enables the parallel computing to have a clear structure for work flow and have a clear restoration points if the computing is interrupted for some reason. The memory maps are used to lower the memory requirement of the analysis as with memory maps all of the data does not have to be in the memory at once. In the parallel implementation, the memory maps are also used to save the partial results of each step to the common network drive from the independent computing nodes.

ISCtoolbox has been designed to work locally on a single computer and on a cluster via available grid engine. The memory maps are used to map the input data to the hard drive and, when needed, read only the necessary parts of the data to the memory: for example, the time series from a single slice of every subject. The same mechanism is used to save the outcome of each step for the later use: different parallel processes can all save the outcome of the process to the same data structure. Further on, the data structure is also used to test if the outcomes of grid engine processes are already computed or not. This prevents the same step getting computed again, if it has already been computed. The mechanism is useful, for example, if the computations crash for some reason: The analysis can be restarted from the crashed process without processing the earlier parts of the analysis again.

In general, the implementation of parallel computations uses three components:

- A system to trigger the sub-processes for the required parts of the analysis.
- A system to wait until the sub-processes, which were triggered, are finished.
- A system to gather the results of the finished sub-processes before the next computing step is allowed to start.

In addition, also the main process must exist to take care of the analysis in general. The main process in ISCtoolbox is the process which is triggered by the GUI. If the analysis is run in cluster environment the main process uses the separated mechanisms to distribute the steps in the computing cluster, otherwise it calls the process steps directly without parallelization.

The first part of the parallelization system is the trigger system which creates the required shell scripts for the detected cluster environment and then triggers them on the grid engine. The currently supported engines are Slurm (Yoo et al., 2003) and the original SGE N1 based systems like Son of a Grid Engine (Love, 2013), Open Grid Scheduler (Scalable Logic, 2013) and Univa Grid Engine (Univa Corporation, 2013). ISCtoolbox works with any operating system where Matlab is available (Windows, Linux, OS X), but all of these grid engines are available only in Unix based environments. This limits the usage of parallel environment within the Unix based operating systems. For this reason, the toolbox detects the available engine automatically by testing the specific environment variables.

The second part of the parallelization system is the waiting mechanism for the triggered processes. The version 2.0 of the toolbox (described in Kauppi et al. (2014)), was waiting that all processes, from the user who had started Matlab, had finished on the cluster and only then the system was allowed to continue the computations. This was noted as a bad idea in practice, because in many cluster environments also the main process would run in its own cluster process. Based on this limitation, the waiting process was modified in the current version 2.1 of the toolbox to wait for only the currently triggered processes. In this process-oriented design it was required to know the exact process identifier numbers from the trigger system and for this reason the trigger process was modified to gather the process identifiers and pass them to the waiting system. In addition, the waiting system is not only pausing the main process, but it also monitors the errors of the sub-processes. If any monitored process creates an error, it kills all of the monitored processes and stops the main process with an error.

The last part of the parallelization system is the gathering process, which is triggered always when the processed step had saved something to the memory maps. If the process had finished without errors, the used map is marked as ready. This is required so that the finished processes are not ran again if the analysis is restarted for some reason. Even though this was the simplest task in principle, the actual implementation of it caused the largest changes for the toolbox implementation. The original functionality for marking the steps finished was inside the individual functions of each analysis step and for parallel environment this was moved into a separated function.

Figure 5.1 illustrates how a single process is run in the Slurm based cluster. At first the main process calls the gridParser function and passes the information about which parameters and which process is triggered. Then gridParser creates and saves the required script file according to the detected grid engine and submits the process to the grid engine which runs it. After this, the gridParser parses the process identifier from the grid engine output and returns it to the main function. When all processes of the current step are submitted, the main process passes the collected process identifiers to the waitGrid function. The waitGrid starts to poll if any of the processes had written anything in the error output files and if all processes in the identifier list would already have been finished. When all processes are finished without error, the waitGrid returns to the main process. As a final phase, the main process calls the gatherGridPointers function if needed. This function marks the results of each ran process write protected if they were run without errors and finally clears the temp files from the current step.

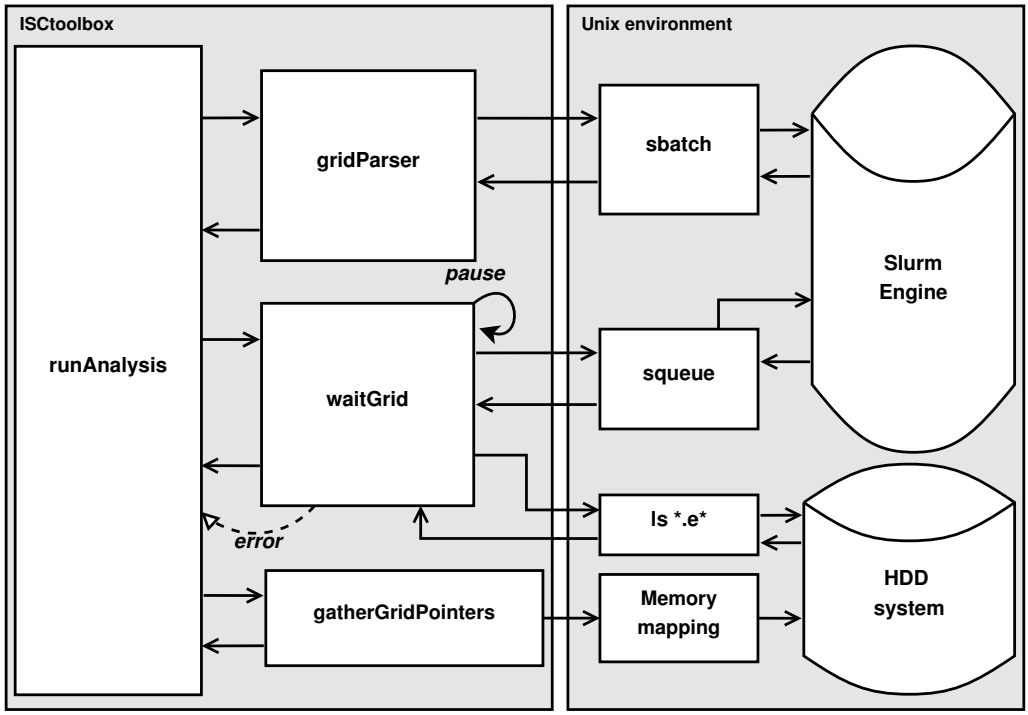



**Figure 5.1:** Illustration how a single process is ran with the Slurm cluster environment in use. First the required scripts are written and submitted by the gridParser function. Then waitGrid function pauses the main function until the whole step is finished. At the end, if needed, the gatherGridPointers function protects the results of the current step from the future write operations.

# 6 Properties and Validation of Inter-subject Correlation Analysis

This chapter describes the materials and methods used to study and validate the ISC analysis. In order to quantitatively validate the analyses, Dice index and correlation measures were used to compare the results from ISC and GLM analyses. These measures are described in Section 6.1. Section 6.2 describes the used fMRI data sets and section 6.3 the used setups with simulated data for comparisons of ISC results against GLM results. At the end, the sections 6.4 and 6.5 bring together the studies to validate the ISC analysis as well as measure the Gaussian smoothing and sample size effects on the ISC analysis results.

## 6.1 Similarity Measures

Most of the comparisons between ISC and GLM analyses in this thesis were applied by using two similarity measures: Pearson's correlation (Pearson, 1896) and Dice index (Dice, 1945). The full statistical 3D maps from ISC and GLM analyses were compared with Pearson's correlation. The Dice index was used to compare the binary maps, which were computed by applying thresholds to the full statistical maps. The Pearson's correlation and Dice index measures were also used with different analysis setups to compare the effects of spatial smoothing and sample size. Pearson's correlation was used to compare the unthresholded statistical maps from the analyses and the Dice index was used to measure the similarity of binary maps representing the significant voxels after applying the statistical thresholds for the ISC and GLM statistics. Here the equations for comparison methods are defined for two statistical maps $M$ and $L$. In studies Pajula and Tohka (2014); Pajula et al. (2012) and Pajula and Tohka (2016) these measures were used within or between the statistical maps from ISC and GLM analyses.

### 6.1.1 Pearson's Correlation

Pearson's correlation coefficient was used to compare the unthresholded statistical 3D maps from ISC and GLM analyses. The equation 6.1 defines Pearson's correlation for two statistical maps M and L:

$$C = \frac{1}{N-1} \sum_{k=1}^{N} \left( \frac{m_k - \bar{M}}{s_m} \right) \left( \frac{l_k - \bar{L}}{s_l} \right), \tag{6.1}$$

where $N$ is the total number of voxels in the comparison. The comparison is applied only for the brain voxels and then $N$ refers to the number of voxels inside the brain. The $m_k$

and the $l_k$ are the $k$th voxels from the statistics in the comparison. The $\bar{M}$ and $\bar{L}$ are the sample means of the volumes $M$ and $L$ respectively and the $s_m$ and $s_l$ are the standard deviations of these volumes.

### 6.1.2 Dice Index

The Dice index (Dice, 1945) was selected for comparing thresholded and binarized statistical maps. The binary maps were created by assigning the voxel to one if the corresponding value of the statistical map passed the threshold and otherwise to zero. If $B_m$ represent the set of the significant voxels of one statistic and $B_l$ the set of second statistic, then the Dice index between these two sets was defined as equation 6.2:

$$I_{Dice} = \frac{2|B_m \cap B_l|}{|B_m| + |B_l|}, \tag{6.2}$$

where the numerator defines the number of significant voxels common to both sets and the denominator the number of significant voxels according to both sets individually. The equation measures the number of voxels where both sets are one against the number of voxels where at least one set has significant voxels. In practice, the Dice index was applied in all studies by reshaping the 3D volumes as N-dimensional binary vectors. For them, the Dice index was redefined according to Equation 6.3:

$$I_{Dice} = \frac{2 \sum_{k=1}^{N} (B_m[k] \cdot B_l[k])}{\sum_{k=1}^{N} (B_m[k]) + \sum_{k=1}^{N} (B_l[k])}, \tag{6.3}$$

where $B_m[k]$ and $B_l[k]$ are the $k$th voxels of binary vectors from the reshaped binary volumes, respectively. The $N$ is the number of voxels in the volume.

The resulting Dice index values vary between 0-1, where 1 denotes the exact similarity and 0 corresponds to situation when there is no overlap between the compared binary maps. To simplify the interpretation of Dice measures in Pajula and Tohka (2014); Pajula et al. (2012) and Pajula and Tohka (2016), the relationship between the Dice index and Kappa coefficient (Cohen, 1968) was utilized. This property was defined according to Zijdenbos et al. (1994) where they showed that under certain assumptions, which are valid in these studies, the Dice index is (asymptotically) equal to Kappa coefficient. Based on equality the categorization of Landis and Koch (1977) for Kappa coefficients was used in all Dice tests to interpret the results.

The categorization of Landis and Koch (1977) has six categories: “No agreement” (index less than 0), “Slight agreement” (index 0 - 0.2), “Fair agreement”(index 0.2 - 0.4), “Moderate agreement” (index 0.4 - 0.6), “Substantial agreement” (index 0.6 - 0.8), “Almost perfect agreement” (index 0.8 - 1.0). Note that the first category is defined also below zero as Kappa coefficient can have negative values. These categories are ad-hoc, but widely used in practice. The relationship between Dice index and Kappa coefficient is described in detail by Finch (2005).

The Dice index was chosen instead of Kappa, because it suited better for purposes of these studies since it ignores the non-activated regions (see Zijdenbos et al. (1994) for more details) and it is widely used as a performance index in the evaluation of medical image segmentation algorithms.

## 6.2 Materials

All validation and testing studies in this thesis are based on data which is acquired from the Laboratory of Neuro Imaging (LONI) image data archive (IDA) database [1]. The selected data includes block design stimuli based on International Consortium for Brain Mapping (ICBM) functional reference battery (FRB)[2] (Mazziotta et al., 2001) definition and it is a part of large fMRI data collection which is free for scientific use for all researchers (Mazziotta et al., 2001). The FRB is developed to reliably find functional landmarks from a group of subjects. The definition of FRB stimulus describes five different stimuli for a typical GLM analysis having a block design structure with six 'OFF-ON' blocks and they are named as auditory naming (AN), external order (EO), hand imitation (HA), oculomotor (OM) and VG tasks according to the type of activity included in the stimuli.

Compared with the button push example in Section 3.1, in every FRB stimulus the 'OFF' blocks are similar: The subjects have to respond with a left mouse button push when they see an arrow pointing to the left. The 'ON' blocks depends on the used stimuli design:

- In the **AN task**, the subjects hear an audio description of some object and they have to think about the answer to the description silently in their mind. This setup is then repeated on each 'ON' blocks with different descriptions. Stimulus of this kind considers the understanding of language from auditory input and according to the FRB definition the auditory cortex with the language areas should be detected in the analysis.

- In the **EO task**, four different abstract images are presented to the subjects at first. Then a fifth image is shown separately and the subjects have to remember if the fifth image was within the earlier four images. If the subjects thought that the image was within the earlier images, they have to press a button. This test is repeated during each 'ON' blocks with different images. This is a working memory task and it is noted that the stimuli should reveal the active memory areas of the brain.

- In the **HA task**, subjects have to mimic the hand gestures with their right hand according to the presented image. The image is different on each 'ON' block. According to FRB definition, this task requires higher order motor coordination and motor planning from the subjects and it is assumed to activate the frontal and parietal areas of the brain.

- In the **OM task**, an abstract figure is shown to the subjects. It includes a cross in the middle surrounded by 10 black boxes. Subjects have to concentrate on the cross and saccade to the surrounding box if it changes to white. After this, they should return their gaze immediately back to the cross. The task is assumed to reveal the visual system and the occipital lobe.

- In the **VG task**, the subjects are watching the images of common objects and as a response to each image they have to silently think a verb associated to the object in their mind. This task is testing the language system with visual input and should activate the language and visual areas of the brain.

[1] https://ida.loni.usc.edu/login.jsp?project=ICBM
[2] http://www.loni.usc.edu/ICBM/Downloads/Downloads_FRB.shtml

The detailed descriptions of the stimulus and the data used in this thesis are found in Pajula et al. (2012). The main reason why subjects had just to think about the response to the tasks where an answer was required, was that the speaking could cause challenging movement artifacts to the data. According to Rosen et al. (2000) thinking causes more or less the same activation in the brain than speaking the same output in loud.

The described stimulus design allows an easy analysis with a boxcar model to find the areas of the brain which are involved in the described activity. The main limitations of the FRB and similar designs are that in everyday life the brain does not have this kind of simple limited stimulus, but an extremely complex combination stimuli of all kind. For this reason these kind of studies cannot exactly show how the individual brains are functioning in everyday life situations and more complex stimulus such as naturalistic stimulus are needed.

## 6.3 Tests with Simulated Data

In Pajula et al. (2012) the results of ISC analysis were compared with the results of GLM analysis. The study was designed to test if the ISC analysis detects similar activations compared with the GLM analysis when the data is selected optimally for the GLM analysis. The agreement between the results of ISC and GLM methods was measured with Pearson's correlation and Dice index described in Section 6.1. In addition to measured fMRI data, a simulated data was used to verify the study setup. The same simulated data was utilized again when the effects of smoothing for ISC analysis results were studied in Pajula and Tohka (2014).

The simulated data included simulated fMRI time series, which were generated by using the GLM model with double gamma HRF (visualized in the third panel of Figure 3.1) as representing the plain time course in the activated regions. The modeled time course was included in all "active" voxels similarly in all simulated volumes with randomly generated $1/f$ pink noise. In each volume within all simulated data sets the regions of activation time courses were exactly the same, only difference between "subjects" was the random noise. Between the simulated data sets, the only difference was the SNR between the modeled signal and the pink $1/f$ noise. In (Pajula et al., 2012) four data sets were used and the SNRs of those were (peak to peak): 0.01, 0.05, 0.02 and 0.001. In (Pajula and Tohka, 2014) only one simulated data set was used with SNR 0.06.

In the comparison studies (Pajula and Tohka, 2014; Pajula et al., 2012), both methods, ISC and GLM, were applied to the same simulated data to test the detection accuracy of the methods compared with the ground truth (voxels locations where the active signal was included in the generation of the simulated data). This was used in both studies to validate the comparison setup between ISC and GLM analyses. In Pajula et al. (2012) the detected areas of significant voxels from both methods were compared with the ground truth using the Dice index measure (see Section 6.1) to verify the validity of the GLM method for the golden standard of the study. In (Pajula and Tohka, 2014) simulated data was denoised spatially with Gaussian kernels and the accuracy of detection in both methods were verified with the specificity and sensitivity measures (compared with the ground truth of the simulation) over different Gaussain kernels. The sensitivity was defined as a percent of true positives against the sum of true positives and false negatives. Similarly, the specificity was defined as a percent of true negatives against the sum of true negatives and false positives.

## 6.4 Comparison Studies

The GLM analysis was justified as a golden standard for the ISC analysis by selecting the data for the analysis optimally according to the GLM analysis. The principle was the same as in the study of Hejnar et al. (2007). It was assumed that with a simple block design data, the GLM analysis can detect precisely the activation triggered by the given strictly controlled stimuli. Hence the block design stimuli were the only common stimuli to all subjects, it was hypothesized that ISC would detect the same activation as GLM. This setup was then used to compare the accuracy of ISC and GLM detection in environment where GLM would be an optimal choice and thus justified as a golden standard. The between-method comparisons were applied for five different fMRI data sets with five different FRB stimuli (see Section 6.2) (Pajula et al., 2012).

To minimize the influence from the other parts of the analysis process, the used pre-processing pipeline was exactly the same in both comparison studies (Pajula and Tohka, 2014; Pajula et al., 2012). In Pajula et al. (2012) the voxelwise correlation measure between the ISC and GLM methods varied between 0.69 and 0.83 depending on the used FRB stimuli. Correspondingly, the Dice measure was varying between 0.56 to 0.76 with q=0.05 FDR corrected thresholds and between 0.56 to 0.86 with q=0.001 FDR corrected thresholds. The highest values were on the category of "Almost perfect agreement" described in previous Section 6.1. On the other hand, the lowest values were on "Moderate agreement" category. The high variance between different tasks was explained as the different nature of the stimulus tasks. Some tasks had a highly similar response between-subjects, in other tasks the subjects could respond to the stimulus in various ways. The results of the correlation and Dice measures indicated that the ISC analysis can detect close to similar activation as GLM and in addition it could detect areas which were not detected by GLM. The areas which were not detected by GLM were hypothesized to belong to areas which had a similar response within most of the subjects, but the response was not following the block design structure of the GLM model.

As described in Chapter 2, the fMRI data is an indirect measure of brain activity and it includes many possible error sources. When considering these in addition to the computational limitations of the pre-processing pipeline, the denoising is a mandatory process for most of the fMRI data analyses. The most common spatial denoising methods is the spatial smoothing with Gaussian kernel. From the ISC analysis point of view, there is no requirement for the smoothness of the data, which is a common requirement within many model based methods. In general, with ISC analysis there are only two main reasons for spatial smoothing. Firstly, denoising is needed to improve the SNR of the data. Secondly, the smoothing improves brain alignment between individual subjects after registration. In practice, the final group alignment is improved by spreading and mixing the time courses between the closest voxels and so even slightly misregistered time courses can still overlap between-subjects and enable the efficient ISC analysis.

The spreading and mixing of signals in the spatial domain can produce also false positive voxels, for example, if too large kernels are used. This is a third aspect which should be taking into account with the implementation of ISCtoolbox. The larger kernel increases the SNR and spreads the activation in the neighboring voxels more than the narrow kernels and thus larger kernel enables higher correlation as well as larger *foci* after applying thresholds (Pajula and Tohka, 2014; Poldrack et al., 2011). As a negative effect, all small areas which are smaller than the kernel itself are removed by the smoothing and some of the areas covered by the larger *foci* might be false positive detection. In general,

this problem should be considered carefully when the pre-processing is designed. The smoothing kernel should be selected large enough that needed SNR is archived but so small that the number of possible false positives are minimized.

## 6.5 The Effects of Sample Size

A typical question with any group fMRI study is how many subjects should be included in the study. This question was investigated comprehensively within other analysis methods earlier (Churchill et al., 2014; David et al., 2013; Murphy and Garavan, 2004; Thirion et al., 2007; Zandbelt et al., 2008), but no study was made for ISC analysis before this thesis. In general, the number of subjects in a typical group fMRI study varies around 20 subjects, even though it has been shown that at least 26 to 30 subjects should be used to gain statistically reliable results (David et al., 2013). Unfortunately, the number of subjects in a typical fMRI study is less than that.

In Pajula and Tohka (2016) the number of subjects required for a stable ISC analysis was studied with a large 130 subjects fMRI data set with a block design stimulus. The ISC analysis was applied for the subsets of 10, 15, ..., 65 subjects from the full 130 subject data set. For each number of subjects, two non-overlapping groups of subjects were randomly selected for 1000 times. This lead to 2000 separated ISC analyses for each number of subjects (24 000 separated ISC analyses in total).

Each of the ISC analyses pairs was compared with Pearson's correlation and the mean absolute error (MAE). The stability of the detected *foci* was tested with the Dice index using the FDR corrected thresholds of q=0.05, q=0.01 and q=0.001. This produced 1000 tests for each number of subjects. These tests were averaged to the mean correlation, the mean MAE and the mean Dice index curves together with corresponding variance curves. In addition, ISC analysis was applied for the whole 130 subjects data set and the results of the analysis was then compared with all other 24 000 ISC analyses results. As earlier the unthresholded maps were compared with Pearson's correlation, but instead of Dice index the thresholded results were compared with the sensitivity and the specificity measures using the thresholded result of 130 subjects as a ground truth. This produced 2000 tests for each number of subjects compared with the result of 130 subjects. Like earlier, all measures were investigated with mean correlation, stability and sensitivity curves with corresponding variance curves.

These tests pointed out that already 20 subjects produces closely the same ISC analysis results as 130 subjects, but more than 30 subjects are required for truly reproducible results. These results were in line with the earlier studies with GLM analysis. In general, it was noted that the required number of subjects in the study depends on the characteristics of the experiment including the effect size.

# 7 Summary of Publications

## 7.1 Publication 1

*Kauppi J-P.*, Pajula J.*, Tohka J. "A Versatile Software Package for Inter-subject Correlation Based Analysis of fMRI", Frontiers In Neuroinformatics, 2014. *Equal contribution*

Publication 1 presents the details behind the ISC analyses implemented in the ISC Toolbox and demonstrates the use cases of the toolbox by summarizing the selected examples. Also the computation time experiments using a single desktop computer and two grid environments were reported for demonstrating the efficiency of parallel implementation. ISC Toolbox version 2.0 was published together with the publication and the ISC Toolbox 2.1 was published shortly after. The version 2.1 was published due to usability problems with parallel environments as was discussed in Section 5.3.

## 7.2 Publication 2

*Pajula J., Kauppi J-P., Tohka J. "Inter-Subject Correlation in fMRI: Method Validation against Stimulus-Model Based Analysis", PlosONE, 2012.*

Publication 2 compares the ISC based analysis with a GLM based analysis using five controlled research setups. The used data was acquired from the LONI IDA database and it included five different FRB tasks described in Chapter 3. It was assumed that GLM could locate the activations accurately with FRB tasks and thus provide a valid setup for the comparison of the ISC and GLM methods. The comparison was done with Pearson's correlation and Dice index, which confirmed that ISC can detect close to similar activation areas as GLM. ISC was also able to detect areas which were not detected by GLM.

## 7.3 Publication 3

*Pajula J., Tohka J., "Effects of Spatial Smoothing on Inter-Subject Correlation Based Analysis of fMRI", Magnetic Resonance Imaging, 2014.*

Publication 3 evaluates the effect of spatial smoothing on the ISC analysis. The study used similar procedures as in Publication 2, comparing the effects of smoothing between and within the ISC and GLM analyses. The GLM analysis was used as a reference as the effect of spatial smoothing had already been studied extensively with GLM. The study was conducted with one simulated and two experimental data sets including block design stimuli. The study confirmed that spatial smoothing had substantial effect on the ISC results. The results of ISC and GLM were highly similar with any smoothing kernel and

the effect of smoothing was slightly milder with ISC. The results indicated also that 2.5 voxel width of Gaussian smoothing kernel can produce appropriate results with the ISC analysis.

## 7.4 Publication 4

*Pajula J., Tohka J., "How Many is Enough? Effect of Sample Size in Inter-subject Correlation Analysis for fMRI Data", Computational Intelligence and Neuroscience, 2016.*

Publication 4 addressed the question: How many subjects should be included in a single analysis to make the analysis results reliable? Another question was how small a sample is enough for an ISC statistic to converge to an ISC statistic obtained with a large sample. The study was conducted with the similar block design data as in Publication 2 and 3, but using a significantly larger sample of 130 subjects. In the study, a split-half resampling based analysis was repeated by iteratively sampling non-overlapping subsets of 10 to 65 subjects for 1000 times for each number of the subjects. In this method, two non-overlapping group of subject were selected for 1000 times from the data set for each number of subjects and ISC analysis was applied for both sets. The results from the paired subsets were compared with Pearson's correlation, MAE and Dice index measures and the final measure was acquired by averaging over the 1000 iterations. The findings suggested that 20 subjects produces closely the same results as 130 subjects, but for reproducible results no less than 30 subjects should be used.

## 7.5 Author's Contribution to the Publications

The author of this thesis is the first author and the main contributor of all of the publications included in the thesis. In the Publication 1, the first authorship was shared with Jukka-Pekka Kauppi (Department of Computer Science and HIIT, University of Helsinki, Helsinki, Finland) who is the original developer of ISCtoolbox. The authorship was shared as the responsibilities of Jukka-Pekka Kauppi and the author were separated but equal to the content of the publication. The author of this thesis was responsible for developing the parallel computing environment support and optimizing the whole toolbox for parallel environment. He also conducted all functional and usability testing to compare the computing environments and conducted other minor improvements to the functionality of the ISCtoolbox.

In the Publication 2, the author of this thesis was the first author and the main contributor of the study. The original idea for the validation of ISC analysis was formed together with supervisor Jussi Tohka, but all the experiments were implemented and conducted by the author of the thesis. The ideas for Publications 3 and 4 as well as all implementation of the studies were designed by the author of this thesis. The statistical testing scheme of Publication 4 was designed together with supervisor Jussi Tohka, but all data processing and most of the manuscript preparation were conducted by the author of this thesis. All data pre-processings for all four publications were conducted by the author of this thesis, which also included the data quality inspections during the data processing.

The author wrote the articles in co-operation with co-authors and has been actively participating in the ISCtoolbox development during the whole thesis work.

# 8 Discussion

Within this thesis, the accuracy of the ISC analysis and properties of ISCtoolbox were investigated from multiple aspects. The validation of the ISC analysis was performed by comparing the accuracy of the ISC analysis with a typical group GLM analysis. The GLM analysis was also applied as a reference when testing the effect of spatial smoothing on the ISC analysis. In order to make it easier to design good-quality analyses with ISC, the effect of sample size was studied with statistical testing. All of the tests required extremely large computing resources for which reason the automated parallel computing support for two grid engines, Slurm and SGE based, were implemented into ISCtoolbox. During the ISC testing the functionality and usability of ISCtoolbox were also examined carefully. In addition to testing and development, a user guide for ISCtoolbox was produced and published in the homepage of ISCtoolbox.

## 8.1 Inter-subject Correlation Analysis

ISC analysis is a fairly new method for fMRI data analysis and it is still rarely used in comparison with GLM or ICA analysis methods. The main application area for ISC is the neurocinematics and other fMRI studies with naturalistic stimuli. The structure of naturalistic stimuli is complex, which causes trouble to separate the effects of the stimuli from the possible effects of the used analysis method. For this reason the studies of this thesis were applied with a simple block design data where it was possible to use the GLM analysis as a reference method. The investigations in this thesis are focusing on the critical questions about the accuracy and properties of the ISC analysis and the implementation of ISCtoolbox.

In Pajula et al. (2012) the ISC analysis was validated with block design data by using the GLM as a golden standard. The first reason for the choice of GLM was the *de facto* status of GLM method especially for block design studies. The second reason for selecting the GLM analysis was that multiple freely available implementations exist for the GLM analysis and they are also comprehensively tested. In studies with a reference method, the methods should behave in similar manner with the used setup or the differences between the methods should be known. With a complex data, the detection of these differences and the validity of the setup is challenging to verify and typically the setup is first tested with a simulated environment or data.

In the studies of this thesis, extremely simple simulations (described in Pajula and Tohka (2014); Pajula et al. (2012)) were applied to test that the ISC and GLM methods can detect the simulated signals with the used setup. An extremely simple simulation setup was used, as it was found that more complex simulations like described by Bellec et al. (2009) could create unwanted correlations to the simulated data and thus produce unintentional

detection with the ISC analysis. The used simple simulations were created by applying the signal with block structure and separately added $1/f$ noise in predefined locations. The setup was found sufficient test environment for both ISC and GLM methods in the applied studies.

In Pajula et al. (2012), the study confirmed that the ISC analysis is applicable also to the traditional block design data. In practice, the ISC analysis detects the similarities in the BOLD response over a group of subjects, but it cannot separate the influence of specific aspects of the stimuli. This property was confirmed in Pajula et al. (2012) when the ISC analysis was able to detect areas which were not detected by the GLM analysis. These areas were speculated to be a result of the time course synchronization between the subjects, which was not following the structure of the stimulus model used in the GLM analysis.

ISC and GLM are both univariate methods and process the data voxel by voxel. The basic assumption within univariate methods is the independence of the voxels and each voxel is analyzed without any concern about the surrounding voxels. As discussed in Section 3.3 this assumption is not true in practice. The independence assumption requires the multiple comparisons corrections to be used and controlling the FWER. Typically, the FWER is balanced between false positive and false negative voxels with no option for a "perfect" threshold without incorrectly identified voxels. Within univariate analysis it is not possible to investigate the similarity or difference within the surrounding voxels which prevents them from being used directly for discovering the neural networks and connectivity of the brain structures. At a single subject level it is possible to select a seed voxel, which time course is then compared with all other voxels. Voxels which have closely similar time courses compared with the seed voxel are then assumed to have a connected activity from the given stimulus (Joel et al., 2011).

When considering the properties of the BOLD signal, the spatial Gaussian smoothing can be proposed as a natural choice of a denoising method for ISC analysis. The area under a single voxel contains hundreds of thousands neurons which together produces indirectly the BOLD signal. From this point of view the BOLD signal itself is an average signal from the area of the voxel and Gaussian smoothing just mixes them more together. The ISC is purely based on Pearson's correlation between these signals and for this reason the ability of Gaussian smoothing to increase the SNR as well as to improve the overlap between individual subjects is beneficial. The drawback of the smoothing is that it can also spread the correlating signals outside the original source area and thus cause false positive voxels. In the analysis pipeline perspective, the applied spatial smoothing is not the only source for smoothing effect on the fMRI data. Other pre-processing steps like spatial registration and motion correction also include certain amount of spatial smoothing into the data. These effects should be taken into account especially in the situations when the ROI is small. In the ISC analysis, the required amount of smoothing depends on the registration accuracy of the subjects and the SNR of the data. To minimize the amount of false positives caused by spatial smoothing a small denoising kernel should be used. If the common registration of subjects is good, less spatial smoothing is required for ensuring the spatial overlap of the subject data and the smoothing kernel size can be selected mainly to ensure the sufficient SNR.

In Pajula and Tohka (2014) it was shown that minimal smoothing is critical for successful ISC analysis. This is natural as for a successful ISC analysis it is critical to have the common stimuli to the whole subject group, high quality alignment to the common spatial space and the sufficient SNR of the data. The common alignment of the subjects

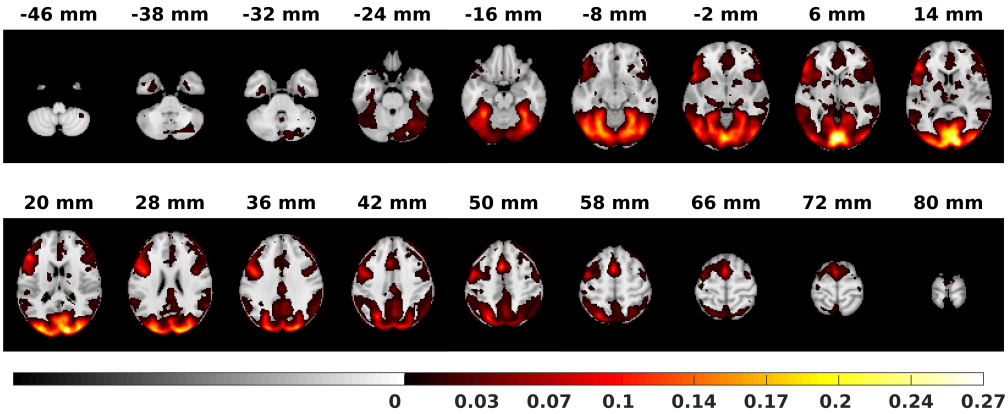
and the SNR of the data are affected by the spatial smoothing. The needed SNR is typically reached by spatial filtering with Gaussian kernel even though other methods like wavelet denoising Khullar et al. (2011a,b) or Local Principal Component Analysis (LPCA) filtering Manjón et al. (2013) could also be used. Beside the improvement with SNR, the Gaussian smoothing also ensures the alignment of the corresponding time courses in a fairly natural way when considering the source of the fMRI signal. A rule of thumb for GLM is that at least two or three voxels FWHM should be used as noted in Chapter 3. In comparison, it was shown that for ISC analysis the Gaussian kernel with FWHM of 2.5 voxles is adequate for ensuring the spatial overlap of individual subject images and SNR of the data (Pajula and Tohka, 2014).

A question about why the Gaussian smoothing has slightly lesser effect after applying statistical thresholds for ISC analysis than GLM was raised in Pajula and Tohka (2014). By definition, ISC does not have any smoothness requirement and the only need for smoothing is the improvement of SNR. Within ISCtoolbox the statistical thresholds based on a resampling test were found to adapt to the used data, which seems to control the increasing false positive voxels from the increased kernel size fairly well. The same FDR correction was applied also to GLM thresholds but with GLM it produced more liberal thresholds. In principle, the FDR corrected thresholds applied to ISC statistics were changing into more conservative when the Gaussian smoothing kernel size was increased. The change was not that fast with GLM which was concluded to be the reason why ISC analysis applied with ISCtoolbox could tolerate the slightly larger smoothing kernels than GLM within the same data and similarly corrected thresholds (Pajula and Tohka, 2014). It should be noted that FDR is not typically used with GLM analysis, more common is the GRF correction. The FDR was used here as GRF is not easily applicable to ISC results. The FDR correction is a natural choice for ISC analysis as the required smoothness of the data for GRF correction cannot be guaranteed.

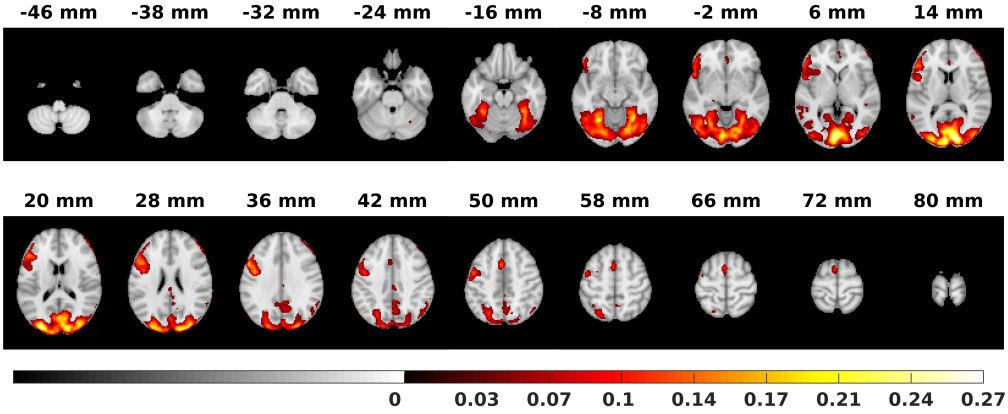
It was shown that the spatial smoothing has a significant effect on the ISC analysis results, but within the study of Pajula and Tohka (2014) a question was raised that how the sample size affects the outcome of the analysis. This topic was investigated in Pajula and Tohka (2016). According to David et al. (2013) most of the fMRI studies are applied with less than 30 subjects and some are conducted even with less than 10 subjects. The sample size in ISC is a critical aspect due to the group analysis nature of the ISC analysis. In Pajula and Tohka (2016) the minimum number of subjects for ISC analysis was proposed to be at least 20 subjects, but it was noted that no less than 30 can ensure the reproducibility of the results. Results of Pajula and Tohka (2016) indicated also that with more conservative thresholds more subjects should be included in the study. This lead to interesting detection that larger sample size increases the SNR of the ISC analysis but it also creates larger *foci* with the same significance level than the same study with a smaller sample. An interesting detection from these results is that the largest p-value in both results is more or less the same. This is visible in Figure 8.1. One possible reason for the larger *foci* is that with a larger number of subjects the border areas of the *foci* have more overlapping subjects which may increase the average ISC in those border area voxels.

## 8.2 ISCtoolbox Implementation

Significant effort has been given to make the ISCtoolbox GUI easy and understandable to use, as well as to prevent the user from selecting conflicting parameters by accident. From




**(a)** ISC analysis results of 130 subjects.




**(b)** ISC analysis results of 37 subjects.

**Figure 8.1:** Comparison of ISC analysis with 130 subjects (image series on the panel (a)) and ISC analysis with 37 subjects (image series on the panel (b)). Both analyses were applied to a similar ICBM data with the VG FRB task and the used threshold was corrected with FDR using q=0.001. The detected *foci* after applying thresholds were significantly larger with 130 subjects analysis than with 37 subjects analysis. Interestingly, the largest p values are still the same for both studies.

the point of view of the users, the graphical GUI with simple controls makes the usage of parallel environment tempting and enables the user to use it even without any true knowledge of the controls of a grid engine in the parallel environment. Even a moderate parallel environment can speed up the overall computing significantly which is presented in the Figure 8.2 (Kauppi et al., 2014). The figure shows a reference how even a modest cluster (Outolintu) can reduce the computing time significantly compared with local computations. In the figure, the desktop computer was equipped with Intel Core2Duo E8400 CPU at 3.00GHz and 5GB RAM, the Outolintu cluster was used with 10 parallel processes running on nodes equipped with Intel Xeon X5450 CPUs at 3.00GHz and 8GB RAM and the Merope cluster was used on average with 32 processes on nodes equipped with Intel Xeon X5650 at 2.67GHz and 8 GB RAM.

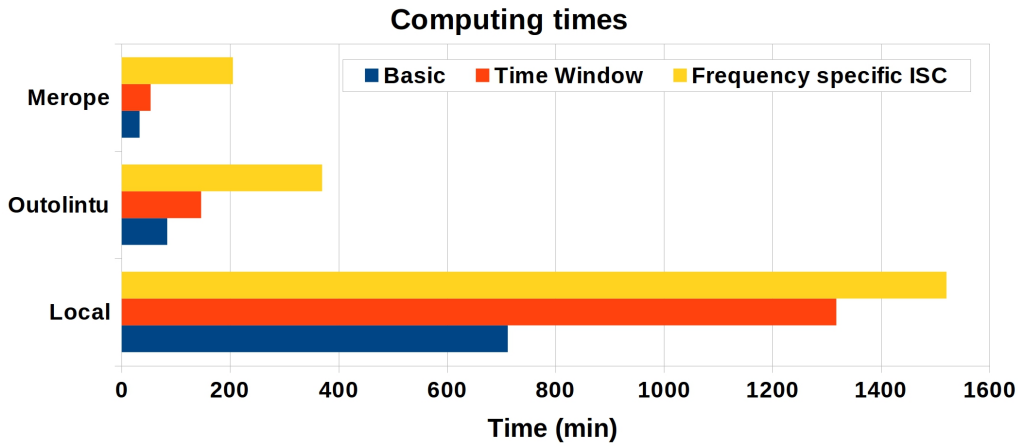



**Figure 8.2:** Reference computing times of ISCtoolbox with three different analyses; basic ISC analysis, time window ISC analysis and frequency specific ISC analysis. Computations were conducted in three different hardware environments; a large Merope cluster, a small Outolintu cluster and a local desktop computer. Reproduced from Kauppi et al. (2014)

Currently, the probably largest number of basic ISC analyses which has been applied in a single study have been in Pajula and Tohka (2016). The testing of the sample size required 24 000 separated full ISC analysis which would had taken almost 4.75 years to compute with a single core in high-end CPU. Computing the analyses in parallel with a cluster environment the computing time was reduced to 2 months. This is a good example of how important the efficient computing environments are for computational analysis like ISC. As presented in Figure 8.2 the parallel cluster environments can speed up the ISC analysis significantly and in the future the parallel support will become even more important when the size and length of the data increases. Increased size and accuracy of the data create increased demand for hardware requirements which are hardly met with common desktop hardware. It is reasonable to assume that the grid engine support of ISCtoolbox will become even more attractive to the researchers in the future. As a future development option, the support for local parallel computations should be considered: The number of cores increases also on the desktop computers which would enable the parallel computations locally.

During the massive computations included in the thesis, it was detected that the paral-

lelization techniques used in ISCtoolbox can reduce the RAM requirement significantly especially on the phase of resampling distribution computation. Even though the parallel patches are applied typically in the cluster environment, they make possible also somewhat large computations run on modern high-end desktop computers locally. For example, 130 subjects ISC analysis in Pajula and Tohka (2016) was at first crashing with 128 GB of RAM because the memory was running out. At the end, the analysis was possible to run with 16 GB of RAM when the resampling testing for thresholds was distributed in 100 patches. In practice, the main difference between cluster environment and a local computer is the way the computations are applied. With the cluster, the individual resampling patches are applied in parallel processes and in the local environment they are computed in series. The memory requirements decrease because the computation of extremely large distribution (for example, 100 000 000 realizations) is computed in multiple smaller pieces and only a single piece of distribution must be handled in memory at once. From this point of view, the parallel mechanisms of ISCtoolbox enable even larger ISC analyses to be run with reasonable hardware.

