Practical and accurate approaches to statistical significance and power for fMRI.

Joke Durnez

Promotor: Prof. dr. Beatrijs Moerkerke
Co-promotor: Prof. dr. Thomas E. Nichols

Proefschrift ingediend tot het behalen van de academische graad van Doctor in de Psychologie

2015

Table of Contents

Acknowledgements

"Every man can, if he so desires, become the sculptor of his own brain" - Santiago Ramón y Cajal

We made it! It is such a pleasure to finally see all these years of work coming together and I'm so stoked to see what the future now has to offer me. But instead of looking forward, I'm going to start by looking backwards. Because when I say we made it, I literally mean **WE** made it. This thesis would not have been possible without the many help and support I got over the last few years.

First of all, I'd like to thank Bieke, my promotor. I remember the first time I ever talked to you. I said your research was so simple and I meant that as a compliment. Now I realize that it takes a lot of work, thought and effort to write simple ideas, so forgive me for my first comment. You were a great teacher, you gave me the freedom to find my own way, but you put just enough pressure upon me to make sure I continued. You pushed me to search deeper, to write better, and to understand everything that we found, even if we sometimes had to look for months. You were a great supervisor, not only professionally, but also personally. I remember that time you nearly punched two guys for harassing me. And the other time when I wasn't able to work anymore for months. You pushed me just enough to keep me going, but at the same time you patiently waited until I would be back. Thank you so much for that!

Also big thanks to Tom, copromotor. You were my hero before I got to know you, I was so happy you accepted me to spend some months in Warwick. These months were probably the time in my life where I maximized the amount of knowledge gathered per time unit. Even though I'm the psychologist and you're a statistician, it was in Warwick where I started to look at the data as 'human' instead of an array of numbers. It was amazing how much I learnt from our lunch meetings together! I'm happy we kept working together afterwards, and I'm grateful for all the interesting people I met.

Next I'd like to thank the members of my guidance committee, Els,

Ruth, Tom and Yves. Your comments were always helpful and you will see that lots of them are included in this thesis. Thank you for all the feedback and support.

Isabelle, you are truly a great person. Always there to help. With everything. You know I'm secretly a bit jealous of your work; probably that is because you are so great at what you are doing. And besides work you were always in for a laugh or complaint, ready to solve both professional and personal problems.

And then our little group of fMRI stats people, thank you! Sanne, you went from unknown to classmate, to friend and colleague. How joyous it was to discover the weird and wonderful world of fMRI analysis. How horrifying it was to discover the dreadful idea of bootstrapping data. Thanks for undergoing that together. Unfortunately, after 11 years our paths separate. Good luck with your work and your beautiful family.

Jasper, what a pleasure it is to work with you. You work so fast, man! And what a pleasure it is just to talk with you. And have drinks, dinner and go to conferences and have the most amazing time together. I hope we will someday get to share an office again.

Ruth, Han and Freya, thanks for the nice talks on brains and science.

As you can see, I didn't only meet colleagues at work, I made some great friends as well. The whole story of my PhD started with you, Celina. You were my first office mate, you told me all the little details on the department, and you led me through the administrative web of our faculty. But you were much more. We shared laughter and tears. I'm happy you still wanted to be my friend after all the complaining I probably did. Amelie, I'm sorry for my many comments on thirty-somethings. I understand now that being around thirty years old (which we will remain forever now) is the best time of your life :).

The story of my PhD ended with you, Lara. Apparently I behaved arrogant at first, but I was actually shy in the beginning. Luckily we spent quite enough time in one office to overcome that weird period. You're one of my best friends now, and I feel we can tell each other everything. We discussed the smallest details on our underwair (the cotton part), but we also talked about the biggest life decisions. You literally have seen almost all of my love life in the few years we've spent together. It was amazing how compassionate you were. Your positivism will make me smile for the

rest of my life.

Everyone else I shared an office with: Bjorn, you are probably never going to read this but it was a pleasure to get to know you. Ines, good luck with your fabulous new house, husband (and future). Leonard, you'll make a great dad, take care of the little one.

The gym team of the department: Haeike, Lara, Jasper. Thank you so much for pushing me to go to the gym. Thanks for undergoing many Azzedine's together. Thanks for understanding when I quit to focus on my thesis. And thanks for the nice moments outside of the gym as well :).

Finally I'd like to thank the other members of the department as well. For the nice tea breaks, the amazing dinners and the really cool team buildings we had together. It was fantastic to work in such a fast growing (not only in number) environment.

Sincerely thank you to my friends to make life more than work alone, who remind me to what is the most important: friendship.

Tibbie, sometimes you just have to let yourself go. And that's what we often did together. Singing teenager songs in your minivan to France, shouting/laughing the hell out of our bodies on the Space mountain in Paris, sweating in the cold on our skis in Switzerland, get tanned/sunburned in Barcelona, sleeping in parks in Protugal, or just have an adult conversation at home. We sometimes don't see or hear each other for months, but that's usually made up in minutes. Thanks for taking the pressure off sometimes. Pieter, thanks for putting the pressure on in Switzerland when playing Machiavelli. Jeroen, thanks for always being ready to go on holidays together. I will never forget your thesis defense. Marjan, Tom, Ampe, Bekkie, Xavier, thanks for sharing these and other stories together.

Anneke, my close friend for the past 16 years. Thank you for always being there with *raad en daad*. Thanks for the coffee's, the talks, the music, the joy. Sometimes we look at old lady friends having coffee together and smile, I'm sure that will be us in 50 years.

And thank you to the GUHO-team and all the friends I've found there. Arne, Matthias, Annemie, thanks for the lovely time we had together. Arno, Pieter, Lisanne, Roel, Steven, everyone-else-I'm-forgetting, thanks!

The loudest shoutout definitely goes to the neighbours of the "steeg". Oh my god! What an amazing group of people, what an incredible bunch of friends. Everyone of you is such a beautiful character in this book someone

has to write one day. Dieter, Kim, Marnix, Eva, Charlotte, Lana, Jesse, Nele, Sebastiaan, Joke, Jarno, Joris, and everyone else who lives/lived there. Meeting you was the second best thing that happened to me in the past six years. The summer of 2013. There aren't even enough words to describe how much all of this means to me. The many many many parties we shared, the one too many drinks we had, the lovely day-long breakfasts we shared in the sun, the discussions we had about mainly everything. I'm extremely sad the days in the cité are over now, but I'm looking forward to the many other shared moments that await us. I wish you every bit of luck and happiness, because I'm pretty sure you all deserve it a little bit more than the rest of the world does. A special mentioning to Marnix, thanks for playlist '17'. It resulted in a boyfriend and a doctoral thesis. Lana, thank you so much for the wonderful cover of this thesis!

Johan, Kim, Sanne, Senne, Wout, Nathalie, Jan en Ben, thanks for the lovely housewarmings. The nice chats about science and crab semen will stay with me.

Annelies, thanks for all this great time together. Even though I complain about your veggie-fascism a lot, I really enjoy trying new vegetarian recipes together. My wish for you is a lifetime of happy traveling and I hope we will encounter each other in every continent of the world.

Merci Nico ! La meilleure chose qui me soit arrivée ces cinq dernières années, c'est toi. Merci pour tout ce que tu as fait pour ce doctorat, directement ou non. Qui aurait cru notre histoire si on l'avait racontée il y a un an ? Moi pas. Mais chaque jour, je te découvre davantage et, chaque jour, je t'aime plus. Quand je me sens triste, tu me fais rire en quelques secondes. Quand je suis trop stressée, tu me calmes. Quand j'élève la voix dans une discussion politique, tu élèves la tienne plus encore. Quand j'ai une idée folle pour une nouvelle aventure, tu en ajoutes une autre, plus folle. Tu es l'homme le plus pur que j'aie rencontré dans ma vie. Chaleureux, sans arrogance, entier, tu es juste toi-même. Je t'aime.

En tenslotte wil ik graag mijn familie bedanken. Reeds bijna 30 jaar staan jullie aan mijn zijde, en het is een ongelofelijke geruststelling te weten dat dit altijd zo zal blijven. Zonder jullie was dit werk niet mogelijk geweest!

Peter, op een moment als dit mis ik je en ik had je graag trots gemaakt met mijn werk. Meter, bedankt om altijd ontzettend fier te zijn op mij.

Bij jullie vond ik altijd een veilige thuis, waarin liefde, moed en rechtvaardigheid centraal staat. Deze waarden hebben me gemaakt tot de persoon die ik vandaag ben. Oma, bedankt om er ook altijd voor me te zijn.

Brecht, mijn grote broer. Al bijna 30 jaar zijn we samen onderweg gegaan. Soms wat dichter bij elkaar, soms wat verder van elkaar, maar nooit ver in gedachten. Bedankt voor je luisterend oor, ontspannende babbels, gezellige etentjes, de mooie concerten waar je me mee naartoe nam. Ik wens je vooral veel geluk en plezier in het leven. Je gaat samen met Ilse een mooi leven tegemoet, geniet ervan!

Mama en papa. 'Zing, vecht, huil, bid, lach, werk en bewonder', zingt Ramses Shaffy. Als we dat bidden even vervangen door onvoorwaardelijk respect voor onze aardse omgeving, komen we tot een levensleuze die in ons huis zeer centraal staat. Zingen en lachen zijn voor mij belangrijk in het leven: met een positieve geest vooruit kijken. Maar huilen is ook toegestaan: er is plaats voor verdriet en troost. Bewonderen heb ik van kindsbeen af geleerd, door overal mee op reis te gaan. Ik heb geleerd kritisch (maar positief) te kijken naar de wereld, een eigenschap die me ook grondig van pas kwam in mijn onderzoek, en in de rest van het leven. Jullie hebben me leren vechten. Vechten voor waar we achter staan, en vechten voor wat we willen. Het resultaat is soms de zoveelste vurige discussie, tot groot ongenoegen van mama. Anderzijds is ook dit boekje een resultaat van onze vechtlust. You win some, you lose some. Tenslotte heb ik van thuis uit een ontzettend werkethos meegekregen. Door doorzettingsvermogen bereikt de slak de ark. Zonder deze waarden en principes was dit werk niet tot een goed einde gekomen. Daarvoor wil ik jullie bedanken. Bedankt om me te maken tot de persoon die ik aan het worden ben.

Joke,
Ghent, 21 April 2015

1 Introduction

1.1 Magnetic Resonance Imaging

1.1.1 A short history of neuroscience

For centuries, people have been interested in the origin of mental processes. The very first record of the term 'brain' dates back to as early as Ancient Egypt and is found in the Edwin Smith Surgical papyrus, the oldest known document dealing with surgery. It was written ca 1700 BCE and describes 48 medical cases of which 7 deal with brain injuries. Using a remarkable rationality for those times, the cranial structures are realistically described and the brain is already linked to speech, paralysis and the occurrence of seizures.

But even a 1000 year later, in Ancient Greece, people didn't seem to know much more about the brain. The general belief at the time, as the philosopher and scientist Aristotle (384-322 BC) defended, was that consciousness is rooted in the human heart, the brain merely being a cooling mechanism for the blood. This view was challenged on one hand by Plato (428-347 BC) who wrote in his book Timaios that the soul is divided into three parts : the immortal soul in the head, reason in the heart, and the roots of emotion lying in the liver. And on the other hand by Hippocrates (460-379 BC), the father of medicine, who already understood, thanks to his experiments, that the brain was actually the seat of mental processes.

Those views remained until the Roman times and were improved by Galen (129-199 AC). This Roman physician dissected and studied numerous brains of sheep, monkeys, dogs, swines and other mammal, and came to the conclusion that the brain, in addition to the mental processes, controlled the bodily functions as well. He also made studies of cranial nerves and the spinal cord and understood the difference between motor and

sensory nerves. All of his new-acquired knowledge led him to believe that there was no intrinsic distinction between the mind and the body and he attempted to attribute particular nervous diseases to dysfunctions of specific brain regions. Even though some of his beliefs would often seem bizarre nowadays, his ideas became influential throughout the world. After him, it was no longer possible to ignore the fact that the brain was the center of our thoughts and emotions.

But during the Middle Ages, it was more important for the Church to focus on the ethereal soul. So instead of promoting the mere materialistic theory of Galen, without rejecting him on other subjects however, they followed Aristotle's classification of the mind into feeling, thinking, remembering, etc. Since the dominant church fathers, like Augustine of Hippo, thought that brain tissue was too mundane to act as an intermediary between the earthly body and the heavenly soul, they put the mind in the empty part of the brain, the ventricles, which had been claimed to be hollow by Galen in the context of his theory of blood circulation in the brain. With such strong views, it was difficult to make any significant progress on the understanding of the brain and the human mind, even though the period was fruitful in philosophical debates about the nature of the mind. While Europe was under the influence of strong religion dogmas, the Arabic civilization was in a Golden Age which allowed the translation, study, and preservation of all the great philosophers and doctors of the Antic World. But there as well, the religious dogmas prevented any real improvement.

Things would radically change with the Renaissance period. Even though the Church would still be strong through the Inquisition, people would start dissecting bodies, driven by scientific or artistic purposes. One of the first to do so was Andreas Vesalius (1514-1564), the father of human anatomy. He gave the most advanced description of the human brain so far, showing the absence of holes in the ventricles, destroying the belief that it was the siege of our soul. He also helped defining the nerves by rejecting ligaments and tendons as part of the nervous system as previously thought, and he showed that the nerves were not hollow. This quest for truth was continued by many scientists of that time, among which Descartes (1596-1650) who was so interested in the subject that it has been written that *if Descartes were alive today, he would be in charge of*

the CAT and PET scan machines in a major research hospital (Watson 2002, p. 15). But in spite of their dedication and studies, and even though they passed on to us a detailed anatomical description, those pioneers could not pin down the relationship between the brain and our mental processes. It became clear that only through a thorough and systematic study in a scientific setting, the secrets of the brain could be further unraveled.

The neuroscientists in the 19th and 20th century focused mostly on the mapping of the brain. Where are the crucial functions such as vision, control, speech, emotions,.. located in the brain? The first widespread theory was the pseudoscience of phrenology. The first phrenologist, Franz Gall (1758-1828), believed the mind was like other muscles in the body: when exercising a specific function, the brain region responsible for this function grows, which leads to bumps on the skull. By carefully inspecting skulls of people, Gall derived a detailed atlas with psychological functions. While we know today that the idea of phrenology is not scientifically grounded, the idea that functions are located in localized parts of the brain is considered an important historical advance in neuroscience.

From then onwards, the brain was studied intensively. Different methods for research can be distinguished: lesion studies, drug studies and recordings of electrical activity. The study of the brain lesions could be done after the subject has died. Despite being useful for anatomical research, it is less suitable for functional research. For that matter one investigated the psychological functions of subjects with lesions in the brain. A famous example is Phineas Gage (1823-1860), who survived a rock-blasting accident that destroyed much of his brain's left frontal lobe. After the accident, his functions remained intact, but his personality and behavior changed largely, which indicates that the frontal lobe is responsible for personality. Another important lesion study was the research of the French physician Pierre Broca (1824-1880). By studying the brain of patients with speech difficulties, he discovered the exact location of the center of language production. Later, this work was extended by Wernicke (1848-1905), who localized precisely the brain area responsible for language comprehension through lesion studies. But even though lesion studies are undoubtedly important and useful for neuroscience, they are limited in their applicability. The patients are only studied after they

obtained the lesion, and it is often difficult to find patients with isolated damage. Another problem with lesion studies relates to the network structure of the brain. From a lesion study, one can conclude that a specific part of the brain relates to a specific function. However, as brain functions act as a network, probably other brain regions relate to the same function. A certain brain region can therefore be necessary but not sufficient for a function.

In contrast, drug studies allow investigation of large brain networks that cannot be associated with a simple task. A disadvantage however is the difficulty to identify functions of specific brain regions.

A third method to study the brain, imaging, was only discovered in the beginning of the 20th century. Electroencephalography, which is up to today the most direct measure of neuronal activity, detects very rapid changes in electrical potential, and is therefore the perfect way to study the timing of brain processes. However, because the measurements can only be performed outside of the brain, its power for localization is very low.

Later, at the beginning of the 20th century, new technological advances and progress in the science of physics led to the discovery of better imaging of the body. In the early 1980s, new techniques using X-rays such as computerised axial tomography (CAT), single photon emission computed tomography (SPECT) and positron emission tomography (PET) allowed to study not only the anatomy but also the function of the brain. A few years later, magnetic resonance imaging (MRI) was applied to study in vivo the anatomy of the brain. Today, the progress in neuroscience is faster than ever and this success can be largely attributed to the advances in neuroimaging, and more specifically to MRI.

To conclude our historical survey of the study of the brain, we can say that even though it has been an object of interest since the beginning of mankind history, it is only very recently that we came to understand, even though in a very limited way, the functions of that amazing and somewhat mysterious organ. After more than 3500 years of philosophical and scientific debates, the question of the relationship between our mind and our brain remains one of the most difficult, but undoubtedly the most interesting, modern medicine has to answer.

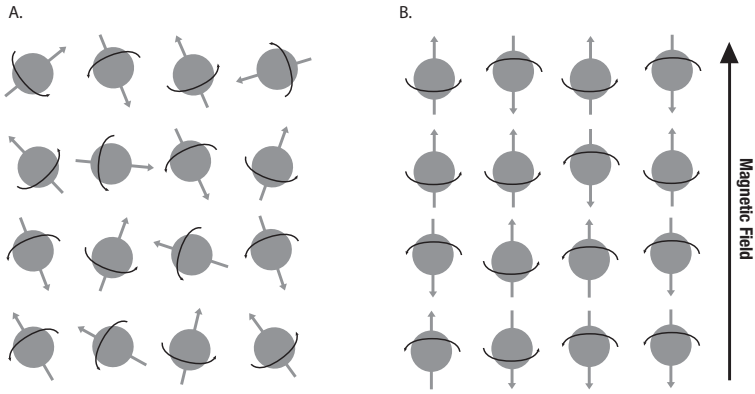


Figure 1.1: Protons in free space with random orientations (A) and protons in a magnetic field (B).

1.1.2 Basic MRI physics

MRI relies on a core set of physical principles. In what follows, a short overview will be given on their physical principles. At the end of this section, it will become clear how brain functions can be measured using MRI.

Nuclear spins MRI focusses on the detection of hydrogen nuclei, which is an atom consisting of a single proton. A single proton under normal conditions spins about itself and are therefore called a spin. This rotational movement causes electrical current because of its positive charge, which in turn causes a magnetic moment, μ. The rotation also induces an angular moment J as a result of its mass. The gyromagnetic ratio γ, defined as μ/J, is crucial for MRI. In the absence of magnetic fields, as in panel A of Figure 1.1, the spins are oriented randomly. When placed in an external magnetic field with strength B_0, the spin axis of all spins align with the magnetic field (Figure 1.1 panel B).

However, the spin axes are not perfectly aligned with the magnetic field. In addition to their spinning motion, their axis of spin itself wobbles

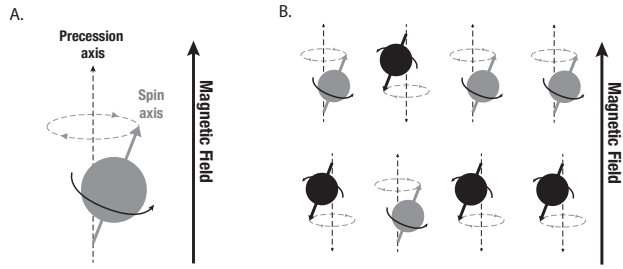


Figure 1.2: Panel A shows the precession of the spin. Panel B shows the difference between high energy (black) and low energy (grey) spins.

around a vertical axis, see panel A in Figure 1.2. This motion is known as precession: the angular moment J precesses arond the external field axis with an angular frequency known as the Larmor frequency $\omega = -\gamma B$, where B is the magnitude of the magnetic field and γ the gyromagnetic ratio. As is shown in panel B in Figure 1.2, recessing protons can be in two states: parallel or antiparallel to the magnetic field. Protons in the parallel state have a low energy level, protons in the antiparallel state have a higher energy level. The aim is to measure the net magnetisation, which is determined by the difference between the number of high energy spins and the number of low energy spins in a certain volume. If an excitation pulse, in the form of another magnetic field, with the right amount of energy is applied, some spins will absorb that energy and jump to the high-energy state, in a process called excitations. Once the excitation pulse is turned off, some high-energy spins return to the low-energy state, thereby releasing the absorbed energy in the reception period. The released electromagnetic energy can be detected by a radio frequency coil. Measurement of this emitted energy, or MR signal, provides the data that go into our images.

Net Magnetisation The net magnetisation M is defined as the sum of the magnetic moments from spins. Without a magnetic field, the spin axes of all nuclei are oriented in random directions, so that the net magnetisation is zero. In the precense of a magnetic field, all magnetic moments align

in either the parallel or antiparallel state. There are always more parallel than antiparallel spins, therefore there is always net magnetisation in the presence of a magnetic field. Based on the net magnetisation, the number of protons within a unit volume can be estimated. As such, in this specific case the amount of hydrogen in a unit volume can be estimated. However, the net magnetisation of a spin system cannot be measured directly. An indirect measure for the net magnetisation is excitation.

Excitation To change the net magnetisation, MRI scanners use radio frequency coils to transmit an electromagnetic excitation pulse (B_1) at the same frequency as the spin precession (i.e. the Larmor frequency), which allows to perturb the spins. By applying an excitation pulse, the net magnetisation vector is tipped downward from the longitudinal to the transverse plane, a process called nutation. Because of the precession, the downward movement of the net magnetisation will follow a spiral motion, as shown in Figure 1.3. As a result of this nutation, the net magnetisation of the spin system is changed and this change can be measured with another receiver coil. Once the excitation pulse is taken away, the spin system gradually loses the energy absorbed during the excitation and the spins return to the parallel (or antiparallel) state. This phenomenon is known as longitudinal relaxation. The time constant associated with this longitudinal relaxation process is called T_1, and the relaxation process is called T_1 recovery. Another process associated with the removal of the excitation pulse is transverse relaxation: at first the spins precess as the same starting point but over time this coherence is gradually lost and they become out of phase. The signal loss by this mechanism is called T_2 decay and is characterised by a time constant known as T_2. Both T_1 recovery and T_2 decay can be measured by the receiver coil. It is exactly the T_2 decay that will provide an indirect way to measure brain activation. After applying Fourier transformation, the data can be transformed to magnitude data and phase data. This magnitude data can be used as the final measure for neural activation.

The BOLD response Under neural activation, neurons immediately need a new supply of energy. Therefore these neurons receive immediately glucose through an increase in oxygenated blood around these neurons. This

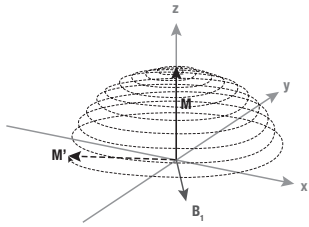


Figure 1.3: This figure shows the nutation of the net magnetisation M, from the longitudinal (z-axis) to the transverse plane (x- and y-axis).

results in a higher rate of oxygenated blood flow and an expansion of blood vessels. The blood-flow change takes place within 2 or 3 mm of where the neural activation is. Oxygenated blood consists mostly of hydrogen atoms. As hydrogen nuclei are known to lose their magnetisation faster via T_2 decay, there will be an increase in MR signal where blood is highly oxygenated. The result is the measurement of the blood oxygenation level dependent (BOLD). The BOLD signal is only a correlate of neural activity. The oxygenation of the blood happens only *after* the neural activation, following a function defined as the haemodynamic response function or HRF (Figure 1.4). The specific form of this response function varies for different brain regions and for different subjects. Still, the BOLD signal with its imperfections is accepted as a plausible measure of neural activity in the brain.

1.2 The statistical analysis of fMRI data

1.2.1 The fMRI experiment

Images The result of a brain scan, is a matrix of numbers, as is shown in Figure 1.5. The brain is divided into many small blocks, which are called voxels, where for each voxel the BOLD response is measured using the principles laid out in the previous section. The size of the voxels depends on the resolution of the scanner, but balances with current scanners some-

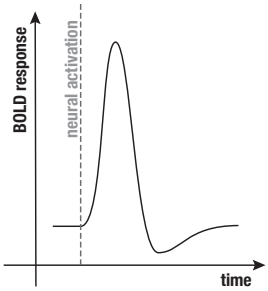


Figure 1.4: Haemodynamic response function.

where between 1 and $3mm^3$. The goal of an fMRI study is to get more insight in the functions of the brain. However, when looking at the BOLD response for a normal awake person, we will see that the whole brain is activated. Therefore, the BOLD response from one scan does not tell us anything about the location of specific psychological functions. We therefore need a range of scans accompanied with a psychological experiment.

Example: the Auditory dataset An example of an fMRI experiment is the auditory experiment, which is one of the first experiments with fMRI data collected and analyzed in 1991 (Friston, Ashburner, Kiebel, Nichols, & Penny, 2007). The goal of the experiment was to detect brain regions involved in auditory perception. In order to find those regions, a subject was placed in a scanner, while auditory stimulation was presented. In the experiment, the subject was presented with a block of auditory stimuli, more specific bi-syllabic words, alternated with silence, where each block lasted for $+/-$ 40 seconds. More general, there are two conditions: the active condition and the inactive condition. During the experiment, MRI scans of the brain are taken every 7 seconds. In the analysis of the data, the scans of the active condition are contrasted with the scans of the inactive condition. More specific, for each voxel, the intensity of the BOLD signal are compared between the two conditions, using statistical tests. If for a certain voxel, the difference in BOLD-intensity between the two conditions

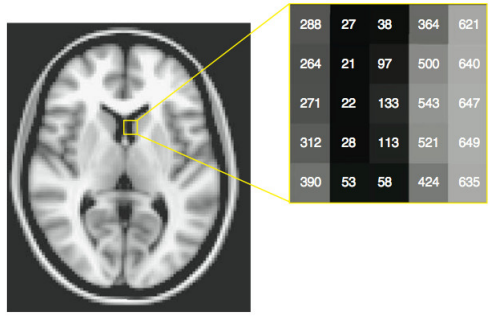


Figure 1.5: Graphical representation of a brain scan divided into voxels. (R. Poldrack et al., 2011)

is sufficiently large compared to the difference within the conditions, this voxel is declared statistically significant related to the auditory function. The result is a map where all voxels are shown that are significantly related to the psychological function of auditory perception. The analysis pipeline of this experiment can be found in Figure 1.6.

fMRI analysis The auditory dataset was collected and analysed in 1991. Of course, in the past 20 years, the analyses of fMRI data have largely been optimised and improved. Nowadays, the procedures and corrections to analyse fMRI data are collected in an analysis pipeline, shown in Figure 1.7. The main steps in the analysis of fMRI data are preprocessing, model fitting and localisation. In the following sections, we will discuss each of these steps in more detail.

1.2.2 Preprocessing

Slice timing correction fMRI data consist of different scans of the brain that are taken for example every 2 seconds. However, the data are taken in different slices, from the neck of the subject to the top of the head. Therefore, the timing between the scans of the two lowest slices is exactly 2 seconds, but the timing between the scans of the lowest and the highest slice are slightly different. The first preprocessing step consists of a

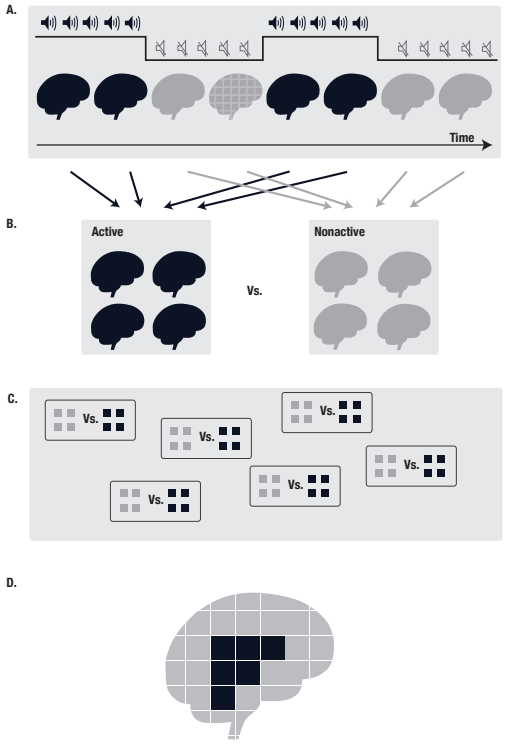


Figure 1.6: Graphical representation of a simple fMRI experiment in which the brain location related to auditory perception is under investigation. In a first step of an fMRI experiment (panel A), a continuous time series of scans is taken, while the subject is presented either with auditory stimulation (active condition) or with silence (inactive condition). In the next step (panel B), a comparison is made between the scans during the active and the nonactive condition. More specific, for each volume unit (voxel) the comparison is made (panel C). Statistical testing procedures enable to draw inference on whether the differences in brain activation are due to the difference in condition, or are rather due to chance. In the last step (panel D), for each voxel, a decision is made whether the voxel is significantly related to task or not.

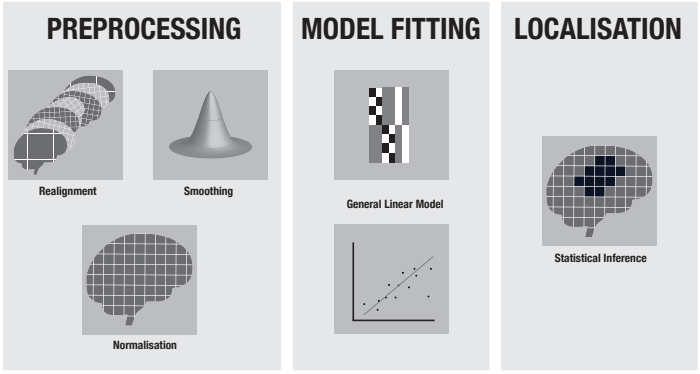


Figure 1.7: Main steps in the analysis of fMRI data.

correction for the difference in slice timing.

Head motion correction The analysis of fMRI data happens typically voxelwise, so the BOLD signal is at first compared between minuscule volume units during the experiment. It is therefore key that these volume units refer to the exact same location in the brain in every scan. A big problem for this is the fact that - even though the head is securely fastened during the experiment - there exist small movements of the brain. By estimating the rotation and shift of the brain, and consequently shifting the BOLD measurements slightly according to the movements, the scans are corrected for head motion. This step is often referred to as realignment.

Coregistration During the experiment, a high resolution anatomical scan is taken, that often has dimensions $256 \times 256 \times 160$. The fMRI data itself have a very low resolution because of the high temporal resolution, often in dimensions $64 \times 64 \times 48$. In the coregistration step of the preprocessing, these two scans are aligned, so that results from the fMRI experiment (with statements about brain functions), can be compared to the anatomical scans (with statements about the anatomical regions).

Normalisation In order to increase the validity of the fMRI experiment, multiple subjects are used. However, different people have different brain shapes and sizes. When we want to make general statements about the brain activation over different subjects, it is important to align the brains of the different subjects to a standard template of the brain. For example, an atlas produced by the Montreal Neurological Institute (MNI) is very popular for the normalisation step.

Spatial smoothing The last big preprocessing step is the spatial smoothing of the brain. The BOLD signal measured in the brain suffers from many noise components and is very rough. However, we can assume that the oxygen in the blood does not stop and completely changes at the border of a voxel. It is the inverse, the oxygen level will not differ much between neighbouring voxels. Therefore, the BOLD value in each voxel is replaced by a weighted average of the BOLD responses in neighbouring voxels, where voxels at small distances have larger weights than voxels at a larger distance.

Other preprocessing steps In the spatial smoothing, the data is blurred across neighbouring voxels. Another preprocessing step consists of temporal smoothing, where the data is blurred in between each scan, as it makes sense that the oxygen level does not change hugely between 2 different time points. However, this induces even more autocorrelation between the timepoints than there already is. As subsequent analyses require different timepoints to be independent, the data is often de-correlated or whitened. Not all noise is due to the subject, also the scanner suffers from noise. One such example is drift: even though the magnetic field should remain constant, it often changes over the course of an experiment. Therefore, the magnetisation is measured separately, and the data are corrected for the drift. Finally, sometimes global differences exist across sessions or subjects that can be easily corrected by applying a grand mean scaling. Consequently, the mean of the BOLD response from each scanning session has the same numerical value. Once all these preprocessing steps are carried out, the actual statistical analysis of the BOLD signal can start.

1.2.3 Model fitting

There are different models plausible to capture the data. Two main families of models exist: the GLM approach and multivariate analyses, such as ICA, PCA,... The GLM method constructs a test for each voxel that is only based on information in the voxel itself. This is a mass-univariate approach as it performs a large number of tests that are univarate of nature. The multivariate approach combines information over voxels. We will use multivariate analyses in chapter 5, but in what follows we focus on univariate analyses.

Single subject analysis

The way in which the conditions are compared in Figure 1.6 is easily understandable, but is unfortunately a very simplified version of the reality. In Figure 1.4 it is shown that there is a lag in the blood flow to the active brain regions, and therefore it is not simply possible to compare the scans in a categorical way (active condition vs. non-active condition). Instead we look at the brain activations as a time series. Panel A of Figure 1.8 shows the design of the experiment. Because of the lag in blood flow in the brain, we would expect the BOLD signal in an activated voxel, without any noise, to be a time series as shown in panel B of Figure 1.8. The BOLD-signal observed in a voxel involved in the experiment is not perfect, as many sources influence the measured signal besides the experimental condition. Instead, we observe a signal as presented in panel C of Figure 1.8. To quantify the overlap between the design and the observed signal in a certain voxel, a linear model as follows is used.

$$Y_i = b_0 + b_1 X_i + e_i \tag{1.1}$$

where Y_i refers to the observed BOLD signal in a certain time point (black line in Figure 1.8 C) and X_i refers to the expected BOLD signal under perfect activation (grey line in panel B of Figure 1.8). $\hat{b}_0$ and $\hat{b}_1$ are estimates for the relation between Y_i and X_i and b_0 and b_1 can be estimated by minimising the difference between Y_i and $b_0 + b_1 X_i$. b_0 refers to the baseline of the BOLD-signal, and b_1 is a scaling factor for the design that quantifies the overlap between the design and the observed BOLD

signal. The part of the observed signal that cannot be predicted by the expected activation, is called the residual error, e_i, which is assumed to be zero on average. When the variance of these errors is large, it is hard to estimate b_1, or in other words the variance on the estimate $\hat{b}_1$, $Var(\hat{b}_1)$ is large. With a lower residual variance, it is easier to observe the relation between Y_i and X_i, and thus $Var(\hat{b}_1)$ will be smaller.

To study whether there is a relationship between the design and the observed BOLD signal, we formulate the following hypothesis:

$$H_0 : b_1 = 0 \tag{1.2}$$

If H_0 cannot be rejected, then we state that the particular voxel is not responsible for the task. However, the estimated b_1 will hardly ever be equal to zero, but what matters is whether this is statistically significant or whether the difference is due to chance. Stated differently, we want to see whether b_1 is large compared to the variance on the estimate b_1. To do so, we compute a statistic that measures the evidence against H_0, such as a T-statistic:

$$T = \hat{b}_1 / Var(\hat{b}_1) \tag{1.3}$$

In a next section, we will discuss how these statistics can help us decide whether we have enough evidence against the hypothesis in Equation (1.2) to conclude that the voxel is active.

In our example on auditory perception, there is only one design component in a blocked design (i.e. 30 seconds on / 30 seconds off). However, many more experimental designs are possible. Figure 1.9 shows two examples of alternative designs. The left panel in the figure shows an event-related design, where the conditions are not alternated in a blocked design, but appear at random time points. An example is an experiment where every now and then a very hard and surprising sound is played. The right panel of Figure 1.9 shows how multiple design components, or regressors, are possible in the same experiment. One example is when no visual stimuli are shown, followed by 20 seconds of seeing faces and thereafter 20 seconds of seeing houses. The linear model then changes to:

$$Y_i = b_0 + b_1 X_{1i} + b_2 X_{2i} + e_i \tag{1.4}$$

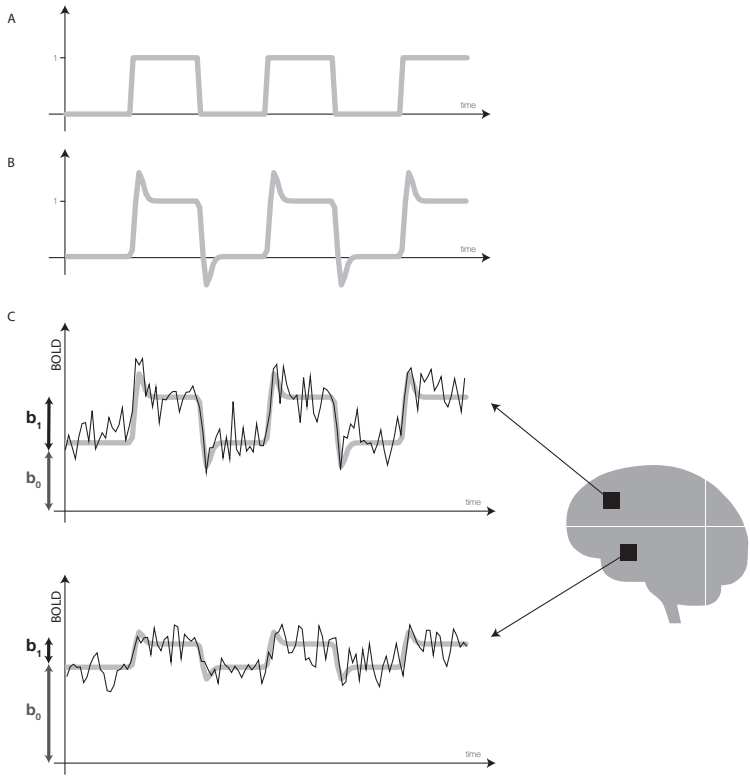


Figure 1.8: This figure shows how to model the voxelwise time series. Panel A shows the box-car design, with an on-off experimental design. Panel B shows the expected activation given the haemodynamic response function. Panel C shows two observed time series, and it shows how the parameters b_0 and b_1 represent the relationship between the observed time series and the expected activation.

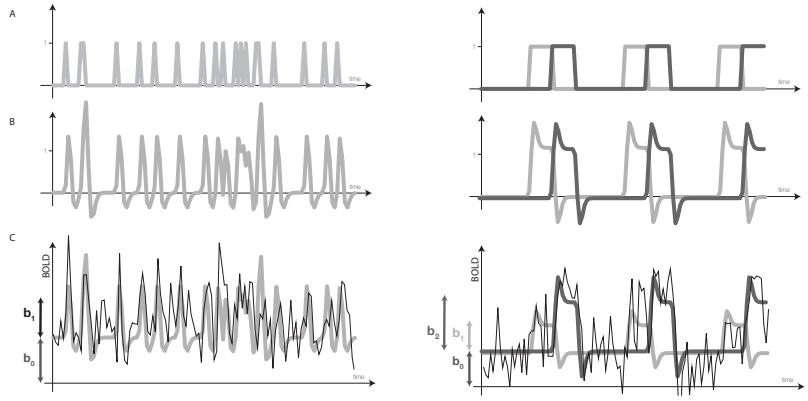


Figure 1.9: This figure shows how different experimental designs are possible. The left panels show an event-related design, while the right panel shows a design with two regressors. The upper panel (A) shows the experimental design, the middle panel (B) shows the expected activation taking into account the HRF. The lower panel (C) shows an example of a time series observed in a certain voxel and the estimates that result from the linear model.

It can be seen that b_1 represents the relationship between the observed BOLD signal and the first regressor (eg. faces), and b_2 quantifies the relationship between the observed BOLD signal and the second regressor (eg. houses). An experiment with multiple regressors allows to compare the active condition with the non-active condition ($H_0 : b_1 = 0$ or $H_0 : b_2 = 0$), but also to compare the different active conditions, for example the difference between seeing faces and houses ($H_0 : b_1 = b_2$).

1.2.4 Group analysis

It is known that brain functions differ between people. Therefore results on brain functions cannot be claimed based on the results from one subject. On the other hand, the interest may lie in the differences between different subjects. Therefore, we often perform the same fMRI experiment on many different subjects. The analysis of a group study always

starts with a single-subject analysis for each subject as described above. For each subject s, and for each voxel i, we obtain an estimated parameter $\hat{b}_{1is}$ and its variance $Var(\hat{b}_{1is})$. This step is often referred to as the first level analysis. In the second level analysis, we perform a new analysis, where we compare the b_1-parameter estimates for each voxel between subject groups. The linear model is the same as in equation 1.1, but now X_{is} refers to group identity. The outcome variable Y_i refers to the b_{1is}-parameter estimates. It is clear from Figure 1.10 that the second level analysis is very similar to the first level analysis, but different conditions now refer to between-subject differences (eg. sick vs. healthy, male vs. female,...) instead of within-subject experimental conditions. In the event that there is only one group, we compare all b_1-parameters with 0, and the design matrix X_{is} will be $(1, 1, ..., 1)$.

1.2.5 Localisation

Statistical inference

In the previous section, we explained how to compute a statistic for each voxel based on a linear model. In this section will discuss how to make binary choices for each voxel, i.e. is a voxel *significantly* related to the experimental design or not? This enables to localise in the brain which area is responsible for a given task. We test for each voxel the null hypothesis H_0 of no task-related activation against the alternative hypothesis of task-related activation H_a. It is unknown whether H_0 or H_a is true. As such two distinct errors are possible, as shown in Table 1.1. If there is no task-related activation in the voxel, we can make a false positive classification, also denoted as a type I error. If the voxel is involved in the experimental task, but we mistakenly declare it non-significant, we make a type II error or a false negative. The goal of any statistical test is to minimise the number of errors made.

To guide the classification, we rely on the probability density function (PDF) of a statistic under the null hypothesis to make inference. The PDF allows us to compute *the one-sided p-value*, where $p = P(T \geq t|H_0)$, the probability to find a statistic equal or larger than the observed statistic when H_0 is true, as is shown in Figure 1.11. For many statistics, such as the T-statistic in Equation 1.3, the PDF is known. To make inferences on

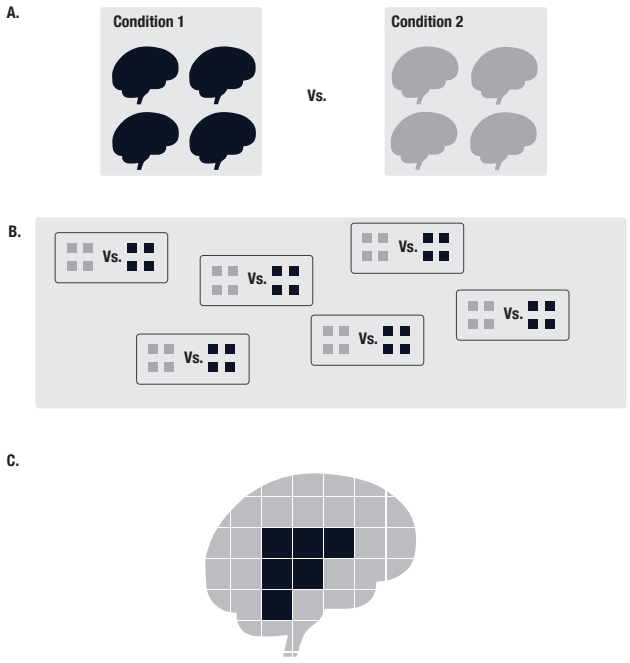


Figure 1.10: This figure shows the second level analysis of a group study.

	Declared active	Declared inactive
Non-active	False positive (Type I error)	True negative
Active	True positive	False negative (Type II error)

Table 1.1: Possible outcomes for a statistical test.

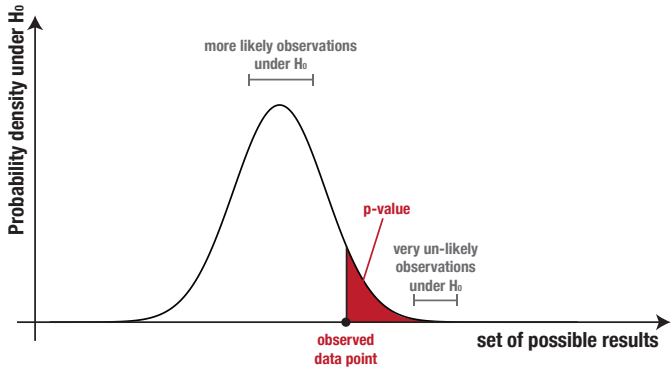


Figure 1.11: Example of a p-value computation. The vertical coordinate is the probability density for each outcome, computed under the null hypothesis. The p-value is the area under the curve past the observed data point.

the activation level for each voxel, we set a threshold α on these p-values, and declare all p-values higher than α non-significant (H_0 not rejected). All p-values below α are declared significant (reject H_0). As a consequence, we find a certain threshold c, for which holds that $\alpha = P(T \geq c|H_0)$, the probability of detecting a false positive when H_0 is true. The rationale behind this is that if one finds enough evidence against the null hypothesis, one can confidently reject it. This procedure controls the probability of a type I error at level α.

