High-dimensional statistical methods for inter-subject studies in neuroimaging

Virgile Fritsch

University Paris-Sud 11 – Faculty of Sciences
Graduate School of Informatics of Paris-Sud – EDIPS
INRIA Saclay – Parietal Team

PhD Thesis

submitted in partial fulfillment
of the requirements for the degree of
DOCTOR OF SCIENCE
Specialized in Computer Science

defended on December 18th 2013

by

Virgile Fritsch

High-dimensional statistical methods for inter-subject studies in neuroimaging

Advisors	Dr Jean-Baptiste Poline	CEA – Unati group, Saclay, France
	Dr Bertrand Thirion	INRIA Saclay – Parietal team, Saclay, France
Reviewers	Dr Florence Forbes	INRIA Grenoble – Mistis team, Grenoble, France
	Pr Tor Wager	Dept. of Psychology and Neuroscience 345 UCB, Uni. of Colorado, Boulder, USA
Examiners	Pr Alain Denise	LRI, IGM, CNRS, INRIA – AMIB team, and Uni. Paris Sud, Orsay, France
	Dr Christine Keribin	Dept. de Mathématiques, Uni. Paris Sud, Orsay, France
	Dr Roberto Toro	CNRS ura2182, Institut Pasteur, Paris, France

Contents

Introduction

Between-subject variability is a prominent effect in many fields of medical imaging, and particularly in brain imaging. While part of this variability can be viewed as normal fluctuations within a population or across repeated measurements, and can be considered as an effect of interest for diagnosis problems, part of it may be a confound, related to scanner instabilities, experimental issues, or acquisition artifacts. Such confounding factors can be much larger than the effects of interest: for instance, in functional neuroimaging, the variability related to acquisition issues (motion, defective experimental setup, scanner spikes) can mask the true effect of interest, which is the variability in brain functional organization related to diseases, psychological or genetic factors. This can undermine the statistical procedures used in group studies as the latter assume that the cohorts are composed of homogeneous samples with anatomical or functional features clustered around a central mode. As the high-dimensional context prevents manual data screening, some outlier detection methods have to be used to provide an automated decision on subjects inclusion. Yet, it remains unclear whether or not outliers should be removed, and, if so, what tolerance to choose. Alternatively, several outlier-resistant methods has been proposed for statistical inference in neuroimaging, although they are still not widely used. Beyond outlier-resistance, such robust methods seem better adapted to real world data since they also compensate for inexact hypotheses (e.g. data normality, homogeneous dataset). Another –partially related– problem is that the lack of stability and of sensitivity of current voxel-based analysis schemes may lead to non-reproducible results.

In this thesis, we first develop statistically-controlled outlier detection procedures especially designed for neuroimaging data: We propose a regularized version of a robust covariance estimator so as it can be used under high-dimensional settings. We compare several regularization schemes and conclude that random projections offer the best compromise. We also present non-parametric outlier detection procedures and show that their accuracy stands high. However, their use is limited in practice because they do not provide a statistical control on outlier detection. We use them as an explanatory tool that provides insight about the data statistical structure. As a second contribution, we propose a new approach –*Randomized Parcellation Based Inference (RPBI)*– to overcome the lack of reproducibility of standard methods. We stabilize parcel-based analysis by considering several independent analyzes that we aggregate together to obtain a final consensus. The method also conveys more sensitivity than state-of-the-art methods, as demonstrated by experiments on synthetic and real datasets. In our third contribution, we apply robust regression to neuroimaging studies, hereby extending some previous work of another group. We focus on large-scale studies that involve cohorts of more than 100 subjects. On both simulated and real data, we show that the use of robust regression improves the sensitivity of the analysis. We also provide evidence of the importance of being robust to deviations from the model assumptions even after careful outlier detection has been performed. We finally combine robust regression with RPBI to obtain even more sensitive statistical tests.

Organization of the manuscript

Chapter 1 introduces functional magnetic resonance imaging (fMRI). After a short technical description of image acquisition, we present how these can be used to understand brain function. We describe the complete data processing framework that starts with raw MRI images of independent subjects and ends with group-level maps that are interpreted by neuroscientists. We give more details on statistical analysis of the images as the complexity of the latter since this defines the setting of our contributions. Chapter 1 is the opportunity to introduce the problematic of this thesis, in which robust and high-dimensional procedures are investigated and developed in order to improve the sensitivity of neuroimaging analyzes: We explain how deviations from the model assumptions can cause the statistical procedures to break down and discuss some solution in order to address this problem when high-dimensional data are considered.

Chapter 2 gives the theoretical background used in the sequel. It includes basics about hypothesis testing, introduces the multiple comparisons problem and discusses permutation testing. Then, methods that are specific to neuroimaging data analysis are introduced. A section of the chapter is dedicated to covariance estimation (including robust estimation), as we develop a regularized version of a robust covariance estimator in chapter 3. We therefore include some explanations about various penalization options. The last section of chapter 2 provides an overview of non-parametric statistics algorithms and procedures that are used within this thesis. Some of these techniques are employed for the sake of synthetic data generation, for validation purpose or as alternative methods that we compare against ours.

Chapter 3 deals with outlier detection. We modify the classical *Minimum Covariance Determinant* estimator by adding a regularization term, that ensures that the estimation is well-posed in high-dimensional settings and in the presence of many outliers. We demonstrate on functional neuroimaging datasets that outlier detection can be performed with small sample sizes and improves group studies, even in situations where the number of dimensions of the data exceeds the number of observations. We also demonstrate that non-parametric outlier detection procedures have a high accuracy in outlier detection tasks. We propose an efficient procedure to summarize the necessary information about the data structure so that the practitioner can find how many observations to discard.

Chapter 4 introduces a new approach –*Randomized Parcellation Based Anal-*

ysis (RPBI)– to overcome the limitations of standard neuroimaging group analysis methods, in which active voxels are detected according to a consensus on several random parcellations of the brain images, while a permutation test controls the false positive risk. Both on synthetic and real data, this approach shows higher sensitivity, better accuracy and higher reproducibility than state-of-the-art methods. In a neuroimaging-genetic application, we find that it succeeds in detecting a significant association between a genetic variant next to the *COMT* gene and the BOLD signal in the left thalamus for a functional Magnetic Resonance Imaging contrast associated with incorrect responses of the subjects from a *Stop Signal Task* protocol.

Chapter 5 investigates the use of robust regression for neuroimaging analyzes, with an emphasis on large-scale studies. While small-sample size studies can hardly be proved to deviate from standard hypotheses (such as the normality of the residuals) due to the low degrees of freedom of the statistical model, large-scale studies (e.g. on more than 100 subjects) give a different picture and encourage the practitioner to use finer models to perform statistical inference. We demonstrate the benefits of robust regression as a tool for analyzing large neuroimaging cohorts. Our first contribution is to design an analytic test based on robust parameters estimates; this procedure makes it possible to forgo permutation testing and thus to perform whole brain analysis in a reasonable time. Then we demonstrate that robust regression yields sensitivity improvements in two real data examples on 392 and 1502 subjects. We finally show that robust regression can be combined with randomized parcellation based analysis to improve whole-brain tests sensitivity.

Chapter 1
Introduction to functional Magnetic Resonance Imaging studies

Contents

1.1 Imaging the brain with Magnetic Resonance Imaging

1.1.1 Magnetic Resonance Imaging

Magnetic Resonance Imaging (MRI) is a medical imaging technique that is used to observe specific types of tissues in a non-invasive fashion. The first MRI image was acquired in 1976 by Damadian et al. [18], but the technique was approved for clinical use almost ten years later [33]. Regarding brain imaging, MRI was then mainly used to diagnosis neurological disorders such as atrophies related to epilepsy, cancerous tumors, or Alzheimer disease [8]. In the early 90's, MRI started to be used to study brain function [48], giving birth to a new domain called *functional MRI.* It has a good spatial resolution, close to the millimeter. Some alternative functional brain imaging techniques are *(i)* Magnetoencephalography (MEG) that has a temporal resolution of 1 ms, but a poor spatial resolution (more than 1 cm) ; *(ii) transcranial magnetic simulation (TMS)* and *Positron Emission Tomography (PET)* that have characteristics equivalent to MRI but are invasive methods ; *(iii) optical imaging*, that has both a better temporal and spatial resolution than MRI but is only available on animals. Functional MRI is a useful technique for human brain imaging because of its non-invasiveness. Besides, it offers a good compromise between temporal and spatial resolution while providing a full brain coverage. A good spatial resolution is obviously useful to observe the brain with enough details while a good temporal resolution is especially useful to functional neuroimaging (see next section).

1.1.2 Short overview of MR images acquisition

1.1.2.1 Magnetization

Most of the atoms that constitute the human body (including the brain) have *nuclear magnetic resonance (NMR) properties*: they behave like small magnets that spin around there axis and thus have a magnetic momentum. When placed into a constant magnetic field, they start precessing around parallel axes in a gyroscopic motion with varying angle and frequency (the atoms are *magnetized*). At the equilibrium, there exist two states for the atoms at this stage: the parallel state (with a lower energy level) and the antiparallel state (higher energy level). As the strength of the magnetic field increases (unit: *Tesla*, T), more and more atoms are in the parallel state. The difference of the numbers of atoms in parallel and antiparallel state is summarized by a vector that represents the sum of the magnetic momenta of all atoms, the *net magnetization*. It is of crucial importance regarding MRI as its norm directly influences an upper bound on the MR signal that can be captured in the sequel. Thus, 7T scanners provide better quality images (in term of signal-to-noise ratio) than 3T scanners, although the latter are still the current standard in neuroimaging. An important fact is that the precessing frequency of the atoms (called the *Larmor frequency*) depends linearly on the magnetic field strength.

1.1.2.2 Excitation and relaxation

In MR image acquisition, the *excitation* step follows the magnetization step. It consists in using *radio-frequency pulse* at the *resonance frequency* of the magnetized atoms so that the net magnetization oscillates between the value obtained at the equilibrium and its opposite. The exact value depends on the time of the excitation and can be computed exactly for a given type of targeted atoms. Another important effect of excitation is that atoms of the same type (e.g. hydrogen nuclei) precess in-phase once excited (i.e. not only do they have the same precessing frequency, but their magnetic momenta are aligned). Each type of atom has its own resonance frequency, that also depends on the magnetic field in which they are placed. MRI targets hydrogen nuclei as the human body is mostly composed of them. Once their magnetization stops, they come back to their original magnetized state (*relaxation*), delivering back the absorbed energy (i.e. the net magnetization recovers). The amount of released energy is captured by the scanner reception coils and corresponds to the *MR signal*. There are two main components in the MR signal: *(i)* The recovery that correspond to the net magnetization returning to its original state (the *T1 recovery*); *(ii)* The phase decay that corresponds to the atoms' phase loosing their coherence due to small magnetic interactions between the nuclei of the system (the *T2 decay*). Local inhomogeneities in the magnetic field add a supplementary effect that modify the phase decay. *T2* decay* is similar to T2 decay, but takes the latter effect into account. Depending on the nature of the observed features, T1-, T2-, or T2*-weighted images are used.

1.1.2.3 Echo-planar imaging

In order to localize where the energy release comes from (i.e. map the MR signal with spatial locations), magnetic field gradients are applied in three orthogonal directions. We have mentioned that the precessing speed of the magnetized nuclei depends on the strength of the magnetic field they are placed into. By choosing the strength of the gradients properly, one can define a one-to-one correspondence between a spatial volume and the MR signal it generates. In practice, MRI images are acquired as sequences of two-dimensional slices: This technique is called *echo-planar imaging*. Depending on the spatio-temporal configuration of the gradients (the *acquisition sequence*), a whole brain volume can be imaged according to various schemes. Modifying the acquisition sequence can help correcting some artifacts that are due for instance to motion (see section 1.2.2.2) but the details of the realization of acquisition sequences is beyond the scope of this thesis. We refer the reader to the excellent book by S.

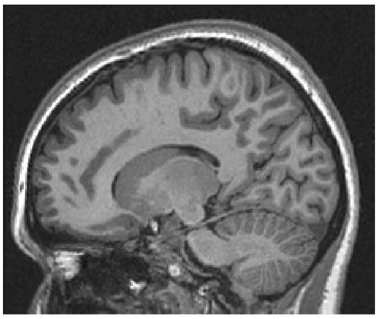


Figure 1.1: **3T structural (T1-weighted) image.** Gray matter is clearly visible on the brain and cerebellum contours and in some part of the middle-brain. White matter is in light gray. Cerebrospinal fluid is in dark black between the brain and the skull. One can note the high resolution (1mm) as compared to the EPI sequence shown in Figure 1.3.