Computing clusters are not the only hardware that can be used to enable parallelization for the computations. During the past decade, the graphics processing units (GPU) hardware has become a new option for the scientific computing. The general-purpose computing on graphics processing units (GPGPU) methods have enabled analyses which were earlier almost impossible to apply due to extreme computing times (Eklund, 2012; Eklund et al., 2012). Although significant success stories have taken place, the GPU hardware has still limitations in available memory and has a restricted instruction set which limits their possible usage with complex analyses. The main challenge for ISC analysis in GPU environment is the limited memory of the devices which causes overhead on the transport of the data between the RAM memory controlled by the central processing unit (CPU) and the onboard GPU memory. The GLM type of analysis is easier to apply with GPU as there is only a single model to be compared with the data, which can be stored to GPU memory. In ISC analysis each subjects acts as a model for each other subjects which make it impossible to hold all the data at once in the GPU memory. When the amount of memory increases and a larger instruction set becomes available in the GPU hardware, it might be possible to utilize the power of GPGPU hardware within the ISC analysis.

In addition to GPU hardware, another interesting option is the Intel Phi (Rahman, 2013) hardware environment. These devices share the idea of GPGPU devices by having a huge number of parallel computing units, but they include the full instruction set of Intel x86 processor specification. This means that they do not have limitations in the available arithmetical operations. Both hardware GPUs and Phi are implemented as extension cards which enables them to be included in a desktop computer or in a node of a computing cluster. The ISC analysis could be applied with the GPU or Phi hardware in the future, but in practice it would require large optimization for the memory usage of the toolbox.

Traditionally, the fMRI analyses are conducted via multiple software packages and the researcher is taking care that the data is passed from a single software to another and possible data format transformations are done between the different software. An emerging trend in neuroinformatics is the use of pipeline software and systems like neuroinformatics Python pipeline (NiPyPe) (Gorgolewski et al., 2011) or LONI Pipeline (Rex et al., 2003). These systems should be taken into account in the future development of ISCtoolbox and possible support for pipeline software should be considered. Using pipeline software could also increase the reproducibility of the study if the pipeline setup would be shared with the

publications. Besides the development of different pipeline software, a significant project called Neuroimaging Data Model (NIDM) is defining standards and formal ways on how the brain imaging data analyses should be reported and shared[1] (Keator et al., 2014). The aim of the NIDM is to unify the terminology and reporting of the neuroimaging studies by defining a World-Wide Web Consortium (W3C) accepted standard for human brain mapping.

## 8.3 Open Data and Data Sharing

All of the fMRI data used in this thesis were downloaded from the LONI IDA database[2] (Mazziotta et al., 2001). Without this kind open data for scientific use, the whole thesis would have been impossible to apply in the current way. The open data is an emerging trend in the computational neuroimaging and already multiple free fMRI data sources exist for scientific usage and even more services are under development. Multiple challenges still exist for these services, for example: linking of separated data sets; physical storage space for the data collections; and missing or varying metadata. These problems are closely connected to the loosely standardized data formats within the brain imaging. An example of a project which is dedicated to help the situation of data sharing in fMRI is the OpenfMRI[3] project (Poldrack et al., 2013), which provides a place for the task-based fMRI data sets for free distribution among the scientific community. The OpenfMRI service is designed to host fMRI data sets including the raw data.

Whereas OpenfMRI provides resources to save the full data sets, the NeuroVault[4] service (Gorgolewski et al., 2015) is a web based visualization service for statistical maps created in analyses of neuroimaging studies. The aim of NeuroVault is that the whole statistical maps could be published with the journal publications and the readers of the publications could investigate the statistical maps by themselves while reading the publication. Another aspect of NeuroVault is that it has direct linking to the NeuroSynth[5] service (Yarkoni et al., 2011), which provides an easy and fast meta-analysis tool for neuroimaging studies. NeuroSynth uses the peak coordinate information published in various journal articles to link the articles which have results of a similar kind. The developers of the NeuroVault service hope that in the future researchers will provide the statistics also from the older publications to the service. This would enable scientists to make extremely large meta analyses based on statistical maps over the different imaging modalities and study topics, which would be impossible to apply otherwise. The statistical maps of Pajula and Tohka (2016) were shared in NeuroVault: `http://www.neurovault.org/collections/WTMVBEZP/`

ISCtoolbox is a free open source software and earlier it has been developed and published via Google code service. The Google code was unfortunately closed in August 2015 and for this reason ISCtoolbox was moved to NITRC[6] service. The NITRC aims to link relevant tools for neuroscientific research as well as provide systems to publish open data for the scientific community. The work of NITRC is needed within the neuroinformatics as typically software from smaller research groups is used only by the group members themselves and nobody else cannot find the software. When the developed software is listed publicly, the global neuroscientific community can save significant amount of time

[1] http://nidm.nidash.org/
[2] https://ida.loni.usc.edu/
[3] http://openfmri.org/
[4] http://neurovault.org/
[5] http://neurosynth.org/
[6] http://www.nitrc.org/

and resources. From the software development point of view, improving the common knowledge can also improve the quality of analyses as a larger number of scientists can give support for the software and not everyone are needed to develop their own implementation of basic analyses. For a single software point of view, it is also important that someone is using the software, as without users there might not be a reason or motivation to develop the software further.

# 9 Conclusion

This thesis investigated the properties of ISC analysis for fMRI. Three major aspects of the ISC analysis were studied: First, the ISC analysis results were validated by comparing them with the results of corresponding GLM analysis. Second, the effects of spatial smoothing on ISC analysis results were investigated using the GLM analysis as a reference. Third, the effect of sample size for the ISC analysis results was tested with comprehensive statistical testing. In addition and due to these studies, the openly available implementation of ISC analysis, ISCtoolbox for Matlab, was improved and a new built-in support for cluster computing environments was included in the toolbox. Altogether, four international journal publications were included in this thesis.

The ISC analysis was verified to be an efficient non-parametric data-analysis method for fMRI data. The ISC analysis was found to be suitable also for block design fMRI studies as it was able to find areas similar to the corresponding GLM analysis. The spatial smoothing was found to be a mandatory pre-processing step for ISC analysis and applied with FDR corrected thresholds the ISC analysis could tolerate mildly larger Gaussian smoothing kernels than GLM analysis. The sample size investigation pointed out that minimum of 20 subjects should be used in the ISC analysis to reach fairly stable results, but only including 30 or more subjects guarantees the reproducibility of the study. Cluster computing environments were found to provide significant improvement for the computing times of ISCtoolbox and the new built-in support for them was a significant improvement for the general usability of the toolbox.

Studies of this thesis support the use of ISC analysis in fMRI studies. The validation and found properties of the ISC analysis improved the credibility of the ISC analysis and the improvements of ISCtoolbox increased the usability of the implementation.

# Bibliography


Abrams, D. A., Ryali, S., Chen, T., Chordia, P., Khouzam, A., Levitin, D. J., and Menon, V., "Inter-subject synchronization of brain responses during natural music listening," *European Journal of Neuroscience*, 2013. [Online]. Available: http://dx.doi.org/10.1111/ejn.12173

Ashby, F. G., *Statistical Analysis Of FMRI Data*. Cambridge, Mass.: MIT Press, 2011. [Online]. Available: http://isbnplus.org/9780262015042

Bartels, A., Zeki, S., and Logothetis, N. K., "Natural vision reveals regional specialization to local motion and to contrast-invariant, global flow in the human brain," *Cerebral Cortex*, vol. 18, no. 3, pp. 705–717, 2008. [Online]. Available: http://dx.doi.org/10.1093/cercor/bhm107

Beckmann, C. F., Jenkinson, M., and Smith, S. M., "General multilevel linear modeling for group analysis in fmri," *NeuroImage*, vol. 20, no. 2, pp. 1052–1063, 2003. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S105381190300435X

Bellec, P., Carbonell, F. M., Perlbarg, V., Lepage, C., Lyttelton, O., Fonov, V., Janke, A., Tohka, J., and Evans, A. C., "A neuroimaging analysis kit for matlab and octave," in *Proceedings of the 17th International Conference on Functional Mapping of the Human Brain*, 2011.

Bellec, P., Perlbarg, V., and Evans, A. C., "Bootstrap generation and evaluation of an fmri simulation database," *Magnetic resonance imaging*, vol. 27, no. 10, pp. 1382–1396, 2009. [Online]. Available: http://dx.doi.org/10.1016/j.mri.2009.05.034

Benjamini, Y. and Hochberg, Y., "Controlling the false discovery rate: A practical and powerful approach to multiple testing," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 57, no. 1, pp. 289–300, 1995. [Online]. Available: http://www.jstor.org/stable/2346101

Bennett, C. M., Wolford, G. L., and Miller, M. B., "The principled control of false positives in neuroimaging," *Social Cognitive and Affective Neuroscience*, vol. 4, no. 4, pp. 417–422, 2009. [Online]. Available: http://scan.oxfordjournals.org/content/4/4/417.abstract

Bennett, C. M., Baird, A. A., Miller, M. B., and Wolford, G. L., "Neural Correlates of Interspecies Perspective Taking in the Post-Mortem Atlantic Salmon: An Argument For Proper Multiple Comparisons Correction," *Journal of Serendipitous and Unexpected Results*, vol. 1, pp. 1–5, 2010. [Online]. Available: http://www.jsur.org/v1n1p1

Biswal, B., Zerrin Yetkin, F., Haughton, V. M., and Hyde, J. S., "Functional connectivity in the motor cortex of resting human brain using echo-planar mri,"

*Magnetic Resonance in Medicine*, vol. 34, no. 4, pp. 537–541, 1995. [Online]. Available: http://dx.doi.org/10.1002/mrm.1910340409

Botzung, A., LaBar, K. S., Kragel, P., Miles, A., and Rubin, D. C., "Component neural systems for the creation of emotional memories during free viewing of a complex, real-world event," *Frontiers in Human Neuroscience*, vol. 4, pp. 1–10, 2010. [Online]. Available: http://dx.doi.org/10.3389/fnhum.2010.00034

Calhoun, V. D. and Pearlson, G. D., "A selective review of simulated driving studies: Combining naturalistic and hybrid paradigms, analysis approaches, and future directions," *NeuroImage*, vol. 59, no. 1, pp. 25–35, 2012. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1053811911006823

Chumbley, J. R. and Friston, K. J., "False discovery rate revisited: {FDR} and topological inference using gaussian random fields," *NeuroImage*, vol. 44, no. 1, pp. 62 – 70, 2009. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1053811908006472

Churchill, N., Yourganov, G., and Strother, S., "Comparing within-subject classification and regularization methods in fmri for large and small sample sizes," *Human Brain Mapping*, vol. 35, no. 9, pp. 4499–4517, 2014. [Online]. Available: http://www.scopus.com/inward/record.url?eid=2-s2.0-84904515523&partnerID=40&md5=a4d125cb3fd57f02567dc471ac123b1f

Cohen, J., "Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit," *Psychological Bulletin*, vol. 70, no. 4, pp. 213–220, 1968. [Online]. Available: http://dx.doi.org/10.1037/h0026256

Cox, R. W., "Afni: software for analysis and visualization of functional magnetic resonance neuroimages." *Computers and Biomedical Research*, vol. 29, no. 3, pp. 162–173, 1996. [Online]. Available: http://dx.doi.org/10.1006/cbmr.1996.0014

Cox, R. W., Ashburner, J., Breman, H., Fissell, K., Haselgrove, C., Holmes, C. J., Lancaster, J. L., Rex, D. E., Smith, S. M., Woodward, J. B., and Strother, S. C., "A (sort of) new image data format standard: NIfTI-1," in *Tenth Annual Meeting of the Organization for Human Brain Mapping*, 2004.

Cox, R. W. and Hyde, J. S., "Software tools for analysis and visualization of fmri data," *NMR in Biomedicine*, vol. 10, no. 4-5, pp. 171–178, 1997. [Online]. Available: http://dx.doi.org/10.1002/(SICI)1099-1492(199706/08)10:4/5<171::AID-NBM453>3.0.CO;2-L

Dale, A. M., Fischl, B., and Sereno, M. I., "Cortical surface-based analysis: I. segmentation and surface reconstruction," *NeuroImage*, vol. 9, no. 2, pp. 179 – 194, 1999. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1053811998903950

David, S. P., Ware, J. J., Chu, I. M., Loftus, P. D., Fusar-Poli, P., Radua, J., Munafò, Marcus, R., and Ioannidis, J. P. A., "Potential reporting bias in fmri studies of the brain," *PLoS ONE*, vol. 8, no. 7, p. e70104, 2013. [Online]. Available: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3723634/

D'Esposito, M., Zarahn, E., and Aguirre, G. K., "Event-related functional mri: implications for cognitive psychology," *Psychological bulletin*, vol. 125, no. 1, pp. 155–164, 1999. [Online]. Available: http://psycnet.apa.org/doi/10.1037/0033-2909.125.1.155

Dice, L. R., "Measures of the amount of ecologic association between species," *Ecology*, vol. 26, no. 3, pp. 297–302, 1945. [Online]. Available: http://www.jstor.org/stable/1932409

Draganski, B., Gaser, C., Busch, V., Schuierer, G., Bogdahn, U., and May, A., "Neuroplasticity: Changes in grey matter induced by training," *Nature*, vol. 427, no. 6972, pp. 311–312, 2004. [Online]. Available: http://dx.doi.org/10.1038/427311a

Eklund, A., "fmri analysis on the gpu—possibilities and challenges," *Computer methods and programs in biomedicine*, vol. 105, no. 2, pp. 145–161, 2012. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0169260711001957

Eklund, A., Andersson, M., Josephson, C., Johannesson, M., and Knutsson, H., "Does parametric fmri analysis with spm yield valid results? — an empirical study of 1484 rest datasets," *NeuroImage*, vol. 61, no. 3, pp. 565–578, 2012. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1053811912003825

Finch, H., "Comparison of distance measures in cluster analysis with dichotomous data," *Journal of Data Science*, vol. 3, no. 1, pp. 85–100, 2005. [Online]. Available: http://www.jds-online.com/file_download/66/JDS-192.pdf

Fjell, A. M. and Walhovd, K. B., "Structural brain changes in aging: courses, causes and cognitive consequences," *Reviews in the neurosciences*, vol. 21, no. 3, pp. 187–221, 2010. [Online]. Available: http://dx.doi.org/10.1515/REVNEURO.2010.21.3.187

Foerde, K., Knowlton, B. J., and Poldrack, R. A., "Modulation of competing memory systems by distraction," *Proceedings of the National Academy of Sciences*, vol. 103, no. 31, pp. 11 778–11 783, 2006. [Online]. Available: http://www.pnas.org/content/103/31/11778.abstract

Friston, K. J., Jezzard, P., and Turner, R., "Analysis of functional mri time-series," *Human brain mapping*, vol. 1, no. 2, pp. 153–171, 1994. [Online]. Available: http://dx.doi.org/10.1002/hbm.460010207

Friston, G. Flandin, K. J., "Statistical parametric mapping (spm)," *Scholarpedia*, vol. 3, no. 3, p. 6232, 2008. [Online]. Available: http://dx.doi.org/10.4249/scholarpedia.6232

Friston, K., Holmes, A., Poline, J.-B., Price, C., and Frith, C., "Detecting activations in pet and fmri: Levels of inference and power," *NeuroImage*, vol. 4, no. 3, pp. 223 – 235, 1996. [Online]. Available: http://dx.doi.org/10.1006/nimg.1996.0074

Genovese, C. R., Lazar, N. A., and Nichols, T., "Thresholding of statistical maps in functional neuroimaging using the false discovery rate," *NeuroImage*, vol. 15, no. 4, pp. 870–878, 2002. [Online]. Available: http://dx.doi.org/10.1006/nimg.2001.1037

Glerean, E., Salmi, J., Lahnakoski, J. M., Jääskeläinen, I. P., and Sams, M., "Functional magnetic resonance imaging phase synchronization as a measure of dynamic functional connectivity," *Brain connectivity*, vol. 2, no. 2, pp. 91–101, 2012. [Online]. Available: http://dx.doi.org/10.1089/brain.2011.0068

Golland, Y., Bentin, S., Gelbard, H., Benjamini, Y., Heller, R., Nir, Y., Hasson, U., and Malach, R., "Extrinsic and intrinsic systems in the posterior cortex of the human brain revealed during natural sensory stimulation," *Cerebral Cortex*, vol. 17, no. 4, pp. 766–777, 2007. [Online]. Available: http://cercor.oxfordjournals.org/content/17/4/766.abstract

Gorgolewski, K., Burns, C. D., Madison, C., Clark, D., Halchenko, Y. O., Waskom, M. L., and Ghosh, S. S., "Nipype: A flexible, lightweight and extensible neuroimaging data processing framework," *Frontiers in Neuroinformatics*, vol. 5, no. 13, 2011. [Online]. Available: http://www.frontiersin.org/neuroinformatics/10.3389/fninf.2011.00013/abstract

Gorgolewski, K. J., Varoquaux, G., Rivera, G., Schwartz, Y., Sochat, V. V., Ghosh, S. S., Maumet, C., Nichols, T. E., Poline, J.-B., Yarkoni, T., Margulies, D. S., and Poldrack, R. A., "Neurovault.org: A repository for sharing unthresholded statistical maps, parcellations, and atlases of the human brain," *NeuroImage*, pp. –, 2015. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1053811915003067

Goswami, J. C. and Hoefel, A. E., "Algorithms for estimating instantaneous frequency," *Signal processing*, vol. 84, no. 8, pp. 1423–1427, 2004. [Online]. Available: http://dx.doi.org/10.1016/j.sigpro.2004.05.016

Hanke, M., Baumgartner, F. J., Ibe, P., Kaule, F. R., Pollmann, S., Speck, O., Zinke, W., and Stadler, J., "A high-resolution 7-tesla fmri dataset from complex natural stimulation with an audio movie," *Scientific Data*, vol. 1, 2014. [Online]. Available: http://dx.doi.org/10.1038/sdata.2014.3

Hasson, U., Nir, Y., Levy, I., Fuhrmann, G., and Malach, R., "Intersubject synchronization of cortical activity during natural vision," *Science*, vol. 303, no. 5664, pp. 1634–1640, 2004. [Online]. Available: http://www.sciencemag.org/content/303/5664/1634.abstract

Hasson, U., Furman, O., Clark, D., Dudai, Y., and Davachi, L., "Enhanced intersubject correlations during movie viewing correlate with successful episodic encoding," *Neuron*, vol. 57, no. 3, pp. 452–462, 2008. [Online]. Available: http://dx.doi.org/10.1016/j.neuron.2007.12.009

Hayasaka, S. and Nichols, T. E., "Validating cluster size inference: random field and permutation methods," *NeuroImage*, vol. 20, no. 4, pp. 2343 – 2356, 2003. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1053811903005020

Hejnar, M. P., Kiehl, K. A., and Calhoun, V. D., "Interparticipant correlations: A model free fmri analysis technique," *Human brain mapping*, vol. 28, no. 9, pp. 860–867, 2007. [Online]. Available: http://dx.doi.org/10.1002/hbm.20321

Hornak, J. P., *The basics of MRI.* Interactive Learning Software, Henietta, NY, 1996-2014. [Online]. Available: http://www.cis.rit.edu/htbooks/mri/index.html

Hupé, J.-M., "Statistical inferences under the null hypothesis: Common mistakes and pitfalls in neuroimaging studies." *Frontiers in Neuroscience*, vol. 9, no. 18, 2015. [Online]. Available: http://www.frontiersin.org/brain_imaging_methods/10.3389/fnins.2015.00018/abstract

Hyvärinen, A. and Oja, E., "Independent component analysis: algorithms and applications," *Neural Networks*, vol. 13, no. 4–5, pp. 411 – 430, 2000. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0893608000000265

Jenkinson, M., Beckmann, C. F., Behrens, T. E., Woolrich, M. W., and Smith, S. M., "{FSL}," *NeuroImage*, vol. 62, no. 2, pp. 782 – 790, 2012. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1053811911010603

Joel, S. E., Caffo, B. S., van Zijl, P., and Pekar, J. J., "On the relationship between seed-based and ica-based measures of functional connectivity," *Magnetic Resonance in Medicine*, vol. 66, no. 3, pp. 644–657, 2011. [Online]. Available: http://dx.doi.org/10.1002/mrm.22818

Jääskeläinen, I. P., Koskentalo, K., Balk, M. H., Autti, T., Kauramäki, J., Pomren, C., and Sams, M., "Inter-subject synchronization of prefrontal cortex hemodynamic activity during natural viewing." *The Open Neuroimaging Journal*, vol. 2, pp. 14–19, 2008. [Online]. Available: http://www.biomedsearch.com/nih/Inter-subject-synchronization-prefrontal-cortex/19018313.html

Kauppi, J.-P., Jääskeläinen, I. P., Sams, M., and Tohka, J., "Inter-subject correlation of brain hemodynamic responses during watching a movie: localization in space and frequency," *Frontiers in Neuroinformatics*, vol. 4, no. 0, p. 5, 2010. [Online]. Available: http://www.frontiersin.org/Journal/Abstract.aspx?s=752&name=neuroinformatics&ART_DOI=10.3389/fninf.2010.00005

Kauppi, J.-P., Pajula, J., and Tohka, J., "A versatile software package for inter-subject correlation based analyses of fmri," *Frontiers in Neuroinformatics*, vol. 8, no. 2, 2014. [Online]. Available: http://www.frontiersin.org/neuroinformatics/10.3389/fninf.2014.00002/abstract

Keator, D. B., Ghosh, S. S., Maumet, C., Flandin, G., Nichols, B. N., Nichols, T. E., Burns, G. A. P. C., Bruehl, R., Craddock, C., Federick, B., Gorgolewski, K., Halchenko, Y. O., Hanke, M., Haselgrove, C., Helmer, K., Klein, A., Marcus, D., Milham, M., Michel, F., Poldrack, R., Steffener, J., Schwartz, Y., Stoner, R. M., Turner, J. A., Kennedy, D. N., and Poline, J.-B., "Developing and using the data models for neuroimaging: the nidash working group," 2014. [Online]. Available: http://www.frontiersin.org/neuroinformatics/10.3389/conf.fninf.2014.18.00030/full

Khullar, S., Michael, A., Correa, N., Adali, T., Baum, S., and Calhoun, V., "Wavelet-based denoising and independent component analysis for improving multi-group inference in fmri data," in *Biomedical Imaging: From Nano to Macro, 2011 IEEE International Symposium on*, 2011, pp. 456–459. [Online]. Available: http://dx.doi.org/10.1109/ISBI.2011.5872444

Khullar, S., Michael, A., Correa, N., Adali, T., Baum, S. A., and Calhoun, V. D., "Wavelet-based fmri analysis: 3-d denoising, signal separation, and validation metrics," *NeuroImage*, vol. 54, no. 4, pp. 2867 – 2884, 2011. [Online]. Available: http://dx.doi.org/10.1016/j.neuroimage.2010.10.063

Kätsyri, J., Hari, R., Ravaja, N., and Nummenmaa, L., "Just watching the game ain't enough: Striatal fmri reward responses to successes and failures in a video game during active and vicarious playing," *Frontiers in Human Neuroscience*, vol. 7, no. 278, 2013. [Online]. Available: http://www.frontiersin.org/human_neuroscience/10.3389/fnhum.2013.00278/abstract

Landis, J. R. and Koch, G. G., "The measurement of observer agreement for categorical data." *Biometrics*, vol. 33, no. 1, pp. 159–174, 1977. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/843571

Lerner, Y., Honey, C. J., Silbert, L. J., and Hasson, U., "Topographic mapping of a hierarchy of temporal receptive windows using a narrated story," *The*

*Journal of Neuroscience*, vol. 31, no. 8, pp. 2906–2915, 2011. [Online]. Available: http://dx.doi.org/10.1523/JNEUROSCI.3684-10.2011

Lindquist, M. A. and Wager, T. D., "Spatial smoothing in fmri using prolate spheroidal wave functions," *Human brain mapping*, vol. 29, no. 11, pp. 1276–1287, 2008. [Online]. Available: http://dx.doi.org/10.1002/hbm.20475

Lindquist, M. A., Loh, J. M., Atlas, L. Y., and Wager, T. D., "Modeling the hemodynamic response function in fmri: Efficiency, bias and mis-modeling," *NeuroImage*, vol. 45, no. 1, Supplement 1, pp. 187–198, 2009. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1053811908012056

Logothetis, N. K., "The underpinnings of the bold functional magnetic resonance imaging signal," *The Journal of Neuroscience*, vol. 23, no. 10, pp. 3963–3971, 2003. [Online]. Available: http://www.jneurosci.org/content/23/10/3963.short

Love, D., "Son of grid engine project," 2013, (Visited 5.11.2013). [Online]. Available: https://arc.liv.ac.uk/trac/SGE

Maguire, E. A., Woollett, K., and Spiers, H. J., "London taxi drivers and bus drivers: A structural mri and neuropsychological analysis," *Hippocampus*, vol. 16, no. 12, pp. 1091–1101, 2006. [Online]. Available: http://dx.doi.org/10.1002/hipo.20233

Malinen, S., Hlushchuk, Y., and Hari, R., "Towards natural stimulation in fmri-issues of data analysis," *NeuroImage*, vol. 35, no. 1, pp. 131–139, 2007. [Online]. Available: http://dx.doi.org/10.1016/j.neuroimage.2006.11.015

Manjón, J. V., Coupé, P., Concha, L., Buades, A., Collins, D. L., and Robles, M., "Diffusion weighted image denoising using overcomplete local pca," *PLoS ONE*, vol. 8, no. 9, p. e73021, 2013. [Online]. Available: http://dx.doi.org/10.1371%2Fjournal.pone.0073021

Mazziotta, J., Toga, A., Evans, A., Fox, P., Lancaster, J., Zilles, K., Woods, R., Paus, T., Simpson, G., Pike, B., Holmes, C., Collins, L., Thompson, P., MacDonald, D., Iacoboni, M., Schormann, T., Amunts, K., Palomero-Gallagher, N., Geyer, S., Parsons, L., Narr, K., Kabani, N., Goualher, G. L., Boomsma, D., Cannon, T., Kawashima, R., and Mazoyer, B., "A probabilistic atlas and reference system for the human brain: International consortium for brain mapping (icbm)," *Philosophical Transactions of the Royal Society of London.Series B: Biological Sciences*, vol. 356, no. 1412, pp. 1293–1322, 2001. [Online]. Available: http://dx.doi.org/10.1098%2Frstb.2001.0915

Mckeown, M. J., Makeig, S., Brown, G. G., Jung, T.-P., Kindermann, S. S., Bell, A. J., and Sejnowski, T. J., "Analysis of fmri data by blind separation into independent spatial components," *Human Brain Mapping*, vol. 6, no. 3, pp. 160–188, 1998. [Online]. Available: http://dx.doi.org/10.1002/(SICI)1097-0193(1998)6:3<160::AID-HBM5>3.0.CO;2-1

Mukamel, R., Gelbard, H., Arieli, A., Hasson, U., Fried, I., and Malach, R., "Coupling between neuronal firing, field potentials, and fmri in human auditory cortex," *Science*, vol. 309, no. 5736, pp. 951–954, 2005. [Online]. Available: http://dx.doi.org/10.1126/science.1110913

Murphy, K. and Garavan, H., "An empirical investigation into the number of subjects required for an event-related fmri study," *NeuroImage*, vol. 22, no. 2, pp. 879–885, 2004. [Online]. Available: http://www.scopus.com/inward/record.url?eid=2-s2.0-2942524074&partnerID=40&md5=27e10c7976bda40ff7f74879334091c8

Nelder, J. A. and Wedderburn, R. W. M., "Generalized linear models," *Journal of the Royal Statistical Society.Series A (General)*, vol. 135, no. 3, pp. pp. 370–384, 1972. [Online]. Available: http://www.jstor.org/stable/2344614

Nichols, T. and Hayasaka, S., "Controlling the familywise error rate in functional neuroimaging: a comparative review," *Statistical Methods in Medical Research*, vol. 12, no. 5, pp. 419–446, 2003. [Online]. Available: http://dx.doi.org/10.1191/0962280203sm341ra

Nichols, T. E. and Holmes, A. P., "Nonparametric permutation tests for functional neuroimaging: a primer with examples," *Human brain mapping*, vol. 15, no. 1, pp. 1–25, 2002. [Online]. Available: http://dx.doi.org/10.1002/hbm.1058

Nir, Y., Fisch, L., Mukamel, R., Gelbard-Sagiv, H., Arieli, A., Fried, I., and Malach, R., "Coupling between neuronal firing rate, gamma lfp, and bold fmri is related to interneuronal correlations," *Current Biology*, vol. 17, no. 15, pp. 1275–1285, 2007. [Online]. Available: http://dx.doi.org/10.1016/j.cub.2007.06.066

Ogawa, S., Lee, T. M., Kay, A. R., and Tank, D. W., "Brain magnetic resonance imaging with contrast dependent on blood oxygenation," *Proceedings of the National Academy of Sciences*, vol. 87, no. 24, pp. 9868–9872, 1990. [Online]. Available: http://www.pnas.org/content/87/24/9868.abstract

Ogawa, S., Menon, R. S., Kim, S.-G., and Ugurbil, K., "On the characteristics of functional magnetic resonance imaging of the brain," *Annual Review of Biophysics and Biomolecular Structure*, vol. 27, no. 1, pp. 447–474, 1998. [Online]. Available: http://dx.doi.org/10.1146/annurev.biophys.27.1.447

Ogawa, S., Lee, T.-M., Nayak, A. S., and Glynn, P., "Oxygenation-sensitive contrast in magnetic resonance image of rodent brain at high magnetic fields," *Magnetic Resonance in Medicine*, vol. 14, no. 1, pp. 68–78, 1990. [Online]. Available: http://dx.doi.org/10.1002/mrm.1910140108

Pajula, J. and Tohka, J., "Effects of spatial smoothing on inter-subject correlation based analysis of fmri," *Magnetic resonance imaging*, vol. 32, no. 9, pp. 1114–1124, 2014. [Online]. Available: http://dx.doi.org/10.1016/j.mri.2014.06.001

Pajula, J. and Tohka, J., "How many is enough? effect of sample size in inter-subject correlation analysis for fmri data," *Computational Intelligence and Neuroscience*, vol. 2016, no. 2094601, 2016. [Online]. Available: http://dx.doi.org/10.1155/2016/2094601

Pajula, J., Kauppi, J.-P., and Tohka, J., "Inter-subject correlation in fmri: Method validation against stimulus-model based analysis," *PLoS ONE*, vol. 7, no. 8, p. e41196, 2012. [Online]. Available: http://dx.doi.org/10.1371%2Fjournal.pone.0041196

Pearson, K., "Mathematical contributions to the theory of evolution. iii. regression, heredity, and panmixia," *Philosophical Transactions of the Royal Society of London.Series A, Containing Papers of a Mathematical or Physical Character*, vol. 187, pp. 253–318, 1896. [Online]. Available: http://www.dx.doi.org/10.1098/rsta.1896.0007

Poldrack, R., Mumford, J., and Nichols, T., *Handbook of Functional MRI Data Analysis*. Cambridge University Press, 2011.

Poldrack, R. A., Barch, D. M., Mitchell, J., Wager, T., Wagner, A. D., Devlin, J. T., Cumba, C., Koyejo, O., and Milham, M., "Towards open sharing of task-based fmri data: The openfmri project," *Frontiers in Neuroinformatics*, vol. 7, no. 12, 2013. [Online]. Available: http://www.frontiersin.org/neuroinformatics/10.3389/fninf.2013.00012/abstract

Raghunathan, T., Rosenthal, R., and Rubin, D. B., "Comparing correlated but nonoverlapping correlations." *Psychological Methods*, vol. 1, no. 2, p. 178, 1996. [Online]. Available: http://dx.doi.org/10.1037/1082-989X.1.2.178

Rahman, R., *Intel Xeon Phi Coprocessor Architecture and Tools: The Guide for Application Developers*, 1st ed. Berkely, CA, USA: Apress, 2013.

Rex, D. E., Ma, J. Q., and Toga, A. W., "The {LONI} pipeline processing environment," *NeuroImage*, vol. 19, no. 3, pp. 1033 – 1048, 2003. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S105381190300185X

Rosen, H. J., Ojemann, J. G., Ollinger, J. M., and Petersen, S. E., "Comparison of brain activation during word retrieval done silently and aloud using fmri," *Brain and Cognition*, vol. 42, no. 2, pp. 201 – 217, 2000. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0278262699911004

Sanchez, C. E., Richards, J. E., and Almli, C. R., "Age-specific mri templates for pediatric neuroimaging," *Developmental neuropsychology*, vol. 37, no. 5, pp. 379–399, 2012. [Online]. Available: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3399736/

Scalable Logic, "Open grid scheduler/grid engine," 2013, http://gridscheduler.sourceforge.net/ (Visited 5.11.2013).

Sladky, R., Friston, K. J., Tröstl, J., Cunnington, R., Moser, E., and Windischberger, C., "Slice-timing effects and their correction in functional {MRI}," *NeuroImage*, vol. 58, no. 2, pp. 588 – 594, 2011. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1053811911007245

Smith, A. M., Lewis, B. K., Ruttimann, U. E., Ye, F. Q., Sinnwell, T. M., Yang, Y., Duyn, J. H., and Frank, J. A., "Investigation of low frequency drift in fmri signal," *NeuroImage*, pp. 526–533, 1999. [Online]. Available: http://dx.doi.org/10.1006/nimg.1999.0435

Smith, S. M., Jenkinson, M., Woolrich, M. W., Beckmann, C. F., Behrens, T. E. J., Johansen-Berg, H., Bannister, P. R., Luca, M. D., Drobnjak, I., Flitney, D. E., Niazy, R. K., Saunders, J., Vickers, J., Zhang, Y., Stefano, N. D., Brady, J. M., and Matthews, P. M., "Advances in functional and structural mr image analysis and implementation as fsl," *NeuroImage*, vol. 23, Supplement 1, pp. 208–219, 2004. [Online]. Available: http://dx.doi.org/10.1016/j.neuroimage.2004.07.051

Talairach, J. and Tournoux, P., *Co-Planar Stereotaxitaxic Atlas of the Human Brain: 3-D Proportional System: An Approach to Cerebral Imaging*. Thieme, 1988.

Thirion, B., Pinel, P., Mériaux, S., Roche, A., Dehaene, S., and Poline, J.-B., "Analysis of a large fmri cohort: Statistical and methodological issues for group analyses," *NeuroImage*, vol. 35, no. 1, pp. 105–120, 2007. [Online]. Available: http://dx.doi.org/10.1016/j.neuroimage.2006.11.054

Univa Corporation, "Univa grid engine," 2013, (Visited 5.11.2013). [Online]. Available: http://www.univa.com/products/grid-engine.php

Vincenzoni, L. and Leone, S., "The Good, the Bad, and the Ugly," 1966.

Vinck, M., van Wingerden, M., Womelsdorf, T., Fries, P., and Pennartz, C., "The pairwise phase consistency: a bias-free measure of rhythmic neuronal synchronization," *Neuroimage*, vol. 51, no. 1, pp. 112–122, 2010. [Online]. Available: http://dx.doi.org/10.1016/j.neuroimage.2010.01.073

Wilson, S. M., Molnar-Szakacs, I., and Iacoboni, M., "Beyond superior temporal cortex: Intersubject correlations in narrative speech comprehension," *Cerebral Cortex*, vol. 18, no. 1, pp. 230–242, 2008. [Online]. Available: http://cercor.oxfordjournals.org/content/18/1/230.abstract

Wink, A. M. and Roerdink, J. B. T. M., "Denoising functional mr images: A comparison of wavelet denoising and gaussian smoothing," *IEEE Transactions on Medical Imaging*, vol. 23, no. 3, pp. 374–387, 2004. [Online]. Available: http://dx.doi.org/10.1109/TMI.2004.824234

Woolrich, M. W., Ripley, B. D., Brady, M., and Smith, S. M., "Temporal autocorrelation in univariate linear modeling of fmri data," *NeuroImage*, vol. 14, no. 6, pp. 1370–1386, 2001. [Online]. Available: http://dx.doi.org/10.1006/nimg.2001.0931

Woolrich, M. W., Behrens, T. E., Beckmann, C. F., Jenkinson, M., and Smith, S. M., "Multilevel linear modelling for fmri group analysis using bayesian inference." *NeuroImage*, vol. 21, no. 4, pp. 1732–1747, 2004. [Online]. Available: http://dx.doi.org/10.1016/j.neuroimage.2003.12.023

Worsley, K. J. and Friston, K. J., "Analysis of fmri time-series revisited—again," *NeuroImage*, vol. 2, no. 3, pp. 173–181, 1995. [Online]. Available: http://dx.doi.org/10.1006/nimg.1995.1023

Worsley, K. J., Evans, A. C., Marrett, S., and Neelin, P., "A three dimensional statistical analysis for cbf activation studies in human brain," *Journal of Cerebral Blood Flow and Metabolism*, pp. 900–918, 1992. [Online]. Available: http://dx.doi.org/10.1038/jcbfm.1992.127

Yarkoni, T., Poldrack, R. A., Nichols, T. E., Van Essen, D. C., and Wager, T. D., "Large-scale automated synthesis of human functional neuroimaging data." *Nature methods*, vol. 8, no. 8, pp. 665–670, 2011. [Online]. Available: http://dx.doi.org/10.1038/nmeth.1635

Yoo, A. B., Jette, M. A., and Grondona, M., "Slurm: Simple linux utility for resource management," in *Job Scheduling Strategies for Parallel Processing*, ser. Lecture Notes in Computer Science, Feitelson, D., Rudolph, L., and Schwiegelshohn, U., Eds. Springer Berlin Heidelberg, 2003, vol. 2862, pp. 44–60. [Online]. Available: http://dx.doi.org/10.1007/10968987_3

Zandbelt, B., Gladwin, T., Raemaekers, M., van Buuren, M., Neggers, S., Kahn, R., Ramsey, N., and Vink, M., "Within-subject variation in bold-fmri signal changes across repeated measurements: Quantification and implications for sample size," *NeuroImage*, vol. 42, no. 1, pp. 196–206, 2008. [Online]. Available: http://www.scopus.com/inward/record.url?eid=2-s2.0-45849117584&partnerID=40&md5=0f16b295b32cf64731d4a384237185d1

Zatorre, R. J., Fields, R. D., and Johansen-Berg, H., "Plasticity in gray and white: neuroimaging changes in brain structure during learning," *Nature neuroscience*, vol. 15, no. 4, pp. 528–536, 2012. [Online]. Available: http://dx.doi.org/10.1038/nn.3045

Zijdenbos, A. P., Dawant, B. M., Margolin, R. A., and Palmer, A. C., "Morphometric analysis of white matter lesions in mr images: method and validation," *Medical Imaging, IEEE Transactions on*, vol. 13, no. 4, pp. 716–724, 1994. [Online]. Available: http://dx.doi.org/10.1109/42.363096

# Publications

# Publication I

Kauppi J-P., Pajula J., Tohka J. "A versatile software package for inter-subject correlation based analysis of fMRI", *Frontiers In Neuroinformatics*, volume 8, number 2, 2014.

frontiers in
NEUROINFORMATICS

METHODS ARTICLE
published: 31 January 2014
doi: 10.3389/fninf.2014.00002

# A versatile software package for inter-subject correlation based analyses of fMRI

***Jukka-Pekka Kauppi[1,2†], Juha Pajula[3†] and Jussi Tohka[3]****

[1] *Department of Computer Science and HIIT, University of Helsinki, Helsinki, Finland*
[2] *Brain Research Unit, O.V. Lounasmaa Laboratory, School of Science, Aalto University, Espoo, Finland*
[3] *Department of Signal Processing, Tampere University of Technology, Tampere, Finland*

***Edited by:***
*Xi Cheng, Lieber Institue for Brain Development, USA*

***Reviewed by:***
*Ralf Schmaelzle, University of Konstanz, Germany*
*Fani Deligianni, University College London, UK*

***Correspondence:***
*Jussi Tohka, Department of Signal Processing, Tampere University of Technology, PO Box 553, 33101 Tampere, Finland*
*e-mail: jussi.tohka@tut.fi*

[†] *These authors have contributed equally to this work.*

In the inter-subject correlation (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. This implies that ISC can be used to analyze fMRI data without explicitly modeling the stimulus and thus ISC is a potential method to analyze fMRI data acquired under complex naturalistic stimuli. Despite of the suitability of ISC based approach to analyze complex fMRI data, no generic software tools have been made available for this purpose, limiting a widespread use of ISC based analysis techniques among neuroimaging community. In this paper, we present a graphical user interface (GUI) based software package, ISC Toolbox, implemented in Matlab for computing various ISC based analyses. Many advanced computations such as comparison of ISCs between different stimuli, time window ISC, and inter-subject phase synchronization are supported by the toolbox. The analyses are coupled with re-sampling based statistical inference. The ISC based analyses are data and computation intensive and the ISC toolbox is equipped with mechanisms to execute the parallel computations in a cluster environment automatically and with an automatic detection of the cluster environment in use. Currently, SGE-based (Oracle Grid Engine, Son of a Grid Engine, or Open Grid Scheduler) and Slurm environments are supported. In this paper, we present a detailed account on the methods behind the ISC Toolbox, the implementation of the toolbox and demonstrate the possible use of the toolbox by summarizing selected example applications. We also report the computation time experiments both using a single desktop computer and two grid environments demonstrating that parallelization effectively reduces the computing time. The ISC Toolbox is available in https://code.google.com/p/isc-toolbox/

**Keywords: functional magnetic resonance imaging, naturalistic stimulus, re-sampling test, Matlab, grid-computing, GUI**

## 1. INTRODUCTION

Most neuroimaging studies, such as those based on functional magnetic resonance imaging (fMRI), have so far utilized relatively simple static stimuli to analyze brain functions (Spiers and Maguire, 2007). However, the human brain has evolved to function in a tremendously stimulating world and the investigation of complex brain functions, including socio-emotional or comprehension-related processes, is limited when using highly controlled/simplistic experimental setups, because these functions are only triggered under highly complex stimuli. There is an increasing interest in studying the human brain function with dynamic, continuous stimuli that are designed to be closer to normal everyday life than in conventional, strictly controlled research paradigms. The used stimuli can be, for example, a movie. This kind of fMRI data cannot be straight-forwardly analyzed based on a general linear model (GLM), because a GLM requires a reference time course of the task that is impossible to obtain for a multi-dimensional stimulus such as a movie, unless focusing the data-analysis on a specific feature of the stimuli. For this reason, new data-driven methodologies are needed. The use of novel experimental setups involving rich stimuli and data-driven analysis methods which are particularly designed to study complex brain functions opens up entire new fields for neuroscience research.