The multiple testing problem

In an fMRI analysis, we test for activation in over 100 000 voxels at the same time. Suppose that there is no effect of the task anywhere in the brain and all voxels are independent. If we would then test using a very common level of $\alpha = 0.05$, we can expect an average of 5000 falsely positives. In other words, we would on average falsely conclude that there is an effect in 5000 voxels. This problematic situation that occurs when

	Declared active	Declared inactive	Total
Non-active	F	$m_0 - F$	m_0
Active	T	$m_1 - T$	m_1
Total	S	$m - S$	m

Table 1.2: Expected numbers in a cross-classification when testing m voxels for significance . m_1 is $m - m_0$. S represents the number of voxels that are declared active. F and T respectively represent the false and true positives.

testing many statistical tests simultaneously, is called the multiple testing problem. Assume that we test H_0 against H_a for m voxels. The possible outcomes when testing m voxels are denoted in the Table 1.2.

In the framework of multiple tests, several definitions of a type I error rate exist:

- PCER: The Per-Comparison Error Rate, equal to $E(F)/m$, the expected number of type I errors divided by the total number of tests.
- FWER: The FamilyWise Error Rate, defined as $P(F \geq 1)$, the probability of making at least one type I error.
- FDR: The False Discovery rate, the expected proportion of true null hypotheses among the rejected ones, E(F/S), where the proportion is set to 0 in case of no rejections (S=0).

Each of these error rates can be controlled using certain procedures. We discuss one procedure per error rate, however many more procedures exist.

PCER control In case we threshold the p-values at level α as we did in a single hypothesis problem, and therefore do not correct for multiple comparisons, we control the PCER at level α.

FWER control The most common FWER controlling procedure is the Bonferroni procedure. Here each of m tests is performed at significance level α/m. This controls the FWER at level α.

FDR control Let $p_1 \leq p_2 \leq ... \leq p_m$ represent the ordered voxelwise p-values. Then the procedure of Benjamini & Hochberg (1995) controls the FDR at level α as follows:

- Find the largest k such that $p_k \geq \frac{k}{m}\alpha$
- Then reject all H_i for $i = 1, ..., k$.

Random Field Theory

The thresholding procedures described above all make the assumption that the tests are independent. But the statistical tests in neighbouring voxels are not independent. When there is strong brain activation in a voxel, there is a high chance that the neighbouring voxels will be active as well. Many testing procedures have been developed where the correlation between the different tests is captured. However, spatial correlation is a specific type of correlation, that calls for specific solutions. One framework for topological inference that allows to test for activation while taking into account the spatial structure is Random Field Theory (Adler, 1981; Friston, Frith, Liddle, & Frackowiak, 1991). Assume that under the null hypothesis of no activation, the three dimensional array of statistics, also called the statistical parametric map (SPM), is a gaussian random field. A gaussian random field is a collection of random variables that have a multivariate Gaussian distribution. One important feature of the random field is the spatial smoothness, translated as the standard deviation in directions $i = x, y, z$ of the multivariate Gaussian distribution, σ_i. Related to this measure of smoothness, we define a resel as a resolution element, which describes the actual spatial image resolution in a volume. The number of resels will be lower or equal to the number of voxels in the image. If we define the full width at half maximum of an image as $\text{FWHM}_i = \sqrt{8 \log 2}\sigma_i$, then the resel size equals:

$$\text{resel} = \text{FWHM}_x \times \text{FWHM}_y \times \text{FWHM}_z \tag{1.5}$$

Assume one only considers statistics that exceed a certain threshold u. The collection of these voxels form an excursion set. For this excursion threshold u, it is possible to compute the Euler characteristic (EC), a topological invariant that describes the topological space's shape. The

precise definition is rather complex, but as an approximation, it can be described as:

$$EC = \#(\text{blobs}) - \#(\text{holes}) \tag{1.6}$$

For example, a doughnut has one blob and one hole (EC=0), a ball has only one blob (EC=1), the galaxy has one hole but many blobs, a block of Swiss cheese has one blob but many holes, etc.

In the context of fMRI, if the excursion threshold u for the SPM is large enough to make sure that there are no holes in the excursion set, then the Euler characteristic will be equal to the number of blobs. Under the null hypothesis of no activation, every blob is a false positive blob. We can redefine all type I error rates described above in this spatial framework:

- PCER: the average number of blobs over all resels
- FWER: the probability of finding one or more blobs

FDR has no meaning under the assumption of the complete null. As such, we can find a new threshold t to control one of these false positive rates voxelwise.

Levels of inference

Within the framework of topological inference, different levels of inference can be defined: voxelwise, clusterwise, peakwise and setwise. All procedures differ in their number of tests n, the feature they focus on, and the probability of the features.

voxel level inference Voxelwise statistical testing refers to all voxels being tested simultaneously, with one of the procedures described above. n is the total number of voxels, the p-values can be computed with classical statistical testing and is derived from the voxelwise test statistic.

peak level inference In peakwise inference, we only consider local maxima (the maxima of all neighbouring voxels). n refers to the number of peaks above excursion threshold u. The p-value for local maxima t is approximated as (for derivation see Durnez, Moerkerke, & Nichols 2014):

$$P(T \geq t | T \geq u, H_0) \approx \exp(-u(t-u)) \tag{1.7}$$

cluster level inference When performing statistical tests on clusters, one considers all clusters above excursion threshold u. The p-value is computed based on the extent of each cluster. The idea is to approximate the shape of the image by a quadratic with a peak at the local maximum. For a Gaussian random field, the spatial extent S is approximated by the volume where the quadratic of height H above u cuts the threshold u. The spatial extent can be expressed as $S \approx rH^{D/2}$ (Worsley, 2007), with D the number of dimensions (i.e. three), with

$$r = \frac{\text{FWHM}^D u^{D/2} P(T \geq u | H_0)}{EC_D(u)\Gamma(D/2+1)}. \tag{1.8}$$

The p-value for a cluster with an observed spatial extent s can be approximated as follows (Worsley, 2007):

$$P(S \geq s | D, FWHM^D, u, H_0) \approx \exp(-u(s/r)^{2/D}) \tag{1.9}$$

set level inference Set level inference aims to make an overall conclusion: is there task-related activation in the brain or not? These tests have not been widely used because their lack of localizing power. The set-level approaches an omnibus test (Friston et al., 2007), with l equal to the cluster number:

$$P(\text{activation}) = 1 - \sum_{i=0}^{l-1} \frac{1}{l!} \text{EC}_D^l \exp\left(-\text{EC}_D\right) \tag{1.10}$$

1.2.6 The power problem in neuroscience

Statistical power

In the previous section, we show how to compute the p-value of a voxel, the probability of a false positive, and how to threshold these, while controlling a fixed false positive rate. While it is important to avoid false positives, it is also important to consider the chance at false negatives as much time and money is invested in scientific research (Figure 1.12). When H_0 is

rejected as soon as the test statistic is larger than or equal to cutoff c, the type II error rate is defined as $\beta = P(T \leq c|H_a)$, the probability of making a false negative error when H_a is true. In panel B of Figure 1.12, both the α- and β-levels are shown. To have a measure how sensitive a testing procedure is, we define *statistical power* as the probability of detecting an effect when it is actually present, $P(T \geq c|H_a)$.

Power in fMRI

We argue that while it is very useful to prevent false positives, it is also necessary to take into account the prevalence of false negatives in a study. Because the more one tries to prevent false positives, the higher the chances at false negatives. This problem is especially present in neuroimaging. The scanners are known to be prone to different sources of noise (Geissler et al., 2007), and the multiple comparisons problem in neuroscience is of a huge dimension. Therefore, neuroscience, more than most other sciences suffer from a huge loss in power (Button et al., 2013). The goal of this doctoral thesis is to assess the power problem for individual fMRI studies, as well as providing answers to specific settings of problems with statistical power.

1.3 Motivation and outline

Different statistical procedures have been introduced that deal with the multiple testing problem. Control of the FDR was introduced in neuroscience and because of its higher power, it replaced FWER control. However, higher power comes with a great loss of specificity, also called the true negative rate. We can ask ourselves: is the gain in power worth the loss in specificity? This question is tackled in the second chapter. We compare control of FWER and FDR on a number of features, such as specificity, statistical power, and stability for a large range of possible scenarios. Stability can for example be measured through the variability of the results, with a lower variability indicating a higher stability. There we find that for a fixed α-level FWER control indeed exhibits very low power, a problem overcome significantly by FDR control. But on the other hand FDR control suffers from a big loss of stability. Consequently FDR

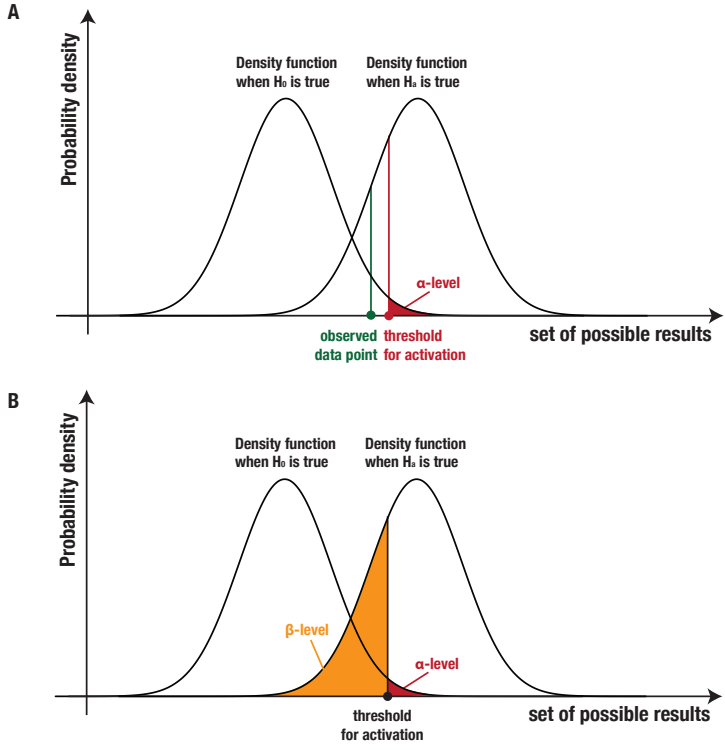


Figure 1.12: Panel A. In noisy contexts, where activation can be hardly separated from non-activation, the distributions under H_0 and H_a are closely together. In this example, the observed data point does not exceed the threshold for activation (i.e. its p-value is higher than the applied α-level. Therefore, for this data point (voxel), we would conclude that there is no significant effect. However, when looking at the probability density function of H_a, it is clear that this observation actually has a high probability of being observed when H_a is true. Even more, the probability of observing this point under H_a is higher than the probability of the observation under H_0. Panel B. For a certain statistical threshold c, we can define α as the probability of a type I error when H_0 is true, and β as the probability of a type II error when H_a is true.

control is not the holy grail of multiple testing procedures. When conducting an fMRI study, careful attention should be given to the thresholding question, thereby considering which thresholding procedure comes with which power.

However, apart from few publications on how to choose an appropriate sample size for fMRI experiments (Friston, 2012; Desmond & Glover, 2002; Mumford, 2012), few is known on statistical power in fMRI. Therefore in the third chapter, we present a concise procedure with which power can be estimated for a given study post-hoc. This enables researchers to evaluate their own studies, and makes it possible to assess to which extent their studies find the task-related brain activation of study.

A drawback of this procedure is that it only functions post hoc and cannot yet predict the power of a study before it is conducted. Therefore, in chapter four, we extend the procedure for this purpose. We present a simple way to characterise the spatial signal in an fMRI study, and a direct way to estimate power based on an existing pilot study. Specifically, using just (1) the proportion of the brain activated and (2) the average effect size in activated brain regions, we can produce closed form power calculations for given sample size, brain volume and smoothness. This procedure allows minimising the cost of an fMRI experiment, while preserving a predefined statistical power.

However sometimes the goal of a study prevents the possibility of simply increasing its sample size. One such example is pre-surgical fMRI, in which one localises certain psychological tasks in a subject prior to undergoing surgery for brain tumors, epilepsy, etc. In that case a researcher has to deal with the power that comes with the single subject data. This fact notwithstanding, a classical stringent focus on preventing false positives is accompanied by a risk of false negatives, which can be detrimental, particularly in clinical settings where false negatives may lead to surgical resection of vital brain tissue. Therefore, in chapter five, we present an alternative based thresholding procedure that incorporates information on false positives and false negatives. This results in a layered statistical map for the brain. One layer marks voxels exhibiting strong evidence against the traditional null hypothesis, while a second layer marks the voxels where activation cannot be confidently excluded. The third layer marks voxels where the presence of activation can be rejected. In this

procedure, control of false positives remains possible but our procedure also takes into account information on the false negative rate. With the procedure in chapter five, we aspire to better delineate the extent of psychological functions in the brain using fMRI. In chapter six, we explore the possibility to evaluate statistical procedures that test for brain activation by combining fMRI data with data from Diffusion Weighted Imaging (DWI). DWI is a technique that rather searches structural connectivity in the brain and separates different brain areas based on their physical connections in the brain. We use data from a patient with a brain tumor who undergoes both an fMRI experiment and DWI. We apply both classical significance testing and our alternative based testing procedure to the fMRI data and compare which of both procedures best predicts the results obtained with DWI. A big overlap between completely independent procedures indicates a better predictive validity of the procedure. We conclude with an overview of the main findings and conclusions in each study in chapter seven. The implications and shortcomings of our results are discussed in a more general framework, while at the same time we provide indications for future research concerning statistical power in neuroscience. Chapter 2, 3, 4, and 5 are originally written as stand-alone articles. As a result, there exists some overlap between these chapters. Notation is introduced per chapter and may not be the same throughout the complete thesis. Chapter 2 is published in Biometrical Journal (Durnez, Moerkerke, & Nichols, 2014), chapter 3 in NeuroImage (Durnez, Roels, & Moerkerke, 2014), and chapter 5 is published in Cognitive and Behavioral Neuroscience (Durnez, Moerkerke, Bartsch, & Nichols, 2013).

2 Multiple testing in fMRI: an empirical case study on the balance between sensitivity, specificity and stability. [1]

Abstract Functional Magnetic Resonance Imaging is a widespread technique in cognitive psychology that allows visualizing brain activation. The data analysis encompasses an enormous number of simultaneous statistical tests. Procedures that either control the familywise error rate or the false discovery rate have been applied to these data. These methods are mostly validated in terms of average sensitivity and specificity. However, procedures are not comparable if requirements on their error rates differ. Moreover, less attention has been given to the instability or variability of results. In a simulation study in the context of imaging, we first compare the Bonferroni and Benjamini-Hochberg procedure. Considering Bonferroni as a way to control the expected number of type I errors enables more lenient thresholding compared to familywise error rate control and a direct comparison between both procedures. We point out that while the same balance is obtained between average sensitivity and specificity, the Benjamini-Hochberg procedure appears less stable. Secondly, we have implemented the procedure of Gordon, Chen, Glazko, & Yakovlev (2009) (originally proposed for gene selection) which includes stability, measured through bootstrapping, in the decision criterion. Simulations indicate that the method attains the same balance between sensitivity and specificity. It improves the stability of Benjamini-Hochberg but does not outperform Bonferroni making this computationally heavy bootstrap procedure less appealing. Third, we show how stability of thresholding procedures can be assessed using real data. In a data set on face recognition, we again find that Bonferroni renders more stable results.

[1] This chapter is based on:
Durnez, J., Roels, S.P. and Moerkerke, B. (2014). Multiple testing in fMRI: an empirical case study on the balance between sensitivity, specificity and stability. *Biometrical Journal*, 56 (4), 649–661

2.1 Introduction

Functional neuroimaging plays an important role in current research that unravels the functioning of the human brain, not only within medical sciences but also within cognitive, clinical and social psychology (Dolan, 2008; DeCharms, 2008). Functional Magnetic Resonance imaging (fMRI) is a neuroimaging technique that allows to measure brain activity in a non-invasive way. An fMRI study leads to the production of enormous amounts of data. The brain is divided in over 200 000 volume units (voxels) for which a signal is measured on a series of time points. In a univariate, voxelwise approach, a statistical test is performed for each voxel, resulting in over 200 000 statistical tests. As such, one is confronted with a multiple testing problem.

Neural activity in the brain leads to an increase in the amount of blood flowing through the active area and the heart sends more oxygenated blood to the activated region resulting in locally higher blood oxygen levels. The signal measured in functional MRI (fMRI) is a correlate of this oxygen level, and is called the blood oxygenation level dependent (BOLD) signal. In an fMRI experiment, a time sequence of brain images is obtained while participants perform a series of tasks under different conditions inside an MRI scanner. Changes in BOLD signal between such different conditions are used to localize task related areas in the brain.

In what follows, we consider testing for activation in a single subject. A moderate resolution of fMRI data is a brain of more than 200 000 voxels, e.g. $m = 64 \times 64 \times 64$ voxels. $\mathbf{Y_i}$ denotes a vector containing the time series (with n timepoints) of measured BOLD signal in voxel i ($i = 1, ..., m$). In practice, preprocessing steps are often performed to correct this signal for artefacts such as head motion etc.

The data is often spatially smoothed as a preprocessing step, as this greatly increases the signal-to-noise ratio which improves sensitivity for detecting true activation. A possible drawback is the reduction of spatial accuracy in detecting activation.

In the classical general linear model (GLM) approach, a time series of BOLD-signal $\mathbf{Y_i}$ for each voxel i is modeled as a linear combination of different signal components, presented by the columns in the design matrix $\mathbf{X}$. Temporal correlation is removed through pre-whitening and a linear

model is fit to test specific parameters evaluating evidence for activation. Subsequently, the effect of the different conditions on the BOLD-signal is expressed by a statistical parametric map (SPM) of the brain with a T-statistic or corresponding p-value for each voxel. A more extensive clarification on the data structure and analysis of an fMRI experiment can be found in the Supplementary Material.

Many procedures have been proposed for voxel based testing that either control the familywise error rate (FWER; Friston et al., 1991) or the false discovery rate (FDR; Genovese, Lazar, & Nichols, 2002). Most of them can be seen as an extension or specific case of the basic methods for FWER and FDR control, which are the Bonferroni correction (BF; Bonferroni, 1936) and the Benjamini-Hochberg correction (BH; Benjamini & Hochberg, 1995). It is often argued that the FWER is too conservative as an error rate in a large-scale test setting and that FDR control is more sensible. However, both BF and BH attain type I error rate control through p-value thresholding. Hence, FDR control only results in a higher sensitivity by imposing a lower specificity. Consequently it can be expected that the average power for comparable p-value cutoffs is equal for both FWER and FDR control. Gordon, Glazko, Qiu, & Yakovlev (2007) state that the conservative nature of the BF procedure is not due to the procedure itself but due to the requirements that are posed on the procedure. They advocate the procedure as a way to control the per family error rate (PFER), defined as the expected number of type I errors. This makes the procedure intrinsically less conservative.

Not only the average performance of testing procedures is important but also their contribution to the variability on test results, which indicates instability of the testing procedure. While studied in the context of genetics (Qiu, Xiao, Gordon, & Yakovlev, 2006; Gordon et al., 2007, 2009; Hommel & Bretz, 2008; Owen, 2005), this aspect has remained largely unexplored in the context of fMRI data analysis until recently (see for example Bellec, Rosa-Neto, Lyttelton, Benali, & Evans, 2010). If a brain region is declared to be active based on statistical criteria, it is key to have a measure of the reproducibility of these results. Stable procedures yield a low variability on the number of selected voxels. Even when the activation is the same in different datasets, the number of selected voxels can be very different due to differences in the noise. Instability in the

testing method is expressed by a high standard deviation on the number of selected voxels in relation to the average number of selected voxels. As the outcomes of statistical tests are random variables, stability can be expressed in terms of the variance of the number of selected voxels (Gordon et al., 2007) or the variance of the number of erroneously accepted or rejected hypotheses (Gordon et al., 2009). In their study on the assessment of stability in the context of gene selection, Qiu et al. (2006) find that FWER controlling procedures tend to be more stable than FDR controlling procedures. Hommel & Bretz (2008) show that the classical BF is more stable than its adjusted forms, more in particular the Bonferroni-Holm procedure (Holm, 1979) as well as the Simes procedure (Benjamini & Hochberg, 1995). Owen (2005) give a theoretical insight in the fact that dependency between tests (which is a characteristic very present in fMRI) increases the instability of the testing procedure.

Also in the context of genetics, Gordon et al. (2009) measure the selection stability of a feature as the frequency of selection of the feature over different bootstrap samples of the data in which the original thresholding is repeated. They propose a new selection criterion that includes this stability measure with the aim to attain a better balance between sensitivity and specificity and a higher stability.

In this paper, we provide new insights into the operating characteristics of FWER and FDR thresholding in fMRI data analyses. We set up a simulation study in an fMRI context to study the balance between sensitivity and specificity for the BF and BH procedure. We also evaluate the procedures in terms of their stability properties. Secondly, we implement the method of Gordon et al. (2009) - GCGY, for application in an fMRI context. Our extensive simulation study enables a thorough evaluation of the performance of the procedure. Simulation results to assess performance are more restrictive in the work of Gordon et al. (2009). We further demonstrate the assessment of stability of selection procedures in a real data set on face recognition.

This paper is organized as follows. In section 2, we briefly introduce the data structure and test setting in an fMRI study. More details on the simulation procedure can be found in the Supplementary Information. In section 3, we describe different methods for evaluating multiple testing procedures. We further outline the testing principle of Gordon et al. (2009)

in an fMRI context. Results of an extensive simulation study comparing BF, BH and GCGY are given in section 4. In section 5, we illustrate how the stability of BF and BH is assessed in a real data set on face recognition.

2.2 Data

2.2.1 Simulation procedure

In each simulation step, a 3-dimensional map consisting of $30 \times 30 \times 30$ voxels is constructed. Every voxel contains a time series of $n = 120$ time-points. A block of $14 \times 14 \times 14$ voxels in the center is activated. A blocked design with 2 conditions (on and off) is used, with 3 blocks of 40 seconds (20 seconds on/20 seconds off). The BOLD signal of the activated voxels is simulated such that it correlates with the ON-OFF status of the blocks. The signal-to-noise ratio is 0.025, such that the effect size is 0.025 and the noise follows a standard normal distribution. Temporal dependence is added in the noise with an auto-correlation coefficient of $\rho = 0.20$. Data are spatially smoothed before analysis. The relation between the smoothness of the data and the spatial correlation is further explored in the Supplementary Information. As this process combines the signal in each voxel with the signal of surrounding voxels, this imposes a spatial correlation among voxels. We use a smoothing kernel with a full width half maximum (FWHM) of 2 times the voxel size. In the Supplementary Material, results are also presented for different smoothing kernels and hence different degrees of spatial correlation. It is important to note that smoothing also spreads out the activation outside the borders of the original activated block. In a simulation setting, this complicates deciding on false and true positives. We have chosen to only consider the original activation block as truly activated. Another option is to look at the extend of true activation in neighboring voxels, however, this typically involves an arbitrary activation cutoff to decide on true activation. To avoid this problem, one could also consider a simulation setting in which only the noise (and not the signal) is smoothed but this renders less realistic data.

After performing the whitening procedure from Glaser & Friston (2007), T-statistics are obtained for each voxel using the classical GLM procedure, where the measured signal is related to the design of the experiment.

Both BF and BH and their bootstrap versions as described in Gordon et al. (2009) are applied as a thresholding technique for each simulated data set. The simulation procedure is repeated 500 times. The data are analyzed in R with the routines used in the software package SPM.

2.2.2 Real data application

The data used for this study are from a single-subject event-related design where faces were presented against a checkerboard (see Henson, Shallice, Gorno-Tempini, & Dolan, 2002). The goal of the analysis is to find the precise location in the brain responsible for recognizing faces. The data set consisted of $64 \times 64 \times 64$ voxels. Two sessions of 356 volumes were acquired, with a repetition time of 2 seconds per volume.

2.3 Methods

2.3.1 Problem setting and aim

A classical BF correction for FWER control at level α rejects all null hypotheses for which the p-value is smaller than α/m resulting in (extremely) low p-value thresholds. While BF is typically advocated as a FWER controlling procedure, the procedure also controls the PFER at the same level (Gordon et al., 2007; Lee, 2004). Based on the fact that the BF controls the PFER, the BF correction can be used to produce p-value thresholds comparable to those in FDR controlling procedures. Therefore, we vary p-value thresholds such that the PFER ranges between 0 and m for BF and the FDR between 0 and 1 for BH. This allows a comparison of the different procedures in terms of sensitivity, but more importantly, this makes it possible to study the operating characteristics of both procedures when requirements on their error rates are comparable (Gordon et al., 2007).

In this paper, our aim is to compare the BF and BH procedure by means of a simulation study in the context of fMRI. As opposed to what is typically done in literature, we will do this for a wide range of requirements on the error rates of the procedures and we will not only focus on average performance but also on the stability. This will help to nuance claims in

literature that one procedure uniformly outperforms another (Gordon et al., 2007).

2.3.2 Evaluation criteria for multiple testing procedures

First, we evaluate the multiple testing procedures based on the receiver-operator characteristic (ROC). Gordon et al. (2007) show that the outcomes of the BF procedure and the FDR controlling BH procedure (Benjamini & Hochberg, 1995) are highly correlated if requirements on their error rates become comparable. Therefore, when comparing different methods, it is important to review the methods for a range of p-value thresholds. This can be done by an ROC analysis, in which the true positive rate (TPR) is plotted against false positive rate (FPR), as one varies a threshold. The quality of a method can be quantified as the area under the curve. The larger the area, the better the balance between sensitivity and specificity (Skudlarski, Constable, & Gore, 1999). To further investigate the findings of Gordon et al. (2007) in the context of fMRI data, we compare an ROC curve for both procedures in a simulation study.

Secondly, we evaluate the multiple testing procedures with respect to their stability. In line with Qiu et al. (2006), we mainly consider two stability measures: the variability on the number of selected voxels and the frequency with which a given voxel is selected over different samples. In a simulation setting, we observe the number of selected voxels and their standard deviation. A high standard deviation on the number of selected voxels in relation to the average number of selected voxels indicates instability in the testing method. In the simulations, we also consider the variance on the different error rates. Selection frequency is studied using a real data set from which we subsample through a bootstrap procedure. The most stable voxels are those that are selected with a high frequency over different bootstrap samples.

2.3.3 Incorporating stability into the decision criterion

As long as we focus on p-value thresholding, we can expect that the average sensitivity is equal when the average specificity is the same. More specific, when thresholding p-values, there is a fixed monotonicity between different hypotheses (i.e. different voxels). Consequently, less specificity

will be accompanied with a fixed increase in sensitivity. Aiming to balance sensitivity and specificity, Gordon et al. (2009) propose to include frequency of selection into the decision criterion. Their approach proceeds as follows: a fixed multiple testing procedure is performed on a number of different sets of features (in their case: genes, in our case: voxels), with each set approximating the original data. The threshold h is a proportion that reflects a predefined selection frequency, i.e. the proportion of selection over the different sets. Only those features are selected that are declared significant in at least h $\times$ 100 % of the different sets based on the used multiple testing procedure. Since this method uses information on reproducibility, one expects higher stability. Gordon et al. (2009) show in their paper how the frequency of selection directly controls the balance of type I and II errors. More in particular, they indicate how h represents the ratio of the penalty for a type I error versus the sum of the penalties for type I and type II errors. Selecting features that are selected in at least h $\times$ 100% the different sets, minimizes the corresponding penalty function of type I and type II error rates. For example, the choice of a threshold h equal to 0.7 implies that only features are selected that are chosen in at least 70% of the different sets. This corresponds to selecting a set of features in such a way that a penalty function is minimized in which the relative costs of type I and type II errors are respectively 70% and 30% of the total cost of an error.

As a test statistic is used other than p-values, which allows to optimize a gain function that balances type I and type II errors, it is of interest to study the overall balance between sensitivity and specificity. To obtain different sets of the same data, we follow the approach of Gordon et al. (2009) and bootstrap from the original data. To implement their procedure, we perform the following steps.

1. *Whitening procedure.* The temporal correlation that disallows to resample the data over time points, is removed. As the data in fMRI research represents a time series of blood flow intensities for each voxel, the data is affected by temporal correlation. When $\mathbf{Y_i}$ represents the measured signal in voxel i, a matrix $\mathbf{W}$ is determined such that WY_i represents the signal without temporal correlation. W can be found using restricted MLE estimation as described in Glaser & Friston (2007).

2. *Bootstrap procedure.* For each voxel, a linear regression model is fitted on the whitened time series, i.e. $\mathbf{WY_i}$ is regressed on $\mathbf{WX}$. Next, a bootstrap sample is taken from the residuals and these residuals are added again to the estimated activation. By 'unwhitening' the data, we obtain a realistic new data set. This second step is repeated 100 times, in order to obtain 100 new data sets.

3. *Activation detection.* For each of the bootstapped data sets, a classical thresholding procedure (here: BF and BH) is used. The critical threshold h selects all voxels that are called significant in more than $(h \times 100)\%$ of the bootstrap samples.

For this procedure, we consider 4 different values for the selection criterion h: 0.10, 0.50, 0.75 and 0.90. A more technical explanation of the resampling procedure can be found in the Supplementary Information.

In what follows, we refer to BF and BH procedures that are applied on the original (i.e. not bootstrapped) data as the classical procedures. The procedures based on the frequency of selection among the bootstrap samples, are referred to as the bootstrapped procedures. We will compare respectively BF and BH with their bootstrapped procedures as proposed by Gordon et al. (2009), on both their average performance and stability.

2.4 Results

2.4.1 Simulated data

The ROC curves for the BF and the BH procedure are displayed in panels A and B in Figure 2.1. For a range of thresholds, an ROC curve can be constructed. As can be seen in panels A and B in Figure 2.1, the BF and BH procedures produce the same ROC curve as expected. That is, the relation between sensitivity (operationalized as the TPR, the true positive rate) and specificity (defined as 1-FPR, the false positive) is on average the same for both methods. We see that FWER and FDR control at a 5% level correspond to different points on the ROC curve. As indicated on panels A and B in Figure 2.1, an FWER control of 0.05 logically leads to a much more stringent result than a FDR control of 0.05, resulting in fewer false positives, but also less power. We can conclude that the difference in

average performance between FWER and FDR control can be explained by the difference in threshold selection, which is an arbitrary choice made by the researcher. This confirms the intuitive findings of Gordon et al. (2007).

Interestingly, the bootstrapped procedure also results in an equal ROC for BH and BF, while deciding on activation is not purely based on p-value thresholding. For procedures that use p-value thresholding, we expect equal ROCs as the order in which voxels are declared significant is equal. However, the bootstrapped procedures are not solely based on p-value thresholding but also incorporate information on stability. Given the ROC remains the same as for the classical procedures, the bootstrapped methods are not able to detect activation with more power than the classical methods. As can be seen on Figure 2.1 and indicated by Gordon et al. (2009), a h-threshold reflects the relative cost of type I and type II errors. By fixing a decision criterion (e.g. FWER=0.05), choosing different h-values allows to move along the ROC-curve, but this can also be achieved by choosing different significance levels for the BF or BH procedures. In the example on Figure 2.1, it can be seen for BH that a higher h-threshold of 0.90 leads to a lower sensitivity but a higher specificity, while a lower h-threshold (0.50) results in higher sensitivity but lower specificity than the classical procedures. According to Gordon et al. (2009), this enables to balance type I and type II errors with pre-specified penalties to both errors which implies that the h-value does not directly reflect the obtained balance between sensitivity and specificity. This makes the results of the bootstrapped procedures more difficult to interpret in practice.

An intuitive interpretation as to why the bootstrapped procedures do not change the overall balance between sensitivity and specificity is as follows. Suppose for each voxel, we have a normally distributed test statistic t. As the bootstrap procedure adds noise obtained by subsampling to the observed effect size, a new distribution of T-values is constructed where the non-centrality parameter equals the original statistic t. If we fix the threshold on the frequency of selection at $h = 0.5$, a voxel is declared to be active as soon as half of the T-values in the bootstrap distribution is larger than t_α with t_α representing the statistical threshold corresponding to either the BF or BH procedure. This implies that the median of the T-values is larger than t_α but in case of a symmetric distribution,

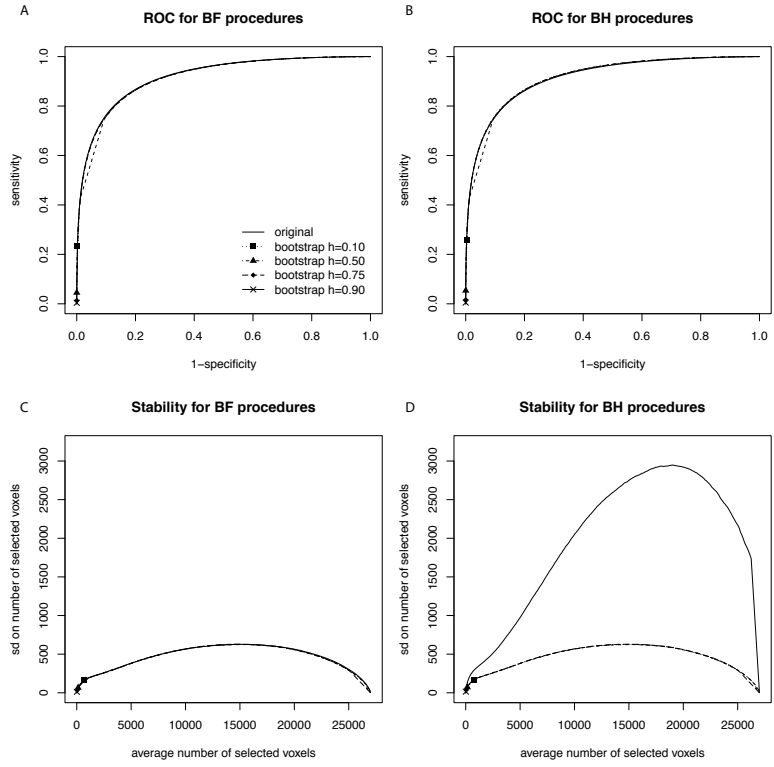


Figure 2.1: Roc curves (upper panel) for FWER / PFER control (BF correction) and FDR control (BH correction). Stability evaluation (lower panel) for FWER control (BF correction) and FDR control (BH correction). A high standard deviation (sd) of the number of selected voxels in relation to the average number of selected voxels points at the instability of the testing procedure. The four different lines in the figures correspond to the four procedures compared: the original procedure and the bootstrap procedure with $h = 0.10$, $h = 0.50$, 0.75 or 0.90.

the median is equal to the observed test statistic t. Hence, a voxel is declared active as soon as $t > t_\alpha$ which coincides with a classical BF criterion. When h is different from 50%, controlling a probability under the bootstrapped distribution of T-values is essentially the same as selecting voxels with $t > t_{\alpha^*}$ where α^* differs from the chosen α in the original BF procedure. With a symmetrical distribution for the test statistics, the bootstrapped procedures asymptotically converge to the BF procedure and hence, choosing different values of h only allows to shift along the ROC curve.

Based on this reasoning, we expect that control with $h = 0.50$ will correspond asymptotically to the original thresholding. In our example, we have a finite number of time points which explains why the original procedure and its bootstrapped versions do not render identical results.

Figure 2.1 also shows the original procedure together with the bootstrapped methods for a specific FWER and FDR control of 0.05. We see that whereas in the original procedure, BH is much more liberal than BF, this difference is smaller to nonexisting for the bootstrapped methods. Consequently, the bootstrapped procedures for BH makes the original specificity higher for all h levels. On the other hand, the specificity for BF is at the same level for the original procedures as for the bootstrapped procedure.

In panel C and D of Figure 2.1, the standard deviation on the number of selected voxels is plotted against the number of selected voxels for the BF and BH procedure. It can be seen that BF is more stable than BH. This could be due to the adaptive character of FDR control. As the number of voxels is the same in each simulation step, the BF threshold remains the same when controlling the PFER at a fixed level. For FDR control using the BH procedure, the threshold is adaptive to the data. As we expect that the bootstrapped procedure asymptotically converges to a classical criterion with fixed t-value threshold, the procedure results in the same level of stability as BF. The instability of BH compared with BF cannot be attributed to a difference in the expected FDR, as is shown in section 4 of the Supplementary Information.

It can further be seen in Figure 2.1 that for the BF procedure, there is no improvement in variability when using the bootstrapped procedure. However, there is a great improvement in stability for the BH procedure.

Table 2.1: Averages and standard deviations (SD) over 500 simulations. The variation coefficient that divides the SD on the number of voxels selected by the mean number of voxels selected is also given (SD/Nr.). m denotes the number of tests (voxels).

	Number of voxels selected (SD, SD/Nr.)	FPR (SD)	TPR (SD)
FWER ≤ 0.05	27.382 (25.024 , 0.914)	0 (0)	0.01 (0.009)
FDR ≤ 0.05	627.128 (270.303 , 0.431)	0.002 (0.001)	0.211 (0.089)
PFER/m ≤ 0.05	3350.398 (299.55 , 0.089)	0.061 (0.009)	0.683 (0.057)

The instability of the procedure, can almost be reduced up to the level of BF. This is in line with what can be expected based on the reasoning above.

The fact that the bootstrapped procedure retains the same balance between average sensitivity and specificity, is computationally heavy, and does not outperform BF in terms of stability, does not make the procedure appealing in an fMRI context.

For the number of selected voxels, the FPR and the TPR, the average and variability over simulations are obtained. Results for specific realistic thresholds are summarized in Table 2.1. In addition, we also show results for controlling the per comparisonwise error rate, i.e. $PFER/m$, at a fixed level. Logically, FWER control is more conservative than FDR control, and PFER control more liberal. We find that the variation coefficient (defined as the standard deviation of the number of selected voxels divided by the number of selected voxels) on the number of selected voxels is higher when controlling the FDR at 5% than controlling the PFER at 5%. This is remarkable, given the fact that control of the PFER is much more liberal and still exhibits a smaller variation coefficient than the FDR control. This reflects what is shown in Figure 2.1 for a range of cutoffs.

However, it should also be noted that FWER control at 5% exhibits a larger variation coefficient than FDR control at 5%, despite that BH overall being less stable than BF. This can be explained by the fact that BF control at 5% is more stringent than BH control at 5%. As can be seen in Figure 2.2, the variation coefficient decreases as procedures gets more liberal.

For now, we have aligned the BF and BH procedure using the number of selected voxels. However, we can show the same results, aligning

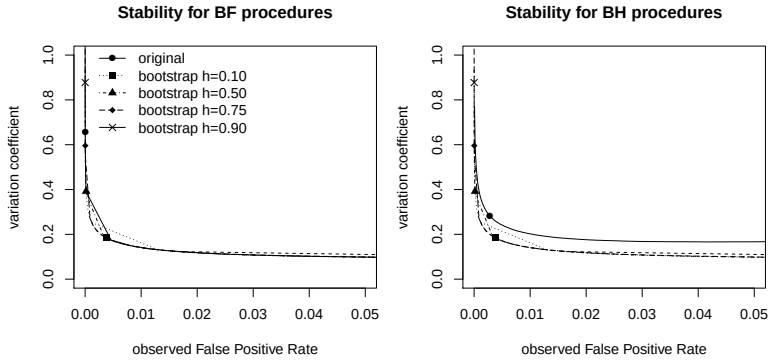


Figure 2.2: The variation coefficient (i.e. standard deviation on number of selected voxels divided by the number of selected voxels) in function of the observed False Positive Rate FPR. Given a fixed level of the observed FPR, BH results in a higher variation coefficient. However, when comparing FWER and FDR control at 5% by respectively BF and BH (indicated by the dots), BF results in a higher variation coefficient than BH.

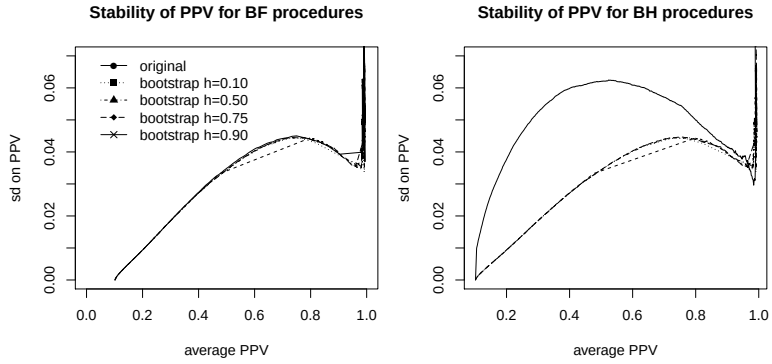


Figure 2.3: Stability evaluation for FWER control (BF correction) and FDR control (BH correction). A high standard deviation (sd) of the number of selected voxels in relation to the posterior predictive value (PPV) points at the instability of the testing procedure.

the procedures using another metric. For example, Figure 2.3 shows the variability of the posterior predictive value as a function of the posterior predictive value PPV which is the proportion of correctly rejected voxels out of total voxels rejected. It can be seen again that BH elicits the highest variability, but logically, using the bootstrap procedure, this instability is decreased.

An important remark should be made on the dependence of the tests. Storey (Storey & Tibshirani, 2003) show that that the threshold of the BH procedure converged almost surely to a fixed number for increasing number of hypotheses. We have done an additional simulation study, where we investigate this further. This can be found in the Supplementary Information.

2.4.2 Real data application

In order to compare the performance of the BF and BH procedure in terms of stability, a GLM is fitted on preprocessed data and an FWER and FDR threshold is applied on the SPM. The thresholds for the BF and BH

procedures are chosen so that both procedures result in the same number of significant voxels (2530 voxels). This leads to a PFER of 154 and an FDR control of 0.001. Subsequently, we obtain 100 replications of the data by bootstrapping from the original data. For each voxel, the frequency of selection over bootstrap samples is determined for both thresholds. We threshold this selection frequency (expressed in % of the number of bootstrap samples) at 80%. Figure 2.4 shows images of the resulting SPMs.

It can be seen that PFER control leads to the selection of more "stable voxels" (with frequency of occurrence $> 80\%$). This is further demonstrated in Figure 2.5, where the average frequency of occurrence is depicted against the uncorrected p-values. We see that the voxels selected by PFER control are uniformly more stable than those selected by FDR control as the voxels with smaller p-values are more frequently selected with the BF procedure.

We have repeated this analysis for FWER control at the 5% level in the Supplementary Information. The same conclusions can be drawn.

2.5 Discussion

Many publications consider the use of multiple testing procedures for voxelwise fMRI data analysis which control either the FWER or the FDR (Logan & Rowe, 2004). While it is often claimed that FDR control is more powerful than FWER control, we have demonstrated through a simulation study that the actual trade-off between sensitivity and specificity for the BF and BH procedure is equal. Hence, BF can be made as powerful as the BH procedure by a proper choice of its threshold, a conclusion also made by Gordon et al. (2007) and Korn, Troendle, Mcshane, & Simon (2004). As such, the procedure should be more generally regarded as PFER controlling. While one could argue that the PFER ignores the multiplicity of tests completely, we advocate a well contemplated choice of threshold since the same (both liberal and conservative) results can be obtained through PFER and FDR control.

A distinct advantage of the BF procedure is its higher stability. Selection stability relates to the reproducibility of scientific results. Simulation results indicate that the variance on the number of selected voxels is overall higher for BH than for BF when considering a range of thresholds.

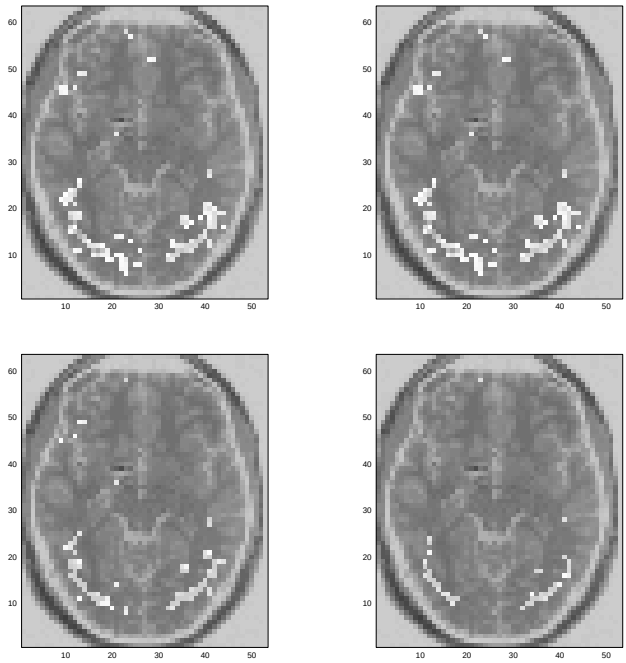


Figure 2.4: Results of the analyses on the Henson data set. On the left are the analyses with BF correction, (PFER=154). On the right are the analyses using the BH procedure (FDR=0.001). Above are the results with classical analyses with the voxels exceeding the significance level highlighted. The plots below display the percentage of selection over bootstrap samples with the "stable voxels" (with frequency of occurence $> 80\%$) highlighted. The thresholds are chosen to make sure that in the images above, the same number of voxels are significant.