A. Huettel [41].

1.1.3 Structural MRI

Structural MRI provides the so-called *anatomical images* on which one can observe the different tissues that constitute the brain (see Figure 1.1). *Gray matter* corresponds to the *synapses* (i.e. heads) and neuron bodies of the billions neurons that are implanted all around the brain (and in some inner nuclei as well). The gray matter is a layer of 2-3mm thickness (depending on the location) that is also called *cortex*. The *axons* (kind of ramified stems) of the neurons dive into the depth of the brain and constitute the *white matter*, as they are covered with white *myelin*. The brain is surrounded by the *cerebrospinal fluid (CSF)*, which separates it from the skull. All these anatomical items can be observed on Figure 1.1. Structural images can be used in longitudinal studies, where the evolution of an individual's brain is observed over a long period of time (up to several years) in several scans. They are also used for diagnosis of epilepsy, or localize cancerous tumors.

1.1.4 Utility to neurosciences

Before the first brain imaging techniques arose, human brain anatomy could only be investigated ex vivo, and brain function could only be observed indirectly by comparing e.g. stroke patients with healthy subjects. In the absence of better experimental protocols, only limited scientific results could be obtained. Most of the actual knowledge about the brain was developed from invasive experimentations on animals, such as electrophysiology on primates [1, 68] or rodents [34]. With the emergence of imaging techniques, and especially non-invasive ones, cohorts of human subjects have been constructed in order to study specific aspects of brain structure or function.

1.2 Analysis of functional MRI data

Cognitive functional neurosciences observe differences of brain activity under varying experimental conditions. The differences can be observed at various time scales, and by various means, including brain imaging. The latter case correspond to *functional neuroimaging*; which is probably the most powerful and promising technique for cognitive neuroscience [56]. In this work, we focus on functional neuroimaging. We have seen that whereas T1-weighted MRI creates neat anatomical images, other types of images can be obtained using alternative properties of the MR signal (e.g. T2- and T2*-weighted images).

1.2.1 Event-related fMRI

1.2.1.1 The BOLD signal

At rest, the oxygen of the blood that comes to the brain is consumed, which transforms the incoming hemoglobin into deoxygenated hemoglobin ; when a cognitive task is performed, we observe an overcompensation of blood flow up to the neurons, resulting in a decrease of the proportion of deoxygenated hemoglobin [60]. This decrease is called the *blood-oxygen-level dependent (BOLD) response*. It is observable on T2*-weighted MRI images. It is interpreted as a correlate of neurons synaptic activity [7], and therefore occur into the gray matter as we have seen is corresponds to the neurons' synapses. In practice, the BOLD response is occurring with a delay, and the corresponding BOLD signal as captured by MRI has its specific pattern, defined by the *Haemodynamic Response Function (HRF)* illustrated in Figure 1.2. Most of the models used in practice consider a canonical HRF [35], although the BOLD response slightly varies according to individuals and brain location [39, 3]. This thesis specifically considers *BOLD functional MRI* [48, 61] (i.e. functional MRI based on the BOLD signal), although we use the more general name *functional MRI (fMRI)* in the sequel.

1.2.1.2 Experimental setup

Functional MRI based on BOLD signal offers a good compromise between the resolution of the images and the acquisition time, hereby providing some measurements of relevant variations in the brain state across experimental conditions. A particular case of functional MRI is *resting state fMRI*, where spontaneous activations patterns are investigated. The main difficulty of resting state MRI is that there is no controlled event that can be associated with signal variations, i.e. there is no difference between brain activity and noise in terms of effect size. This thesis only considers event-related fMRI, that consists in the observations of BOLD signal variations in response to timed experimental events such as finger-tipping, sounds listening, or viewing faces. Simple cognitive tasks are thus performed by a subject into the MRI scanner while the BOLD signal fluctuations are recorded as T2*-weighted images sequences. Every confounding parameter is controlled so that only the performance of the task is supposed to create the measured BOLD signal fluctuations (e.g. the subject can be asked to close his eyes while performing a listening task, in order to be sure that no spurious visual event would pollute its mental state). More complex confounding factors such as breathing or heart beat can be recorded too and used in image preprocessings (see section 1.2.2.2. The ultimate goal of functional MRI is to be able to match the mental state of subjects with their behavior, their clinical status (diseased or not) and some genetic variable that could cause the observed difference. There is therefore a need to define population standards as reference/control measures. In neuroscience, they take the form of population-level brain maps that result from a two-stage statistical analysis. This is based on an

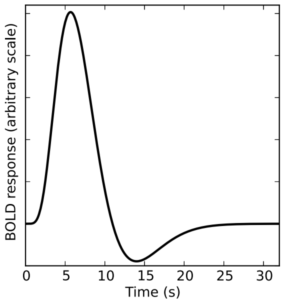


Figure 1.2: **Canonical haemodynamic response function as defined by Glover [35]**. It corresponds to the timed BOLD response relative to a discrete event occurring at "time = 0". A peak occurs approximately 5 seconds after the event. Between 10 seconds and 15 seconds from the original event, an undershoot of the BOLD response is observed. The HRF is taken into account in event-related designs via a convolution with the design matrix (see section 1.2.2.4).

intra-subject analysis, the results of which are embedded in a subsequent inter-subject analysis. The next sections describe the standard statistical framework associated with fMRI neuroimaging studies.

1.2.2 Intra-subject (or first-level) analysis

1.2.2.1 Input images and their characteristics

The raw functional images coming out from the scanner are organized as 4D volumes, i.e. a set of voxel time courses arranged in a 3D array. Each volume represents the BOLD signal at a precise time stamp. The sampling rate of the time serie lies between one and three seconds, instead of about ten minutes for the acquisition of an anatomical image. This comes at the cost of a poorer resolution: 2-3 mm versus less than 1 mm. An example of fMRI sequence is given in Figure 1.3. At the 3T resolution, each fMRI volume is a grid of $\simeq 200,000$ voxels, that can be masked to focus the analysis to the voxels that actually are part of the brain ($\simeq 60,000$ voxels). The most important characteristic of functional MRI images is their *signal-to-noise* ratio, that is the relative strength of a signal compared with other sources of variability in the data. For an effect to be detected, the amplitude of the task-related BOLD signal has to be larger than that of non-task-related variability, the so-called *noise*. Noise can come from Magnetic Resonance Imaging itself, independently of brain imaging. This includes *thermal noise* (noise related to temperature changes within the scanner electronics or the imaged object) and *system noise* (variations due to the imaging hardware, e.g. slow changes in voxel intensity over time known as *scanner drift*). In functional neuroimaging, additional noise sources are observed: *motion artifacts* and *physiological noise*. Motion artifacts can be reduced when preprocessing the data (see next section), while physiological noise reflects natural variations in the regional brain blood flow level related to cardiac rhythm fluctuations arousal. It is important to notice that the subject images include some variability across trials that can be of two orders: task-related variability and non-task-related variability. The first one is a variability in terms of direct performance of the task (varying time response, accommodation to the task), while the second one comes from differences of mental state that can occur across trials (e.g. fluctuations of attention, tiredness or pain). There exist specific methods to accommodate or prevent these variability sources, although the inter-subject variability is much more of an issue regarding reproducibility of the results (see next section).

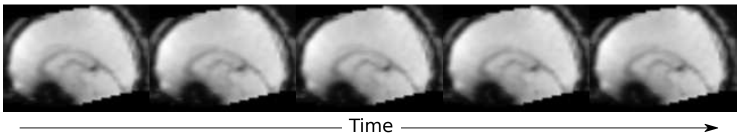


Figure 1.3: **A BOLD fMRI sequence.** Small changes between images cannot be observed by eye. Only statistical analysis can reveal significant changes and relate them to the experimental conditions. The data have already been preprocessed.

1.2.2.2 Image preprocessing

There are some corrections that need to be applied to raw functional MRI sequences prior to data analysis. First, raw images can contain artifacts, defined as "features appearing in an image that are not present in the original object" [22]. Second, some processing is necessary to improve the quality of the images and improve subsequent analysis. Just as a photographer would apply corrections, filters and magnify raw photographs, the signal-to-noise ratio of the raw MRI images can be improved by a set of specific processing. This section introduces the most common artifacts and distortions but do not systematically explain how these can be corrected, for most of them involve a lot of technical details that are beyond the scope of this manuscript.

Artifacts and distortions can be categorized into three groups:

Scanner-related artifacts/distortions are of various types and result from hardware failures or imprecision. *Scanner inhomogeneities* [42] is the most famous example of scanner-related artifact, they are due to the difficulty to maintain perfectly stable and homogeneous high magnetic fields within the scanner. The constant magnetic field as well as the gradient magnetic fields (see section 1.1.2 are subject to inhomogeneities [62]. Radiofrequences used at the excitation time also yield artifacts: They may also be inhomogeneous and they are sensitive to interferences with the neighboring hardware (e.g. monitors or computers) [19].

Signal-processing artifacts/distortions [11] depends on the acquisition scheme. In echo-planar imaging (see section 1.1.2), *slice-timing correction* [74] is intended to drop the discrepancies that come from the fact that different slices of each volume are imaged at a different times (with differences of the order of the second) and would therefore correspond to different response levels. *Partial volume artifacts* correspond to the disappearance of an object because it is smaller than the size of an image voxel [84]. *Chemical shift artifacts* and *magnetic susceptibility artifacts* come from the fact that the resonance frequency of the hydrogen nuclei varies slightly according to the type of tissue on which they are [49]. As a result, spurious layers can appear at the interface of different tissues. *Ringing artifacts* [64] are caused by the *Gibbs phenomenon*, which occurs at the image discontinuous intensity regions. That phenomenon was first describe in the context of Fourier analysis, which we think is a good idea to read about in order to understand the corresponding MRI artifact into details[36].

Patient-related artifacts/distortions are caused by the imaged subject. *Motion correction* [30] is one of the most important preprocessing steps to take into account potential movements of the subject. Some of these movements are related to breathing or heart beat, while some are mere movements that are due to the difficulty for the subject to remain steady during the whole acquisition. To correct for subject motion, the volumes are registered to a reference slice in order to improve the spatial correspondence between voxels. Ideally, there should be a perfect correspondence, but this is not the case in practice since motion cinetics is only approximated and does not accommodate non-rigid deformations related to field inhomogeneities [93, 30]. One common practice is to register functional images to an anatomical image of the subject. This is necessary to relate function to anatomical structures and it is a first step towards the *spatial normalization* [27] of the images (i.e. registration to a reference stereotaxic space and scaling of the brain size). *Spatial normalization* is presented in the next subsection, as it is a key point to group analysis [46]. There are other types of patient-related artifacts than motion. Ferromagnetic medical devices or make-up can create *metal artifacts* because they locally distort the magnetic field [54]. Finally, liquid flows within the human body (e.g. blood flow) also alter local magnetic fields and may yield artifacts [20].

It is possible to apply a Fourier transform to a voxel's time course in order to obtain a frequency-domain representation of that time course. In this domain, one can easily observe a large power at the frequencies corresponding to the task-related events, while some fluctuations correspond to noise frequencies. Thus, one can identify low- and high-pass filters that can be applied to the original image in order to improve the signal-to-noise ratio. Such a filtering is called *temporal filtering*. *Spatial filtering* is also a standard preprocessing: it is used to spatially smooth the images in order to maximize to signal-to-noise ratio. In practice, a Gaussian filter is applied to the images, where the smoothing extent is controlled by the bandwidth of the filter. The matched filter theorem [92] states that the optimal bandwidth corresponds to the size of the observed effect. The pros and cons of smoothing are discussed in chapter 2.

1.2.2.3 Spatial normalization

Although optional in intra-subject studies, spatial normalization makes it possible to use the subject image in inter-subject analyzes. The images of one subject are registered within the same stereotaxic space. The most famous one is the *Talairach space* [81], that was defined from a set of anatomical landmarks of one particular brain. Importantly, these landmarks were chosen by Jean Talairach as being the most reproducible ones across individuals. The Montreal Neurological Institute (MNI) derived a more stable template by registering and averaging the structural brain images of more than one hundred subjects [23]. The procedure to register one brain image to the template is done in four steps: *(i)* a translation so that the point at the middle of the anterior commisure is at the origin of the stereotaxic space ; *(ii)* a rotation so that the inter-hemispheric plane is orthogonal to the y axis (with positive values of the z axis corresponding to the top of the skull) ; *(iii)* a rotation so that the middle point of the posterior commisure lies on the x axis ; *(iv)* a scaling of the brain according to others specific anatomical landmarks so that the brain as a normalized size. There exist

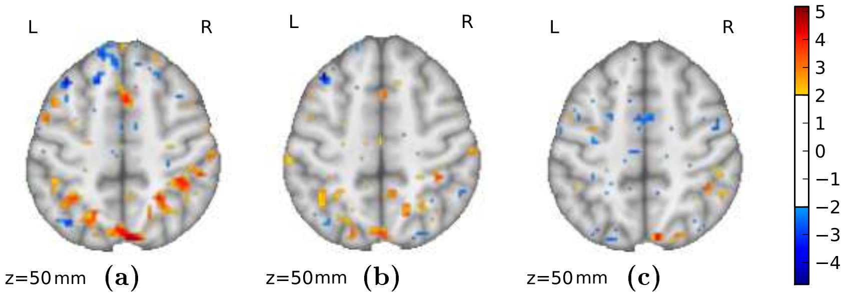


Figure 1.4: **Contrast maps associated with a computation task (axial cut, z maps).** The arithmetic operations were presented as written instructions. **(a)** The activation map associated with the complete computation task; **(b)** The activation map associated with sentence reading ; **(c)** The contrast map associated with the difference between the complete computation task and sentence reading.

a lot of registration algorithms that can register brain images to a pre-defined template (see for example [79] for a review). Some of them require user intervention [53], some others use non-rigid transformations [46], other approaches use surface-based landmarks [25] or high-level landmarks such as sulci [15].