Inter-subject correlation (ISC) based analysis, originally introduced by Hasson et al. (2004), is a conceptually simple approach to analyze fMRI data acquired under naturalistic stimuli. In the ISC based analysis, the extent of shared processing across subjects during the experiment is determined by calculating correlation coefficient between the fMRI time series of the subjects in the corresponding brain locations. This way, ISC based analyses effectively avoid the modeling of the stimuli.

ISC based analyses have been previously applied to analyze fMRI data collected during complex stimuli or tasks, including movies (Hasson et al., 2004; Jääskeläinen et al., 2008; Kauppi et al., 2010b; Nummenmaa et al., 2012), TV news reports (Schmälzle et al., 2013), auditory and audiovisual narratives (Wilson et al., 2008), pieces of music (Abrams et al., 2013) and aesthetic

performances (Jola et al., 2013). ISC based analysis has also been used for feature selection as a part of multivariate pattern analysis of data collected during a movie experiment (Kauppi et al., 2011). There can be different motivations to apply ISC based analysis for fMRI data. One can address specific neuroscientific research questions (some examples are provided in section 3) or simply try to make sense of highly complex fMRI data to generate new hypotheses. Whatever the motivation, it is important to keep in mind that the ISC is primarily a measure of *shared* hemodynamic activity across subjects and not a measure of hemodynamic activity *per se*. However, as shown by Pajula et al. (2012), when equipped with proper nonparametric statistical procedures (Kauppi et al., 2010b), ISC based methods can be used for detecting traditional fMRI activations without requiring specific, *a-priori* stimulus time course models.

Despite of the suitability of the ISC based approach to analyze complex fMRI data, no generic software tools have been made available for this purpose, limiting a widespread use of ISC based analysis techniques among neuroimaging community. Reliable and sophisticated ISC based analysis requires management of several nontrivial methodological, computational, and visualization related issues (such as heavy computational and memory load of the analysis, the choice of a proper ISC measure, handling non-standard statistical significance testing, and the visualization of multidimensional time-varying ISC maps). Hence, it is obvious that a toolbox solving these issues would be highly beneficial and can substantially simplify the use of the ISC based analysis among neuroscientists, consecutively advancing our understanding of complex human brain functions.

We have previously introduced a framework for the basic ISC based analysis (Kauppi et al., 2010b) and started building an open source, graphical user interface (GUI) based Matlab toolbox, termed the ISC toolbox, for a generic, ISC based analysis of fMRI. A set of visualization tools—particularly designed for the ISC analyses—are integrated to the GUI. In this paper, we describe the methods behind of the ISC toolbox that implements, in addition to the basic ISC analysis, many advanced ISC based computations such as phase ISC, time-windowed ISC, and comparison of ISCs between different stimuli. We will describe the analysis methods, explain the rationales behind them and demonstrate their potential use by reviewing selected example application studies.

As the ISC based analyses are data and computation intensive, the ISC toolbox is equipped with mechanisms to execute the parallel computations in a cluster environment automatically and with an automatic detection of the cluster environment in use. Currently, SGE-based environments [Unity Grid Engine (Univa Corporation, 2013), Son of a Grid Engine (Love, 2013), or Open Grid Scheduler (Scalable Logic, 2013)] and Slurm environment (Yoo et al., 2003) are supported. As there are ISC method-specific challenges in the parallelization, we will describe the automatic parallelization mechanisms in the paper. The ISC toolbox (the current version is 2.0) is available in https://code.google.com/p/isc-toolbox/

The organization of the paper is as follows. In section 2, after providing an overview of the toolbox, we will detail the ISC methods (section 2.2), describe the implementation of the toolbox (section 2.3), and briefly describe a set of visualization tools, customized to the ISC analyses (section 2.4). In section 3, we demonstrate the use of ISC-based analyses by reviewing selected studies. In section 4, as we consider cluster computing features of the toolbox important, we present the computation time experiments demonstrating the added value of parallel computing. Section 5 discusses current limitations and future directions of the toolbox and section 6 concludes the paper.

## 2. MATERIALS AND METHODS

### 2.1. OVERVIEW AND USAGE OF ISC TOOLBOX

The ISC toolbox is designed for generic ISC based analysis of fMRI data. No information about the stimulus is required to carry out the analysis, making the toolbox suitable to analyze nearly any kind of fMRI data. Naturally, data from at least two subjects are needed for the analysis because the analysis procedure is based on voxel-wise correlations of fMRI time-series across subjects. A normal desktop computer equipped with the Matlab is sufficient to carry out the basic ISC analysis in many situations. However, in certain situations it is recommended to utilize a computer cluster to carry out the analysis. For instance, the use of cluster can be meaningful if the number of subjects is high (tens of subjects), advanced ISC analyses need to be computed, or reliable re-sampling based nonparametric statistical inference is needed to construct ISC maps. The toolbox can efficiently and automatically utilize cluster environment, allowing easy and fast ISC based analysis.

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. The main window of the startup GUI is shown in **Figure 1** to demonstrate the main features of the ISC toolbox. Using the startup GUI, a user can easily select the appropriate analyses and their parameters. In the left side of the panel, a user chooses a descriptive project name and the destination folder of the analysis. For a large textbox ("Subject source files"), a user adds the names of the files containing fMRI time-series of the subjects used in the analysis. The toolbox assumes that fMRI signals have been preprocessed and preferably registered to a standard template. Preprocessing and registration algorithms are not implemented in the ISC toolbox because well developed free software packages exist for these purposes. Preprocessed and registered fMRI data sets of the subjects should be given either in nifti- or mat-format as 4-dimensional (a 3-dimensional position coordinate and time) matrices. If several acquisitions are available for each subject or acquisitions for more than one group are available, a user can analyze them all by adding more sessions to the project. The left side of the panel also contains buttons for parameter validation and for launching the main program which computes ISC maps once the parameters have been successfully validated. After running the main program, the visualization GUI to analyze results can be launched from the separate button. There is also an option to export parameters to Matlab's workspace (using the button "Export to workspace"). Automatic postprocessing

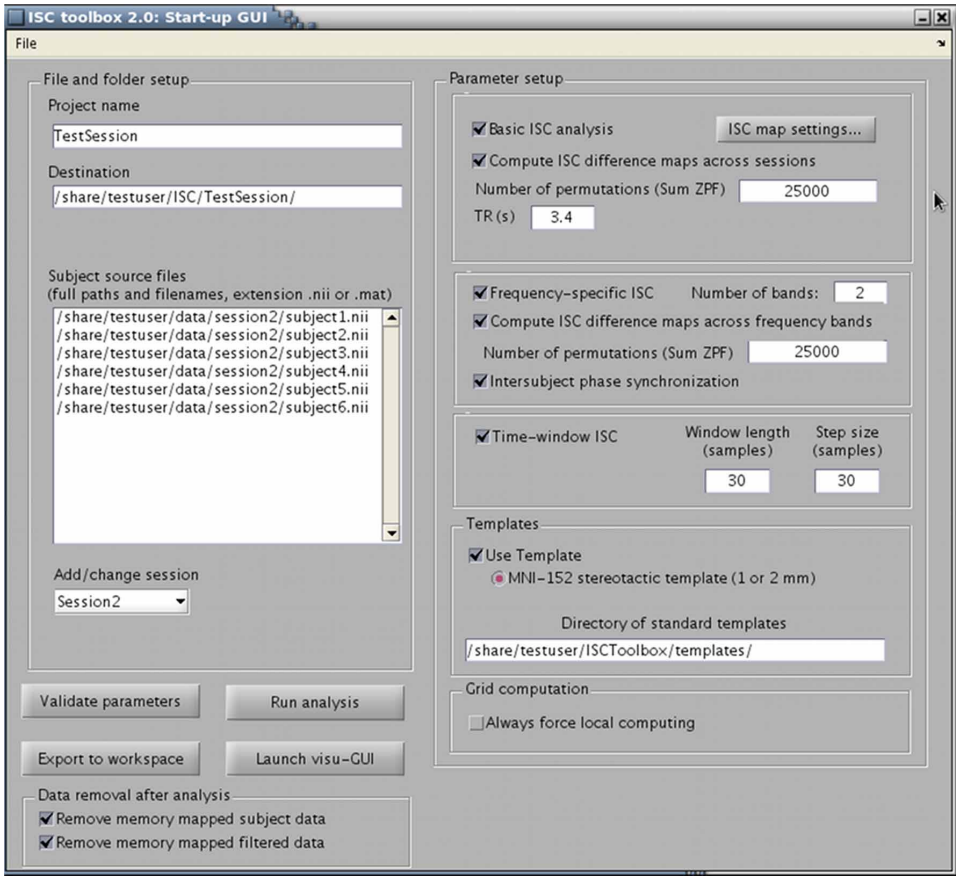



**FIGURE 1 | ISC Toolbox's startup GUI where user can define the parameters for analysis, test them, run the ISC based computations and launch a separate GUI for visualization of the results.**

operations can also be used to remove portions of data generated during the analysis to free disk space.

Different ISC analysis options are selected from the right side of the panel. A *basic ISC analysis* includes the generation of the ISC maps including thresholding of the maps based on a nonparametric statistical test. Details of this analysis can be specified from a separate panel under a button "ISC map settings." If more than one session is added to the project, it is possible to compute *ISC difference maps* to investigate whether ISCs in some of the sessions (conditions) are higher than in the others. *Frequency-specific ISC* decomposes fMRI time-series of the subjects to frequency sub-bands and computes and thresholds ISC maps for each sub-band. *Time-window ISC* computes ISC maps for several consecutive time-frames. *Inter-subject phase synchronization* combines the localization of inter-subject similarities in space, time, and frequency. These analyses are explained in section 2.2.

An arbitrary volume size can be used to compute ISC maps as long as the volume is same across subjects. However, the GUI built for the visualization of the results assumes that all fMRI data sets have been registered to a common MNI152 template. The toolbox also assumes that Harvard-Oxford cortical and sub-cortical brain atlases are available to compute and visualize inter-subject similarities for selected brain regions. Hence, to allow convenient analysis of the results, it is highly recommended to register the data to the MNI template prior to ISC analysis as well as to have the Harvard-Oxford brain atlases available. The anatomical template, atlases, and the brain mask for limiting ISC computations only for the voxels within the brain are freely provided with the FSL software package. The directory including the corresponding nifti-files should be provided in the startup GUI (subpanel "Templates"). The use of a computational cluster can be disabled under the panel "Grid computation" if needed.

After all parameters have been set and validated, they are automatically saved under the project directory in a single structure array called "Params" (the parameters can also be saved or the existing parameters can be loaded by a user from the file-menu in the upper left corner). The main program performs all ISC based computations defined in this parameter structure array. The program saves intermediate and final results of the computations to the project folders. The visualization GUI allows flexible analysis of ISC maps over an anatomical template together with the brain atlases. It also allows exporting interesting data to Matlab's Workspace for customized analysis.

## 2.2. ISC METHODS

### 2.2.1. Generation and visualization of the ISC maps

A correlation coefficient [1] is a natural measure of similarity between fMRI time-courses of two subjects. ISC toolbox allows an analysis of the similarities in the time-courses across multiple subjects. We compute the mean of the voxel-wise correlation coefficients across all possible subject pairs as (Kauppi et al., 2010b):

$$\bar{r} = \frac{1}{N(N-1)/2} \sum_{i=1}^{N} \sum_{j=2, j>i}^{N-1} r_{ij}, \tag{1}$$

where $\bar{r}$ denotes a group-level ISC in a given voxel (a voxel index is omitted for clarity), $N$ is the total number of subjects, and $r_{ij}$ is the correlation coefficient between fMRI time-courses of subjects $i$ and $j$. Note that because $r_{ii} = 1$ and $r_{ij} = r_{ji}$, it is sufficient to compute correlation coefficients across $N(N-1)/2$ subject pairs (instead of $N^2$ pairs). However, because the number of subject pairs increases approximately quadratically with $N$ and Equation (1) is computed for every voxel within the brain, it may be necessary to compute extremely high number of correlation coefficients (in the order of $10^8$) even for the most basic ISC analysis, rendering the analysis procedure computationally demanding.

We briefly explain our preference to $\bar{r}$ as the test-statistic, particularly over a related one used by Lerner et al. (2011). The main reason is that the test statistic $\bar{r}$ can be seen as an estimator of the true (but unknown) population ISC $\rho$ under the model that $\rho_{ij} = \rho + \epsilon_{ij}$, where $\rho_{ij}$ is the true correlation between subjects $i$ and $j$ and $\epsilon_{ij}$, with zero-expectation, models the between subject-pair variation. More specifically, if $r_{ij}$ approaches $\rho_{ij}$ and $\rho_{ij}$ approaches $\rho$, then $\bar{r}$ approaches $\rho$. Lerner et al. (2011) computed the average correlation of the subject time course and average time course of remaining subjects. This is closely related to $\bar{r}$ statistic[2] and neither one seems to be quantitatively better than the other. However, the statistic in Lerner et al. (2011) cannot be straight-forwardly interpreted as an estimator of the population ISC in an above sense, which results in our preference of $\bar{r}$.

### 2.2.2. Nonparametric re-sampling test

The correlation coefficients $r_{ij}$ in Equation (1) are not independent because each subject is present in more than one subject pair (e.g., $r_{ij}$ and $r_{kj}$ are overlapping because they both depend on the same time-series measured from subject $j$). Also, it is well known that BOLD-fMRI signals are temporally correlated. Therefore, the standard tests for assessing the significance of $\bar{r}$ are not valid. We use a fully nonparametric re-sampling based method to evaluate the significance of $\bar{r}$ (Kauppi et al., 2010b). In this method, we perform a test against a null hypothesis that $\bar{r}$ statistic is the same as for data with no specific time-structure. To compute a "null" re-sampling distribution, we circularly shift each subjects time-series by a random amount so that they are no longer aligned in time across the subjects, and then calculate $\bar{r}$ statistic. This way we can account for temporal autocorrelations present in the fMRI data. In practice, calculation of all the possible time shift combinations is computationally prohibitive and the distribution is approximated with finite number of realizations, randomizing the experiment across voxels and time-points, by default 100 million realizations are generated. To obtain critical thresholds for significant ISCs, we first compute $p$-values of the true realizations for each voxel based on the null distribution and then correct the values using the false discovery rate (FDR) based multiple comparisons correction (Benjamini and Hochberg, 1995). Using our visualization tool, it is possible to investigate thresholded ISC maps over an anatomical template with different critical thresholds.

### 2.2.3. Parametric t-test

The ISC toolbox contains an option to threshold group-level ISC maps also based on a simple parametric test proposed by Wilson et al. (2008). For this test, correlation coefficients are first transformed to z-scores using a Fisher's z transformation:

$$z_{ij} = \frac{1}{2} \log \left( \frac{1 + r_{ij}}{1 - r_{ij}} \right). \tag{2}$$

Then, a one-sample $t$-test with $N(N-1)/2 - 1$ degrees of freedom is performed under a null hypothesis that the ISC is zero. Note that the independence assumption of the observations made by the test is violated in practice.

[1] By correlation coefficient, we refer to a standard Pearsons correlation coefficient.

[2] Let $\mathbf{s}_i$ denote the time course of the subject $i$ that is de-meaned and normalized to unit length so that $||\mathbf{s}_i|| = 1$ (the conclusion of this analysis does not depend on the normalization to the unit length but the analysis is simplified by that assumption). Now, $r_{ij}$ can be written as an inner-product $r_{ij} = \mathbf{s}_i^T \mathbf{s}_j$. Define a test statistic similarly to Lerner et al. (2011)

$$\bar{l} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{s}_i^T \left( \frac{1}{Z_{-i}(N-1)} \sum_{j=1,\, j \neq i}^{N} \mathbf{s}_j \right),$$

where $Z_{-i} = ||(1/(N-1)) \sum_{j=1,\, j \neq i}^{N} \mathbf{s}_j||$. A straight-forward computation yields

$$\bar{l} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{s}_i^T \left( \frac{1}{(N-1)Z_{-i}} \sum_{j=1,\, j \neq i}^{N} \mathbf{s}_j \right) = \frac{1}{N} \frac{1}{N-1} \sum_{i=1}^{N} \sum_{j \neq i} \frac{\mathbf{s}_i^T \mathbf{s}_j}{Z_{-i}}$$

$$= \frac{1}{N^2 - N} \sum_{i=1}^{N} \sum_{j \neq i} \frac{r_{ij}}{Z_{-i}} = \frac{1}{\frac{N^2-N}{2}} \sum_{i=1}^{N} \sum_{j=2,\, j>i}^{N} \frac{r_{ij}}{Z_{-i}},$$

since $r_{ij} = r_{ji}$. It can be seen that this is a weighted version of $\bar{r}$, where the weights are proportional to the standard deviations of the average time courses.

#### 2.2.4. Generation and visualization of the ISC difference maps

With the ISC toolbox, it is also possible to generate and visualize ISC *difference maps* to investigate if there are significant differences in the ISCs between two conditions. For instance, in studies where same subjects are scanned twice under different stimuli, it can be highly interesting to analyze whether or not ISC was stronger in one of the conditions. We use a modified Pearson-Filon statistic based on Fisher's z-transformation (ZPF; Raghunathan et al., 1996) for this purpose, which is a recommended statistic for testing if two nonoverlapping but dependent correlation coefficients are different (Krishnamoorthy and Xia, 2007). Consider four time-series $\mathbf{t}_i^a$, $\mathbf{t}_j^a$, $\mathbf{t}_i^b$ and $\mathbf{t}_j^b$ measured from two subjects $i$ and $j$ in two conditions $a$ and $b$. The corresponding correlation coefficients $r_{ij}^a$ and $r_{ij}^b$ are nonoverlapping, because they have been computed using different time-series. However, stimuli used in two conditions $a$ and $b$ may not be independent, making a dependency assumption plausible. We extend the pairwise ZPF statistic for group-level analysis by combining the pairwise statistic from all subject pairs, and design a fully nonparametric test to assess the significance of the resulting group-level statistic (Reason et al., under review). Our final "sum ZPF" statistic is given by:

$$\mathrm{ZPF}_{\Sigma ij}^{ab} = \sum_{i=1}^{N} \sum_{j=2,\, j>i}^{N-1} \frac{(z_{ij}^a - z_{ij}^b)\sqrt{(T-3)/2}}{\sqrt{1-\mathrm{cov}(r_{ij}^a, r_{ij}^b)} \Big/ \left[\left(1-(r_{ij}^a)^2\right)(1-(r_{ij}^b)^2)\right]}, \quad (3)$$

where $z_{ij}^a$, $z_{ij}^b$ are the Fisher's z transforms [see Equation (2)] of the correlation coefficients $r_{ij}^a$, $r_{ij}^b$, respectively, $T$ is the length of a time-course and $\mathrm{cov}(r_{ij}^a, r_{ij}^b)$ is a large scale covariance (Raghunathan et al., 1996). The test is performed under the null hypothesis that each ZPF value is drawn from a distribution with zero mean, which occurs when there is no difference in ISC between the conditions. The approximate permutation distribution is generated by randomly flipping the sign of pairwise ZPF statistics before calculating Equation (3) using a subsample of all possible random labelings. Maximal and minimal statistics over the entire image corresponding to each labeling are saved to account for multiple comparisons by controlling family-wise error rate (FWER; Nichols and Holmes, 2002). Due to the symmetry of the distribution, thresholds for both directions are obtained with this procedure. The default number of random permutations over the whole image is 25,000.

Note that we cannot readily confirm the full exchangeability under the null hypothesis for the permutation test since: (1) fMRI time series are autocorrelated and (2) the subject pairs are not independent. Assuming temporal independence and normality, the ZPF-statistic can be shown to be distributed according to the standard normal distribution under the null hypothesis of no correlation difference (Raghunathan et al., 1996), which is enough to ensure the correctness of the test (Good, 2005). However, it is unclear to what extent this distributional result holds for the ISC analysis. We performed here a simple Monte Carlo simulation that verified that the ZPF statistics are normally distributed with a constant variance, not dependent on the (true) values of $r_{ij}^a = r_{ij}^b$, thus partially verifying the permutation test. The experiment and its results are summarized in **Figure 2**.

#### 2.2.5. Frequency-specific ISC analysis

The ISC toolbox contains an option to analyze ISCs in distinct frequency sub-bands. The approach is well-motivated because real-world events and stimuli unfold over multiple time-scales (Kauppi et al., 2010b). For instance, features of visual stimuli, spoken sentences, or the development of social interaction may unfold over very different time-scales. Thus, it is plausible to assume that the brain processes information in distinct frequency sub-bands. In the frequency-specific ISC analysis, we first filter the original time-series of each voxel (and subject) to multiple frequency sub-bands using an octave filter bank based on stationary wavelet transformation (SWT; Kauppi et al., 2010b). After band-pass filtering each fMRI time-series, we compute ISCs using the Equation (1) voxel-wise separately within each frequency sub-band and threshold the ISC maps using the same test as described in section 2.2.1.

It has been shown previously that wavelets are well-suited to analyze fMRI data because of certain properties of the cortical fMRI time-series, such as 1/f -like frequency characteristics (Bullmore et al., 2004). Moreover, the SWT algorithm is specifically suited to our analysis because it performs a time-invariant (Bradley, 2003) transformation unlike the discrete wavelet transform (DWT). In practice, this property means that a small difference in the hemodynamic delays of two fMRI time series transforms into a similar small difference in the filtered signals, allowing consistent estimation of the correlation coefficients between the subjects time-series after performing the filtering. For the DWT, even a minor delay between two identical input signals might cause a large difference in the filtered signals, making it much less-suited algorithm for frequency-specific ISC analysis. The SWT algorithm can be efficiently implemented using a sub-band coding scheme based on successive decimations of so called quadrature mirror filters (QMFs) and convolution operations (Vetterli and Kovačević, 1995).

We use Daubechies scaling and wavelet functions as a default filter option as they satisfy a necessary QMF relationship (Vetterli and Kovačević, 1995) and have been successfully applied to fMRI data earlier (Bullmore et al., 2001; Achard et al., 2006). The maximum degree of the polynomials the scaling function can reproduce is called the number of the vanishing moments. The number of Daubechies filter coefficients are associated with the number of the vanishing moments by the equation $K = 2V$, where $K$ is the number of filter coefficients and $V$ is the number of vanishing moments. We use short filters of length $K = 4$ as a default analysis option which are flexible enough to encode polynomials with two coefficients (both constant and linear signal components). In principle, the localization in the frequency domain could be improved by using higher filter lengths, but the use of long filters increases computation time (SWT needs to be computed separately for the time-series of every subject for each brain voxel) and makes the detection of rapid signal changes less accurate. In addition, because typical fMRI measurements


**FIGURE 2 | ZPF Monte Carlo simulation.** We simulated fMRI time series for two subjects and two conditions as $\mathbf{x}_s^a = \beta \mathbf{t}^a + \mathbf{n}_s^a$, where $\mathbf{x}_s^a(s,\ a = 1, 2)$ are the simulated time series, $\mathbf{t}^a$ is the common time-series between the two subjects in the condition $a$, $\mathbf{n}_s^a$ is pink noise [generated as in Pajula et al. (2012)], and $\beta$ is selected so that the true correlation between the subjects' time-series is $\rho$, which is varied during the simulation. $\mathbf{t}^a$ was generated by smoothing white Gaussian noise with a box filter and convolving the resulting time series by a hemodynamic response function. The times series length was 100 with modeled TR of 2*s*. Note that $\rho$ has the same value for both conditions, according to the null-hypothesis. The simulation does not model for the dependence between the two conditions. The simulation was repeated 100000 times for each $\rho = -0.5, -0.4, \ldots, 0.5$. **(A)** Shows the histogram of ZPF values when $\rho = 0.3$, **(B)** Shows the QQ-plot for the same case, and **(C)** Shows the average value and the standard deviation of the ZPF-statistics as a function of $\rho$, where a slight dependence of the standard deviation on $\rho$ is observed. As it is visible (1) the distribution of the ZPF was Gaussian (with a larger variance than 1), and (2) its parameters did not markedly depend on the value of $\rho$.

contain relatively low number of time points, short filters are preferred to minimize boundary artifacts. Daubechies basis functions are optimal in the sense that they provide the shortest filter length for the given number of vanishing moments. However, also other basis functions have been proposed for fMRI data analysis. For instance, Ruttimann et al. (1998) used symmetric spline wavelets because of their phase-preserving property.

#### *2.2.6. Time window ISC analysis*

When analyzing complex fMRI data sets such as those collected during a movie watching, it is likely that ISCs vary drastically over the experiment. To analyze how ISC varies over time, it can be highly useful to compute ISC maps for several consecutive possible overlapping time windows. With the ISC toolbox, a user can specify suitable time window parameters (window length, step length between two consecutive windows) and compute "short-time ISC maps" for each window. To obtain these maps, we compute $\bar{r}$ statistic (across all voxels and subjects) within each time-window and assess the significance of the ISCs as described above. We randomize the generation of the null distribution across all time windows which leads to a common threshold for all windows. The length of the time window has to be sufficient to obtain reliable estimates of $\bar{r}$ for each time window. The choice depends on the number of subjects and the type of the stimulus. Therefore, it is not straight-forward to give exact suggestions about the minimal time-window length. However, window lengths as short as 10 samples have been used (Nummenmaa et al., 2012).

The toolbox allows the visualization of the time window ISC maps over an anatomical template. It also automatically computes the mean of $\bar{r}$-values across voxels within different brain region-of-interest (ROIs), allowing plotting ROI-averaged ISCs over time. These curves can be correlated with the features of the stimuli, behavioral ratings or other variables of interest.

#### *2.2.7. Intersubject phase synchronization*

Time window ISC and frequency-specific ISC analyses can provide neuroscientifically meaningful insights into complex fMRI data. An obvious way to combine benefits of both approaches is to compute frequency-specific ISC maps in several time windows to investigate temporal evolution of the ISCs in specific time-scales. ISC toolbox automatically computes also these maps if the user performs both time window ISC and frequency-specific ISC analyses. A limitation of this approach is that the temporal resolution of the analysis can be modest because each time window must contain several time points to allow meaningful interpretation of the correlation coefficient. This problem is most prominent in the lowest frequency sub-bands because the temporal resolution of slow fluctuations is inherently poor as stated by the time-frequency uncertainty principle (Cohen, 1995). To increase the temporal resolution of the time-varying analysis in distinct frequency sub-bands, we propose using phase synchronization between subjects as a measure of inter-subject similarity. A similarity measure based on instantaneous phase allows the analysis of the band-pass filtered signals on the basis of inherent temporal resolution of the time series. This is in contrast to the time window ISC analysis for which the resolution is further limited by the length of the time-window.

Many phase synchronization measures have been designed to analyze functional neuroimaging signals (Vinck et al., 2011) but they are mainly used to analyze electroencephalography and magnetoencephalography signals. Unlike these signals, fMRI time-series may not be characterized by oscillatory activity. However, the analysis of the instantaneous phases still remains a valid method to characterize a specific interrelation between phases (Pikovsky et al., 2000; Laird et al., 2002). To extract phase information, complex-valued analytic time-series must be available. Hence, we apply the Hilbert transform (Goswami and Hoefel, 2004) to the fMRI time-series to obtain their corresponding

analytic signals [3]. We take the *absolute angular distance* (Vinck et al., 2010) between the time-series of two subjects as a dissimilarity measure:

$$p_{ij}(t) = |\theta_i(t) - \theta_j(t)| \mod \pi, \quad (4)$$

where $t$ is a time-point index and angles $\theta_i$, $\theta_j$ are computed based on the analytical time-series measured from subjects $i$ and $j$. This is an intuitive measure of phase interrelationship between the fMRI time-series of two subjects: If fluctuations of the (band-limited) time-series between subjects are highly similar, it is expected that the absolute phase difference is smaller than when fluctuations are different. There are different possibilities to extend this measure to group-level analysis (Glerean et al., 2012). We use a comparable definition to our ISC measure [Equation (1)] and compute the average of all subject-pairwise absolute angular distances as:

$$\overline{p}(t) = \frac{1}{N(N-1)/2} \sum_{i=1}^{N} \sum_{j=2,\, j>i}^{N-1} p_{ij}(t). \quad (5)$$

Our final measure of inter-subject phase synchronization (IPS) is the normalized version of $\overline{p}$:

$$\hat{\overline{p}}(t) = 1 - \frac{\overline{p}(t)}{\pi}. \quad (6)$$

This measure has its values always within the range [0 1], where the value 1 indicates a complete phase similarity and the value 0 corresponds to a complete *absence* of phase similarity across subjects. Similarly to time window ISCs, the ISC Toolbox allows different plotting options for IPS results. For instance, averaged IPS values within selected ROIs can be plotted over time. These curves can be then correlated with the features of the stimuli or other variables of interest.

### 2.3. IMPLEMENTATION

As explained in section 2.1, the use of the ISC Toolbox starts from the startup GUI where a user defines requested analyses and their parameters (see **Figure 1**). The GUI automatically detects the operating system and checks that all necessary software and files are available. After a user has selected desired analysis options, the GUI validates them. After a successful validation, the parameters are set in a structure array called *Params* which is saved in a mat-file. The GUI also generates the destination directory and all necessary sub-directories for the analysis results.

The computational analysis is controlled inside the main function named *runAnalysis*. The Matlab code of this function is grouped in six computational stages to clarify how the computations can be distributed across a computer cluster:

Stage 1 Binary data files for the analysis results as well as the memory map pointers to access these files are initialized. The pointers are saved in the structure called *memMaps* which is saved in the analysis destination directory. In the later stages of the program, the files are repeatedly accessed and modified using these pointers (see more information about the Matlab's memory mapping feature below).

Stage 2 The wavelet filtering for the frequency-specific ISC analysis is performed.

Stage 3 Average ISC maps are computed, including the generation of the re-sampling distributions for the assessment of statistical thresholds.

Stage 4 Critical thresholds are calculated based on the re-sampling distributions including threshold correction for multiple comparisons. In the FDR-based correction, $p$-values for statistically significant (before a multiple comparison correction) samples need to be available. These are estimated in a nonparametric fashion from the observations of the re-sampling distribution using a linear interpolation.

Stage 5 Inter-subject synchronization curves over time are computed for the time window ISC and IPS for all the brain regions and thresholds defined in the Harvard-Oxford sub-cortical and cortical atlases.

Stage 6 All the generated statistical maps in the previous stages are saved to the analysis destination folder as nifti files. This stage is always computed locally even if a grid enviroment would be available.

The grouping of the code is based on the dependencies of the analysis pipeline: the execution of the functions within any of the stages is always dependent on the results of the preceding stage and therefore cannot be performed before all previous stages have been completed and their intermediate results have been saved to the analysis destination directories. However, computations inside the loop structures *within* each computational stage are independent of each other, meaning that functions repeatedly called inside these loops can be equally well run in parallel. In practice, a user does not need to understand how the code is written because the program can automatically parallelize computations across a computer grid/cluster.

Only those stages corresponding to ISC based analyses that are requested by the user are run when executing *runAnalysis*. For example, if the frequency-specific ISC analysis is not chosen by a user, the stage 2 is skipped.

Matlab's memory mapping is a mechanism that maps a portion of a file, or an entire file, on disk to a range of addresses within an application's address space. The application can then access files on disk in the same way it accesses dynamic memory (The Mathworks Inc., 2013). This memory mapping mechanism is employed in the ISC Toolbox for three main reasons:

1. Because of a large memory demand, all the data cannot be held in the central memory all the time.
2. The traditional file I/O can be very slow especially in cluster computing environments.

[3] The analytic signal $\mathbf{x}_a(t) = x(t) + jy(t) = A(t)e^{j\theta(t)}$ can represent both the instantaneous amplitude envelope $A(t)$ and phase $\theta(t)$ of the time-series, but only phase information $\theta(t) = \arctan\left(\frac{y(t)}{x(t)}\right)$ is used to derive our phase similarity measure.

3. The memory mapping provides a mechanism for sharing the memory between multiple processes that is important for the cluster computing abilities in the ISC Toolbox.

The disadvantage of the used memory mapping mechanism is that it is highly hardware and also somewhat operating system and Matlab version dependent. The memory mapped data can become corrupt or unreadable if the used hardware or the Matlab version is changed. In the ISC Toolbox, the problem is circumvented by saving the important results out from the memory maps to nifti files. The corrected statistical thresholds are saved as Matlab's mat-file and also as a text file. Therefore, the visualization of the thresholded maps can be done afterwards easily with any visualization software. The memory mapping has been previously used in the SurfStat software within brain imaging (Worsley, 2008).

A heavy computational burden is one of the major issues when using the ISC Toolbox. Computations require large memory as mentioned already and they also take a long time to compute. Currently, the ISC Toolbox supports cluster computing in SGE-based (Oracle Grid Engine, Son of a Grid Engine, or Open Grid Scheduler) and Slurm (Simple Linux Utility for Resource Management) environments. Generally, the SGE based parallelization (Love, 2013; Scalable Logic, 2013; Univa Corporation, 2013) has been used extensively within brain imaging software such as FSL. The Slurm grid engine (GE) (Yoo et al., 2003) is currently becoming more common and for this reason also Slurm based parallelization was selected to be supported in the ISC Toolbox. The only requirements to use parallelization procedures in the toolbox are that the operating system of the used computer must be Linux and the user must have access to system running on one of these two GEs.

In both cases (SGE or Slurm), separate shell scripts must be generated for each computational stage before distributing them to the GE. The script generation and submission to GE is handled with the function *gridParser*. The *gridParser* function generates separate shell scripts for each process stage of each possible parallel process and submits these to the current GE. Simplified examples from the shell scripts generated by *gridParser* for the first stage of execution are presented in Listings 2.3 and 2.3. In this example, the project name is "ISC_test_analysis," which defines the mat-file name for the Params struct. *memMapData* function implements the stage 1 of the analysis. The only input for the function is the Params struct. The number of generated scripts varies from 4, for the basic analysis using a single CPU, to hundreds depending on the selected analyses and the degree of parallelization.

**Listing 1.** Bash script example for the Stage 1 of the analysis generated by the *gridParser* function for the SGE environment.

```
matlab -nosplash -nodisplay -nojvm -n
  odesktop -r "addpath(genpath('/home
  /testuser/ISCofficial/isc-toolbox/'));
  load('/home/testuser/ISCtest/ISC_test
  _analysis'); memMapData(Params); exit"
exit
```

**Listing 2.** Bash script example for the Stage 1 of the analysis generated by the *gridParser* function for the Slurm environment.

```
#!/bin/sh
module load matlab
matlab -nosplash -nodisplay -nojvm -n
  odesktop -r "addpath(genpath('/home
  /testuser/ISCofficial/isc-toolbox/'));
  load('/home/testuser/ISCtest/ISC_test
  _analysis'); memMapData(Params); exit"
exit
```

The monitoring of the submitted tasks is handled with the function *waitGrid*. The function requests the running processes in the GE in defined time interval and prevents the main function to continue before all submitted sub-processes are finished.

The data integrity is always a critical question within parallel computing. It must be ensured that any two processes are not interfering each other and all data are saved safely. The function *freeToWrite* was developed to maintain the data integrity. It handles a specific lock system to ensure that only one process updates the memory maps at once. The lock is based on a simple lock-file which is generated before the data are going to be saved and deleted when the saving process has been finished. Every process which updates the memory maps are using the lock system. To simplify the debugging, every lock file has its own identifier based on the name of the process and the current process ID from the GE.

### 2.4. VISUALIZATION GUI

To simplify the investigation of the ISC analysis results a separate visualization GUI, shown in **Figure 3**, was developed to interactively show the statistics maps and other results computed by the ISC Toolbox. The visualization GUI can show all the statistical maps resulting from the analyses by the toolbox. There are several software tools for high quality, interactive visualizations of the statistical maps from neuroimaging analyses. However, as far as we know, none of these is suitable for the visualization of advanced ISC analysis such as 4-D statistical maps of the time window ISC. In addition, a specialized visualization application provides additional convenience by allowing user to switch between different analysis results by a quick button press instead of a cumbersome reloading of the statistical maps one-by-one from a disk. We avoid the re-loading of the statistical maps by directly accessing the data portion of interest from the disk. The fast random access to the data is possible because the ISC results were mapped to a disk during the main analysis procedure with the aid of memory-mapping. Most of the data which are presented or used for creating the visualizations in the GUI are precomputed by the main analysis procedure and mapped to a memory. The memory-mapping minimizes the need of RAM, which enables the efficient interactive visualization and exploration of the analysis results also with slower computers. A price to pay for this added flexibility are possible cross-platform incompatibility issues, mentioned already in section 2.3, if the actual analysis is carried out with a different hardware than with which the analysis results are viewed. The simplest of these issues is the endianness, which can be changed by ticking the checkbox "Swap bytes."

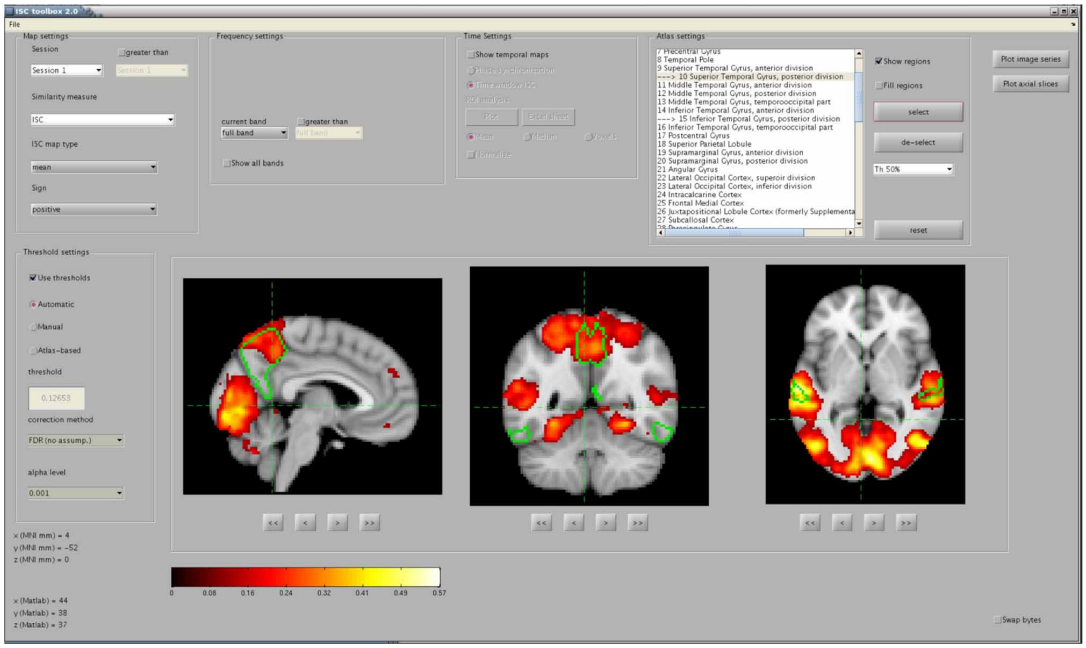



**FIGURE 3 | The main window of the visualization GUI.** In the shown analysis example, a user has located significant ISCs in several brain areas, including the precuneus cortex and the posterior division of the superior temporal gyrus, whose perimetries are shown in green color over the anatomical template. The map was thresholded and FDR corrected ($q$ <0.001) over the whole brain using the re-sampling test of section 2.2.1.

In addition to minimizing the memory consumption and the access time to the data in a disk, it is important to minimize the time that is spent for plotting accessed data on the screen. In Matlab, the plotting of the images is much faster when using indexed images (in an integer format) than using true color images or intensity images in a floating point format. Indexed images are fast to visualize because they use direct mapping of pixel values to colormap values. Hence, to maximize the browsing speed, the GUI converts ISC maps and anatomical templates from a floating point format to an integer format and combines these data into a single matrix of integer values. An appropriate colormap is then created to allow visualization of the indexed image on the screen in multiple colors. The colormap involves hot (yellow and red), cold (magenta and blue), and gray colors which allows the visualization of positive and negative ISCs as well as anatomical intensity values over a single image.

The exact appearance of the Visualization GUI on the screen depends which analyses, described in section 2.2, the user has run. For example, if only the basic ISC analysis has been run the visualization GUI enables only the analysis of ISCs across a whole session and frequency-spectrum by disabling "temporal settings" and "frequency settings" -panels. Statistical maps are shown in sagittal, coronal and axial views. The MNI coordinates of the views can be changed via the buttons below the axis. An additional option is to visualize several axial slices across the whole brain volume in a single figure. The exploration of the volume along a fourth dimension (time interval or frequency range) currently requires a button press. A user can also select Harvard-Oxford probabilistic atlas regions for the visualization over the statistical map and it is possible to view average ISCs for selected ROIs as a function of time. In addition to these visualizations, the GUI contains more advanced visualization options which allow detailed localization of ISCs in spatial, temporal and spectral dimensions.