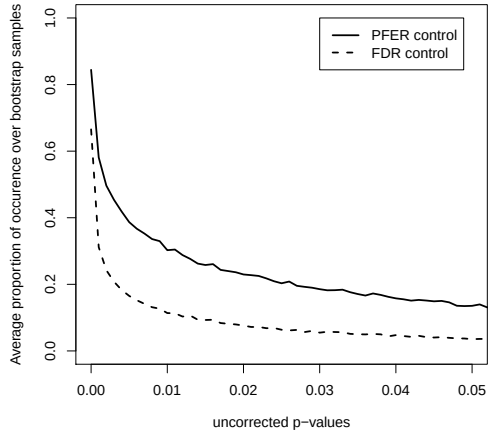


Figure 2.5: Henson data: Average selection proportion as a function of uncorrected p-values.

In real data, we find that PFER (or FWER) control leads to more stable voxels (with frequencies of more than 80%) than FDR control when selection criteria are comparable.

Gordon et al. (2009) use the frequency of occurrence as a selection procedure in its own right. We have applied this technique in the context of fMRI for a large range of possible cutoffs, thereby providing more insight into the characteristics of the procedure. In our simulation study, we find an ROC curve that is virtually identical to the ROC curve of the original BF and BH procedure. For a fixed threshold (e.g. FDR=0.05) however, a different point on the ROC is obtained when taking selection frequency into account. The finding of identical ROC's for the original and bootstrapped versions of the procedures is because the two procedures will asymptotically converge, and therefore again the balance between sensitivity and specificity cannot be improved. We further find an advantage for the bootstrapped procedures in terms of the stability. Combining FDR control with frequency of selection renders more stable results, almost up to the level of FWER control.

While including frequency of selection improves selection stability, we conclude that the procedure is not practically appealing in an fMRI context. First, the improvement in stability is small to non-existent for the BF procedure. Secondly, the procedure is very computer intensive given the dimension of the data sets in brain imaging. Finally, the ROC curve is the same as for the BF and BH procedure. In this paper, we have shown that using the BF procedure as a PFER controlling procedure allows a different trade-off between specificity and sensitivity. The procedure also demonstrates a high stability, excluding the need for bootstrapped procedures. We show that these results hold for different levels of smoothness and are therefore independent of the correlation between the measured signal in different voxels.

These conclusions however can only be drawn for Bonferroni and Benjamini-Hochberg procedures. Other procedures controlling the FWER and FDR have been proposed in literature. For example for FDR control, some authors have discussed controlling quantiles of the false discovery proportion, rather than controlling it in expectation, leading typically to more conservative inference (Wasserman & Genovese, 2004). We do not claim that the same conclusions can be drawn for these procedures controlling the FDR.

In this paper, we only consider voxelwise detection of activation through a general linear model approach. Statistical techniques that take into account the spatial correlation are also available. Some of these methods (e.g. smoothing raw test statistics, Logan, Geliazkova, & Rowe, 2008) also apply voxelwise thresholding. Hence, the findings in this study also translate to these more general settings. Random field theory (Friston et al., 1991; Chumbley, Worsley, Flandin, & Friston, 2010) is often used in fMRI to obtain a more precise estimate of the FWER by taking into account the statistical dependence between voxels as opposed to BF which may become overly conservative. Nevertheless, Random field theory suffers from many burdensome assumptions (Nichols & Hayasaka, 2003) and can only be used for low p-value thresholds which excludes the possibility to study performance for a wide range of thresholds. Other spatial methods perform inference on topological features, such as peaks or clusters, rather than on voxels (Chumbley et al., 2010). Further research is needed to see if the conclusions in this study are also tenable for topological testing

procedures. Chumbley et al. (2010) point out the difficulty to assess error rates for such procedures. Morever, Nichols (2012) argues that peaks or cluster thresholding makes strong assumptions on the smoothness of the data, and therefore voxelwise activation detection remains a useful tool in fMRI.

For bootstrapping fMRI time series, we use a residual bootstrap on decorrelated data. Other procedures, such as Fourier, Wavelet (Friman & Westin, 2005) and block bootstrap (Lahiri, 2003) resampling are possible, but few is known on the properties and results of different resampling techniques for the specific case of fMRI data. Another issue in this study is the use of smoothing within the resampling setting. In an fMRI data analysis, the signals are smoothed before analysis, leading to a 'smeared out' image. In this paper, resampling is done on this smoothed data. However, the order of smoothing and resampling can be inversed. Further research is needed to compare the different smoothing models when resampling fMRI data.

Acknowledgements

The computational resources (Stevin Supercomputer Infrastructure) and services used in this work were provided by Ghent University, the Hercules Foundation and the Flemish Government - department EWI.

Appendices

2.A fMRI analysis

Neural activity in the brain leads to an increase in the amount of blood flowing through the active area and the heart sends more oxygenated blood to the activated region resulting in locally higher blood oxygen levels. The signal measured in functional MRI (fMRI) is a correlate of this oxygen level, and is called the blood oxygenation level dependent (BOLD) signal. During an fMRI experiment, subjects are placed inside the scanner. Their brain is divided in a large number of volume units or voxels, where $64 \times 64 \times 64$ voxels corresponds with an average resolution. For each voxel, the BOLD signal is measured during the timcourse of the experiment. For a single subject, a 4D image is obtained containing the BOLD signal, with the first 3 dimensions the spatial dimensions in the brain, and the fourth dimension the different timepoints at which measurements have been taken. $\mathbf{Y_i}$ denotes the vector of time series of the BOLD signal for voxel i $(i = 1, ..., m)$ with n time points. As the neural activation invokes the blood flow, the BOLD-signal represents the neural activation with a certain lag following a specific haemodynamic response function (HRF). The true HRF is often unknown and voxelspecific, but the canonical HRF is closely related to the true HRF.

During the experiment, subjects perform a task evoked by stimuli, under specific conditions. The most simple and straightforward method is a block design, in which stimuli are presented continuously for a longer period of time, alternated with a block without stimulation. An example is listening to sounds alternated with silence. The design is expressed in the design matrix $\mathbf{X}$. To take into account the lag with which the blood runs, $\mathbf{X}$ is convolved with the canonical HRF.

In order to relate $\mathbf{Y_i}$ to the design, a linear regression analysis of the time series of the BOLD signal on the design matrix is performed for each voxel, resulting in a three-dimensional T-map, which is called the statistical parametric map (SPM).

2.B Simulations

Simulated data A 3-dimensional map consisting of $30 \times 30 \times 30$ voxels is constructed. For each voxel, a time series of $n = 120$ timepoints is constructed. The image in this stage is 4 dimensional, with the first 3 dimensions referring to space, while the fourth dimension contains the timing information. A blocked design with 2 conditions (on and off) is used, with 3 blocks of 40 seconds (20 seconds on/20 seconds off), the design is represented in Figure 2.6.

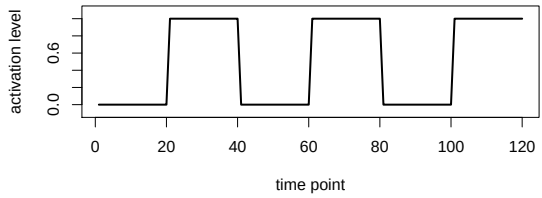


Figure 2.6: Graphical representation of the design used in the simulated experiment.

About 10% of the voxels are actively related to the design, all gathered in an activated region of $14 \times 14 \times 14$ voxels in the center of the image. For the activated voxels, the activation level is presented in the upper left panel of Figure 2.7. The height of the activation depends on the signal-to-noise ratio, which is defined as $\bar{S}/\sigma_{tnoise}$, with $\bar{S}$ the mean signal intensity and σ^2_{tnoise} the non-task-related variability over time. The SNR is set to 0.2. Consequently, we have an activation height of 0.2 and σ^2_{tnoise} is equal to 1. The activation is convolved with a canonical hemodynamic response function,

$$f(x) = \frac{1}{28.48909}\left(\left(\frac{x}{6*0.9}\right)^6 e^{\frac{-(x-6*0.9)}{0.9}} - 0.35\left(\frac{x}{12*0.9}\right)^{12} e^{\frac{-(x-12*0.9)}{0.9}}.\right) \tag{2.1}$$

This results in a BOLD-signal without noise $\mathbf{Y}_{bold}$ for each voxel, as

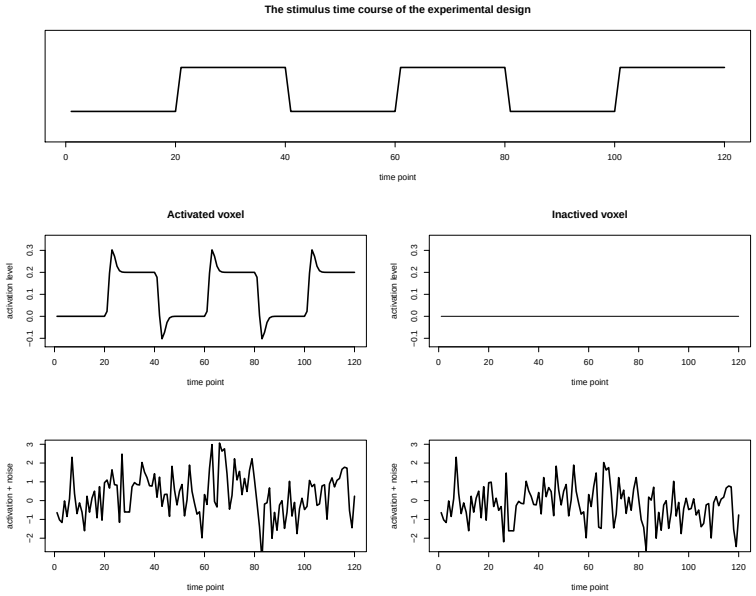


Figure 2.7: Graphical representation of the simulated time series in an active (left) and an inactive voxel (right). The upper panel represents the raw activation. The middle panels represent the convoluted activation, and the lower panels represent the resulting time series after adding the noise (before smoothing).

shown in the middle left panel of Figure 2.7.

In each simulation, noise is added to the image, consisting of temporal correlated noise. The noise follows an autoregressive model AR(1), where the error term for one time point is partially dependent on the error term of the previous timepoint. The temporal dependence is fixed to $\rho = 0.2$. The error term for each voxel i on timepoint j is created using Equations 2.2 and 2.3.

$$\epsilon_{ij} = \rho\epsilon_{i,j-1} + \xi_{ij} \tag{2.2}$$

$$\xi_{ij} \sim N(0, \sigma^2_{tnoise}) \tag{2.3}$$

The resulting time series for an active and inactive voxel are shown in Figure 2.7. Before analysis, the data are smoothed with a Gaussian kernel with a full width at half maximum (FWHM) 2 times the size of the voxel dimension.

Analysis Once the simulated data are obtained, the data are analyzed. The data are analyzed in R (R Core Team, 2014) with the routines used in the software package SPM (Ashburner, Friston, & Penny, 2012). First the whitening procedure used in the software package SPM estimates and removes the temporal dependence. Consequently, T-statistics are obtained for each voxel using a classical GLM procedure. Both BF and BH are applied as a thresholding technique for each simulated data set, as well as the bootstrapped procedure.

P-value thresholds are varied such that the PFER ranges between 0 and m for BF and the FDR between 0 and 1 for BH. For the bootstrap procedures, the selection criterion h is fixed at 3 different values: 0.50, 0.75 and 0.90.

Each simulation procedure is repeated 500 times.

2.C Resampling

After whitening the data to remove temporal correlation, a linear regression model is fit for each voxel i, using a least squares algoritm, $\mathbf{Y}_i = \mathbf{X}\hat{\beta}_i + \mathbf{e}_i$, with $E(e_i) = 0$. As a result, $\mathbf{Y_i}$ is split up into the activation $\mathbf{X}\hat{\boldsymbol{\beta}}_i$ and the residuals $\mathbf{e_i}$. From the residuals $\mathbf{e_i}$, a random bootstrap sample with replacement $\mathbf{e_i}^* = (e_{i1}^*, e_{i2}^*, \cdots, e_{in}^*)$ is taken, with n the number of time points. These residuals are then added to the fitted values:

$$\tilde{\mathbf{Y}}_\mathbf{i}^{\,*} = \tilde{\mathbf{X}}_\mathbf{i}\hat{\beta}_i + \mathbf{e_i}^*.$$

Resampling is performed for all voxels simultaneously by bootstrapping time points from entire images. This means that for a new resampled data set, Y_i^* is the same as in the original data for each voxel i. e_i is constructed by resampling e_i over time points such that the resampled time points are equal for all voxels. In this way, the spatial structure in the original data is accounted for. This results in a replicate of the whitened vector $\mathbf{Y_i}$. To construct a replicate of the original data matrix, temporal correlation is again added to the data by performing the inverse operation of the original whitening procedure.

We can now repeat the bootstrap procedure ℓ times. This results in ℓ replicates of the original data set. For every replicate, a thresholding procedure is performed. A voxelwise GLM can be performed to test for activation, resulting in a T-statistic for each voxel. p-values are obtained under the assumption of normally distributed error terms. Each statistical parametric map (SPM) is then thresholded according to a specified procedure. After thresholding all ℓ replicates, there are ℓ decisions on whether a voxels is selected or not. In that way, the frequency of occurrence across subsamples can be obtained.

High frequencies of occurrence (e.g. 90%) indicate stability: the voxels are chosen in most of the resamples. Lower frequencies indicate a lower stability: these voxels are chosen a few times in the different resamples.

2.D Difference in observed FDR as possible explanation for the difference in stability.

We compared both methods in terms of stability by plotting the variability on the number of selected voxels against the average number of selected voxels. Therefore, we assume that the average number of selected voxels is the base for a fair comparison between both methods. However, one could reason that the false discovery rate for a fixed average number of selected voxels is different for BF and BH. In this case, fixing the average number of selected voxels would not allow to compare both procedures. Consequently, it is useful to not only compare the methods based on the average number of selected voxels, but also based on the average number of false discoveries. Therefore we plotted the false discovery rate observed in our simulations against the average number of selected voxels. It can be seen in Figure 2.8 that the false discovery rates are equal for BF and BH when the average number of selected voxels is equal. Consequently, we cannot attribute the difference in stability to a difference in the number of false discoveries.

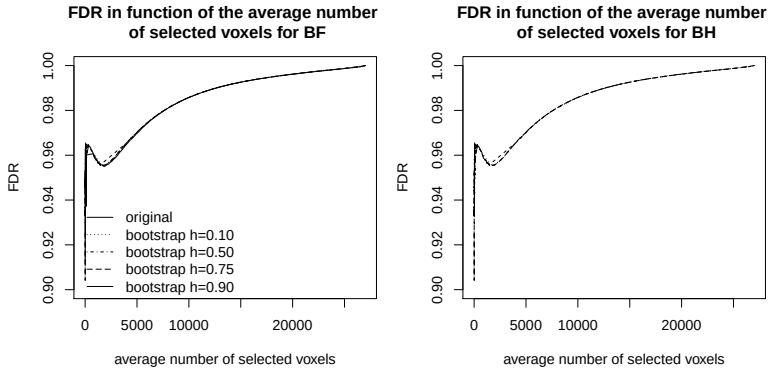


Figure 2.8: A plot of the average observed false discovery rate (FDR) against the average number of selected voxels. Observed FDR is defined as the proportion of the number of false discoveries amongst all selected voxels.

2.E Additional simulation study

We have done a small simulation study to further investigate the observed differences in stability between Benjamini-Hochberg and Bonferroni. We replicate the simulation study of Qiu et al. (2006), where outcome values for m genes or features are compared between two groups. Each group has a sample size of $n = 15$. To vary the number of tested hypotheses, we perform simulations for different values of m ($m = 100, 500, 1000, 10000$). The number of true alternative hypotheses is fixed to 10%.

The observed outcome values for the features are generated from the standard normal distribution. For the features that stem from the alternative, the values within one group stem from a normal distribution with mean 2 and variance 1.

For data sets generated in a setting where correlation is present, an exchangeable pairwise correlation structure with correlation coefficient ρ is imposed on the outcome values of the features. Following Qiu et al. (2006), we generate a vector of length $2n = 30$ for each feature i ($i = 1, \ldots, m$) with components x_{ij} ($j = 1, \ldots, 2n$) drawn from the standard normal distribution. Then we add correlation between the values of the same sample by creating the outcome value y_{ij} as follows:

$$y_{ij} = \sqrt{\rho} a_j + \sqrt{1 - \rho} x_{ij},$$

where a_j is again drawn from the standard normal distribution. At the end, a value of 2 is added to the outcome values in the first group of the features that stem from the alternative. In that way, for every $i_1 \neq i_2$, $\text{cor}(y_{i_1 j}, y_{i_2 j}) = \rho$. In our simulations, ρ is chosen either equal to 0 or 0.5. Qiu et al. (2006) argue that they do not attempt to model actual biological correlation (which can be more complicated) as their purpose is mainly to investigate the effect of correlation on results of testing procedures. For each feature, a two-sample t-test is performed which is corrected for multiple testing, either using Bonferroni or Benjamini-Hochberg.

Results of the simulation study are shown in Figure 2.9 and 2.10.

It can be seen on Figure 2.9 that stability increases with the number of hypotheses when the features are uncorrelated ($\rho = 0$). We find that in case the proportion of selected hypotheses is small (which corresponds to respectively a small imposed FDR level and Type I error rate),

both procedures coincide in terms of stability when $m = 10\,000$. For a larger proportion of selected hypotheses, BF appears to remain more stable. The difference becomes smaller when m increases and is expected to disappear when m is increased even further. This effect is also shown in Figure 2.10 where the specific uncorrected p-value cutoff indeed converges to a fixed number with increasing tests, but converges quicker when the cutoff is smaller. Results drastically change as soon as correlation is introduced ($\rho = 0.5$). We find that increasing the number of hypotheses does not increase stability and that overall, Bonferroni is more stable than Benjamini-Hochberg (Figure 2.9).On Figure 2.10, it can be seen that the variance of the p-value cutoff remains more or less the same when the number of tested hypotheses increases.

We acknowledge that in this simulation setting, we did not consider the case of weak dependency in which case the performance of test procedures is asymptotically the same as if tests were independent. On the contrary, we simulated scenarios where there is a strong correlation between each pair of features. This enables to assess how procedures react in case of strong (and often complex) correlations among tests.

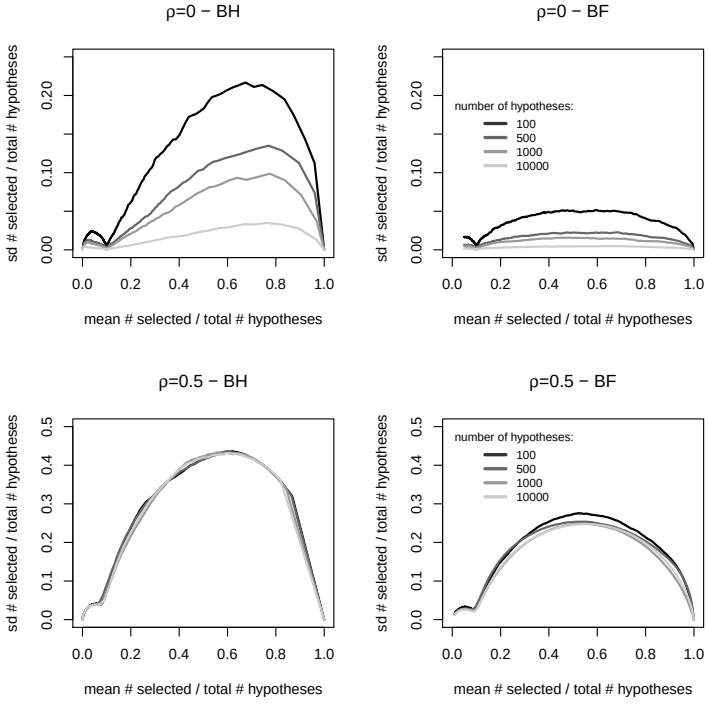


Figure 2.9: Stability of Benjamini-Hochberg and Bonferroni. Instability is assessed by the standard deviation on the number of selected voxels relative to the total number of hypotheses.

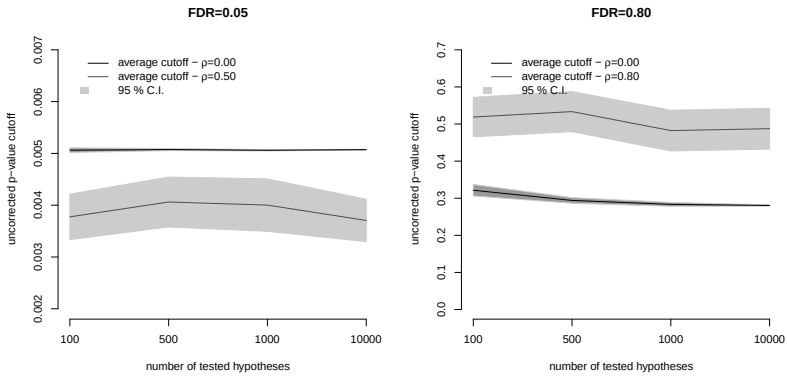


Figure 2.10: The figure shows the uncorrected p-value cutoff when using BH without ($\rho = 0$) and with ($\rho = 0.5$) correlation among the tests. Results are shown for FDR control at 5% (left) and at 50% (right). The figure shows that indeed the uncorrected p-value threshold converges to a fixed point for an increasing number of tests. This can be seen as the variance decreases when the number of tests increase. This variance however depends on the FDR cutoff. When thresholding at FDR = 0.80, the variance is much higher. Although this cutoff is unrealistic, it shows how the procedure performs at the extremes. However, the convergence to a fixed p-value cutoff completely disappears when correlation is present among the test statistics.

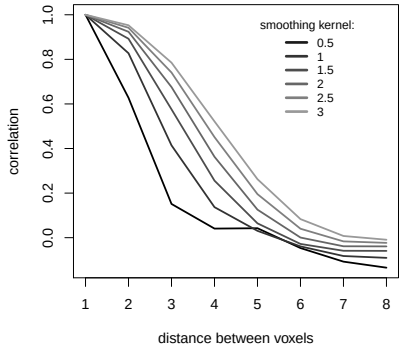


Figure 2.11: The figure shows the spatial correlation between the voxels.

2.F Spatial correlation

Spatial correlation among test statistics is key in fMRI data. By smoothing the data, we impose a spatial correlation on the data. To visualise the relation between smoothness and spatial correlation, we have computed the correlation between a certain voxel, and voxels at different distances. We have averaged this over time points. The result can be seen in Figure 2.11 in which we show the spatial correlation in terms of classical point-by-point correlation. On this figure, we demonstrate that a spatial correlation with a fixed size is induced by smoothing the data.

2.G Results of the simulations with varying degrees of smoothness

An often-used preprocessing step in the analysis of fMRI data are smoothing the data. As this has a large positive effect on the sensitivity of the analysis, it is interesting to see how our results can be extended under different degrees of smoothness. In the original paper, we used a smoothing kernel with full-width at half maximum (FWHM) of 2 times the voxel size. We have also used 5 other possible values of the FWHM: 0.5, 1, 1.5, 2.5 and 3. In Figures 2.12 to 2.16, the results can be found under different degrees of smoothness.

It can be seen in the figures below that indeed the area under the curve of the ROC increases for higher values of smoothness, thereby indicating that the power of the analysis increases for equal levels of specificity. The main conclusions of our simulation study however do not differ for different ranges of smoothness, i.e. the different multiple testing procedures produce equal ROCs. BF elicits a much higher stability then BH. Using the bootstrap procedure by Gordon et al. (2009) can improve the stability of BH up to the level of BF, whereas the stability of BF remains the same.

The observant reader will see that the maximum standard deviation of the number of selected voxels is not monotonic in the smoothing parameter FWHM as is shown in Figure 2.17. In this Figure, the smoothness parameter σ is related to the FWHM as follows: $FWHM = \sqrt{\sigma^2 8log2}$. We have investigated this further by looking at the standard deviation in function of the smoothing parameter when fixing the average number of selected voxels in Figure 2.18. It can be seen that for BF, the standard deviation increases for increasing smoothing values. Another pattern is observed for BH procedures. There, we see that for rather stringent FPR control (i.e. selecting 5000 voxels), instability increases with the smoothing parameter. On the other hand, at the more liberal thresholds, the instability for FDR control decreases with the smoothing parameter. Due to this pattern, we expect that peak of the standard deviation corresponds to a different average number of selected voxels, when varying the smoothing parameter, and therefore it is not monotonic.

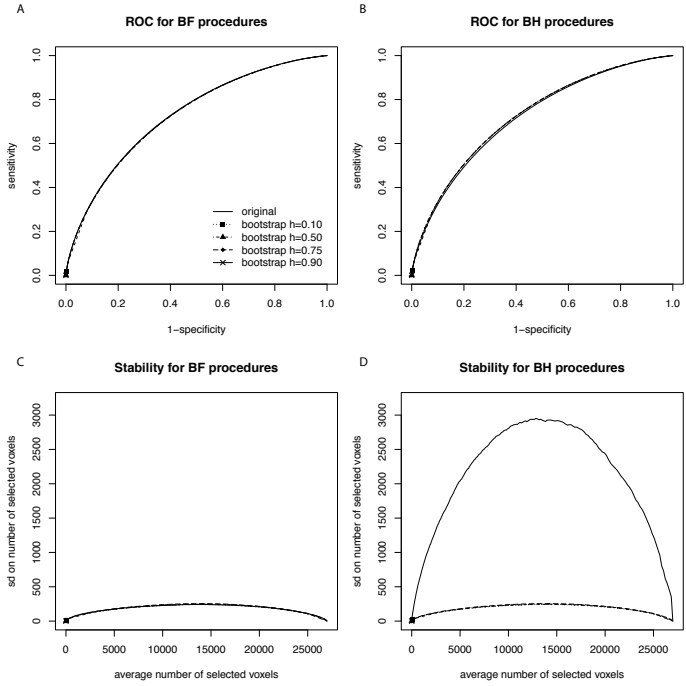


Figure 2.12: Roc curves (upper panel) and stability plot when the data was smoothed with a Gaussian kernel with a FWHM of 0.5.

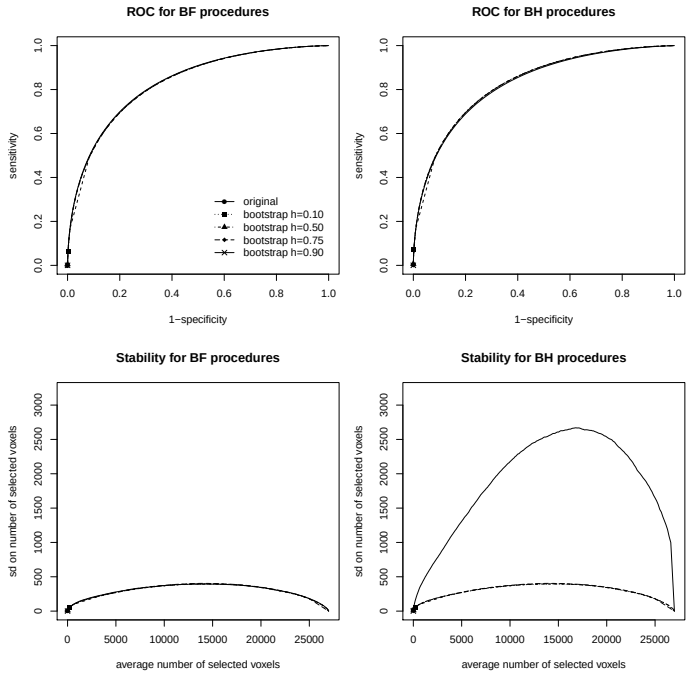


Figure 2.13: Roc curves (upper panel) and stability plot when the data was smoothed with a Gaussian kernel with a FWHM of 1.

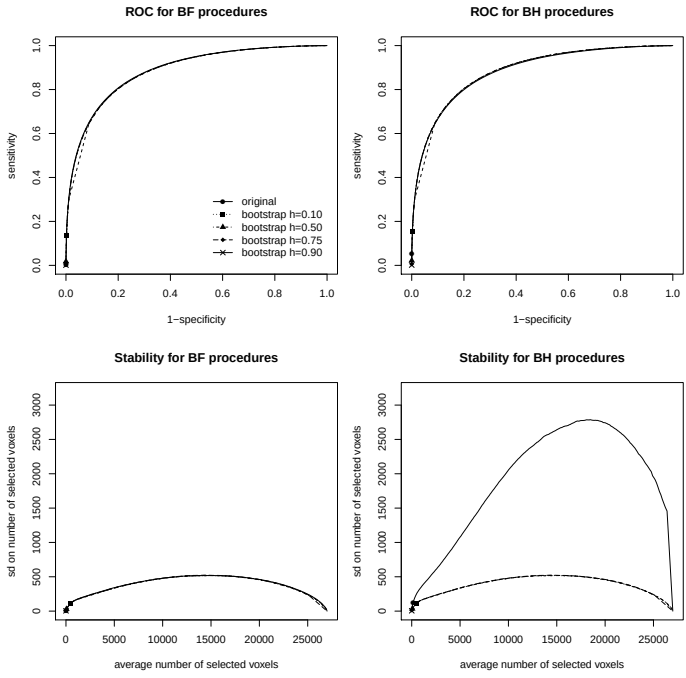


Figure 2.14: Roc curves (upper panel) and stability plot when the data was smoothed with a Gaussian kernel with a FWHM of 1.5.

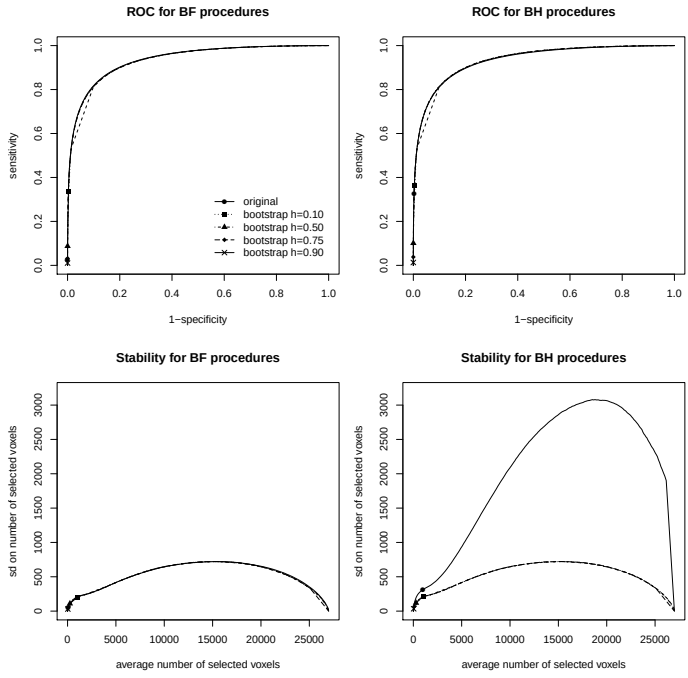


Figure 2.15: Roc curves (upper panel) and stability plot when the data was smoothed with a Gaussian kernel with a FWHM of 2.5.

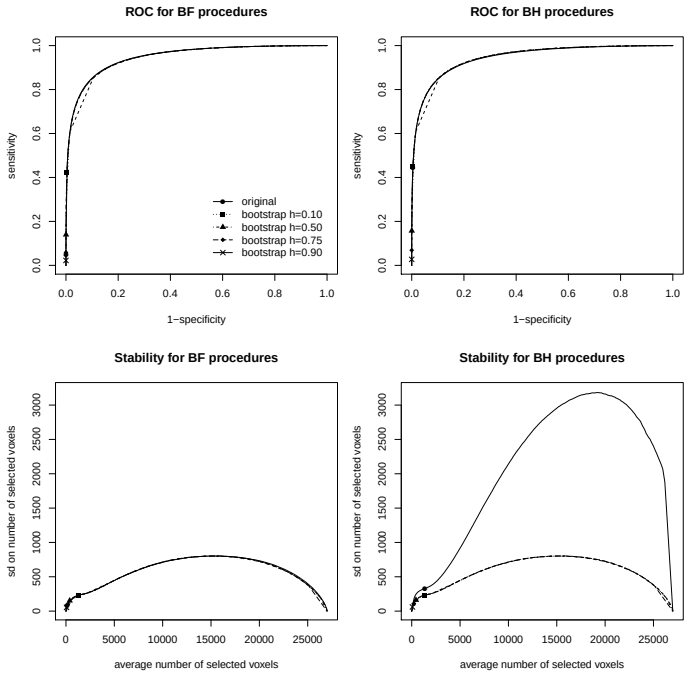


Figure 2.16: Roc curves (upper panel) and stability plot when the data was smoothed with a Gaussian kernel with a FWHM of 3.

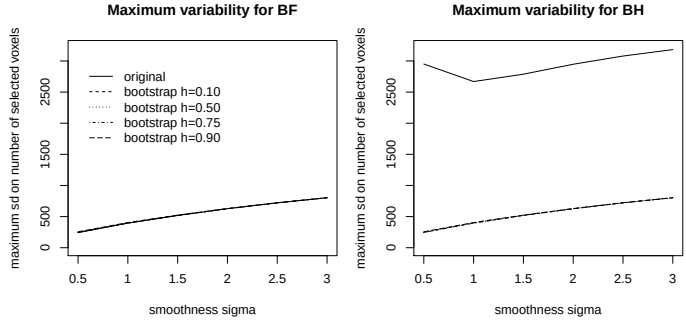


Figure 2.17: Plot which shows the maximum variability of a procedure as a function of the smoothness parameter sigma. The parameter sigma σ is related to the full width half maximum up to a constant, $FWHM = \sqrt{\sigma^2 8log2}$.

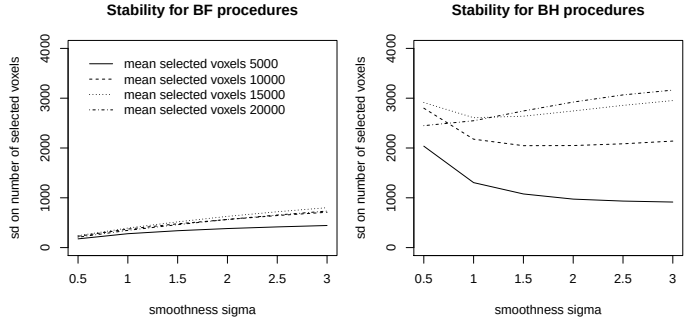


Figure 2.18: The standard deviation of the number of selected voxels as a function of the smoothness parameter sigma, for fixed average numbers of selected voxels. The parameter sigma σ is related to the full width half maximum up to a constant, $FWHM = \sqrt{\sigma^2 8log2}$.

2.H Real data application with FWER control at 0.05

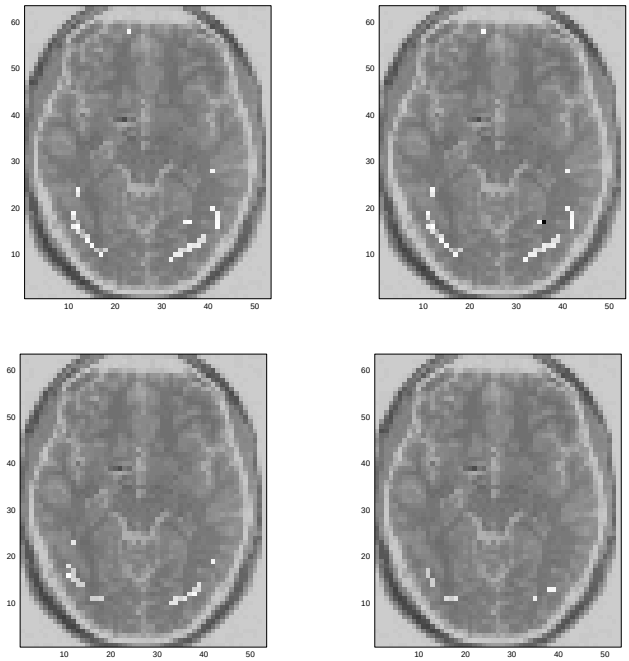


Figure 2.19: Results of the analyses on the Henson data set. On the left are the analyses with BF correction, (FWER=0.05). On the right are the analyses using the BH procedure (FDR=3×10^{-7}). Above are the results with classical analyses with the voxels exceeding the significance level highlighted. The plots below display the percentage of selection over bootstrap samples with the "stable voxels" (with frequency of occurence $> 80\%$) highlighted. The thresholds are chosen to make sure that in the images above, the same number of voxels are significant.

3 Post-hoc power estimation for topological inference in fMRI[1]

Abstract When analyzing functional MRI data, several thresholding procedures are available to account for the huge number of volume units or features that are tested simultaneously. The main focus of these methods is to prevent an inflation of false positives. However, this comes with a serious decrease in power and leads to a problematic imbalance between type I and type II errors. In this paper, we show how estimating the number of activated peaks or clusters enables to estimate post-hoc how powerful the selection procedure performs. This procedure can be used in real studies as a diagnostics tool, and raises awareness on how much activation is potentially missed. The method is evaluated and illustrated using simulations and a real data example. Our real data example illustrates the lack of power in current fMRI research.

3.1 Introduction

Traditional multiple testing procedures for neuroimaging data, such as the Bonferroni procedure (Bonferroni, 1936), control the family-wise error rate (FWER) which is the probability of at least one false positive. When dealing with a huge number of tests, as is the case in fMRI data analysis, controlling the FWER typically results in low power or sensitivity to detect important effects. As a result interest has grown in procedures that control the number of false positives relative to the total number of positive statistical tests. This error rate is called the False Discovery Rate (FDR) (Benjamini & Hochberg, 1995). Both FWER and FDR control have been succesfully applied to voxelwise inference for fMRI data (Nichols & Hayasaka, 2003; Genovese et al., 2002).

As smooth fMRI data represents a continuous signal, some procedures

[1]This chapter is based on:
Durnez, J., Moerkerke, B. and Nichols, T.E. (2014). Post-hoc power estimation for topological inference in fMRI. *NeuroImage*, 84, 45–64.

focus on inference for topological features as clusters or peaks, instead of making inference on individual voxels (Worsley, Evans, Marrett, & Neelin, 1992; Hayasaka, Phan, Liberzon, Worsley, & Nichols, 2004; Perone Pacifico, Genovese, Verdinelli, & Wasserman, 2004; Worsley, Taylor, Tomaiuolo, & Lerch, 2004). Random field theory (Adler, 1981) or permutation procedures (Nichols & Holmes, 2002) enable cluster and peak level inference. For peak-level inference, the statistic values at the peaks (or local maxima) are then transformed to peak p-values, which can be interpreted as *"The probability of finding a peak at least as high as this peak when the null hypothesis is true"*. Cluster-level inference constitute of cluster p-values based on the extent of the cluster. The p-value can consequently be interpreted as *"The probability of finding a cluster at least as large as this cluster when the null hypothesis is true"*. Techniques have been developed to use these peak and cluster p-values to control both the FWER (Friston et al., 2007) and the FDR (Chumbley et al., 2010; Chumbley & Friston, 2009) when testing topological features.

Whichever method one uses, the emphasis is typically on controlling an excess of type I errors. This comes with a price: when making the threshold for statistical significance more stringent, many activated brain regions will inevitably not reach the level of significance, leading to an increase in type II errors or a (possible a dramatic) drop in power. As Lieberman & Cunningham (2009) point out, the considerable focus on measures taken to avoid type I errors is in stark contrast to the scant attention paid to power. Lieberman & Cunningham (2009) show with a small omnibus power analysis that, for a large effect, the probability of detection drops to less than 15% when a voxelwise significance level of $\alpha = 0.001$ is used, a very liberal significance level when measured in terms of the FWER.

While stringent control of type I errors minimizes false positive results in literature, the accompanying limits on power comes with negative consequences. One example is the biasing of results of meta-analyses of fMRI studies: type II error rates are never reported and will consequently not be considered in aggregate analyses (Lieberman & Cunningham, 2009). Another case in which little power may lead to dramatic results is presurgical fMRI, where patients undergoing brain surgery are scanned in order to find vital brain tissue. In this setting, significant activations are

used to mark "eloquent" brain tissue, neccessary for vital function such as speech or walking. A type II error can therefore lead to removal of vital brain tissue. Several new methods have been developed with the specific goal of avoiding type II errors around the tumor (Johnson, Liu, Bartsch, & Nichols, 2012; Gorgolewski, Storkey, Bastin, & Pernet, 2012).

To meet an acceptable level of power in fMRI studies, several *a priori* power calculation methods are available (Mumford & Nichols, 2008; Desmond & Glover, 2002; Zarahn & Slifstein, 2001). However, while the type I error rate of a study is mostly clearly defined and easily interpretable (e.g. when controlling FDR at 5%, one can estimate that 1 in 20 significant peaks will be false positives and therefore non-active), the existing power estimators often involve complex equations and many unknown parameters, depending on how *activation* is defined. In contrast, the aim of this study is to present a method to estimate the power of an fMRI study *post-hoc*. One family of existing post-hoc power estimators makes use of the proportion of non-active features (Zehetmayer & Posch, 2010). Zehetmayer & Posch (2010) have extensively studied the estimation of the proportion of non-activated features (π_0) in the field of statistical genetics where one aims to detect genetic markers that are associated with a phenotype. Within brain imaging, estimates for π_0 have been introduced in voxel-based morphometry for adaptive FDR control (Chen, Wang, Eberly, Caffo, & Schwartz, 2009). Chen et al. (2009) argue that within fMRI, it is impossible and needless to estimate the proportion of non-active voxels, since this proportion is typically very close to 1. Post hoc approaches to power that use the proportion of non-active voxels can therefore not be used in the context of voxelwise inference. However, when considering topological features, the number of statistical tests decreases and the proportion of active features increases. Consequently, existing techniques can be used to estimate the number of inactive peaks or clusters in respectively peak-level or cluster-level inference for fMRI which enables post-hoc power estimation.

In section 2, we show how the number of inactive features (π_0) for both peaks and clusters can be estimated. As in Zehetmayer & Posch (2010) in the context of genetics, we use these estimates for π_0 to estimate the true positive rate (TPR, classical definition of power) and the false non-discovery rate (FNR) (Genovese et al., 2002) in section 3. This leads to the

construction of a study-specific free-response receiver operator characteristics (FROC) curve, in which the true positive rate is plotted against the false positive rate (Smith & Nichols, 2009) in section 4. Finally, section 5 illustrates the proposed methods using a real data set on the neuropsychological underpinning of face recognition (Henson et al., 2002).

3.2 Prevalence of activation

3.2.1 Estimation procedures

Topological inference The two most widely used topological features in neuroimaging are clusters and peaks. These features are defined on a test statistic image, where a linear model is fit and a test statistic t computed at each voxel. A well-considered threshold u is applied to the test statistic image to create an excursion set of supra-threshold voxels. This primary threshold eliminates very small peak values and defines clusters.

Peaks are local maxima in the excursion set and are defined as having the maximal t-value within their neighbourhood; we use a 26-point neighborhood defined by voxels sharing a common face, edge or corner with a given voxel. The test statistic for the peak is the t-value of the peak itself. When performing inference on a peak, the goal is to find a second threshold t_α for which all suprathreshold peaks are considered to be active. This is shown in Figure 3.1. Uncorrected p-values for peak values can be found that give the probability that a peak equal or higher can be found under the null hypothesis of no activation. These peak p-values can be submitted to any multiple testing strategy, resulting in a controlled type I or false positive rate.

Clusters are defined as a neighbouring set of voxels in the excursion set; we again use a 26-point neighborhood to define clusters. The size of each cluster is quantified as its extent, or the number of voxels within the cluster. Cluster-level inference consists of finding an extent threshold, where all clusters encompassing more than a certain number of voxels are called active. Analogously to peak-level inference, uncorrected p-values for cluster size are computed which give the probability that a cluster as large or larger can be found when there is no activation. As is the case for peaks, corrections for multiple testing can be used.

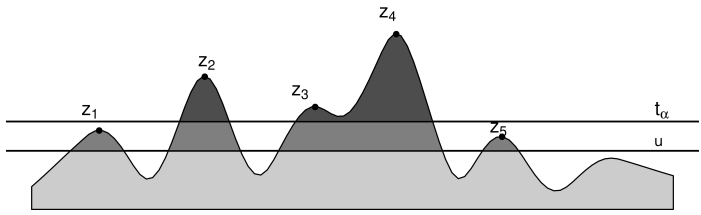


Figure 3.1: Peaks and clusters are defined by first applying a first threshold u, which reveals several clusters and eliminates small peaks. Thereafter, based on the t-values of the peaks $\{z_1, z_2, ..., z_m\}$ or the size of the clusters $\{s_1, s_2, ..., s_m\}$, a second threshold t_α is chosen, to control the false positive rate at a predefined level.

An important characteristic of peaks and clusters is that their appearance is a stochastic process, defined by the spatial distribution of the t-values. Consequently, peaks and clusters are randomly located and when an experiment is replicated, clusters and peaks may comprise different voxels. Therefore interpretation of results cannot be directly translated between different replications of the same experiment (Nichols, 2012).

A number of procedures have been presented that provide p-values for clusters or peaks (Heller, Stanley, Yekutieli, Rubin, & Benjamini, 2006; Hayasaka et al., 2004; Zhang, Nichols, & Johnson, 2009; Chumbley & Friston, 2009; Friston et al., 2007; Forman et al., 1995; Hayasaka et al., 2004; Perone Pacifico et al., 2004; Nichols & Holmes, 2002). In what follows, we consider the two most popular techniques: random field theory (Chumbley et al., 2010; Friston et al., 2007) and non-parametric permutation tests (Nichols & Holmes, 2002).