1.2.2.4 Contrasts estimation

As we already mentioned, the goal of functional MRI is to capture (and explain) task-specific brain activity patterns. These can be observed as differences between two experimental conditions. For example, a computation task not only involves brain networks that are specific to arithmetic operations, but also activates the visual or auditory network (depending on how the stimuli are presented to the subject, see Figure 1.4). Thus, brain activity patterns related to computation are observed as the difference between a computation task and a visualization (or listening) task, as illustrated in Figure 1.4. In a functional MRI protocol, the subjects involved generally perform several cognitive tasks, that need to be combined in order to yield interpretable activity maps, the so-called *contrast maps*.

Formally, the first step towards contrast maps estimation is to relate the experimental events with the subject's time course images with a *General Linear Model (GLM)* [29]:

$$\boldsymbol{Y} = \boldsymbol{X}_1\boldsymbol{\beta} + \boldsymbol{\epsilon}_1. \tag{1.1}$$

$\boldsymbol{Y}$ is a $(t \times v)$ matrix that encodes the t images of the time course of v voxels each (note that a subject-specific brain mask is generally applied to the images in order to reduce the number of voxels). $\boldsymbol{\beta}$ is the $(m \times v)$ coefficients matrix that has to be estimated. $\boldsymbol{X}_1$ is the *design matrix* of shape $(t \times m)$ that encodes the m experimental conditions (e.g. binary variates that indicate the presence or the absence of an experimental event such as "sound heard" or "button pressed"). The design matrix is convolved with the *Haemodynamic Response Function (HRF)* filter (see Figure 1.5) that relates the synaptic activity that causes the BOLD response to the observed blood flow variations. $\boldsymbol{\epsilon}_1$ is a $t \times v$ noise model, with $\boldsymbol{\epsilon}_{1_{(\cdot,i)}} \sim \mathcal{N}(\boldsymbol{0}, \sigma^2\boldsymbol{V}_i)$, where $\boldsymbol{\epsilon}_{1_{(\cdot,i)}}$ is the vector that corresponds to the ith column of $\boldsymbol{\epsilon}_1$. A first step is to fit the linear model of Equation 1.1, yielding estimates of the model coefficients $\hat{\boldsymbol{\beta}}$ and the noise scale

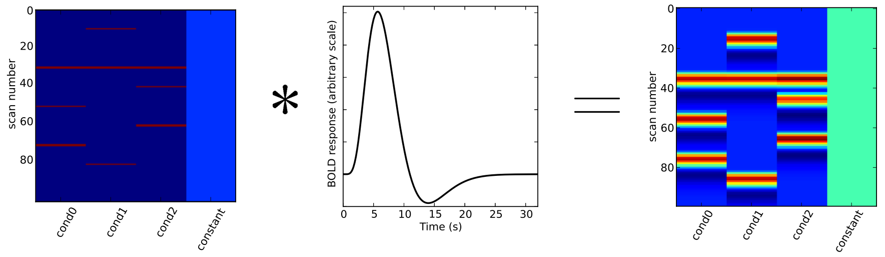


Figure 1.5: **Design matrix and Haemodynamic Response Function (HRF).** The design matrix (left) encodes the experimental conditions and their occurrence over time. The HRF helps taking into account the neurons response delay and the spontaneous variations of the BOLD signal. It is integrated to the design matrix via a convolution.

$\hat{\sigma}$. The differences between experimental conditions arise when testing for the statistical significance of the estimated model coefficients: one uses a *contrast vector* $\boldsymbol{c}$ to perform a statistical test on a combination of variates rather than a simple test on one single coefficient. For instance, let us assume than the first column of the design matrix corresponds to viewing letters on a screen while the second column corresponds to performing a computation task from visually presented stimuli. To obtain the contrast map corresponding to the specific activity pattern of performing an arithmetic operation, we compute the F-statistic at the ith voxel:

$$F_i = \frac{Tr(\boldsymbol{c}\hat{\boldsymbol{\beta}}_{(\cdot,i)}\hat{\boldsymbol{\beta}}^{\mathsf{T}}_{(\cdot,i)}\boldsymbol{c}^{\mathsf{T}})}{\hat{\sigma}^2 Tr(\boldsymbol{c}^{\mathsf{T}}(\boldsymbol{X}_1^{\mathsf{T}}\boldsymbol{V}_i\boldsymbol{X}_1)^{-1}\boldsymbol{c})}, \tag{1.2}$$

with $\boldsymbol{c} = (-1, 1, 0, \ldots)$. The F-statistic measures the signal-to-noise ratio. A t-statistic can be obtained in the same fashion. A common practice is to convert the test statistics into z-values (i.e. a normalized statistic). Such *decision statistics* reflect an effect size, that has to be compared to what would be obtained by chance (i.e. if no effect were present). Rather than the size of an effect, neuroscientists are concerned with its *significance*[1], which is measured by a conversion of F- or t-statistics into p-values. The theoretical decision of the distribution statistics varies according to the number of covariables in the design matrix, m. The quantity $t - m$ is called the *degrees of freedom* of the model and should be greater than 1.

1.2.3 Inter-subject (or second-level) analysis

Group comparisons is one of the most powerful scientific approach to assess the effect of a given feature on brain function. Indeed, pattern variations in the activation maps of several subjects can be associated with an external variate. If groups can be defined according to that variate, it is easy to perform statistical tests that confirm the source of the pattern variations. The more subjects, the more powerful is the method. *Inter-subjects* (or *second-level*) analysis aim at uncovering such relationships between imaging features and various types of other variates, that can be discrete or continuous.

[1] Although this practice might not be the only relevant practice regarding the analysis of biological data, as discussed in [58].

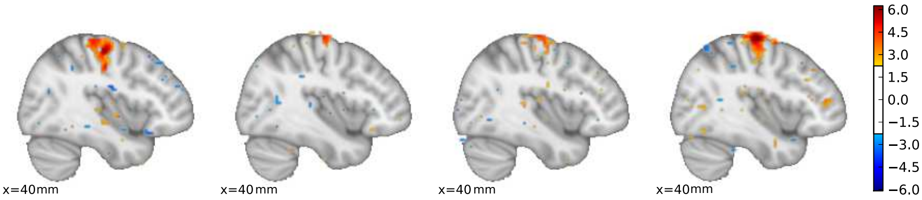


Figure 1.6: **Contrast maps associated with a motor task in different subjects (sagittal cut, z maps, *[left-right]* contrast).** The contrast maps are all different, although similar activations are observed. *Spatial smoothing* is the most straightforward response in order to improve the between-subject correspondence (not applied here).

1.2.3.1 Variability of neuroimaging data

Two types of variability can be distinguished between subjects: the anatomical variability [32] and the functional variability [67].

Anatomical variability corresponds to a variation of brain size, shape, folding and structure across subjects. A better between-subject correspondence is attained by considering surface-based registration of the subject anatomical images [25] (see section 1.2.2.3). The residual variability is commonly considered to be as large as 1cm. Smoothing the images or considering local signal averages as images descriptors improves again the anatomical correspondence but results in a loss of statistical power for the subsequent functional analysis and the amount of smoothing needs to be adapted to the analysis method (see for instance [92]).

Functional variability corresponds to between-subject variations in the performance of a given cognitive task. It has been shown that various cognitive strategies could lead to the same result, as in the well-known reading task example, where a subject can read a word by recognizing it as a whole, or by considering its constitutive letters one after each other [13]. In both cases, the task is performed correctly, but the existence of alternative strategies is reflected on fMRI contrast maps by distinct activation patterns, as illustrated in Figure 1.6. Without being so extreme, small differences can arise between subjects, especially when complex tasks are performed. As a result, inter-subject analysis require fine-tuned statistical procedures that model various levels of variance [90, 70].

Surface-based analyzes increase the inter-subject overlap of active regions and are more sensitive [85], but it is probably only because they reduce the anatomical variability.

1.2.3.2 Statistical inference at the population level

Let us assume that we have the contrast images of several subjects, and that these images correspond to the same underlying task. *Second-level analyzes*, or *group studies* aim at finding relationships between imaging features (the *targets*) and non-imaging variates (the *covariates*), conditionally to other optional variates (the *confounds*). Here again, a linear model is considered:

$$\boldsymbol{B} = \boldsymbol{X}_2\boldsymbol{\gamma} + \boldsymbol{\epsilon}_2, \tag{1.3}$$

where $\boldsymbol{X}_2$ is the second-level design matrix (that encodes the covariates and the confounds), $\boldsymbol{\gamma}$ is the matrix of the model coefficients, $\boldsymbol{B}$ is a $n \times v$ matrix corresponding to first-level analysis maps of n subjects, that contain v voxels each. By combining equations 1.1 and 1.3, we obtain the so-called *mixed effects model (MFX)*, in which two levels of variance are used ($\boldsymbol{\epsilon}_1$ and $\boldsymbol{\epsilon}_2$):

$$\boldsymbol{Y} = \boldsymbol{X}_1\boldsymbol{X}_2\boldsymbol{\gamma} + \boldsymbol{X}_1\boldsymbol{\epsilon}_2 + \boldsymbol{\epsilon}_1. \tag{1.4}$$

Equation 1.4 models the fact that a subject response matches a population-level pattern (which varies according to $\boldsymbol{\epsilon}_2$) modulated by some subject-specific variations (as modeled by $\boldsymbol{\epsilon}_1$). To solve equation 1.4, a two-stage procedure called the *summary statistic procedure* [40] is generally used: The maps that are used in for the second-level analysis (equation 1.3) are the contrast maps, the t-values maps, or any kind of maps that carry information from the first-level analysis (see section 1.2.2.4 and equation 1.1). In general, the contrast maps are used (i.e. the maps corresponding to a linear combination of the coefficients $\hat{\boldsymbol{\beta}}$). One also retains the subject-specific variance information from the first-level analysis – namely $\hat{\sigma}$ –, although it can be replaced by a common estimate for all subjects, yielding the *random effects model (RFX).* Statistical tests are then conducted in order to quantify the strength of the associations and eventually perform inference at the population level. The tests are similar to those presented for the first-level analysis (section 1.2.2.4). Section 2.1 of chapter 2 reviews the state-of-the-art methods for group analysis.

The simplest kind of group analysis is to test if the mean signal across subjects is significantly non-null. This corresponds to taking a constant variate (the *intercept*) as the unique covariate. A significant effect at a particular location indicates an average group-level activation of the brain at that location when the underlying functional task is performed. Statistical inference can be done with other types of brain images, for instance with gray matter probability maps as in *Voxel Based Morphometry (VBM)* [2] (see the experiment of chapter 3, Figure 3.18): gray matter probability maps are computed with a segmentation algorithm (see [12] for a review); left and right hemispheres are averaged so as to obtain symmetric maps; the images are spatially normalized and smoothed so as to improve between-subject correspondences; finally a statistical analysis is performed in the same fashion than in functional MRI analysis.

1.2.3.3 Application to cognitive neuroscience

Most of the neuroimaging studies are performed on groups of 10 to 50 subjects. Thirion et al. [82] showed that 20 was a prerequisite for the results to be reproducible. In some particular cases, mainly clinical studies of rare diseases, less than 10 subjects are involved. For 15 years, neuroscientists have tried to map the brain regions with functional tasks by performing group analysis on the intercept (as described above), resulting in high-level brain function atlases that recall the phrenologists work – although more complex –, with for instance temporal regions associated with auditory tasks [37, 5, 94], or occipital regions with visual tasks [21, 78]. At Neurospin imaging center, a functional MRI protocol has been created as a routine experiment to map the main global cognitive brain functions [66]. Overall, a large number of increasingly complex functional tasks have been associated with various brain locations, e.g. reaction to autobiographical events [10] or violation of social norms [6]. We also remark a growth in the average size of the studies, as images are easier to acquire. The trend is actually more oriented towards the analysis of brain networks, which seems to be a more realistic approach to understand the brain deep functions (see e.g. [86]). Indeed, there are pieces of evidence against the validity of the phrenologists model [76]. Yet, regular group analyzes studies are still relevant, especially when subject-specific covariates are considered. Such variates can be used as diagnosis indicators regarding the prevention of disease (e.g. [77, 51]) or risky behaviors [50, 59]. The next section gives examples of such studies.