The GUI allows fast and comprehensive visualization of the ISC analysis results in an exploratory manner. However, to address specific research questions, a further analyses not supported by the ISC Toolbox may be needed. Moreover, it may also be meaningful to customize the way how the results are visualized. For these purposes, the GUI has an option to export ISC maps and other results to the Matlab's workspace as variables. Although the ISC analysis results are also saved in a disk as Nifti-files and are freely accessible for a user, the export option allows quick and easy visualization of the threholded maps over an anatomical image and selected atlas regions for the dimensions (spatial, temporal, and spectral) of interest.

## 3. APPLICATIONS

Next, we shortly exemplify how the toolbox has been successfully used to analyze fMRI data.

### 3.1. BASIC ISC ANALYSIS FOR ACTIVATION DETECTION

A primary interest in many fMRI based imaging studies is to detect brain locations associated with a task related neural activity. Traditionally, this is achieved by a GLM based analysis,

where voxel time courses are compared to the task-derived reference time course. The application of the GLM requires explicit knowledge how stimuli are varied during the experiment and cannot therefore be used to detect activations from experiments involving complex naturalistic stimuli. Pajula et al. (2012) showed that our "basic" ISC analysis described in section 2.2.1 is a suitable method to detect task related neural activation without making any assumptions about the applied stimuli. In this study, fMRI data from 37 right-handed subjects who all had performed the same five blocked design tasks [4] were analyzed with both ISC Toolbox and a GLM based method. The idea is that the GLM-detected activations with this kind of strictly controlled and well-known tasks can be assumed to be reliable and can be treated as a gold standard. Interestingly, the comparison of the statistical maps of ISC and GLM revealed high agreement of the findings. This demonstrates that the ISC analysis can detect truly active brain regions in a manner that is completely "blind" to stimuli, making it highly promising method for detecting activity in data sets collected under naturalistic stimuli experiments.

### 3.2. ISC DIFFERENCE MAPS FOR ANALYSIS OF AESTHETIC EXPERIENCES

Understanding how spectators' brains process information during an aesthetic performance, such as a dance performance, is an interesting topic in neuroscience. To investigate this, videos of aesthetic performances can be shown to subjects while their brain activity is being measured using the fMRI. Stimuli in these experiments are very rich, making ISC based methods a natural choice for data analysis. Reason et al. (under review) used ISC toolbox to study whether auditory stimulation have an effect on the kinesthetic experience and/or the aesthetic appreciation of the spectator while watching dance. In the study, fMRI signals were acquired from 22 subjects under two different stimulus conditions: (1) a full audiovisual dance performance accompanied by the soundscapes of Bach (condition = "Bach"), and (2) the same dance performance without the music, including only visual stimuli as well as sounds of breathing and footfalls of the dancer (condition = "Breathing"). ISC toolbox was used to construct individual ISC maps of both conditions as described in section 2.2.1 as well as to construct ISC difference maps "Bach" < "Breathing" and "Bach" > "Breathing" as described in section 2.2.4.

The individual ISC maps showed large overlap in the visual and auditory cortices for both conditions. However, the analysis of the ISC difference maps revealed clusters in the temporal cortex that were unique to the different audio conditions, indicating also clear differences between the processing of the sound in the "Bach" and "Breathing" conditions. Based on detailed investigation of the ISC difference maps, Reason et al. (under review) suggested several possibilities how the presence or absence of music may influence spectators' experience. For instance, the postcentral gyrus of parietal cortex (BA 7) showed significantly greater ISC in the "Breathing" condition. The area is known for simultaneously processing multiple sensory modalities, in particular the somesthetic modality that includes touch. This somesthetic connection implies a form of motor cognition and could suggest that the "Breathing" elicited greater engagement of action understanding within body-specific mechanisms.

### 3.3. FREQUENCY-SPECIFIC ISC FOR ANALYSIS OF TEMPORAL BRAIN HIERARCHY

In our previous study (Kauppi et al., 2010b), we performed frequency-specific ISC analysis to investigate processing of movie events that occur over multiple time-scales. We analyzed fMRI data collected from the experiment (Jääskeläinen et al., 2008) where 12 subjects watched the 36 min clip of an Academy Award winning drama movie Crash (Lions Gate Films, 2005, directed by Paul Haggis; the movie was presented with sound). We constructed both frequency-specific ISC maps described in section 2.2.5 as well as ISC difference maps to compare differences in ISCs between distinct frequency subbands (see section 2.2.4).

The frequency-specific ISC analysis provided novel and interesting insights into the highly complex fMRI data. For instance, the analysis revealed that visual cortical ISC was present across the whole frequency spectrum of the fMRI signal, ISC in temporal areas occurred in all but the highest frequency band, and frontal cortical ISC was present only in the two lowest frequency bands. Hence, the frequency range showing significant ISC *contracted* when moving from lower-order sensory areas toward higher-order cortical areas. There are several possible explanations for the mappings found in this study. For instance, the findings might reflect the hierarchy of temporal receptive windows (TRWs) in the human brain, with sensory visual cortical areas showing short TRWs, and the TRWs becoming progressively longer as one ascends to functionally higher-order cortical areas (Hasson et al., 2008).

### 3.4. TIME WINDOW ISC FOR ANALYSIS OF HIGHER-ORDER BRAIN FUNCTIONS

The use of movies as stimuli in neuroimaging studies offers new possibilities to understand higher-order brain functions, such as those related to social cognition and emotions. Nummenmaa et al. (2012) used the time window ISC (which they call moment-to-moment ISC) to analyze how ISC is associated with events that elicit emotions in movies. Functional MRI data from 16 subjects were collected while they watched movies depicting unpleasant, neutral, and pleasant emotions. After scanning, participants watched the movies again and continuously rated their experience of pleasantness–unpleasantness (i.e., valence) and of arousal–calmness. Short-time ISCs for each voxel were then computed using the ISC toolbox as described in section 2.2.6, using a 17-s sliding window (a step size of the time-window was one time point). Time series of valence and arousal ratings were then used to predict temporal variation of ISCs within each voxel.

Negative valence was associated with increased ISC in the emotion-processing network (thalamus, ventral striatum, insula) and in the default-mode network (precuneus, temporoparietal junction, medial prefrontal cortex, posterior superior temporal sulcus). High arousal was associated with increased ISC in the somatosensory cortices and visual and dorsal attention networks

[4]Functional MRI data from the measurements with Functional Reference Battery tasks developed by the International Consortium for Human Brain Mapping (ICBM) were used (Mazziotta et al., 2001): http://www.loni.ucla.edu/ICBM/Downloads/Downloads_FRB.shtml.

comprising the visual cortex, bilateral intraparietal sulci, and frontal eye fields. It was proposed that negative valence synchronizes individuals brain areas supporting emotional sensations and understanding of anothers actions, whereas high arousal directs individuals attention to similar features of the environment.

### 3.5. IPS FOR TIME-VARYING ANALYSIS OF NATURALISTIC fMRI DATA

IPS described in section 2.2.7 is an alternative option for time window ISC to analyze complex fMRI data over time. Glerean et al. (2012) applied both time window ISC and IPS analysis for naturalistic fMRI data collected from 12 subjects while they watched a feature movie (for details of the experiment, see Lahnakoski et al., 2012). IPS was computed within a frequency-band of 0.04–0.07 Hz and a time window ISC was computed for several window sizes from 4 to 32 samples (corresponding to window lengths from 8 to 64-s with the TR of 2-s) using a sliding window.

A major conclusion of the study was that the IPS approach provided improved temporal resolution as compared with the time window ISC. In addition, an anatomical mapping of the whole-brain temporal average of the IPS was highly consistent with the anatomical mapping of the ISC computed across the whole movie experiment (without using time windows), indicating that the IPS is a realiable measure of inter-subject similarity.

## 4. COMPUTATION TIME

The computation time was measured in three different hardware setups utilizing the both the local and distributed computing abilities of the ISC Toolbox. The local computations were tested with Dell Optiplex 755 desktop computer equipped with Intel Core2Duo E8400 CPU @ 3.00 GHz and 5GB read access memory (RAM). The distributed computations were tested in two computing clusters. The larger cluster, called Merope, had nodes running on HP ProLiant SL390s G7 equipped with Intel Xeon X5650 CPU 2,67 GHz and minimum of 4 GB RAM / core. The GE was Slurm. The smaller of the tested computing clusters, called Outolintu, was running with SGE and had nodes running on IBM System x3550 equipped with two Intel Xeon X5450 CPUs 3.0 GHz and 32 GB RAM (with 10 GB swap) for each node.

In the Merope cluster, on average 32 processes were run simultaneously. With the Outolintu cluster the maximum of parallel processes was limited to 10 due to global usage limitations for a single user of this cluster. The computing times of cluster environments were averaged from three separated runs as in the cluster the computing time can be affected from the current load of the cluster as well as the implementation of the distributing system causes a small variation on computing time.

The computing time was measured from "the user perspective": Starting from the moment when user pushes the "Run Analysis" -button of the startup GUI to the moment when the analysis was finished. In a cluster environment, this means that the processing times of the GE were included to the total processing time.

The analyses were performed for the same measurement data which was used in earlier studies with the ISC Toolbox (Jääskeläinen et al., 2008; Kauppi et al., 2010b). The data was acquired from 12 subjects ($TR = 3.4$ s, 244 time points) and was registered to MNI152 space (for details see Kauppi et al., 2010b). The image dimensions were $91 \times 109 \times 91 \times 244$ (X × Y × Z × time) which resulted in an 840 MB file size for each subject and 9.8 GB total size of the analysis data set.

The tested ISC Toolbox setups were "basic ISC", "basic ISC + time window ISC," and "basic ISC + frequency-specific ISC." The first setup computed the ISC map across the entire length of the time-series and constructed a re-sampling distribution based on 100 million random shufflings of the time-series as in our earlier study (Kauppi et al., 2010b). The second setup was similar to the first setup except that the time window ISC with the window length and window step of 30 samples was used in addition to the basic ISC analysis. The third setup used frequency-specific ISC with three frequency sub-bands instead of time window ISC. The number of randomizations to threshold the ISC $\bar{r}$-maps was the same as in setups 1 and 2, and 25,000 random permutations for each brain voxel was used to construct a null permutation distribution of the sum ZPF statistic to allow thresholding of ISC difference maps between frequency bands.

The computing times are presented in **Figure 4**. On a single desktop computer, the computing time varied from 11 h 52 min to 25 h 20 min depending from the selected analysis. On the smaller Outolintu cluster, the corresponding times varied from 1 h 24 min to 6 h 10 min and, on the larger Merope cluster from 33 min to 3 h 25 min. Comparing local and distributed systems, the speed up factor was 10 with Outolintu and 24 with Merope in the first two setups. For the final setup with the frequency band analysis, the speed up factor was 4 with Outlintu and 7.5 with Merope.

The smaller speed up factors for the frequency band analysis was probably due to a higher number of hard drive interactions involved in this analysis as compared with the other tested analyses. In a cluster computing environment, a high number of hard drive interactions slows down the computations as the data is commonly located on a network drive and the speed of the data transfer in a network is usually clearly slower than the speed of data transfer via the internal bus of a desktop computer.

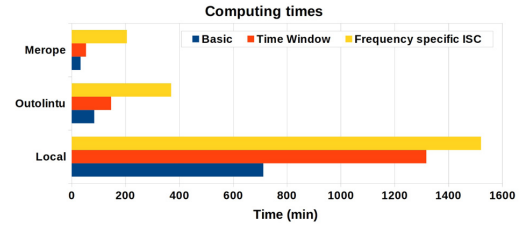



**FIGURE 4 | The computing times from desktop computer and two cluster environments.** The desktop computer was equipped with Intel Core2Duo E8400 CPU 3.00 GHz and 5 GB RAM. Ten parallel processes were run on Outolintu cluster with nodes equipped with Intel Xeon X5450 CPUs 3.0 GHz. On average, 32 parallel processes were run on Merope cluster with nodes equipped with Intel Xeon X5650 CPUs 2,67 GHz.

## 5. DISCUSSION

Branches of neuroscience investigating brain functions in experiments mimicking real-world conditions are growing rapidly and the development of data analysis methods must address an increasing diversity of research questions. The ISC based approaches can address key questions such as how processing differs between two groups (e.g., healthy vs. nonhealthy) exposed to identical complex stimuli or between two conditions (e.g., silent vs. nonsilent video). The new analyses methods incorporated in the ISC toolbox, described in sections 2.2.4–2.2.7, are one of the first attempts to help neuroscientists to address these and other aspects of neural processing. In addition to these new features, the toolbox will be continuously updated in the future to allow even more versatile analyses. For instance, the toolbox is currently limited to analyze between-subject correlations in a voxel-wise manner and does not allow more general investigations of functional correspondence between spatially disjoint brain areas across subjects. Features to analyze ISCs between different brain areas both within- and across subjects will be incorporated in the future versions of the toolbox.

One limitation of the current analysis approach is that the used ISC measure does not capture any information about the variability of the ISCs among subjects as it is simply the average of the upper-triangular (or lower-triangular) elements of the between-subject correlation matrix computed separately for each voxel [Equation (1)]. The consequence of the averaging is that interesting features of brain processing may be missed especially in higher-order brain regions where inter-subject variability is expected to be very high. To increase the sensitivity of the existing method to localize interesting brain areas as well as to perform more fine-grained ISC based analyses, it can be highly useful to preserve and analyze the entire structure of the between-subject correlation matrices. We have already taken steps toward this direction (Kauppi et al., 2010a) and will equip the toolbox with matrix-based analysis methods in the future.

One of the key issues in ISC based analyses is how to select a suitable threshold to distinguish meaningful ISC values from spurious ones. Because of the restrictive assumptions made by standard parametric statistical procedures, such as the ordinary $t$-test, we have decided to use fully nonparametric re-sampling based methods to determine the critical thresholds to improve reliability of the analysis. Despite of the flexibility of the nonparametric methods, it is important to keep in mind that also they provide only approximations of true, underlying null re-sampling distributions. This is due to finite number of realizations drawn as well as certain assumptions required by the tests which may not be fulfilled by real fMRI time-series. However, as shown by the results, our easy and fully automated mechanism which distributes calculations across a computational cluster allows drawing huge number of realizations in a relatively short time, making the generation of accurate re-samplings distributions feasible. Moreover, we showed with a simple Monte-Carlo simulation that certain critical assumptions made by the sum ZPF test are not violated in practice. In any case, further validation and improvement of our current statistical procedures is another important topic of future research.

## 6. CONCLUSIONS

We have presented a software package, named ISC Toolbox, implemented in Matlab for computing various ISC based analyses. The computations can be launched from a GUI making the use of the toolbox easy. Many advanced techniques such as time window ISC analysis, frequency-specific ISC analysis, IPS analysis and the comparison of ISCs between different stimuli are supported by the toolbox. The analyses are coupled with nonparametric re-sampling based statistical inference methods. As these analyses are computationally intensive, the ISC Toolbox is equipped with automated cluster computing mechanisms to reduce the computation time via parallelization and a marked reduction in computation time was achieved by cluster computing. The ISC Toolbox is available in https://code.google.com/p/isc-toolbox/ under the MIT open source licence.

## ACKNOWLEDGMENTS

### FUNDING

This research has been supported by the Academy of Finland under the grants 130275, 263785 to Jussi Tohka and under the National Centers of Excellence Programme in Algorithmic Data Analysis, and Computational Sciences Program.

## REFERENCES

Abrams, D. A., Ryali, S., Chen, T., Chordia, P., Khouzam, A., Levitin, D. J., et al. (2013). Inter-subject synchronization of brain responses during natural music listening. *Eur. J. Neurosci.* 37, 1458–1469. doi: 10.1111/ejn.12173

Achard, S., Salvador, R., Whitcher, B., Suckling, J., and Bullmore, E. (2006). A resilient, low-frequency, small-world human brain functional network with highly connected association cortical hubs. *J. Neurosci.* 26, 63–72. doi: 10.1523/JNEUROSCI.3874-05.2006

Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. B* 57, 289–300.

Bradley, A. P. (2003). "Shift-invariance in the discrete wavelet transform," in *Proceedings of VIIth Digital Image Computing: Techniques and Applications* (Sydney).

Bullmore, E., Fadili, J., Maxim, V., Şendur, L., Whitcher, B., Suckling, J., et al. (2004). Wavelets and functional magnetic resonance imaging of the human brain. *Neuroimage* 23, S234–S249. doi: 10.1016/j.neuroimage.2004.07.012

Bullmore, E., Long, C., Suckling, J., Fadili, J., Calvert, G., Zelaya, F., et al. (2001). Colored noise and computational inference in neurophysiological (fMRI) time series analysis: resampling methods in time and wavelet domains. *Hum. Brain Mapp.* 12, 61–78. doi: 10.1002/1097-0193(200102)12:2<61::AID-HBM1004>3.0.CO;2-W

Cohen, L. (1995). *Time-Frequency Analysis*, Vol. 778. Englewood Cliffs, NJ: Prentice Hall PTR.

Glerean, E., Salmi, J., Lahnakoski, J. M., Jääskeläinen, I. P., and Sams, M. (2012). Functional magnetic resonance imaging phase synchronization as a measure of dynamic functional connectivity. *Brain Connect.* 2, 91–101. doi: 10.1089/brain.2011.0068

Good, P. (2005). *Permutation, Parametric and Bootstrap Tests of Hypotheses*. 3rd Edn. New York, NY: Springer.

Goswami, J. C., and Hoefel, A. E. (2004). Algorithms for estimating instantaneous frequency. *Signal Process.* 84, 1423–1427. doi: 10.1016/j.sigpro.2004.05.016

Hasson, U., Nir, Y., Levy, I., Fuhrmann, G., and Malach, R. (2004). Intersubject synchronization of cortical activity during natural vision. *Science* 303, 1634–1640. doi: 10.1126/science.1089506

Hasson, U., Yang, E., Vallines, I., Heeger, D. J., and Rubin, N. (2008). A hierarchy of temporal receptive windows in human cortex. *J. Neurosci.* 28, 2539–2550. doi: 10.1523/JNEUROSCI.5487-07.2008

Jääskeläinen, I. P., Koskentalo, K., Balk, M. H., Autti, T., Kauramäki, J., Pomren, C., et al. (2008). Inter-subject synchronization of prefrontal cortex

hemodynamic activity during natural viewing. *Open Neuroimag. J.* 2, 14–19. doi: 10.2174/1874440000802010014

Jola, C., McAleer, P., Grosbras, M.-H., Love, S. A., Morison, G., and Pollick, F. E. (2013). Uni-and multisensory brain areas are synchronised across spectators when watching unedited dance recordings. *Iperception* 4, 265–284. doi: 10.1068/i0536

Kauppi, J., Jääskeläinen, I., Sams, M., and Tohka, J. (2010a). "Clustering inter-subject correlation matrices in functional magnetic resonance imaging," in *Information Technology and Applications in Biomedicine (ITAB), 2010 10th IEEE International Conference on* (Corfu), 1–6.

Kauppi, J.-P., Jääskeläinen, I. P., Sams, M., and Tohka, J. (2010b). Inter-subject correlation of brain hemodynamic responses during watching a movie: localization in space and frequency. *Front. Neuroinform.* 4:5. doi: 10.3389/fninf.2010.00005

Kauppi, J.-P., Huttunen, H., Korkala, H., Jääskeläinen, I. P., Sams, M., and Tohka, J. (2011). "Face prediction from fMRI data during movie stimulus: strategies for feature selection," in *Artificial Neural Networks and Machine Learning–ICANN 2011,* eds T. Honkela, W. Duch, M. Girolami, and S. Kaski (Espoo: Springer), 189–196. doi: 10.1007/978-3-642-21738-8_25

Krishnamoorthy, K., and Xia, Y. (2007). Inferences on correlation coefficients: one-sample, independent and correlated cases. *J. Stat. Plan. Infer.* 137, 2362–2379. doi: 10.1016/j.jspi.2006.08.002

Lahnakoski, J. M., Glerean, E., Salmi, J., Jääskeläinen, I. P., Sams, M., Hari, R., et al. (2012). Naturalistic fMRI mapping reveals superior temporal sulcus as the hub for the distributed brain network for social perception. *Front. Hum. Neurosci.* 6:233. doi: 10.3389/fnhum.2012.00233

Laird, A. R., Rogers, B. P., Carew, J. D., Arfanakis, K., Moritz, C. H., and Meyerand, M. E. (2002). Characterizing instantaneous phase relationships in whole-brain fMRI activation data. *Hum. Brain Mapp.* 16, 71–80. doi: 10.1002/hbm.10027

Lerner, Y., Honey, C. J., Silbert, L. J., and Hasson, U. (2011). Topographic mapping of a hierarchy of temporal receptive windows using a narrated story. *J. Neurosci.* 31, 2906–2915. doi: 10.1523/JNEUROSCI.3684-10.2011

Love, D. (2013). *Son of Grid Engine Project*. Available online at: https://arc.liv.ac.uk/trac/SGE (Visited 5.11.2013).

Mazziotta, J., Toga, A., Evans, A., Fox, P., Lancaster, J., Zilles, K., et al. (2001). A probabilistic atlas and reference system for the human brain: International consortium for brain mapping (ICBM). *Philos. Trans. R Soc. Lond. B Biol. Sci.* 356, 1293–1322. doi: 10.1098/rstb.2001.0915

Nichols, T. E., and Holmes, A. P. (2002). Nonparametric permutation tests for functional neuroimaging: a primer with examples. *Hum. Brain Mapp.* 15, 1–25. doi: 10.1002/hbm.1058

Nummenmaa, L., Glerean, E., Viinikainen, M., Jääskeläinen, I. P., Hari, R., and Sams, M. (2012). Emotions promote social interaction by synchronizing brain activity across individuals. *Proc. Natl. Acad. Sci. U.S.A.* 109, 9599–9604. doi: 10.1073/pnas.1206095109

Pajula, J., Kauppi, J.-P., and Tohka, J. (2012). Inter-subject correlation in fMRI: Method validation against stimulus-model based analysis. *PLoS ONE* 7:e41196. doi: 10.1371/journal.pone.0041196

Pikovsky, A., Rosenblum, M., and Kurths, J. (2000). Phase synchronization in regular and chaotic systems. *Int. J. Bifurcat. Chaos* 10, 2291–2305. doi: 10.1142/S0218127400001481

Raghunathan, T., Rosenthal, R., and Rubin, D. B. (1996). Comparing correlated but nonoverlapping correlations. *Psychol. Methods* 1:178. doi: 10.1037/1082-989X.1.2.178

Ruttimann, U. E., Unser, M., Rawlings, R. R., Rio, D., Ramsey, N. F., Mattay, V. S., et al. (1998). Statistical analysis of functional MRI data in the wavelet domain. *IEEE Trans. Med. Imaging* 17, 142–154. doi: 10.1109/42.700727

Scalable Logic. (2013). *Open Grid Scheduler/Grid Engine*. Available online at: http://gridscheduler.sourceforge.net/ (Visited 5.11.2013).

Schmälzle, R., Häcker, F., Renner, B., Honey, C., and Schupp, H. (2013). Neural correlates of risk perception during real-life risk communication. *J. Neurosci.* 33, 10340–10347. doi: 10.1523/JNEUROSCI.5323-12.2013

Spiers, H., and Maguire, E. (2007). Decoding human brain activity during real-world experiences. *Trends Cogn. Sci.* 11, 356–365. doi: 10.1016/j.tics.2007.06.002

The Mathworks Inc. (2013). *Overview of Memory-Mapping.* Matlab R2013b Documentation.

Univa Corporation. (2013). *Univa Grid Engine*. Available online at: http://www.univa.com/products/grid-engine.php (Visited 5.11.2013).

Vetterli, M., and Kovačević, J. (1995). *Wavelets and Subband Coding*, Vol. 87. New Jersey: Prentice Hall PTR Englewood Cliffs.

Vinck, M., Oostenveld, R., van Wingerden, M., Battaglia, F., and Pennartz, C. (2011). An improved index of phase-synchronization for electrophysiological data in the presence of volume-conduction, noise and sample-size bias. *Neuroimage* 55, 1548–1565. doi: 10.1016/j.neuroimage.2011.01.055

Vinck, M., van Wingerden, M., Womelsdorf, T., Fries, P., and Pennartz, C. (2010). The pairwise phase consistency: a bias-free measure of rhythmic neuronal synchronization. *Neuroimage* 51, 112–122. doi: 10.1016/j.neuroimage.2010.01.073

Wilson, S. M., Molnar-Szakacs, I., and Iacoboni, M. (2008). Beyond superior temporal cortex: intersubject correlations in narrative speech comprehension. *Cereb. Cortex* 18, 230–242. doi: 10.1093/cercor/bhm049

Worsley, K. (2008). *SurfStat. A Matlab Toolbox for the Statistical and Multivariate Surface and Volumetric Data Using Linear Mixed Effects Models and Random Field Theory*. Available online at: http://www.math.mcgill.ca/keith/surfstat/ (Visited 5.11.2013).

Yoo, A., Jette, M., and Grondona, M. (2003). "SLURM: simple linux utility for resource management," in *Job Scheduling Strategies for Parallel Processing*, *Lecture Notes in Computer Science*, Vol. 2862, eds D., Feitelson, L., Rudolph, and U., Schwiegelshohn (Berlin; Heidelberg: Springer), 44–60.

**Conflict of Interest Statement:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

*Received: 07 November 2013; accepted: 09 January 2014; published online: 31 January 2014.*

*Citation: Kauppi J-P, Pajula J and Tohka J (2014) A versatile software package for inter-subject correlation based analyses of fMRI. Front. Neuroinform.* ***8****:2. doi: 10.3389/fninf.2014.00002*

*This article was submitted to the journal Frontiers in Neuroinformatics.*

# Publication II

Pajula J., Kauppi J-P., Tohka J. "Inter-Subject Correlation in fMRI: Method Validation against Stimulus-Model Based Analysis", *PlosONE*, volume 7, number 8, page e41196, 2012.

OPEN ACCESS Freely available online

PLoS one

# Inter-Subject Correlation in fMRI: Method Validation against Stimulus-Model Based Analysis

**Juha Pajula[1]*, Jukka-Pekka Kauppi[2], Jussi Tohka[3]**

**1** Department of Signal Processing, Tampere University of Technology, Tampere, Finland, **2** Department of Computer Science, University of Helsinki, Helsinki, Finland, **3** Department of Signal Processing, Tampere University of Technology, Tampere, Finland

## Abstract

Within functional magnetic resonance imaging (fMRI), the use of the traditional general linear model (GLM) based analysis methods is often restricted to strictly controlled research setups requiring a parametric activation model. Instead, Inter-Subject Correlation (ISC) method is based on voxel-wise correlation between the time series of the subjects, which makes it completely non-parametric and thus suitable for naturalistic stimulus paradigms such as movie watching. In this study, we compared an ISC based analysis results with those of a GLM based in five distinct controlled research setups. We used International Consortium for Brain Mapping functional reference battery (FRB) fMRI data available from the Laboratory of Neuro Imaging image data archive. The selected data included measurements from 37 right-handed subjects, who all had performed the same five tasks from FRB. The GLM was expected to locate activations accurately in FRB data and thus provide good grounds for investigating relationship between ISC and stimulus induced fMRI activation. The statistical maps of ISC and GLM were compared with two measures. The first measure was the Pearson's correlation between the non-thresholded ISC test-statistics and absolute values of the GLM Z-statistics. The average correlation value over five tasks was 0.74. The second was the Dice index between the activation regions of the methods. The average Dice value over the tasks and three threshold levels was 0.73. The results of this study indicated how the data driven ISC analysis found the same foci as the model-based GLM analysis. The agreement of the results is highly interesting, because ISC is applicable in situations where GLM is not suitable, for example, when analyzing data from a naturalistic stimuli experiment.

**Citation:** Pajula J, Kauppi J-P, Tohka J (2012) Inter-Subject Correlation in fMRI: Method Validation against Stimulus-Model Based Analysis. PLoS ONE 8(8): e41196. doi:10.1371/journal.pone.0041196

**Editor:** Ben J. Harrison, The University of Melbourne, Australia

**Received** February 6, 2012; **Accepted** June 18, 2012; **Published** August 8, 2012



**Funding:** This research was supported by the Academy of Finland (application number 129657, Finnish Programme for Centres of Excellence in Research, 2006–2011 and grant number 130275). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

**Competing Interests:** The authors have declared that no competing interests exist.

* E-mail: juha.pajula@tut.fi

## Introduction

Inter-subject correlation (ISC) analysis method provides an opportunity for the functional magnetic resonance imaging (fMRI) analysis under naturalistic research paradigms. In these paradigms, the stimuli are designed to be closer to normal everyday life than in conventional research paradigms. The used stimuli can be, for example, a movie or a 3D video game [1].

One of the major benefits of the ISC analysis is that it can be used to locate activations without *a priori* knowledge of the temporal composition of processes contributing to the neuronal activation. In the ISC analysis, the hemodynamic activity of a subject is used to quantify the hemodynamic activity of another subject by calculating the correlation coefficient between the corresponding fMRI time series of the subjects. Inferences about the locations of activations are solely based on the similarities in hemodynamic responses across the subjects. Instead, a massively univariate stimulus-model-based analysis in fMRI predominantly relies on the theory of general linear models that provide a framework of analyzing subjects fMRI responses with respect to the model of the known and fixed stimulus type, typically appearing as the columns of the design (or predictor) matrix in the GLM. This often restricts the application of these GLM-based analyses to strictly controlled research setups as the parametric model for the BOLD signal changes related to the activation have to be defined *a priori*. The major difference between ISC and GLM based analyses is that the former is completely non-parametric in the sense it does not require any parametric form for the stimulus time-course while the latter requires a model for the stimulus time course. We note that there is a direct connection between the statistical analysis of a slope parameter in a simple regression, i.e., a simplified version of a single subject GLM-based analysis and a correlation coefficient. In what follows, we will use the terms ISC and GLM analysis rather loosely, referring to the major difference explained above rather than to the technical details of computations and statistics involved.

Hasson et al. [2] introduced the concept of ISC in fMRI and demonstrated that a simple movie stimulus produced significant correlations between the voxel-wise fMRI time series of the subjects, especially in visual and auditory cortices. Since then ISC analysis has been applied to investigate speech comprehension [3], auditory abnormalities [4], memory encoding [5] and brain functions during movie watching [2,6–8]. In a particular relation to this work, Kauppi et al. [9] developed a new ISC based method by adding an option to compute the frequency specific ISC and designed novel non-parametric resampling tests to make inferences about ISCs. Resampling tests were designed, since the data was not guaranteed to be uncorrelated as Heijnar et al. [4] had earlier noted. Significant ISCs were found in visual and auditory areas in

line with earlier neurocinematics studies and additionally in prefrontal cortical areas when studying low frequency bands.

One of the main questions concerning the ISC analysis is how to interpret correlations between subjects. Because the ISC measures the similarity of subjects' Blood Oxygenation Level Dependent (BOLD) fMRI responses during the same stimulus, a high ISC does not directly imply a high degree of task (or stimulus) related activation [4]. However, it has been shown by comparing intracranial single-unit and local field potential recordings of epilepsy patients and fMRI of healthy subjects experiencing the same movie stimulus that the correlated firing rate in a local population of neurons correlates with the BOLD response ([10], [11]). Further, Hanson et al. [12] argued that if there are correlations between individual subjects, who all experience the same stimuli, most of the correlated activations should be caused by the stimuli and so it might be possible to find the activity patterns of the brain even in complex situations.

A parametric GLM-based analysis is a standard method for detecting task-related activations in fMRI. Therefore, a potential way to investigate whether an ISC analysis method can locate activated brain regions due to stimulus presentation is to compare the results of the ISC analysis with those of the GLM analysis for the same fMRI data.

In this case, the data must be acquired under strictly controlled experimental setting so that the GLM analysis can be performed reliably. Previously, Heijnar et al. [4] studied ISCs of 20 subjects with the fMRI data acquired during the auditory oddball task and compared the results with those of the GLM. Multi-subject ISC maps were thresholded empirically, as it was noted that statistical thresholds cannot be obtained using standard statistical approaches due to dependencies between the correlations. The comparison was limited to the visual analysis of the activation maps. The conclusion was that the ISC analysis could find the same activation foci as GLM but ISC also found foci which were not visible in the model-based results.

Also in this work, we compare the ISC analysis results with those of the model-based GLM method to investigate the accuracy of the non model-based ISC analysis method detecting activated brain regions. We considerably extend the study of Hejnar et al. by incorporating more tasks and subjects to the comparative analysis. Moreover, we evaluate the similarity of the analysis results quantitatively and use a resampling-based method to obtain statistical thresholds for the ISC brain maps. It is important to use automatic thresholding scheme instead of a manual threshold selection to avoid a possible user-dependent bias in the comparison.

We use the GLM as a reference method in the comparison since it is a standard data analysis tool for locating brain activations in fMRI. The key difference between ISC and GLM methods is presented in Figure 1. ISC analysis combines voxel-wise correlations between several subject pairs in a fully non-parametric way to a single multi-subject statistical measure. Instead, GLM first compares voxel-wise the fMRI time series of each individual with a predefined model of the hemodynamic activity and then combines the results to a single multi-subject statistic. It is obvious that unlike the ISC method, where the model is not needed, GLM is not easily applicable to analyzing fMRI datasets acquired under complex stimuli for which the construction of the parametric model is far too difficult. Thus, it is necessary to use fMRI datasets which are acquired under strictly controlled experimental settings in order to carry out reliable validation, where the parametric model is guaranteed to succeed extremely well and this way provide the ground-truth for the non-parametric study.

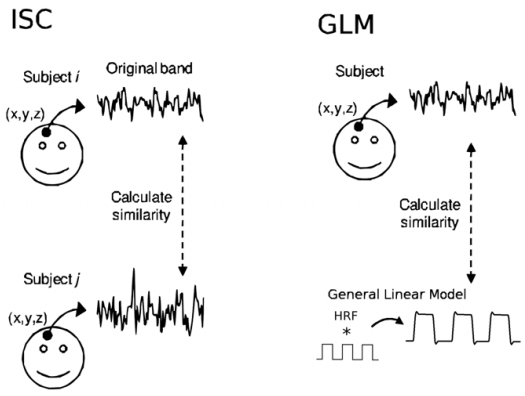



**Figure 1. General conceptual difference between the non-parametric ISC and parametric GLM analysis.** The ISCs are computed voxel-wise over the measured time series of every possible subject pair and then the results are combined to a single statistic. The GLM analysis fits the mathematical model (Here: boxcar function convolved with the canonical hemodynamic response function (HRF)) to the measured time-series of every subject and the group level results are then combined from the results of the individual subjects' analyses.
doi:10.1371/journal.pone.0041196.g001

## Materials and Methods

### ICBM functional reference battery data

For this study, we used fMRI data from the measurements with Functional Reference Battery tasks developed by the International Consortium for Human Brain Mapping (ICBM) [13] (http://www.loni.ucla.edu/ICBM/Downloads/Downloads_FRB.shtml) The data was obtained from ICBM database in the Image Data Archieve (IDA) of the Laboratory of Neuro Imaging (LONI) (http://www.loni.ucla.edu/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 co-investigators from UCLA, Montreal Neurologic Institute, University of Texas at San Antonio, and the Institute of Medicine, Juelich/Heinrich Heine University - Germany.

The selected data included measurements from 37 healthy right-handed subjects (19 men and 18 women; average age was 28.2 years from the range of 20–36 years), who had all performed the five selected tasks from FRB. 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.

Similarly to Bellec et al. [14], we restricted the age range of the subjects to 20–38 years. In the database, this resulted to 41 right-handed subjects who had fMRI measurements from all five different FRB tasks: auditory naming (AN), external ordering (EO), hand imitation (HA), oculomotor (OM) and verbal generation (VG). The image data was pre-screened before analysis to ensure high quality of the data. According to pre-screening, fMRI data from four subjects were discarded because of a poor data quality for at least one task in the battery.

**FRB tasks.** The detailed task definitions of the functional reference battery are included in the FRB software package and

they are next explained briefly here. All the five FRB task designs had the same block-structure in their implementations and they consisted of consecutive 'off' and 'on' blocks. There were 12 blocks per run (6 'off-on') and 3 volumes at the beginning of the run to wait for magnetisation stabilisation. The blocks lasted 28 s so that 'off-on' phases lasted totally 56 s. This created finally 5 min 48 s duration for the whole experiment where there were 12 blocks (six 'off' and six 'on' blocks) for each run with 7 volumes in each block.

In every task, the 'off' block instruction was the same: the subjects had to respond with the left mouse button press every time they saw an arrow pointing to the left. The different 'on' blocks were defined separately for each task.

In the first 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. The stimulus had first 2 s of silence, then 1.5 s of description and finally again 2 s of silence. This is a language task with an auditory input modality and the FRB definition noted that auditory cortex should be activated here (in addition to language areas).

In the EO task, which is a working memory 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. The designs were visible for 450 ms and the screen was blank 50 ms between the designs. The subjects responded via a button press whether the final stimulus was among the four previously shown. This test was repeated five times during each 'on' block.

In the HA task subjects were instructed to imitate the presented hand configuration with their right hand. The example hand configurations were presented to them with pictures on the screen. Each hand position was presented for 3.5 s. This is a task requiring higher order motor coordination and motor planning and in the FRB description, it was noted that this task should activate the frontal and parietal areas.

In the OM task subjects were watching 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 should have returned their gaze immediately to the central cross. In each 'on' block there were 20 fixation trials and 20 target trials. There were four fixations of each of the following durations: 800 ms, 1000 ms, 1200 ms, 1400 ms, and 1600. These were randomized and each were followed by a 200 ms target trial. This way the task was supposed to activate the visual system and the occipital lobe.

Finally, 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. During the 'on' blocks, line drawings were presented for 0.5 s. This task is a language task with visual input and was noted to activate the language and visual areas.

**Pre-processing.** Pre-processing and the GLM part of statistical analysis were performed by using the program FSL (version 4.1.6) [Oxford Centre for Functional Magnetic Resonance Imaging of the Brain (FMRIB), Oxford University, Oxford, U.K.] [15]. The data processing in FEAT (version 5.98) was done in three phases. First, motion correction was performed using the FSL's MCFLIRT by maximizing the correlation ratio between each time point and the middle volume, using linear interpolation [16,17]. Second, the Brain extraction tool (BET) [18] was applied to to extract the brain volume from functional data. Finally, the images were temporally high-pass filtered with a cutoff period of 60 s and the spatial smoothing was applied with a Gaussian kernel with full width at half maximum (FWHM) of 5 mm. The original data had 87 volumes with three stabilization volumes, which were discarded from the analysis. The brain extraction from the anatomical T1 images was also performed by BET, but this was done manually for each T1-weighted image separately from the FEAT procedure as the parameters of BET required individual tuning.

The image registration was performed in two phases using FSL Linear Registration Tool (FLIRT) [16,17]. First, the skull-stripped functional images were aligned (6 degrees of freedom, full search) to the skull-stripped high-resolution T1-weighted image of the same subject, and then the results were aligned to the standard (brain only) ICBM-152 template (12 degrees of freedom, full search).

## Analysis Methods

**General Linear Model with FEAT.** After preprocessing, the GLM was performed at the single subject level with the FSL (FEAT, fMRI Expert Analysis Tool) [19,20]. Most of the processing options were chosen according to the defaults of FEAT. The model was defined for 84 volumes where each block had the length of seven volumes. The length of the block in volumes was computed from the timing of the tasks and the scanning parameters (28 s divided by 4 s). The boxcar model was designed with the three-column format of FEAT. In this format it was possible to define separately for every block the current value of the model (one for each 'on' block), starting point in time from the beginning of experiment, and duration of the current block from the starting point. Then, the model was convolved with the canonical hemodynamic response function (HRF) (a single $\gamma$-function modeling: phase 0 s, standard deviation 3 s, mean lag 6 s) along to its temporal derivative. Finally, the same default high pass filtering as applied to experimental data (with a cutoff of 60 s) was applied to the model. The analysis itself was performed with the FILM prewhitening procedure [21].

Higher-level mixed effects group analyses were performed for each contrast by using FSL's FLAME (FMRIB's Local Analysis of Mixed Effects) module with two stages (1+2), where the second stage estimation was performed using MH MCMC (Metropolis-Hastings Markov Chain Monte Carlo) sampling [19]. Voxel-wise False Discovery Rate (FDR) based multiple comparison correction [22,23] under the independence or positive dependence assumption was used to threshold the z-statistic volumes. As argued in [9], the FDR based multiple comparison correction is a natural option for ISC and for this reason also the GLM thresholds were corrected with the FDR method. The used thresholding levels were $q=0.05$, $q=0.005$, $q=0.001$ and the FDR corrected GLM thresholds are presented in the Table 1 for reference.

**Inter-Subject Correlation Analysis.** The ISC analysis was performed using ISCtoolbox for Matlab by Kauppi et al. [9] (

**Table 1.** FDR corrected GLM thresholds for different tasks.

| | q = 0,05 | q = 0,005 | q = 0,001 |
|---|---|---|---|
| AN | 0.0025 | $0.1483 \cdot 10^{-3}$ | $0.2186 \cdot 10^{-4}$ |
| EO | 0.0063 | $0.4530 \cdot 10^{-3}$ | $0.7259 \cdot 10^{-4}$ |
| HA | 0.0049 | $0.3574 \cdot 10^{-3}$ | $0.5872 \cdot 10^{-4}$ |
| OM | 0.0040 | $0.2723 \cdot 10^{-3}$ | $0.4233 \cdot 10^{-4}$ |
| VG | 0.0039 | $0.2626 \cdot 10^{-3}$ | $0.4104 \cdot 10^{-4}$ |
| Average | 0.0043 | $0.2987 \cdot 10^{-3}$ | $0.4832 \cdot 10^{-4}$ |

doi:10.1371/journal.pone.0041196.t001

http://code.google.com/p/isc-toolbox/). This implementation can perform the ISC analysis over the specific frequency bands of the time series and threshold the results via voxel-wise resampling with the selected significance level. In this study, the analysis was performed only across the full frequency band.