Random Field Theory (RFT) To precisely state the random field theory results we need a small amount of notation for the continuous random process that approximates our observed statistic image. Let $\Omega \subset \mathbb{R}^D$ be the D-dimensional search region of interest (e.g., the brain), and $T(x) \in \mathbb{R}$ be a random variable, the test statistic at voxel $x \in \mathbb{R}^D$. Then to control the FWER over the brain we approximate the maximum distribution of

$T(x)$ as

$$P(\max_{x\in\Omega} T(x) \geq t|H_0) \approx \sum_{d=0}^{D} \text{Resels}_d \text{EC}_d(t) \tag{3.1}$$

with H_0 representing the hypothesis of no activation, $\max(T)$ the value of the maximum statistic searched over the brain, D the number of dimensions, Resels_d the number of d-dimensional resolution elements of the search region, and $\text{EC}_d(t)$ the d-dimensional Euler characteristic density of the test statistic. The Euler characteristic counts the number of clusters above t if the region has no holes, which is likely to be the case if t is large (Worsley, 2007).

While $P(\max_{x\in\Omega} T(x) \geq t|H_0)$ is usually used to approximate the FWER-corrected p-value for a voxel with statistic value t, RFT can also be used to obtain uncorrected p-values for a local maximum in the image, as well as clusters. The result for peak uncorrected p-values, surprisingly, does not depend on the search region *or* smoothness and is poorly described, so we detail it here.

Let $T = u + H$ be the height of a local maximum above a threshold, that is, conditional on $T > u$. The conditional probability of H can be seen to be a ratio of two extrema

$$\begin{aligned} P(H \geq m/u|T > u) &= P(M \geq u + m/u|T > u) \\ &= \frac{P(M \geq u + m/u)}{P(T > u)} \end{aligned}$$

for some fixed $m > 0$. The numerator and denominator of this expression can each be approximated with the $d = 3$ term of Equation 3.1, and then Resels_3 terms cancel and one is left with an expression that, as $u \to \infty$, goes to e^{-m}. To find an uncorrected p-value for observed local maximum t, we solve $t = u + m/u$ for m, substitute and get

$$P(T \geq t|T \geq u, H_0) \approx \exp(-u(t-u)).$$

For clusters, the spatial extent can be expressed as $S \approx cH^{D/2}$ (Worsley, 2007) with

$$c = \frac{FWHM^D u^{D/2} P(T \geq u|H_0)}{EC_D(u)\Gamma(D/2+1)} \tag{3.2}$$

The p-value for a cluster with an observed spatial extent s is approximated as follows:

$$P(S \geq s|H_0) \approx \exp(-u(s/c)^{2/D}) \tag{3.3}$$

The non-parametric permutation framework RFT results depend on assumptions of smooth Gaussian multivariate noise, and even when these assumptions are met RFT results may not be accurate (Nichols & Hayasaka, 2003; Hayasaka, 2003). Hence we also consider the use of permutation methods to provide emperical null distributions upon which to base p-values (Dwass, 1957). A permutation test is based on an assumption of exchangeable data under the null hypothesis of no effect. For example, in a 2-sample t-test, under the null hypothesis all the data are equivalent and can be randomly assigned to either group. For fMRI, drift and temporal autocorrelation violate the assumption of exchangeability; thus we whiten the residuals before permutation (Bullmore et al., 1996) and, as a further precaution, only shuffle condition labels within adjacent blocks.

Data is repeatedly analyzed to create a null distribution and, to control FWER, a maximal distribution is created and used to compute FWER-corrected p-values (Nichols & Hayasaka, 2003).

Choice of p-values In what follows, we use p-values on topological features to estimate power. We used simulated and real null data to verify the accuracy of RFT and permutation p-values. We used temporally and spatially correlated synthetic data as well as real resting-state fMRI data, for each emperically measuring the probability of finding a feature greater than or equal to the observed feature under the null of no activation. See Appendix A for details.

To summarize the results shown in Appendix A, for peaks both RFT and non-parametric p-values are very close to their observed p-values under the null hypothesis. In this paper, we therefore only show results for RFT p-values because of computational simplicity, though of course our methods are equally applicable to permutation p-values. For clusters, RFT p-values show a deviation from the observed p-values, while non-parametric p-values provide good estimates. Therefore, we only consider

	Declared active	Declared inactive	Total
Non-active	F	$m_0 - F$	m_0
Active	T	$m_1 - T$	m_1
Total	S	$m - S$	m

Table 3.1: Expected numbers in a cross-classification of selected features with active features. m_1 is $m - m_0$. S represents the number of features that are declared active. F and T respectively represent the false and true positives.

non-parametric p-values for clusters. This does not imply that the proposed procedures are not suitable for RFT cluster p-values.

In the remainder of this paper, we generically denote peaks and clusters as "topological features", or just "features".

Estimation of π_0 If a set of m topological features is subjected to statistical testing, a certain number m_0 of these features is expected to be non-active. In this study, we presume that a topological feature can be binary classified as either active or non-active. As in Schwartzman, Gavrilov, & Adler (2011), we define an active peak as a peak situated in active brain tissue, while a non-active peak lies in a non-active brain region. We consider a cluster to be active if it contains at least one voxel that lies in an active area. Other definitions of active clusters are possible such as in Heller et al. (2006) and Chumbley & Friston (2009).

Standard practice is to reject the null hypothesis when the p-value is lower than a pre-specified significance level α (with α the per comparison error rate). Table 3.1 gives a two-way classification of true activation status and declared status according to a statistical test. For example, from the m features, S features can be declared as active while $m - S$ features do not reach the level of significance to be declared as active. Assuming p-values are uniformly distributed under the null hypothesis H_0, we expect to select $\alpha \times m_0$ non-active features with a p-value threshold equal to α. Table 3.2 shows the expected counts, where only the number of true hypotheses m_0 is unknown, and is the basis of power calculations made further in the text (Delongchamp, Bowyer, Chen, & Kodell, 2004).

	Declared active	Declared inactive	Total
Non-active	$\alpha m_0 = F$	$m_0 - \alpha m_0$	m_0
Active	$S - \alpha m_0 = T$	$m_1 - (S - \alpha m_0)$	m_1
Total	S	$m - S$	m

Table 3.2: Expected numbers in a classification of m features (Delongchamp et al., 2004), where only m_0 (and hence $m_1 = m - m_0$) is unknown.

Estimating m_0 enables estimating the total number of true positives and false positives (Zehetmayer & Posch, 2010). Zehetmayer & Posch (2010) also outline assumptions with respect to dependency among tests. As we consider topological as opposed to voxelwise inference, we are not dealing with highly correlated test statistics across the brain.

The number of inactive features m_0 is equal to $\pi_0 \times m$, where π_0 is the proportion of inactive features. In the literature, there are several different methods available that estimate the proportion of inactive features in a large-scale testing setting (Benjamini & Hochberg, 2000; Storey, 2002; Storey & Tibshirani, 2003; Pounds & Morris, 2003; Pounds & Cheng, 2004). These methods all rely on the fact that p-values are uniformly distributed under the null hypothesis H_0 but differ in the way they estimate the proportion of null p-values. An overview within statistical genetics has been given by Broberg (2005). We reconsider these results however, as the topological features of statistical parametric maps are random in number, and number fewer than the genes or proteins used in these other works. We apply and compare the following methods to estimate π_0 using simulations, in a similar way as Broberg (2005) and Zehetmayer & Posch (2010):

1. BH: The adaptive lowest slope estimator (Benjamini & Hochberg, 2000)

2. S: The p-value smoother (Storey, 2002)

3. ST: The p-value smoother with bootstrap resampling (Storey & Tibshirani, 2003)

4. PM: The beta-uniform model (Pounds & Morris, 2003)

5. PC: The spacing loess histogram (Pounds & Cheng, 2004)

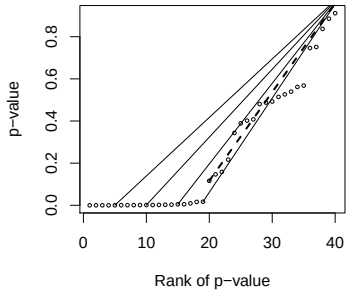
An overview and an example of these estimators can be found in Figure 3.2.

3.2.2 Simulations

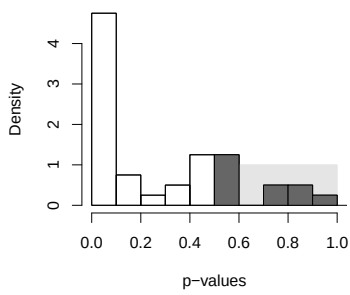
We simulate a blocked design with 10 blocks of the active condition, altered with 10 blocks of non-activity. Each block consists of 20 time points, thus leading to 400 time points in total. The image is comprised of $40 \times 40 \times 40$ voxels, which corresponds to a brain volume in moderate resolution data. The data are generated using neuRosim (Welvaert, Durnez, Moerkerke, Verdoolaege, & Rosseel, 2011) with a mixture of Gaussian white noise and temporal noise. NeuRosim is a statistical toolbox in R, consisting of different data generation functions for neuroimaging data. The temporal dependence is fixed to $\rho = 0.20$. First, for 5 regions of $7 \times 7 \times 7$ voxels, a signal is added in half of the time points with signal-to-noise ratio (SNR) of 0.015 before smoothing. The images are convolved with a 3D isotropic Gaussian kernel with $\sigma = 2.5$. A general linear model (GLM) is fit to the data resulting in a T-statistic image with 398 degrees of freedom. The excursion threshold u is set such that $P(T_{398} \geq u) = 0.01$. This threshold needs to be high enough for random field theory assumptions to hold, but low enough to preserve as much information as possible. Apart from some vague advise, no clear vision has been given to the exact choice of the threshold (Smith & Nichols, 2009). The permutation based approach works for any excursion threshold. After the excursion set is defined, the peaks and clusters on the T-image are computed with 26-point neighborhood and their uncorrected p-values are calculated.

Next, the different parameters (number of activated regions, SNR, smoothing parameter σ, u) are varied in the study over a range of values to investigate the influence they exert on the proposed estimation procedure.

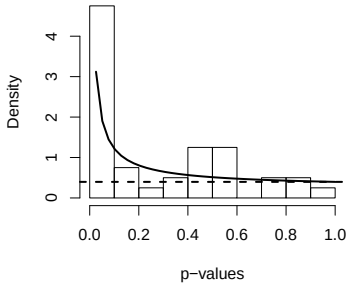
The statistic values of the topological features (i.e. peak height and cluster size) are converted to uncorrected p-values, the basis for the estimation of π_0. This simulation routine is repeated 500 times.



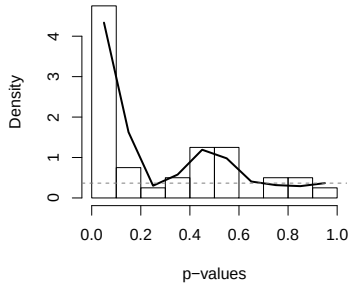
Benjamini & Hochberg (2000) - BH: After sorting the p-values, the slope S_i is calculated between the ith p-value and the point $(m+1, 1)$ for ascending i (solid lines). The first i for which $S_i < S_{i-1}$ is considered the first p-value from the alternative hypothesis (dashed line). Subsequently $\hat{m}_0$ is estimated as $\min(m, (1/S - i + 1))$.



Storey & Tibshirani (2003) - ST; Storey (2002) - S : For a certain λ (here: 0.5), π_0 is estimated as the density of p-values greater than λ (dark grey) divided by the expected density under H_0, $1 - \lambda$ (light grey). Storey & Tibshirani (2003) and Storey (2002) differ in the way they optimize the choice of λ.



Pounds & Morris (2003) - PM: The probability density function is assumed to be a mixture of a uniform distribution (dashed line) and a beta-distribution (solid line), with weight π_0 to the uniform distribution. Through maximal likelihood estimation, the optimal weights and parameters for the beta-distribution are estimated.



Pounds & Cheng (2004) - PC: The probability density function $\hat{f}(p_i)$ is estimated by applying a loess smoother through the histogram of p-values (solid line). π_0 is then estimated as $\hat{f}(\max(p_i))$.

Figure 3.2: Overview of the different methods to estimate π_0 under consideration. All methods are applied to the same example, with $H_0 = x \sim N(0,1)$ and $H_a = x \sim N(3,1)$ with $df = 30$. The true π_0 is equal to 0.5. The estimates from the different estimators are: $\hat{\pi}_0(BH) = 0.60$, $\hat{\pi}_0(S) = 0.29$, $\hat{\pi}_0(ST) = 0.31$, $\hat{\pi}_0(PM) = 0.40$, $\hat{\pi}_0(PC) = 0.21$.

	Peaks		Clusters	
	Mean $\hat{\pi}_0$ (sd)	RMSE (sd)	Mean $\hat{\pi}_0$ (sd)	RMSE (sd)
True π_0	0.77 (0.04)		0.87 (0.03)	
Benjamini-Hochberg (BH)	0.87 (0.06)	0.12 (0.11)	0.95 (0.04)	0.08 (0.08)
Storey (S)	0.84 (0.19)	0.2 (0.20)	0.94 (0.14)	0.15 (0.22)
Storey-Tibshirani (ST)	0.72 (0.17)	0.17 (0.25)	0.84 (0.16)	0.15 (0.25)
Pounds-Morris (PM)	0.74 (0.08)	0.08 (0.10)	0.83 (0.06)	0.07 (0.08)
Pounds-Cheng (PC)	0.59 (0.12)	0.21 (0.24)	0.62 (0.11)	0.28 (0.25)

Table 3.3: Results of π_0 estimators when applied to neuroimaging data based on n =500 simulations with the average π_0 (standard deviation) and the RMSE (standard deviation). The RMSE is defined as $\sqrt{1/n \sum_{i=1}^{n} (\widehat{\pi_{0i}} - \pi_{0i})^2}$. Results are obtained with the basic scenario with $P(T_{398} \geq u) = 0.05$, SNR $= 0.015$, $\sigma = 2.5$ and with 5 blocks of activation.

3.2.3 Comparison of different π_0 estimators

The different π_0 estimators are applied to simulated data sets. A comparison between the estimators for a specific combination of parameters can be found in Table 3.3. We compare the estimators with the true underlying values and compute the root of the mean squared error (RMSE). With π_{0i} ($\widehat{\pi}_{0i}$) the (estimated) proportion of non-active features in simulation step i and n the number of simulations, the RMSE is defined as $\sqrt{1/n \sum_{i=1}^{n} (\widehat{\pi_{0i}} - \pi_{0i})^2}$. It should be noted that due to the random character of peak and cluster appearance and location, the true π_0 varies in each simulation step and therefore, the estimates for π_0 are evaluated against the true π_0 in each simulation step separately. We see that PM produces the smallest bias and RMSE, with a small variance on the results. In Appendix B, it is shown that this result is consistent for varying SNR, smoothing parameter σ, u and for a varying number of activated regions. We decide that PM has the best overall performance, and use this estimator for the following steps. We stress, though, that our method can be used with any accurate π_0 estimation method.

3.3 Estimation of the number of true positives and false negatives

Having derived a list of significantly 'active' features, it is of interest to estimate power. As is the case for defining a measure of the number of false positives, there are several possible measures of the number of true positives when performing many tests simultaneously. We will enlist two measures of power that can be estimated based on Table 3.1.

3.3.1 True Positive Rate

As a counterpart of the false positive rate (FPR = $E(F)/m_0$), the classical way to denote the power of a test is the true positive rate (TPR = $E(T)/m_1$). This measure quantifies the expected proportion of the significant active peaks out of all active peaks. The TPR can be estimated based on the results of Table 3.2 as $(S - \alpha m_0)/m_1$. Typically, one considers a range of p-value thresholds. For example, the TPR corresponding to thresholds equal to the observed p-values p_i $(i = 1, \ldots, m)$ can be estimated. The TPR should be monotonously increasing with increasing threshold, and needs to be smaller than 1. This is imposed as follows in the estimation procedure:

$$\widehat{TPR}(p_i) = \max_{p^* < p_i} \left(\frac{S - p^* m_0}{m_1}, \frac{S - p_i m_0}{m_1} \right) \tag{3.4}$$

$$\widehat{TPR}(p_i) = \max \left(\frac{S - p_i m_0}{m_1}, 1 \right) \tag{3.5}$$

The results of the simulation study with respect to estimating the TPR are presented in Table 3.4 for p-value thresholds that are chosen to respectively control the FWER and the FDR at the 5% level.

Figure 3.3 shows that the average quality of the estimation of the TPR is independent of the amount of activation and the SNR for peaks. For clusters, bias shrinks as more activated clusters are present. The signal-to-noise ratio and the smoothness have a small impact on the estimation procedure. For peak-level thresholding, a low to moderate excursion threshold u leads to a minimal bias on the results, while for cluster-level

		$FWER = 0.05$		$FDR = 0.05$	
		Mean (sd)	RMSE (sd)	Mean(sd)	RMSE (sd)
TPR	peaks: observed TPR	0.22 (0.12)		0.41 (0.18)	
	peaks: estimated TPR	0.21 (0.11)	0.06 (0.08)	0.39 (0.16)	0.1 (0.13)
	clusters: observed TPR	0.64 (0.22)		0.82 (0.17)	
	clusters: estimated TPR	0.53 (0.18)	0.18 (0.21)	0.68 (0.13)	0.22 (0.24)
FNR	peaks: observed FNR	0.22 (0.06)		0.18 (0.06)	
	peaks: estimated FNR	0.25 (0.09)	0.08 (0.11)	0.21 (0.09)	0.09 (0.11)
	clusters: observed FNR	0.07 (0.04)		0.04 (0.04)	
	clusters: estimated FNR	0.1 (0.05)	0.06 (0.08)	0.07 (0.04)	0.06 (0.08)

Table 3.4: For the estimated TPR and FNR for peak and cluster level inference, the average (standard deviation) and RMSE (standard deviation) is shown, based on 500 simulations. The RMSE is defined as respectively $\sqrt{1/n \sum_{i=1}^{n} (\widehat{TPR_i} - TPR_i)^2}$ and $\sqrt{1/n \sum_{i=1}^{n} (\widehat{FNR_i} - FNR_i)^2}$ with n the number of simulations. Results are obtained with the basic scenario with $P(T_{398} \geq u) = 0.05$, SNR $= 0.015$, $\sigma = 2.5$ and with 5 blocks of activation.

inference, the best results are obtained for a moderate to high excursion threshold u. Roughly spoken, we find biases ranging from -0.20 to 0.20, which is comparable to the results of Zehetmayer & Posch (2010), where absolute biases on the TPR were found ranging from 0.11 to 0.70.

A low true positive rate indicates that some suprathreshold clusters that overlap with true activated areas remain undiscovered. However, it is possible that multiple suprathreshold clusters overlap with truly activated areas. For neuroscientists, missing some active clusters is acceptable if other clusters around the same activated area are discovered. In that case, the activation in the brain is discovered by at least one significant cluster. The average percentage of missed areas is shown in Figure 3.4. It can be seen that this percentage is rather high, but shrinks with higher levels of signal-to-noise ratio. It can also be seen that the percentage missed areas drops when the excursion threshold is liberal. This is because suprathreshold clusters in this case are very large, and one suprathreshold cluster often extends to multiple activated areas.

3.3.2 False Non-discovery Rate

When controlling the false discovery rate (FDR), it is of interest to define a measure for power equivalent to the FDR, named the false non-discovery rate (FNR) (Genovese et al., 2002) which is defined as $E[(m_1 - T)/(m - S)]$

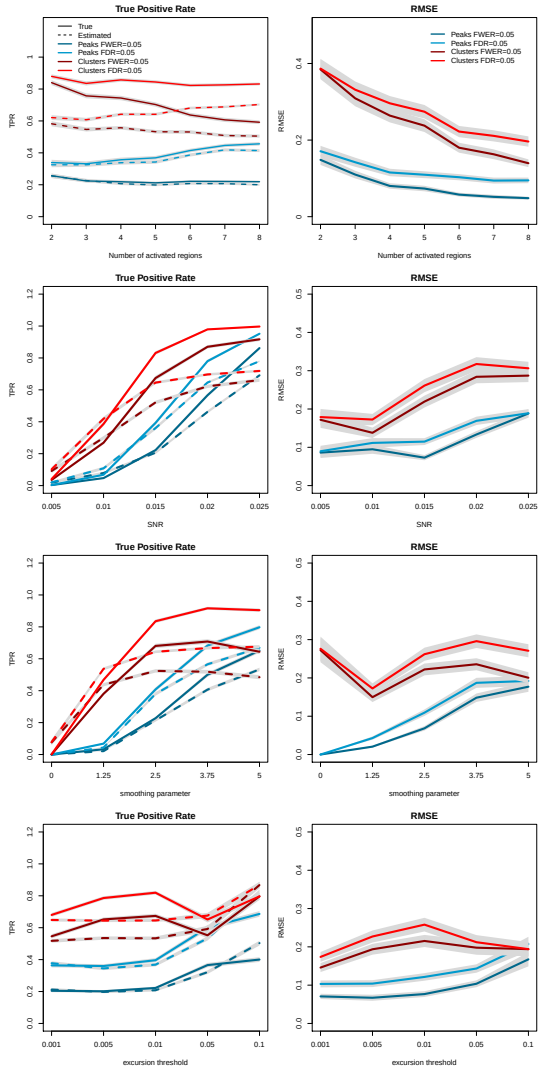


Figure 3.3: Bias and RMSE of the TPR estimator with 95% confidence interval. For 500 simulations, in each of the graphs, one of the parameters is varied. When the parameter is not varied, it takes the following values: number of activated regions = 5, SNR= 0.015, smoothing parameter $\sigma = 2.5$, excursion threshold $P(T_{398} \geq u) = 0.01$. On the left side, the mean estimated TPR (dashed line) is compared with the true mean TPR (full line). On the right side, the RMSE is plotted.

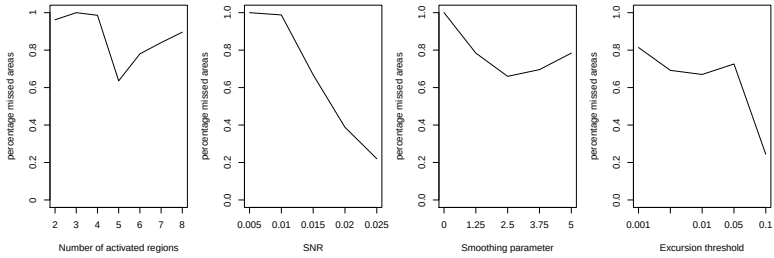


Figure 3.4: The average percentage of activated areas that are not discovered for cluster-level inference with $FWER = 0.05$ over 500 simulations.

if $S \neq m$ and 0 otherwise. It represents the proportion of false non-discoveries in relation to the total number of non-discoveries. This measure can again be estimated using Table 3.2.

Equivalent with the restrictions on the TPR, the FNR should be monotonously decreasing with increasing FPR, and needs to be smaller than 1. This is imposed as follows in the estimation procedure:

$$\widehat{FNR}(p_i) = \min_{p^* < p_i} \left(\frac{m_1 - (S - p^* m_0)}{m - S}, \frac{m_1 - (S - p_i m_0)}{m - S} \right) \quad (3.6)$$

$$\widehat{FNR}(p_i) = \max \left(\frac{m_1 - (S - p^* m_0)}{m - S}, 1 \right) \quad (3.7)$$

The results of the simulation study are presented in Table 3.4 and Figure 3.5. As opposed to the results of the TPR, the FNR is consistently over-estimated. The estimates are robust against variations in the different parameters.

3.3.3 Effect of the number of time points

Our simulated design represents an fMRI experiment of 400 time points. The duration of the experiment influences the power obtained in the analysis. This is however reflected in the signal-to-noise ratio: a longer experiment increases the signal-to-noise ratio. In Appendix 3.C we have separately investigated the effect of the duration on the TPR and the FNR

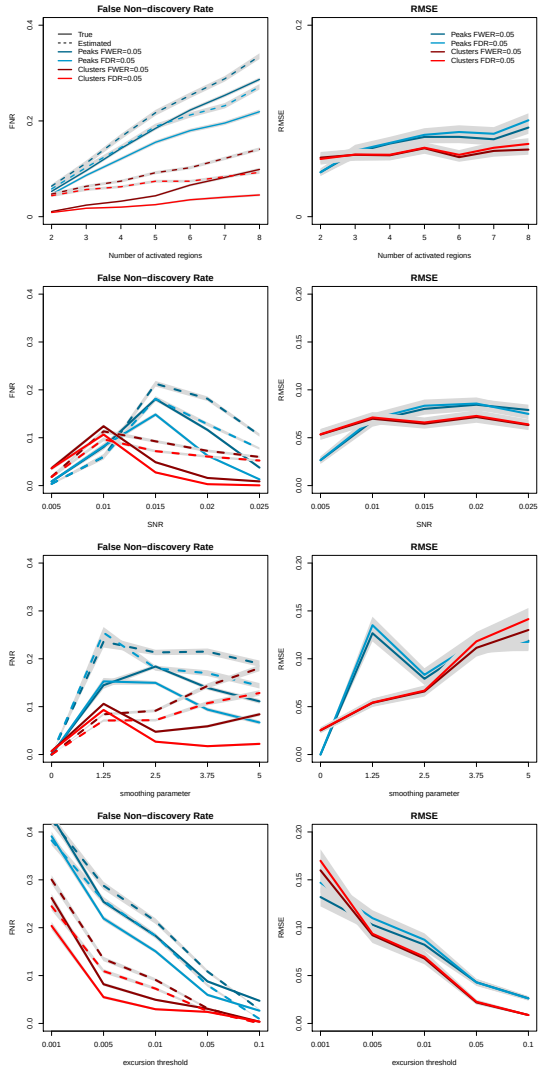


Figure 3.5: Bias and RMSE of the FNR estimator with 95% confidence interval. For 500 simulations, in each of the graphs, one of the parameters is varied. When the parameter is not varied, it takes the following values: number of activated regions = 5, SNR= 0.015, smoothing parameter $\sigma = 2.5$, excursion threshold $P(T_{398} \geq u) = 0.01$. On the left side, the mean estimated FNR (dashed line) is compared with the true mean FNR (full line). On the right side, the RMSE is plotted.

(and its estimates). Results confirm that the effect of the number of time points can be studied through the SNR.

3.4 Estimating the Free-Receiver Operator Curve (FROC)

3.4.1 Procedure

In a Free-Receiver Operator Curve, the TPR as in Equation 3.4 and 3.5 for a range of p-value thresholds α is plotted against the false positive rate ($FPR = \alpha$). The FROC displays the relation between sensitivity and specificity.

3.4.2 Simulation results

This estimation procedure has been applied to the simulated data for both peaks and clusters. A figure of the estimated and true FROC can be found in Figure 3.6. It can be seen that there is a substantial deviation between the true and the estimated FROC for high values for α. However, realistic values of α are between 0 and 0.05, and within this range, the estimation procedure performs well. An overview of the different FROCs under different conditions can be found in the Appendix D.

3.5 Real Data Application

To further demonstrate the principle of the method for peak-level inference, we apply the presented technique to real fMRI data stemming from a single-subject event-related design (Henson et al., 2002). The study is 2 $\times$ 2 factorial with factors 'fame' and 'repetition'; famous and non-famous faces are presented twice against a checkerboard (see Henson et al. 2002 for more details). For this example, we look at the average effect of presenting faces, averaged over the levels of repetition.

As in the original paper by Henson et al. (2002), we use an uncorrected p-value threshold of 0.01 as excursion threshold, which reveals 97 peaks. Using the post-hoc power estimation of this procedure, we obtain

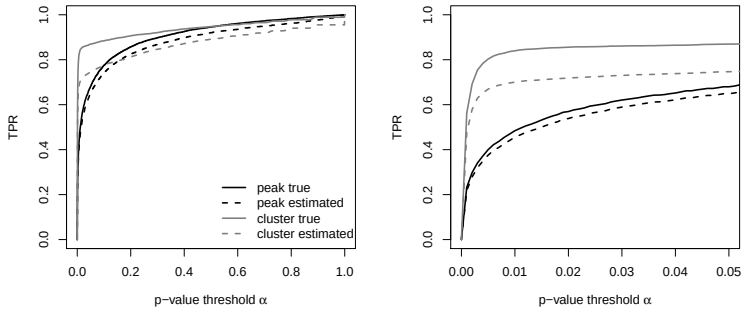


Figure 3.6: Free-Receiver Operator Curve (FROC). The left-hand side shows the complete FROC, the right-hand side shows the part of this FROC for realistically small α-values. Results are obtained with the basic scenario with $P(T_{398} \geq u) = 0.05$, SNR $= 0.015$, $\sigma = 2.5$ and with 5 blocks of activation.

an estimate for the proportion of null peaks (π_0) in the image of 36%. With this estimate, we are able to construct the FROC for this study, as can be seen in Figure 3.7.

As a first example, we use the thresholding procedure performed in the original paper by Henson et al. (2002), where they threshold these peaks to control the FWER at a level 5%. There are 35 peaks considered as active, i.e. these peaks are considered to lie within a brain region that becomes active when performing the experimental task involving faces (see Table 3.5).

We find that thresholding peaks to control the FWER at level 5% coincides with an estimated TPR of 55%. This means that we estimate that only 55% of the truly active peaks in this study are detected. The estimated FNR is equal to 44%, meaning that among 62 peaks we did not select, we estimate that 44% are actual active peaks. It is obvious that results indicate that the post-hoc power for an FWER control of 5% is too low as there is only a 55% chance to detect true activation.

Next, we consider results for thresholding peaks to control the FDR at level 5%, using the procedure described by Chumbley et al. (2010). This

procedure is the default in the popular software package SPM. Using this threshold, we find 52 significantly active peaks. With the procedure presented in this paper, we find an estimate for the TPR equal to 81% and for the FNR equal to 26%.

When taking a closer look at Figure 3.7, it can be seen that FDR control of 5% allows more false positives than FWER control at 5%, hence the increase in power. By using a different thresholding strategy, one shifts on the FROC. Therefore, when one aims at a power of 80%, which is a reasonable request in social sciences, the FDR procedure reaches this goal, while FWER control fails at providing enough power in this fMRI study. However, estimates for the FNR indicate that, even with FDR control, 26% of truly active peaks are not detected.

When thresholding p-values uncorrected for multiple comparisons at 5%, i.e. at a per-family error rate (PFER) of 5%, an estimated power of 84% is obtained with a FNR of 23%.

Region of activation	L/R	x	y	z	T	Region of activation	L/R	x	y	z	T
Fusiform gyrus	R	42	-43	-20	13.92	Inferior parietal lobe	R	27	-58	55	6.90
	R	36	-67	-11	14.38	Cuneus	L	3	-91	28	6.11
	L	-42	-64	-17	11.15	Insula	R	30	23	1	8.50
	L	-36	-76	-11	8.91		R	30	62	22	5.17
	L	-30	-76	-2	6.98	Inferior frontal gyrus	R	48	23	-11	5.43
Inferior occipital gyrus	R	36	-85	-5	6.86		R	51	20	4	5.79
Middle occipital gyrus	L	-30	-91	7	8.92		R	33	14	31	6.85
	R	27	-94	13	7.69		L	-36	26	1	7.15
	R	18	-94	22	5.96	Middle frontal gyrus	R	45	26	25	9.87
Inferior temporal gyrus	L	-45	-52	-17	12.53	Superior frontal gyrus	L	-24	-10	79	5.52
	L	-42	-31	-17	6.78	Precentral gyrus	L	-42	-13	67	6.23
Middle temporal lobe	R	45	-76	7	11.18		R	48	2	55	5.15
Superior temporal gyrus	R	48	-34	10	6.03	Postcentral gyrus	L	-54	-25	58	6.13
Lingual gyrus	R	9	-91	-8	7.91		L	-51	-13	58	5.54
	L	0	-79	1	7.99		L	-48	-31	64	6.63
Cerebellum	R	15	-67	-8	9.47		L	-57	-19	25	7.91
	R	27	-79	-14	10.47	Superior motor area	L	-3	14	55	7.71
Calcarine gyrus	R	21	-97	7	8.04						

Table 3.5: Peaks of regions showing effects of presenting faces. The coordinate system used is MNI.

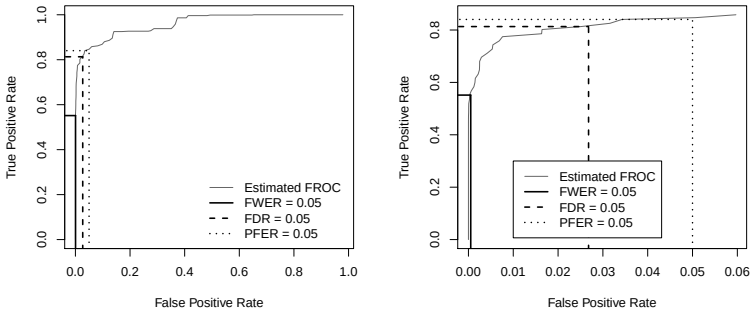


Figure 3.7: Estimated FROC for the Henson data set. The panel on the left shows the a range of FPR and TPR from 0 to 100%, while the right hand panel zooms in on the obtained values for FPR and TPR.

3.6 Discussion

We investigate estimating power for inference in fMRI research, based on the estimated proportion π_0 of truly non-active peaks or clusters. We show that for different signal-to-noise ratio conditions, π_0 can be estimated with a reasonable accuracy. Different estimators are available, but our results favor the method of Pounds & Morris (2003), since this estimator seems robust against differences in signal-to-noise ratio, smoothness, the size of π_0 and the excursion threshold u. We show that one can easily estimate a threshold's post-how power for detecting activation.

In social sciences, power studies are often performed a priori, i.e. before the actual experiment is performed. In these studies, the standard is to reach a power of 80%. For fMRI, such procedures are available for group-level analyses (Mumford & Nichols, 2008; Zarahn & Slifstein, 2001; Smith, Jenkinson, Beckmann, Miller, & Woolrich, 2007; Hayasaka, Peiffer, Hugenschmidt, & Laurienti, 2007; Desmond & Glover, 2002), but these mostly include an arbitrary standard error or effect size estimation. We presented a technique that can estimate power post hoc, without any assumptions on effect sizes. Our real data example shows that when one aims to see the activation related to viewing faces, only 50% of the ac-

tual activity can be discovered. Such findings will mostly impair the most vulnerable studies, such as studies involving social cognition, which are often characterised by small effect sizes and low power to detect the effects (Lieberman & Cunningham, 2009). We do not make an explicit comparison here between the a priori defined power and the obtained post-hoc power, as we mainly focused on single-subject designs, and power calculations present in literature compute power for group-level inference.

In the real data example, we can see that if we must attain a power of 80%, the type I error rate cannot be controlled at a FWER level, but should be controlled using a more lenient measure of false positives such as the false discovery rate (FDR) or the per family error rate (PFER). Controlling the FDR or PFER at 5% leads to more acceptable levels of power.

It is argued in a general setting that, post-hoc power estimation procedures make no sense since, post-hoc, test results are no longer random: whether the peak is active or not is fixed (Hoenig & Eisey, 2001). However, in a multiplicity setting, post-hoc power does not relate to the result of one statistical test, but many different tests. Therefore it is useful to consider how many activation is missed on average. Using π_0, we can estimate how many of the peaks are actually active. Of course, this does not tell us where the activity can be found. Hoenig & Eisey (2001) argue against the use of post-hoc power estimation, as this often involves calculating the probability of finding the observed effect *'if it were true'*. In that case, the alternative hypothesis H_a is always centered around this observed effect. As for non-significant results, the applied T-value threshold is higher than this observed effect, the T-threshold will always be on the upper half of H_a for a non- significant result. Therefore, non-significant tests will always have a power lower than 50%. Inversely for significant results, the used T-threshold will always be on the lower half of H_a and significant tests always lead to powers higher than 50%. As the power estimation procedure presented in this paper does not involve test-specific power estimation based on the observed effect, but rather the overall performance in a multiplicity context, the argument of Hoenig & Eisey (2001) is no longer valid.

We show that the used estimators are vulnerable to differences in smoothness: higher smoothness leads to lower bias. Peaks and clusters

are spatial features, and therefore require a certain amount of spatial continuousness. The spatial continuousness is improved by pre-smoothing the data. The higher the level of smoothness, the more accurate the p-values for the features, and therefore, the more accurate our estimation procedure becomes.

Two important words of caution should be made on the use of peaks and clusters. The first is on the random character of the topological features. The importance of this study has been evaluated with simulated data sets. However, in each simulation step of a given scenario, the features appear at random locations. Secondly, peak and cluster p-values can only be estimated conditional on the excursion threshold u. Therefore, for power estimation, features below the threshold are not considered. In this study, we did not account for any true or false negatives below the excursion threshold. This is common practice in real studies as topological features below the excursion threshold are not considered. The problem of choosing the excursion threshold is acknowledged in literature and solutions have been provided (Adler, Bobrowski, & Borman, 2010; Smith & Nichols, 2009). The techniques shown in this paper are also applicable to threshold-free techniques, as the only required input are p-values on the topological features. All techniques to obtain p-values for topological features, such as threshold-free cluster enhancement (Smith & Nichols, 2009) or cluster mass inference (Zhang et al., 2009) can be combined with the procedure presented here.

It should also be pointed out that while the procedure presented here was mainly applied to single-subject fMRI designs, it is also suitable for group-level inference. When a statistical parametric map is obtained for any design, it is possible to derive the topological features and corresponding p-values enabling power estimation. Consequently, the procedure can also be used for other neuroimaging techniques such as voxel-based morphometry and diffusion tensor imaging.

The computational complexity of this procedure is very low. Estimating $\widehat{\pi}_0$ using the method provided by Pounds & Morris (2003) includes a computational minimization of a two-parameter function, but as the number of peaks or clusters is much smaller than the number of voxels, the minimization takes less than one minute on an Intel Core i7 3,4GHz with 4 GB of memory. Once $\widehat{\pi}_0$ has been estimated, estimating the power

is only a matter of simple calculus.

To conclude, we argue that post-hoc power estimation gives unique and insightful information on a completed study. Our real data example shows that the estimated power is dramatically low when controlling the FWER at 5 %. This means that for this study, no clear conclusions can be drawn for areas where the null of no activation is not rejected since only half of the true activation is detected. We expect that these results are not unique, since the effect of faces in the Henson data set has been replicated many times. We expect that many activations in less powerful experiments, such as social cognition experiments, remains unexplored. Low power of 8% on average was also reported by Button et al. (2013) on average over 91 different studies. They argue for more attention improving reproducibility in neuroscience. One possibility to increase sensitivity in neuroimaging studies, could be multivariate analyses that consider several voxels simultaneously, such as multi-voxel pattern analysis. (Kriegeskorte, Goebel, & Bandettini, 2006) or canonical correlation analysis (Friman, Borga, Lundberg, & Knutsson, 2003). These multivariate approaches yield a higher power than univariate GLM analyses. We argue that a call needs to be made to statistical testing procedures with a better balance between sensitivity and specificity (Lieberman & Cunningham, 2009). By raising awareness on the power of existing studies, this procedure is a first step in this direction.

Acknowledgements

The computational resources (STEVIN Supercomputer Infrastructure) and services used in this work were kindly provided by Ghent University, the Flemish Supercomputer Center (VSC), the Hercules Foundation and the Flemish Government - department EWI.

We sincerely thank three reviewers for their helpful and insightful comments. We are grateful to dr. Jonathan Taylor for guidance on the formulation of Random Field Theory. We thank dr. Ruth Seurinck and Han Bossier for useful comments and Sanne Roels for help with debugging the code.

Appendices

3.A Comparison of topological random field theory and non-parametric p-values

3.A.1 Simulations under the null hypothesis

Time series of 400 time points are constructed, with for each time point an image of $40 \times 40 \times 40$ voxels, which corresponds to moderate resolution data. Images are generated using neuRosim (Welvaert et al., 2011) with temporal correlated noise. The temporal dependence is fixed to $\rho = 0.20$. The error term for each voxel on timepoint i is created using the following equation.

$$\epsilon_i = \rho\epsilon_{i-1} + \xi_i \tag{3.8}$$

$$\xi_i \sim N(0, \sigma^2) \tag{3.9}$$

The images are convolved with a 3D isotropic Gaussian kernel of width 9 and $\sigma = 2.5$.

A design matrix is constructed for a blocked design with 10 blocks of the active condition, altered with 10 blocks of non-activity. The active condition does not have any effect on the brain images, and therefore, the null hypothesis of no activation is true for the whole brain map. A GLM is fit to the data and a T-statistic image results with 398 degrees of freedom. The excursion threshold u is set at $P(T_{398} \geq u) = 0.01$.

After the excursion set is defined, peaks and clusters on the T-image are sought with a 26-point clustering algorithm. p-values for peaks and cluster size are calculated using both random field theory and permutations. RFT cluster p-values are computed using FSL (Smith et al., 2004). From the simulated data we also calculate the empirical or observed p-values, the percentage of features that are as large or larger. In the following section, these observed p-values are compared to the RFT and permutation p-values.

In each simulation, on average 86.29 peaks and 68.99 clusters were discovered. The average cluster size is 23.29 voxels, and the average peak height is equal to 3.00. 500 simulations were performed.

3.A.2 Real data under the null hypothesis

Analogous to what is described in Eklund, Andersson, Josephson, Johannesson, & Knutsson (2012), we have analysed 500 resting state fMRI data sets with an arbitrary (non-resting) experimental design. As subjects were at rest, the null hypothesis is true and this data can be use to compare with RFT and permutation inference. We have used the freely available resting state fMRI data sets in the Neuroimaging Informatics Tools and Resources Clearinghouse (NITRC) 1000 functional connectomes project (Biswal et al., 2010). The data sets are described in Table 3.6. All data sets were analyzed with FSL. Each data set was motion corrected and smoothed with a kernel with FWHM 8mm. No motion regressors are included. A block based design was used with altering periods of 10 seconds of activity and rest. The design is convolved with the hemodynamic response function and its temporal derivative. The analyses result in t-statistics for each voxel. These are subject to topological inference with an excursion threshold of $u = 2.65$. The data sets result on average in 23.01 clusters and 28.42 peaks. Using RFT and permutation inference, p-values are computed for each suprathreshold cluster or peak. As we did not wish to assume that the peak values or cluster sizes were comparable over different datasets, we did not compute observed p-values. Therefore, we only analyse the distribution of the topological p-values over data sets.

For the permutation tests, only 20 permutations were performed for each data set due to its computational complexity, but as the data sets emit an average of at least 20 clusters/peaks per data set, the permutation distribution is based on over 400 values, which is acceptable.

3.A.3 Results

Simulated data sets

Cluster p-values Figure 3.8 shows that the bias of the RFT p-values is larger than the bias of the permutation p-values, with a much wider

Institution	Persons	Number of subjects
AnnArbor	Monk, C.S., Seidler, R.D., Peltier, S.J.	25
AnnArbor	Monk, C.S., Seidler, R.D., Peltier, S.J.	36
Atlanta	Mayberg, H.S.	28
Baltimore	Pekar, J.J., Mostofsky, S.H.	23
Bangor	Colcombe, S.	20
Beijing	Zang, Y.F.	198
Berlin	Margulies, D.	26
Cambridge	Buckner, R.I.	146

Table 3.6: Overview of the rest data sets used for the empirical study.

variance. When looking at the absolute deviation for each estimate, the results are more striking: on average, an RFT cluster p-value has a deviation of 40% from the real p-value. This deviation is much smaller for permutation p-values.

On the left panel of Figure 3.8, it can be seen that the deviation of the RFT p-values follows a specific pattern. p-values estimated with RFT mostly overestimate the true p-value. The permutation p-values shows a straight pattern, as can be expected. There is also a large mean bias for the RFT estimates, while these quantities are rather low for the permutation p-values.

These results are also reflected in Figure 3.9. Permutation p-values and observed p-values are uniformly distributed, as expected under the null hypothesis. The histogram of the RFT p-values however show a deviation from this uniform distribution.

We can conclude from this small simulation study that permutation p-values are preferrred to obtain p-values for clusters.

Peak p-values For peak p-values, better results for RFT are obtained as can be seen in Figure 3.10 and 3.11. The variance on the estimation bias is also smaller, and the histogram shows a more uniform distribution. We therefore conclude that both estimation procedures perform equally well.

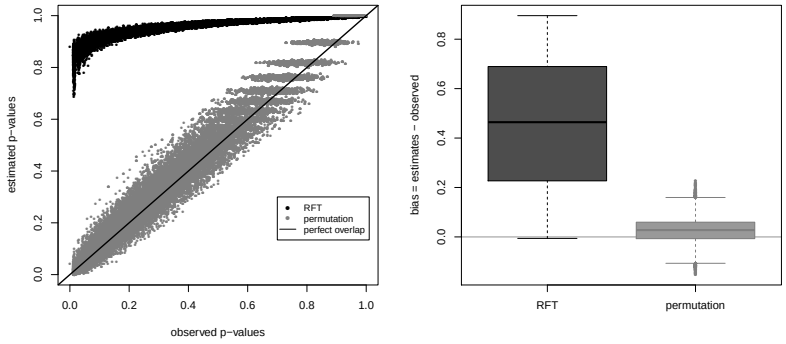


Figure 3.8: (left) Plot of estimated p-values against observed p-values for the simulated data, (right) boxplot of bias for cluster p-values.