1.2.3.4 Neuroimaging genetic studies

Neuroimaging genetic studies are a particular kind of neuroimaging study that involves genetic variates in their design matrix (see for instance [16, 9]). Most of the time, these variates encode the variations of a nucleotide at a specific location in the ADN, a so-called *Single Nucleotide Polymorphism (SNP)* [55]. The inclusion of SNP(s) in the design often creates imbalanced classes or ill-posed designs as some genetic variations only occur in a few portion of the population (sometimes less than 5). SNPs are therefore associated with complex statistical structures for which no universal model exists. The major difficulty related to SNPs is that there are up to millions of them in the whole human genome [72], resulting in as many potential variates of interest to be included in neuroimaging experiments. *Genome Wide Associations Studies (GWAs)* [80, 17] are especially designed for the purpose of screening the whole genome in search of a significant association between brain images phenotypes and candidate genetic variates (genes or SNPs). The more SNPs are tested, the higher the probability there is that one of them would be found to have a significant correlation with neuroimaging features. This is the *multiple comparisons problem* well known to statisticians and discussed in further details in chapter 2. Neuroimaging genetic studies are especially subject to multiple comparisons problems because the dimension of the genetic data is multiplied by the dimension of the brain images (up to tenth thousands of voxels). In most of specialized journals, any publication involving genetic variates is considered to have significant results if these have an uncorrected p-value of 10^{-8} at most[2] for one single phenotype [4]. To reduce the dimension of genetic data, experts in genetics build a prior knowledge about the genome that helps targeting some specific associations between SNPs and brain characteristics, but they mainly do it at the gene scale. At best, a few dozen of SNPs are subject to analysis. Vounou et al. [88] propose a classification of neuroimaging genetic studies into four groups: *brain-wide, candidate-gene association (BW-CGA) studies*; *candidate phenotype–genome-wide association (CP–GWA) studies*; *candidate phenotype–candidate gene association (CP–CGA) studies*; and *brain-wide, genome-wide association (BW-GWA) studies*. Several examples of BW-CGA studies can be found in chapter 4. GWAs that me mentioned above encompass CP–GWA and BW-GWA studies.

1.2.4 Main difficulties encountered in neuroimaging group studies

[2] Corrected for multiple comparisons according to the non-genetic variates.

1.2.4.1 Outliers and their influence

Medical image acquisitions are prone to a wide variety of errors such as scanner instabilities, acquisition artifacts, or issues in the underlying bio-medical experimental protocol. In addition, due to the high variability observed in populations of interest, these datasets may also contain uncommon, yet technically correct, observations. Manual screening of the images is often performed in order to discard the observations that can potentially drive out subsequent analysis of the data, the so-called *outliers*. Such a quality-check is also required to ensure that the dataset meets some technical prerequisites, such as design balance, unimodal data, Gaussian distribution of the data or homogeneous variability. As the number of design factors and image descriptors improves, manual screening is no longer possible, and automated tools have to be used. Unfortunately, these tools needs to be parametrized according to a prior knowledge about the dataset statistical structure, that is not known in practice. It is difficult to measure the influence of outliers on the results of a neuroimaging study. First, there is no formal definition of what an outlier is, and it is therefore impossible to compare the results obtained with and without the inclusion of those in experiments performed on real data. Second, outliers may in fact reflect a particular, seldom property of the statistical structure of the whole dataset [71] (e.g. a tiny cluster of similar outliers corresponding to a population caste), which suggests considering another model rather than discarding observations. Third, some observations may be outliers according to a limited number of their features, and still contribute to improving the statistical power of the analysis in some brain regions. However, the literature (including our work) reports some examples of neuroimaging studies that were performed in a non-robust and a robust version, and for which the results were extremely different in both cases [31, 89, 63]. These examples do not give an universal solution to control the influence of potential outliers, but they clearly point out its danger.

Chapter 3 deals with the detection of multivariate outliers in neuroimaging datasets and mainly targets clinical studies where subjects inclusion plays an important role. However, one has to remain aware that the potential presence of outliers is not the only reason why robust procedures are needed and outlier detection therefore only addresses a limited part of the much more general problem discussed in the previous Subsection.

1.2.4.2 Data scarcity and lack of reproducibility

Due to the difficulty of acquiring good quality functional images, most of the neuroimaging are limited to a few dozens of subjects. Regarding the large number of images descriptors, such high-dimensional studies are hindered by the large estimation variance and no subtle effect can be investigated (i.e. the statistical power is poor). In order to perform neuroimaging genetic studies, large cohorts of more than 1000 subjects start to emerge as part of major projects (e.g. Imagen [73], Human Connectom Project [87], ADNI [43]). From a theoretical point of view, it seems however than more subjects (10 to 100 times more) would be needed in order to reproduce neuroimaging genetic findings [80], especially regarding genome-wide association studies.

Before the statistical regime mentioned above is reached, methods that reduce the variability of the studies results are necessary. This variability may come from the presence of abnormal data (see the next paragraph about outliers). Robust statistical procedures can be used to drop that source of variability. Some statistical variability is yet unavoidable but can be reduced by using appropriate methods. Chapter 4 introduces a method for neuroimaging group analysis that has the property of being more *stable* than the state-of-the-art methods, i.e. we observe a drop in the variability of the results of a same analysis performed on random subgroups of the same cohort. Our method moreover comes with increased sensitivity, which affords a more powerful statistical inference and the detection of more subtle effects.

1.2.4.3 Data quality is hard to control

A standard assumption in univariate voxel-wise analyzes is that the data are Gaussian distributed. Figure 1.7 demonstrates that this assumption does almost never holds by showing the histogram of the Shapiro-Wilk Gaussianity test applied to *[Angry faces viewing - Control]* fMRI contrast images and gray matter probability maps of 1500 subjects. Similar results (not shown) were obtained with parcel-level images descriptors, for brain parcellations with 100, 1000 and 10000 parcels. The skewness and the kurtosis of the voxels across subjects are shown in Figure 1.8 for fMRI data and in Figure 1.9 for gray matter intensity maps. In both cases, almost all voxels have skewness and kurtosis values that do not correspond to Gaussian distributed data (only such values are colored). We observe smooth regionalized patterns, which suggests that the statistical properties of neighboring voxels are similar. We also see a spatial correspondence between skewness and kurtosis. The pattern does not correspond yet to the group-level activation pattern (not shown). Strong positive values of the skewness are observed on the brain contours for gray matter density maps, and strong negative values lie in deep brain structures. The former correspond to high-variability regions (we have seen that the folding structure is highly variable) where only a few subjects may actually have gray matter, while the latter correspond to the opposite situation where only a few subjects –probably outliers– do not have a gray matter probability of 1 at the corresponding locations. Thirion et al. [82] also demonstrated that the Gaussianity assumption was not always enforced. Kherif et al. [45] stresses the importance of good choice of model and discuss the homogeneity assumption.

The literature about *robust statistics* provides several techniques and tools to perform reliable statistical inference on datasets that do not fulfill the model assumptions, potentially because of a contamination with outliers. Thus, robust regression is a useful tool regarding neuroimaging studies, especially when complex designs are involved (e.g. neuroimaging genetic studies) or when large cohorts are studied. Chapter 5 demonstrates this statement and quantifies the sensitivity improvements brought by robust regression.

1.3 Some neuroimaging and data-analysis softwares

We focus on the softwares that have been used in this thesis, which are mostly Python packages for neuroimaging data analysis and handling. The *Image Processing* subsection introduces the three main neuroimaging softwares. Although

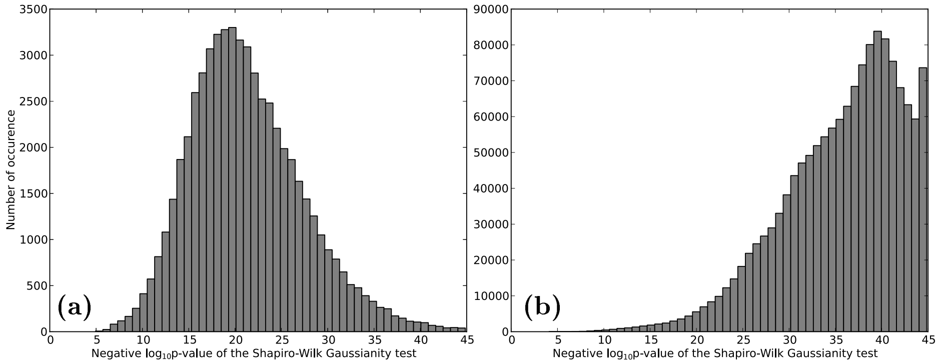


Figure 1.7: **Negative $\log_{10}$p-values of a voxel-wise Shapiro-Wilk Gaussianity test for (a) fMRI contrast images; (b) gray matter probability maps.** The histograms clearly demonstrate that the data are not Gaussian distributed, although this assumption is often made in practice. Considering signal averages within parcels does not help.

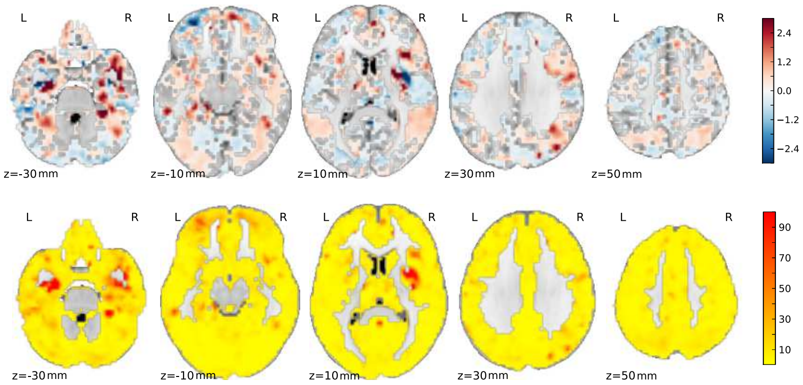


Figure 1.8: **Skewness (first row) and kurtosis (second row) across the fMRI contrast maps of 1500 subjects.** Positive and negative values of the skewness appear separated. Voxels with the highest kurtosis values seem to be organized in clusters that spatially match regions with the highest absolute skewness values. Such observations suggest local homogeneities in statistical structure. However, nothing seems to explain the observed patterns.

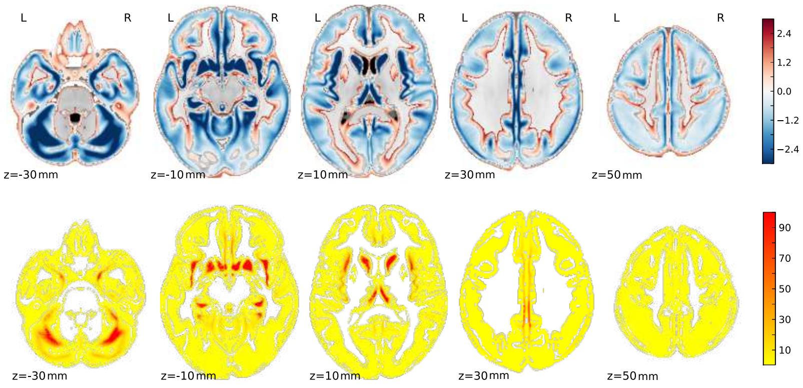


Figure 1.9: **Skewness (first row) and kurtosis (second row) across the gray matter probability maps of 1500 subjects.** Positive values of the skewness statistic correspond to the brain contours, while deeper brain regions corresponding to white matter have strong negative skewness values. The latter regions are also associated with strong kurtosis values. Unlike in Figure 1.8, the patterns observed in this examples seem to be well explained by the brain anatomical organization.

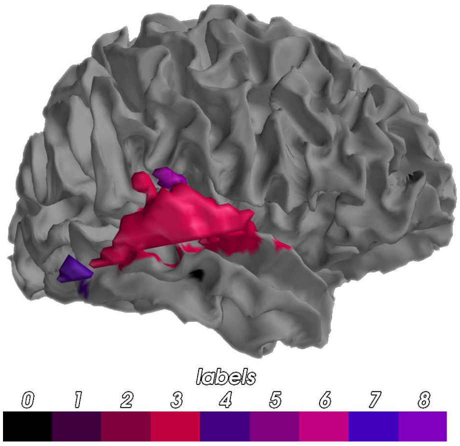


Figure 1.10: **Data visualization with Mayavi.** Activations obtained in a surface-based analysis are embedded in a three-dimensional space and matched to the activations obtained in the corresponding volume-based analysis.

they all can perform visualization and statistical analysis as well as image processing, their source-code cannot be edited as easily as Python packages, resulting in a less practical use for methodological research. The three softwares however implements specific methods and treatments that are sometimes not available elsewhere.