In [9], the ISC is defined as a multi-subject similarity measure as follows. First, Pearson's correlation coefficient is calculated voxel-wise between every pair of subjects as:

$$r_{ij}=\frac{\sum_{n=1}^{N}[(s_i[n]-\bar{s}_i)(s_j[n]-\bar{s}_j)]}{\sqrt{\sum_{n=1}^{N}(s_i[n]-\bar{s}_i)^2\sum_{n=1}^{N}(s_j[n]-\bar{s}_j)^2}}, \quad (1)$$

where $r_{ij}$ is the sample correlation coefficient between the time series, N is the total number of samples in time series, $s_i$ and $s_j$ are time series obtained from the $i$th and $j$th subject, respectively, and $\bar{s}_i$ and $\bar{s}_j$ denote the means of $s_i$ and $s_j$.

To obtain the final multi-subject measure, the $r_{ij}$ values from all subject pairs were combined into a single ISC statistic by averaging:

$$\bar{r}=\frac{1}{\frac{m^2-m}{2}}\sum_{i=1}^{m}\sum_{j=2,j>i}^{m}r_{ij}, \quad (2)$$

where $m$ is the number of subjects. Since $m$ was 37 in our study, the correlation coefficients were averaged from $(37^2-37)/2=666$ subject pairs.

The statistical inference with this measure is complicated by the dependency of 666 correlation coefficients. To account for this problem a fully non-parametric voxel-wise resampling test is implemented in the ISC toolbox. This test accounts for temporal correlations inherent to fMRI data (for details of the test, see [9]). Similar to [9], we approximated resampling distribution with 1,000,000 realizations and corrected the resulting p-values using an FDR-based multiple comparison correction with independence or positive dependence assumption [22,23].

## Simulated Data

In order to obtain quantitative validation results against a known ground truth, we generated four sets of simulated imaging data with different noise levels mimicking the real data which was used in the study. Each set contained 37 simulated functional images in the standard ICBM-152 space. Each voxel in these images was either activated or not activated. Activation regions were selected according to the binarized GLM analysis results of AN task with the threshold level of $q=0.05$. A hemodynamic signal was included in the timeseries of the voxels in the activated regions. The signal was selected to be exactly the same which was used as a model in the GLM analysis, i.e., a boxcar convolved with a canonical HRF. Finally, pink 1 f noise generated as described in [24] (https://ccrma.stanford.edu/~jos/sasp/Example_Synthesis_1_F_Noise.html) was added to every timeseries in the volume. The power of the noise was 100, 200, 500 and 1000 times stronger than the power of the included hemodynamic signal resulting to signal to noise ratios (SNR) of 0.01, 0.005, 0.002 and 0.0001. The areas outside the activated regions contained only the noise signal. The simulation procedure was exactly the same for every 37 simulated images, that is, we ignored the anatomical and effect size variations between subjects.

The pink noise was chosen in the simulations due to empirical evidence that fMRI noise time-series contains 1 f-like noise [25]. As the data was generated directly in MNI-152 coordinates no registration or motion correction was needed for the simulated data and pre-processing included only temporal and spatial filtering which were performed exactly as described for FRB data.

## Method comparison

We compared the results of the ISC analysis and GLM with two performance measures. The first measure was suitable for comparing non-thresholded statistical images and was based on Pearson's correlation coefficient:

$$C=\frac{1}{K-1}\sum_{k=1}^{K}\left(\frac{|Z_k|-\bar{Z}}{s_Z}\right)\left(\frac{\bar{r}_k-\bar{R}}{s_{\bar{r}}}\right), \quad (3)$$

Here $K$ is the total number of brain voxels in the image ($K=228453$ voxels) and $Z_k$, $\bar{r}_k$ are the GLM and ISC statistics of the $k$th voxel, respectively. The absolute value of the $Z$ statistic was taken before computing the correlation measure because it was expected that both large negative and large positive $Z$-values relate to high $\bar{r}$ values. The $\bar{Z}$ and $\bar{R}$ are the corresponding sample means and $s_Z$, $s_{\bar{r}}$ the corresponding standard deviations ($\bar{Z}$ and $s_Z$ are computed from $\{|Z_k|\}_{k=1}^{K}$).

Our second performance measure was the Dice index [26] which was suitable for comparing thresholded and binarized GLM and ISC maps. The binarized maps were created by assigning the value of one to a voxel if the statistic value passed the threshold and otherwise assigning the value of zero to it. Let $B_Z$ denote the set of activated voxels of GLM and $B_{\bar{r}}$ the set of those of the ISC. The Dice index between two sets was defined as:

$$I_{Dice}=\frac{2|B_Z\cap B_{\bar{r}}|}{|B_Z|+|B_{\bar{r}}|}, \quad (4)$$

where the numerator measures the size of common activation occurrence and the denominator measures the sizes of activated areas according to individual methods. In other words, the equation measures the areas where both binaries are true against the areas where at least one binary is true. In practice the Dice index was computed from the binary vectors. The thresholded and binarized statistic volumes of the GLM and ISC analyses were vectorized by reshaping them to M-dimensional vectors. Then, the Dice index was computed as follows:

$$I_{Dice}=\frac{2\sum_{l=1}^{M}(B_Z[l]\cdot B_{\bar{r}}[l])}{\sum_{l=1}^{M}(B_Z[l])+\sum_{l=1}^{M}(B_{\bar{r}}[l])}, \quad (5)$$

where $B_Z[l]$ and $B_{\bar{r}}[l]$ are the $l$th voxels of binary vectors reshaped from binarized GLM and ISC statistic volumes, respectively. The sums were computed over the whole volumes ($M=91\times109\times91=902629$ voxels).

The resulting Dice index values vary between 0–1, where 1 denotes the exact similarity and 0 denotes no overlap. To further ease the interpretation of Dice indices, we can utilize the relationship between the Dice index and Kappa coefficient. Zijdenbos et al. showed that under certain assumptions [27], which are valid here, the Dice index is (asymptotically) equal to Kappa coefficient. According to Landis et al. [28] the Kappa coefficient values can be divided into six categories: less than 0,

"No agreement"; 0–0.2, "Slight agreement"; 0.2–0.4, "Fair agreement"; 0.4–0.6, "Moderate agreement"; 0.6–0.8, "Substantial agreement"; 0.8–1.0, "Almost perfect agreement". These categories are ad-hoc, but widely used. The relationship between Dice index and Kappa coefficient is further described by Finch [29]. Dice index was chosen instead of Kappa, because it is better suited to for our purposes since it ignores the non-activated regions (see [27] for more details) and it is widely used as the performance index in the evaluation of medical image segmentation algorithms.

## Results

Pearson's correlations, Eq. (3), between the absolute values of the $Z$-statistic of GLM and ISC are presented in Table 2. The values of the correlation coefficients were between 0.69 and 0.83, where the lowest correlation was from the task EO and the highest from the task HA. The average of the correlation coefficients across all of the tasks was 0.74. These values indicate a high similarity between the test statistics of GLM and ISC.

The Dice index, Eq. (4), between binary maps resulted in the average value of 0.73 across the tasks and the thresholds. The average Dice index values across the three thresholds for the specific tasks ranged from 0.65 to 0.81. The average over the tasks varied from 0.72 to 0.74 depending on the threshold. The results are presented in the Table 3. When comparing these with the Kappa categories discussed earlier, the similarity of the thresholded statistical maps of ISC and GLM had a moderate (0.4–0.6, 3 values), substantial (0.6–0.8, 9 values) or almost perfect (0.8–1.0, 3 values) agreement. Most of the Dice indices were at the level of substantial agreement. The Dice index values of the VG task were most stable across the thresholds (0.77, 0.81, 0.77), whereas the corresponding values of the EO task were most variable (0.76, 0.66, 0.56). With the tasks AN, HA, and OM, the values of the Dice indices with the two tightest threshold levels were close to each other but the values were notably lower with the most liberal level. These results indicated that the $q = 0.05$ level might be too liberal for this kind of study. The correlation and Dice index results are visualized together in Figure 2.

The Figure 3 presents all three threshold levels $q = 0.05$ (a), $q = 0.005$ (b), and $q = 0.001$ (c) of the AN task. The Figure 4 (a) presents a voxel-wise scatter plot between GLM (horisontal axis) and ISC values (vertical axis). Figure 4 (b) presents the corresponding histogram, which shows more clearly how the mass of the values is distributed with respect to the thresholds. The red lines in the Figure 4 denotes the three thresholds. The scatterplots and histograms of the other tasks are present in the Figures S2, S4, S6 and S8 of Supplement. The thresholded statistical maps of GLM and ISC with the threshold level $q = 0.001$ are presented in Figure 3 (c) for AN task and Figures 5 and 6 for EO and HA tasks. The threshold images from tasks EO and HA with threshold levels $q = 0.05$ and $q = 0.005$ are presented in Figures S1 and S3 of Supplement. Similarly to Figure 3, the Figures S5 and S7 of Supplement presents all three threshold levels for the tasks OM and VG respectively. In the figures, the red color indicates those voxels, which are activated according to both methods, the blue color indicates activated voxels according to GLM analysis only and the green color denotes activated voxels according to ISC analysis only. The images are in neurological orientation.

**Table 2.** Voxel-wise correlation measures, Eq. (3).

| TASK | AN | EO | HA | OM | VG | Average |
|---|---|---|---|---|---|---|
| C | 0,69 | 0,69 | 0,83 | 0,76 | 0,75 | 0.74 |

doi:10.1371/journal.pone.0041196.t002

With the AN task, both methods detected activations in auditory cortex, visual cortex, and cingulate gyrus (see Figure 3). This was as expected based on the FRB task definition and comparison to the previous fMRI studies with the AN task through a meta-analysis tool Pubbrain (http://www.pubbrain.org). With the EO task, the activations according to both methods were in lateral occipital cortex, inferior frontal gyrus, precentral gyrus and supplementary motor cortex (see Figure 5). As we expected, these results were highly similar to the detected activations of the healthy control subjects in the study of Hamilton et al. [30] which studied the same EO task as we were using here. With the HA task, there were activations in multiple parietal areas and inferior frontal gyrus and cingulate gyrus in the frontal lobe (see Figure 6). These were as expected (the FRB description noted that this task should activate at least frontal and parietal areas). With the HA task, ISC (but not GLM) detected activation in precuneous cortex. The activation remained visible even with the tightest threshold presented in Figure 6. Based on a review [31], it seems plausible that the precuneous is active during the hand imitation task. With the OM task, there were activations present at precentral gyrus, occipital pole, supplementary motor cortex and lateral occipital cortex (see Figure S5 of Supplement). These were as expected as the FRB description noted that the task should activate the visual system and the occipital lobe. With the VG task, activations at inferior temporal gyrus, inferior frontal gyrus, temporal occipital fusiform cortex, lingual gyrus, occipital pole, lateral occipital cortex and supplementary motor cortex were detected (see Figure S7 of Supplement). These were as expected as the FRB definition noted that the task should activate language and visual areas.

Two general trends were noticeable from the overlay images. First, with the EO (Figure 5) and VG tasks (Figure S7 of Supplement), the ISC analysis was generally more conservative than the GLM analysis for detecting activation areas, because the number of voxels detected only by GLM (blue) was high and common areas (red) were surrounded by these (blue) areas. Second, with the tasks AN, HA and OM, ISC tended to find more activated voxels than the GLM when the most liberal threshold ($q = 0.05$) was used. Thus, GLM analysis was more conservative of the two methods. However, the situation was reversed when the most tightest threshold ($q = 0.001$) was used, i.e., ISC analysis became more conservative than the GLM analysis. This is also visible in the Figure 7, which presents the voxels that were consistently detected as activated up by one method and not the other method for the AN task. Corresponding images for other tasks are presented in Figures S9, S10, S11 and S12 of Supplement.

The correlation measure was computed between $\bar{r}$-statistics and $|Z|$-statistics instead of signed Z-statistics. This was done because it was expected that both high negative (de-activations) and high positive (activations) $Z$-values relate to high positive $\bar{r}$ values. To validate this hypothesis, we computed the correlation between signed $Z$-values and $\bar{r}$-values. In that case, the correlation measures dropped to 0.50, 0.53, 0.74, 0.70 and 0.57 for AN, EO, HA, OM and VG tasks, respectively. By comparing these values to the values in Table 2, we can see that the decrease was larger with the low correlation tasks (0.19 (AN), 0.16(EO) and 0.18 (VG)) and smaller with high correlation tasks (0.10 (HA), 0.06(OM)).

With the simulated data, the Dice indices between the activations detected (either by ISC or GLM) and the ground truth are presented in the Figure 8 for different noise and thresholding levels. Average Dice index was 0.76 for ISC and 0.81

**Table 3.** Dice Indices, Eq. (4).

| Task/ Threshold | q = 0.05 | q = 0.005 | q = 0.001 | Average |
|---|---|---|---|---|
| AN | 0.56 | 0.69 | 0.7 | 0.65 |
| EO | 0.76 | 0.66 | 0.56 | 0.66 |
| HA | 0.71 | 0.86 | 0.86 | 0.81 |
| OM | 0.54 | 0.71 | 0.73 | 0.66 |
| VG | 0.77 | 0.81 | 0.77 | 0.78 |
| Average | 0.72 | 0.74 | 0.72 | 0.73 |

According to Landis et al. [28] the results can be categorized as following: less than 0, "No agreement"; 0–0.2, "Slight agreement"; 0.2–0.4, "Fair agreement"; 0.4–0.6, "Moderate agreement"; 0.6–0.8, "Substantial agreement"; 0.8–1.0, "Almost perfect agreement". By comparing the results with these categories the HA task can be nominated to have "Almost Perfect" agreement and the EO task, which had the lowest results as "Substantial agreement" even it also has values from "Moderate agreement" level.
doi:10.1371/journal.pone.0041196.t003

for GLM. The non-parametric ISC method detected simulated activations very accurately when the SNR was 0.002 or greater. Only with the highest noise level and especially with the most conservative thresholding level, the accuracy of ISC was poor (Dice index below 0.4) as it failed to detect the truly activated voxels. The lower Dice indices for GLM with the two lowest levels of noise were due to enlargening of the activation regions due to filtering. In other words, the GLM-based analysis was too sensitive in this highly idealized setting. Overall, we consider that the performance of the two methods was similar at the three lowest noise levels and only at the highest noise level the advantages of using stimulus model derived information as in GLM became clearly apparent.

## Discussion

We have compared activations detected by two different fMRI data analysis methods: a standard model-based GLM method and a non-parametric ISC method. The major difference between these two flavours of analyses is that the former requires a model for the stimulus time course while the latter is completely non-parametric in the sense it does not require any parametric form for the stimulus time-course. This means that the ISC can be used to analyze fMRI data acquired from the experiments of complex multi-dimensional stimuli, e.g., a movie. The used datasets were deliberately chosen so that they were optimized for the GLM type analysis to maximize the accuracy of the GLM analysis. The data was acquired from the ICBM research database, which contains fMRI acquisitions during highly standardized FRB stimuli. The data was pre-processed and separately analyzed with GLM (FSL) and ISC [9]. The Pearson's

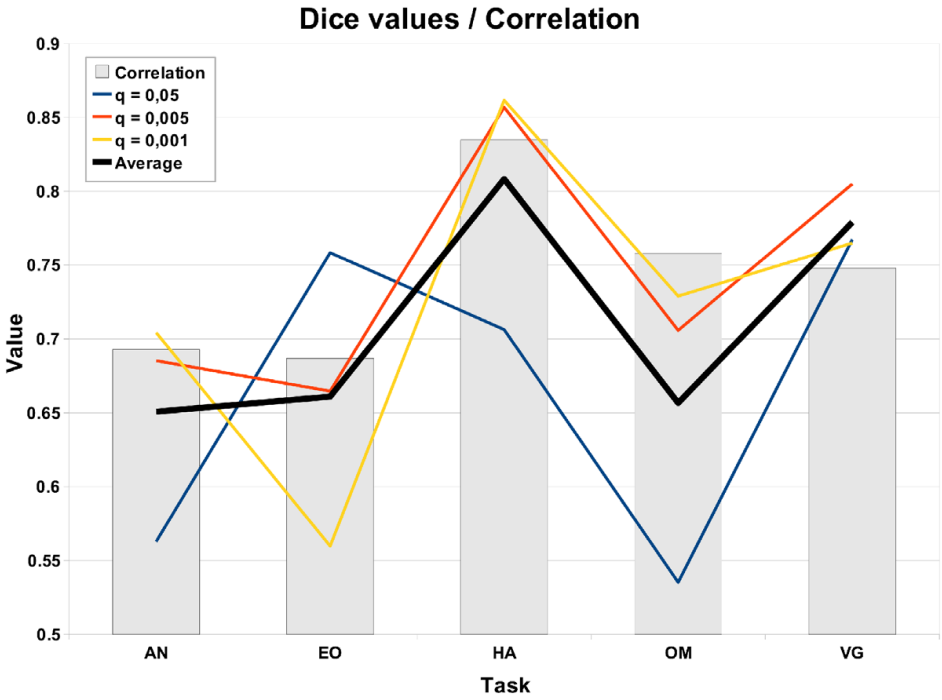
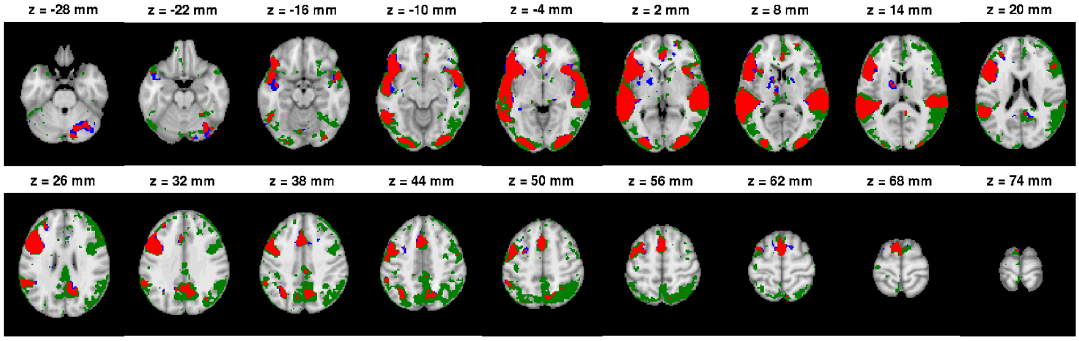
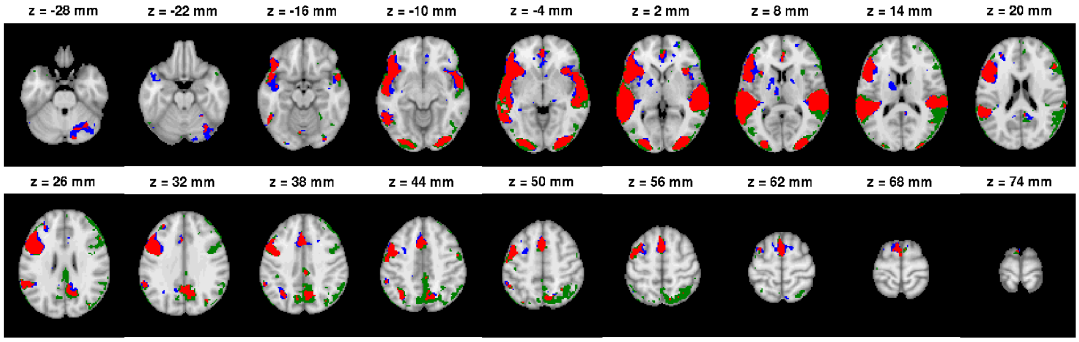


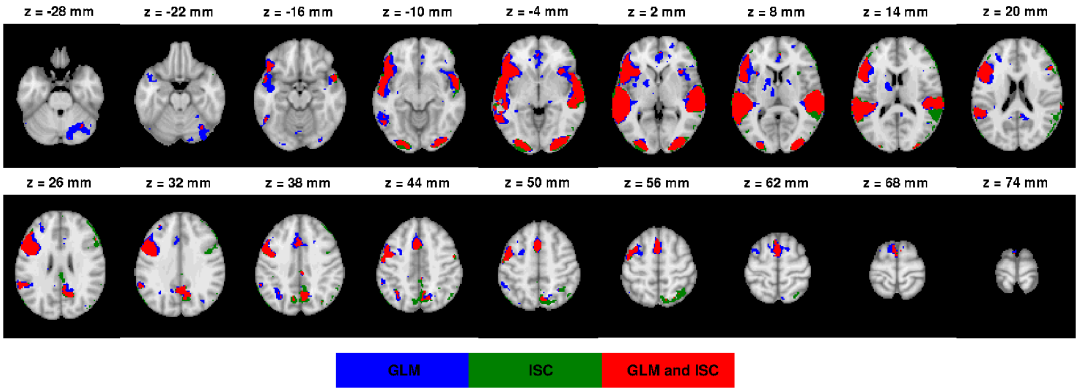

**Figure 2. The correlation measure and the Dice index.** The bars show the correlation measure between ISC and GLM and the lines present the Dice index values from different significance levels. The continuous black line presents the average over the Dice values within the current task. The HA task has higher correlation measure than other tasks and a high Dice index value. The EO task has the lowest correlation measure and the Dice index is also lower and varies the most with the thresholds. This suggests that a high correlation measure predicts a high Dice index value. We note that the values used as the basis for this figure are of higher numerical precision than those reported in Tables 2 and 3.
doi:10.1371/journal.pone.0041196.g002




(a) Auditory naming task, $q < 0.05$

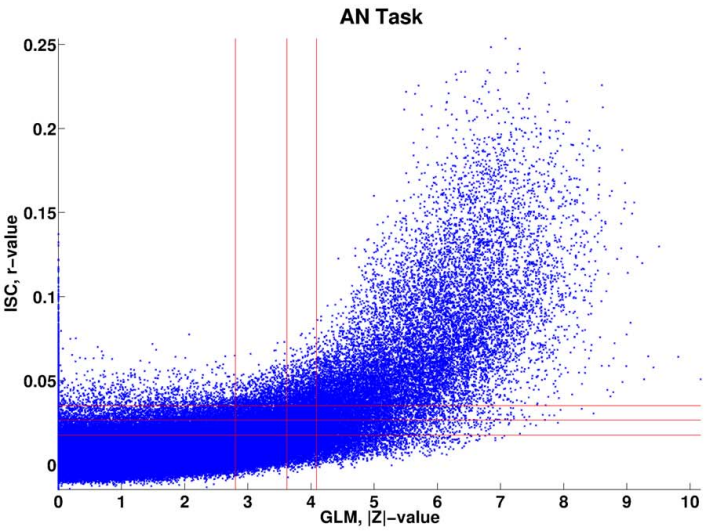



(b) Auditory naming task, $q < 0.005$

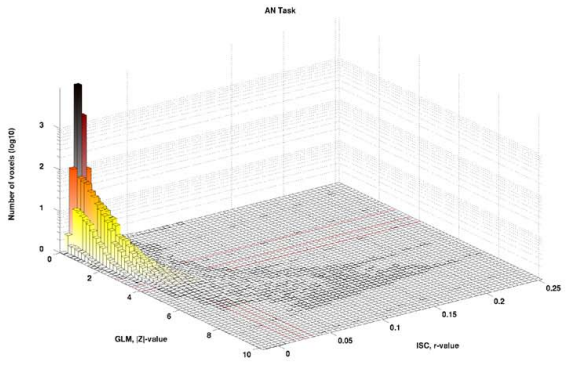



(c) Auditory naming task, $q < 0.001$

**Figure 3. GLM and ISC analysis results for the AN task (thresholded and FDR corrected, q = 0.05 (a), q = 0.005 (b), q = 0.001 (c) ).** In the images, the red color indicates voxels which are activated according to both ISC and GLM methods, blue indicates voxels activated according to GLM but not according to ISC and green indicates voxels activated according to ISC but not with GLM. The images are in neurological orientation. There is a notable correspondence between the ISC and GLM maps especially in auditory cortex, visual cortex, and cingulate gyrus. We can also see that the ISC analysis was clearly more liberal than the GLM analysis with a loose threshold (q = 0.05), but became more conservative when the thresholds became tighter (q = 0.005 and q = 0.001).
doi:10.1371/journal.pone.0041196.g003




(a) Scatterplot between ISC and GLM of the Auditory Naming task




(b) Histogram between ISC and GLM of the Auditory Naming task

**Figure 4. GLM and ISC analysis results for the AN task.** The scatterplot (a) presents the voxel-wise statistic values of GLM (horisontal axis) and ISC (vertical axis). Red lines define the thresholds with levels q = 0.05, q = 0.005 and q = 0.001. The second image (b) displays the corresponding histogram, which shows more clearly how the mass of the values is distributed with respect to the thresholds defined by the red lines. Most of the values are focused close to the origin which is not visible in the scatterplot.
doi:10.1371/journal.pone.0041196.g004

correlation was computed between corresponding statistics of ISC and GLM. The statistical maps from both methods were thresholded while accounting for the multiple comparisons based on FDR. The resulting binarized thresholded maps were compared by computing Dice index between them.

The correlations between GLM and ISC statistics validated the original assumption of the similarity of the measures used to quantify the activations. The average correlation value over all five tasks was 0.74, which can be considered as a high correlation. The average Dice-index over all five tasks varied between 0.72 and 0.74 depending on the task. As noted earlier, nine of the 15 Dice values were classified as substantial agreement (0.6–0.8) and three of the 15 as almost perfect agreement according to a widely used Landis and Koch categorization. Not surprisingly, the tasks with the highest Pearson's correlations featured the highest (and the most stable) Dice index values.

Accordingly, the activations detected by ISC matched well with the activations detected by GLM. The activation maps presented in Figure 3 and Figures 5 and 6 illustrate that ISC method was slightly more conservative than GLM method especially at the most conservative thresholding level q = 0.001 presented in the figures. The development is easiest to see from the Figure 3 where

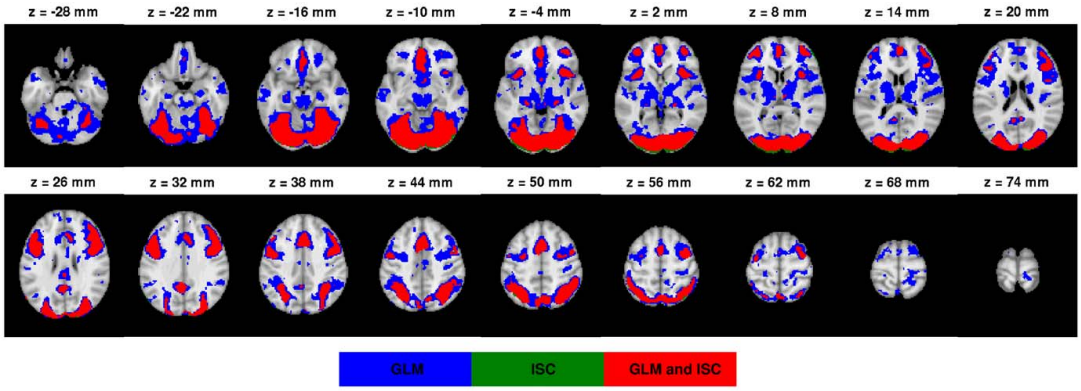



**Figure 5. GLM and ISC analysis results for the EO task.** In the image the thresholded (FDR corrected, q = 0.001) results for EO task are presented as a binary overlay image. The color coding in the image is the same as in Figure 3. The threshold images from the levels q = 0.05 and q = 0.005 are visible in the Figure S1 of the Supplement. Both methods find the same activation areas widely across the brain, including lateral occipital cortex, inferior frontal gyrus, precentral gyrus and supplementary motor cortex. Note also how ISC only (green) and commonly detected areas (red) are vanishing faster than GLM only areas (blue) when the threshold becomes more conservative. Thus, the ISC analysis was more conservative of the two methods especially with the lowest q-value. This tendency explains relatively high variation in the Dice index values with different significance levels for this particular task.
doi:10.1371/journal.pone.0041196.g005

all threshold levels are present (See also the Figures S5 and S7 of Supplement). In most of these cases, the area of common activation (in red) was surrounded by GLM only activation area (in blue) indicating that ISC had found the same overall activation location as GLM method. This result is promising from the fMRI data analysis point of view under naturalistic paradigms, because it suggests that the nonparametric ISC method can locate true sources of BOLD signal activity well and yet it is not susceptible to spurious findings, easily leading to overinterpretation of the results.

The variation in the correlation measure (range 0.69–0.83) and Dice index (range 0.54–0.86) could have resulted from the differences in the nature of the behavioral tasks. Especially, the EO task had lower correlation value and Dice index than other tasks, probably because it is the most complex task in FRB designed to activate working memory. Surprisingly, the Dice and correlation measures of the AN and the VG tasks were different although the tasks are similar.

The simulation study demonstrated that the ISC could in principle accurately detect activations even when the signal to noise ratio was as low as 0.002. The lower Dice index values of GLM than those of ISC with the simulated databases with low noise levels (SNR 0.01 and 0.005) could be largely attributed to the spatial smoothing applied to the data before analysis. (With higher noise levels, the leakage of the activation to the voxels surrounding the true activation region by smoothing became harder to detect and thus GLM detected more accurately true activation areas.) As

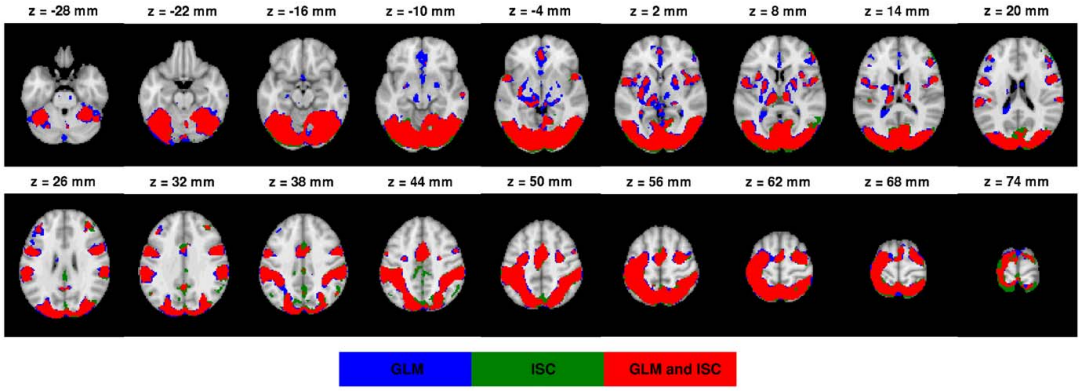



**Figure 6. GLM and ISC analysis results for the HA task.** In the image the thresholded (FDR corrected, q = 0.001) results for HA task are presented as a binary overlay image. The threshold images from the levels q = 0.05 and q = 0.005 are visible in the Figure S3 of the Supplement. The color coding in the image is the same as in Figure 3. Here it is clear that commonly detected areas (red) are dominant. There are also a notable number of ISC only detections (green), which might indicate that ISC can detect activations which are not detectable by GLM. On the other hand, some GLM only activations were located in cerebrospinal fluid, which suggested that there might exist measurement artifacts.
doi:10.1371/journal.pone.0041196.g006

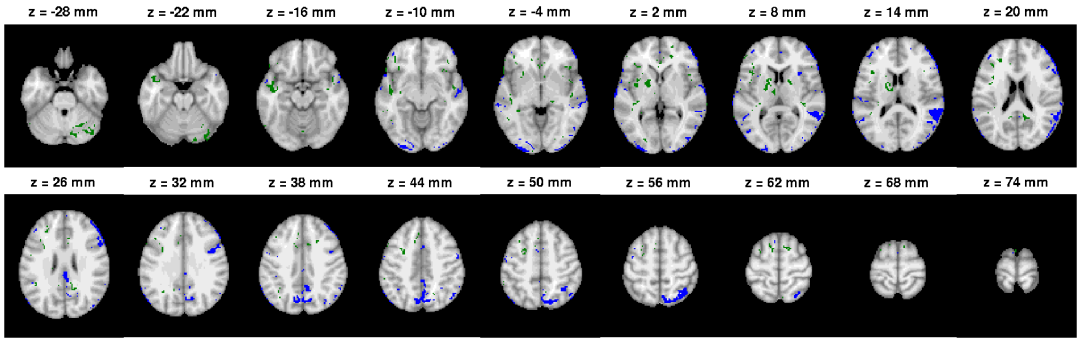



**Figure 7. The voxels consistently detected as activated by one method and not by the other with AN task.** Green color indicates voxels which were detected as activated by GLM in all thresholding levels, but not detected as activated by ISC in even the most liberal thresholding level ($q = 0.05$). Viceversa, blue color indicates voxels which were detected as activated by ISC in all of the thresholding levels, but not detected as activated by GLM with even the most liberal thresholding level ($q = 0.05$). Mostly these are isolated voxels or voxels lying near the boundary of the activation area. However, the ISC detected activations in Posterior and Anterior cingulate cortex and Precuneus as well as Occipital lobe that were not detected by the GLM. These areas are suspected to overlap with the default mode network in several studies, e.g., [34–36].
doi:10.1371/journal.pone.0041196.g007

the FWHM of the smoothing kernel was the same for both methods this indicates ISC was more conservative (or less sensitive) than GLM. This phenomenon was observed also with experimental data - albeit to a lesser extent. As the simulation model was idealized and greatly simplified ignoring all between-subject variability, the results with simulated data should be interpreted

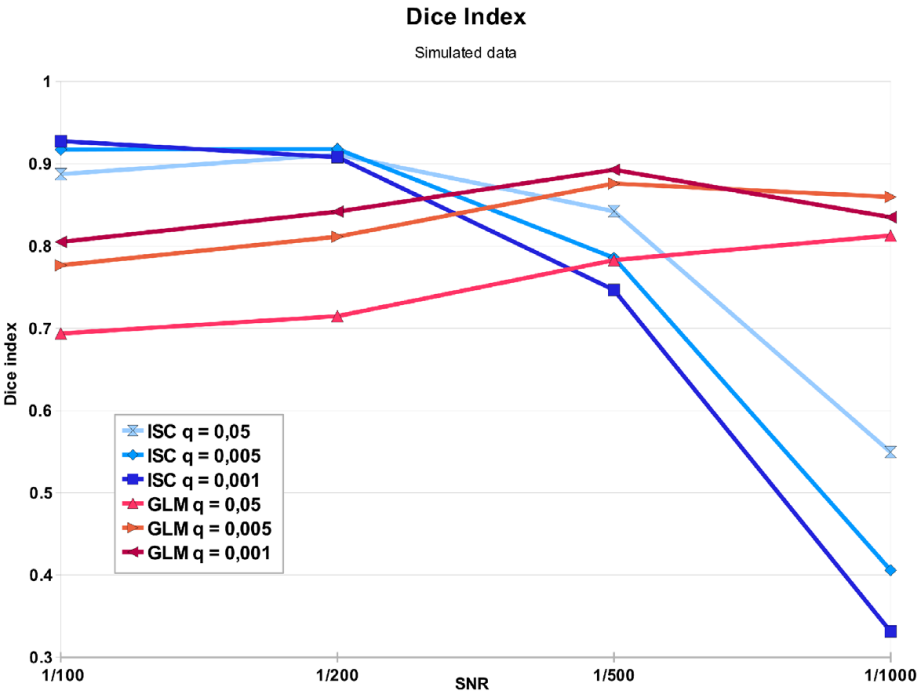



**Figure 8. Similarity of the detected activation region and ground truth activation region in the simulation study.** The lines present the Dice index values between the simulated versus detected activation area by ISC with different thresholding levels (blue lines) and by GLM with different thresholding levels (red lines). The ISC performed well with lower noise levels (SNR 1/100 and 1/200) but failed with the highest noise level (SNR 1/1000). The GLM performed overall well, but has a lower detection rate at low noise levels compared to ISC. This is due to false positive detections on the areas nearby ground-truth activation areas due to the effects of the spatial smoothing.
doi:10.1371/journal.pone.0041196.g008

with caution expecially regarding the exact noise levels that ISC could tolerate.

In this study, we used a relatively large database of 37 subjects. One interesting topic for future research would be to test comprehensively how the number of subjects affects the ISC analysis and what is the minimum number of subjects for the ISC analysis. Some results in this direction were presented by Hanson et al. [12] who demonstrated (but did not quantify) the stability of Roy's largest root statistic based ISC analysis after six or more subjects with a video stimulus of length of 156 s. However, for example, the reproducibility of ISC across subject samples remains an almost untouched research topic. Another slightly unusual aspect of the data is rather long TR of 4 seconds. It is difficult to speculate what effects this would have to the results of the method comparison as the two methods might react differently to the shortening of repetition time. However, it is important to note that recent ISC applications have typically used shorter TRs from 1.5 to 2 seconds.

Certain methodological choices warrant commenting. The GLM was used as the reference method because it is the standard method for analyzing fMRI studies acquired under a strictly controlled stimulus. The particular implementation of the multi-subject GLM (FSL's FLAME using MCMC) was selected because it is widely used and properly evaluated [15,20]. In particular, a computationally heavy MCMC approach was selected due to its accuracy [20]. Obviously, activations detected by GLM cannot be considered as ground truth and we therefore verified that our GLM analysis results to matched to the prior expectations based on fMRI literature. This was done by comparing our analysis results with the information available through a meta-analysis tool Pubbrain. In the GLM-based fMRI analysis, it is often recommendable and more typical to apply a family-wise error rate based multiple comparisons correction (either in voxel or cluster level) instead of a more liberal FDR-based criterion adopted by us (see [32] for a comparison of different multiple comparison options in fMRI). We adopted it, since FDR is a natural choice for ISC analysis and it is essential to compare detected activations at the same significance level. Indeed, as can be noted based on Figure 3, especially the FDR level $q=0.05$ was liberal (technically, we could expect 5% of the activated voxels to be false positives) and some of the activations were likely to be due to imaging artefacts. In visual inspection, both ISC and GLM seemed to detect activations that could be suspected to be artefactual at the most liberal thresholding level while at the most conservative thresholding level activations that could be easily labeled as artefactual were almost non-existent.

Obviously, there are also methodological choices related to the ISC analysis although the methodological literature about ISC is scarce compared to that of the model-based GLM analysis. The first choice is that of the test statistic, in this work given in Eq. 2. Alternatives to this statistic include average of Z-transformed correlation coefficients [3], Roy's largest root [12], and average correlation coefficient between subjects response time-course and an averaged response time course [33]. In the latter, the order of the averaging and normalization to unit variance is reversed compared to our test statistic leading to a different (but related) test-statistic. Our preference of the test statistic selected in this work relate to its easy interpretation in the simple case that the true correlation between all subjects' time series has an equal value (see [9]). However, we speculate that the choice of test statistic is not critical unless the number of subjects or time-points is much smaller than here and, in particular, the qualitative results of this work do not rely on a particular choice of test statistic. The second, we think more critical, choice is that of the thresholding procedure. The important question here is if the hypothesis testing relying on parametric models (e.g. [3]) could replace more computationally heavy resampling procedures (e.g. [33], [9], and this work). In this work, we have experimentally shown that a time-domain resampling test produces inference results comparable to model-based activation detection. Further work is required to identify the most optimal thresholding scheme.

An interesting detail can be observed by studying activations detected only by ISC colored in blue in Figure 7. These activations detected by solely by ISC included voxels from Posterior and Anterior cingulate cortex and Precuneus as well as Occipital lobe. Similar patterns of activations detected solely by ISC can also be found by inspecting the Figures S10 and S11 in the Supplement. These areas are suspected to overlap with the default mode network in several studies, e.g., [34–36]. In a wider scope, [8] suggested that that naturalistic stimulation may provide a complementary tool to the resting state protocol for studying the default mode network.

Both the ISC- and GLM-based statistics presented here focus on shared responses across subjects while allowing some intersubject variablity in the models via mixed effects modelling (GLM) or how the hypothesis testing is performed (ISC). This seems to be a reasonable assumption in the tasks presented here, but under other kind of experiments intersubject variability can be considerably higher and harder to model due to individual differences in information processing. The investigation of these differences requires the use of more sensitive methods which take better into account the variability across subjects. For instance, clustering approach presented in [37] preserves the entire structure of the intersubject correlation matrices, making it a suitable method for investigating differences and similarities in brain responses in data-driven manner even for a large group of subjects simultaneously. Another approach was presented in [38], where individual differences were investigated by comparing the results of group-level ISC analysis and intra-subject correlation analysis computed across repeated presentations.

Our results indicate that the ISC analysis can be used to find the same activation areas as the stimulus model-based GLM analysis when the parametric form of the stimulus is known. The motivation for this study is that ISC-based methods do not require the model of the stimulus time course and therefore they can be used in many research settings where the parametric modeling of the stimulus is not applicable. For example, movies provide an interesting form of a more naturalistic stimulus that is impossible to model completely and where the applicability of the parametric model based methods for activation detection is therefore limited.