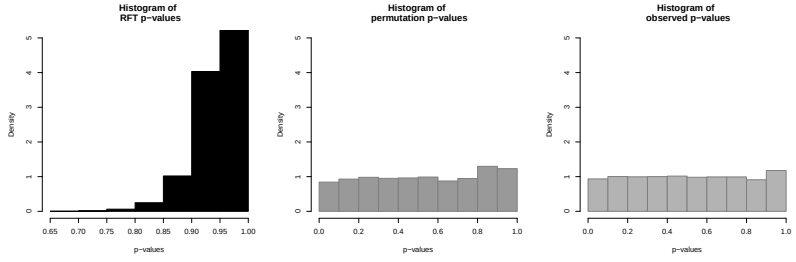


Figure 3.9: Histogram of estimated and observed p-values for clusters for the simulated data.

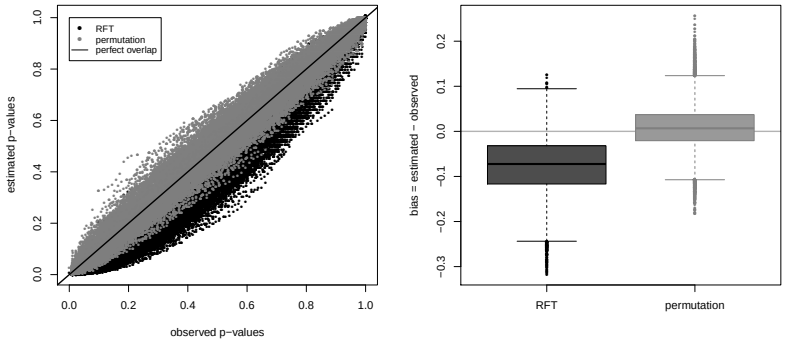


Figure 3.10: (left) Plot of estimated p-values against observed p-values, (right) boxplot of bias for peak p-values for the simulated data.

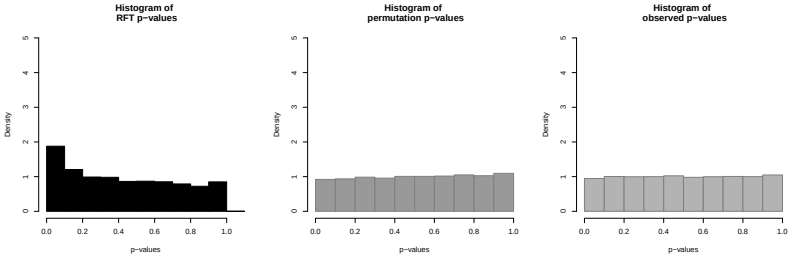


Figure 3.11: Histogram of estimated and observed p-values for peaks for the simulated data.

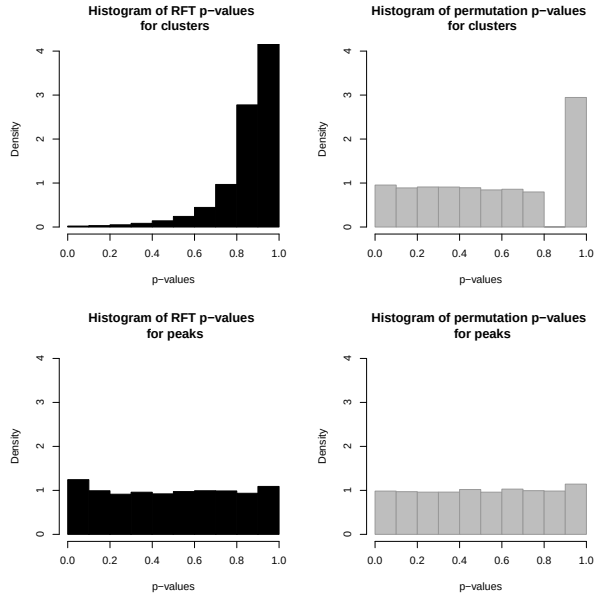


Figure 3.12: Histograms of the estimated p-values for clusters (upper panels) and for peaks (lower panels) for both RFT (left) and permutation (right) for the resting state data.

Resting state data sets

Cluster p-values In Figure 3.12, it can be seen that the p-values obtained with RFT for clusters overestimate the real p-value. This conclusion is in line with the conclusions drawn with the simulation study. We can conclude that permutation based p-values perform much better, except for the very small clusters of size 1, that all result in an equal p-value very close to 1.

For peak level inference based on permutation, we can draw the same conclusions as with the simulated data sets, namely that both RFT and permutation based methods result in a uniform distribution, in concordance with the expectance.

3.B Influence of different parameters on the estimation of π_0

As can be seen in Figure 3.13, increasing the number of activated fields in the brain, and therefore decreasing π_0 does not change the performance of the estimators for peaks. $\hat{\pi}_0$ remains conservatively overestimated for a different number of activated regions with the BH and S estimator. The bias and the RMSE of the PC estimator are small and independent of the number of activated regions. PM provides good estimates with in each simulation an average deviation of less than 5%.

The same results can be seen in cluster size thresholding, where PM provides the least biased estimates with the smallest RMSE.

When the signal-to-noise ratio is varied, as is shown in Figure 3.14, results are ambiguous. For peaks, low SNR-values produce large deviations between the estimated π_0 and its true value. Larger values for SNR result in good estimates for PC and PM, while ST, S and BH largely overestimate π_0. This is also reflected in the RMSE. PM provides a low RMSE for high SNR-values, S, ST and BH provide a higher RMSE regardless of the SNR.

SNR influences the estimation of π_0 to a less extent for cluster thresholding. Here, again PM provides the best estimates.

When varying the smoothness (Figure 3.15), the same pattern appears for peaks and clusters. When the smoothing parameter σ is very low, the results are very unstable and volatile. After a certain point ($\sigma \sim 2.5$), the performance of the estimators become stable. For peaks, this can be due to the assumptions of RFT, where the data are required to be continuous fields. When no smoothing is applied, the brain image is not spatially continuous, and with this assumption broken, the peak p-values are unreliable. For clusters, when no smoothing is applied, less consistent clusters appear and the excursion set follows a speckled pattern. Conseqently this could cause the non-parametric p-values to become unreliable. Again PM shows the smallest bias and the lowest RMSE for both clusters and peaks.

In Figure 3.16, we varied the excursion threshold u. We find that the higher the threshold, the smaller the RMSE for all estimators for both peaks and clusters. Varying the excursion threshold only influences the bias of the PM estimator to a small extent for peaks. For clusters, all methods are dependent on the excursion threshold.

When summarizing all results, we find that PM is rather stable over smoothness and SNR.

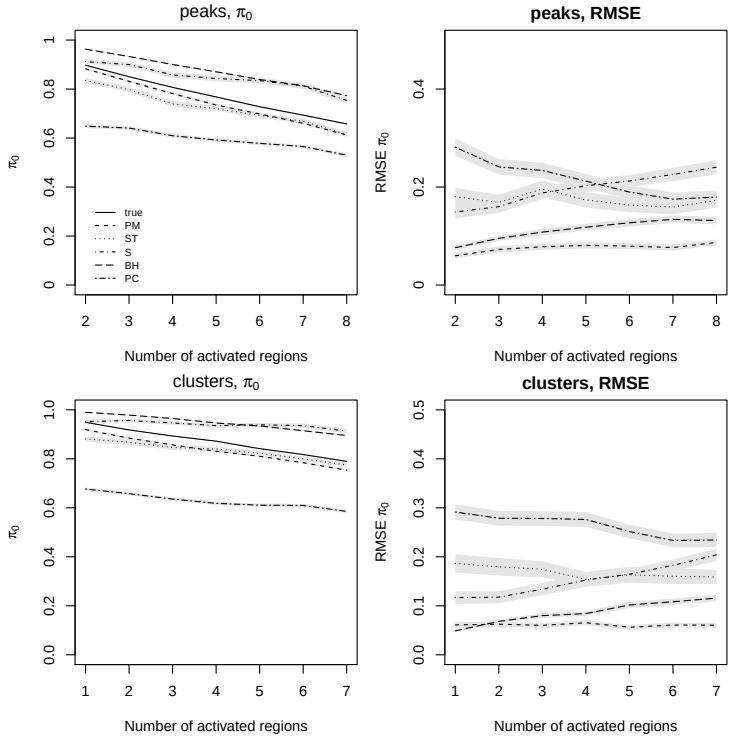


Figure 3.13: Bias and RMSE of π_0 estimators with 95% confidence interval. For 500 simulations, the number of activated regions is varied, while all other parameters are held constant (SNR= 0.04, smoothing parameter $\sigma = 2.5$, excursion threshold $P(T_{398} \geq u) = 0.01$). The RMSE is defined as $\sqrt{1/n \sum_{i=1}^{n} (\widehat{\pi_{0i}} - \pi_{0i})^2}$.

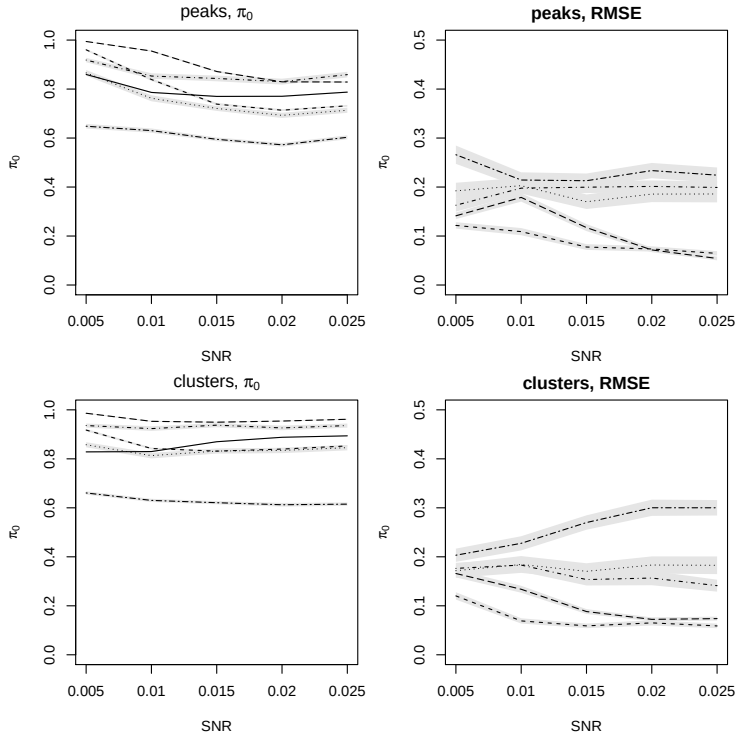


Figure 3.14: Bias and RMSE of π_0 estimators with 95% confidence interval. For 500 simulations, the signal-to-noise ratio is varied, while all other parameters are held constant (5 activated fields, smoothing parameter σ = 2.5, excursion threshold $P(T_{398} \geq u) = 0.01$)). The RMSE is defined as $\sqrt{1/n \sum_{i=1}^{n} (\widehat{\pi_{0i}} - \pi_{0i})^2}$.

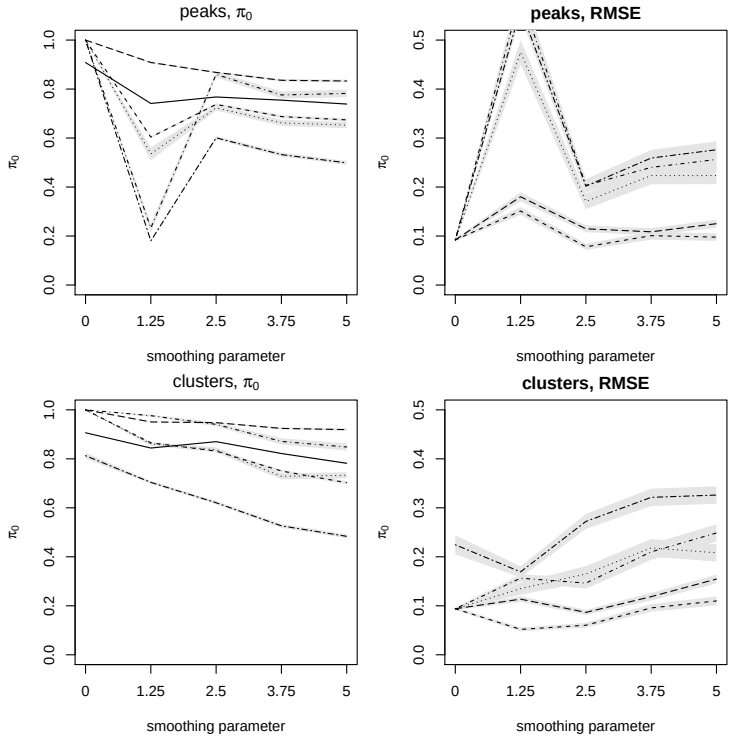


Figure 3.15: Bias and RMSE of π_0 estimators with 95% confidence interval. For 500 simulations, the smoothness of the image is varied, while all other parameters are held constant (5 activated fields, SNR = 0.04, excursion threshold $P(T_{398} \geq u) = 0.01$). The RMSE is defined as $\sqrt{1/n \sum_{i=1}^{n} (\widehat{\pi_{0i}} - \pi_{0i})^2}$.

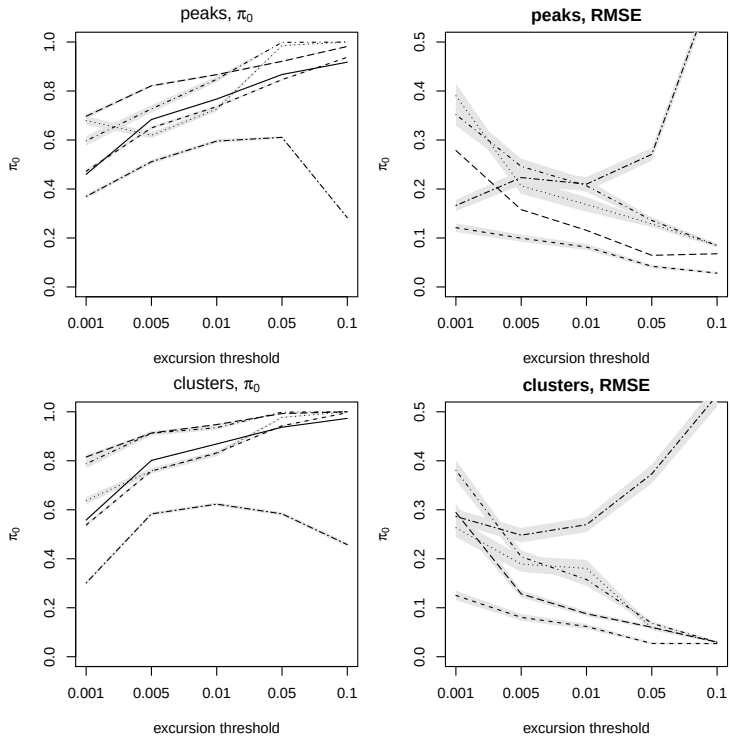


Figure 3.16: Bias and RMSE of π_0 estimators with 95% confidence interval. For 500 simulations, u is varied, while all other parameters are held constant (5 activated fields, smoothing parameter $\sigma = 2.5$, SNR=0.04). The RMSE is defined as $\sqrt{1/n \sum_{i=1}^{n} (\widehat{\pi_{0i}} - \pi_{0i})^2}$.

3.C The influence of the duration of the experiment on the power of the study

We have repeated the simulation paradigm described in section 3.2.2, while varying the total number of time points. We have simulated a blocked design with blocks of 20 time points. The image is made of $40 \times 40 \times 40$ voxels, which corresponds to moderate resolution data. Images are generated using neuRosim (Welvaert et al., 2011) with Gaussian white noise. First, for 5 regions of $7 \times 7 \times 7$ voxels, a signal is added in half of the time points with signal-to-noise ratio (SNR) of 0.015 before smoothing. The images are convolved with a 3D isotropic Gaussian kernel of width 9 and $\sigma = 2.5$. A general linear model (GLM) is fit to the data resulting in a T-statistic image with 398 degrees of freedom. The excursion threshold u is set at $P(T_{398} \geq u) = 0.01$. After the excursion set is defined, the peaks and clusters on the T-image are sought with a 26-point clustering algorithm and their uncorrected topological p-values are calculated.

We have considered 4 different scenario's with the number of time points equal to 100, 200, 300 or 400. Consequently the experiment consisted of 5, 10, 15 or 20 blocks. The resulting TPR and FNR can be seen in Figure 3.17. It can be seen that with an increasing number of timepoints, the power of the study as measured by the TPR also increases. There is no clear relationship with the resulting FNR. These results are in concordance with the results of the influence of the SNR, as can be seen in Figures 3.3 and 3.5.

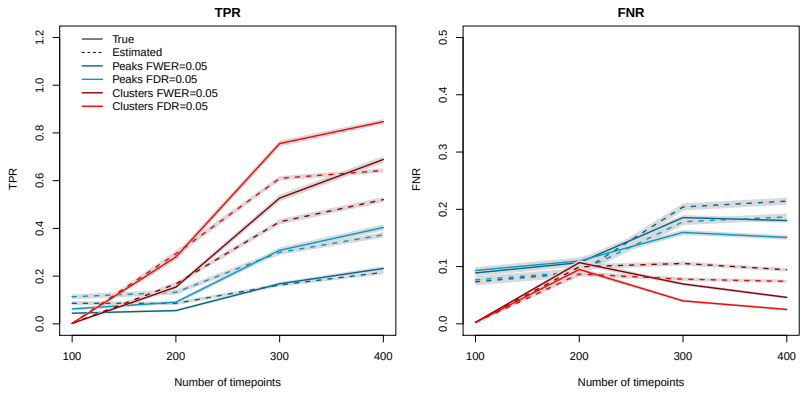


Figure 3.17: The influence of the number of time points on the observed and estimated TPR and FNR.

3.D FROCs under different conditions

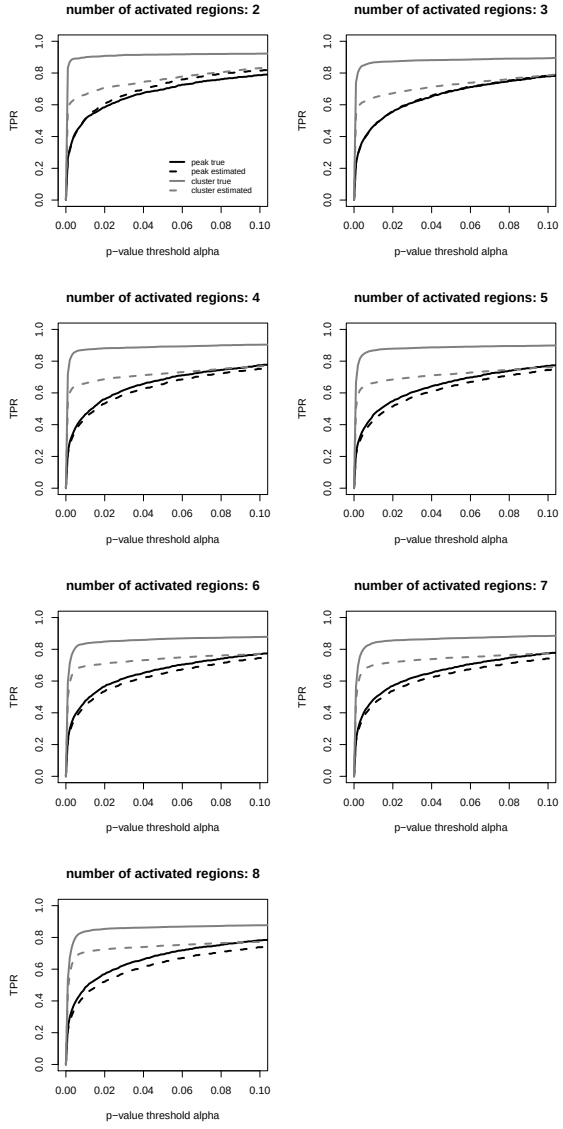


Figure 3.18: FROCs when varying the number of activated fields.

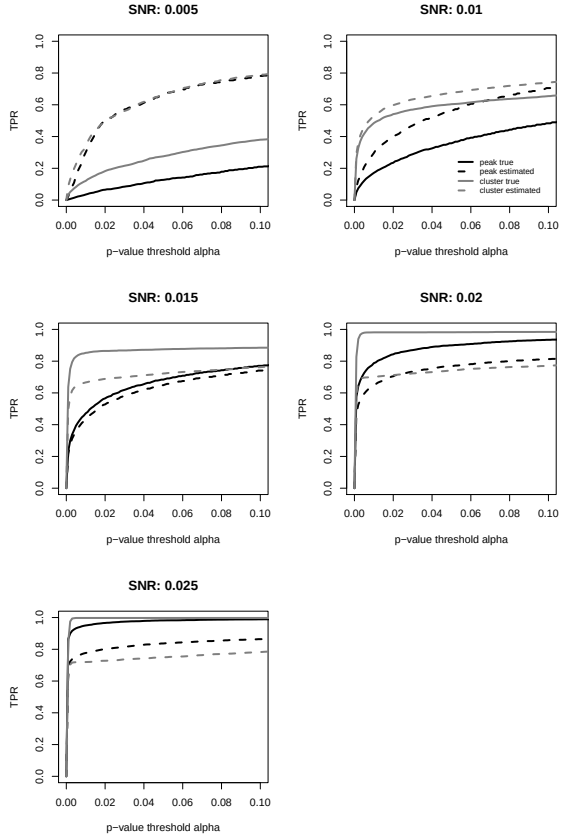


Figure 3.19: FROCs when varying the SNR.

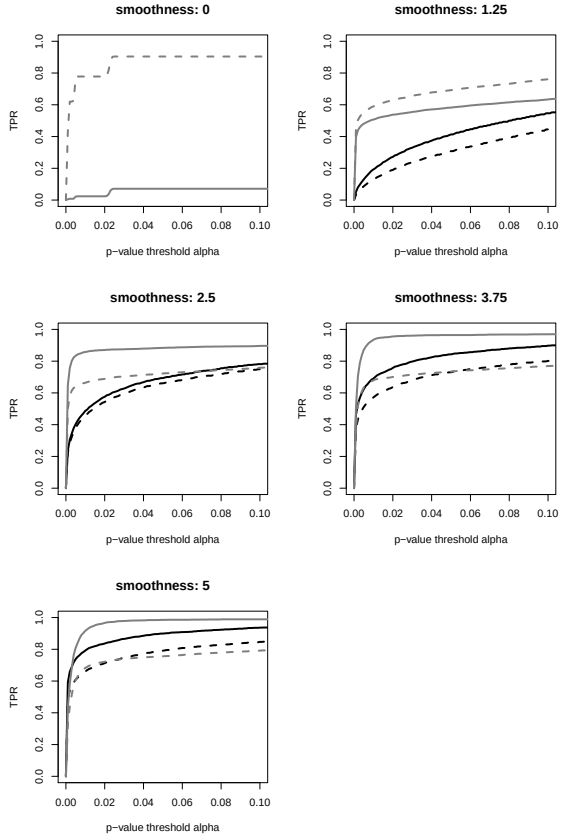


Figure 3.20: FROCs when varying the smoothing parameter σ.

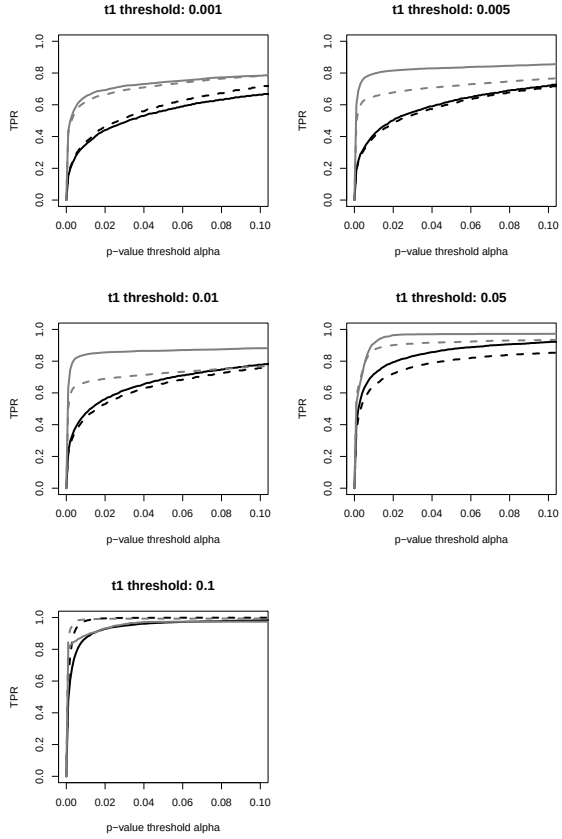


Figure 3.21: FROCs when varying the excursion threshold.

4 Simplified power and sample size calculations using prevalence & magnitude of active peaks.[1]

Abstract There is increasing concern about statistical power in published neuroscience research, and not the least for published fMRI experiments. Not only does low power decreases the chance of detecting a true effect, but also reduces the chance that a statistically significant result indicates a true effect (Ioannidis, 2005). Put another way, studies with the least power will be the least reproducible, and thus a (prospective) power analysis is a critical component of any paper. In this work we present a simple way to characterize the spatial signal in a fMRI study, and a direct way to estimate these two parameters based on an existing study. Specifically, using just the (1) the proportion of the brain activated and (2) the average effect size in activated brain regions, we can produce closed form power calculations for given sample size, brain volume and smoothness. This procedure allows to minimize the cost of an fMRI experiment, while preserving a predefined statistical power. The method is evaluated and illustrated using simulations a real data example, and an application to data from the Human Connectome Project. The procedures presented in this paper are made publicly available in a toolbox.

4.1 Introduction

The goal of a prospective power analysis for fMRI experiments is twofold. A first task to be performed is to optimization of the experimental design to ensure maximal statistical power for the study goal. Methods have been developed to find optimal experimental designs (Wager & Nichols, 2003; Friston, Zarahn, Josephs, Henson, & Dale, 1999; Smith et al., 2007). The second aim is to guarantee an optimal number of subjects, which has been discussed by Desmond & Glover (2002) and Mumford & Nichols (2008). One typically aims for a statistical power of 80% implying that a true effect in the population is detected with a 80% chance. Given statistical

[1]This chapter represents collaborative work with following authors: Durnez J., Burgess G., Degryse J., Seurinck R., Barch D., Moerkerke B. and Nichols T.

procedures, it is possible to compute the minimal number of subjects to obtain the aimed statistical power. As such, power calculations prevent researchers to spend time and money on studies that are underpowered, but on the other hand also prevents researchers to invest too much time and money in scanning extra subjects, while the accompanying increase in power is only minimal.

While it is straightforward to compute power for a single, univariate response, determining the power of an fMRI study is a formidable task as many parameters need to be specified, such as a within and between subject variance, the first and second level design, the temporal autocorrelation and the size of the hypothesised effect. Many of these parameters are estimated based on a pilot study. But the most difficult parameter to specify will be the region-of-interest, the ROI, the specific location where activations are expected. It is dangerous to base this on the results of the pilot study, as this results in circularity (Vul, Harris, Winkielman, & Pashler, 2009; Kriegeskorte, Simmons, Bellgowan, & Baker, 2010).

In this work we present a simple way to characterize the spatial signal in an fMRI study, and a direct way to estimate power based on an existing pilot study. Specifically, using just (1) the volume of the brain activated and (2) the average effect size in activated brain regions, we can directly calculate power for given sample size, brain volume and smoothness. This procedure allows minimising the cost of an fMRI experiment, while preserving a predefined statistical power.

The procedure is an extension of the procedure presented in Durnez, Moerkerke, & Nichols (2014), where we estimate the peakwise prevalence of activation π_1 in an experiment. The estimate of π_1 allows to predict the number of false negatives and the total number of activated peaks. The ratio between these two, gives a post hoc estimate of the average power. In the new procedure we present here, we start with the estimation of π_1 based on the pilot data set. Using this parameter, we fit peak heights above cluster forming threshold u as a mixture of two distributions, one for null peaks and one for peaks in activated regions. Once an estimate for the alternative distribution H_a is known, this distribution can be transformed to an alternative distribution for a larger sample size. As such, not only the power of the pilot study can be estimated, but also a new study with the same experiment and a larger sample size. This allows to perform

sample size calculations.

In this paper, we present the procedure, and we evaluate the procedure based on simple simulations for different combinations of fMRI characteristics, such as spatial extent of the signal, peak forming threshold and signal intensity. Next, we show the procedure on a typical example of an fMRI experiment using an fMRI dataset (Seurinck, de Lange, Achten, & Vingerhoets, 2011). Third, we present results of the procedure applied to the data from the Human Connectome Program (Van Essen et al., 2012). First, these data is of a very high quality, resulting in a very high power when taking all subjects. Therefore, when analysing all data, we have a high level of certainty of the location of the effect. When we use subsamples of these data, we recreate smaller fMRI studies. The results thereof can be compared to the results of the full dataset, as is the case in simulations. The added value of the HCP data is that many unknown noise sources in fMRI are present that cannot be added to simulated data. Finally we conclude with a discussion on the topic and the implementation of the procedure in an `R`-package.

4.2 Methods

4.2.1 Measures of power

Consider a single univariate test for which the null hypothesis H_0 is rejected in favor of the alternative hypothesis H_a when the test statistic Z exceeds threshold z with z chosen to control the type I error rate. The power of this test is defined as $P(Z \geq z|H_a)$. To calculate this power, we need the distribution of Z under H_a.

In a multiplicity context where many tests are performed simultaneously such as voxelwise testing, several definitions for power exist (see for example Dudoit, Shaffer, & Block, 2003). Let $\mathcal{I}_1$ denote the set of indices for voxels that are truly activated. For each voxel i with $i \in \mathcal{I}_1$ with corresponding test statistic Z_i, the power is $P(Z_i \geq z)$. Under the assumption of an equal distribution for all Z_i with $i \in \mathcal{I}_1$, this voxelwise power can be written as $P(Z_i \geq z|i \in \mathcal{I}_1)$ and can be given the interpretation of average power over all activated voxels. Alternatively, as a counterpart for controlling the familywise error rate (FWER) or the

probability of at least one type I error, familywise power is defined as $P(Z_i \geq z \text{ for at least one } i \in \mathcal{I}_1)$.

Equally, we can define average and familywise power for peak-wise tests where only peaks larger than the excursion threshold u are considered and where we have a test statistic Z_j^u in each peak j. Let $\mathcal{J}_1$ denote the set of indices for peaks above u corresponding to a voxel containing true signal. Average power is then defined as $P(Z_j^u \geq z | j \in \mathcal{J}_1)$ while familywise power is defined as $P(Z_j^u \geq z \text{ for at least one } j \in \mathcal{J}_1)$.

In what remains, we focus on average power. First, it is the most intuitive measure of power as it reflects the average probability of a rejection within the set of false null hypotheses. Secondly, familywise power may not be a satisfactory measure as a high probability of at least one rejection within the set of truly activated peaks may still be accompanied by a large proportion of false negatives.

4.2.2 Estimation Procedure

To calculate average peakwise power, it is key to know the distribution of peak heights for active peaks (distribution under H_a). Using excursion threshold u, peak heights of non-active peaks follow an exponential distribution with mean $1/u$ (Worsley, 2007). We further assume that the distribution of active peak heights $\mathcal{J}_1$ is a normal distribution truncated at excursion threshold u. We define Z as the peak heights, and Z^u the peak heights above clusterforming threshold u. With π_1 the proportion of activated peaks among the total number of peaks above u, the distribution of peak heights Z^u above u can be written as the following mixture distribution:

$$f(z^u|\pi_0, \mu_1, \sigma_1, u) = (1-\pi_1) \underbrace{u \exp(-u(z^u - u))}_{\text{null distribution}} + \pi_1 \underbrace{\frac{\frac{1}{\sigma_1}\varphi\left(\frac{z^u-\mu_1}{\sigma_1}\right)}{1-\Phi\left(\frac{u-\mu_1}{\sigma_1}\right)}}_{\text{alternative distribution}} \tag{4.1}$$

with ϕ and Φ respectively representing the density and cumulative distribution function of the standard normal distribution.

To obtain π_1 for a certain dataset, different estimators have been pro-

posed in literature (Benjamini & Hochberg, 2000; Storey & Tibshirani, 2003; Storey, 2002; Pounds & Morris, 2003; Pounds & Cheng, 2004). All of these use the observed distribution of the p-values to determine what proportion of p-values stem from the null. A comparison of the estimators for peak inference in fMRI data analysis have been discussed in Durnez, Moerkerke, & Nichols (2014).

Once π_1 is estimated for a dataset with sample size n, we estimate the remaining parameters in Equation 4.1 using maximum likelihood on the same data. This renders the parameters μ_1 and σ_1 of the truncated peak distribution under the alternative H_a. Let $Z_j^u = (\overline{b_j}/\sigma)\sqrt{n}$ denote an active peak ($j \in \mathcal{J}_1$) with $\overline{b_j}$ the average experimental effect for voxel j over n subjects and σ the standard deviation of b_j over n individuals. The expected value of the peak height under the alternative before truncation equals $\mu_1 = E(\overline{b_j}/\sigma)\sqrt{n} = (\mu/\sigma)\sqrt{n}$. We define $\delta = \mu/\sigma$.

Consequently, for a new sample of size n^*, the distribution of activated peaks before truncation will be $\mathcal{N}(\mu_1^*, \sigma_1)$ with $\mu_1^* = \delta\sqrt{n^*}$. We hereby assume that the variance of the distribution σ_1^2 remains constant for different sample sizes. Given a statistical threshold z such that the null hypothesis of no activation is rejected for a peak if the height of the peak is larger than z, the average peakwise power with truncation at peakforming threshold u can be calculated as $P(Z^u \geq z)$ where Z^u follows again the truncated normal distribution described above. $P(Z^u \geq z)$ is computed using the cumulative density function of a truncated normal distribution, given by

$$P(Z^u \geq z) = \frac{\Phi\left(\frac{z-\mu_1^*}{\sigma_1}\right) - \Phi\left(\frac{u-\mu_1^*}{\sigma_1}\right)}{1 - \Phi\left(\frac{u-\mu_1^*}{\sigma_1}\right)} \tag{4.2}$$

The statistical threshold z can either be uncorrected or corrected for multiple testing.

4.2.3 Simulations

For a given sample size n, we generate n statistical parametric maps, summarizing evidence for activation. These represent for each subject the average effect of the design in a first-level analysis, i.e. a three-dimensional

map of b-values. We simulate n single $208 \times 208 \times 208$ white standard normal noise Gaussian images with voxel sizes of $3 \times 3 \times 3$ mm. The sample size n varies between 10 and 30 subjects. Images are smoothed with a 3D Gaussian kernel with a full-width at half maximum (FWHM) of 8 mm. We truncate the outer 72 voxels from the smoothed images to avoid non-uniform smoothness at the edge, resulting in a $64 \times 64 \times 64$ image. In each b-map, true activation is imposed on a block of $33 \times 33 \times 33$ voxels, which spans 14 % of the total brain volume. Within this activation block, the effect size varies between 0.25 and 2.25 units for all n subjects. In a next step, a t image is obtained by calculating a one-sample t test statistic on the subject-specific b-maps. In this t image, peaks are defined. All peaks are transformed from T to Z statistics and only peaks above cluster forming threshold u are considered, where u varies between 0 and 4 in the simulations. This allows to use random field theory for the computation of peak p-values:

$$P(Z \geq z | Z \geq u, H_0) \approx \exp\left(-u(z-u)\right) \tag{4.3}$$

Peaks located on the border of the activation (2 voxels at the inner or outer border) are discarded, as voxels at the inner or outer border have respectively a higher or lower probability to become a peak than elsewhere in the image. Voxels in the non-active area that border the activated are adjacent to maximal three voxels that are active, and are therefore higher on average. Therefore these voxels have a lower probability of becoming a peak. Likewise, voxels in the inner border of the active area neighbour at least three voxels that are on average lower. These voxels thus have a higher probability of being a peak.

We use these simulations to study the performance of the procedure described in section 3.1. More specifically, in a given study, the so-called *pilot study*, we estimate π_1 - using the procedure presented by Pounds & Morris (2003) - and the distribution of the truncated normal distribution $N(\mu_1, \sigma_1)$ to predict power of future studies in function of sample size. This predicted power is compared to the actual power in simulated images for which the underlying truth is known. The true power in these images is obtained as the ratio of the number of discovered peaks in activated areas and the total number of peaks in activated areas. In total, 100

simulations are performed. Hence, power for a sample size n_1 is estimated in each of 100 simulated pilot studies of sample size n_0 and the average over these simulations is compared to the average of the actual power in 100 simulated images of sample size n_1. In the same manner, we compare the estimated π_1 to the corresponding true values. The true underlying π_1 is obtained by calculating the percentage of peaks that are located in activated areas.

For the true effect size, $E(Z_j^u)$ is computed as the average peak height of peaks in activated areas. However, we need to take into account that the estimated effect size μ_1 is the effect size before truncation, and does not equal $E(Z_j^u)$. The expected value of the peak height under the truncated alternative equals $E(Z_j^u|j \in \mathcal{J}) = \mu_1 + \sigma_1\tau$, with τ

$$\tau = \frac{\varphi\left(\frac{u-\mu_1}{\sigma_1}\right)}{1-\Phi\left(\frac{u-\mu_1}{\sigma_1}\right)} \tag{4.4}$$

In order to provide a good comparison, we compare $E(Z_j^u)$ to the expected peak height of the estimated truncated normal distribution, which equals $\hat{\mu_1} + \hat{\sigma}_1\hat{\tau}$.

We consider statistical thresholds z for peak height that are (a) uncorrected for multiple testing and correspond to α-levels of 0.001 and 0.05, (b) corrected to control the FWER at level 0.05 using a Bonferroni correction and (c) corrected to control the false discovery rate (FDR) at level 0.05. False discovery rate control is done with the correction of Benjamini & Hochberg (1995) and of Storey (2002). These thresholds are computed based on the null distribution of peaks as shown in Equation 4.1. For (b), it is assumed that the number of peaks will be the same in a new sample as in the pilot study. For (c), it is important to note that FDR correction uses an adaptive, data dependent threshold, but in a priori power calculations, we use a prediction of this threshold for future studies with typically larger sample sizes. We use the FDR threshold from the pilot study as a cutoff in the larger sample sizes. The rationale behind this is that the average power will increase with sample size, and consequently the FDR cutoff will decrease. Therefore, using the original FDR threshold will result in conservative results.

4.2.4 Data example

We apply our estimation procedure in a study on the role of higher ordered visual areas for imagined visual motion of Seurinck et al. (2011). The original study consists of 13 subjects and was analysed voxelwise with an FWER control at the 5% significance level. We re-analyse the data as follows: first level analyses were carried out with a standard voxelwise GLM-approach using FSL's feat function (Jenkinson, Beckmann, Behrens, Woolrich, & Smith, 2012). Consequently, the second level analysis is performed in R with ordinary least squares, resulting in a 3D T-statistic map. In this map, peaks are defined with an excursion threshold $u = 2$ and a 18-points clustering algorithm. Topological peak p-values are computed using Random Field Theory as in Equation 4.3.

The prevalence of activated peaks is computed with the procedure presented by Pounds & Morris (2003). Consequently, we transform the distribution for activated peaks based on 13 subjects to distributions for sample sizes between 10 and 40 subjects. Power is calculated for different statistical thresholds.

4.2.5 HCP data

Simulations provide a good way to evaluate a procedure, but they often fail to capture all subtleties of real data. Instead of adjusting our simulations to capture as much as noise components as possible, we also evaluate the procedure on real data. Therefore we use task data from the human connectome project (HCP) (Van Essen et al., 2012). We have data from 80 subjects for a simple working memory task. With such a large pool of participants, we assume that any group analysis of > 50 participants results in high powered results and can reflect population results.

The procedure used to analyse these data is showed in Figure 4.1 and is described below.

We start with 80 first-level b-maps. From this, we take a subsample of 10 to 30 subjects (the held in data). On these data we perform an OLS group analysis and define peaks based on the resulting t image. We transform the peaks to Z statistics and apply an excursion threshold u. Using these peaks, we apply the estimation procedure as described above. On the remaining subjects (the held out data), we also perform an OLS

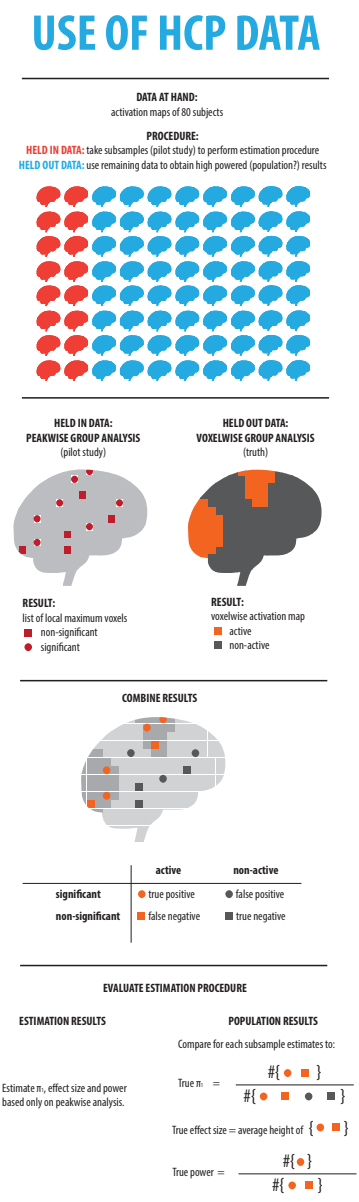


Figure 4.1: Overview of the procedure used to evaluate power calculations on the HCP-data

group analyses, but once we obtain the t image, we threshold voxelwise with false discovery rate control (Benjamini Hochberg procedure) and continue with voxelwise results.

To validate the estimation procedure, we combine the held-in data and the held-out data as follows: when a peak from the held-in data corresponds to a significant voxel in the powerful held-out data, we consider this as an *active* peak. When a peak corresponds to a non-significant voxel in the held-out data, it is considered *inactive*. True π_1 is then defined as the ratio of the number of activated peaks and the total number of peaks. The true effect size is the average peak height of all peaks that are located in activated area. Finally, true power is defined as the ratio of the number of significant peaks (for given thresholding procedure) and the total number of activated peaks. The true power is computed for subsamples between 10 and 30 subjects. As such the power predictions on small sample sizes can be compared to true power on larger samples.

This resampling procedure is repeated 100 times. For the evaluation of the procedure, the power and the estimates are averaged over all 100 resamples.

Some tweaking however is needed to use these real data for evaluation purposes. While we assume that the analysis of the held-out data is powerful enough to represent results on a population level, it is well known that fMRI data suffer from some artefacts that need to be taken into account. First of all, the null distribution of statistics within the t image are more dispersed than under the expected t distribution. This affects the voxelwise analysis of the held-out data, in which we use the false discovery rate. Therefore, instead of thresholding the maps based on the observed t image, we first estimate the null distribution for the held-out data using the empirical null procedure of Efron (2004). Using a maximal likelihood estimator, the mean (μ_0) and the variance (σ_0) of the null distribution are estimated. Next, all maps (both held-in and held-out) are normalized as $(T - \mu_0)/\sigma_0$, with the estimates of this empirical null distribution before continuing with the voxelwise statistical analysis.

Another problem arises from the fact that we assume to have *knowledge* on the true activation level based on the results for the held out data while these data were analyzed using the false discovery rate control. Therefore, the set of voxels (and consequently peaks) that we call *acti-*

vated will be contaminated with about 5% false positives. On the other hand, within the voxels (and consequently peaks) that appear *inactive*, there might be false negatives. This poses a problem for our measures of true π_1, true effect size δ and true power. Table 4.1 states the problem and introduces some notation.

Estimated Null (P_N)		Estimated Alternative (P_A)	
Truly Null	Truly Alternative	Truly Null	Truly Alternative
P_N^0	P_N^1	P_A^0	P_A^1
$\pi_0^N P_N$	$(1-\pi_0^N)P_N$	$\pi_0^A P_A$	$(1-\pi_0^A)P_A$

Table 4.1: Table to clarify the problem of the miscalculation of the true π_0.

We define P_N as the number of peaks that we estimate to be non-active (null) and P_A the number of peaks that we estimate to be active (alternative). In the set of peaks that are labeled inactive, P_N^0 peaks are truly null, while P_N^1 peaks are actually active. We can estimate the proportion of null peaks in the estimated null distribution π_0^N using the model of Pounds & Morris (2003), as such we can estimate P_N^0 and P_N^1. In a similar way, for the peaks that are labeled active, we define P_A^0 as the peaks that are null and P_A^1 as the peaks that are alternative. Here as well we define the proportion of null peaks in the estimated alternative distribution π_0^A.

Now we can define π_0^c as the proportion of nonactive peaks, corrected for the errors in what we estimate to be null and alternative:

$$\pi_0^c = \frac{\pi_0^N P_N + \pi_0^A P_A}{N} \tag{4.5}$$

with N the total number of peaks.