1.3.1 Visualization

1.3.1.1 Anatomist

Anatomist [14] is a viewer written in C++. It can interpret Python code, but we use it for its simple graphical interface. We particularly appreciate how it is easy to dynamically change the color map of one or several images and adjust their bounds. Basics operations such as superimposing images or synchronizing the view of independent windows are very handy when one wants to look at the results of any neuroimaging algorithm, the drawback being that they are difficult to reproduce.

1.3.1.2 Mayavi

Images vizualisations can be embedded in Python scripts through calls to the *Mayavi* [69] Python package routines (output example given at Figure 1.10). Although not specifically designed for neuroimaging, Mayavi suits the neuroscientists needs as few code lines can generate and interactive 3D rendering, providing real-time viewing of the analysis results. However, obtaining fine-tuned visualizations requires a good understanding of Mayavi data structures and API. As a result, its use remains limited to specific cases where complex visualizations are needed.

1.3.2 Image processing

1.3.2.1 Freesurfer

The *Freesurfer* software [24], from Harvard University, is mainly famous for its robust and good quality reconstruction pipeline: input raw anatomical images are preprocessed, registered, resampled, segmented, and a reconstruction of the surface is computed. Specific neuroimaging algorithms can be chained with the reconstruction pipeline (e.g. fibers reconstruction). Other softwares such as Brainvisa or Caret have a similar reconstruction pipeline, but their use is less standard in the community. Freesurfer also provides good quality preprocessings. In particular, it implements Boundary-Based Registration (BBR), which yields a better quality registration, and hence improves the overall quality of the preprocessings.

1.3.2.2 SPM

Statistical Parametric Mapping (SPM) is a software that is designed to find local changes in the brain activity under specific experimental conditions from brain images [26, 28]. SPM can perform preprocessing and first-level analysis as well as second-level analysis. It is particularly useful to estimate contrast images in experiments involving several contrast and more generally to deal with design matrices. The dependency on Matlab is a drawback. All the brain images used in this thesis were preprocessed with SPM8 (see section 1.4 about the Imagen database).

1.3.3 Note on *all-in-one* softwares

The two softwares presented above look more like meta-softwares that encompass various methods and that are used to distribute new ones. Another important one is *FSL (FMRIB Software Library)* [44, 91, 75], from Oxford University. For each of these softwares, a default behavior is set up so that one can run an entire analysis without needing to choose any detailed parameter, but this feature is also the software's weakness as the default framework may change across versions and yield results that are version-dependent. Interestingly, indication of the software version may be as important in a research report as the detailed description of the analysis framework.

1.3.4 Statistical analysis

1.3.4.1 Scikit-learn

Scikit-learn [65] is an open source machine learning Python package in which a particular emphasis is put on high-dimensional data analysis. Its development originated from the Inria Parietal team. Scikit-learn is useful in neuroimaging since a lot of standard algorithms such as clustering, model fitting or data transformation can be used "out of the box" before more specific algorithms are applied. Indeed, only well-known algorithms are available in Scikit-learn, i.e. algorithms for which the code can be understood and maintained by a sufficient number of people. This ensures the long-term support of the project as well as its quality, but limits the application scope of the package, that needs to be combined with more specific packages regarding neuroimaging-specific applications.

1.3.4.2 Statsmodels

Fine-tuned statistical analysis can be performed with the *Statsmodels*[3] Python package. Unlike Scikit-learn, statsmodels implements an impressive number of statistical tests and estimators, including specific variants that seem to have a poor practical interest. The Statsmodels package is yet advantageously combined with neuroimaging-oriented packages to explore new analysis techniques (e.g. robust regression in group analysis or covariance estimation for outlier detection).

1.3.4.3 Nipy / Nilearn

Data treatments that are specific to neuroimaging are available as parts of the *Nipy* [57] Python software, or its (still young but promising) alternative *Nilearn*[4]. Nipy also provides algorithms to perform preprocessings. As open-source softwares written in Python, Nipy and Nilearn are very helpful to the researcher as their code serves as a basis to investigate new methodological tools. These tools can thus directly be shared as routines that have limited dependence on the original code base.

1.4 The Imagen database

This thesis uses the data from the *Imagen database*. Imagen is a European multicentric study involving adolescents [73]. It contains a large functional neuroimaging database with fMRI associated with 99 different contrast images for 4 protocols in more than 2000 subjects, who gave informed signed consent. Regarding the functional neuroimaging data, we notably used the faces protocol [38] and its *[angry faces - control]* contrast, i.e. the difference between watching angry faces and non-biological stimuli (concentric circles). We also use the Stop Signal Task (SST) protocol [52], with the activation during a *[go wrong]* event, i.e. when the subject pushes the wrong button. Images from the Modified Incentive Delay (MID) protocol [47] were also used.

Eight different 3T scanners from multiple manufacturers (GE, Siemens, Philips) were used to acquire the data. Standard preprocessing, including slice timing correction, spike and motion correction, temporal detrending (functional data), and spatial normalization (anatomical and functional data), were performed using the SPM8 software and its default parameters; functional images were resampled at 3mm resolution. All images were warped in the MNI152 coordinate space using a study-specific template. Obvious outliers detected using simple rules such as large registration or segmentation errors or very large motion parameters were removed after this step. BOLD time series was recorded using Echo-Planar Imaging, with TR = 2200 ms, TE = 30 ms, flip angle = 75° and spatial resolution 3mm × 3mm × 3mm. Gaussian smoothing at 5mm-FWHM was finally added[5]. Contrasts were obtained using a standard linear model, based on the convolution of the time course of the experimental conditions with the canonical haemodynamic response function, together with standard high-pass filtering

[3] http://statsmodels.sourceforge.net/.

[4] http://nilearn.github.io.

[5] Smoothing is only applied in the first-level analysis in order to improve the sensitivity of the first-level analysis that yields the contrast maps.

(period = 120s) and temporally auto-regressive noise model. The estimation of the first-level was carried out using the SPM8 software. T1-weighted MPRAGE anatomical images were acquired with spatial resolution 1mm × 1mm × 1mm, and gray matter probability maps were available for 1986 subjects as outputs of the SPM8 "New Segmentation" algorithm applied to the anatomical images. A mask of the gray matter was built by averaging and thresholding the individual gray matter probability maps. More details about data preprocessing can be found in [83]. Genotyping was performed genome-wide using Illumina Quad 610 and 660 chips, yielding approximately 600,000 autosomic SNPs. 477,215 SNPs are common to the two chips and pass *plink* standard parameters (Minor Allele Frequency > 0.05, Hardy-Weinberg Equilibrium $P < 0.001$, missing rate per SNP < 0.05).

References

[1] Arezzo, J., Vaughan Jr, H., Kraut, M., Steinschneider, M., Legatt, A.: Intracranial generators of event-related potentials in the monkey. Frontiers of clinical neuroscience 3, 174–189 (1986)

[2] Ashburner, J., Friston, K.J.: Voxel-based morphometry – the methods. Neuroimage 11(6), 805–821 (2000)

[3] Badillo, S., Vincent, T., Ciuciu, P.: Group-level impacts of within- and between-subject hemodynamic variability in fMRI. NeuroImage (2013)

[4] Barsh, G.S., Copenhaver, G.P., Gibson, G., Williams, S.M.: Guidelines for genome-wide association studies. PLoS genetics 8(7), e1002812 (2012)

[5] Berry, I., Demonet, J., Warach, S., Viallard, G., Boulanouar, K., Franconi, J., Marc-Vergnes, J., Edelman, R., Manelfe, C.: Activation of association auditory cortex demonstrated with functional MRI. Neuroimage 2(3), 215–219 (1995)

[6] Berthoz, S., Armony, J., Blair, R., Dolan, R.: An fMRI study of intentional and unintentional (embarrassing) violations of social norms. Brain 125(8), 1696–1708 (2002)

[7] Boynton, G.M., Engel, S.A., Glover, G.H., Heeger, D.J.: Linear systems analysis of functional magnetic resonance imaging in human V1. The journal of neuroscience 16(13), 4207–4221 (1996)

[8] Brant-Zawadzki, M., Fein, G., Van Dyke, C., Kiernan, R., Davenport, L., De Groot, J.: MR imaging of the aging brain: patchy white-matter lesions and dementia. American journal of neuroradiology 6(5), 675–682 (1985)

[9] Brown, S.M., Hariri, A.R.: Neuroimaging studies of serotonin gene polymorphisms: exploring the interplay of genes, brain, and behavior. Cognitive, Affective, & Behavioral Neuroscience 6(1), 44–52 (2006)

[10] Cabeza, R., Prince, S.E., Daselaar, S.M., Greenberg, D.L., Budde, M., Dolcos, F., LaBar, K.S., Rubin, D.C.: Brain activity during episodic retrieval of autobiographical and laboratory events: an fMRI study using a novel photo paradigm. Journal of cognitive neuroscience 16(9), 1583–1594 (2004)

[11] Chen, N.k., Wyrwicz, A.M.: Correction for EPI distortions using multi-echo gradient-echo imaging. Magnetic resonance in medicine 41(6), 1206–1213 (1999)

[12] Clarke, L., Velthuizen, R., Camacho, M., Heine, J., Vaidyanathan, M., Hall, L., Thatcher, R., Silbiger, M.: MRI segmentation: methods and applications. Magnetic resonance imaging 13(3), 343–368 (1995)

[13] Cohen, L., Dehaene, S., Naccache, L., Lehéricy, S., Dehaene-Lambertz, G., Hénaff, M.A., Michel, F.: The visual word form area spatial and temporal characterization of an initial stage of reading in normal subjects and posterior split-brain patients. Brain 123(2), 291–307 (2000)

[14] Cointepas, Y., Geffroy, D., Souedet, N., Denghien, I., Rivière, D., Roses, F.: The BrainVISA project: a shared software development infrastructure for biomedical imaging research. Proceedings 16th HBM (2010)

[15] Collins, D.L., Le Goualher, G., Evans, A.C.: Non-linear cerebral registration with sulcal constraints. In: Medical Image Computing and Computer-Assisted Interventation – MICCAI98, pp. 974–984. Springer (1998)

[16] Congdon, E., Canli, T.: The endophenotype of impulsivity: reaching consilience through behavioral, genetic, and neuroimaging approaches. Behavioral and cognitive neuroscience reviews 4(4), 262–281 (2005)

[17] Da Mota, B., Frouin, V., Duchesnay, E., Laguitton, S., Varoquaux, G., Poline, J.B., Thirion, B., et al.: A fast computational framework for genome-wide association studies with neuroimaging data. In: 20th International Conference on Computational Statistics (COMPSTAT 2012) (2012)

[18] Damadian, R., Goldsmith, M., Minkoff, L.: NMR in cancer: XVI. FONAR image of the live human body. Physiological chemistry and physics 9(1), 97–100 (1976)

[19] Duzenli, C., Robinson, D.: Correcting for RF inhomogeneities in multiecho pulse sequence MRI dosimetry. Medical Physics 22, 1645 (1995)

[20] Ehman, R.L., Felmlee, J.P.: Flow artifact reduction in MRI: a review of the roles of gradient moment nulling and spatial presaturation. Magnetic resonance in medicine 14(2), 293–307 (1990)

[21] Engel, S.A., Glover, G.H., Wandell, B.A.: Retinotopic organization in human visual cortex and the spatial precision of functional MRI. Cerebral cortex 7(2), 181–192 (1997)

[22] Erasmus, L., Hurter, D., Naudé, M., Kritzinger, H., Acho, S.: A short overview of MRI artefacts. South African Journal of Radiology 8(2) (2004)

[23] Evans, A., Collins, D., Milner, B.: An MRI-based stereotactic atlas from 250 young normal subjects. In: Soc. neurosci. abstr. vol. 18, p. 408 (1992)

[24] Fischl, B.: Freesurfer. NeuroImage 62(2), 774–781 (2012)

[25] Fischl, B., Sereno, M.I., Dale, A.M.: Cortical surface-based analysis: II: Inflation, flattening, and a surface-based coordinate system. Neuroimage 9(2), 195–207 (1999)

[26] Frackowiak, R.S.: Human Brain Function. Academic Press, Burlington (1997)

[27] Friston, K., Ashburner, J., Frith, C.D., Poline, J.B., Heather, J.D., Frackowiak, R.S., et al.: Spatial registration and normalization of images. Human brain mapping 3(3), 165–189 (1995)

[28] Friston, K.J.: Statistical parametric mapping. In: Kötter, R. (ed.) Neuroscience Databases, pp. 237–250. Springer US (2003)

[29] Friston, K.J., Holmes, A.P., Worsley, K.J., Poline, J.P., Frith, C.D., Frackowiak, R.S.: Statistical parametric maps in functional imaging: a general linear approach. Human brain mapping 2(4), 189–210 (1994)

[30] Friston, K.J., Williams, S., Howard, R., Frackowiak, R.S., Turner, R.: Movement-related effects in fMRI time-series. Magnetic resonance in medicine 35(3), 346–355 (1996)