## Supporting Information

**Figure S1 GLM and ISC analysis results for the EO task.** In the image the thresholded (FDR corrected, $q=0.05$ (a) and $q=0.005$ (b)) results for EO task are presented as a binary overlay image. The color coding in the images is the same as in Figure 3 of the article. The image of $q=0.001$ is presented in the Figure 5 of the article. Both methods find the same activation areas widely across the brain, including lateral occipital cortex, inferior frontal gyrus, precentral gyrus and supplementary motor cortex. Note also how ISC only (green) and commonly detected areas (red) are vanishing faster than GLM only areas (blue) when the threshold becomes more conservative. Thus, the ISC analysis was more conservative of the two methods especially with the lowest q-value. This tendency explains relatively high variation in the Dice

index values with different significance levels for this particular task.
(TIFF)

**Figure S2 GLM and ISC analysis results for the EO task.** The scatterplot (a) presents the voxel-wise statistic values of GLM (horisontal axis) and ISC (vertical axis). Red lines define the thresholds with levels $q = 0.05$, $q = 0.005$ and $q = 0.001$. The second image (b) displays the corresponding histogram, which shows more clearly how the mass of the values is distributed with respect to the thresholds defined by the red lines. Most of the values are focused close to the origin which is not visible in the scatterplot.
(TIFF)

**Figure S3 GLM and ISC analysis results for the HA task.** In the image the thresholded (FDR corrected, $q = 0.05$ (a) and $q = 0.005$ (b)) results for HA task are presented as a binary overlay image. The color coding in the images is the same as in Figure 3 of the article. The image of $q = 0.001$ is presented in the Figure 6 of the article. Here it is clear that commonly detected areas (red) are dominant. There are also a notable number of ISC only detections (green), which might indicate that ISC can detect activations which are not detectable by GLM. On the other hand, some GLM only activations were located in cerebrospinal fluid, which suggested that there might exist measurement artifacts.
(TIFF)

**Figure S4 GLM and ISC analysis results for the OM task.** The scatterplot (a) presents the voxel-wise statistic values of GLM (horisontal axis) and ISC (vertical axis). Red lines define the thresholds with levels $q = 0.05$, $q = 0.005$ and $q = 0.001$. The second image (b) displays the corresponding histogram, which shows more clearly how the mass of the values is distributed with respect to the thresholds defined by the red lines. Most of the values are focused close to the origin which is not visible in the scatterplot.
(TIFF)

**Figure S5 GLM and ISC analysis results for the OM task.** In the image the thresholded (FDR corrected, $q = 0.05$ (a), $q = 0.005$ (b) and $q = 0.001$ (c)) ) results for OM task are presented as a binary overlay image. The color coding in the images is the same as in Figure 3 of the article. As earlier with the HA task in Figure S3, also here ISC was first very liberal $q = 0.05$ and there was mainly common (red) and ISC only (green) areas. When the threshold gets tighter $q = 0.005$ the ISC only areas becomes smaller like with AN task and with the tightest threshold $q = 0.001$ ISC becomes more conservative than GLM. Here some ISC only areas remained visible even with the tightest significance level $q = 0.001$.
(TIFF)

**Figure S6 GLM and ISC analysis results for the OM task.** The scatterplot (a) presents the voxel-wise statistic values of GLM (horisontal axis) and ISC (vertical axis). Red lines define the thresholds with levels $q = 0.05$, $q = 0.005$ and $q = 0.001$. The second image (b) displays the corresponding histogram, which shows more clearly how the mass of the values is distributed with respect to the thresholds defined by the red lines. Most of the values are focused close to the origin which is not visible in the scatterplot.
(TIFF)

**Figure S7 GLM and ISC analysis results for the VG task.** In the image the thresholded (FDR corrected, $q = 0.05$ (a), $q = 0.005$ (b) and $q = 0.001$ (c)) results for VG task are presented as a binary overlay image. The color coding in the images is the same as in Figure 3 of the article. Here we can see the similar progress than with the task EO. There were merely a few ISC only areas (green) without GLM areas next to them and most of the common (red) areas were surrounded by GLM only areas (blue). When the threshold tightened from $q = 0.05$ to $q = 0.001$ both ISC and GLM detections contracted, but ISC contracted somewhat faster, which again suggested that ISC was more conservative than GLM.
(TIFF)

**Figure S8 GLM and ISC analysis results for the VG task.** The scatterplot (a) presents the voxel-wise statistic values of GLM (horisontal axis) and ISC (vertical axis). Red lines define the thresholds with levels $q = 0.05$, $q = 0.005$ and $q = 0.001$. The second image (b) displays the corresponding histogram, which shows more clearly how the mass of the values is distributed with respect to the thresholds defined by the red lines. Most of the values are focused close to the origin which is not visible in the scatterplot.
(TIFF)

**Figure S9 The voxels consistently detected as activated by one method and not by the other with EO task.** Green color indicates voxels which were detected as activated by GLM in all thresholding levels, but not detected as activated by ISC in even the most liberal thresholding level ($q = 0.05$). Viceversa, blue color indicates voxels which were detected as activated by ISC in all of the thresholding levels, but not detected as activated by GLM with even the most liberal thresholding level ($q = 0.05$). Mostly these are isolated voxels or voxels lying near the boundary of the activation area.
(TIFF)

**Figure S10 The voxels consistently detected as activated by one method and not by the other with HA task.** The color coding of the image is the same as in Figure S9. Mostly these are isolated voxels or voxels lying near the boundary of the activation area. However, the ISC detected activations in Precuneous cortex that were not detected by the GLM.
(TIFF)

**Figure S11 The voxels consistently detected as activated by one method and not by the other with OM task.** The color coding of the image is the same as in Figure S9. Mostly these are isolated voxels or voxels lying near the boundary of the activation area. However, the ISC detected activations in middle frontal gyrus that were not detected by the GLM.
(TIFF)

**Figure S12 The voxels consistently detected as activated by one method and not by the other with VG task.** The color coding of the image is the same as in Figure S9. Mostly these are isolated voxels or voxels lying near the boundary of the activation area. However, the ISC detected activations in middle temporal cortex and in superior cortex that were not detected by the GLM.
(TIFF)

## Acknowledgments

Data collection and sharing for this project was 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 California, Los Angeles.

## Author Contributions

Conceived and designed the experiments: JP JPK JT. Performed the experiments: JP. Analyzed the data: JP JPK JT. Wrote the paper: JP JPK JT.

## References


1. Spiers HJ, Maguire EA (2007) Decoding human brain activity during real-world experiences. Trends in cognitive sciences 11: 356–365.
2. Hasson U, Nir Y, Levy I, Fuhrmann G, Malach R (2004) Intersubject synchronization of cortical activity during natural vision. Science 303: 1634–1640.
3. Wilson SM, Molnar-Szakacs I, Iacoboni M (2008) Beyond superior temporal cortex: Intersubject correlations in narrative speech comprehension. Cerebral Cortex 18: 230–242.
4. Hejnar MP, Kiehl KA, Calhoun VD (2007) Interparticipant correlations: A model free fmri analysis technique. Human brain mapping 28: 860–867.
5. Hasson U, Furman O, Clark D, Dudai Y, Davachi L (2008) Enhanced intersubject correlations during movie viewing correlate with successful episodic encoding. Neuron 57: 452–462.
6. Jääskeläinen IP, Koskentalo K, Balk MH, Autti T, Kauramäki J, et al. (2008) Inter-subject synchronization of prefrontal cortex hemodynamic activity during natural viewing. Open Neuroimag J 2: 14–19.
7. Golland Y, Bentin S, Gelbard H, Benjamini Y, Heller R, et al. (2007) Extrinsic and intrinsic systems in the posterior cortex of the human brain revealed during natural sensory stimulation. Cerebral Cortex 17: 766–777.
8. Hasson U, Malach R, Heeger DJ (2010) Reliability of cortical activity during natural stimulation. Trends in cognitive sciences 14: 40–48.
9. Kauppi JP, Jääskeläinen IP, Sams M, Tohka J (2010) Inter-subject correlation of brain hemodynamic responses during watching a movie: localization in space and frequency. Frontiers in Neuroinformatics 4: 5.
10. Mukamel R, Gelbard H, Arieli A, Hasson U, Fried I, et al. (2005) Coupling between neuronal firing, field potentials, and fmri in human auditory cortex. Science 309: 951–954.
11. Nir Y, Fisch L, Mukamel R, Gelbard-Sagiv H, Arieli A, et al. (2007) Coupling between neuronal firing rate, gamma lfp, and bold fmri is related to interneuronal correlations. Current Biology 17: 1275–1285.
12. Hanson SJ, Gagliardi AD, Hanson C (2009) Solving the brain synchrony eigenvalue problem: Conservation of temporal dynamics (fmri) over subjects doing the same task. Journal of computational neuroscience 27: 103–114.
13. Mazziotta J, Toga A, Evans A, Fox P, Lancaster J, et al. (2001) A probabilistic atlas and reference system for the human brain: International consortium for brain mapping (icbm). Philosophical Transactions of the Royal Society of LondonSeries B: Biological Sciences 356: 1293–1322.
14. Bellec P, Perlbarg V, Evans AC (2009) Bootstrap generation and evaluation of an fmri simulation database. Magnetic resonance imaging 27: 1382–1396.
15. Smith SM, Jenkinson M, Woolrich MW, Beckmann CF, Behrens TEJ, et al. (2004) Advances in functional and structural mr image analysis and implementation as fsl. Neuro Image 23, Supplement 1: 208–219.
16. Jenkinson M, Smith S (2001) A global optimisation method for robust affine registration of brain images. Medical image analysis 5: 143–156.
17. Jenkinson M, Bannister P, Brady M, Smith S (2002) Improved optimization for the robust and accurate linear registration and motion correction of brain images. Neuro Image 17: 825–841.
18. Smith SM (2002) Fast robust automated brain extraction. Human brain mapping 17: 143–155.
19. Beckmann CF, Jenkinson M, Smith SM(2003) General multilevel linear modeling for group analysis in fmri. Neuro Image 20: 1052–1063.
20. Woolrich MW, Behrens TE, Beckmann CF, Jenkinson M, Smith SM (2004) Multilevel linear modelling for fmri group analysis using bayesian inference. Neuro Image 21: 1732–1747.
21. Woolrich MW, Ripley BD, Brady M, Smith SM (2001) Temporal autocorrelation in univariate linear modeling of fmri data. Neuro Image 14: 1370–1386.
22. 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: 289–300.
23. Genovese CR, Lazar NA, Nichols T (2002) Thresholding of statistical maps in functional neuroimaging using the false discovery rate. Neuro Image 15: 870–878.
24. Smith J (Accessed 02.05.2012) Spectral Audio Signal Processing. http://ccrma.stanford.edu/~jos/sasp/. Online book.
25. Zarahn E, Aguirre GK, D'Esposito M (1997) Empirical analyses of bold fmri statistics. i. spatially unsmoothed data collected under null-hypothesis conditions. Neuro Image 5: 179–197.
26. Dice LR (1945) Measures of the amount of ecologic association between species. Ecology 26: 297–302.
27. Zijdenbos AP, Dawant BM, Margolin RA, Palmer AC (1994) Morphometric analysis of white matter lesions in mr images: method and validation. Medical Imaging, IEEE Transactions on 13: 716–724.
28. Landis JR, Koch GG (1977) The measurement of observer agreement for categorical data. Biometrics 33: 159–174.
29. Finch H (2005) Comparison of distance measures in cluster analysis with dichotomous data. Journal of Data Science 3: 85–100.
30. Hamilton LS, Altshuler LL, Townsend J, Bookheimer SY, Phillips OR, et al. (2009) Alterations in functional activation in euthymic bipolar disorder and schizophrenia during a working memory task. Human brain mapping 30: 3958–3969.
31. Cavanna AE, Trimble MR (March 2006) The precuneus: a review of its functional anatomy and behavioural correlates. Brain 129: 564–583.
32. Logan BR, Rowe DB (2004) An evaluation of thresholding techniques in fmri analysis. Neuro Image 22: 95–108.
33. Lerner Y, Honey CJ, Silbert LJ, Hasson U (2011) Topographic mapping of a hierarchy of temporal receptive windows using a narrated story. The Journal of Neuroscience 31: 2906–2915.
34. Raichle ME, MacLeod AM, Snyder AZ, Powers WJ, Gusnard DA, et al. (2001) A default mode of brain function. Proceedings of the National Academy of Sciences 98: 676–682.
35. Greicius MD, Srivastava G, Reiss AL, Menon V (2004) Default-mode network activity distinguishes alzheimer's disease from healthy aging: Evidence from functional mri. Proceedings of the National Academy of Sciences of the United States of America 101: 4637–4642.
36. Calhoun VD, Pearlson GD (2012) A selective review of simulated driving studies: Combining naturalistic and hybrid paradigms, analysis approaches, and future directions. Neuro Image 59: 25–35.
37. Kauppi J, Jaaskelainen I, Sams M, Tohka J (2010) Clustering inter-subject correlation matrices in functional magnetic resonance imaging. In: Information Technology and Applications in Biomedicine (ITAB), 2010 10th IEEE International Conference on. pp. 1–6. doi:10.1109/ITAB. 2010.5687650.
38. Hasson U, Avidan G, Gelbard H, Vallines I, Harel M, et al. (2009) Shared and idiosyncratic cortical activation patterns in autism revealed under continuous real-life viewing conditions. Autism Research 2: 220–231.

# Publication III

Pajula J., Tohka J., "Effects of spatial smoothing on inter-subject correlation based analysis of FMRI", *Magnetic Resonance Imaging*, volume 32, number 9, pages 1114-1124, 2014.

Magnetic Resonance Imaging 32 (2014) 1114–1124

Contents lists available at ScienceDirect

Magnetic Resonance Imaging

journal homepage: www.mrijournal.com

# Effects of spatial smoothing on inter-subject correlation based analysis of FMRI

Juha Pajula *, Jussi Tohka *

*Department of Signal Processing, Tampere University of Technology, Tampere, Finland*

ARTICLE INFO

*Article history:*
Received 28 August 2013
Revised 6 April 2014
Accepted 15 June 2014

*Keywords:*
General linear model
Gaussian filter
FWHM
Filtering
Functional magnetic resonance imaging
Block-design

ABSTRACT

This study evaluates the effects of spatial smoothing on inter-subject correlation (ISC) analysis for FMRI data using the traditional model based analysis as a reference. So far within ISC analysis the effects of smoothing have not been studied systematically and linear Gaussian filters with varying kernel widths have been used without better knowledge about the effects of filtering. Instead, with the traditional general linear model (GLM) based analysis, the effects of smoothing have been studied extensively.

In this study, ISC and GLM analyses were computed with two experimental and one simulated block-design datasets. The test statistics and the detected activation areas were compared numerically with correlation and Dice similarity measures, respectively. The study verified that 1) the choice of the filter substantially affected the activations detected by ISC analysis, 2) the detected activations according to ISC and GLM methods were highly similar regardless of the smoothing kernel and 3) the effect of spatial smoothing was mildly smaller on ISC than GLM analysis. Our results indicated that a good selection of the full width at half maximum of the Gaussian smoothing kernel for ISC was slightly larger than double the original voxel size.



## 1. Introduction

Typical stimulus paradigms used with the functional magnetic resonance imaging (FMRI) are strictly controlled having precisely defined task and baseline periods. This makes it possible to define a parametric model for the activation time-course. This kind of FMRI data can be analyzed using a general linear model (GLM), where the stimulus time course needs to be modeled parametrically. However, these strictly controlled stimulus paradigms are necessarily simplified and perhaps limited in their facilities to reveal brain activity that occurs in real-life, and therefore more naturalistic stimulus such as movies are increasingly used instead. However, the parametric model for multidimensional stimulus such as a movie is challenging to define and for this reason alternatives to the GLM based analysis are required. The inter-subject correlation (ISC) is one potential analysis method for these more naturalistic stimulus paradigms. The ISC analysis is a completely data-driven method, based on voxel-wise correlation between the corresponding time series of the subjects where the high correlations are typically interpreted as activations. ISC does not require any parametric model of the stimulus time course and has no assumptions about the smoothness of the data. Albeit ISC methods do not require a model for the stimulus, they require the data to be task based. In the absence of the stimulus, we would expect no significant correlations due to the between subject similarities in processing the stimulus.

Since its introduction [1], the ISC methodology has been further developed [2–4]. The ISC method of [4] was validated in our earlier study [5] by comparing its results with the results of GLM. In [5], the used data were based on simple block-design tasks, ideal for the GLM based analysis, which justified the use of GLM as a gold standard. The validation confirmed that ISC detects highly similar activation areas as GLM within strictly controlled research setups. This indicated that ISC could find reliable activations in more complex, naturalistic stimulus paradigms. In [5], the Gaussian filter of 5 mm full width at half maximum (FWHM) was used for spatial smoothing according to the requirements of the GLM analysis. However, it is well known that the selection of filter width has a strong effect on the activations detected by GLM, which leads to interesting follow-up questions: Does the similarity between the analysis results of these methods depend on the applied filtering and how does the filtering influence the results of the ISC based analysis? This paper sets to answer these questions and evaluates the effects of spatial smoothing on the ISC analysis of FMRI data.

The effects of spatial smoothing in FMRI are known well within a GLM based analysis, but smoothing effects have not been studied within ISC. Typically, linear Gaussian filters with varying widths (see Table 1) have been used in ISC studies. In Table 1, the width of the filters varied from 5 to 12 mm and the ratio between voxel size and filter width from 1.67 to 4. A general rule of thumb in FMRI studies is that FWHM of the Gaussian filtering kernel is twice [6,7] or three

* Corresponding authors at: Department of Signal Processing, Tampere University of Technology, P.O. Box 553, FIN-33101, Finland.
*E-mail addresses:* juha.pajula@tut.fi (J. Pajula), jussi.tohka@tut.fi (J. Tohka).

http://dx.doi.org/10.1016/j.mri.2014.06.001
0730-725X/

**Table 1**
Used filter widths with ISC analysis.

| Pixel Resolution | FWHM | Ratio | Article |
|---|---|---|---|
| 3 mm | 6 mm | 2 | Hasson et al. [3] |
| 3 mm | 12 mm | 4 | Hasson et al. [1] |
| 3 mm | 8 mm | 2.67 | Wilson et al. [36] |
| 3 mm | 8 mm | 2.67 | Jääskeläinen et al. [37] |
| 3.44 mm | 12 mm | 3.49 | Hejnar et al. [2] |
| 3 mm | 8 mm | 2.67 | Golland et al. [38] |
| 3 mm | 5 mm | 1.67 | Kauppi et al. [4] |
| 3 mm | 8 mm | 2.67 | Nummenmaa et al. [39] |

times [8,9] the voxel size: For example, if the original voxel size is 2 mm, a kernel with at least 4 mm or 6 mm FWHM should be used. These considerations are based on the requirements of the random field theory (RFT) based multiple comparisons correction, which is commonly used to determine thresholds with the GLM based massively univariate analysis. However, in the ISC based analysis, non-parametric multiple comparison correction procedures, which do not have any formal requirements concerning data smoothness, are far more commonly used than RFT based multiple comparison correction.

In the GLM based analysis of FMRI data, spatial smoothing improves the signal to noise ratio (SNR), helps meeting the RFT requirements for smoothness, and increases the overlap of single subject statistics after spatial normalization in the group-level studies [6]. Nevertheless, using too large filters can cause several errors in statistical analysis: merge small activations together, spread the activation signal out from the original area and shift the activation centers to arbitrary positions [10]. Respectively, too small filters can compromise the spatial overlap of the activations of individual subjects in the group-level studies [10,11], as well as lead to a low SNR possibly resulting in a loss in statistical power [12]. The optimal spatial filter width is study dependent and thus the original voxel size of the data, the region of interest and inter-subject variability should be considered when selecting it [13–15]. The selected filter width should also correspond to the assumed size of the region of interest [16].

In this paper, we studied the effects of spatial smoothing on the ISC analysis by comparing the analysis results between ISC and GLM methods in a similar manner as in [5]. Both ISC and GLM analyses were computed with various Gaussian filter widths and a wavelet denoising method in two experimental blocked-design (external order and hand imitation tasks from the ICBM Functional Reference Battery) data sets and a simulated data set. Equal pre-processing was performed to experimental data within ISC and GLM methods as well as the resulting statistical maps were thresholded at the same significance level. This allows us to draw conclusions about the relative effects of the different filtering levels on the ISC and GLM analysis and contrast the differences (due to spatial filtering) in the ISC detected activations to these of GLM. We primarily consider Gaussian smoothing because the vast majority of the ISC based studies had applied it. However, we additionally consider a wavelet based spatial smoothing method to see if the conclusions reached for Gaussian smoothing extend to other denoising methods. The results of this work showed clearly how the choice of filter width can substantially affect activation areas detected by the ISC analysis. Interestingly, the effects of spatial smoothing differed between the GLM and ISC analysis methods.

## 2. Methods

A part of the materials and methods is as in [5], but the descriptions of the used data and main methods are briefly repeated here.

### 2.1. ICBM functional reference battery data

The used FMRI data included stimuli from the Functional Reference Battery (FRB) tasks developed by the International Consortium for Human Brain Mapping (ICBM) [17].[1] The data were obtained from ICBM database in the Image Data Archive of the Laboratory of Neuro Imaging.[2] 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 co-investigators from UCLA, Montreal Neurologic Institute, University of Texas at San Antonio, and the Institute of Medicine, Juelich/Heinrich Heine University-Germany.

The age range of the subjects was restricted to 20–38 years. In the ICBM database, this resulted in 41 right-handed subjects who had FMRI measurements from all five different FRB tasks: auditory naming (AN), external ordering (EO), hand imitation (HA), oculomotor (OM) and verbal generation (VG). The image data were pre-screened before analysis to ensure the high quality of the data. According to pre-screening, FMRI data from four subjects were discarded because of a poor data quality for at least one task in the battery. Instead of using all five tasks in this study the analysis was limited to the tasks with the highest (HA) and lowest (EO) similarity between ISC and GLM on [5]. The detailed definitions of the tasks are available in [5] and in the FRB software package.

The selected data included measurements from 37 healthy right-handed subjects (19 men and 18 women; average age was 28.2 years from the range of 20–36 years). The functional data were collected with a 3 tesla Siemens Allegra FMRI scanner and the anatomical $T_1$ weighted MRI data with a 1.5 tesla Siemens Sonata scanner. The TR/TE times for the functional data were 4 s/32 ms, flip angle 90°, pixel spacing 2 mm and slice thickness 2 mm. The parameters for the anatomical T1 data were 1.1 s/4.38 ms, 15°, 1 mm and 1 mm, correspondingly.

### 2.2. Pre-processing

The pre-processing and the GLM part of statistical analysis were performed by using FSL (version 4.1.6) from the Oxford Centre for Functional Magnetic Resonance Imaging of the Brain, Oxford University, Oxford, U.K. [18]. The data pre-processing was done in three phases. First, motion correction was performed using the FSL's MCFLIRT by maximizing the correlation ratio between each time point and the middle volume, using linear interpolation [19,20]. Second, the Brain extraction tool (BET) [21] was applied to extract the brain volume from functional data. Finally, the images were temporally high-pass filtered with a cutoff period of 60 s and the spatial smoothing was applied with an isotropic three dimensional Gaussian kernel with the FWHM of 0 mm, 2 mm, 4 mm, 5 mm, 8 mm and 12 mm in each direction. As an alternative to Gaussian smoothing, we also applied wavelet denoising with Wavelet Denoising Toolbox [22,23]. Following the recommendations in [22,23], two decomposition levels and wavelet family Symlet 2 (sym2) were applied. The original data had 87 volumes with three stabilization volumes, which were discarded from the analysis. The brain extraction from the anatomical $T_1$ images was also performed by BET, but this was done manually and separately from the main procedure for each $T_1$ weighted image as the parameters of BET required individual tuning.

The image registration was performed in two phases using FSL Linear Registration Tool (FLIRT) [19,20]. First, the skull-stripped functional images were aligned (6 degrees of freedom, full search) to

[1] http://www.loni.ucla.edu/ICBM/Downloads/Downloads_FRB.shtml.
[2] http://www.loni.ucla.edu/ICBM.

the skull-stripped high-resolution $T_1$ weighted image of the same subject, and then the results were aligned to the standard (brain only) ICBM-152 template (12 degrees of freedom, full search).

### 2.3. Simulated data

In addition to experimental data, a set of simulated imaging data was generated. The data set contained 37 simulated functional images in the standard MNI-152 space. Every voxel in these images was either activated or non-activated. Voxels were selected as activated according to the binarized statistical maps of the AN task (thresholded with $q = 0.05$ false discovery rate (FDR) corrected thresholds) from the GLM analysis in [5]. A hemodynamic signal was included in the time series of the activated voxels. The signal was selected to be exactly the same which was used as a model in the GLM analysis, i.e., a boxcar convolved with a canonical hemodynamic response function (HRF). Finally, pink 1/f noise was generated as described in [24][3] and added to every time series in the volume. The contrast to noise ratio of the noisy data was approximately 0.06 when computed against the amplitude of the boxcar before convolved with canonical HRF. The areas outside the activated regions contained only the noise signal. The simulation procedure was exactly the same for every 37 simulated images, that is, we ignored the anatomical and effect size variations between subjects. As the data were generated directly in MNI-152 coordinates no registration or motion correction was performed and pre-processing included only temporal and spatial filtering, exactly as described for experimental data.

### 2.4. Analysis methods

The ISC analysis was computed with ISCtoolbox [4,25]. The ISC is based on Pearson's correlations between the corresponding time series of all subject pairs and we denote the correlation coefficient between subjects $i$ and $j$ by $r_{ij}$. To obtain the final multi-subject test statistic, correlation values of all subject pairs $r_{ij}$ are combined into a single ISC statistic by averaging:

$$\bar{r} = \frac{1}{\frac{m^2-m}{2}} \sum_{i=1}^{m} \sum_{j=2, j>i}^{m} r_{ij}, \tag{1}$$

where $m$ is the number of subjects. The $m$ was 37 in our study and thus the correlation coefficients were averaged from $(37^2 - 37)/2 = 666$ subject pairs. The statistical inference was accomplished by a fully non-parametric voxel-wise resampling test implemented in the ISCtoolbox. This test accounts for temporal correlations inherent to FMRI data (for details of the test, see [4]). The resampling distribution was approximated with 1,000,000 realizations and the resulting p-values were corrected voxel-wise over the whole brain using an FDR based multiple comparisons correction with independence or positive dependence assumption [26,9]. The used threshold was $q = 0.001$.

The GLM was computed first with a single subject level and then the group-level statistic was computed from the single subject level results. Both phases were performed with FMRI Expert Analysis Tool (FEAT, version 5.98) from FSL [27,28]. The particular implementation of the multi-subject GLM (FSL's FLAME using MCMC) was selected because it is widely used and properly evaluated. The details of the analysis procedure were as described in [5] and as earlier FDR was chosen also for GLM as it is essential to compare the detected activations at the same significance level. More detailed discussions about the choice of thresholding methodology are available in [5,29].

[3] https://ccrma.stanford.edu/~jos/sasp/Example_Synthesis_1_F_Noise.html.

### 2.5. Methods comparison

The activation maps were compared with different numerical measures: Dice index between two sets of activated voxels, Pearson's correlations between test statistics and sensitivity and specificity measures between the detected activations and the ground truth of simulations. Dice index [30] measures the similarity of the detected activation areas between two thresholded and binarized activation maps and it is closely related to Kappa coefficient [5,31]. The binarized maps were created by assigning the value of voxel as one if the statistic value passed the threshold and otherwise as zero. The Dice index between two sets $B_1$ and $B_2$ of voxels was then defined as in Eq. (2):

$$d_{B_1,B_2} = \frac{2|B_1 \cap B_2|}{|B_1| + |B_2|}, \tag{2}$$

The numerator in Eq. (2) measures the size of common activation occurrence and the denominator measures the sizes of activated areas according to individual methods.

Pearson's correlations $C$ were computed between $\bar{r}$-values of ISC and absolute $Z$-values of GLM [5].

The absolute value of the $Z$-statistic was taken before computing the correlation measure because it is expected that both large negative and large positive z-statistics (so called activations and de-activations) relate to high $\bar{r}$-values (as activations and de-activations both lead to (large) positive inter-subject correlation values). It should be noted that large negative $\bar{r}$-value should not be interpreted as activation or similarity, and therefore, we do not take absolute value of the $\bar{r}$ statistic before computing the correlation. The ISC statistic $\bar{r}$ has a maximal value of 1 while the $Z$-statistic is not bounded. This difference has very small practical meaning since the observed $Z$-statistic values were finite and the observed $\bar{r}$ values were considerably smaller than the maximal value of one [5].

With simulated data, the Dice index was computed between the ground truth of simulated activations and the binarized results of ISC and GLM. In addition, we computed the sensitivity (the percentage of the activated voxels detected as activated) and specificity (the percentage of non-activated voxels detected as non-activated) measures with the simulated data. The measures were computed for both methods against the ground truth of the simulation across all the filtering levels.

Certain numerical methods were used to compare and display the differences between the methods over the used filtering levels and, to simplify the inspection of the results, specific detection matrices were generated. The Dice index values were collected in the Dice matrix according to Eq. (3):

$$D_{(k,l)} = \begin{bmatrix} d_{k,1} & d_{k,2} & \cdots & d_{k,k-1} & d_{k,l} \\ d_{k-1,1} & & & & d_{k-1,l} \\ \vdots & & \ddots & & \vdots \\ d_{2,1} & d_{2,2} & & & d_{2,l} \\ d_{1,1} & d_{1,2} & \cdots & d_{1,l-1} & d_{1,l} \end{bmatrix}, \tag{3}$$

where values $d_{k,l}$ ($k,l = 1,\ldots, n$ corresponding to the filtering levels) denote Dice indexes computed according to Eq. (2) between the thresholded activation maps with filtering levels $k$ and $l$. For experimental data $n$ was 6 (*FWHM = 0 mm, 2 mm, 4 mm, 5 mm, 8 mm, 12 mm*) and for the simulated data it was 5 (*FWHM = 0 mm, 2 mm, 4 mm, 5 mm, 8 mm*). It should be noted that in the detection matrix the vertical indexes increase from bottom to top for more intuitive display of the results. Filtering with 12 mm FWHM was not applied for simulated data as it was considered unnecessary.

We computed these matrices over a single method resulting in the intra-method Dice matrices as well as between the two methods

**Table 2**
Comparison of the detected activations against the ground truth of the simulation.

| Filter (FWHM) | Dice | | Specificity | | Sensitivity | |
|---|---|---|---|---|---|---|
| | GLM | ISC | GLM | ISC | GLM | ISC |
| 0 mm | 0.42 | 0.00 | 0.9980 | 1.0000 | 0.2714 | 0.0015 |
| 2 mm | 0.89 | 0.86 | 0.9792 | 0.9979 | 0.9432 | 0.7693 |
| 4 mm | 0.92 | 0.88 | 0.9841 | 0.9992 | 0.9652 | 0.7890 |
| 5 mm | 0.86 | 0.91 | 0.9654 | 0.9964 | 0.9779 | 0.8574 |
| 8 mm | 0.70 | 0.87 | 0.9029 | 0.9815 | 0.9844 | 0.8896 |
| Average | 0.76 | 0.70 | 0.9659 | 0.9950 | 0.8284 | 0.6614 |
| Average (2–8 mm) | 0.84 | 0.88 | 0.9579 | 0.9938 | 0.9677 | 0.8263 |

resulting in the inter-method Dice matrix. For example, in the intra-method Dice matrix for ISC, $d_{3,2}$ refers to the Dice index between activations detected by ISC with 2 mm and 4 mm filtering. Similarly, in the inter-method Dice matrix, $d_{3,2}$ refers to the Dice index between activations detected by ISC with 4 mm and GLM with 2 mm.

When the matrix was computed over a single method, i.e., ISC (GLM) activations were compared with ISC (GLM) activations, the main cross diagonal ($d_{1,1}$,$d_{2,2}$,…, $d_{n,n}$) contained only ones corresponding perfect similarity (since a set was compared with itself). When the matrix was computed over a single method, it was also symmetric over the main cross diagonal.

The voxel was defined as unique when it was detected as activated with a filtering level $k$ but not with a level $l$. The numbers of unique voxels $v_{k,l}$ across all filtering levels were counted and collected in a unique voxel count matrix $V$ in the same manner as matrix $D$. The matrix $V$ was computed for both studied analysis methods independently. Now the main cross diagonal of the matrix $V$ contained only zeros, which correspond to the full similarity in the sense of unique detected voxels (no unique voxel exists when the data are compared with themselves).

In order to measure the development in detected activation clusters, the 26-connectivity measure [32] was used to search for the clusters in the binarized statistics.

### 2.6. Jackknife estimates

We used the leave one out jackknife standard error estimator [33] to evaluate the variability of correlation and Dice measures of similarity with respect to subject sample. Denote the fMRI data from subject $i$ as $x_i$, $i = 1,\ldots, n$, where $n = 37$ is the number of subjects. The jackknife estimate of standard error for a measure $\hat{\rho}$ is defined as

$$\hat{\sigma}_J = \sqrt{\frac{(n-1)}{n}\sum_{i=1}^{n}\left(\hat{\rho}_{(i)} - \hat{\rho}_{(\cdot)}\right)^2}, \tag{4}$$

where $\hat{\rho}_{(i)} = \hat{\rho}(x_1, x_2, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n)$ denotes the measure $\hat{\rho}$ computed without the data from $i^{th}$ subject and $\hat{\rho}_{(\cdot)} = \frac{1}{n}\sum_{i=1}^{n}\hat{\rho}_{(i)}$. In our case $\hat{\rho}$ can be either the correlation measure $C$ between the test statistic values or Dice index between the two sets of voxels.

### 2.7. Peak shift measurements

To quantify the changes in the coordinates of the peaks over the used filtering levels, we followed a technique adapted from [13]. First, three spatial locations of the peaks (i.e., local maxima of the test statistic) from the EO and HA tasks were manually selected so that the peaks in the $Z$-statistics of GLM and the $\bar{r}$ statistics of ISC matched. This was done by first selecting three clusters so that corresponding clusters were approximately on the same location after thresholding in both, ISC and GLM, statistics with non-smoothed data set. Then, local maxima were selected from each cluster so that the spatial distances between corresponding peaks in ISC and GLM statistics were as small as possible. Only a single peak per cluster and method was selected. These peaks were then tracked over the smoothing levels (2 mm to 12 mm filtering levels) so that, for each smoothing level, the local maximum nearest to the selected peak in non-smoothed data set was searched. Each search was further limited to a search area with the radius corresponding to the used FWHM. For example, if the data were smoothed by 4 mm FWHM kernel, it was assumed that the local maximum could move a maximum of 4 mm from the original location. This restriction was necessary to ensure the tracing of the same peak in practice. Finally, the absolute changes for each filtering level were computed as Euclidean distance between the original location and the new location and these distances were averaged over successive filtering levels.

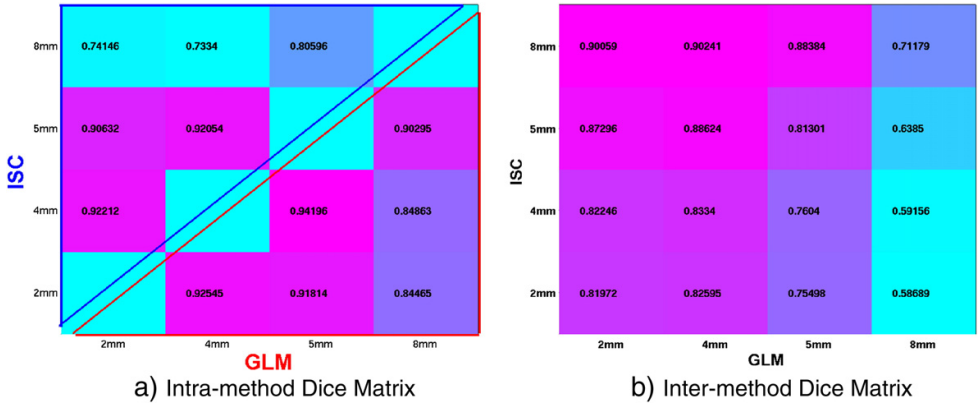



**Fig. 1.** Dice matrices for the simulated data. The Dice matrices were constructed according to Eq. (3). The intra-method Dice matrix (a) is constructed from two cross diagonal symmetric matrixes by selecting either upper (ISC, blue) or lower (GLM, red) triangle for the final matrix. The inter-method Dice matrix (b) contains the Dice similarity between the detected activations of ISC and GLM methods. The matrix in panel (a) shows that ISC had the lowest similarity (Dice value 0.73) between thresholded areas with 4 mm and 8 mm filters and GLM had the lowest similarity (Dice value 0.84) with 2 mm and 8 mm filters. The matrix in panel (b) shows that ISC and GLM had the most similar detection areas (Dice value 0.90) when ISC had 8 mm and GLM had 4 mm kernel in use.

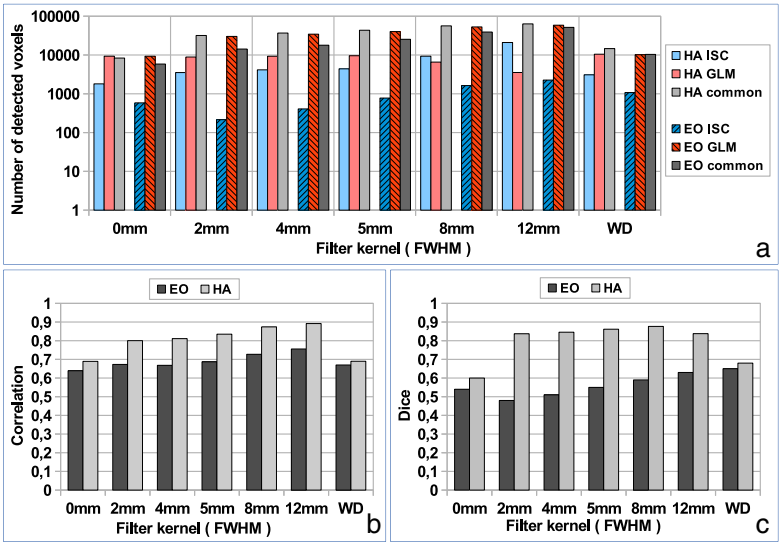



**Fig. 2.** Unique voxels (a), Correlation measures (b) and Dice measures (c) between ISC and GLM. Panel (a) presents unique voxels between methods after FDR corrected ($q = 0.001$) thresholding. The voxel is defined as unique if it is detected by a single analysis method (ISC or GLM) but not by both. The columns present the number of unique voxels, which were activated according to ISC (blue columns) or GLM (red columns) analysis only. The gray bars present the number of common voxels detected by both ISC and GLM methods. The abbreviation WD refers to wavelet denoising. Panel (b) presents the correlation measures between ISC and GLM statistics and panel (c) presents the Dice values between ISC and GLM detections.

## 3. Results

### 3.1. Simulated data

Table 2 presents the Dice indexes, sensitivities, and specificities of the ISC and GLM methods against the ground truth of the simulated data. With the simulated data, both methods performed poorly when no spatial filtering was applied. With both methods, the average Dice value over 2 mm to 8 mm (0.84 for GLM and 0.88 for ISC) against the ground truth was on the level of "Almost Perfect Agreement" (0.8–1.0) according to the categorization of Landis and Koch [34]. The sensitivity of GLM was on average higher (0.97 for GLM versus 0.83 for ISC) and vice versa the specificity of ISC was slightly higher on average (0.96 for GLM versus 0.99 for ISC).

Fig. 1 presents the Dice matrices defined in Eq. (3) for the simulated data (this figure concerns only differences of the activations detected by GLM and ISC methods and makes no reference to ground-truth of the simulation). The matrix in panel (a) of Fig. 1 presents the intra-method Dice similarity across the used filtering levels by combining two symmetric Dice matrices. The upper half of the matrix represents ISC (blue triangle) and the lower half represents GLM (red triangle). Both methods had the highest values in the first cross diagonal, which was natural as filters with similar FWHM should lead to similar detected activations. The matrix in panel (b) of Fig. 1 shows the Dice similarity between the methods. The highest similarity (0.90) was found when ISC had 8 mm filtering and GLM 4 mm filtering.

**Table 3**
Correlation values ($C$) between ISC and GLM statistics and their jackknife standard error estimates ($\hat{\vartheta}_{JC}$).

| Filter (FWHM) | EO | | HA | |
|---|---|---|---|---|
| | $C$ | $\hat{\vartheta}_{JC}$ | $C$ | $\hat{\vartheta}_{JC}$ |
| 2 mm | 0.67 | 0.013 | 0.80 | 0.024 |
| 4 mm | 0.67 | 0.013 | 0.81 | 0.026 |
| 5 mm | 0.69 | 0.012 | 0.83 | 0.028 |
| 8 mm | 0.73 | 0.016 | 0.87 | 0.024 |
| 12 mm | 0.76 | 0.017 | 0.89 | 0.020 |
| Average | 0.70 | 0.014 | 0.84 | 0.024 |

### 3.2. Experimental data

Fig. 2(a) presents the number of voxels that were detected as activated by both methods and the number of unique voxels detected as activated by one method but not by the other. According to the figure, ISC was more conservative than GLM, because ISC found fewer unique voxels than GLM with all studied filtering methods.

Table 3 and Fig. 2(b) present Pearson's correlations $C$ between the test statistics over different filtering levels. Table 4 and Fig. 2(c) present the Dice index values between activation maps of the ISC and GLM methods. The average correlation for the EO task was 0.70 with jackknife standard error estimates of 0.014 averaged over filtering levels and for the HA task the average correlation was 0.84 with average error standard estimate of 0.024. The average Dice measures were 0.55 for EO and 0.85 for HA with average standard error estimates of 0.13 and 0.064. Corresponding Dice measures were 0.54 (EO) and 0.60 (HA). For the wavelet denoised (WD) data the correlation values were 0.67 (EO) and 0.69 (HA) and Dice values were 0.65 (EO) and 0.68 (HA).