Likewise, we want to correct for misestimating the true effect size δ. The problem is stated in table 4.2. For the set of peaks that are labeled as inactive, E_N^0 denotes the effect size of the true null effects and E_N^1 the effect size of the peaks that are actually activated. For the set of peaks that are labeled as active, E_A^0 denotes the effect size of the effects of the peaks that are actually inactive while E_A^1 represents the effect size of the true alternative effects. From Random Field Theory, we know that the

effect size under the null distribution is $u + 1/u$, where u is the cluster forming threshold. We compute E_N as the average peak height in the null distribution and E_A as the average peak height in the alternative distribution. As such we can compute E_N^1 and E_A^1 as follows:

$$E_N^1 = \frac{E_N - \pi_0^N(u + 1/u)}{(1 - \pi_0^N)} \tag{4.6}$$

$$E_A^1 = \frac{E_A - \pi_0^A(u + 1/u)}{(1 - \pi_0^A)} \tag{4.7}$$

This enables to compute the effect size E^c corrected for the computation error as a weighted average of E_N^1 and E_A^1 with the weights according to the number of peaks in the estimated null and alternative distribution.

$$E^c = \frac{P_N^1 E_N^1 + P_A^1 E_A^1}{P_N^1 + P_A^1} \tag{4.8}$$

Estimated Null (E_N)		Estimated Alternative (E_A)	
Truly Null	Truly Alternative	Truly Null	Truly Alternative
$E_N^0 = u + 1/u$	E_N^1	$E_A^0 = u + 1/u$	E_A^1
$E_N = \pi_0^N(u + 1/u) + (1 - \pi_0^N)E_N^1$		$E_A = \pi_0^A(u + 1/u) + (1 - \pi_0^A)E_A^1$	

Table 4.2: Table to clarify the problem of the miscalculation of the true effect size.

4.3 Results

4.3.1 Simulations

Results of the estimation procedure for the prevalence of activation π_1 are shown in Figure 4.2. It can be seen that for all sample sizes, the prevalence of activation is overestimated. Biases range from 5 to 10%, with a small variance on the estimates. The prevalence of activation on the other hand is estimated with a small bias, the estimation of the expected peak height is very accurate with no bias and a very small variance on the estimation, which can be seen in Figure 4.3.

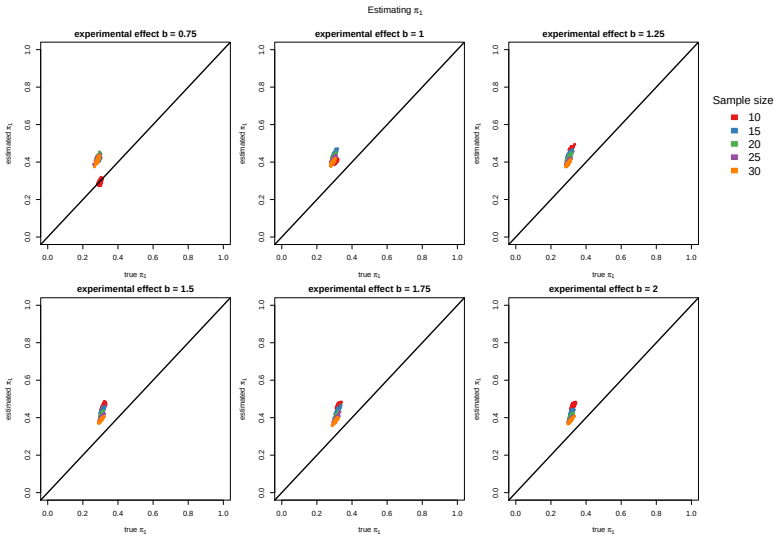


Figure 4.2: Plot of estimated π_1 against true π_1 for different sample sizes and different values for μ_1.

Figure 4.4 shows that not only the effect magnitude estimation but also the estimation of the complete distribution also matches the true

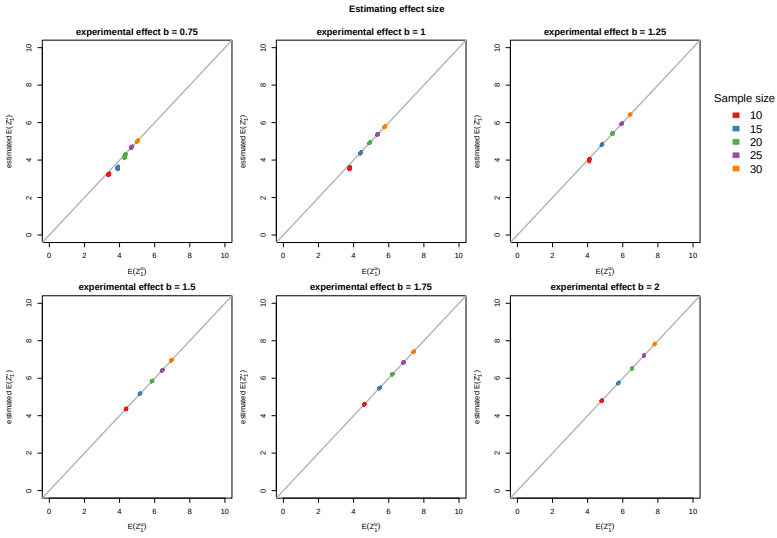


Figure 4.3: Plot of estimated expected peak height against true expected peak height for different effect sizes.

underlying distribution. More in particular, it can be seen that there is a good overlap of the average estimated alternative distribution and the observed alternative distribution. The difference in height of the distribution is due to the bias in the estimation of the prevalence of activation π_1.

Finally, when looking at the results for the power estimation in Figure 4.5, it can be seen that we obtain good estimates for power, not only in the given sample size, but also predictive for larger sample sizes. In low effect sizes, where only uncorrected thresholding ($p < .05$ and $p < .001$) leads to significant results (power > 0), the estimation is somewhat biased. The reason why the procedure is not able to estimate power for FDR thresholding in larger sample sizes, is that in the pilot study no significant effects are found. Therefore, there is no cutoff, and as such power cannot be estimated in larger sample sizes. For effect sizes, uncorrected thresholding easily results in a power of 100%, but this is not the case for corrected thresholding procedures for which power estimations are very accurate.

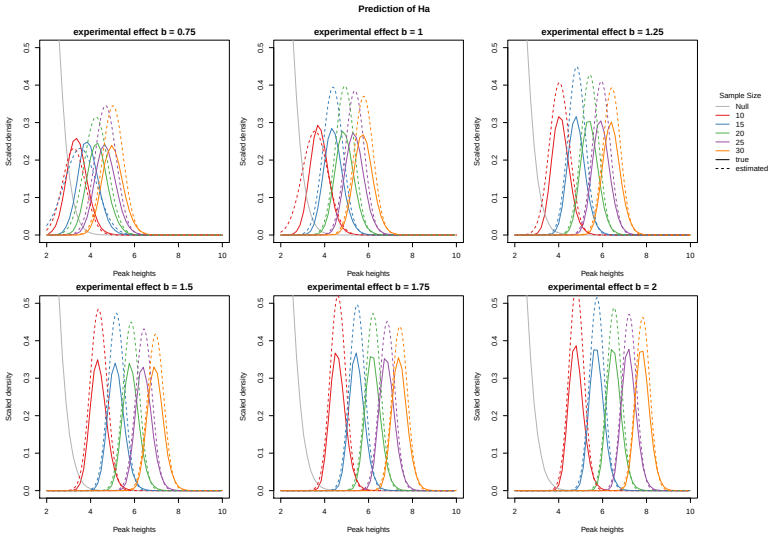


Figure 4.4: Plot of estimated and observed peak height distributions for different effect sizes. The solid line represents the truly observed distribution, the dashed line is the estimated distribution. Different colours refer to different sample sizes.

One assumption for our method is that π_1 remains constant for different sample sizes and that the number of peaks is constant for different sample sizes. It can be seen in Figure 4.6 and 4.7 that when the peak forming threshold u is held constant at 2, the emission of peaks is constant for different sample sizes, as well for non-active and active regions. The prevalence of activation π_1 is also constant for different sample sizes for a fixed u.

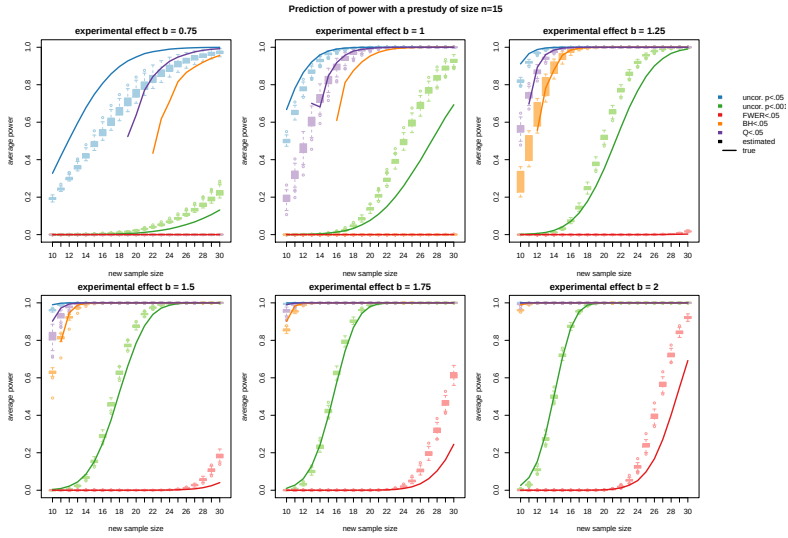


Figure 4.5: Plot of estimated and observed power for different effect sizes and different testing procedures. The power is estimated from a pilot study with 15 subjects.

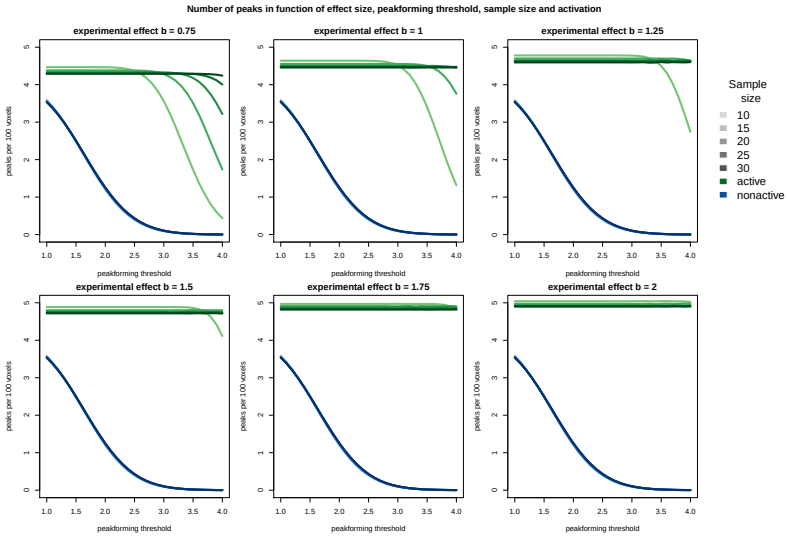


Figure 4.6: Plot of number of peaks per 100 voxels for different effect sizes, sample sizes and peak forming thresholds.

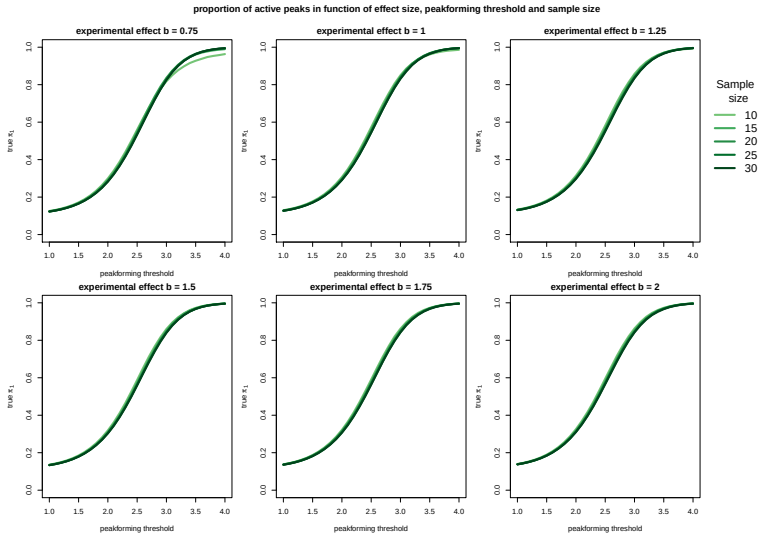


Figure 4.7: Plot of the true π_1 for different effect sizes, sample sizes and peak forming thresholds.

4.3.2 Example

After the second level analysis, we find 253 peaks. We compute p-values based on the exponential null distribution (see Equation 4.3). Using the method of Pounds & Morris (2003), we find an estimate for the prevalence of activation $\pi_1 = 0.55$. The estimated distributions of p-values can be seen in the left panel of Figure 4.8. Once we obtain this estimate, we proceed with estimating the distribution of the peak heights under the alternative hypothesis. We estimate that the distribution of active peaks is $N(4.24, 2.5)$ truncated at $u = 2$ which corresponds to an effect size δ of 1.18. The results can be seen in the right panel of Figure 4.8.

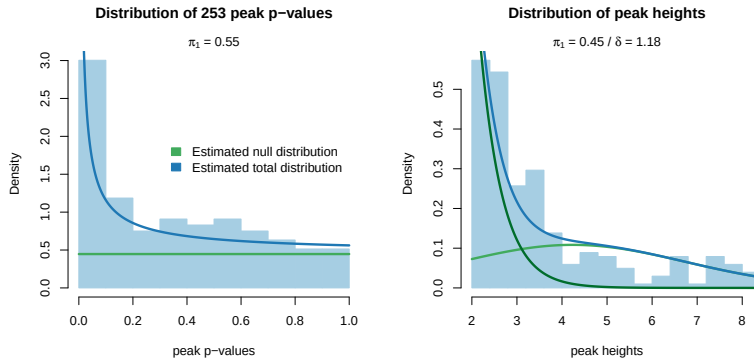


Figure 4.8: Left: Estimated distribution of peak p-values. The histogram of peak p-values is shown in light blue, the lines show the estimated part of the histogram stemming from the null distribution (green) and the total distribution (blue). Right: Estimated distribution of peak heights. The histogram of the peak heights is shown in light blue, the lines show the estimated distributions for the null (dark green), the alternative (light green) and the total distribution (blue)

The results of the power estimation procedure can be found in Figure 4.9. It can be seen that in the current study, with false discovery rate control with the method of (Storey, 2002), further denoted as Q, the

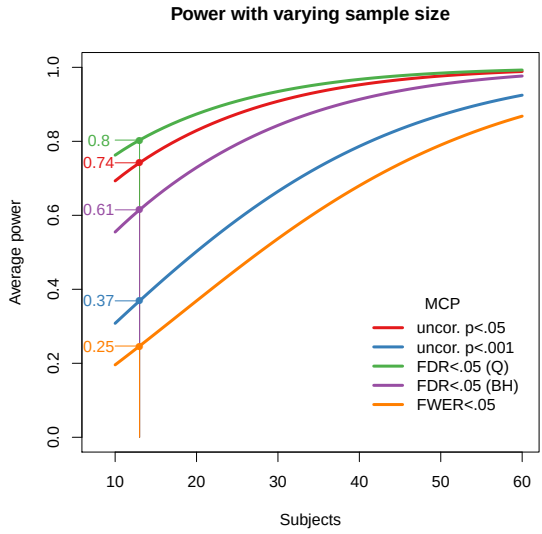


Figure 4.9: The figure shows the estimated power for different multiple testing procedures in function of sample size. The big dot represents the current study as it was performed, on the left hand side is the power in the current study printed.

highest power is obtained (80 %). A false discovery rate control with the method of Benjamini & Hochberg (1995), further denoted as BH, results in a power of 61%. Uncorrected testing at $\alpha = 0.05$ leads to a power of 74%, with $\alpha = 0.001$ the power will be 37%. Lastly, the power with familywise error rate control at 0.05 is the lowest, and equals 25%.

4.3.3 HCP data

Results of the estimation procedure for the prevalence of activation π_1 and the effect size are shown in Figure 4.10. It can be seen that the prevalence is overestimated in all cases while the effect size δ is underestimated. When looking at the estimation of the power (Figure 4.11), we see that the power is mostly overestimated for the low powers, which can be explained by the underestimation of the effect size. For larger sample sizes (and consequently larger powers) we obtain good predictions for the power based on a pilot study with 15 subjects.

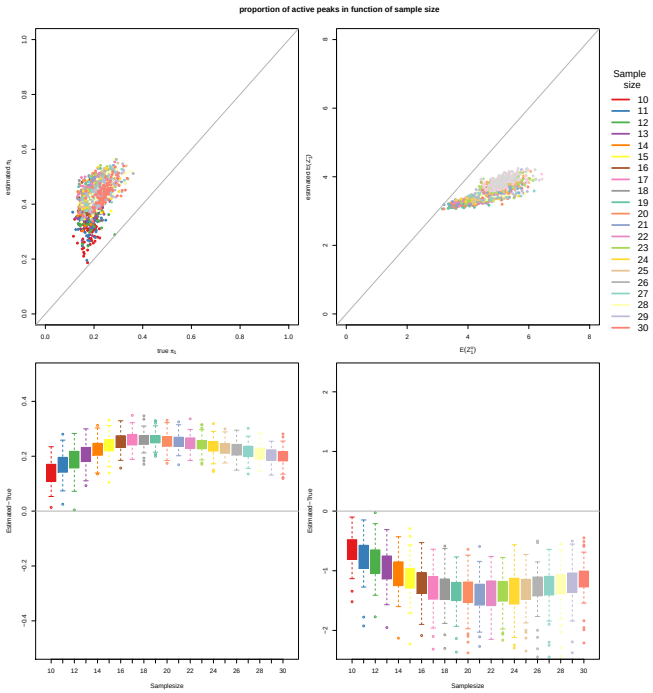


Figure 4.10: The upper panels show an evaluation of the estimation of π_1 (left) and the effect size (right). The dots refer to the different subsamples taken from the data. The lower panels show the bias to these estimates.

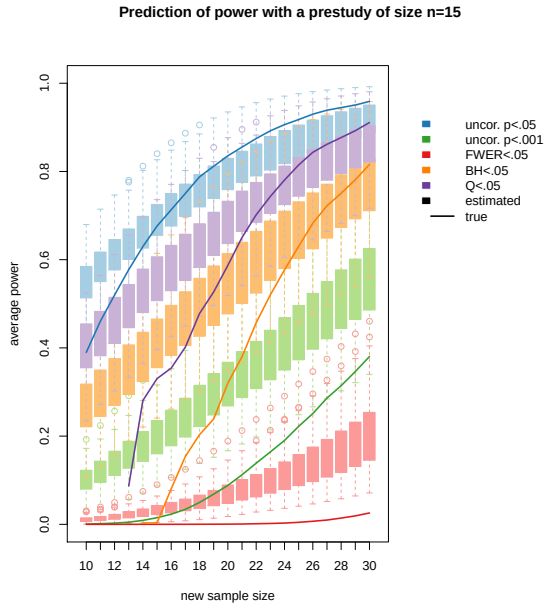


Figure 4.11: Plot of the estimated power from a pilot study with 15 subjects for different multiple testing procedures.

4.4 Discussion

In neuroscience, scientific results are often based on fMRI studies that suffer from a lack of power. In order to safe costs and effort, while preserving sufficient power for detecting important effects, we presented a method to predict power for different sample sizes. While other methods for power calculations in fMRI often require the estimation of many different parameters that are often difficult to estimate (Mumford & Nichols, 2008; Desmond & Glover, 2002), our method is based on only peak values that require Random Field Theory assumptions for the computation of p-values.

We focus on peaks for several reasons. We have shown in previous work that the assumption of a uniform distribution of the p-values under the null is mainly attained with peak p-values, but not with cluster p-values (Durnez, Moerkerke, & Nichols, 2014). As this is an assumption crucial for the procedure presented here, we opt for peak inference rather than cluster inference. Moreover, problems with localisation and stability have been reported with cluster inference (Roels, Bossier, Loeys, & Moerkerke, 2014; Woo, Krishnan, & Wager, 2014). However, when a user wants to infer power for cluster inference, this procedure on peaks can be used as a lower bound, as the power of cluster inference should be generally higher than peak inference (Friston et al., 2007). We do not choose to apply the method for voxelwise testing as π_1 will be very close to 0 in that case which makes its estimation hard if not impossible. On the other hand, it is possible to use the method if the prevalence of activation is widespread in the brain. It is up to the researcher to decide whether this shall be the call in the new study.

We have evaluated the procedure using simulated data. The data represent a simplified fMRI experiment, but we still vary a number of parameters, like the effect size, the thresholding procedure,... to ensure that the findings are generalisable to a range of different possible fMRI experiments. In our simulations, we have used a constant effect size of activation over different subjects. We have not applied subject-specific effect sizes, as we believe this would not alter the average effect size, but it would only inflate the between subjects variance, which will lead to a smaller normalised effect size. We believe it is not necessary to both vary the average

effect size, and the between subjects variance.

This method is only a first step in developing a means to obtain more powerful fMRI studies. Many different extensions are possible. One of these possibilities is the development of a testing procedure that would allow to use the pilot data in the final study without harming the false positive rate. Another possibility is to use the pilot data not only for a power calculation, but also for a variance optimisation procedure, and thereby improving power not only by increasing the sample size, but also by minimising the variance on the results. Thirdly, the estimated effect size could, besides sample size, incorporate other characteristics, that allow the optimisation of future studies without the restriction that all characteristics are identical to the pilot study.

Although the evaluation on this method was performed on whole-brain analyses, it is also possible to only apply it to a certain part of the brain, when a region of interest is specified.

We have made the procedure available to the community in an online toolbox.

Acknowledgements

The computational resources (STEVIN Supercomputer Infrastructure) and services used in this work were kindly provided by Ghent University, the Flemish Supercomputer Center (VSC), the Hercules Foundation and the Flemish Government - department EWI.

5 Alternative based thresholding with application to pre-surgical fMRI.[1]

Abstract Functional Magnetic Reasonance Imaging (fMRI) plays an important role in pre-surgical planning for patients with resectable brain lesions such as tumors. With appropriately designed tasks, the results of fMRI studies can guide resection, thereby preserving vital brain tissue. The mass univariate approach to fMRI data analysis consists of performing a statistical test in each voxel, which is used to classify voxels either as active or inactive, i.e. related, or not, to the task of interest. In cognitive neuroscience, the focus is on controlling the rate of false positives while accounting for the severe multiple testing problem of searching the brain for activations. However, stringent control of false positives is accompanied by a risk of false negatives which can be detrimental, particularly in clinical settings where false negatives may lead to surgical resection of vital brain tissue. Consequently, for clinical applications we argue for a testing procedure with a stronger focus on preventing false negatives. We present a thresholding procedure that incorporates information on false positives and false negatives. We combine 2 measures of significance for each voxel: a classical p-value which reflects evidence against the null hypothesis of no activation and an alternative p-value which reflects evidence against activation of a pre-specified size. This results in a layered statistical map for the brain. One layer marks voxels exhibiting strong evidence against the traditional null hypothesis, while a second layer marks voxels where activation cannot be confidently excluded. The third layer marks voxels where the presence of activation can be rejected.

5.1 Introduction

A common treatment for patients suffering from a brain tumor is surgical resection of the tumor. In order to minimize the risk of resecting brain tissue involved in essential brain functions, such as speech or language com-

[1] This chapter is based on:
Durnez, J., Moerkerke, B., Bartsch, A., Nichols, T.E. (2013). Alternative-based thresholding with application to pre surgical fMRI. *Cognitive, Affective & Behavioral Neuroscience*, 13 (4), 703–713.

prehension, these patients often undergo pre-surgical functional Magnetic Resonance Imaging (fMRI). This is a technique that shows subject-specific neural activity changes in the brain. The resulting fMRI data can assist the surgeon in performing the tumor resection while preserving the brain tissue involved in important cognitive and sensorimotor functions (Bartsch, Homola, Biller, Solymosi, & Bendszus, 2006), and can even be used to predict the outcome of post-operative cognitive functioning (Richardson et al., 2004).

To analyse fMRI data, a huge number of statistical tests are performed simultaneously. In cognitive neuroscience, this technique is used to link neurological and neuropsychological functions with their respective location in the brain, supporting different theories of brain function. To be confident that a brain area is associated with a task it is essential to account for the multiple testing problem. This can be done using corrections for either the familywise error rate (Worsley et al., 1996; Friston et al., 1991) or the false discovery rate (Genovese et al., 2002). These multiple testing corrections result in a more stringent control of the null hypothesis of no activation, and consequently, the probability of a false negative increases (Logan & Rowe, 2004; Lieberman & Cunningham, 2009). In cognitive neuroscience, a false positive means fallacious support for a given cognitive theory. While false positives can often be discovered by unsuccessfully trying to replicate the study, much time, effort and money can be expended. As a result the scientific discipline generally deems stringent control of false positives necessary, accepting the concomitant sacrifices in sensitivity.

In a clinical setting such as pre-surgical fMRI however, a loss in power means true activation is not discovered, and this might result in the resection of vital brain tissue (Haller & Bartsch, 2009). Inversely, false positives have a less negative impact on the surgical result (Gorgolewski et al., 2012). The goal of classical hypothesis testing is to prevent the null hypothesis from being rejected, by only considering voxels as being active when enough evidence against the null of no activation is found. This asymmetrical way of penalising errors in statistical inference is undesirable in this context (Johnson et al., 2012), and instead the focus should be on protecting the alternative hypothesis: one only wants to exclude activation when enough evidence against activation is found. We therefore

present a new hypothesis thresholding procedure that incorporates both information on false positives and false negatives and thus is ideally suited for pre-surgical fMRI.

In classical hypothesis testing, the evidence against null hypothesis is measured with the p-value, the null hypothesis probability of data as or more extreme than that observed. Thresholding a p-value at α produces a statistical test that controls of false positive rate at α. To allow direct control of false negative risk, we present a symmetrical measure which quantifies evidence against the alternative hypothesis (Moerkerke, Goetghebeur, De Riek, & Roldan-Ruiz, 2006). Correspondingly, thresholding this probability measure at β ensures control of the false negative rate at β.

By combining thresholds on the classical and alternative p-value, we use information on the probability of false positives and false negatives. We show that thresholding both error measurements results in a layered statistical map for the brain, each layer marking voxels with evidence (or lack there of) against the null and/or alternative hypothesis. One layer consists of voxels exhibiting strong evidence against the null of no activation, while a second layer is formed by voxels for which activation cannot be confidently excluded. The third level then consists of voxels for which the presence of activation can be rejected.

fMRI data can be analyzed in different ways. The most popular method is a confirmatory mass-univariate GLM-analysis, where the measured time series in each voxel is regressed onto the design of the experiment, resulting in an estimate of the effect, for which a T-statistic with corresponding classical p-value can be computed for each voxel. This method has shown to be very effective and robust, but its downside is the mass-univariate character. While many attempts have been made to take into account the spatial character of the data with data smoothing and peak- and cluster-thresholding, the GLM fails to recognize patterns of activation or noise. In this light, statistical techniques for multivariate data have been successfully applied to fMRI data. Independent Component Analysis (ICA, Beckmann & Smith, 2004) is an exploratory method used to find hidden source signals, modelling the observed data as a (unobserved) linear mixture of (unobserved) sources. ICA therefore allows to discover spatially and temporally structured noise. Given the popularity of the GLM and

the upcoming interest for ICA, especially in a clinical context, we will introduce the thresholding procedure for both techniques. We show how the ideas can be translated to different statistical techniques.

In section 2, we introduce and combine quantities to measure significance when testing for activation. To this end, we start with a simple setting in which test statistics are assumed to be Gaussian distributed and take the general form of the ratio of an observed effect and its standard error. These settings directly translate to the case of univariate linear modeling that make use of T-distributions. We further demonstrate how to use the principle for independent component analysis (ICA). In section 3, we present the results of the procedure applied to pre-surgical fMRI data.

5.2 Methods

5.2.1 Measures of evidence against the null and alternative

At each voxel i, $i = 1, \ldots, I$, we assume that a linear model is fit and produces $\widehat{\Delta}_i$, an unbiased estimate of the BOLD effect of interest Δ_i, and an estimate of the standard deviation of $\widehat{\Delta}_i$, its "standard error" $\mathrm{SE}(\widehat{\Delta}_i)$. We henceforth suppress the voxel subscript unless needed for clarity. We assume that the degrees-of-freedom are sufficiently large so that $\mathrm{SE}(\widehat{\Delta})$ has neglible variability, as is the case for fMRI time series. We further assume that the data, model and contrast has been scaled appropriately so that $\widehat{\Delta}$ has units of percent BOLD change (or, at least approximately, as when global brain intensity is scaled to 100[2]).

The null and the alternative hypothesis The null hypothesis $H_0 : \Delta = 0$ states that the true effect magnitude is zero, and an underlying difference between conditions Δ is equal to 0. Classical statistical inference involves computing a test statistic, converted to a p-value, that measures the evidence against this null hypothesis. The decision procedure to reject H_0 is calibrated to maintain the Type I error at α. However, failing to reject

[2]Note that, as of SPM8, the global brain intensity after intensity normalisation, is scaled to 200 or greater.

H_0 does not allow one to conclude that H_0 is true. The reason is that the probability calculation of the p-value is based on the assumption that the null hypothesis is true. It is a logical fallacy, "affirming the consequent" or "reasoning to a forgone conclusion," to begin by assuming something and then, eventually, conclude that the initial thing is true. More concretely, when we fail to reject H_0 it could simply be because there are only subtle deviations from H_0 that are not detected or because the precision on the observed effect is too small to reach statistical significance. Scientists frequently make this mistake and there have been various guidelines on reporting study results (see e.g. Meehl 1978; Schmidt & Hunter 2002) which all stress the importance of complementing p-values with effect sizes.

Our procedure considers an "alternative hypothesis" p-value, p_1, that measures the evidence against $H_a : \Delta = \Delta_1$, the non-zero effect magnitude expected under activation. fMRI-studies are often preceeded with power analyses for sample size calculations which also require the specification of Δ_1. In literature, different approaches to choose a meaningful Δ_1 have been presented (Desmond & Glover, 2002; Hayasaka et al., 2007; Mumford & Nichols, 2008; Zarahn & Slifstein, 2001). Alternatively, in pre-surgical fMRI, one can estimate Δ_1 based on data in previous patients.

Measures of significance At a given voxel we have a test statistic T with observed value

$$t = \frac{\widehat{\Delta}}{\mathrm{SE}(\widehat{\Delta})}. \tag{5.1}$$

We assume that T has a known distribution under H_0 (e.g. Student's t with given degrees-of-freedom, or Gaussian), so that we can compute the classical p-value

$$p_0 = P(T \geq t|H_0). \tag{5.2}$$

That is, p_0 quantifies the evidence against the null hypothesis H_0 of no task-related activation.

In a symmetrical fashion, the alternative p-value is defined as in Moerkerke et al. (2006):

$$p_1 = P(T \leq t|H_a). \tag{5.3}$$

Correspondingly, p_1 measures the evidence against H_a, and corresponds to the classical p-value for testing a "null" H_1 versus "alternative" H_0. In

generally, as the evidence in favor of H_1 grows, p_0 becomes smaller and p_1 becomes larger.

In order to compute p_1 we need the distribution of T under H_a, which requires specification of Δ_1. However, we don't expect a single magnitude of true activation, but a distribution of different true values (Desmond & Glover, 2002). Therefore, in a Bayesian spirit, we specify a distribution of likely values of Δ_1 instead of fixed value:

$$\Delta_1 \sim \mathcal{N}\left(\mu, \tau^2\right) \tag{5.4}$$

where μ is the expected magnitude of effect under true activation while acknowledging variation among voxels, specifically Gaussian variation with standard deviation τ.

Assuming that T also follows a Gaussian distribution, it has the following distribution under H_a at voxel i:

$$T_i \sim \mathcal{N}\left(\frac{\mu}{\mathrm{SE}(\widehat{\Delta}_i)}, \frac{\mathrm{SE}(\widehat{\Delta}_i)^2 + \tau^2}{\mathrm{SE}(\widehat{\Delta}_i)^2}, \right) \tag{5.5}$$

where voxel subscripts are used to emphasize that the values of μ and τ are *fixed* for the entire brain, and based on prior knowledge or other experiments, while $\mathrm{SE}(\widehat{\Delta}_i)$ is from each individual voxel. With this distribution we can compute p_1 at each voxel. An illustration of both measures of significance can be seen in Figure 5.1. As the alternative distribution depends on the voxel-specific standard error, the distance between the null and alternative distributions will be voxel-specific. In particular a large standard error results in a large overlap between H_0 and H_a, while small standard errors lead to a large distance and little overlap between H_0 and H_a.

5.2.2 Combining measures of significance

In classical null hypothesis significance testing, a threshold α on p_0 can be translated into a threshold t_α for the test statistic in equation (5.1). In parallel, a threshold β on p_1 can be translated into a test statistic threshold t_β. While t_α is determined by α (and degrees-of-freedom if not using a Gaussian), t_β further depends on β, μ, τ and $\mathrm{SE}(\widehat{\Delta}_i)$. Thus t_β

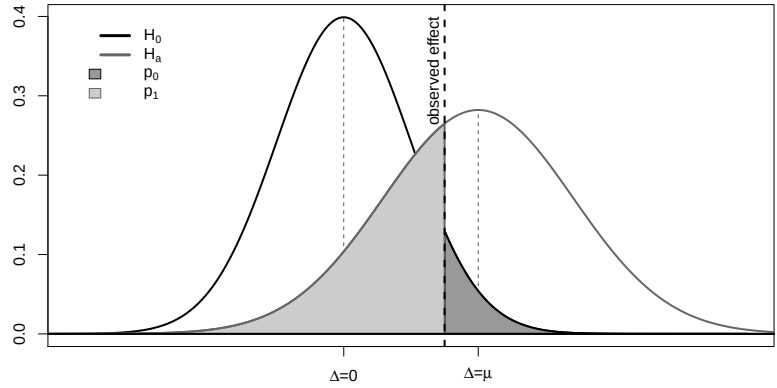


Figure 5.1: The distributions of an effect under H_0 and H_a are displayed for an observed effect of $t = 1.5$, $\text{SE}(\widehat{\Delta}) = 1$, $\Delta_1 = 2$ and $\tau = 1$. Note that H_a has a wider distribution than H_0 due to the uncertainty on Δ_1.

varies over the brain depending on the (estimated) standard error.

Figure 5.2 shows the possible results of this testing procedure, with α and β relatively small. In what is expected to be the typical scenario, with standard error that is large relative to the true effect magnitude, $t_\beta < t_\alpha$ and three possible outcomes can be distinguished.

One outcome is when voxels exhibit evidence against H_0 and at the same time are consistent with H_a ($p_0 < \alpha$ and $p_1 > \beta$; red in Figure 5.2). This is the most compelling case for the presence of true activation ($\Delta > 0$). The opposite outcome is a large p_0 and a small p_1 ($p_0 > \alpha$ and $p_1 < \beta$; grey in figure 5.2). Here the data are consistent with the null and there is evidence to reject the alternative, and is the most compelling case for true absence of activation ($\Delta = 0$). The third outcome is when the data are compatible with both the null and alternative, and neither can be excluded ($p_0 > \alpha$ and $p_1 > \beta$; yellow in Figure 5.2).

A less frequent, albeit possible scenario appears when the standard error is small relative to the true effect magnitude, $t_\alpha < t_\beta$ and H_0 and H_a can be clearly distinguished. Voxels with no effect or strong effects will be identified as before ($p_0 > \alpha$ and $p_1 < \beta$, no activation; $p_0 < \alpha$ and $p_1 > \beta$, activation). However, for certain data there is both evidence

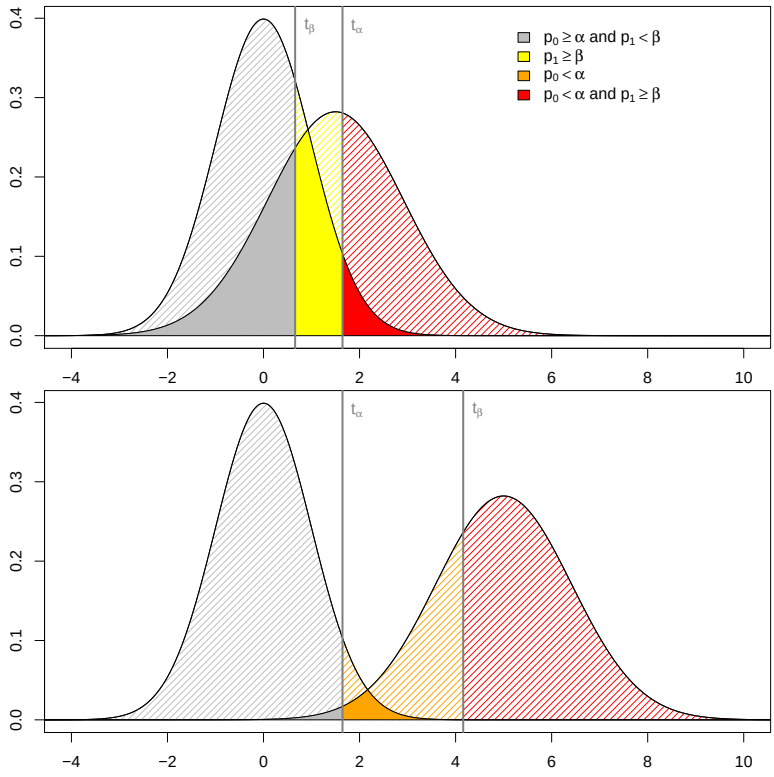


Figure 5.2: When thresholding p_0 and p_1 at significance levels α and β, two possibilities arise: $t_\beta < t_\alpha$ (upper panel) or $t_\alpha < t_\beta$ (lower panel).

against H_0 as well as against H_a ($p_0 < \alpha$ and $p_1 < \beta$; orange in Figure 5.2). It indicates a case where the effect is so small as to lack *practical significance*.

For pre-surgical fMRI, this procedure provides information on which areas are confidently safe to be resected (grey areas), which area's should absolutely be avoided when resecting brain tissue (red areas) and in which area's the surgeon should take care because neither hypotheses can be rejected (yellow areas). When the fourth type of voxel is found, meaning both hypotheses can be rejected (orange areas), an abundance of caution suggests that again care be taken, since rejection of H_0 does suggest some association with the task, just at a possibly very small magnitude. The specific application to real data is shown in the section 3.

5.2.3 Alternative Thresholding of Independent Component Analysis (ICA)

Above we described the classical and alternative p-value for a traditional setting, where a test statistic is the ratio of an observed effect and its standard error as is the case for T-statistics when using the GLM. Here we demonstrate that the technique is also applicable in more general settings, in particular with maps from Independent Component Analysis. Exact implimentation details of ICA methods differ; our development here follows the FSL[3] software's implementation, MELODIC (Beckmann & Smith, 2004), but should be readily applicable to other ICA software.

Independent Component Analysis (ICA) is a technique for multivariate data-driven analysis of fMRI-data. It does not require the specification of the experimental design and produces spatio-temporal patterns that explain the variability in the data. ICA transforms the four-dimensional fMRI data into K pairs of spatial and temporal modes. Each spatial mode, or Independent Component (IC) image, is associated with one IC time series. The variation explained by each component is the IC time series scaled by the weights at each voxel in the IC image; equivalently, it is the spatial pattern in an IC image scaled by each value of the IC time series. Stated simply, the weights represent the association between the temporal activation pattern observed in the voxel and the temporal pattern in the

[3] `http://www.fmrib.ox.ac.uk/fsl`

K different components.

Let $\boldsymbol{Y}$ represent the $J \times I$ data matrix, where J is the number of time points and I is the number of voxels. We assume that the data at each voxel have been mean-centered, that is, the column means of $\boldsymbol{Y}$ are zero. ICA decomposes the data as per

$$\boldsymbol{Y} \approx \boldsymbol{M}\, \boldsymbol{S}_1\, \boldsymbol{C}\, \boldsymbol{S}_2\, \boldsymbol{S}_0 \tag{5.6}$$

where $\boldsymbol{M}$ is a $J \times K$ matrix with one temporal mode in each column, and $\boldsymbol{C}$ is a $K \times I$ matrix with one spatial mode in each row; $\boldsymbol{S}_1$ ($K \times K$) is a diagonal scaling matrix that ensures that the temporal modes have unit variance, and $\boldsymbol{S}_0$ and $\boldsymbol{S}_2$ (both $I \times I$) are diagonal scaling matrices that ensure that background noise in the spatial modes have unit variance. See Appendix 5.A for detailed definitions of these scaling factors.

In the presentation and interpretation of ICA results, each of the K spatial modes in $\boldsymbol{C}$ are visualised and explored. Since they have been noise-normalised, they are often treated as z-score images and thresholded to control a nominal false positive rate. The end result is an inference that quantifies the relation between the corresponding temporal mode in $\boldsymbol{M}$ and $\boldsymbol{Y}$. We seek to apply our alternative hypothesis thresholding procedure to these maps, but first we need to define a meaningful effect size in percent BOLD change, and transform this to the scale of $\boldsymbol{C}$.

Meaningful BOLD effect sizes with ICA Consider a particular IC of interest, $k \in \{1, \ldots, K\}$, and a particular voxel $i \in \{1, \ldots, I\}$ of interest in the spatial mode. Specifically, consider the contribution of the k-th IC to the time series at voxel i:

$$\boldsymbol{m}_k\, s_{1,k}\, c_{ki}\, s_{2,i}\, s_{0,i} \tag{5.7}$$

where $\boldsymbol{m}_k$ is the k-th column of $\boldsymbol{M}$, $c_{ki} = (\boldsymbol{C})_{ki}$, and $s_{1,k}$, $s_{2,i}$, and $s_{0,i}$ are the indicated diagonal elements of the scaling matrices.

As previously mentioned, the rows of $\boldsymbol{C}$ are normalised to have noise variance of 1, so c_{ki} has z-score (and not BOLD data) units. We need to compute a meaningful percent BOLD change effect. We will first compute this for a fixed Δ_1, in the units of c_{ki}, and will later impose a distribution on the effect size. Equation (5.7) shows that the temporal variation from

IC k is determined by not only c_{ki} and the scaling factors, but $\boldsymbol{m}_k$ as well. But $\boldsymbol{m}_k$ is scaled to unit variance, and will not induce a unit BOLD change in the data. We propose scaling $\boldsymbol{m}_k$ so that it (roughly) expresses a unit BOLD effect, and as a result preserves the units of the other terms. Specifically, we introduce h_k

$$\boldsymbol{m}_k \, h_k h_k^{-1} \, s_{1,k} \; c_{ki} \, s_{2,i} \, s_{0,i}. \tag{5.8}$$

so that $\boldsymbol{m}_k \, h_k$ expresses a unit BOLD effect in the data. One way to set the factor h_k is so that $\boldsymbol{m}_k \, h_k$ has baseline to peak range of 1. Another way is to regress $\boldsymbol{m}_k$ on a covariate $\boldsymbol{d}$ that expresses the anticipated (unit) experimental effect; setting h_k to the inverse of the regression coefficient will ensure that $\boldsymbol{m}_k \, h_k$ corresponds to an approximate unit BOLD effect.

Finally, we correct for the attenuation of the hypothesized effect based on the mismatch between $\boldsymbol{m}_k$ and $\boldsymbol{d}$. That is, even if we choose h_k well, $\boldsymbol{m}_k \, h_k$ may only be weakly correlated with $\boldsymbol{d}$. As a result we scale the expected BOLD effect of IC k at voxel i by $\rho_{\boldsymbol{m}_k \boldsymbol{d}}$, the correlation between $\boldsymbol{m}_k$ and $\boldsymbol{d}$; see Appendix 5.B for details.

Now we can relate the expected (attenuated) percent BOLD change, $\rho_{\boldsymbol{m}_k \boldsymbol{d}} \, \Delta_1$, to the units of IC temporal mode. Let Δ_1^* be the expected alternative mean effect in the z-score statistic c_{ki}, then

$$\rho_{\boldsymbol{m}_k \boldsymbol{d}} \, \Delta_1 \approx h_k^{-1} \, s_{1,k} \; \Delta_1^* \, s_{2,i} \, s_{0,i} \tag{5.9}$$

and thus we can translate BOLD units into c_{ik} units with $\Delta_1^* \approx s_{ik}^* \; \Delta_1$, where

$$s_{ki}^* = \rho_{\boldsymbol{m}_k \boldsymbol{d}} \, h_k \, s_{1,k}^{-1} \, s_{2,i}^{-1} \, s_{0,i}^{-1}. \tag{5.10}$$

Finally, this implies that our distribution of alternative effects in c_{ki} units is

$$\Delta_1^* \sim \mathcal{N}\left(s_{ki}^* \, \mu, \; s_{ki}^{*2} \, \tau^2\right) \tag{5.11}$$

(c.f. Eqn. (5.4)).