[31] Fritsch, V., Da Mota, B., Varoquaux, G., Frouin, V., Loth, E., Poline, J.B., Thirion, B., et al.: Robust group-level inference in neuroimaging genetic studies. In: Pattern Recognition in Neuroimaging (2013)

[32] Galaburda, A.M., Rosen, G.D., Sherman, G.F.: Individual variability in cortical organization: its relationship to brain laterality and implications to function. Neuropsychologia 28(6), 529–546 (1990)

[33] Gangarosa, R.E., Minnis, J.E., Nobbe, J., Praschan, D., Genberg, R.W.: Operational safety issues in MRI. Magnetic Resonance Imaging 5(4), 287–292 (1987)

[34] Gilmour, R.F., Zipes, D.P.: Electrophysiological characteristics of rodent myocardium damaged by adrenaline. Cardiovascular research 14(10), 582–589 (1980)

[35] Glover, G.H.: Deconvolution of impulse response in event-related BOLD fMRI. Neuroimage 9(4), 416–429 (1999)

[36] Gottlieb, D., Shu, C.W.: On the Gibbs phenomenon and its resolution. SIAM review 39(4), 644–668 (1997)

[37] Grasby, P., Frith, C., Friston, K., Bench, C., Frackowiak, R., Dolan, R.: Functional mapping of brain areas implicated in auditory-verbal memory function. Brain 116(1), 1–20 (1993)

[38] Grosbras, M.H., Paus, T.: Brain networks involved in viewing angry hands or faces. Cereb Cortex 16(8), 1087–1096 (Aug 2006)

[39] Handwerker, D.A., Ollinger, J.M., D'Esposito, M.: Variation of BOLD hemodynamic responses across subjects and brain regions and their effects on statistical analyses. Neuroimage 21(4), 1639–1651 (2004)

[40] Holmes, A., Friston, K.: Generalisability, random effects & population inference. Neuroimage 7, S754 (1998)

[41] Huettel, S.A., Song, A.W., McCarthy, G.: Functional magnetic resonance imaging, vol. 1. Sinauer Associates Sunderland (2004)

[42] Hutton, C., Bork, A., Josephs, O., Deichmann, R., Ashburner, J., Turner, R.: Image distortion correction in fMRI: a quantitative evaluation. Neuroimage 16(1), 217–240 (2002)

[43] Jack, C.R., Bernstein, M.A., Fox, N.C., Thompson, P., Alexander, G., Harvey, D., Borowski, B., Britson, P.J., L Whitwell, J., Ward, C., et al.: The Alzheimer's disease neuroimaging initiative (ADNI): MRI methods. Journal of Magnetic Resonance Imaging 27(4), 685–691 (2008)

[44] Jenkinson, M., Beckmann, C.F., Behrens, T.E., Woolrich, M.W., Smith, S.M.: FSL. NeuroImage 62(2), 782–790 (2012)

[45] Kherif, F., Poline, J.B., Mériaux, S., Benali, H., Flandin, G., Brett, M.: Group analysis in functional neuroimaging: selecting subjects using similarity measures. Neuroimage 20(4), 2197–2208 (2003)

[46] Klein, A., Andersson, J., Ardekani, B.A., Ashburner, J., Avants, B., Chiang, M.C., Christensen, G.E., Collins, D.L., Gee, J., Hellier, P., et al.: Evaluation of 14 nonlinear deformation algorithms applied to human brain MRI registration. Neuroimage 46(3), 786–802 (2009)

[47] Knutson, B., Westdorp, A., Kaiser, E., Hommer, D.: fMRI visualization of brain activity during a monetary incentive delay task. Neuroimage 12(1), 20–27 (Jul 2000)

[48] Kwong, K.K., Belliveau, J.W., Chesler, D.A., Goldberg, I.E., Weisskoff, R.M., Poncelet, B.P., Kennedy, D.N., Hoppel, B.E., Cohen, M.S., Turner, R.: Dynamic magnetic resonance imaging of human brain activity during primary sensory stimulation. Proceedings of the National Academy of Sciences 89(12), 5675–5679 (1992)

[49] Langlois, S., Desvignes, M., Constans, J.M., Revenu, M., et al.: MRI geometric distorsion: A simple approach to correcting the effects of non-linear gradient fields. Journal of magnetic resonance imaging 9, 821–831 (1999)

[50] Lee, T.M., Leung, A.W., Fox, P.T., Gao, J.H., Chan, C.C.: Age-related differences in neural activities during risk taking as revealed by functional MRI. Social cognitive and affective neuroscience 3(1), 7–15 (2008)

[51] Lieb, W., Beiser, A.S., Vasan, R.S., Tan, Z.S., Au, R., Harris, T.B., Roubenoff, R., Auerbach, S., DeCarli, C., Wolf, P.A., et al.: Association of plasma leptin levels with incident Alzheimer disease and MRI measures of brain aging. JAMA: the journal of the American Medical Association 302(23), 2565–2572 (2009)

[52] Logan, G.D.: On the ability to inhibit thought and action: A users' guide to the stop signal paradigm. Psychological Review 91(3), 295–327 (1994)

[53] Lombaert, H., Sun, Y., Cheriet, F.: Landmark-based non-rigid registration via graph cuts. In: Image Analysis and Recognition, pp. 166–175. Springer (2007)

[54] Lu, W., Pauly, K.B., Gold, G.E., Pauly, J.M., Hargreaves, B.A.: SEMAC: slice encoding for metal artifact correction in MRI. Magnetic Resonance in Medicine 62(1), 66–76 (2009)

[55] Marth, G.T., Korf, I., Yandell, M.D., Yeh, R.T., Gu, Z., Zakeri, H., Stitziel, N.O., Hillier, L., Kwok, P.Y., Gish, W.R.: A general approach to single-nucleotide polymorphism discovery. Nature genetics 23(4), 452–456 (1999)

[56] Menon, R.S., Luknowsky, D.C., Gati, J.S.: Mental chronometry using latency-resolved functional MRI. Proceedings of the National Academy of Sciences 95(18), 10902–10907 (1998)

[57] Millman, K., Brett, M.: Analysis of functional magnetic resonance imaging in Python. Computing in Science Engineering 9(3), 52–55 (2007)

[58] Nakagawa, S., Cuthill, I.C.: Effect size, confidence interval and statistical significance: a practical guide for biologists. Biological Reviews 82(4), 591–605 (2007)

[59] Norman, A.L., Pulido, C., Squeglia, L.M., Spadoni, A.D., Paulus, M.P., Tapert, S.F.: Neural activation during inhibition predicts initiation of substance use in adolescence. Drug and alcohol dependence 119(3), 216–223 (2011)

[60] Ogawa, S., Lee, T.M., Nayak, A.S., Glynn, P.: Oxygenation-sensitive contrast in magnetic resonance image of rodent brain at high magnetic fields. Magnetic resonance in medicine 14(1), 68–78 (1990)

[61] Ogawa, S., Tank, D.W., Menon, R., Ellermann, J.M., Kim, S.G., Merkle, H., Ugurbil, K.: Intrinsic signal changes accompanying sensory stimulation: functional brain mapping with magnetic resonance imaging. Proceedings of the National Academy of Sciences 89(13), 5951–5955 (1992)

[62] Orzada, S., Maderwald, S., Poser, B.A., Bitz, A.K., Quick, H.H., Ladd, M.E.: RF excitation using time interleaved acquisition of modes (TIAMO) to address B1 inhomogeneity in high-field MRI. Magnetic Resonance in Medicine 64(2), 327–333 (2010)

[63] Ousdal, O.T., Anand Brown, A., Jensen, J., Nakstad, P.H., Melle, I., Agartz, I., Djurovic, S., Bogdan, R., Hariri, A.R., Andreassen, O.A.: Associations between variants near a monoaminergic pathways gene (PHOX2B) and amygdala reactivity: A genome-wide functional imaging study. Twin Research and Human Genetics 15, 273–285 (6 2012)

[64] Parker, D.L., Gullberg, G.T., Frederick, P.R.: Gibbs artifact removal in magnetic resonance imaging. Medical physics 14, 640 (1987)

[65] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, É.: Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research (Oct 2011)

[66] Pinel, P., Thirion, B., Mériaux, S., Jobert, A., Serres, J., Le Bihan, D., Poline, J.B., Dehaene, S.: Fast reproducible identification and large-scale databasing of individual functional cognitive networks. BMC neuroscience 8(1), 91 (2007)

[67] Poline, J.B., Thirion, B., Roche, A., Meriaux, S.: Intersubject variability in fMRI data: Causes, consequences, and related analysis strategies. edited by Stephen José Hanson and Martin Bunzl p. 173 (2010)

[68] Porter, R.: Somato-sensory projections to the motor complex. Neurology and neurobiology 56, 157–167 (1990)

[69] Ramachandran, P., Varoquaux, G.: Mayavi: 3D visualization of scientific data. Computing in Science & Engineering 13(2), 40–51 (2011)

[70] Roche, A., Mériaux, S., Keller, M., Thirion, B.: Mixed-effect statistics for group analysis in fMRI: a nonparametric maximum likelihood approach. Neuroimage 38(3), 501–510 (2007)

[71] Rousseeuw, P.J., Leroy, A.M.: Robust regression and outlier detection, vol. 589. Wiley. com (2005)

[72] Sachidanandam, R., Weissman, D., Schmidt, S.C., Kakol, J.M., Stein, L.D., Marth, G., Sherry, S., Mullikin, J.C., Mortimore, B.J., Willey, D.L., et al.: A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. Nature 409(6822), 928–933 (2001)

[73] Schumann, G., Loth, E., Banaschewski, T., Barbot, A., Barker, G., BÃ$\frac{1}{4}$chel, C., Conrod, P.J., Dalley, J.W., Flor, H., Gallinat, J., Garavan, H., Heinz, A., Itterman, B., Lathrop, M., Mallik, C., Mann, K., Martinot, J.L., Paus, T., Poline, J.B., Robbins, T.W., Rietschel, M., Reed, L., Smolka, M., Spanagel, R., Speiser, C., Stephens, D.N., Ströhle, A., Struve, M., IMAGEN consortium: The IMAGEN study: reinforcement-related behaviour in normal brain function and psychopathology. Mol Psychiatry 15(12), 1128–1139 (Dec 2010)

[74] Sladky, R., Friston, K.J., Tröstl, J., Cunnington, R., Moser, E., Windischberger, C.: Slice-timing effects and their correction in functional MRI. Neuroimage 58(2), 588–594 (2011)

[75] Smith, S.M., Jenkinson, M., Woolrich, M.W., Beckmann, C.F., Behrens, T.E., Johansen-Berg, H., Bannister, P.R., Luca, M.D., Drobnjak, I., Flitney, D.E., Niazy, R.K., Saunders, J., Vickers, J., Zhang, Y., Stefano, N.D., Brady, J.M., Matthews, P.M.: Advances in functional and structural MR image analysis and implementation as FSL. NeuroImage 23, Supplement 1(0), S208–S219 (2004)

[76] Sohrabi, A., Brook, A.: Functional neuroimaging and its implications for cognitive science: Beyond phrenology and localization. In: Proceedings of the Twenty-Seventh Annual Meeting of the Cognitive Science Society. pp. 2044–2049 (2005)

[77] Soininen, H.S., Scbeltens, P.: Early diagnostic indices for the prevention of Alzheimer's disease. Annals of medicine 30(6), 553–559 (1998)

[78] Somers, D.C., Dale, A.M., Seiffert, A.E., Tootell, R.B.: Functional MRI reveals spatially specific attentional modulation in human primary visual cortex. Proceedings of the National Academy of Sciences 96(4), 1663–1668 (1999)

[79] Sotiras, A., Christos, D., Paragios, N., et al.: Deformable medical image registration: A survey (2012)

[80] Stein, J.L., Hua, X., Lee, S., Ho, A.J., Leow, A.D., Toga, A.W., Saykin, A.J., Shen, L., Foroud, T., Pankratz, N., et al.: Voxelwise genome-wide association study (vGWAS). Neuroimage 53(3), 1160–1174 (2010)

[81] Talairach, J., Tournoux, P.: Co-planar stereotaxic atlas of the human brain. 3-dimensional proportional system: an approach to cerebral imaging (1988)

[82] Thirion, B., Pinel, P., Mériaux, S., Roche, A., Dehaene, S., Poline, J.B.: Analysis of a large fMRI cohort: Statistical and methodological issues for group analyses. NeuroImage 35(1), 105 – 120 (2007)

[83] Thyreau, B., Schwartz, Y., Thirion, B., Frouin, V., Loth, E., Vollstädt-Klein, S., Paus, T., Artiges, E., Conrod, P.J., Schumann, G., Whelan, R., Poline, J.B., Consortium, I.M.A.G.E.N.: 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 (May 2012)

[84] Tohka, J., Zijdenbos, A., Evans, A.: Fast and robust parameter estimation for statistical partial volume models in brain MRI. Neuroimage 23(1), 84–97 (2004)