According to Landis and Koch [34] categorization, all Dice values for the EO task were categorized as "Moderate Agreement". Instead, most of the Dice values for the HA task were categorized as "Almost Perfect Agreement" (see [5] about the details of the categorization). The highest correlation (0.76) and Dice value (0.63) were reached using the largest 12 mm filter in the EO task. With the HA task, the highest correlation (0.89) was reached with 12 mm filter also, but the highest Dice similarity (0.88) was reached with 8 mm filter. The correlations $C$ increased with the filter size and for both tasks the increase in $C$ across filtering levels was several times larger than the

**Table 4**
Dice measurements ($d$) between the activation areas detected by ISC and GLM methods and their jackknife standard error estimates ($\hat{\sigma}_{Jd}$).

| Filter (FWHM) | EO | | HA | |
|---|---|---|---|---|
| | $d$ | $\hat{\sigma}_{Jd}$ | $d$ | $\hat{\sigma}_{Jd}$ |
| 2 mm | 0.48 | 0.15 | 0.84 | 0.075 |
| 4 mm | 0.51 | 0.15 | 0.85 | 0.065 |
| 5 mm | 0.55 | 0.13 | 0.86 | 0.048 |
| 8 mm | 0.59 | 0.10 | 0.88 | 0.036 |
| 12 mm | 0.63 | 0.11 | 0.84 | 0.097 |
| Average | 0.55 | 0.13 | 0.85 | 0.064 |

standard error estimate. The Dice values for the HA task were rather stable across filtering levels (within one standard error margin). For the EO task, a trend of increasing Dice values could be noticed, however, these increases are small with respect to standard error estimates. The jackknife standard error estimates of the correlations were all small for all tested filtering levels in both tasks (the average standard error was 0.014 for EO and 0.024 for HA). For Dice measures, the standard errors were small for the HA task (average 0.064) but larger for the EO task (average 0.13).

Fig. 3 shows the detected activation areas in the axial slice at the $Z = 32$ mm (anteroposterior axis in MNI-152 coordinates) over ICBM-152 template. It can be observed that large smoothing kernels spread the activation area. The values of the test statistics along the lines from left to right at MNI coordinates $Y = 76$ mm, $Z = 32$ mm of the EO task and at $Y = 26$ mm, $Z = 32$ mm of the HA task are presented in Fig. 4 for both methods. The corresponding lines are marked by the blue color in Fig. 3.

In Fig. 4, it is easy to notice how the shapes of activation profiles are similar between the methods. However, with ISC, the

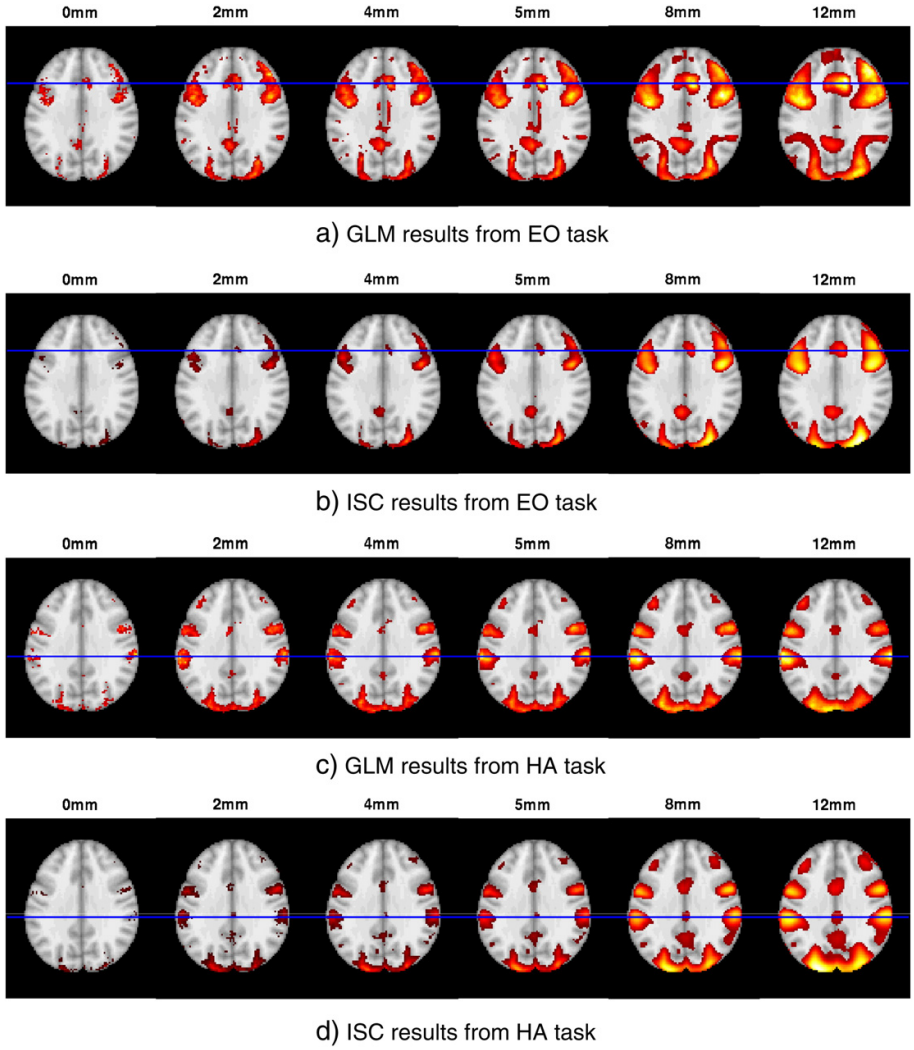



**Fig. 3.** Thresholded activation maps for the EO (panels a, b) and HA (panels c, d) task (FDR corrected, $q = 0.001$). The images in panels (a) and (c) present the results of the GLM analysis and the images in panels (b) and (d) present the corresponding results for the ISC method. The images are from the axial slice of $Z = 32$ mm anteroposterior axis in MNI-152 coordinate system. The blue line in the images presents the position of line visualized in Fig. 4. It can be observed that the ISC analysis was on average more conservative than the GLM analysis and the spreading effect was less obvious with ISC.

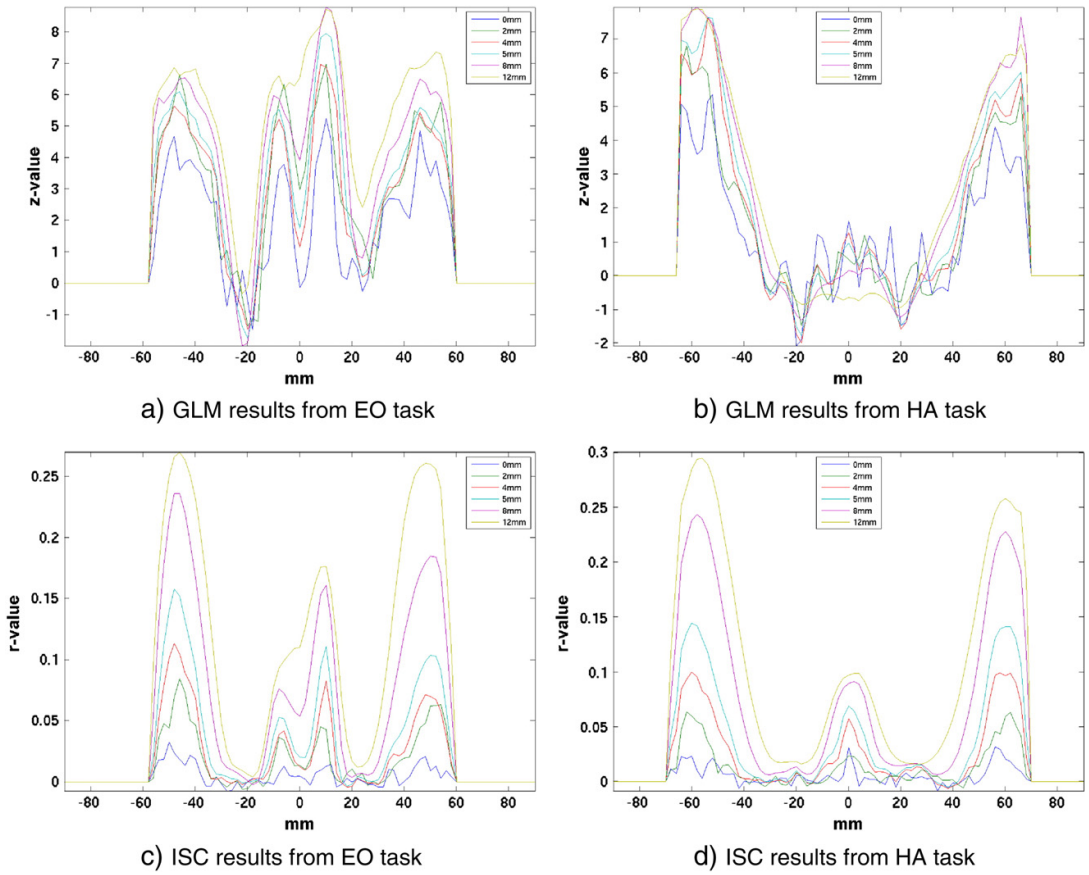



**Fig. 4.** The values of the test statistics for GLM (a, b) and ISC (c, d) tasks. The images present the values of test statistics along the left to right lines, which are marked with the blue color in the image series of Fig. 3. The images in panels (a) and (c) present the value of test statistics along the line from the EO task in the position Y = 24 mm, Z = 32 mm in MNI-152 coordinate system. Panels (b) and (d) present the lines from HA task in the position of Y = −26 mm, Z = 32 mm in MNI-152 coordinate system. The shapes of the lines are similar. In the case of ISC (panels (c) and (d)) the peaks remained in one position; instead with GLM (panels (a) and (b)) the peaks shifted across the filtering levels. The phenomenon is easy to notice for example from the line of the EO task (a) at the approximated x-coordinate of −50 mm.

activation peaks were staying here approximately in a single position although smaller peaks were merged together (see, for example, 0 mm position in Fig. 4(d)). The peak shifting in this single direction for GLM is easy to notice here, for example, in Fig. 4(a) at the approximate x-coordinate location of −50 mm. Other effects of spatial smoothing on GLM analysis are noticeable in Fig. 4(b) which demonstrates how the small activation peaks were merged together (for example, compare the peaks in ±65 mm positions).

**Table 5**
Average changes of the locations of local maximum peaks the tested Gaussian denoising levels. Differences are in millimeters of the Euclidean distance Δ from the local maximum peak of the corresponding test with non-smoothed data. For EO task the selected peaks were from clusters in Paracingulate Gyrus (ParCin), Precentral Gyrus at Left (PreCL) and Precentral Gyrus at right (PreCR) areas. Correspondingly for the HA task the three peaks were selected from clusters in areas of Supplementary Motor Cortex (SMC), Precentral Gyrus Left (PreCL) and Precentral Gyrus Right (PreCR).

| Task | | EO | | | HA | | | |
|---|---|---|---|---|---|---|---|---|
| Method | | ParCin | PreCR | PreCL | SMC | PreCL | PreCR | Average |
| Average Δ (mm) | ISC | 3.0 | 4.9 | 1.9 | 1.8 | 2.5 | 3.2 | 2.9 |
| | GLM | 3.3 | 4.8 | 1.9 | 3.6 | 3.8 | 1.9 | 3.2 |

To obtain a quantitative measure of peak movement the local maxima in three clusters were traced in both tasks. The average movement distances of local maxima are listed in Table 5. The average movement distance for ISC method was 2.9 mm and 3.2 mm for GLM method (averaged over three peaks). The initial local maxima for EO task were in the vicinity of MNI coordinates (−2,18,44) in Paracingulate Gyrus, (50,12,32) in Precentral Gyrus (right) and (−46,6,30)in Precentral Gyrus (left). For HA task the initial peak locations were (2,−6,50) in Supplementary Motor Cortex, (−36,−4,62) in Precentral Gyrus (left) and (36,−22,66) in Precentral Gyrus (right). Supporting the visual evidence in Fig. 4 the peak movement across the filtering levels was slightly smaller in ISC analysis.

Figs. 5 and 6 present the Dice and unique voxel count matrices relating to EO and HA tasks. The Dice matrices in panels (a) and (b) of these figures are as explained in the context of Fig. 1. The matrices in panels (c) and (d) of these figures are the unique voxel count matrices. The values in the Dice matrix for EO task (Fig. 5(a)) increased towards main cross diagonal as it could be expected based on the simulations. The similarity between methods in Fig. 5(b) was not linear as the highest values between methods were not on the main cross diagonal of the matrix: The highest Dice similarity between methods (0.77) was reached when ISC had 8 mm filter and

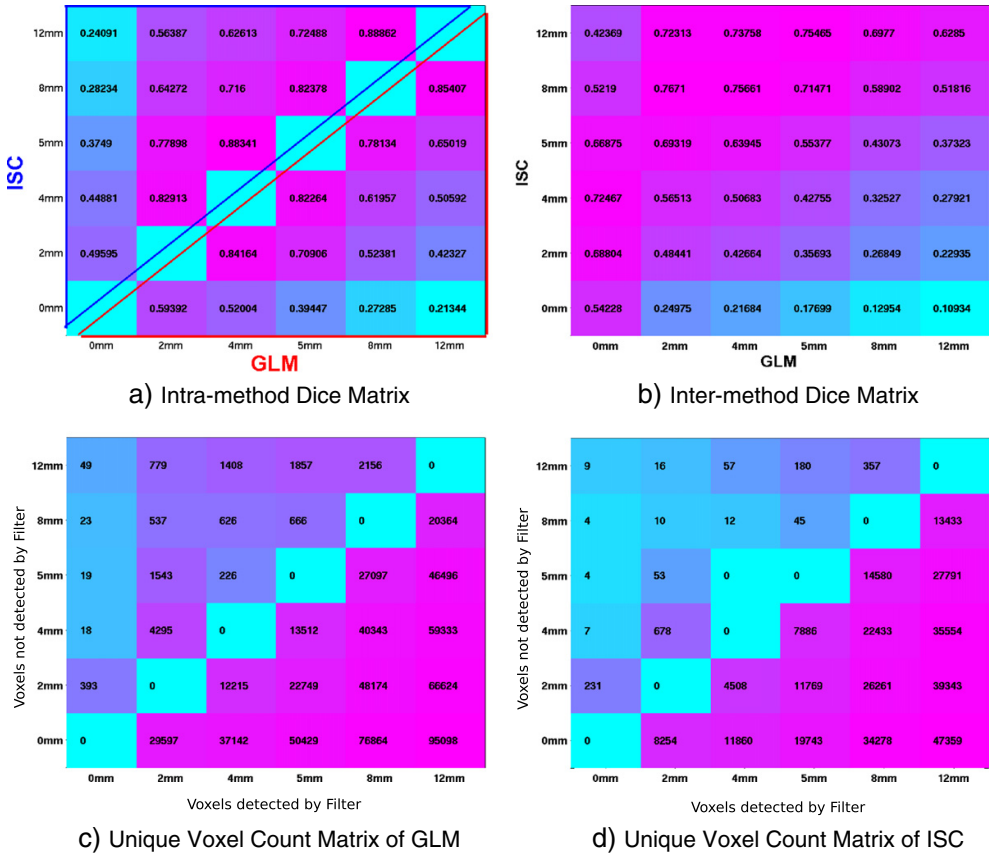



**Fig. 5.** The Dice and the unique voxel count matrices of the EO task. See Fig. 1 for the details of the Dice matrices in panels (a) and (b). In panel (a), intra-method Dice values are similar between the methods. In panel (b), the best similarity (0.77) between the methods was reached when ISC had a large 8 mm filter in use and GLM had a small 2 mm filter in use. Panels (c) and (d) present the unique voxel count matrices for GLM and ISC. When comparing the unique voxel counts of GLM (c) and ISC (d), it is noticeable that GLM had clearly more voxels which were detected by larger filters but not by smaller filters as indicated by relatively large values in the upper left triangle in the unique voxel count matrix in panel (c). This means that increased filter width led to the loss of some activation areas with GLM. Instead, ISC had relatively smaller values in the upper left triangle in panel (d), which means that with ISC there was no such loss and the detected activation area merely grew with filter width.

GLM had 2 mm filter in use. With the HA task, the results of inter-method Dice matrix differed slightly as the between methods similarity was close to linear in Fig. 6(b) (the highest values were on the main cross diagonal) and the highest similarity (0.88) was found when both methods had 8 mm filter in use.

When comparing the intra-method unique voxel count matrices in Figs. 5 and 6 (c) and (d), it can be observed that ISC had lower values than GLM in the upper left triangle. This points out that, in comparison to GLM, there were fewer such voxels for ISC that were detected by smaller filters but not by larger filters. The significance of the unique voxel count proportion differences between ISC and GLM in Figs. 5 (c–d) and 6 (c–d) was tested with Fisher's exact test. The test confirmed that all differences were significant (one-sided $p < 0.025$ in all cases).

The filter size had a clear effect on the size of detected activations. Fig. 7 presents the number of detected clusters and the average size of the clusters over the different filtering levels. The lines in the figure present clusters computed from the voxels detected by ISC (turquoises) and GLM (grays) methods. The bars present the size of the detected clusters in voxels by ISC (green and blue) and GLM (yellow and red) for both tasks. As predicted, the size of clusters increased at the same time as their number decreased: The smaller clusters were merged together when the filter size was increased.

## 4. Discussion

In this study, the effects of smoothing on the ISC based analysis of fMRI data were evaluated by comparing the analysis results of ISC and GLM methods in standardized block design tasks, where one can expect the GLM analysis to be accurate. Both ISC and GLM analyses were performed using smoothing with varying FWHM kernel widths for the same real and simulated datasets. Pre-processing steps and the FDR based false positive thresholding were the same for both methods. Our results indicated that the choice of filter width affects the ISC analysis results, as it is well demonstrated in the case of standard GLM analysis. Interestingly, the effects of filtering were slightly different between the ISC and GLM analyses.

According to the earlier research [6,8,9,13,10,11,16,14,15], multiple strategies exist for selecting the spatial filtering kernel for a FMRI study. These strategies are highly dependent on the used statistical analysis methods and reasons for spatial smoothing differ, for example, between the ISC and GLM analyses. First, ISC is

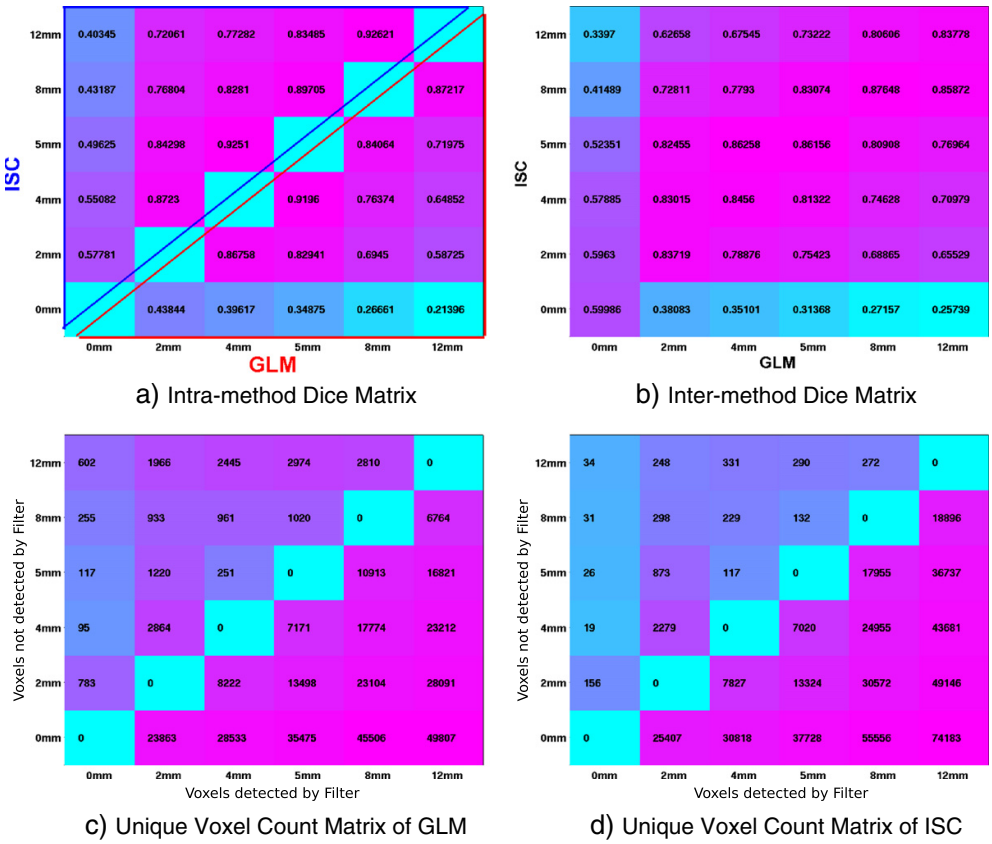



**Fig. 6.** The Dice and the unique voxel count matrices of the HA task. See Fig. 1 for the details of the Dice matrices in panels (a) and (b). In panel (a), intra-method Dice values of ISC were higher than the intra-method Dice values of GLM. The matrix in panel (b) shows how the results of both methods were close the same with all filtering levels and the highest values were in the cross diagonal. The highest similarity (0.88) was reached when both methods had the filter kernel of 8 mm in use. When comparing the unique voxel counts of GLM in panel (c) and ISC in panel (d) the similar behavior was present as with the EO task in Fig. 5: ISC had lower values in the upper triangle.

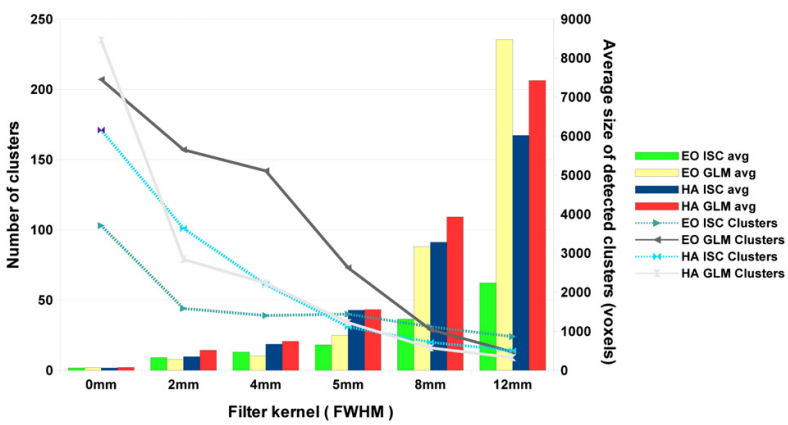



**Fig. 7.** The number of detected activation clusters versus the average sizes of clusters. The bars present the average size of detected clusters in voxels and the lines present the number of the detected clusters. The size of clusters increases approximately at the same speed as the number of clusters decreases when filter kernel width increases. When comparing the number of voxels in clusters, cluster sizes and number of detected voxels in Fig. 2(a), one observation is that small clusters are merged together when filter kernel width increases. For example, the average cluster size with GLM in the EO task is almost three times larger with 12 mm filter than 8 mm filter.

computed over the group of subjects, which makes the registration to the common stereotactic space the most critical part of the pre-processing. With an inaccurate registration, ISC fails as the FMRI time-courses that are correlated do not match spatially. For this reason spatial smoothing is often used to counterbalance the inaccuracy of spatial registration. Second, ISC does not apply any parametric stimulus model and therefore it has no parametric assumptions about the data. The statistical inference in ISC is commonly based on non-parametric tests unlike in GLM, where satisfying the assumptions by the parametric model is one of the major reasons for spatial smoothing. For these reasons, spatial smoothing guidelines devised for GLM might be inapplicable to ISC.

Earlier we had demonstrated that ISC found the same activation areas as GLM in strictly controlled research setups [5]. In this paper, we showed that both methods found similar activations regardless of the used filtering method. This is visible in Fig. 2(b) and (c) where the correlations between test statistics and Dice values, were high and fairly stable across the filtering levels. This was also confirmed in Fig. 3 where the detected activation areas changed in a similar manner with both methods. The wavelet denoising led to similar correlation values and Dice indices as the Gaussian smoothing, which indicates that similarity between the ISC and GLM analysis did not depend on Gaussian smoothing. The maximal similarity between the methods was not achieved by using the same filtering in the EO task. The maximum Dice index was 0.77 when ISC used 8 mm filter and GLM used 2 mm filter (see Fig. 5(b)). Instead, for the HA task, the highest similarity (0.88) was found using 8 mm filter kernel with both methods (see Fig. 6(b)).

The widely used rule of thumb that the filter width should be at least twice the size of the voxel (here the voxel size was 2 mm) seemed to hold also with the ISC analysis: ISC analysis results lacked sensitivity and were noisy with smaller filter widths. The unwanted smoothing effects such as merging different activation areas together, moving the centers of activations and spreading the activation in the surrounding areas were visible in the results of both ISC and GLM analysis methods. Based on Fig. 4, it appears that movements of activation peaks were a more severe problem with GLM than with ISC. As an example, the peak-movement problem with GLM is visible with larger filters (8 mm and 12 mm) in Fig. 4(a) around the position of 50 mm in x-axis. With ISC in Fig. 4(c) and (d) the movement was mainly caused by the merger of smaller peaks and thus ISC suffered mainly from the spreading and merging effects. These observations are also supported by peak shift measurements in Table 5, where the changes in peaks were slightly smaller with ISC than with GLM.

When examining the detected activations within a single method across filtering levels, the detected activations changed less with ISC than with GLM. For example, when successive filtering levels were compared (see values next to the cross diagonal in Fig. 5(a)) ISC had on average 3.7 % larger values with the smoothed data (2 mm–12 mm kernels) in the EO task and, in the HA task (see Fig. 6(a)), the average values of ISC were 3.3% larger. The unique voxel count matrices in Fig. 5(c) and (d) and Fig. 6(c) and (d) suggested that within ISC analysis there existed less voxels which were detected with smaller filter widths but not with larger filter widths. These observations combined with the peak shift measurements suggest that the variation on spatial filtering had smaller effects on the ISC analysis than on the GLM analysis. In consequence, ISC could be argued to tolerate slightly larger Gaussian smoothing kernels than GLM.

Caution is needed when investigating the values in Table 2, as the simulation was extremely simple and thus the results of the table are merely suggestive. The simulation was observed to be almost too perfect for GLM, which detected activation also from the areas where the signal had spread from the true activation area. However, also with simulated data, ISC seemed to be affected less by the change of filter size than GLM as was the case with experimental data. According to the Dice values in the Table 2, the optimal filtering level with the simulated data was 5 mm for ISC (Dice value 0.91) and 4 mm for GLM (Dice value 0.92).

The spatial filtering was mandatory for both analysis methods: GLM was not able to perform whitening and the sensitivity of ISC was minimal with 0 mm filter with simulated data (see Table 2). The main reason for GLM to fail on simulated data with 0 mm kernel was that GLM failed to compute estimates for the whitening process and for this reason the GLM analysis was computed without whitening for 0 mm kernel. This led to a severe number of false positive voxels in the thresholded statistics. Because of the poor performance of both ISC and GLM, the results of 0 mm kernel in the Table 2 are presented only as a reference and are not comparable to the other results. As both ISC and GLM were able to detect reasonable activations without filtering with the experimental data (see Fig. 2), the simulation pointed out an important property of the pre-processing pipeline: the experimental data had multiple pre-processing steps, such as motion correction and image registration, which resulted in a certain amount of smoothing as a side product [35].

The ISC analysis requires spatial smoothing to ensure the spatial registration accuracy and decent SNR. Our results indicated that, for ISC, a good selection of FWHM for Gaussian smoothing kernel was slightly larger than double the original voxel size. Furthermore, this study verified that 1) the choice of the filter width substantially affected the activations detected by ISC analysis, 2) the detected activations according to ISC and GLM methods were highly similar regardless of the used spatial smoothing method and 3) the used Gaussian filter for the spatial smoothing can be slightly wider with ISC than with GLM analysis.

## Acknowledgments

This research was supported by the Academy of Finland (grant numbers 130275 and 263785).

Data collection and sharing for this project were 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 California, Los Angeles.

## References

[1] Hasson U, Nir Y, Levy I, Fuhrmann G, Malach R. Intersubject synchronization of cortical activity during natural vision. Science 2004;303(5664):1634–40.
[2] Hejnar MP, Kiehl KA, Calhoun VD. Interparticipant correlations: a model free fMRI analysis technique. Hum Brain Mapp 2007;28(9):860–7.
[3] 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.
[4] Kauppi J-P, Jääskeläinen IP, Sams M, Tohka J. Inter-subject correlation of brain hemodynamic responses during watching a movie: localization in space and frequency. Front Neuroinform 2010;4:5.
[5] Pajula J, Kauppi J-P, Tohka J. Inter-subject correlation in fMRI: method validation against stimulus-model based analysis. PLoS ONE 2012;7(8):e41196. http://dx.doi.org/10.1371/journal.pone.0041196.
[6] Ball T, Breckel TPK, Mutschler I, Aertsen A, Schulze-Bonhage A, Hennig J, et al. Variability of fMRI-response patterns at different spatial observation scales. Hum Brain Mapp 2012;33(5):1155–71.
[7] Friston K, Holmes A, Poline J-B, Price C, Frith C. Detecting activations in PET and fMRI: levels of inference and power. NeuroImage 1996;4(3):223–35. http://dx.doi.org/10.1006/nimg.1996.0074.
[8] Petersson KM, Nichols TE, Poline J-B, Holmes AP. Statistical limitations in functional neuroimaging II: signal detection and statistical inference. Philos Trans R Soc Lond B 1999;354:1261–81.
[9] Nichols T, Hayasaka S. Controlling the familywise error rate in functional neuroimaging: a comparative review. Stat Methods Med Res 2003;12(5):419–46.

[10] White T, O'Leary D, Magnotta V, Arndt S, Flaum M, Andreasen NC. Anatomic and functional variability: the effects of filter size in group fMRI data analysis. NeuroImage 2001;13(4):577–88.
[11] Fransson P, Merboldt K, Petersson KM, Ingvar M, Frahm J. On the effects of spatial filtering — a comparative fMRI study of episodic memory encoding at high and low resolution. NeuroImage 2002;16(4):977–84.
[12] Scouten A, Papademetris X, Constable RT. Spatial resolution, signal-to-noise ratio, and smoothing in multi-subject functional MRI studies. NeuroImage 2006;30(3):787–93.
[13] Mikl M, Marecek R, Hluštík P, Pavlicová M, Drastich A, Chlebus P, et al. Effects of spatial smoothing on fMRI group inferences. Magn Reson Imaging 2008;26(4):490–503.
[14] Shaw ME, Strother SC, Gavrilescu M, Podzebenko K, Waites A, Watson J, et al. Evaluating subject specific preprocessing choices in multisubject fMRI data sets using data-driven performance metrics. NeuroImage 2003;19(3):988–1001.
[15] Weibull A, Gustavsson H, Mattsson S, Svensson J. Investigation of spatial resolution, partial volume effects and smoothing in functional MRI using artificial 3d time series. NeuroImage 2008;41(2):346–53. http://dx.doi.org/10.1016/j.neuroimage.2008.02.015.
[16] Clare S. Functional MRI: methods and applications. [Ph.D. thesis] University of Nottingham; 1997.
[17] Mazziotta J, Toga A, Evans A, Fox P, Lancaster J, Zilles K, et al. A probabilistic atlas and reference system for the human brain: International Consortium for Brain Mapping (ICBM). Philos Trans R Soc Lond Ser B Biol Sci 2001;356(1412):1293–322.
[18] Smith SM, Jenkinson M, Woolrich MW, Beckmann CF, Behrens TEJ, Johansen-Berg H, et al. Advances in functional and structural MR image analysis and implementation as fsl. NeuroImage 2004;23(Suppl 1):208–19.
[19] Jenkinson M, Smith S. A global optimisation method for robust affine registration of brain images. Med Image Anal 2001;5(2):143–56.
[20] Jenkinson M, Bannister P, Brady M, Smith S. Improved optimization for the robust and accurate linear registration and motion correction of brain images. NeuroImage 2002;17(2):825–41.
[21] Smith SM. Fast robust automated brain extraction. Hum Brain Mapp 2002;17(3):143–55.
[22] Khullar S, Michael A, Correa N, Adali T, Baum S, Calhoun V. Wavelet-based denoising and independent component analysis for improving multi-group inference in fMRI data. Biomedical imaging: from nano to macro, 2011 IEEE International Symposium on; 2011. p. 456–9. http://dx.doi.org/10.1109/ISBI.2011.5872444.
[23] Khullar S, Michael A, Correa N, Adali T, Baum SA, Calhoun VD. Wavelet-based fMRI analysis: 3-D denoising, signal separation, and validation metrics. NeuroImage 2011;54(4):2867–84. http://dx.doi.org/10.1016/j.neuroimage.2010.10.063.
[24] Smith JO. Spectral audio signal processing. http://-ccrma.stanford.edu/~jos/-sasp/. [Accessed 24.05.2013, online book].
[25] Kauppi J-P, Pajula J, Tohka J. A versatile software package for inter-subject correlation based analyses of fMRI. Front Neuroinform 2014;8:5.
[26] Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Ser B Methodol 1995;57(1):289–300.
[27] Beckmann CF, Jenkinson M, Smith SM. General multilevel linear modeling for group analysis in fMRI. NeuroImage 2003;20(2):1052–63.
[28] 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.
[29] Logan BR, Geliazkova MP, Rowe DB. An evaluation of spatial thresholding techniques in fMRI analysis. Hum Brain Mapp 2008;29(12):1379–89.
[30] Dice LR. Measures of the amount of ecologic association between species. Ecology 1945;26(3):297–302.
[31] Zijdenbos AP, Dawant BM, Margolin RA, Palmer AC. Morphometric analysis of white matter lesions in MR images: method and validation. IEEE Trans Med Imaging 1994;13(4):716–24.
[32] Bellec P, Perlbarg V, Jbabdi S, Pélégrini-Issac M, Anton J-L, Doyon J, et al. Identification of large-scale networks in the brain using fMRI. NeuroImage 2006;29(4):1231–43.
[33] Efron B, Gong G. A leisurely look at the bootstrap, the jackknife, and cross-validation. Am Stat 1983;37(1):36–48.
[34] Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977;33(1):159–74.
[35] Strother S, Conte SL, Hansen LK, Anderson J, Zhang J, Pulapura S, et al. Optimizing the fMRI data-processing pipeline using prediction and reproducibility performance metrics: I. A preliminary group analysis. NeuroImage 2004;23(Suppl 1):196–207.
[36] Wilson SM, Molnar-Szakacs I, Iacoboni M. Beyond superior temporal cortex: intersubject correlations in narrative speech comprehension. Cereb Cortex 2008;18(1):230–42.
[37] Jääskeläinen IP, Koskentalo K, Balk MH, Autti T, Kauramäki J, Pomren C, et al. Inter-subject synchronization of prefrontal cortex hemodynamic activity during natural viewing. Open Neuroimag J 2008;2:14–9.
[38] Golland Y, Bentin S, Gelbard H, Benjamini Y, Heller R, Nir Y, et al. Extrinsic and intrinsic systems in the posterior cortex of the human brain revealed during natural sensory stimulation. Cereb Cortex 2007;17(4):766–77.
[39] Nummenmaa L, Glerean E, Viinikainen M, Jääskeläinen IP, Hari R, Sams M. Emotions promote social interaction by synchronizing brain activity across individuals. Proc Natl Acad Sci 2012;109(24):9599–604.

# Publication IV

Pajula J., Kauppi J-P., Tohka J. "How Many is Enough? Effect of Sample Size in Inter-Subject Correlation Analysis of fMRI", *Computational Intelligence and Neuroscience*, volume 2016, ID 2094601, pages 1-10, 2016.

Hindawi Publishing Corporation
Computational Intelligence and Neuroscience
Volume 2016, Article ID 2094601, 10 pages
http://dx.doi.org/10.1155/2016/2094601

*Research Article*

# How Many Is Enough? Effect of Sample Size in Inter-Subject Correlation Analysis of fMRI

**Juha Pajula[1] and Jussi Tohka[2,3]**

[1]*Department of Signal Processing, Tampere University of Technology, P.O. Box 553, 33101 Tampere, Finland*
[2]*Department of Bioengineering and Aerospace Engineering, Universidad Carlos III de Madrid, Avenida de la Universidad 30, 28911 Leganes, Spain*
[3]*Instituto de Investigacion Sanitaria Gregorio Marānon, Calle de Doctor Esquerdo 46, 28007 Madrid, Spain*

Correspondence should be addressed to Juha Pajula; juha.pajula@tut.fi

Received 8 September 2015; Revised 9 December 2015; Accepted 14 December 2015

Academic Editor: Thomas DeMarse



Inter-subject correlation (ISC) is a widely used method for analyzing functional magnetic resonance imaging (fMRI) data acquired during naturalistic stimuli. A challenge in ISC analysis is to define the required sample size in the way that the results are reliable. We studied the effect of the sample size on the reliability of ISC analysis and additionally addressed the following question: How many subjects are needed for the ISC statistics to converge to the ISC statistics obtained using a large sample? The study was realized using a large block design data set of 130 subjects. We performed a split-half resampling based analysis repeatedly sampling two nonoverlapping subsets of 10–65 subjects and comparing the ISC maps between the independent subject sets. Our findings suggested that with 20 subjects, on average, the ISC statistics had converged close to a large sample ISC statistic with 130 subjects. However, the split-half reliability of unthresholded and thresholded ISC maps improved notably when the number of subjects was increased from 20 to 30 or more.

## 1. Introduction

Inter-subject correlation (ISC) [1, 2] is a widely used method for detecting and comparing activations in functional magnetic resonance imaging (fMRI) acquired during complex, multidimensional stimuli such as audio narratives, music, or movies [3–9]. Instead of trying to model the stimulus as in the standard general linear model (GLM) based fMRI analysis ISC computes voxel-by-voxel correlations of the subjects' fMRI time courses, assuming that the images have been registered to a common stereotactic space. The activation maps can then be formed by thresholding the average correlation coefficient values. The ISC method has been shown to produce activation maps closely matching those of the standard GLM based analysis when the stimuli are simple and can be modelled [10]. Note, however, that while not using a model time course of the stimulus, ISC expects that all the subjects are exposed to the same stimulus and it is not a method for an analysis of resting state fMRI.

A common challenge in any fMRI group analysis, including ISC analysis, is to define the required number of subjects in such a way that the analysis results are reliable and have enough statistical power, but the costs of the data acquisition are minimized. In principle, a larger sample size provides a more reliable analysis and more statistical power [11, 12]. Obviously, the sample size is not the only factor contributing to reliability (or the statistical power) of the study, but ideally the whole study design should be done to reach the desired limits of statistical power [13–15]. However, between-subject variability in fMRI data is generally much higher than within-subject variability and consequently choosing a large enough sample size is essential [16].

While there are no general methods for the optimal experimental design using naturalistic stimuli, the generalizability of the analysis results, necessarily with a limited sample size, to the population level is an important consideration. Particularly, it is important to know how many subjects are required for a reproducible (or reliable) analysis, so that small

variations in the subject sample do not cause too large variations in the analysis results. This is the question we ask in this paper and to our knowledge it has not been addressed previously in the context of the ISC analysis. Similar studies on the reliability of fMRI group studies with general linear model (GLM) analyses have been reported earlier in [16–18]. All of these studies have concluded that closer to 30 subjects should be included in a group level studies in fMRI data analysis. The sample size issue has been studied also with independent component analysis [19], where the reproducibility of the results was noticed to improve with an increased number of subjects. Critically, David et al. [20] reported that the average number of subjects in their meta-analysis was 13 and 94% of all studies were applied with less than 30 subjects, which suggests that typically fMRI group studies based on GLM might not reach the required level of reliability.

In this study, we examined how the number of subjects included in the study affects the reliability of the statistical ISC maps and the FDR corrected binary thresholded maps. We used a large 130-subject data set with a simple block design task and performed a split-half resampling based analysis (similar to [16]) while varying the number of subjects in each split-half. The resampling procedure was repeated 1000 times. This setup enables us to address the reproducibility of the studies with the maximum of 65 subjects. We compared the statistical ISC maps formed using independent subjects samples and also the thresholded ISC maps. In addition and similarly to [17] we compared statistical ISC maps with the subsets of 130 subjects with the statistical ISC map derived from the whole 130-subject data set.

## 2. Materials and Methods

*2.1. fMRI Data.* The fMRI data used in the preparation of this work were obtained from the ICBM database (https://ida.loni.usc.edu/login.jsp?project=ICBM) in the Image Data Archive of the Laboratory of Neuro Imaging. 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 Neurologic Institute, University of Texas at San Antonio, and the Institute of Medicine, Juelich/Heinrich Heine University, Germany.

We selected all subjects from the ICBM database who had fMRI measurements with the verb generation (VG) task and the structural MR image available. This produced 132 subjects' data set. After a quality check by visual inspection two subjects were discarded due to clear artifacts in their fMRI data. This led to a final data set of 130 subjects: 61 males, 69 females; age range 19–80 years, mean 44.35 years; 117 were right-handed, 10 were left-handed, and 3 were ambidextrous. The data was acquired during the block design VG task (a language task with a visual input) from Functional Reference Battery (FRB) developed by the International Consortium for Human Brain Mapping (ICBM) [21]. The FRB holds a set of behavioral tasks designed to reliably produce functional landmarks across subjects and we have previously used fMRI data extracted from the ICBM FRB database for other experiments [10, 22]. The details of the data and VG task are provided in [10]. The VG task contained the largest number of subjects with fMRI measurements in the ICBM database among the five FRB tasks and therefore we selected it for this study.

The functional data was collected with a 3-Tesla Siemens Allegra fMRI scanner and the anatomical $T_1$ weighted MRI data was collected with a 1.5-Tesla Siemens Sonata scanner. The TR/TE times for the functional data were 4 s/32 ms, with flip angle 90 degrees, pixel spacing 2 mm, and slice thickness 2 mm. The parameters for the anatomical $T_1$ data were 1.1 s/4.38 ms, 15 degrees, 1 mm, and 1 mm, correspondingly.