Significance procedure As c_{ki} has unit noise variance, with an assumption of Gaussianity the null distribution is given by

$$c_{ki}|H_0 \sim \mathcal{N}(0,1)\,. \tag{5.12}$$

Under the alternative, we consider the addition of effect Δ_1^* to c_{ki} yielding the alternative distribution

$$c_{ki}|H_a \sim \mathcal{N}\left(s_{ki}^*\,\mu,\, 1 + s_{ki}^{*2}\,\tau^2\right). \tag{5.13}$$

5.2.4 Data

We consider data from a patient suffering from a left prefrontal brain tumor. The study design was a box-car design, where the patient was asked to alternate between recitation of tongue-twisters and quiescence. Figure 5.3 shows a sagittal slice of the T2 image, with the tumor visible in inferior prefrontal frontal cortex. For the application to mass univariate linear modeling, the data were analyzed wit FEAT in FSL 4.1 (Smith et al., 2004). The application to independent component analysis was performed using MELODIC in FSL 4.1 (Beckmann & Smith, 2004).

5.3 Results

5.3.1 Univariate linear modeling

We applied these techniques to the data described in Section 5.2.4. We derive the expected effect magnitude for Δ_1 and the variability of that effect τ from 5 patients who underwent the same fMRI paradigm. We threshold the image of each individual using an FDR-control at 0.05 and look at the average percent BOLD change units in each individual. The results are shown in table 5.1. Therefore we specify the expected effect magnitude for Δ_1 of $\mu = 0.73$ percent BOLD change units, and variability of that effect as $\tau = \sqrt{\widehat{\tau^2}} = 0.21$ percent BOLD change. These results are consistent with others in the literature (see e.g. Desmond & Glover (2002), Figure 7A)

Results are shown in Figure 5.4 with thresholds $\alpha = 0.001$ and $\beta = 0.20$. In other words, we specified a p_0 threshold for declaring an activation

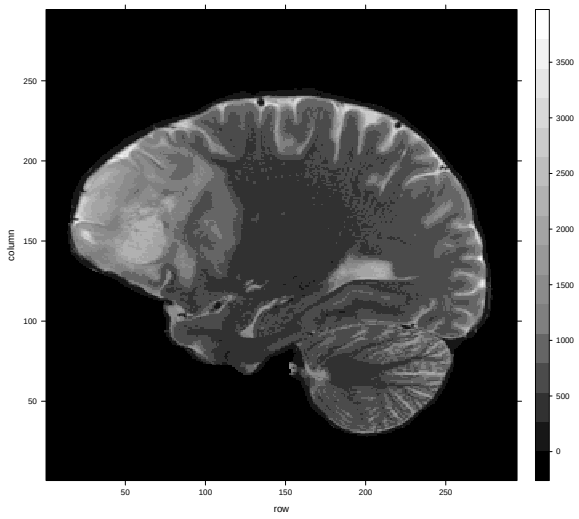


Figure 5.3: Anatomical scan of the patient. The tumor can be clearly seen in the prefrontal cortex.

	Average $\hat{\mu}$	Average $\hat{\tau}^2$
Patient 1	0.59	0.20
Patient 2	0.55	0.26
Patient 3	0.68	0.35
Patient 4	0.75	0.46
Patient 5	1.08	0.84
Average	0.73	0.43

Table 5.1: Average effect sizes in five previously tested patients in percent BOLD change units.

when there is none at 1-in-1000; and we set the p_1 threshold for declaring the absence of activation when in fact the specified activation magnitude is present at 1-in-5. The red and the (scant) orange voxels show where H_0 can be confidently rejected, and, if presurgical planning was done only on the basis of classical null hypothesis testing, all other tissue would be regarded as "safe". Considering information on the alternative, we have the red voxels where, specifically, H_0 can be rejected and H_a cannot be rejected; i.e. the red voxels are incompatible with the null and compatible with the alternative, and thus are strong evidence for the effect. The yellow areas are areas are where neither H_0 *nor* H_a can be rejected; here the data is compatible with both the null and alternative, and suggest a lack of confidence in ruling out activation. Finally, for voxels with no coloration, the H_0 cannot be rejected but H_a can; the data are compatible with the null and incompatible with the alternative, and thus have good evidence for a lack of activation and suggest that these brain regions can be safely resected. This shows the key strength of the procedure: among voxels traditionally classified as "nonactive", i.e. those with insufficiently small p_0's, it distinguishes between voxels where there is compelling evidence for non-activation (not colored) and those voxels where we cannot rule out the possibility of activation (yellow).

The orange voxels represent voxels for which the observed effect size is between the null hypothesis of no activation and the expected effect size. In these voxels, both the null and the alternative hypothesis are rejected which corresponds to very low residual noise in the GLM.

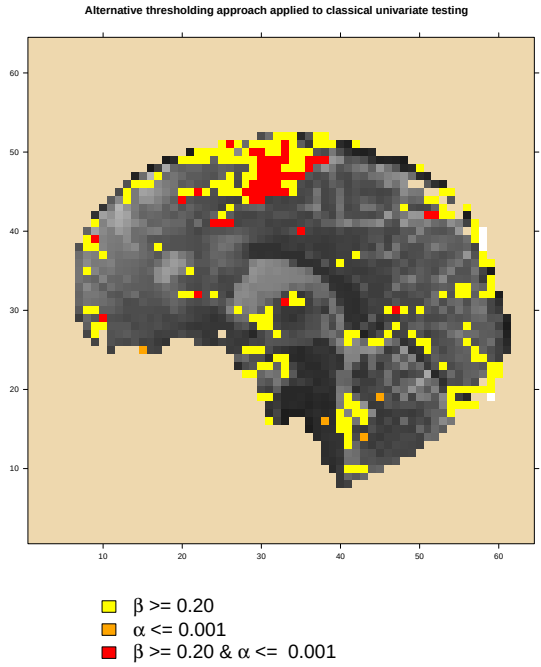


Figure 5.4: Sagittal slice of "layered" activation inference overlaying grayscale T2* reference image, threshold values of $\alpha = 0.001$ and $\beta = 0.20$. Red areas show areas of high confidence of activation (H_0 rejected, H_a not rejected), while yellow areas show areas where activation cannot be ruled out (neither H_0 nor H_a rejected); uncolored areas have high confidence of no activation (H_0 not rejected, H_a rejected), while the few orange voxels indicate voxels with significant but surprisingly small BOLD response magnitude (H_0 and H_a rejected).

5.3.2 Independent Components Analysis Results

We applied these techniques to the data described in Section 5.2.4. We use the same effect size and uncertainty as in Section 5.3.1, i.e. $\mu = 0.73$ and $\tau = 0.18$ percent BOLD change units. MELODIC's automated dimensionality estimation method in MELODIC found 52 components. We chose one IC whose time series corresponded to the design matrix, shown in Figure 5.5. Regressing this temporal mode on the design gives a coefficient of $\hat{\beta} = 1.48$, and thus $h = \hat{\beta}^{-1} = 0.677$ is the scaling factor used to have the temporal mode express a unit-BOLD effect (see Eqn. (5.8)). The pointwise correlation between the design and the chosen component is $\rho_{\boldsymbol{md}} = 0.63$, which is used to attenuate he expected effect magnitude (see Eqn. (5.9)).

The layered thresholding procedure for this IC is shown in Figure 5.6, for $\alpha = 0.001$ and $\beta = 0.20$. There is a set of voxels with strong evidence (red, H_0 rejected, H_a accepted) but also additional voxels where both hypotheses are rejected (orange). As mentioned above, in the setting of presurginal planning, these orange regions are best regarded as regions of possible activation, and thus excluded from resection.

This result is quite different from the GLM results, and is a reflection of the dramatically lower voxel-wise variance in the IC spatial mode relative to the GLM statistic image. The explanation is that the GLM result accounts for all noise variance, while the IC spatial map only reflects the noise in the subspace corresponding to the IC temporal mode (Beckmann & Smith, 2004).

Crucially, we stress that our thresholding procedure only seeks to improve the interpretability of the ICA result, and does not produce confirmatory inferences; IC selection is intrincally *post hoc* and subsequent inferences circular, and all we attempt to do here is improve the thresholding of a selected IC spatial map.

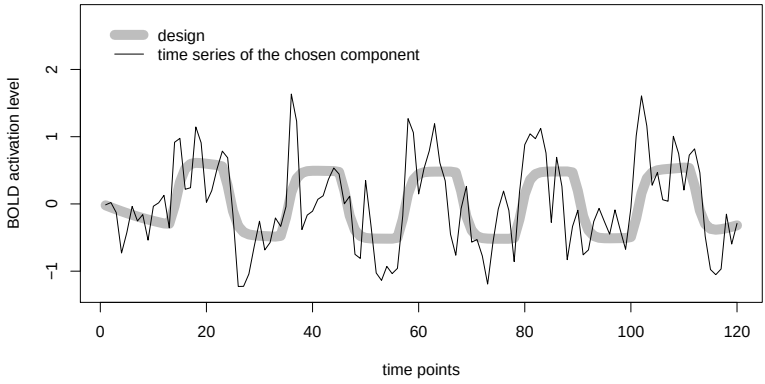


Figure 5.5: The time series of a selected IC, with a least squares fit of regressing the series on the design shown in gray. The estimated response height is used to normalise the component to have unit BOLD effect, and the pointwise correlation between the design and the selected IC is used to attenuate the expected BOLD response manitude.

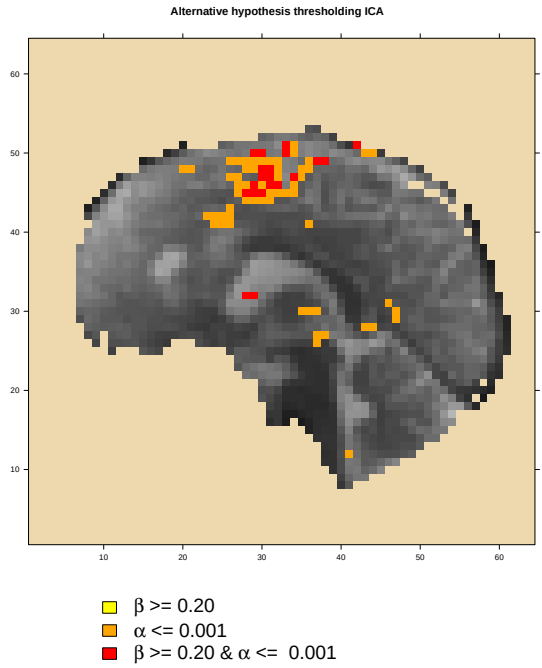


Figure 5.6: Results of the alternative thresholding procedure when using ICA. Sagittal slice of "layered" activation inference overlaying grayscale T2* reference image, threshold values of $\alpha = 0.001$ and $\beta = 0.20$. Red areas show areas of high confidence of activation (H_0 rejected, H_a not rejected), orange areas show voxels with significant but surprisingly small BOLD response magnitude (H_0 and H_a rejected); uncolored areas have high confidence of no activation (H_0 not rejected, H_a rejected).

5.4 Discussion

Statistical thresholding in the context of multiple tests is generally driven by the need to limit false positives. These stringent testing procedures in fMRI research leads to an abundance of false negatives (Lieberman & Cunningham, 2009) and are therefore less useful in the context of pre-surgical fMRI where a false negative can have dire consequences. While many attempts have been made to propose more liberal testing criteria for example by controlling the FDR instead of the FWER (Genovese et al., 2002), the focus is still on protecting the type I error rate. The unilateral focus on preventing false positives leads to a bias towards large obvious effects and against complex cognitive and affective effects (Lieberman & Cunningham, 2009). We therefore propose a measure that quantifies the evidence against the alternative hypothesis as introduced in Moerkerke et al. (2006). We use this quantity p_1 in addition with the classical p_0-value in a procedure that results in a thresholding procedure with multiple layers of significance. One layer consists of voxels exhibiting strong evidence of activation (red in Figure 5.4 & 5.6), while a another layer shows voxels with ambiguous evidence (yellow and orange), and a final layer then consists of voxels for which the presence of activation can be confidently rejected (an absense of overlaid statistic values). Thereby we offer a more symmetrical interest towards both false positives and false negatives.

We have chosen to focus on voxel-wise inference instead of other topological features, such as peaks (Chumbley et al., 2010) or clusters (Chumbley & Friston, 2009). These topological inference methods have reduced spatial specificity relative to voxel-wise inference, and are therefore less suitable for pre-surgical fMRI where maximal spatial precision is needed.

To use the procedure discribed in this paper, an expected effect size and its variance needs to be defined on a BOLD-scale. This is an arbitrary choice, however many possibilities are available. Desmond & Glover (2002) show for a specific experimental paradigm the distribution of percent signal change with its distribution. They show on average a BOLD effect size of 0.48 percent BOLD change. Another possibility to estimate the expected effect size can be based on previous research. As in pre-surgical fMRI, the same experiment is repeated over most patients, the effect size can be derived from patients that already underwent the experiment and

surgery. The degree to which brain activation in patients is representative for the particular setting for which estimates are needed, highly depends on the context and should be carefully judged. It can be expected that different methods will affect the estimates for μ and τ, however we found that the estimates we obtained by averaging over voxels is close to the effect sizes that can be found in literature.

The two different analytical approaches we used, the GLM and ICA, show somewhat different results. While both GLM and ICA analyses found similar sets of voxels that were confidently activated (H_0 rejected, H_a not), in the GLM analysis many voxels were found that do not show evidence against the null nor against the alternative (yellow, in Figure 5.4). The explanation for this outcome is the high level of noise present in the data and thus confusion about the veracity of either H_0 or H_a. In contrast, in the ICA analysis, almost no voxels have this ambiguity, and instead we find voxels that have evidence against both the null and the alternative. As ICA is a good tool to identify structured noise in a data-driven manner, it can be expected that the residual voxelwise variance will be smaller. Low variances results in a large distance between the null and the alternative distribution functions. Whereas the difference between ICA and the GLM seem contradictory at first, we argue that the differences in our approach reflect real differences between the two analysis tools.

The quantity p_1 shows a relationship with the voxel-based statistical power defined by Van Horn, Ellmore, Esposito, & Berman (1998). The voxelwise power by Van Horn et al. (1998) translates to the complement of the alternative p-value, p_1, in our study. However, the use of the quantity is fundamentally different. Whereas Van Horn et al. (1998) use the voxelwise power to visualise and interpret the results of a certain study, we explicitly threshold the quantity. Moreover, the interpretation of both quantities is not so straightforward. When a high power is encountered in a certain voxel, with the method of Van Horn et al. (1998), it is interpreted as follows: "If the observed effect in the voxel would be used as cutoff when testing from H_0, we have a high probability to reject H_0 when H_0 is indeed false". However, a large voxelwise power translates to a small p_1 and is in our study interpreted as follows: "When the alternative hypothesis is true, there is a small probability of observing this effect", and we will

interpret this effect as evidence against the alternative hypothesis. This interpretation is much more straightforward and usable.

This procedure has been developed in light of pre-surgical fMRI, as false negatives can have harmful consequences for the patient. However the lack of power is omnipresent in fMRI-analyses (Lieberman & Cunningham, 2009) and therefore this procedure is also very useful in all branches of cognitive neuroscience. For example, negative results (i.e. voxels that are not significantly related to the task) are sometimes regarded as evidence against activation. However, these conclusions are not provided by null hypothesis significance testing. The presented procedure, on the other hand, quantifies the evidence for no activation at each voxel, and is therefore perfectly suited to interpret negative results.

We would like to stress that this procedure does not abandon null hypothesis significance testing. The classical significance testing framework is still included in the procedure, represented by one layer of significance. The method is merely an extension of the thresholded statistical parametric map, thereby providing a new layer with information on type II error rate control. Mixture modelling is similiar in spirit to this method, in that null and the alternative distribution are used, however mixture model applications usually focus on only controlling type I errors. With a fitted mixture model you could also apply our method and find p_0 and p_1 values, however we take pains to estimate alternative effect magnitudes *a priori*, from separate data, to remove any circularity.

In this procedure, control of false positives remains possible but our procedure also takes into account information on the false negative rate. We do not assert that our method alleaviates all concerns with multiplicity, and one possible direction of future work is a multiplicity correction that adjusts both null and alternative hypothesis inferences for the number of tests.

Appendices

5.A Scaling steps in ICA

Let $\boldsymbol{Y}$ be the $J \times I$ data matrix for time points $j = 1, \ldots, J$ and voxels $i = 1, \ldots, I$. Then the normalised datamatrix $\boldsymbol{Y}^*$ can be expressed as

$$\boldsymbol{Y}^* = \boldsymbol{Y}\boldsymbol{S}_0^{-1},$$

where $\boldsymbol{S}_0$ the $I \times I$ diagonal matrix with voxelwise robust variance estimates on the diagonal. Then the independent component analysis results in the following decomposition

$$\boldsymbol{Y}^* \approx \boldsymbol{M}\boldsymbol{C}^*, \tag{5.14}$$

where $\boldsymbol{M}$ represents the $J \times K$-mixing matrix, where K is the number of components. When $K < J$, equation 5.14 is only an approximation. $\boldsymbol{C}^*$ is the $K \times I$-matrix with the original image component loadings.

In the probabilistic ICA framework Beckmann & Smith (2004), $\boldsymbol{C}^*$ is Gaussian distributed, and can therefore be used as a test statistic. To normalise $\boldsymbol{C}^*$ to a standard Gaussian distribution,

$$\boldsymbol{C} = \boldsymbol{S}_1^{-1}\boldsymbol{C}^*\boldsymbol{S}_2^{-1} \tag{5.15}$$

where the diagonalmatrix $\boldsymbol{S}_1$ scales the components (over voxels) and the diagonalmatrix $\boldsymbol{S}_2$ scales the voxels (over components):

$$\boldsymbol{S}_1^2 = \text{diag}\{\boldsymbol{M}^{-1}(\boldsymbol{M}^{-1})'\} \tag{5.16}$$

$$\boldsymbol{S}_2 = \text{diag}\{\text{SD}(\boldsymbol{Y}^* - \boldsymbol{M}\boldsymbol{C}^*)\}\frac{\sqrt{I-K}}{\sqrt{I-1}}, \tag{5.17}$$

where $\text{SD}(\boldsymbol{Y}^* - \boldsymbol{M}\boldsymbol{C}^*)$ is the column-wise standard deviation of the residuals of the ICA approximation. Consequently, we can approximate $\boldsymbol{Y}$ as $\boldsymbol{M}\boldsymbol{S}_1\boldsymbol{C}\boldsymbol{S}_2\boldsymbol{S}_0$ in equation 5.6.

5.B Attenuation of Anticipated BOLD Effect for a Given IC

A basic result in psychometrics (Spearman, 1904) gives that the correlation between unreliable measures is attenuated by the the test-retest reliability of each measure. For example, if measure A imperfectly measures variable A^*, and B imperfectly measures B^*, then the correlation of A and B is attenuated relative to the uncorrupted measures:

$$\mathsf{Corr}(A, B) = \mathsf{Corr}(A^*, B^*)\ \sqrt{\rho_{AA}}\ \sqrt{\rho_{BB}}, \tag{5.18}$$

where ρ_{AA} and ρ_{BB} are the test-retest correlations of A and B.

In our setting, let A be the BOLD response, and B be the IC time course $\boldsymbol{m}$. The reproducibilty of BOLD (ρ_{AA}) is respectible, with Vul et al. (2009) reporting reliabilities between .66 and .94; however most procedures for fMRI data analysis do not take this into account, and hence we only consider $\rho_{AA} = 1$.

We are interested in the "reproducibility" of a IC time course *relative* to the experimental design $\boldsymbol{d}$. Of course obtaining two replicates of an IC time course that both equally reflect $\boldsymbol{d}$ is not feasible, but we can indirectly estiamte ρ_{BB} as follows.

Consider a general test-retest setting, where measurement B_1 is made at one time, and, later, a "retest" gives measurement B_2. Assuming additive error, we can relate the 'corrupted' measures to the 'uncorrupted' measures as

$$B_1 = B^* + \epsilon_1 \tag{5.19}$$
$$B_2 = B^* + \epsilon_2, \tag{5.20}$$

where ϵ_j, $j = 1, 2$, are the measurement-specific errors, and we assume $\mathsf{Var}(B^*) = \sigma^2$ is the variance of the perfectly reproduciable measure and $\mathsf{Var}(\epsilon_j)$, is the variance of the corrupting noise. The test-retest correlation is then

$$\rho_{BB} = \mathsf{Corr}(B_1, B_2) = \frac{\sigma^2}{\sigma^2 + \tau^2}. \tag{5.21}$$

If, instead, one corrupted measure was correlated with the uncorrupted 'true' measure, you find

$$\mathsf{Corr}(B_1, B^*) \quad = \quad \frac{\sigma}{\sqrt{\sigma^2 + \tau^2}} = \sqrt{\rho_{BB}}. \tag{5.22}$$

Or, equivalently, $\rho_{BB} = \mathsf{Corr}(B_1, B^*)^2$. In short, these results show that we can estimate the "reproducibility" of $\boldsymbol{m}_k$ as a noisy sample of the true $\boldsymbol{d}$ as $\rho^2_{\boldsymbol{m}_k\boldsymbol{d}}$.

Finally, from Equation (5.18), taking $\rho_{AA} = 1$ and $\rho_{BB} = \rho^2_{\boldsymbol{m}_k\boldsymbol{d}}$, we see that the attenuation factor needed to account for the mismatch between $\boldsymbol{m}_k$ and $\boldsymbol{d}$ is just $\rho_{\boldsymbol{m}_k\boldsymbol{d}}$.

6 Evaluating statistical procedures using different signal sources: a case study.[1]

Abstract Statistical inference in cognitive neuroscience focuses on stringent control of false positives, accepting the concomitant sacrifices in sensitivity. However, this is accompanied by a risk of false negatives, which can be detrimental, for example, in clinical settings where false negatives may lead to surgical resection of vital brain tissue. We have recently presented a new hypothesis thresholding procedure that incorporates information on both false positives and false negatives (Durnez et al., 2013). The result is a layered statistical map, marked by voxels exhibiting (i) strong evidence against the null hypothesis, (ii) evidence against the null but at practically insignificant effect sizes, (iii) responses where activation cannot be confidently excluded and finally (iv) responses where activation can be rejected.
In this chapter, some first steps are taken towards evaluation of statistical testing procedures by assessing the overlap between functional activations and structural connectivity. To compare our procedure with classical significance testing, we assess the difference between alternative-based testing (ABT) and classical hypothesis testing (CHT) using cross-correlations and overlap between activation and structural connectivity profiles (Homola, Jbabdi, Beckmann, & Bartsch, 2012). The approach is exemplified in a patient undergoing pre-surgical mapping and tractography.

6.1 Introduction

When surgically resecting brain tumors, one wants to minimize risk of resecting brain tissue involved in important cerebral functions. Pre-surgical fMRI probes such functions to localize eloquent brain tissue. Statistical inference in cognitive neuroscience focuses on control of false positives. The scientific discipline deems stringent control of false positives necessary, accepting the concomitant sacrifices in sensitivity. In a clinical setting, a loss

[1]This chapter represents collaborative work with following authors: Durnez J., Homola G., Jbabdi S., Nichols T., Moerkerke B. and Bartsch A.

in power means true activation is not discovered, and this might result in the resection of vital brain tissue. This asymmetrical way of penalising errors in statistical inference is undesirable in this context. We therefore presented a new hypothesis thresholding procedure that incorporates information on both false positives and false negatives and is as such ideally suited for pre-surgical fMRI (Durnez et al., 2013).

To test for brain activation, we test for each voxel $H_0 : \Delta = 0$ against $H_a : \Delta = \Delta_1$, where Δ is BOLD effect of interest in units of percent BOLD change and Δ_1 the non-zero effect magnitude expected under activation. In classical hypothesis testing, the evidence against H_0 is measured with the p-value, the null hypothesis probability of data at least as extreme than the observed result. Thresholding a p-value at α produces a statistical test that controls the false positive rate at α. To allow direct control of false negative risk, we present a symmetrical measure p_1 which quantifies evidence against the H_a. Thresholding this probability measure at β ensures control of the false negative rate at β.

We measure evidence against H_0 with $p_0 = P(T \geq t|H_0)$ and the evidence against H_a with $p_1 = P(T \leq t|H_a)$. As we do not expect a single magnitude of true activation but a distribution of different true values, we impose the following distribution on $\Delta_1 : \Delta_1 \sim \mathcal{N}\left(\mu, \tau^2\right)$. Let $\hat{\Delta}$ represent an estimator for the BOLD effect in a single voxel with corresponding standard error $SE(\hat{\Delta})$. With $T = \hat{\Delta}/SE(\hat{\Delta})$ the test statistic for testing H_0 against H_a, we find:

$$T_i \sim \mathcal{N}\left(\frac{\mu}{\mathrm{SE}(\widehat{\Delta}_i)}, \frac{\mathrm{SE}(\widehat{\Delta}_i)^2 + \tau^2}{\mathrm{SE}(\widehat{\Delta}_i)^2},\right) |H_a \tag{6.1}$$

When thresholding p_0 and p_1 at respectively level α and β, for given value of μ (expected activation) and τ (its uncertainty), the result is a layered statistical map, marked by voxels exhibiting (i) strong evidence against the null hypothesis, (ii) evidence against the null but at practically insignificant effect sizes, (iii) responses where activation cannot be confidently excluded and finally (iv) responses where activations can be rejected. In this chapter, we use data from a single patient undergoing pre-surgical mapping and tractography. We use these data to inspect the relation between activation and connectivity. More specifically, we assess the difference between alternative-based testing (ABT) and classical

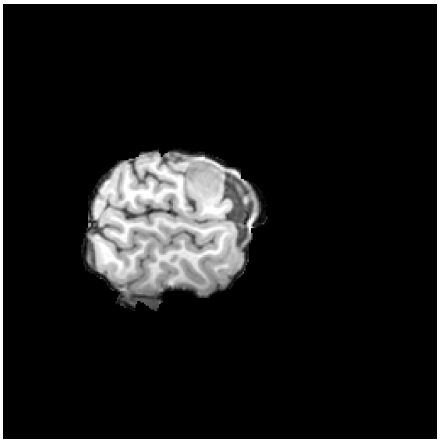


Figure 6.1: T_1-weighted scan of the patient.

hypothesis testing (CHT) using cross-correlations and overlap between activation and connectivity (Homola et al., 2012).

The approach is exemplified in a patient undergoing pre-surgical mapping and tractography.

6.2 Methods

6.2.1 Data

We consider data from a patient with left frontal grade II oligodendroglioma. The lesion is an intra-axial space-occupying lesion of 26 (A-P) $\times$ 29 (L-R) $\times$ 32 (V-D) mm extension behind the coronary suture entered to the left percentile sulcus. Figure 6.1 shows the tumor in the left hemisphere of the brain.The patient is right-dominant for hand, foot and eye. Data are obtained with a 3T TimTrio (Siemens, Erlangen, Germany) , 32 channel head coil.

6.2.2 Preprocessing and statistical analysis of fMRI data

In order to locate anterior and posterior language regions, the patient underwent two fMRI experiments: (1) reading nonfinal embedded clause sentences versus rest (semantic language-comprehension task and (2) detection of words from pseudowords versus blocks of tones (phonological language task. The first contrast aims to detect the posterior language area, while the second contrast refers to the anterior language area. The data are processed using FSL 5.0.6 (http://www.fmrib.ox.ac.uk/fsl/). As preprocessing, the data are motion-corrected using mcflirt, pre whitened using film and smoothed with a Gaussian kernel with 5mm full width at half maximum (voxelsize $\times$ 3 $\times$ 3.45 mm). The data are transformed to T_1-space before analysis using FSL's FLIRT tool, where the transformation matrix is computed based on the transformation from the mean EPI-image to the T_1-image. First, regions-of-interest (ROI) are defined. For the posterior language area (1st paradigm), we restrict analysis to the anterior devision of the middle and superior temporal gyrus and the supramarginal gyrus. The analysis of the anterior language area (2nd paradigm) is restricted to the inferior and middle frontal gyrus. The first-level analysis is carried out by applying a GLM within FEAT within these regions-of-interest. From the GLM we derive T-statistic images and a p-value for each voxel. We apply the testing procedure presented above, in which four layers of activation are defined. For the α-parameter, we choose the cutoff that controls the false discovery rate at 5% within the ROI. The β-parameter is set to 0.20, to provide an average statistical power of 80%. For the alternative distribution, we aim at effects of size $\mu = 0.50$ (0.5 % BOLD change) with variation $\tau = 0.01$. After analysis with the alternative-based procedure, four voxels are labeled according to the following labeling scheme:

- **NON-ACTIVE LABEL:** $p_0 > \alpha \cap p_1 < \beta$**:** Activation can be confidently excluded.
- **ACTIVE LABEL:** $p_0 \leq \alpha \cap p_1 \geq \beta$. The voxels show strong evidence against the null hypothesis.
- **UNCERTAINTY LABEL:** $p_0 \geq \alpha \cap p_1 \geq \beta$**:** Voxels cannot be confidently declared inactive.

- **PRACTICAL INSIGNIFICANT LABEL: $p_0 \leq \alpha \cap p_1 \leq \beta$:** These voxels show evidence against the null but at practically insignificant effect sizes.

To compare classical hypothesis testing (CHT) with alternative-based testing (ABT), we define the significant result based on these layers. Classical testing results comprise the **active** layer and the **practically insignificant** layer (which just corresponds to using a FDR corrected threshold at 5%. The *significant area* for the alternative-based thresholding procedure is the **active** layer and the **uncertain** layer, which corresponds to voxels with $p_1 \geq \beta$.

6.2.3 Preprocessing and statistical analysis of DWI data

The diffusion weighted images are taken in 2x160 directions, with resolution $120 \times 120 \times 60$. DWI data are corrected for eddy-currents (including motion) and geometric distortions and brain-extracted. The DWI data are processed using FSL's FDT toolbox. Two fiber orientations are modeled and the probabilistic distributions of diffusion parameters are built up at each voxel (using bedpostx, part of FDT). After computing the fiber orientations, the fiber anisotropy results are transformed to T_1-space using FSL's FLIRT tool, where the transformation matrix is computed based on the translation from the mean functional isotropy image to the T_1-image. For the tractography in T_1-space, we use probabilistic modelling of multiple fiber orientations using probtrackx (part of FDT). The goal of the DWI analysis is to track which voxels in the anterior language area have tracts to the posterior language area and vice versa. To that end, two masks are defined. The anterior language area is defined as the intersection of the fMRI results for the words-tones-contrast and the anterior anatomical mask described above (inferior and middle frontal gyrus). The posterior language area is defined as the intersection of the fMRI results for reading nonfinal embedded clause sentences and the posterior anatomical map used for fMRI analysis (anterior devision of middle and superior temporal gyrus and the supra marginal gyrus). Probabilistic streamlines are seeded from one of these masks. A total of 5000 samples is sent out from each tracking point. Stop masking is used to exclude indirect routes. The result is a map containing for each voxel the number of samples seeded from that

voxel reaching the relevant target mask. These probabilistic pathways are thresholded at $\geq 1\%$ connecting samples passing through each voxel. We measure connectivity score as the number of connecting samples divided by the total number of samples.

6.2.4 Combining fMRI and DWI

We compare the connectivity scores and the activation scores (p-values) using a minimum intersection map. The idea is that peaks in the structural connectivity profile should predict peaks in the functional activation profile.To generate voxel-wise minimum intersection maps (see Figure 6.2), the distributions of activation probabilities and average connectivity scores are shifted to zero minima and normalized to their robust maximum values (i.e. the 95th percentile).

6.3 Results

6.3.1 fMRI results

In Figure 6.3, we show the results of the fMRI data analysis in the four different layers. There is an indication for activation in the tumor based on the deviation from the null hypothesis of no activation. However, adding information on the alternative indicates that the result might be statistically significant, but is not of practical significance: the effect size in these voxels is too small to represent real activation. Electrocortical stimulation mapping and the postoperative patient condition after gross tumor resection confirmed that there were no essential intratumoral activations. We show how the uncertain layer is in this case mainly an extension of the width of the active region and can be seen as a 'safe' boundary delineation.

6.3.2 DWI results

We show in Figure 6.4 how high connectivity values are indeed related with the posterior language area. The anterior language area is not correctly discovered, as the connectivity values in frontal mask are highest inside the tumor, which is adjacent to the expected region according to the mask.

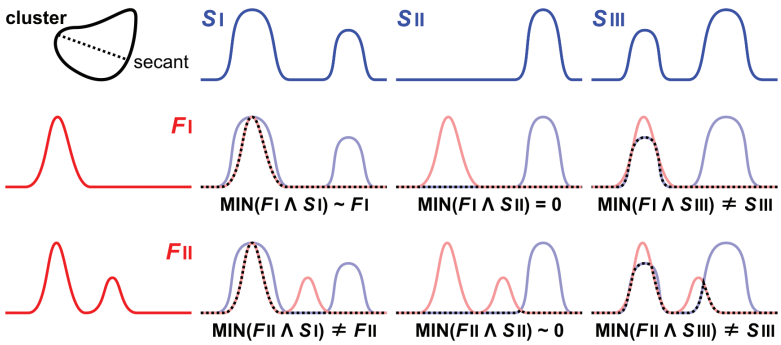


Figure 6.2: From Homola (2012) *Schematic illustration of minimum intersection maps. Minimum intersection maps are generated between different profiles of functional activation (red) and structural connectivity (blue). The profiles are normalized, i.e. scaled to the same min/max range. To build the minimum intersection (dotted), the minimum (MIM) of the two is considered at each point along the profile. Minimum intersection peaks indicate different degrees of spatial correspondence between high structural connectivity (S) and functional probability (F) values: Minimum intersection maps resembling F signify concordant presence of F- and S-peaks (left and upper right minimum intersects). Note that when F and S are too dissimilar, the minimum intersection is flat (middle). A non-flat minimum intersect with a sharp peak and displaced compared to F indicates a close but out-of-center overlap of F- and S-peaks (bottom right).*

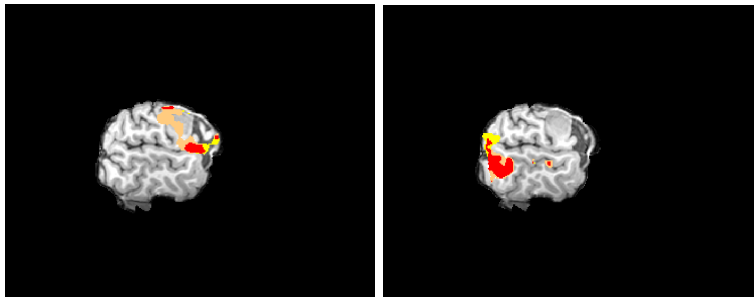


Figure 6.3: The results from the fMRI analysis with alternative based thresholding. The results for the anterior language area is shown in the left panel, the right panel represents the experiment for the posterior language area. The copper color refers to the practically insignificant voxels, red refers to active voxels, the yellow voxels show uncertainty

It should be noted that the specific value of the connectivity values is dependent on the size of the masks and is therefore not interpretable.

6.3.3 Minimum intersection maps

For the posterior language area, we find reasonable overlap between connectivity measures and fMRI results in a comparable way, as is shown in Figure 6.5. For the anterior language area, we find slight overlap between the classical hypothesis testing results and connectivity **in** the tumor, whilst no overlap between alternative-based testing and classical hypothesis testing.

6.3.4 Measures of activation and connectivity scores

More insight in the connectivity scores in relation to the testing procedures is given in Figure 6.6. An important remark is that the connectivity values in the posterior language area are overall smaller than the connectivity values in the anterior language area. This can be explained by the fact that the anterior mask is smaller than the posterior mask, and thus there is a smaller chance that sent samples arrive in the anterior mask than in

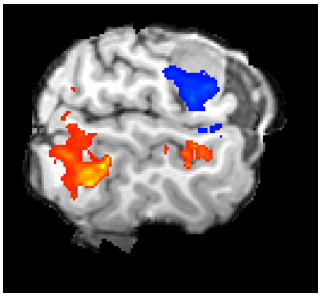


Figure 6.4: The results from the DWI analysis. The red voxels are the connectivity values in the posterior language area mask, while the blue voxels represent the connectivity values in the anterior language area.

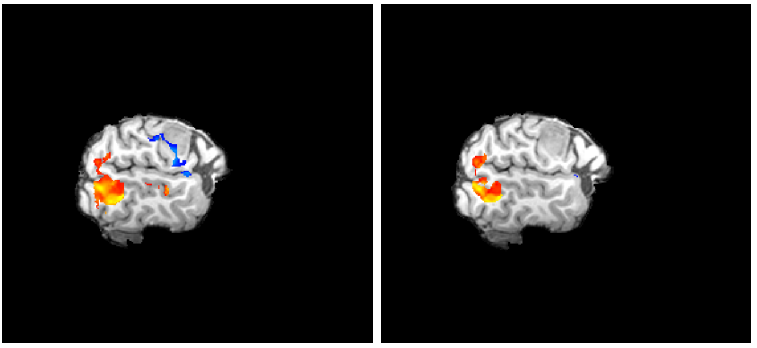


Figure 6.5: The minimum intersection maps. The red voxels represent the overlap between fMRI and DWI for the posterior language region, while the blue voxels represent the overlap for the anterior language area. The left image shows the classical testing results, the right figure are the results from the alternative based thresholding procedure.

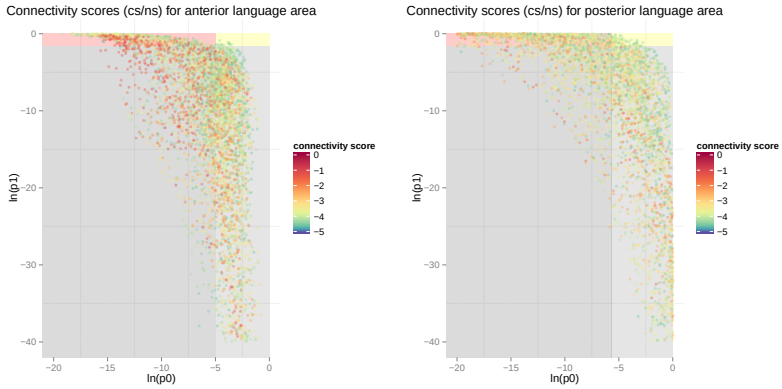


Figure 6.6: The connectivity scores with respect to the voxelwise p_0 and p_1-value. Each dot represents a voxel with a connectivity score higher than 1%. Furthermore the testing procedures are shown as background colors. The labeling is as follows: red refers to the active label, yellow represents the uncertain label, practical insignificance is shown in dark grey and light grey is for non-significant voxels.

the posterior mask. This difference has no intrinsic meaning with respect to the connectivity in both regions.

The main finding is the apparent relation between p_0 and connectivity scores. Low connectivity scores are related to high p_0 values, as expected. But furthermore we also see that higher p_1-values are also linked with higher connectivity scores. This indicates that adding the p_1-value to the testing criterion provides useful information.

6.3.5 Spatial cross-correlations

We show the relationship between p_0-values and connectivity scores in Figure 6.7. For the posterior language area, we show that higher p_0-values are accompanied by lower connectivity scores as expected. However, for the anterior language area, there is a clear bimodal distribution visible and therefore the smoother cannot be interpreted in a straightforward way. The same figure in relation with p_1-values instead of p_0-values is shown

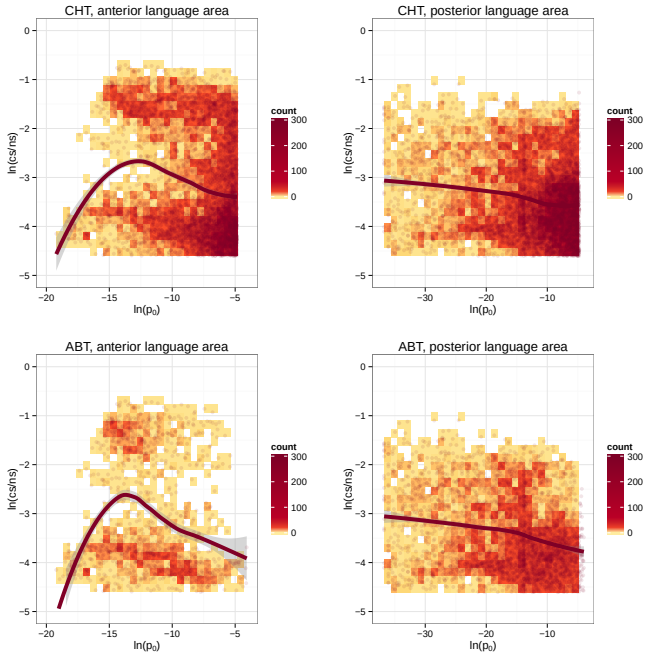


Figure 6.7: Two-dimensional histograms for the p_0-values against the connectivity values. Darker colors indicate more datapoints in the given bin. The line through the datapoints is a non-parametric loess smoother.

in Figure 6.8. We expect that higher p_1-values indicate higher functional activation, and should be accompanied by higher connectivity values. This is observable in the posterior language area, but again because of the bimodality in the anterior language area, we make no interpretations of the smoother. Finally, we show the connectivity values in relation with $(1 - p_0)/(1 - p_1)$ in Figure 6.9. Higher values of the numerator indicate more activation, higher values of the denominator indicate less activation. As such, higher values of the fraction indicate more activation. Again, we find expected results in the posterior language area but not in the anterior language area. In the posterior language area, we find a drop in connectivity values for high activation.

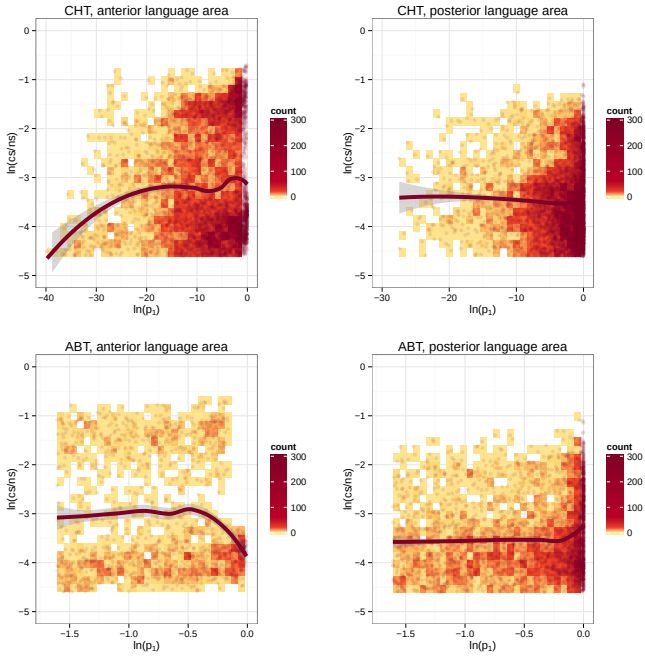


Figure 6.8: Two-dimensional histograms for the p_1-values against the connectivity values. Darker colors indicate more datapoints in the given bin. The line through the datapoints is a non-parametric loess smoother.

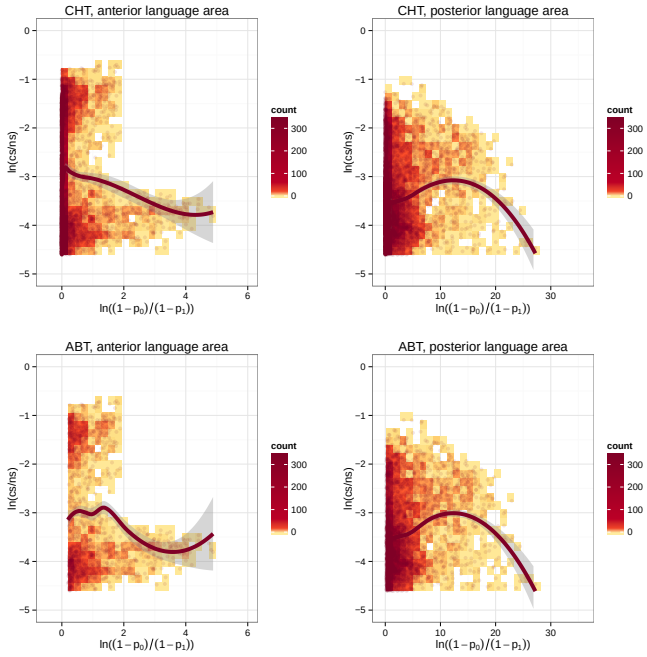


Figure 6.9: Two-dimensional histograms for $(1 - p_0)/(1 - p_1)$ against the connectivity values. Darker colors indicate more datapoints in the given bin. The line through the datapoints is a non-parametric loess smoother.

6.4 Discussion

In this work, we have taken some first steps towards evaluating statistical procedures by relating the fMRI results to the structural connectivity profile. More precise, we localise the two main language areas in the brain, of which we know there is a strong connection between both. We measure voxelwise connectivity measures, indicating the strength of connection from one region to the other region, and relate these to the voxelwise testing procedure.