[85] Tucholka, A., Fritsch, V., Poline, J.B., Thirion, B.: An empirical comparison of surface-based and volume-based group studies in neuroimaging. Neuroimage 63(3), 1443–1453 (2012)

[86] Van Den Heuvel, M.P., Hulshoff Pol, H.E.: Exploring the brain network: a review on resting-state fMRI functional connectivity. European Neuropsychopharmacology 20(8), 519–534 (2010)

[87] Van Essen, D.C., Ugurbil, K., Auerbach, E., Barch, D., Behrens, T., Bucholz, R., Chang, A., Chen, L., Corbetta, M., Curtiss, S., et al.: The human connectome project: a data acquisition perspective. Neuroimage 62(4), 2222–2231 (2012)

[88] Vounou, M., Nichols, T.E., Montana, G.: Discovering genetic associations with high-dimensional neuroimaging phenotypes: a sparse reduced-rank regression approach. Neuroimage 53(3), 1147–1159 (2010)

[89] Wager, T.D., Keller, M.C., Lacey, S.C., Jonides, J.: Increased sensitivity in neuroimaging analyses using robust regression. NeuroImage 26, 99 (2005)

[90] Woolrich, M.W., Behrens, T.E., Beckmann, C.F., Jenkinson, M., Smith, S.M.: Multilevel linear modelling for fMRI group analysis using Bayesian inference. Neuroimage 21(4), 1732–1747 (2004)

[91] Woolrich, M.W., Jbabdi, S., Patenaude, B., Chappell, M., Makni, S., Behrens, T., Beckmann, C., Jenkinson, M., Smith, S.M.: Bayesian analysis of neuroimaging data in FSL. NeuroImage 45(1, Supplement 1), S173–S186 (2009)

[92] Worsley, K.J., Marrett, S., Neelin, P., Evans, A.: Searching scale space for activation in PET images. Human brain mapping 4(1), 74–90 (1996)

[93] Wu, D.H., Lewin, J.S., Duerk, J.L.: Inadequacy of motion correction algorithms in functional MRI: Role of susceptibility-induced artifacts. Journal of Magnetic Resonance Imaging 7(2), 365–370 (1997)

[94] Yoo, S.S., Lee, C.U., Choi, B.G.: Human brain mapping of auditory imagery: event-related functional MRI study. Neuroreport 12(14), 3045–3049 (2001)

Chapter 2
Statistical inference for group neuroimaging datasets

Contents

This chapter introduces the main theoretical tools that are used in the sequel of the manuscript. We have seen in chapter 1 that statistical tests are the basis of functional studies as these try to uncover differences in brain activity across varying experimental conditions. Section 2.1 gives a formal description of hypothesis testing and its applications to neuroimaging. Then we discuss covariance estimation in section 2.2 as it is an essential ingredient in many statistical models used within this thesis (e.g. outlier detection 3 or F-tests on the coefficients estimated in a linear regression framework). We present the main concepts and tools of non-parametric statistics within the last section of this chapter. These are used throughout this thesis for various purpose such as building sampling schemes, computing consensus statistics or approximating statistical structure with no prior knowledge.

2.1 Hypothesis testing

2.1.1 Inferential statistics

As soon as practical applications are concerned, theoretically-grounded descriptive statistics or models are subject to interpretation by domain experts. *Hypothesis testing* is a sub-domain of applied statistics that encompasses every rule, procedure or technique aiming at turning statistical measures performed on data into practical decisions while taking into account the probability of potential errors. One important role of hypothesis testing is thus to limit subjective interpretation by providing means of quantifying decision errors.

In practice, experts observe a phenomenon occurring under precise experimental conditions that they design for this purpose, and they wonder if the observation of the phenomenon could have happened by chance (the *null hypothesis*). If so, they conclude that their observation is not related to the experimental conditions, or at least not specifically: We say that the null hypothesis *cannot be rejected*[1]. If not, that is the phenomenon has much greater chance to be observed under the specific experimental conditions, the conclusion is therefore that the phenomenon is not a spurious observation reveals a true effect and we say that experts *reject* the null hypothesis.

2.1.1.1 Formal description

If we consider the "observed phenomenon" of the previous paragraph as a random variable X, we denote $H_{0,X}$ the distribution of X under no specific experimental conditions (i.e. the null hypothesis), while H_1 is the distribution of X under those (i.e. the *alternative hypothesis*). Thus, the experts' concern would be to test whether H_0 or H_1 is more likely (we omit the X index when no confusion is possible). However, only one measurement of X, say $X(\omega)$, is performed in general and the question boils down to determining whether $X(\omega)$ can be considered as a regular measurement of X with respect to H_0 or whether there this measurement is so exceptional that it obviously corresponds to H_1. The more the measurement seems exceptional, the more one is convinced that H_1 is likely.

[1] In the sequel, we use the term *accept* (resp. *acceptance*) as a misnomer for *fail to reject* (resp. *failure to reject*).

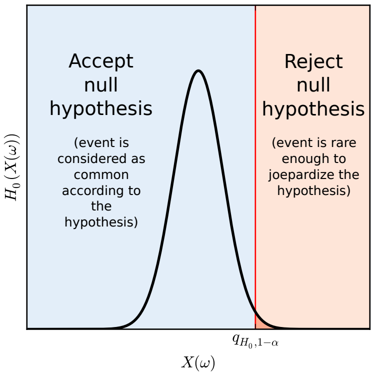

Accept null hypothesis
(event is considered as common according to the hypothesis)
Reject null hypothesis
(event is rare enough to joepardize the hypothesis)
$H_0(X(\omega))$
$q_{H_0,1-\alpha}$
$X(\omega)$

Figure 2.1: **Simple hypothesis testing.** If the observed event $X(\omega)$ is exceptional regarding to what we expect from the null distribution H_0, we reject that hypothesis because it is likely that the event has been generated according to another rule. The α parameter (type I error rate) controls the level of confidence of the rejection.

One generally considers the likelihoods of $X(\omega)$ under H_0 and H_1, respectively $L(H_0|X(\omega))$ and $L(H_1|X(\omega))$. The ratio $\Lambda(X(\omega)) = \frac{L(H_0|X(\omega))}{L(H_1|X(\omega))}$ is a realization of another random variable that represents the trade-off between the probabilities of $X(\omega)$ to be observed under H_0 or H_1. The final decision on accepting/rejecting the null hypothesis can therefore be made according to the value of $\Lambda(X(\omega))$. It is possible to control the confidence with the which we accept/reject the null hypothesis with a parameter $0 < \alpha < 1$: Let us define η so that $\mathrm{P}(\Lambda(X) \leq \eta | H_0) = \alpha$. Then the null hypothesis should be rejected if $\Lambda(X(\omega)) \leq \eta$. According to Neyman-Pearson lemma, this testing procedure yields the most discriminative power over the set of all potential tests. Equivalently to the Neyman-Pearson procedure, if $q_{D,\delta}$ is the δ quantile of the distribution D, then the null hypothesis should be rejected (with $100(1-\alpha)\%$ of error) if $X(\omega) > q_{H_0,1-\alpha}$ (see Figure 2.1). Indeed, $\mathrm{P}(X > q_{H_0,1-\alpha}|H_0) = 1 - \alpha$. $\mathrm{P}(X > X(\omega)|H_0)$ is called the *p-value* associated with the observation of $X(\omega)$ under the null hypothesis. Decision about accepting/rejecting the null hypothesis can be made directly from the p-value. More detailed information about hypothesis testing can be found in [30].

When n several measurements $\{X(\omega_1), ..., X(\omega_n)\}$ of X are made under the same specific experimental conditions, one can consider a transformation of the multiple measures into a summary *test statistic* $T(X(\omega_1), ..., X(\omega_n))$ (e.g. the mean of the measurements). The value of that test statistic is the realization of a new random variable that can be compared to its null distribution $H_{0,T}$ following the framework described above. The choice of T is crucial as the quality of the final decision will highly depend upon it. A *sufficient statistic* is a statistic that contains as much information as the whole sample with respect to the testing procedure that is applied. Ideally, the chosen statistic *(i)* must be sufficient; *(ii)* its null distribution should be known, or at least estimable (see section 2.1.3); *(iii)* it must respect some practical criteria such as computation cost, interpretability, or theoretical validity under the practical problem constraints (e.g. minimum number of observations required). In neuroimaging, the F-statistic presented in section 1.2.2.4 is used. A review of the main testing procedures can be found in the same textbook than cited in the previous section [30].

2.1.1.2 Measuring the quality of a test

The ability of a statistical test to guaranty the $100\alpha\%$ bound on the false null hypothesis rejections if called the *specificity* of the test, or the *Type I error control*, while its ability to reject the null hypothesis when it actually should is called the *sensitivity*, or *Type II error control.* Sensitivity can be thought of as the discriminative power offered by the test, which is equivalent to the area under $H_{1,T}$ on the interval $[q_{(H_0,1-\alpha)}; +\infty[$. The larger the area, the more sensitive is the test.

Sensitivity, specificity, precision, recall, accuracy. The final binary decision that comes out from a statistical test yield two types of errors (type I – wrong rejection of H_0, and type II – wrong acceptance of H_0) and two types of correct outcomes (correct acceptance or rejection of H_0). The terms *True Positive (TP)*, *False Positive (FP)*, *True Negative (TN)* and *False Negative* are employed for correct rejection, uncorrect rejection, correct acceptance, uncorrect acceptance of H_0, respectively. Considering a decision procedure that performs several tests on different experimental conditions or observations, the *precision* of that procedure is defined as TP / (TP + FP), and is the probability that the detection of an "unordinary phenomenon" in one given test is actually relevant. In the same fashion, the *recall* is defined as TP / (TP + FN) and is the probability that a relevant "unordinary phenomenon" is actually detected by the procedure. Finally, the *accuracy* of the procedure if defined as (TP + TN) / (TP + TN + FP + FN) and measures the overall performance of the procedure, i.e. how close it is to the ideal procedure. Another widely used measure is the *False Discovery Rate (FDR)* that corresponds to "1 − precision".

Graphical methods. Regarding practical applications, it is important to choose the right α threshold so as to guaranty a given specificity/sensitivity or precision/recall trade-off. The *Receiver Operating Characteristic (ROC)* curve [20] (see for instance Figure 3.9) plots the paired values of specificity and sensitivity for a range of α potential values. Thus, the ROC curve corresponding to a test with random decision about the rejection or acceptance of the null hypothesis would be the identity line ($f(x) = x, \forall x \in [0, 1]$). The area under the ROC curve can be seen as a measure of the testing procedure accuracy. Similarly, the *precision-recall* curve [42] plots the paired precision and recall values associated with various α. Unlike sensitivity and specificity, there is no smooth variation of the precision and recall values together with α. Therefore, there is not straightforward interpretation of the precision-recall curve in terms of area under the curve. One has to fix the parameter α in order to obtain a precision (resp. recall) level that correspond to an acceptable associated recall (resp. precision).

Graphical methods are often used to compare testing procedures and methods: The curves corresponding to several methods are displayed on the same plot, and one can easily see if one method dominates the others, at least in a given regime of type I error control. Another appreciated property of graphical methods is that they can be built directly from any test statistic, without turning it into p-values. α is indeed an hidden parameter that can be replaced by any varying threshold.

2.1.2 Challenges in neuroimaging

Neuroimaging group analyzes are used to relate inter-subject signal differences observed in brain imaging with behavioral or genetic variables and to assess risks factors of brain diseases. There is therefore a major interest in being able to point out significant differences of brain activity occurring between varying experimental conditions. The major difficulty with neuroimaging studies lies in the inter-subject variability of brain shape and vasculature. In functional studies, a task-related variability of subject performance is also observed. The standard analytic approach is to register and normalize the data in a common reference space. However a perfect voxel-to-voxel correspondence cannot be attained, and the impact of anatomical variability is tentatively reduced by smoothing [14]. This problem holds for voxel-based statistical tests and multivariate methods that consider the similarity between brain images. In the absence of ground truth, choosing the best procedure to analyze the data is a challenging problem. Practitioners as well as methodologists tend to prefer models that maximize the sensitivity of a test under a given control for false detections. While this significance level is arbitrary, the level of sensitivity conditional to this control is indeed informative on the appropriateness of a model. As mentioned in chapter 1, the reference approach in neuroimaging is to fit and test a model at each voxel (univariate voxel wise method) that can be written as follows:

$$\boldsymbol{B} = \boldsymbol{X}_2\boldsymbol{\gamma} + \boldsymbol{\epsilon}_2, \tag{2.1}$$

where $\boldsymbol{X}_2$ is a design matrix (that encodes the covariates and the confounds), $\boldsymbol{\gamma}$ is the matrix of the model coefficients, $\boldsymbol{B}$ is a $n \times v$ matrix corresponding to the maps of n subjects, that contain v voxels each. We test the null hypothesis $H_0 : \boldsymbol{\gamma} = 0$. As mentioned in [1], an appropriate test statistic for the two-tailed test is the square of the correlation coefficient between $\boldsymbol{B}$ and $\boldsymbol{X}_2$, r_S^2, that can be compared to its theoretical distribution. If the design matrix $\boldsymbol{X}_2$ has only one column, r_S^2 is equivalent to the F-statistic that we described in chapter 1. For technical reasons, we consider r_S^2 throughout the whole section instead of the standard F used in neuroimaging. This both simplify the notations and make them consistent with that of [15] in case the reader refers to it. The theoretical results that follow can be readily extended to the case where $\boldsymbol{X}_2$ has several colums and a contrast $\boldsymbol{c}$ has to be used.