*2.2. Preprocessing.* The preprocessing of the data was performed with FSL (version 5.0.2.2) from Oxford Centre for Functional Magnetic Resonance Imaging of the Brain, Oxford University, Oxford, UK [23]. The data preprocessing, which was identical to [10], included motion correction with FSL's MCFLIRT and the brain extraction for the functional data was done with FSL's BET [24]. The fMRI images were temporally high-pass filtered with a cutoff period of 60 s and the spatial smoothing was applied with an isotropic three-dimensional Gaussian kernel with the full-width half-maximum (FWHM) 5 mm in each direction. The brain extraction of the structural $T_1$ images was also performed by BET, but this was done separately from the main procedure for each $T_1$ weighted images as the parameters of BET required individual tuning for the images.

The image registration was performed with FSL Linear Registration Tool (FLIRT) [25, 26] in two stages. At the beginning, the skull-stripped functional images were aligned (6 degrees of freedom, full search) to the skull-stripped high-resolution $T_1$ weighted image of the same subject, and then the results were aligned to the standard (brain only) 2 mm ICBM-152 template (12 degrees of freedom, full search).

*2.3. ISC Analyses.* All of the ISC analyses were computed with ISCtoolbox for Matlab [2]. ISCtoolbox computes the ISC statistic by first computing Pearson's correlations between the corresponding time series of all subject-pairs. Then, to obtain the final multisubject test statistic, correlation values of all subject-pairs are combined into a single ISC statistic by averaging. This is the ISC statistical map.

The statistical inference was accomplished by a fully nonparametric voxel-wise resampling test implemented in the ISCtoolbox [27]. The resampling test constructs the null-distribution of the ISC values by circularly shifting the time series of each subject by a random amount. This test resembles the circular block bootstrap test [28] and it accounts for temporal correlations inherent to fMRI data. For a more detailed description of the test, we refer to [29]. For thresholding each ISC map, the resampling distribution was approximated with 10 000 000 realizations, sampling randomly across the brain voxels for each realization and generating a new set of time-shifts (one for each subject) for each realization. The resulting $p$-values were corrected voxel-wise over the whole brain using a false discovery rate (FDR) based multiple comparisons correction [30].

*2.4. Experimental Procedure.* We performed a split-half resampling type of the analysis for the ISC method. The process consisted of randomly drawing (without replacement) two independent subsets of $P = 10, 15, \ldots, 65$ subjects from the total pool of 130 subjects. Then, the full ISC analysis (including resampling distribution approximation and computation of corrected thresholds) was performed for both subsets and the full ISC analysis results from both sets were saved. This process was repeated 1000 times meaning that the ISC analysis was performed separately and independently 2000 times for each number of subjects $P = 10, 15, \ldots, 65$.

We compared the ISC statistical maps of the split-half analysis with the following criteria.

(1) Pearson's correlation coefficient $C_n$ for comparing the nonthresholded statistical maps was defined as

$$C_n = \frac{1}{K-1}\sum_{k=1}^{K}\left(\frac{\bar{l}_k - \overline{L}}{s_{\bar{l}}}\right)\left(\frac{\bar{r}_k - \overline{R}}{s_{\bar{r}}}\right), \quad (1)$$

where $K$ is the total number of brain voxels in the volume. $\bar{l}_k$ and $\bar{r}_k$ are the two ISC statistics of the $k$th voxel, respectively. $\overline{L}$ and $\overline{R}$ are the sample means of $\{\bar{l}_k\}$ and $\{\bar{r}_k\}$ across the brain volume, and $s_{\bar{l}}$ and $s_{\bar{r}}$ are the standard deviations of $\{\bar{l}_k\}$ and $\{\bar{r}_k\}$ across the brain volume. The final measure was computed by averaging the correlation measures $C_n$ according to

$$C_{\text{avg}} = \frac{1}{N}\sum_{n=1}^{N} C_n, \quad (2)$$

where $N$ is the number of resampling replications, which was 1000 in this study.

(2) The mean absolute error (MAE) between paired ISC maps was defined according to

$$M_n = \frac{1}{K}\sum_{k=1}^{K}\left|\bar{r}_k - \bar{l}_k\right|, \quad (3)$$

where $K$ is the total number of brain voxels in the volume. $\bar{r}_k$ and $\bar{l}_k$ are the two ISC statistics of the $k$th voxel, respectively. The final measure was computed by averaging the MAE measures $M_n$ according to

$$M_{\text{avg}} = \frac{1}{N}\sum_{n=1}^{N} M_n, \quad (4)$$

where $N = 1000$ is the number of resampling replications.

We used Dice index to compare the thresholded paired binary ISC activation maps [31]. The justification for the use of Dice index can be found in [10]. The Dice index between two sets ($A_n$ and $B_n$, $n = 1, \ldots, 1000$ refers to resampling replication) of activated voxels was defined as

$$D_n = \frac{2\left|A_n \cap B_n\right|}{\left|A_n\right| + \left|B_n\right|} \quad (5)$$

and it takes values between 0 and 1. The tested thresholds were corrected with a false discovery rate (FDR) over the whole brain using $q = 0.05$, $q = 0.01$, and $q = 0.001$ (no correlation assumptions). The Dice indexes were computed for 1000 times for each number of subjects and the reported average Dice index was computed by averaging 1000 Dice indexes $D_n$ in the same way as with correlation and MAE measures.

The Dice index defines the binary similarity between two binary images and it can be categorized with Landis and Koch categorization for Kappa coefficients [10]. According to [32] the categories are

(i) ≤0, no agreement,

(ii) 0–0.2, slight agreement,

(iii) 0.2–0.4, fair agreement,

(iv) 0.4–0.6, moderate agreement,

(v) 0.6–0.8, substantial agreement,

(vi) 0.8–1.0, almost perfect agreement.

As Landis and Koch themselves note these categories are highly subjective [32] but are maybe useful as a reference.

Similarly to [17], we considered how fast the statistic maps converge to a large sample statistic map with 130 subjects. For this, we repeated Pearson's correlation analyses described above by comparing statistic maps resulting from resampling to the statistic map obtained using all 130 subjects as in (1) and averaging over 2000 resampling iterations. More specifically, $\bar{r}$ and $\overline{R}$ in (1) were from the same statistic map with 130 subjects and in (2) $N$ was then 2000. We computed also the sensitivity and specificity of thresholded ISC maps by using the thresholded 130 subjects ISC statistic with the corresponding threshold ($q = 0.05$, $q = 0.01$, and $q = 0.001$ with no correlation assumptions) as the ground truth. The final sensitivity and specificity (for each number of subjects) were averaged from 2000 sensitivity and specificity measures that resulted from 1000 split-half resampling replications.

*2.5. Implementation.* This study was computationally demanding. For each number of subjects, 2000 ISC analyses with 10 000 000 realizations for corrected thresholds were computed. This was repeated with 12 different numbers of subjects and the whole analysis required 24 001 ISC analyses (one extra analysis was for the whole data set of 130 subjects). For implementing the computations, parallel computing environment Merope of Tampere University of Technology, Finland, was used. It has nodes running on HP ProLiant SL390s G7 equipped with Intel Xeon X5650 CPU 2,67 GHz and minimum of 4 GB RAM/core. The used grid engine was Slurm. The equivalent computing time would have been 4.75 years if they had been computed with a single high end CPU.

## 3. Results

Figure 1 presents the thresholded (voxel-wise FDR corrected over the whole brain $q = 0.001$) results from the ISC analysis with the whole 130 subjects' data set. Significant ISC values were found around occipital and temporal lobes, lateral occipital cortex, and paracingulate gyrus as well as on

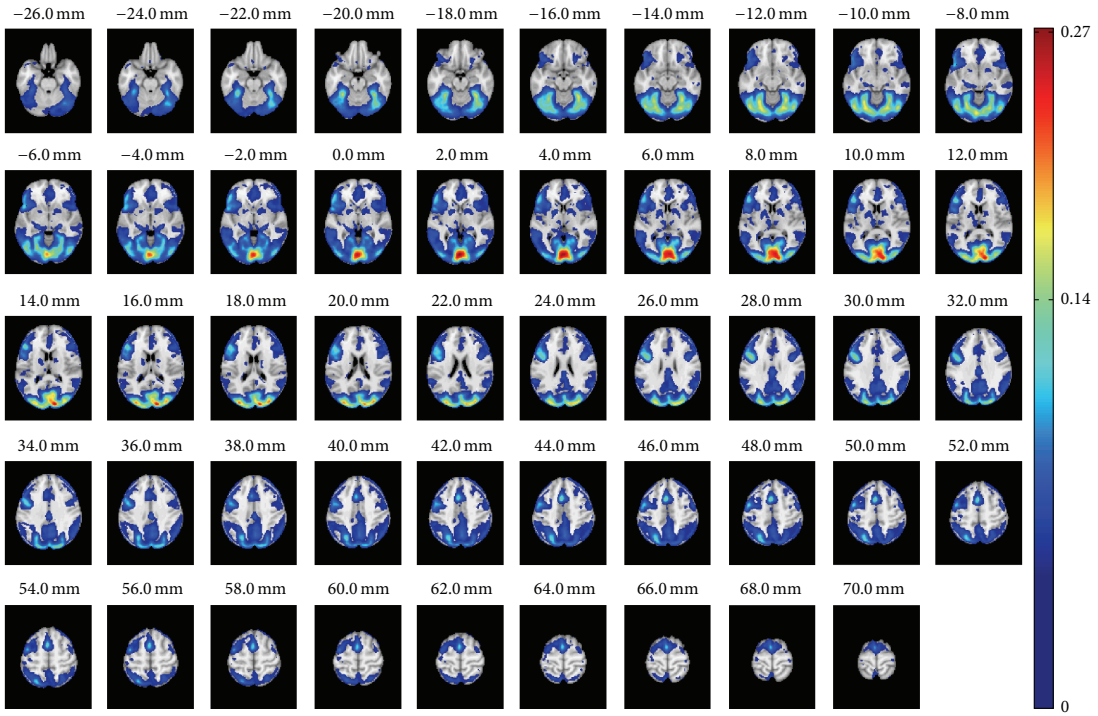



Figure 1: The ISC analysis based on 130 subjects. The figure presents the axial slices of the ISC analysis results of the whole 130 subjects' data set after applying FDR corrected $q = 0.001$ thresholding. The full statistical map is visible and available in NeuroVault: http://www.neurovault.org/collections/WTMVBEZP/images/11576/.

middle frontal and inferior frontal gyri. The 130-subject ISC map was highly similar to ISC map presented earlier with partially the same data but with smaller number of subjects ($P = 37$) [10]. The most noticeable difference compared with the 37-subject analysis was that with 130 subjects a larger number of voxels survived from the threshold and significant ISCs formed a more symmetric pattern over the hemispheres. One specific note concerning ISC map of Figure 1 is in order: There appears to be an artifact, which can be seen as a thin activation line in the left frontal cortex (e.g.) in the axial slice $z = 50$ mm. The investigation of the data at that location revealed a slight signal drop in time series of majority of subjects, buried under the noise in any single subject data, which increased ISC values with the large data set to level of statistical significance. The temporal location of the drop was in the middle of the time series ($t = 172$ s, while not counting the stabilization volumes). The statistical ISC map from 130 subjects is available in the NeuroVault service [33] at http://www.neurovault.org/collections/WTMVBEZP/images/11576/.

Figure 2 presents the correlation criteria resulting from the split-half resampling analysis. Figure 2(a) presents the average correlation $C_{\text{avg}}$ (2) and Figure 2(b) presents the corresponding variance of $C_n$, $n = 1, \ldots, 1000$ (see (1)). As expected the average correlation between nonoverlapping samples increased when the number of subjects increased and, at the same time, the variance decreased. The average correlation curve was not linear with respect to the number of subjects and stabilized after 30 subjects finally reached the value of 0.95 as the number of the subjects reached the value of 65.

Figure 3 presents the MAE criteria resulting from the split-half resampling analysis. Figure 3(a) presents the average MAE (3) and Figure 3(b) presents the corresponding variance of $M_n$, $n = 1, \ldots, 1000$ (see (3)). Again, as expected, the average MAE between nonoverlapping samples decreased when the number of subjects increased and at the same time the variance decreased, largely replicating the correlation based curves in Figure 2. With 20 subjects the average MAE was 0.015 and with 30 subjects it was 0.011 indicating that, on average, ISC with 20 or 30 subjects already provided a high degree of reproducibility when averaged over the whole brain. However, this does not reveal whether there were variations in the reproducibility in voxel-wise ISC values across the brain. Figure 4 presents how the MAEs were distributed over the brain volume with 30 subjects. We note that the spatial shape of MAE distribution across the brain was highly similar to all numbers of subjects, and only the magnitude of the average MAE changed. Comparing Figure 4 with Figure 1 revealed that the highest

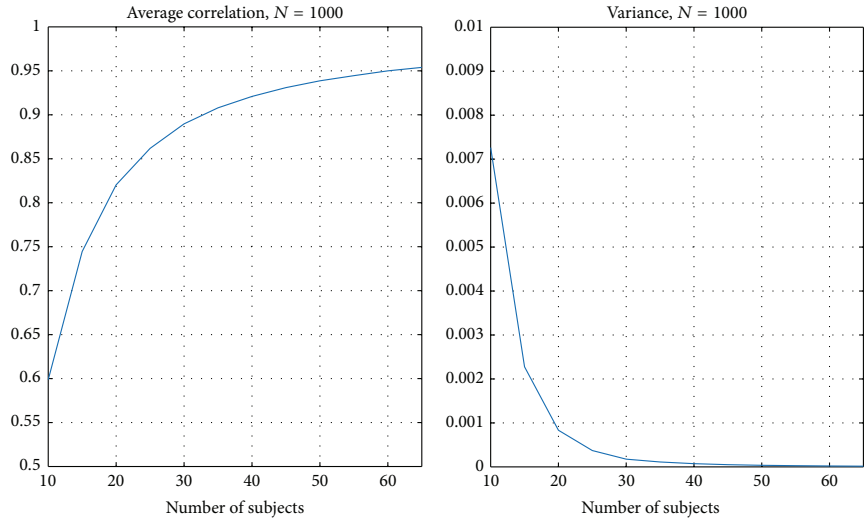



Figure 2: Average correlation $C_{\text{avg}}$ over 1000 resampling replications. (a) presents the average correlation over and (b) the corresponding variance of $C_n$, $n = 1, \ldots, 1000$. The correlation increased when the sample size increased and at the same time the variance decreased.

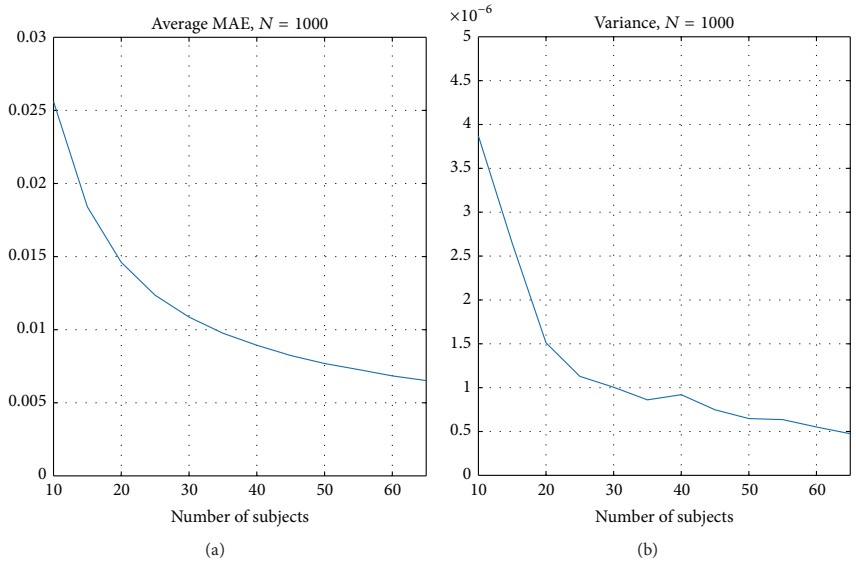



Figure 3: Average MAE $M_{\text{avg}}$ over 1000 resampling replications. (a) presents the average MAE and (b) the corresponding variance of $M_n$, $n = 1, \ldots, 1000$.

variations in the ISCs coincided with the highest ISC values. The three-dimensional MAE maps with all numbers of subjects are available in the NeuroVault service [33] at http://www.neurovault.org/collections/WTMVBEZP/.

Figure 5 presents Dice indexes over the 1000 resampling replications. Figure 5(a) presents the average of Dice indexes $D_n$ for three threshold levels (voxel-wise FDR corrected over the whole brain with $q = 0.05$ (blue), $q = 0.01$ (red), and $q = 0.001$ (yellow)). Figure 5(b) presents the corresponding variance of the Dice indexes $D_n$. Again, as expected the Dice similarity between thresholded ISC maps increased when the number of subjects increased and the variance of Dice

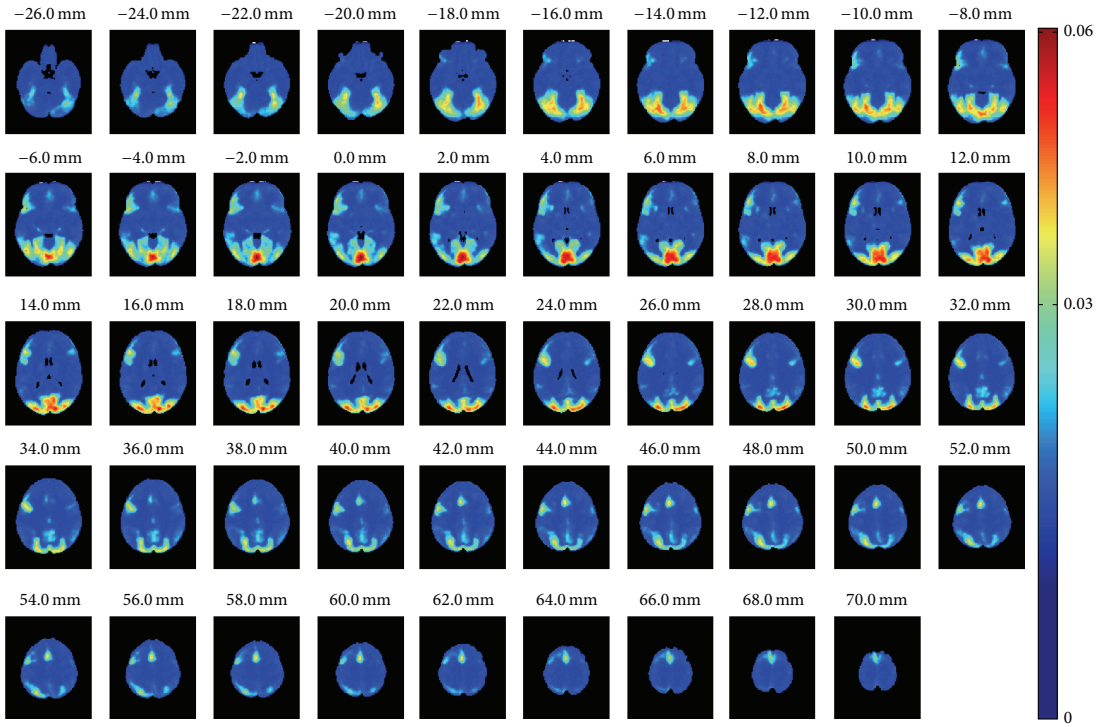



Figure 4: Average MAE computed voxel-wise over 1000 resampling replications with 30 subjects. The average voxel-wise MAE map had similar spatial shape with every tested number of subjects. The only clear difference was the magnitude of MAE values.

indexes decreased when the number of subjects increased. Based on Figure 5(a), it is noticeable that more conservative thresholds required slightly more subjects to stabilize. The most liberal threshold $q = 0.05$ had all average Dice indexes within the category "substantial agreement" but stays under the level of "almost perfect agreement" even with 65 subjects. The more conservative $q = 0.01$ reached the "almost perfect agreement" level with 45 subjects and $q = 0.001$ had the Dice index over the required 0.8 already with 35 subjects.

Figure 6 presents the average of correlation when ISC maps with resampled subsets of subjects were compared with the ISC map computed with the whole set of 130 subjects (average over 2000 resampling replications). In Figure 6, (a) presents the average correlation and (b) presents the corresponding variance. Again, the correlation increased when the number of subjects increased and the variance decreased when the number of subjects increased. The variance was close to zero and the correlation to the full 130-subject ISC map was 0.95 with 30 subjects. The sensitivity and specificity curves, using 130-subject thresholded ISC map as the ground truth, are presented in Figure 7. The sensitivity increased when the number of subjects increased and the specificity stayed close to 1 with all numbers of subjects. Figure 7 also shows that the more liberal the threshold the higher the sensitivity value at a slight expense of the specificity value.

## 4. Discussion

In this study, we evaluated the reliability of the ISC analysis for fMRI data and studied the effect of the sample size on the reliability of the ISC analysis. This was accomplished by using a split-half resampling based design, similar to that of [16]. We randomly sampled two nonoverlapping subsets of subjects from the 130-subject ICBM-fMRI data set with a verb generation task. We iterated the paired resampling procedure 1000 times for each number of subjects varying from 10 to 65 and compared the ISC analysis results obtained based on two nonoverlapping subsets of subjects. We compared both the raw ISC statistic maps and the thresholded statistical maps.

Previously, we have validated the ISC analysis against a gold standard set by GLM analysis in [10] and investigated the effect of smoothing to the ISC analysis results in [22]. Both of these studies used a relatively large fMRI data set of 37 subjects, which was larger than the data sets typically applied in the naturalistic stimulus experiments. Therefore, in addition to the question concerning the reliability of the ISC analysis, it was important to study how many subjects are needed for the ISC analysis in order for statistical maps to stabilize. When comparing the ISC results of our earlier study applied for 37 subjects [10] with the current study of 130 subjects, it is not surprising that the statistical power of

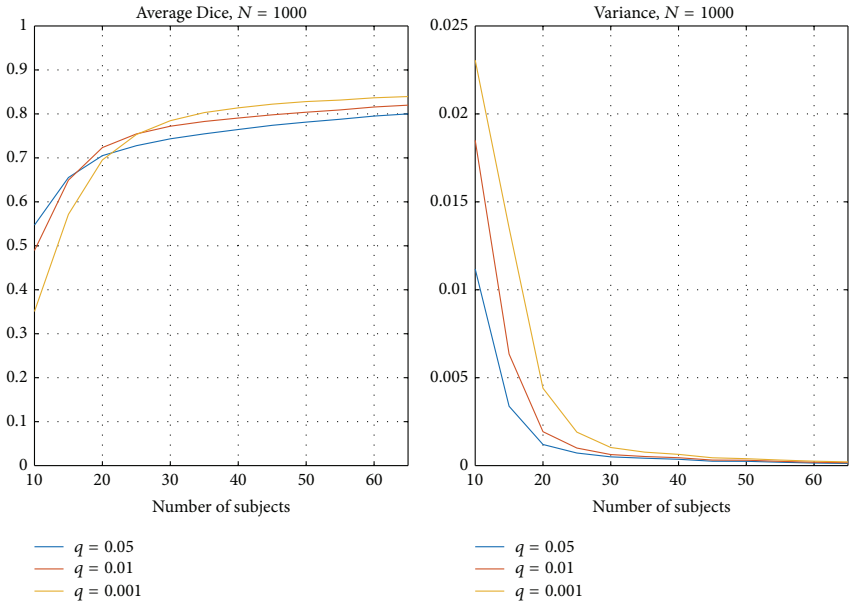



Figure 5: Average Dice index over 1000 resampling replications with three FDR levels: $q = 0.05$, $q = 0.01$, and $q = 0.001$. (a) presents the average Dice indexes $D_n$ over 1000 replications and (b) presents the corresponding variance. The curve corresponding to the most conservative threshold $q = 0.001$ (yellow) shows that more subjects are required for greater similarity after applying the threshold to the data. The more liberal thresholds $q = 0.01$ (in red) and $q = 0.05$ (in blue) required fewer subjects to stabilize than the most conservative threshold $q = 0.001$ (yellow) but on the other hand the highest similarity was reached with the most conservative threshold.

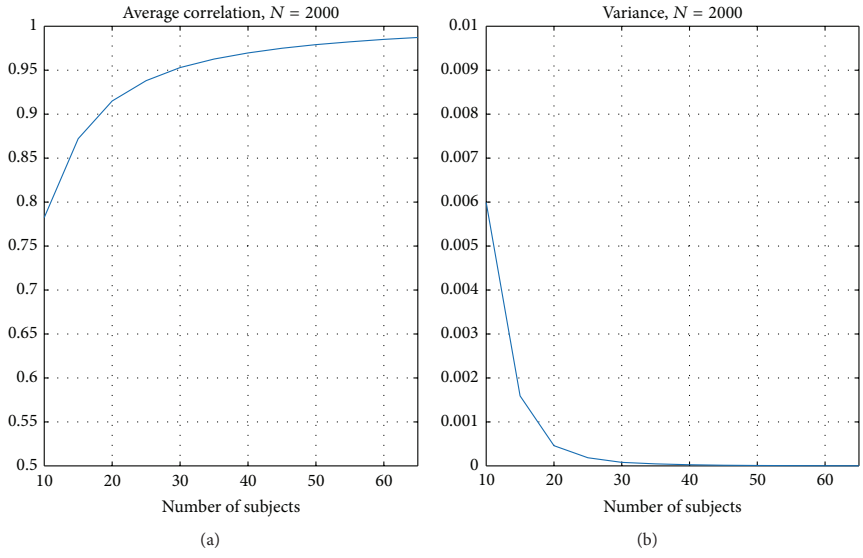



Figure 6: Average correlation comparing subsampled ISC maps with the ISC statistic map of the whole 130 subjects. (a) presents the average correlation over 2000 replications and (b) presents the corresponding variance. Again, the correlation increased when the number of subjects increased. With 30 subjects or more, the average correlation was greater than 0.95 and the variance was less than 0.0002.

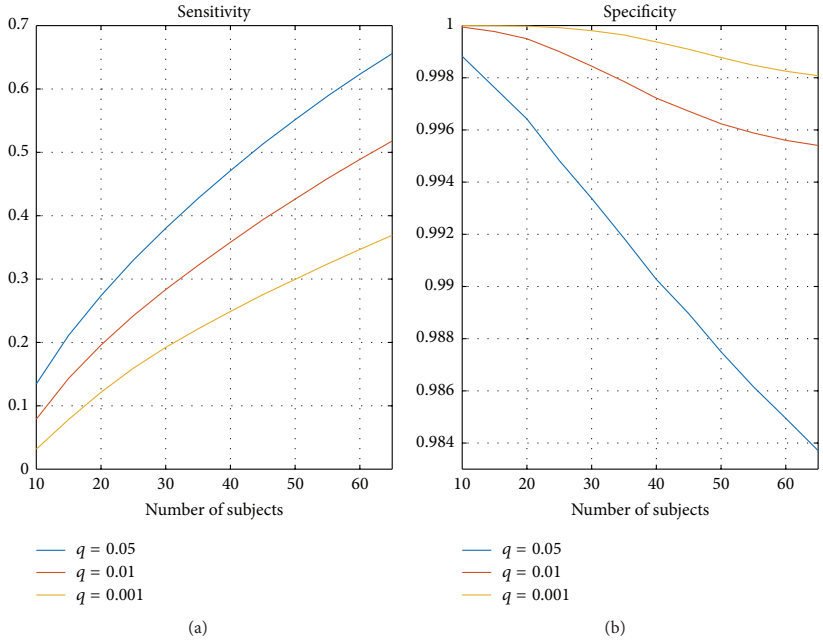



Figure 7: Average sensitivity and specificity from thresholded binary maps compared with the thresholded ISC statistic map of the full 130-subject sample size. (a) presents the average sensitivity over $N = 2000$ replications and (b) presents the corresponding specificity. The sensitivity increased when the number of subjects increased. The specificity was close to 1 with conservative thresholds and even with the most liberal threshold with $q = 0.05$ the specificity was over 0.98 with any number of subjects.

the analysis had been increased with the increased number of subjects; that is, the activated areas were larger with 130 subjects.

When examining the voxel-wise MAE values shown in Figure 4, it was clear that the largest MAE coincided with the strongest ISCs in Figure 1. This is an interesting phenomenon because purely technically the sample variance of the correlation coefficients decreases when the true correlation increases [34]. Thus, the increase in the voxel-wise MAE values with the average ISC means that subject-pair-to-subject-pair variability of ISC generally increases with increasing average ISC. We note that this phenomenon was independent of the applied sample size and particularly all the MAE maps, uploaded to http://www.neurovault.org/collections/WTMVBEZP/, were virtually identical except for the scale of MAE values.

The data in this study was based on a traditional block design stimulus while the ISC analysis is typically applied for fMRI data with naturalistic stimuli. This choice was made out of necessity since no large enough naturalistic stimulation studies exist. In principle, the block design data might have limitations not to reveal all sources of variation involved in the ISC analysis. In particular, the data involves the replication of the same task/stimulus pattern and therefore might lead to positively biased reliability measures for the naturalistic stimulation fMRI. On the other hand, we have shown that ISC is applicable to block design data [10, 22], which partially justifies the use of block design data. Also, it should be noted that the naturalistic stimuli themselves are highly varied and therefore using one type of naturalistic stimuli might have the same limitations as our use of the block design stimulus. Due to high computational demands of the analysis, we chose to only consider fMRI time series of certain length albeit the minimal length of the time series is an important consideration especially to the so-called time-window ISC analysis [2, 35]. To render the analysis more targeted towards the naturalistic stimulation studies, where one may stipulate that individual reactions to the used stimuli may differ more among the participants than with traditional fMRI setups, we included subjects with a wide age range spanning from 19 to 80 years to our analysis (see [36] for the age-effects on the verb generation task). We also included left-handed and ambidextrous subjects, which may be slightly controversial due to greater prevalence of right-lateralized language among the left-handed subjects (see [37] and references therein). However, most left-handers have left-lateralized language and there exist multiple other reasons not to exclude left-handers from neuroimaging studies [37].

The results of our split-half resampling analysis indicated that 20 subjects were the minimum number of subjects to achieve somehow reproducible ISC statistical maps, but for a good reproducibility it would be preferred to have 30 subjects or more. With 20 subjects, the correlation measure ($C_{\text{avg}}$ (2)) was 0.82 (see Figure 2), the average MAE ($M_{\text{avg}}$ (4)) was 0.015 (see Figure 3), and the average Dice coefficient was 0.71,

0.72, and 0.70 for $q = 0.05$, 0.01, and 0.001, respectively (see Figure 5). When the number of subjects was below 20, our analysis indicated weak reproducibility (see Figures 2, 5, and 3). The reproducibility improved clearly when the number of subjects was incremented from 20 to 30 ($C_{\text{avg}}$ increased to 0.89, $M_{\text{avg}}$ decreased to 0.010, and the average Dice coefficient increased to 0.74, 0.77, and 0.78, resp.), but adding more than 30 subjects did not improve the reproducibility so steeply any more. The average correlation between the subsample ISC statistical map and the whole sample ISC statistical map was 0.92 already with 20 subjects and 0.95 with 30 subjects indicating that ISC statistics maps converged rapidly towards the whole sample ISC maps. As seen in Figure 7, the average sensitivity of the ISC detection, when compared to the thresholded ISC map with 130 subjects, was not particularly high even with 30 subjects. However, the specificity of ISC detections was close to 1 indicating that nearly all voxels detected with small sample sizes were also detected in the full 130-subject sample. This is not surprising and largely replicates the findings for the GLM based analysis of the event related GO/NOGO task in [17]. Also, our results were in line with the studies on the reproducibility in the GLM based analysis [16] recommending that more than 20 or even more than 30 subjects should be used in fMRI group analysis. Obviously, how many subjects are required for a particular fMRI study ultimately depends on the experiment and the guidelines provided by this work may not be applicable for all experiments involving ISC analysis.

## 5. Conclusions

We studied the effect of sample size for ISC analysis to determine how many subjects are needed for a reliable ISC analysis. We also investigated how small sample is enough for the ISC statistic to converge to ISC statistic obtained with a large sample. We found that with 20 subjects the ISC statistics were converged close to a large 130 subjects' ISC statistic. However, the reliability of unthresholded and thresholded maps improved notably when the number of subjects was increased to 30 subjects, which indicated that with this data 30 subjects or more should be used with ISC analysis for truly reproducible results. Finally, we emphasize that the required number of subjects depends on the specific characteristic of the experiment, including the expected effect size.

## Additional Material

Three-dimensional statistical maps are available in the NeuroVault service: http://www.neurovault.org/collections/WTMVBEZP/.

## Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

## Acknowledgments

This project has received funding from Universidad Carlos III de Madrid, the European Union's Seventh Framework Programme for Research, Technological Development and Demonstration under Grant Agreement no. 600371, el Ministerio de Economía y Competitividad (COFUND2013-40258), and Banco Santander. Computing resources for the study was provided by the Signal Processing Department of Tampere University of Technology with the Merope computing cluster of Tampere University of Technology. Data collection and sharing for this project were provided by the International Consortium for Brain Mapping (ICBM; Principal Investigator: John Mazziotta, M.D., Ph.D.). 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.

## References

[1] U. Hasson, Y. Nir, I. Levy, G. Fuhrmann, and R. Malach, "Intersubject synchronization of cortical activity during natural vision," *Science*, vol. 303, no. 5664, pp. 1634–1640, 2004.

[2] J. P. Kauppi, J. Pajula, and J. Tohka, "A versatile software package for inter-subject correlation based analyses of fMRI," *Frontiers in Neuroinformatics*, vol. 8, article 2, 2014.

[3] R. Schmälzle, F. E. Häcker, C. J. Honey, and U. Hasson, "Engaged listeners: shared neural processing of powerful political speeches," *Social Cognitive and Affective Neuroscience*, vol. 10, no. 8, pp. 1137–1143, 2015.

[4] G. Bernardi, L. Cecchetti, G. Handjaras et al., "It's not all in your car: functional and structural correlates of exceptional driving skills in professional racers," *Frontiers in Human Neuroscience*, vol. 8, article 888, 2014.

[5] J. M. Lahnakoski, E. Glerean, I. P. Jääskeläinen et al., "Synchronous brain activity across individuals underlies shared psychological perspectives," *NeuroImage*, vol. 100, pp. 316–324, 2014.

[6] R. Schmälzle, F. Häcker, B. Renner, C. J. Honey, and H. T. Schupp, "Neural correlates of risk perception during real-life risk communication," *Journal of Neuroscience*, vol. 33, no. 25, pp. 10340–10347, 2013.

[7] D. A. Abrams, S. Ryali, T. Chen et al., "Inter-subject synchronization of brain responses during natural music listening," *European Journal of Neuroscience*, vol. 37, no. 9, pp. 1458–1469, 2013.

[8] I. P. Jääskeläinen, K. Koskentalo, M. H. Balk et al., "Inter-subject synchronization of prefrontal cortex hemodynamic activity during natural viewing," *The Open Neuroimaging Journal*, vol. 2, no. 1, pp. 14–19, 2008.

[9] Y. Golland, S. Bentin, H. Gelbard et al., "Extrinsic and intrinsic systems in the posterior cortex of the human brain revealed during natural sensory stimulation," *Cerebral Cortex*, vol. 17, no. 4, pp. 766–777, 2007.

[10] J. Pajula, J.-P. Kauppi, and J. Tohka, "Inter-subject correlation in fMRI: method validation against stimulus-model based analysis," *PLoS ONE*, vol. 7, no. 8, Article ID e41196, 2012.

[11] J. Suckling, A. Barnes, D. Job et al., "Power calculations for multicenter imaging studies controlled by the false discovery rate," *Human Brain Mapping*, vol. 31, no. 8, pp. 1183–1195, 2010.

[12] J. E. Desmond and G. H. Glover, "Estimating sample size in functional MRI (fMRI) neuroimaging studies: statistical power analyses," *Journal of Neuroscience Methods*, vol. 118, no. 2, pp. 115–128, 2002.

[13] B. Maus, G. J. P. van Breukelen, R. Goebel, and M. P. F. Berger, "Optimal design of multi-subject blocked fMRI experiments," *NeuroImage*, vol. 56, no. 3, pp. 1338–1352, 2011.

[14] J. A. Mumford and T. E. Nichols, "Power calculation for group fMRI studies accounting for arbitrary design and temporal autocorrelation," *NeuroImage*, vol. 39, no. 1, pp. 261–268, 2008.

[15] S. Hayasaka, A. M. Peiffer, C. E. Hugenschmidt, and P. J. Laurienti, "Power and sample size calculation for neuroimaging studies by non-central random field theory," *NeuroImage*, vol. 37, no. 3, pp. 721–730, 2007.

[16] B. Thirion, P. Pinel, S. Mériaux, A. Roche, S. Dehaene, and J.-B. Poline, "Analysis of a large fMRI cohort: statistical and methodological issues for group analyses," *NeuroImage*, vol. 35, no. 1, pp. 105–120, 2007.

[17] K. Murphy and H. Garavan, "An empirical investigation into the number of subjects required for an event-related fMRI study," *NeuroImage*, vol. 22, no. 2, pp. 879–885, 2004.

[18] B. B. Zandbelt, T. E. Gladwin, M. Raemaekers et al., "Within-subject variation in BOLD-fMRI signal changes across repeated measurements: quantification and implications for sample size," *NeuroImage*, vol. 42, no. 1, pp. 196–206, 2008.

[19] N. W. Churchill, G. Yourganov, and S. C. Strother, "Comparing within-subject classification and regularization methods in fMRI for large and small sample sizes," *Human Brain Mapping*, vol. 35, no. 9, pp. 4499–4517, 2014.

[20] S. P. David, J. J. Ware, I. M. Chu et al., "Potential reporting bias in fMRI studies of the brain," *PLoS ONE*, vol. 8, no. 7, Article ID e70104, 2013.

[21] J. Mazziotta, A. Toga, A. Evans et al., "A probabilistic atlas and reference system for the human brain: International Consortium for Brain Mapping (ICBM)," *Philosophical Transactions of the Royal Society of London Series B: Biological Sciences*, vol. 356, no. 1412, pp. 1293–1322, 2001.

[22] J. Pajula and J. Tohka, "Effects of spatial smoothing on inter-subject correlation based analysis of FMRI," *Magnetic Resonance Imaging*, vol. 32, no. 9, pp. 1114–1124, 2014.

[23] S. M. Smith, M. Jenkinson, M. W. Woolrich et al., "Advances in functional and structural MR image analysis and implementation as FSL," *NeuroImage*, vol. 23, supplement 1, pp. S208–S219, 2004.

[24] S. M. Smith, "Fast robust automated brain extraction," *Human Brain Mapping*, vol. 17, no. 3, pp. 143–155, 2002.

[25] M. Jenkinson and S. Smith, "A global optimisation method for robust affine registration of brain images," *Medical Image Analysis*, vol. 5, no. 2, pp. 143–156, 2001.

[26] M. Jenkinson, P. Bannister, M. Brady, and S. Smith, "Improved optimization for the robust and accurate linear registration and motion correction of brain images," *NeuroImage*, vol. 17, no. 2, pp. 825–841, 2002.

[27] J.-P. Kauppi, I. P. Jääskeläinen, M. Sams, and J. Tohka, "Inter-subject correlation of brain hemodynamic responses during watching a movie: localization in space and frequency," *Frontiers in Neuroinformatics*, vol. 4, article 5, 2010.

[28] D. N. Politis and J. P. Romano, "A circular block-resampling procedure for stationary data," in *Exploring the Limits of Bootstrap*, R. LePage and L. Billard, Eds., pp. 263–270, John Wiley & Sons, New York, NY, USA, 1992.

[29] J. Tohka, "Non-parametric test for inter-subject correlations," Tech. Rep., 2015, https://www.nitrc.org/docman/view.php/947/2017/parametric-test-inter.pdf.

[30] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: a practical and powerful approach to multiple testing," *Journal of the Royal Statistical Society B: Methodological*, vol. 57, no. 1, pp. 289–300, 1995.

[31] L. R. Dice, "Measures of the amount of ecologic association between species," *Ecology*, vol. 26, no. 3, p. 297, 1945.

[32] J. R. Landis and G. G. Koch, "The measurement of observer agreement for categorical data," *Biometrics*, vol. 33, no. 1, pp. 159–174, 1977.

[33] K. J. Gorgolewski, G. Varoquaux, G. Rivera et al., "NeuroVault.org: a repository for sharing unthresholded statistical maps, parcellations, and atlases of the human brain," *NeuroImage*, vol. 124, pp. 1242–1244, 2016.

[34] A. L. Bowley, "The standard deviation of the correlation coefficient," *Journal of the American Statistical Association*, vol. 23, pp. 31–34, 1928.

[35] L. Nummenmaa, E. Glerean, M. Viinikainen, I. P. Jääskeläinen, R. Hari, and M. Sams, "Emotions promote social interaction by synchronizing brain activity across individuals," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 109, no. 24, pp. 9599–9604, 2012.

[36] J. Persson, C.-Y. C. Sylvester, J. K. Nelson, K. M. Welsh, J. Jonides, and P. A. Reuter-Lorenz, "Selection requirements during verb generation: differential recruitment in older and younger adults," *NeuroImage*, vol. 23, no. 4, pp. 1382–1390, 2004.

[37] R. M. Willems, L. van der Haegen, S. E. Fisher, and C. Francks, "On the other hand: including left-handers in cognitive neuroscience and neurogenetics," *Nature Reviews Neuroscience*, vol. 15, no. 3, pp. 193–201, 2014.

Tampereen teknillinen yliopisto
PL 527
33101 Tampere

Tampere University of Technology
P.O.B. 527
FI-33101 Tampere, Finland

ISBN 978-952-15-3721-9
ISSN 1459-2045