For the case study at hand, an important finding from the fMRI results is that only the classical testing procedure detects activation within the tumor while removing the tumor did not have any significant effect, and moreover electrocortical stimulation mapping confirmed that there were not intramural activations. However, we also found high connectivity measures in the tumor and not in the area where it is to be expected. Furthermore we observed an unexpected bimodal distribution of the connectivity values in this area. There might be several reasons for these results, such as distortion because of the tumor. Another reason could be that co-registration (between fMRI - structural T_1-scan and DWI data) resulted in spatial deplacement of crucial information. Co-registration is optimal in terms of finding the global minimal deviation from the template, but around tumors, registration often fails locally. A third reason for the connectivity pattern in the tumor could be due to our mask definition. We have used large masks, which allows large pathways to be discovered. To avoid contamination between different pathways, we used exclusion masks. However, it is still possible that there is contamination of the dorsal pathway from the ventral pathway, which could explain the high intratumoral connectivity values. Based on these data and analyses, a unique cause for the results cannot be identified.

One possible way to further inspect the relationship between functional and structural measures, is to define unrelated regions for negative control. High connectivity values in unrelated regions (either close or far from the tumor) could disentangle possible explanations for the results.

Furthermore, we would like to remark that while we found a drop in connectivity profile for high activation values, this drop has also been observed with Homola et al. (2012) in a different setting.

This work aims to show how evaluation of statistical procedure could be validated using real data. We showed an example of such a validation with data from a single person. However, these results are not answering all questions. It is clear that on the one hand more research should be done on the relation between fMRI and connectivity in general. On the other hand, interesting results have emerged from the current validation but to draw conclusions on the performance of both thresholding procedures, the validation requires more data and deeper analyses.

Acknowledgements

The computational resources (STEVIN Supercomputer Infrastructure) and services used in this work were kindly provided by Ghent University, the Flemish Supercomputer Center (VSC), the Hercules Foundation and the Flemish Government - department EWI.

7 General discussion

7.1 Introduction

During the past five years, we studied practical and accurate approaches to statistical significance and power for fMRI. We investigated and compared different methods for detecting brain activation based on statistical significance in terms of reproducibility, statistical specificity and power (Chapter 2). We presented a procedure to estimate power for a given study, and demonstrated the lack of statistical power in current fMRI studies (Chapter 3). We extended the procedure to enable the computation of the minimal sample size in an fMRI experiment that is required to achieve a predefined statistical power (Chapter 4). However, the goal of an fMRI study may prevent the possibility of simply increasing its sample size. Therefore, we presented an alternative based thresholding procedure for pre-surgical fMRI (Chapter 5) and some first steps towards evaluating the method by combining functional (fMRI) and connectivity (DTI) data have been demonstrated (Chapter 6).

In this chapter, we present a short overview of our results and their implications. We integrate these findings with existing literature and put the frameworks of this thesis in a broader context, both with more historical important work and with recent advances in the field. Finally, we will present some new interesting research directions, through which the power problem in fMRI and neuroimaging in general could be further investigated.

7.2 Multiple testing framework

We dedicated the first chapter of this dissertation to the multiple testing problem in fMRI. The multiple testing problem emerges when many statistical tests are performed simultaneously; in that case there is an inflation of false positive results. This problem is typically addressed by imposing a more stringent statistical threshold, though the extent of this adjustment is up for debate. The thresholding problem is not new to neuroimaging. Ever since the first fMRI experiment, the multiple testing problem was acknowledged in determining appropriate thresholds, and specific solutions were provided. The multiple testing problem is an old problem, for which many solutions were available at the time fMRI analyses were starting to develop. These procedures aim to control the familywise error rate, the chance of at least one false positive over all tests performed (e.g. the Bonferroni correction, Bonferroni 1936). At the same time there was already a well established framework for randomness in spatial fields, the random field theory (Adler, 1981). Combining these two, Worsley et al. (1992) and (Friston et al., 1991) presented a way to take both the multiple testing problem in fMRI and the spatial character of the fMRI data. Gradually, focus shifted to topological inference in which voxelwise testing is replaces by cluster- or peakwise testing. This approach also builds on random field theory.

Friston, Holmes, Poline, Price, & Frith (1996) was the first to discuss power issues in neuroimaging, by performing a power test for different levels of inference. However, controlling the familywise error rate was an idea put forward when the multiple testing problem mostly comprised a small number of tests. With the rise of big datasets (neuroscience, genetics, etc.), it became clear that controlling the family-wise error rate was too strict as a testing criterion. Therefore the control of the false discovery rate, the proportion of false discoveries out of all discoveries, was introduced (Genovese et al., 2002). Up until today, many of the statistical questions from that time are currently still hot topics.

Recently, the problem of thresholding received a storm of attention when Bennett, Baird, Miller, & Wolford (2009) performed a real psychological experiment with one subject: a dead atlantic salmon, with an uncorrected statistical threshold. The dead salmon appeared to have brain

activation. This experiment was designed to show that when the end user community prefers the use of uncorrected thresholds, despite many available statistical procedures to control the multiple testing problem in fMRI, results are prone to discover an abundance of false brain activation. In another publication, Bennett, Wolford, & Miller (2009) argued for a principled control of false positives. *"A procedure that places appropriate limits on the rate of false positives across the whole brain gives readers the information they need to properly evaluate the results."* (Bennett, Wolford, & Miller, 2009). Luckily, under the impulse of a few popular publications about publishing fMRI studies (R. A. Poldrack et al., 2008; Friston, 2012), today nearly all fMRI publications correctly report how they handle the multiple testing problem.

Still, there is no golden standard to deal with the multiple testing problem. Many effort has been put into providing a good comparison between different thresholding procedures (Marchini & Presanis, 2004; Logan & Rowe, 2004; Nichols & Hayasaka, 2003; Eklund et al., 2012). However, through FDR control, power is only gained by allowing more false positives. We demonstrated this in our first study (Durnez, Roels, & Moerkerke, 2014). In the second chapter of this thesis, we found no new trade-off between false positives and false negatives for FWER and FDR control. In other words, for a given specificity, FWER and FDR control performance is equal in terms of sensitivity. Therefore, we compared FDR and FWER control (operationalised as the procedure by Benjamini & Hochberg 1995 and Bonferroni 1936) on a third operation characteristic, namely its stability. We found a much bigger variance on the results of the FDR procedure than on the results of the FWER controlling procedure, because of its adaptive nature. Including stability into the decision criterion through bootstrapping made FDR control more stable but did not alter the trade-off between sensitivity and specificity. As a conclusion, like (Bennett, Wolford, & Miller, 2009), we advocate a well-contemplated choice of threshold since the same (both liberal and conservative) results can be obtained through FWER and FDR control.

7.3 Statistical power framework

Currently there is still a huge imbalance between false positives and false negatives in fMRI with typically less attention for the latter. In cognitive neuroscience, a false positive means fallacious support for a given cognitive theory. While false positives can often be discovered by unsuccessfully trying to replicate the study, much time, effort and money can be spent. As a result, the scientific discipline generally deems stringent control of false positives necessary, accepting the concomitant sacrifices in sensitivity. However, in chapter five, we discussed pre-surgical fMRI, where false negatives can be detrimental, for example, in clinical settings where false negatives may lead to surgical resection of vital brain tissue. We presented a procedure procedure that incorporates information on both false positives and false negatives [2]. The result was a layered statistical map, marked by voxels exhibiting (i) strong evidence against the null hypothesis, (ii) evidence against the null but with practically insignificant effect sizes, (iii) responses where activation cannot be confidently excluded and finally (iv) responses where activations can be rejected. This procedure can greatly decrease the risk of missing activation in a surgical setting.

Nonetheless, the problematic imbalance between false positives and false negatives is not restricted to clinical settings only. Lieberman & Cunningham (2009) also argue for a better balance between false positives and false negatives in cognitive neuroscience. In 2013, cognitive neuroscience was criticised in the important publication of Button et al. (2013), where the author tackles the reliability of all neurosciences. They stated that most neuroscience studies suffer from low statistical power. They argued that low power not only decreases the chance of detecting a true effect, but it also reduces the chance that a statistically significant result indicates a true effect. To meet an acceptable level of power in fMRI studies, several a priori power calculation methods have been made available (Desmond & Glover, 2002; Mumford & Nichols, 2008; Zarahn & Slifstein, 2001; Hayasaka et al., 2007; Smith et al., 2007; Zarahn & Slifstein, 2001). However, while the type I error rate of a study is mostly clearly defined and easily interpretable, the existing power estimators often involve complex equations and many unknown parameters, depending on how activation is defined. In our third chapter, we presented a method to estimate the

number of activated features (Durnez, Moerkerke, & Nichols, 2014). We considered peaks or clusters of activation as a topological feature. Knowledge of the number of active features enables to estimate post-hoc how powerful the selection procedure performs. In our data example in the third chapter, we showed that when one aims to see the activation related to viewing faces, only 50% of the actual activity can be found. Such findings will mostly impair the most vulnerable studies, such as studies involving social cognition, which are often characterised by small effect sizes and low power to detect the effect (Lieberman & Cunningham, 2009).We argued that a call needs to be made to statistical testing procedures with a better balance between sensitivity and specificity. By raising awareness on the power of existing studies, this procedure is a first step in this direction. A drawback of this procedure is that it only functions post-hoc and cannot predict the outcome of a study before it is conducted. As such the lack of power can be made more explicit, but not overcome with this procedure.

While power is straightforward to compute for a single, univariate response, determining the power of a fMRI study is a formidable task as the magnitude, spatial extent and location of a hypothesized effect are difficult to specify. In the fourth chapter of this thesis, we presented a simple way to characterize the spatial signal in a fMRI study, and a direct way to estimate power based on an existing pilot study. Specifically, using just (1) the volume of the brain activated and (2) the average effect size in activated brain regions, we directly calculated power for given sample size, brain volume and smoothness. This procedure allows minimizing the cost of an fMRI experiment, while preserving a predefined statistical power.

7.4 Future directions

This dissertation's main merit lies in a thorough investigation of the problem of statistical power for fMRI, which has deepened our understanding of the driving factors of statistical power in fMRI and serves as a reference point for further research. We end this dissertation by presenting possible ways as to how significance and power in neuroscience can be further examined.

Integration within the field: meta-analysis

The number of human fMRI studies has risen dramatically over the last 15 years. The growth of the field deems integration of findings necessary. Quantitative meta-analyses can be used to localise the brain regions most consistently activated by a particular type of task (Wager, Lindquist, Nichols, Kober, & Van Snellenberg, 2009). Moreover, meta-analyses can help to separate the wheat from the chaff in imaging studies, identifying consistent activations and those that do not replicate (Kober & Wager, 2010). In other words, false positives will be filtered as they will likely appear in only a small fraction of the studies while a false negative can still be picked up as this activation will likely show in some of the performed studies. While we focused in this thesis mainly on single (multi-subject) studies, it is clear that integrating different fMRI studies to the same cognitive function can largely improve statistical power as well - especially given the large body of studies that are already available. An overview of statistical procedures to perform meta-analyses is given in Wager, Lindquist, & Kaplan (2007).

Integration beyond the field: multimodal data integration

Meta-analyses are available to integrate different studies in the same modality (i.e. fMRI). However, the growth of neuroscience extends far beyond fMRI, as nowadays many different techniques for brain activation and structure exist, such as diffusion weighted imaging, resting state fMRI, structural MRI. The power problem in neurosciences extends beyond the fMRI studies we have targeted in this thesis. It is therefore a logical step to advance towards combining and integrating data from different sources (R. Poldrack, 2012). We have touched this topic in chapter six, where we combine DWI with fMRI. The possible combinations of information sources is endless, and this line of research is only starting to develop.

An open source character is typical and deemed important in neuroscience. Open databases are perfectly suited for multimodal data integration. Many repositories with large datasets, such as the Human Connectome Project (Van Essen et al., 2012), or repositories collecting datasets from many different studies, such as OpenfMRI (R. Poldrack, 2012) are

available. Bringing together already existing information from different sources will not only be very cost-efficient, but more importantly it will advance a better understanding of the human brain.

Including practical significance into the testing criterion

Finally I would like to point out the importance of a better inspection of the observed effect size when analysing fMRI data. Statistical analyses focus mainly on T or Z statistics, which roughly equals as the observed effect size divided by its standard error. This means that voxels with a very low variance are often seen as voxels carrying important effects, even if the effect size is small, and not biologically relevant. Likewise, effects in voxel may be of scientific interest but may not reach the threshold of statistical significance because of a low precision with which the effect is measured, and are therefore ignored. Therefore it is key to make a distinction between practical and statistical significance. Statistical significance is concerned with whether a research result is due to chance or sampling variability; practical significance is concerned with whether the result is of scientific interest (Kirk, 1996). To define what is practically relevant, we should look at the unit of measurement. The MRI signal is measured is units of BOLD signal and the contrast between different conditions will be measured in percent BOLD change. This unit is difficult to interpret, and as such it is hard to make statements on the expected percent BOLD change, but Desmond & Glover (2002) reported that effect sizes between 0.25 and 0.75 percent BOLD change are valid estimates for relevant effects. Therefore we can assume that effect that are much smaller than these estimates are not biologically relevant. The idea of mapping and interpreting effect sizes has also been suggested in literature by Van Horn et al. (1998) and Nichols & Holmes (2002).

We approached this topic in chapter five, where we included the expected effect size that is of practical interest in the testing criterion. Instead of labeling voxels as 'active' versus 'non-active', we added two new labels: 'practically non significant' and 'uncertain'. These two labels refer to the link between the observed effect size and the expected effect size. If an effect showed no statistically significant deviance from the null hypothesis, but the observed effect size was still close to the expected effect

size, we declared the voxel as 'uncertain'. If the effect showed substantial deviance from the null hypothesis of no activation, but the effect size was small (and biologically non-relevant), the voxel was declared practically insignificant. We have shown the merit of this classification system in chapter six. There we found deviance from the null hypothesis in a brain tumor, but with our alternative-based testing procedure we find that this activation is practically insignificant, a label that was proven to be correct by the intact psychological functions after removal of the tumor.

We argued for the use of effect sizes in the testing criterion in a presurgical context. Two other settings where including the effect size is of importance are the following:

Functional Regions-of-Interest (fROI) When testing for significance for a certain brain task, all measurements of the brain are considered. However, sometimes there may be strong arguments as to why only a portion of the brain needs to be investigated. For example for a language experiment, the present knowledge of the language network can restrict the analysis to a certain region of interest (ROI). This reduces greatly the number of tests and thus increases sensitivity. For certain experiments, researchers perform a single subject fMRI experiment to detect for example the subject-specific language network. The ROI is defined based on a functional task, and is therefore called a functional ROI. In such a task localiser, it is also of key to avoid false negatives, and thus this setting will benefit from statistical analysis including pre-defined effect sizes.

Large datasets Most of the published fMRI studies involve 15 to 25 subjects. However, subtle effects are not observable in small samples. Consequently, large consortia of research groups are emerging to obtain larger sample sizes that are able to inspect small effects, and they reach sample sizes up to 1000 subjects. But by increasing sample size, statistically non-significant results may become significant, independent from the observed effect size. Therefore Quinlan (2013) argues in favour of small-scale science. Thyreau et al. (2012) tested a cohort study of 1326 subject and found that the larger the sample size, the less anatomically relevant the results tend to be. These facts notwithstanding, we believe that returning to small scale science would be a serious degression in the growth of neu-

roscience. A better solution to the significance problem in large databases would be to include pre-defined effect sizes in the testing criterion, and as such provide an answer biologically implausible, but significant results.

8 Nederlandstalige samenvatting

Eén van de belangrijkste trends van de afgelopen 20 jaar op gebied van psychologisch onderzoek, is de opkomst van neurologische beeldvorming. De koploper in dit domein is functionele magnetische resonantie imaging (fMRI), waarmee de verhouding van zuurstofarm en zuurstofrijk bloed gemeten kan worden (het BOLD signaal), wat toelaat om breinactivatie op een niet-invasieve manier te meten. In een fMRI-experiment wordt een opeenvolging van hersenbeelden bekomen terwijl de participanten een cognitieve taak uitvoeren. Wanneer een verandering van signaal tussen de beelden optreedt, kan men op die manier de taakgerelateerde gebieden in de hersenen lokaliseren. Een fMRI-studie leidt tot de productie van enorme hoeveelheden data. Om deze data adequaat te analyseren, worden de hersenbeelden in een groot aantal volume-eenheden (voxels) opgedeeld. Vervolgens wordt per voxel de tijdsreeks van het gemeten signaal gemodelleerd als een lineaire combinatie van verschillende signaalcomponenten. Op die manier kan per voxel onderzocht worden of er aanwijzingen voor activatie zijn.

Nagaan of er activatie is, houdt in dat er een gigantisch aantal statistische toetsen simultaan uitgevoerd worden (meer dan 100 000 voxels). Daardoor wordt men geconfronteerd met het multiple testing probleem: het veelvuldig toetsen leidt tot een explosie van valse positieven. Om hiervoor afdoende te corrigeren, zijn procedures ontwikkeld om het aantal type I fouten (het verkeerdelijk oppikken van activatie) binnen de perken te houden. Zo werd er aanvankelijk voor gezorgd dat men de kans op ten minste 1 vals positief resultaat (de family-wise error rate, FWER) onder een bepaald niveau α kan houden. Deze aanpak heeft tot gevolg dat de sensitiviteit of de power van de toetsen heel laag wordt. Dit betekent dat er een grote toename is van het aantal type II fouten (het verkeerdelijk niet oppikken van activatie). Omwille van het conservatieve karakter van de

FWER als foutmaat, verschoof de focus naar een foutmaat die het relatief aantal vals positieven ten opzichte van het aantal geselecteerde voxels onder controle houdt. Er werden procedures ontwikkeld om deze foutmaat, de false discovery rate FDR, lager dan een vooropgesteld niveau α te houden (Benjamini & Hochberg, 1995; Storey, 2002; Genovese et al., 2002; Chumbley & Friston, 2009). Bij een vast α-niveau levert dit meer power op dan controle van de FWER. Deze hogere power gaat steeds gepaard met een daling in specificiteit, met andere woorden, er wordt enkel aan power gewonnen door het aantal valse positieven te laten toenemen. We kunnen onszelf afvragen of de winst in power het verlies in specificiteit wel waard is? We bekijken dit in meer detail in hoofdstuk twee. Daarin vergelijken we controle van FWER en FDR op een aantal eigenschappen, zoals specificiteit, statistische power en stabiliteit voor een breed scala aan mogelijke scenario's. We vinden dat voor een vast α-niveau FWER-controle inderdaad een heel lage power heeft. Dit probleem vermindert sterk bij het gebruik van FDR-controle. Als we echter het α-niveau variëren, leveren FDR en FWER-controle bij eenzelfde specificiteit gemiddeld dezelfde power op. Dit betekent in essentie dat de FDR-procedures enkel minder streng zijn bij de selectie maar geen nieuwe trade-off tussen sensitiviteit en specificiteit opleveren. Bovendien vinden we dat FDR-controle gepaard gaat met een lagere stabiliteit van de testprocedure en dus een beperktere reproduceerbaarheid. Bijgevolg concluderen we dat FDR-controle niet de ultieme oplossing biedt voor het multiple testing probleem. Bij het uitvoeren van een fMRI studie moet men daarom voldoende aandacht schenken aan het probleem van thresholding en aan de power die hierbij hoort.

Behalve een beperkt aantal publicaties over hoe een gepaste steekproefgrootte kan gekozen worden in functie van power (Friston, 2012; Desmond & Glover, 2002; Mumford, 2012), heeft statistische power in fMRI tot nu toe relatief weinig aandacht gekregen. In het derde hoofdstuk stellen we daarom een procedure voor waarmee we post-hoc power kunnen schatten voor een gegeven studie. Dit stelt onderzoekers in staat hun eigen studies te evalueren, en bovendien is het met deze manier mogelijk na te gaan in welke mate deze studies in staat zijn om de taak-gerelateerde hersenactivatie te detecteren.

Een nadeel van deze procedure is dat ze enkel post-hoc functioneert, en daardoor de uitkomst van een studie niet kan voorspellen voordat deze

uitgevoerd wordt. Daarom breiden we de procedure uit in hoofdstuk vier. We stellen een eenvoudige manier voor die het spatiale signaal in een fMRI studie karakteriseert, en een directe manier om power te schatten op basis van een bestaande pilootstudie. Meer bepaald gebruiken we enkel (1) de proportie van het brein dat geactiveerd is en (2) de gemiddelde effectgrootte in geactiveerde breinregio's om powerberekeningen te doen voor een gegeven steekproefgrootte, breinvolume en smoothness. Deze procedure laat toe om de kost van een fMRI experiment te minimaliseren, terwijl men een vooraf te bepalen niveau van statistische power kan handhaven.

Anderzijds laat het doel van een studie niet altijd toe om de steekproefgrootte te vergroten. Pre-chirurgische fMRI is daar een voorbeeld van. Bijvoorbeeld bij hersentumoren en epilepsie worden bij een patiënt via prechirurgische fMRI psychologische taken gelokaliseerd in de hersenen, voordat deze patiënt hersenchirurgie ondergaat. Zo probeert men te vermijden dat cruciale psychologische functies beschadigd worden bij de operatie. In dit geval moet de onderzoeker de statistische power aanvaarden die hoort bij data afkomstig van slechts één subject. Echter gaat een strikte focus op de preventie van valse positieven steeds samen met een risico op valse negatieven. Deze focus kan vooral in klinische settings ongunstig zijn, waar valse negatieve kunnen leiden tot het chirurgisch verwijderen van belangrijk hersenweefsel. Daarom stellen we in hoofdstuk vijf een alternatieve thresholding procedure voor, waar rekening gehouden wordt met de kans op zowel valse positieven als valse negatieven. We combineren twee maten voor significantie in elke voxel: een klassieke p-waarde die bewijskracht tegen de nulhypothese van geen effect meet en een alternatieve p-waarde die bewijskracht tegen activatie van een welbepaalde grootte meet. Het resultaat is een gelaagde statistische map van het brein. Een eerste laag representeert voxels die sterke evidentie uiten tegen de traditionele nulhypothese, terwijl een tweede laag voxels bevat waar activatie niet kan uitgesloten worden met een zekere betrouwbaarheid. In de derde laag van voxels kan activatie uitgesloten worden. In deze procedure blijft controle op valse positieven mogelijk, maar onze procedure neemt daarbij ook informatie over de kans op valse negatieven in overweging. Met de procedure van hoofdstuk vijf hopen we beter de psychologische functies in het brein af te bakenen.

In hoofdstuk zes geven we de aanzet om een manier te ontwikkelen waarmee statistische procedures voor fMRI gevalideerd kunnen worden. We passen dit toe op een fMRI experiment op een patiënt met een hersentumor. De patiënt onderging eveneens Diffusion Weighted Imaging (DWI), een techniek die eerder de structurele connecties in de hersenen blootlegt, en die verschillende breinregio's kan afbakenen op basis van hun fysieke connecties. We passen zowel de klassieke testprocedure als onze alternatieve testprocedure toe op de fMRI data, en we vergelijken welke van beide procedures best de resultaten van DTI kan voorspellen. Een grote overlap tussen volledig onafhankelijke procedures (fMRI en DWI) wijst op een betere predicatieve validiteit van de procedure.

We besluiten dit doctoraat met een overzicht van de belangrijkste bevindingen en conclusies van de verschillende studies in hoofdstuk zeven. De gevolgen en tekortkomingen van onze resultaten worden besproken in een meer algemeen denkkader. Daar bovenop geven we indicaties voor toekomstig onderzoek voor wat betreft statistische power in neurowetenschappen. Hoofdstuk twee, drie, vier en vijf werden origineel geschreven als wetenschappelijke papers. Hoofdstuk twee werd gepubliceerd in Biometrical Journal (Durnez, Roels, & Moerkerke, 2014), hoofdstuk drie in NeuroImage (Durnez, Moerkerke, & Nichols, 2014) en hoofdstuk vijf werd gepubliceerd in Cognitive and Behavioral Neuroscience (Durnez et al., 2013).

Bibliography

Adler, R. J. (1981). *The geometry of random fields*. Wiley, London.

Adler, R. J., Bobrowski, O., & Borman, M. S. (2010). Persistent homology for random fields and complexes. *IMS Collections. Borrowing SStrength: Theory Powering Applications - A Festschrift for Lawrence D. Brown*, *6*, 124–143.

Ashburner, J., Friston, K. J., & Penny, W. D. (2012). *SPM*. London, UK.

Bartsch, A. J., Homola, G., Biller, A., Solymosi, L., & Bendszus, M. (2006). Diagnostic functional MRI: illustrated clinical applications and decision-making. *Journal of magnetic resonance imaging : JMRI*, *23*(6), 921–32.

Beckmann, C. F., & Smith, S. M. (2004). Probabilistic independent component analysis for functional magnetic resonance imaging. *IEEE transactions on medical imaging*, *23*(2), 137–152.

Bellec, P., Rosa-Neto, P., Lyttelton, O. C., Benali, H., & Evans, A. C. (2010). Multi-level bootstrap analysis of stable clusters in resting-state fMRI. *NeuroImage*, *51*(3), 1126–39.

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, Statistical methodology*, *57*(1), 289–300.

Benjamini, Y., & Hochberg, Y. (2000). On the Adaptive Control of the False Discovery Rate in Multiple Testing With Independent Statistics. *Journal of Educational and Behavioral Statistics*, *25*(1), 60–83.

Bennett, C. M., Baird, A. A., Miller, M. B., & Wolford, G. L. (2009). Neural correlates of interspecies perspective taking in the post-mortem atlantic salmon: an argument for proper multiple comparisons corrections. In *15th annual meeting of the organisation for human brain mapping*.

Bennett, C. M., Wolford, G. L., & Miller, M. B. (2009). The principled control of false positives in neuroimaging. *Social cognitive and affective neuroscience*, *4*(4), 417–22.

Biswal, B. B., Mennes, M., Zuo, X.-N., Gohel, S., Kelly, C., Smith, S. M., ... Milham, M. P. (2010). Toward discovery science of human brain function. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(10), 4734–9.

Bonferroni, C. (1936). Teoria statistica delle classi e clcolo delle probabilita. *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze*, *8*, 3–62.

Broberg, P. (2005). A comparative review of estimates of the proportion unchanged genes and the false discovery rate. *BMC bioinformatics*, *6*, 199.

Bullmore, E., Brammer, M., Williams, S. C., Rabe-Hesketh, S., Janot, N., David, a., ... Sham, P. (1996). Statistical methods of estimation and inference for functional MR image analysis. *Magnetic resonance in medicine : official journal of the Society of Magnetic Resonance in Medicine / Society of Magnetic Resonance in Medicine*, *35*(2), 261–77.

Button, K. S., Ioannidis, J. P. a., Mokrysz, C., Nosek, B. a., Flint, J., Robinson, E. S. J., & Munafò, M. R. (2013). Power failure: why small sample size undermines the reliability of neuroscience. *Nature reviews. Neuroscience*, *14*(5), 365–76.

Chen, S., Wang, C., Eberly, L. E., Caffo, B. S., & Schwartz, B. S. (2009). Adaptive control of the false discovery rate in voxel-based morphometry. *Human brain mapping*, *30*(7), 2304–11.

Chumbley, J. R., & Friston, K. J. (2009). False discovery rate revisited: FDR and topological inference using Gaussian random fields. *NeuroImage*, *44*(1), 62–70.

Chumbley, J. R., Worsley, K., Flandin, G., & Friston, K. J. (2010). Topological FDR for neuroimaging. *NeuroImage*, *49*(4), 3057–64.

DeCharms, R. C. (2008). Applications of real-time fMRI. *Nature reviews. Neuroscience*, *9*(9), 720–9.

Delongchamp, R. R., Bowyer, J. F., Chen, J. J., & Kodell, R. L. (2004). Multiple-testing strategy for analyzing cDNA array data on gene expression. *Biometrics*, *60*(3), 774–82.

Desmond, J. E., & Glover, G. H. (2002). Estimating sample size in functional MRI (fMRI) neuroimaging studies: statistical power analyses. *Journal of neuroscience methods*, *118*(2), 115–28.

Dolan, R. J. (2008). Neuroimaging of cognition: past, present, and future. *Neuron*, *60*(3), 496–502.

Dudoit, S., Shaffer, J. P., & Block, J. C. (2003). Multiple Hypothesis Testing in Microarray Experiments. *Statistical Science*, *18*(1), 71–103.

Durnez, J., & Degryse, J. G. U. (2015). NeuroPower. *R-News*.

Durnez, J., Moerkerke, B., Bartsch, A. J., & Nichols, T. E. (2013). Alternative-based thresholding with application to presurgical fMRI. *Cognitive, affective & behavioral neuroscience*, *13*(4), 703–13.

Durnez, J., Moerkerke, B., & Nichols, T. E. (2014). Post-hoc power estimation for topological inference in fMRI. *NeuroImage*, *84*, 45–64.

Durnez, J., Roels, S. P., & Moerkerke, B. (2014). Multiple testing in fMRI: An empirical case study on the balance between sensitivity, specificity, and stability. *Biometrical journal. Biometrische Zeitschrift*, *00*, 1–13.

Dwass, M. (1957). Modified randomization tests for nonparametric hypotheses. *Annals of Mathematical Statistics*, *28*, 181–187.

Efron, B. (2004). Large-Scale Simultaneous Hypothesis Testing: The Choice of a Null Hypothesis. *Journal of the American Statistical Association*, *99*(465), 96–104.

Eklund, A., Andersson, M., Josephson, C., Johannesson, M., & Knutsson, H. (2012). Does parametric fMRI analysis with SPM yield valid results? An empirical study of 1484 rest datasets. *NeuroImage*, *61*(3), 565–78.

Forman, S. D., Cohen, J. D., Fitzgerald, M., Eddy, W. F., Mintun, M. A., & Noll, D. C. (1995). Improved Assessment of Significant Activation in Functional Magnetic Resonance Imaging (fMRI): Use of a Cluster-Size Threshold. *Magnetic Resonance in Medicine*, *33*(5), 636–647.

Friman, O., Borga, M., Lundberg, P., & Knutsson, H. (2003). Adaptive analysis of fMRI data. *NeuroImage*, *19*(3), 837–845.

Friman, O., & Westin, C.-F. (2005). Resampling fMRI time series. *NeuroImage*, *25*(3), 859–67.

Friston, K. J. (2012). Ten ironic rules for non-statistical reviewers. *NeuroImage*, *61*(4), 1300–10.

Friston, K. J., Ashburner, J., Kiebel, S. J., Nichols, T. E., & Penny, W. D. (2007). *Statistical parametric mapping: the analysis of functional brain images.*

Friston, K. J., Frith, C. D., Liddle, P. F., & Frackowiak, R. S. (1991). Comparing functional (PET) images: the assessment of significant change. *Journal of cerebral blood flow and metabolism*, *11*(4), 690–9.

Friston, K. J., Holmes, A., Poline, J. B., Price, C. J., & Frith, C. D. (1996). Detecting activations in PET and fMRI: levels of inference and power. *NeuroImage*, *4*(3 Pt 1), 223–35.

Friston, K. J., Zarahn, E., Josephs, O., Henson, R. N., & Dale, a. M. (1999). Stochastic designs in event-related fMRI. *NeuroImage*, *10*(5), 607–19.

Geissler, A., Gartus, A., Foki, T., Tahamtan, A. R., Beisteiner, R., & Barth, M. (2007). Contrast-to-noise ratio (CNR) as a quality parameter in fMRI. *Journal of magnetic resonance imaging : JMRI*, *25*(6), 1263–70.

Genovese, C. R., Lazar, N. a., & Nichols, T. E. (2002). Thresholding of statistical maps in functional neuroimaging using the false discovery rate. *NeuroImage*, *15*(4), 870–8.

Glaser, D., & Friston, K. J. (2007). Covariance components. In K. J. Friston, J. Ashburner, S. J. Kiebel, T. E. Nichols, & W. D. Penny (Eds.), *Statistical parametric mapping* (pp. 140–147). Amsterdam: Elsevier.

Gordon, A., Chen, L., Glazko, G., & Yakovlev, A. (2009). Balancing Type One and Two Errors in Multiple Testing for Differential Expression of Genes. *Computational statistics & data analysis*, *53*(5), 1622–1629.

Gordon, A., Glazko, G., Qiu, X., & Yakovlev, A. (2007). Control of the mean number of false discoveries, Bonferroni and stability of multiple testing. *Annals of Applied Statistics*, *1*(1), 179–190.

Gorgolewski, K. J., Storkey, A. J., Bastin, M. E., & Pernet, C. R. (2012). Adaptive thresholding for reliable topological inference in single subject fMRI analysis. *Frontiers in Human Neuroscience*, *6*(August), 1–14.

Haller, S., & Bartsch, A. J. (2009). Pitfalls in FMRI. *European radiology*, *19*(11), 2689–706.

Hayasaka, S. (2003). Validating cluster size inference: random field and permutation methods. *NeuroImage*, *20*(4), 2343–2356.

Hayasaka, S., Peiffer, A. M., Hugenschmidt, C. E., & Laurienti, P. J. (2007). Power and sample size calculation for neuroimaging studies by non-central random field theory. *NeuroImage*, *37*(3), 721–30.

Hayasaka, S., Phan, K. L., Liberzon, I., Worsley, K. J., & Nichols, T. E. (2004). Nonstationary cluster-size inference with random field and permutation methods. *NeuroImage*, *22*(2), 676–87.

Heller, R., Stanley, D., Yekutieli, D., Rubin, N., & Benjamini, Y. (2006). Cluster-based analysis of FMRI data. *NeuroImage*, *33*(2), 599–608.

Henson, R. N. a., Shallice, T., Gorno-Tempini, M. L., & Dolan, R. J. (2002). Face repetition effects in implicit and explicit memory tests as measured by fMRI. *Cerebral cortex*, *12*(2), 178–86.

Hoenig, J. M. H., & Eisey, D. M. H. (2001). The Abuse of Power : The Pervasive Fallacy of Power Calculations for Data Analysis. *The Americal Statistician*, *55*(1), 1–6.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, *6*(2), 65–70.

Hommel, G., & Bretz, F. (2008). Aesthetics and power considerations in multiple testing–a contradiction? *Biometrical journal. Biometrische Zeitschrift*, *50*(5), 657–66.

Homola, G. a., Jbabdi, S., Beckmann, C. F., & Bartsch, A. J. (2012). A brain network processing the age of faces. *PloS one*, *7*(11), e49451.

Ioannidis, J. P. a. (2005). Why most published research findings are false. *PLoS medicine*, *2*(8), e124.

Jenkinson, M., Beckmann, C. F., Behrens, T. E. J., Woolrich, M. W., & Smith, S. M. (2012). Fsl. *NeuroImage*, *62*(2), 782–90.

Johnson, T. D., Liu, Z., Bartsch, A. J., & Nichols, T. E. (2012). A Bayesian non-parametric Potts model with application to pre-surgical FMRI data. *Statistical methods in medical research*.

Kirk, R. E. (1996). Practical significance: a concept whose time has come. *Educational and psychological measurement*, *56*, 746—-759.

Kober, H., & Wager, T. D. (2010). Meta-analysis of neuroimaging data. *Wiley Interdisciplinary Reviews: Cognitive Science*, *1*(2), 293—-300.

Korn, E., Troendle, J., Mcshane, L., & Simon, R. (2004). Controlling the number of false discoveries: application to high-dimensional genomic data. *Journal of Statistical Planning and Inference*, *124*(2), 379–398.

Kriegeskorte, N., Goebel, R., & Bandettini, P. A. (2006). Information-based functional brain mapping. *Proceedings of the National Academy of Sciences of the United States of America*, *103*(10), 3863–8.

Kriegeskorte, N., Simmons, W. K., Bellgowan, P. S. F., & Baker, C. I. (2010). Circular Analysis in systems neuroscience - the dangers of double dipping. *Nature Neuroscience*, *12*(5), 535–540.

Lahiri, S. (2003). *Resampling methods for dependent data*. Springer.

Lee, M.-L. (2004). *Analysis of microarray gene expression data*. Kluwer,Boston.

Lieberman, M. D., & Cunningham, W. a. (2009). Type I and Type II error concerns in fMRI research: re-balancing the scale. *Social cognitive and affective neuroscience*, *4*(4), 423–8.

Logan, B. R., Geliazkova, M. P., & Rowe, D. B. (2008). An evaluation of spatial thresholding techniques in fMRI analysis. *Human brain mapping*, *29*(12), 1379–89.

Logan, B. R., & Rowe, D. B. (2004). An evaluation of thresholding techniques in fMRI analysis. *NeuroImage*, *22*(1), 95–108.

Marchini, J., & Presanis, A. (2004). Comparing methods of analyzing fMRI statistical parametric maps. *NeuroImage*, *22*(3), 1203–13.

Meehl, P. E. (1978). Theoretical Risks and Tabular Asterisks: Sir Karl, Sir Ronald, and the Slow Progress of Soft Psychology. *Journal of Consulting an Clinical Psychology*, *46*, 806–834.

Moerkerke, B., Goetghebeur, E., De Riek, J., & Roldan-Ruiz, I. (2006). Significance and impotence: towards a balanced view of the null and the alternative hypotheses in marker selection for plant breeding. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *169*(1), 61–79.

Mumford, J. A. (2012). A power calculation guide for fMRI studies. *Social cognitive and affective neuroscience*, *7*(6), 738–42.

Mumford, J. A., & Nichols, T. E. (2008). Power calculation for group fMRI studies accounting for arbitrary design and temporal autocorrelation. *NeuroImage*, *39*(1), 261–8.

Nichols, T. E. (2012). Multiple testing corrections, nonparametric methods, and random field theory. *NeuroImage*, 10–14.

Nichols, T. E., & Hayasaka, S. (2003). Controlling the familywise error rate in functional neuroimaging: a comparative review. *Statistical Methods in Medical Research*, *12*(5), 419–446.

Nichols, T. E., & Holmes, A. P. (2002). Nonparametric permutation tests for functional neuroimaging: a primer with examples. *Human brain mapping*, *15*(1), 1–25.

Owen, A. B. (2005). Variance of the number of false discoveries. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *67*(3), 411–426.

Perone Pacifico, M., Genovese, C., Verdinelli, I., & Wasserman, L. (2004). False Discovery Control for Random Fields. *Journal of the American Statistical Association*, *99*(468), 1002–1014.

Poldrack, R. (2012). The future of fMRI in cognitive neuroscience. *Neuroimage*, *62*(2), 1216–1220.

Poldrack, R., Mumford, J. A., & Nichols, T. E. (2011). *Handbook of functional MRI data analysis*.

Poldrack, R. A., Fletcher, P. C., Henson, R. N., Worsley, K. J., Brett, M., & Nichols, T. E. (2008). Guidelines for reporting an fMRI study. *NeuroImage*, *40*(2), 409–14.

Pounds, S., & Cheng, C. (2004). Improving false discovery rate estimation. *Bioinformatics (Oxford, England)*, *20*(11), 1737–45.

Pounds, S., & Morris, S. W. (2003). Estimating the occurrence of false positives and false negatives in microarray studies by approximating and partitioning the empirical distribution of p-values. *Bioinformatics*, *19*(10), 1236–1242.

Qiu, X., Xiao, Y., Gordon, A., & Yakovlev, A. (2006). Assessing stability of gene selection in microarray data analysis. *BMC bioinformatics*, *7*, 50.

Quinlan, P. T. (2013). Misuse of power : in defence of small-scale science. *Nature reviews. Neuroscience*, *237*, 2013.

R Core Team. (2014). *R: A language and environment for statistical computing*. Vienna, Austria: R foundation for statistical computing.

Richardson, M. P., Strange, B. a., Thompson, P. J., Baxendale, S. a., Duncan, J. S., & Dolan, R. J. (2004). Pre-operative verbal memory fMRI predicts post-operative memory decline after left temporal lobe resection. *Brain : a journal of neurology*, *127*(Pt 11), 2419–26.

Roels, S., Bossier, H., Loeys, T., & Moerkerke, B. (2014). Data-analytical stability of cluster-wise and peak-wise inference in fMRI data analysis. *Journal of Neuroscience Methods*, 1–11.

Schmidt, F., & Hunter, J. (2002). Are there benefits from NHST. *American Psychologist*, *57*(1), 65–66.

Schwartzman, A., Gavrilov, Y., & Adler, R. J. (2011). Multiple testing of local maxima for detection of peaks in 1D. *The Annals of Statistics*, *39*(6), 3290–3319.

Seurinck, R., de Lange, F. P., Achten, E., & Vingerhoets, G. (2011). Mental rotation meets the motion aftereffect: the role of hV5/MT+ in visual mental imagery. *Journal of cognitive neuroscience*, *23*(6), 1395–404.

Skudlarski, P., Constable, R. T., & Gore, J. C. (1999). ROC analysis of statistical methods used in functional MRI: individual subjects. *NeuroImage*, *9*(3), 311–29.

Smith, S. M., Jenkinson, M., Beckmann, C. F., Miller, K., & Woolrich, M. (2007). Meaningful design and contrast estimability in FMRI. *NeuroImage*, *34*(1), 127–36.

Smith, S. M., Jenkinson, M., Woolrich, M. W., Beckmann, C. F., Behrens, T. E. J., Johansen-Berg, H., ... Matthews, P. M. (2004). Advances in functional and structural MR image analysis and implementation as FSL. *NeuroImage*, *23 Suppl 1*, S208–19.

Smith, S. M., & Nichols, T. E. (2009). Threshold-free cluster enhancement: addressing problems of smoothing, threshold dependence and localisation in cluster inference. *NeuroImage*, *44*(1), 83–98.

Spearman, C. (1904). The Proof and Measurement of Association between Two Things. *The American Journal of Psychology*, *15*(1), 72–101.

Storey, J. D. (2002). A direct approach to false discovery rates. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *64*(3), 479–498.

Storey, J. D., & Tibshirani, R. (2003). Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences of the United States of America*, *100*(16), 9440–5.

Thyreau, B., Schwartz, Y., Thirion, B., Frouin, V., Loth, E., Vollstädt-Klein, S., ... Poline, J.-B. (2012). Very large fMRI study using the IMAGEN database: sensitivity-specificity and population effect modeling in relation to the underlying anatomy. *NeuroImage*, *61*(1), 295–303.

Van Essen, D. C., Ugurbil, K., Auerbach, E., Barch, D. M., Behrens, T. E. J., Bucholz, R., ... Yacoub, E. (2012). The Human Connectome Project: a data acquisition perspective. *NeuroImage*, *62*(4), 2222–31.

Van Horn, J. D., Ellmore, T. M., Esposito, G., & Berman, K. F. (1998). Mapping voxel-based statistical power on parametric images. *NeuroImage*, *7*(2), 97–107.

Vul, E., Harris, C., Winkielman, P., & Pashler, H. (2009). Puzzlingly High Correlations in fMRI Studies of Emotion, Personality, and Social Cognition. *Perspectives on Psychological Science*, *4*(3), 274–290.

Wager, T. D., Lindquist, M., & Kaplan, L. (2007). Meta-analysis of functional neuroimaging data: current and future directions. *Social cognitive and affective neuroscience*, *2*(2), 150–8.

Wager, T. D., Lindquist, M. a., Nichols, T. E., Kober, H., & Van Snellenberg, J. X. (2009). Evaluating the consistency and specificity of neuroimaging data using meta-analysis. *NeuroImage*, *45*(1 Suppl), S210–21.

Wager, T. D., & Nichols, T. E. (2003). Optimization of experimental design in fMRI: a general framework using a genetic algorithm. *NeuroImage*, *18*(2), 293–309.

Wasserman, L., & Genovese, C. (2004). A stochastic process approach to false discovery control. *The Annals of Statistics*, *32*(3), 1035–1061.

Welvaert, M., Durnez, J., Moerkerke, B., Verdoolaege, G., & Rosseel, Y. (2011). neuRosim: An R package for generating fmri data. *Journal of Statistical Software*, *44*(10), 1–18.

Woo, C.-W., Krishnan, A., & Wager, T. D. (2014). Cluster-extent based thresholding in fMRI analyses: pitfalls and recommendations. *NeuroImage*, *91*, 412–9.

Worsley, K. J. (2007). Random Field Theory. In K. Friston, J. Ashburner, S. J. Kiebel, T. E. Nichols, & W. Penny (Eds.), *Statistical parametric mapping* (pp. 232–236). London: Academic Press.

Worsley, K. J., Evans, a. C., Marrett, S., & Neelin, P. (1992). A three-dimensional statistical analysis for CBF activation studies in human brain. *Journal of cerebral blood flow and metabolism*, *12*(6), 900–18.

Worsley, K. J., Marrett, S., Neelin, P., Vandal, a. C., Friston, K. J., & Evans, a. C. (1996). A unified statistical approach for determining significant signals in images of cerebral activation. *Human brain mapping*, *4*(1), 58–73.

Worsley, K. J., Taylor, J. E., Tomaiuolo, F., & Lerch, J. (2004). Unified univariate and multivariate random field theory. *NeuroImage*, *23 Suppl 1*, S189–95.

Zarahn, E., & Slifstein, M. (2001). A reference effect approach for power analysis in fMRI. *NeuroImage*, *14*(3), 768–79.

Zehetmayer, S., & Posch, M. (2010). Post hoc power estimation in large-scale multiple testing problems. *Bioinformatics (Oxford, England)*, *26*(8), 1050–6.

Zhang, H., Nichols, T. E., & Johnson, T. D. (2009). Cluster mass inference via random field theory. *NeuroImage*, *44*(1), 51–61.