2.1.2.1 Multiple comparisons

A major problem with the testing procedure is that the large number of tests performed yields a multiple comparison problem. The sensitivity of a statistical test guarantees that false rejections will not happen too often, e.g. $\alpha = 5\%$ times. But actually, when a lot of tests are performed, say m tests, the probability of rejecting at least one true null hypothesis out of m is equal to $1-(1-\alpha)^m \simeq 1-\alpha\, m$ if $\alpha\, m \ll 1$ (see Figure 2.2). The expected number of false detections is $\alpha\, m$. In neuroimaging voxel-level analysis, it is common to have $m > 40,000$. This has a strong impact on the actual specificity of the detections. Yet, the statistical significance of the voxel intensity test can be corrected with various statistical procedures.

Bonferroni correction consists in adjusting the significance threshold by dividing it by the number of tests performed (m). As a result, the probability

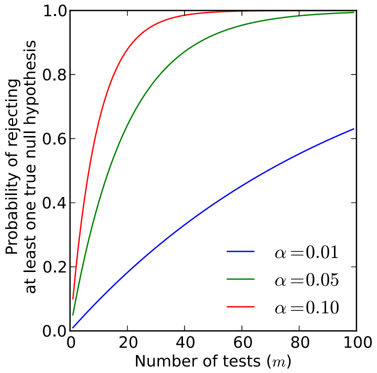


Figure 2.2: **Multiple comparisons.** The probability of rejecting at least one true null hypothesis rapidly grows with the number of performed tests (m). This is a major issue in neuroimaging since spurious detections may result in misinterpretations. Bonferroni correction solves the problem but it is accurate only in the case of independent tests (i.e. no correlation exist between the tested variables).

of rejecting at least one true null hypothesis is close to α. This approach is known to be too conservative, especially when non-independent tests are involved, which is the case of neighboring voxels in neuroimaging. Bonferroni correction yields an approximation of $\mathrm{P}(\mathrm{FP} > 0)$. Any correction of the specificity that controls this quantity is called a *Family Wise Error Rate (FWER)* correction. The *False Detection Rate* is also employed as a solution to deal with multiple comparisons. It is defined as $\mathbb{E}[\mathrm{FP} \,/\, (\mathrm{TP} + \mathrm{FP})]$ with the convention that $\mathrm{FDR} = 0$ when $(\mathrm{TP} + \mathrm{FP}) = 0$. More correction procedures specific to neuroimaging have been proposed in the literature. A good compromise between computation cost and sensitivity can be found in analytic corrections based on Random Field Theory (RFT), in which the smoothness of the images is estimated [59]. However, this approach requires both high threshold and data smoothness to be really effective [22].

2.1.2.2 Test dependence

Spatial models try to overcome the lack of correspondence between individual images at the voxel level. The most straightforward and widely used technique consists in smoothing the data to increase the overlap between subject-specific activated regions [60]. The main drawback is that the tests performed at the voxel level are not independent anymore. Under those conditions, correction procedures that control the FWER have been shown to be a bit more conservative[2]. In the literature, several approaches propose more elaborate techniques to model the noise in neuroimaging, like Markov Random Fields [37], wavelets decomposition [56], spatial decomposition or topographic methods [17, 12] and anatomically informed model [26]. These techniques are not widely used probably because they are computationally costly and not always well-suited to analysis of a group of subjects. A popular approach consists in working with subject-specific Regions of Interest (ROIs), that can be defined in a way that accommodates inter-subject variability [36] (see Chapter 4).

2.1.3 Permutations testing

Going back to model 2.1, the theoretical distribution of the test statistic r_S^2 may only be known under particular assumptions that may be be violated in

[2] http://jpktd.blogspot.fr/2013/04/multiple-testing-p-value-corrections-in.html.

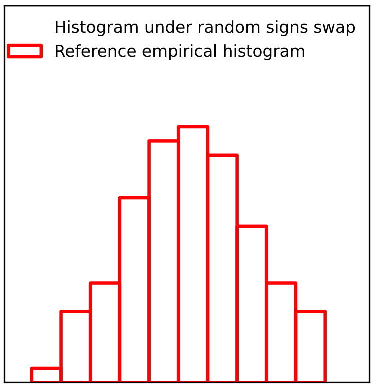


Figure 2.3: **Permutation testing.** Under symmetry assumption, the empirical distribution of the decision statistic (their histogram, in red) stays the same when the data distribution is randomly flipped with respect to 0 by swapping signs (blue shaded histograms). Here, 10 such random swaps have been performed and the corresponding new histograms have been superimposed in background. In average, one rovers the reference distribution. As a result, is is possible to generate 2^n artificial datasets from the original data while preserving the statistics distribution under the null.

practice. The specificity of the test is then impacted. Alternatively, one may want to use a specific test statistic, the theoretical distribution of which is not known under the null hypothesis. This is likely to happen for computational reasons for instance. The Monte-Carlo method [33] can be used to empirically approximate the unknown distribution: N artificial datasets are created on the model of the original data under the null hypothesis, so that N realizations of the test statistic are observed. However, it is difficult to reproduce the real-data problem with simulations. Permutation testing can be seen as a way to build the unknown H_0 distribution from the observed data. Those are transformed in such a way the decision statistics' distribution remains the same under the null hypothesis, while the tested effect is removed in the process. For instance, under the assumption $\boldsymbol{\gamma} = 0$, equation 2.1 boils down to $\boldsymbol{B} = \boldsymbol{\epsilon}_2$. Randomly swapping the sign of the data b_i does not change the distribution of the entire vector B and one can therefore generate $N = 2^n$ artificial datasets from which the null hypothesis of the statistic can be drawn (see Figure 2.3). Computing the test statistic of these 2^n datasets yields an approximation of H_0. Another standard permutation scheme is to shuffle the observations b_i while keeping the features fixed. This scheme is used by Freedman & Lane [15]: They denote π a permutation of the set of indices $\{1, \dots, n\}$. When used as a lowerscript, π refers to a variable computed using a transformed (i.e. permuted) dataset, while π as an upperscript means that the upperscripted variable is itself permuted. Thus, the usual test statistic r_S^2 becomes

$$r_\pi^2 = \frac{(\sum_{i=1}^n \boldsymbol{B}_i^\pi \boldsymbol{X}_{2_i})^2}{\sum_{i=1}^n \boldsymbol{B}_i^2 \sum_{i=1}^n \boldsymbol{X}_{2_i}^2}$$

when computed on a π-permuted dataset. The p-value associated with that test is $\mathrm{P}(r_\pi^2 \geq r_S^2)$.

2.1.3.1 Permutation schemes

The design of the permutation scheme is crucial when confounding variables (i.e. variables that are fitted but not tested) are included in the model. More precisely, the transformation that is applied to the data must preserves the statistical structure between the observations and the confounding variables, while it breaks that between the observations and the target variables. However, one also has to take into account a potential correlation between the confounding variables and the target variables. Let us consider the following linear model:

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{\alpha} + \boldsymbol{\epsilon}. \tag{2.2}$$

The observations in $\boldsymbol{Y}$ are not exchangeable under $\boldsymbol{H}_0$ because they include some portion of variability explained by $\boldsymbol{Z}$. A valid permutation test must takes this into account as the power of the associated test may indeed depend on the permutation scheme [1]. However, the difference is significant only for small sample size or when the design matrix is low rank or ill-conditioned. Note that this is likely to be the case in neuroimaging genetic studies. According to Anderson & Robison's work [1], only the permutation scheme proposed by Freedman & Lane [15] provides an exact test. They consider *(i)* the relationship between $\boldsymbol{Y}$ and $\boldsymbol{Z}$: $\boldsymbol{Y} = \boldsymbol{Z}\boldsymbol{\alpha} + \boldsymbol{R}_{\boldsymbol{Y}|\boldsymbol{Z}}$; *(ii)* the relationship between $\boldsymbol{X}$ and $\boldsymbol{Z}$: $\boldsymbol{X} = \boldsymbol{Z}\boldsymbol{\gamma} + \boldsymbol{R}_{\boldsymbol{X}|\boldsymbol{Z}}$, and they define the test statistic of no relationship between $\boldsymbol{Y}$ and $\boldsymbol{X}$, beyond their potential shared correlation with $\boldsymbol{Z}$:

$$r^2 = \frac{\left(\sum_{i=1}^n \boldsymbol{R}_{\boldsymbol{Y}|\boldsymbol{Z}_i} \boldsymbol{R}_{\boldsymbol{X}|\boldsymbol{Z}_i}\right)^2}{\sum_{i=1}^n \boldsymbol{R}^2_{\boldsymbol{Y}|\boldsymbol{Z}_i} \sum_{i=1}^n \boldsymbol{R}^2_{\boldsymbol{X}|\boldsymbol{Z}_i}},$$

Then, the permuted statistic associated with the model 2.2 would be:

$$r_E^2 = \frac{\left(\sum_{i=1}^n (\boldsymbol{Y}_\pi - \boldsymbol{Z}\boldsymbol{\alpha}_\pi)_i \boldsymbol{R}_{\boldsymbol{X}|\boldsymbol{Z}_i}\right)^2}{\sum_{i=1}^n (\boldsymbol{Y}_\pi - \boldsymbol{Z}\boldsymbol{\alpha}_\pi)_i^2 \sum_{i=1}^n \boldsymbol{R}^2_{\boldsymbol{X}|\boldsymbol{Z}_i}},$$

but since $\boldsymbol{\alpha}$ and $\boldsymbol{R}_{\boldsymbol{Y}|\boldsymbol{Z}}$ are not known in practice, Freedman & Lane proposed to replace them by their least-squares estimates $\boldsymbol{a} = \sum_{i=1}^n \boldsymbol{Y}_i \boldsymbol{Z}_i / \sum \boldsymbol{Z}_i^2$ and $\boldsymbol{R}_{\boldsymbol{Y}|\boldsymbol{Z}}$, yielding $\boldsymbol{Y}_{\pi(F)} = \boldsymbol{Z}\boldsymbol{a} + \boldsymbol{R}^\pi_{\boldsymbol{Y}|\boldsymbol{Z}}$ and the test statistic under permutation:

$$r_F^2 = \frac{\left(\sum_{i=1}^n (\boldsymbol{Y}_{\pi(F)} - \boldsymbol{Z}\boldsymbol{a}_{\pi(F)})_i \boldsymbol{R}_{\boldsymbol{X}|\boldsymbol{Z}_i}\right)^2}{\sum_{i=1}^n (\boldsymbol{Y}_{\pi(F)} - \boldsymbol{Z}\boldsymbol{a}_{\pi(F)})_i^2 \sum_{i=1}^n \boldsymbol{R}^2_{\boldsymbol{X}|\boldsymbol{Z}_i}},$$

where $\boldsymbol{a}_{\pi(F)} = \sum_{i=1}^n \boldsymbol{Y}_{\pi(F)_i} \boldsymbol{Z}_i / \sum_{i=1}^n \boldsymbol{Z}_i^2$.

An interesting point of that permutation scheme is that splitting the design matrix into testing covariates ($\boldsymbol{X}$) and confounds ($\boldsymbol{Z}$) can reduce the computation time because the effect of the confounds is computed once and for all.

2.1.3.2 Cluster-size inference

A widely used method is a test on clusters size, which aims to detect spatially extended effects [43, 18, 40]. The statistical significance of the size of an activation cluster can be obtained with theoretical corrections based on the RFT [61, 22] or with a permutation test [13, 23, 35]. Cluster-size tests tend to be more sensitive than voxel- intensity test, especially when the signal is spatially extended [34, 16, 41] at the expense of a strong statistical control on all the voxels within such clusters. This approach however suffers from several drawbacks. First, such a procedure is intrinsically unstable and its result depends strongly on an arbitrary cluster-forming threshold [16]. Second, the correlation between neighboring voxels varies across brain images, which makes detection difficult where the local smoothness is low. Combining permutations and RFT to adjust for spatially-varying smoothness leads to more sensitive procedures [22, 47]. A more complete discussion of the limitations and comparisons of these techniques can be found in [39, 34].

2.1.3.3 TFCE

The threshold-free cluster enhancement (TFCE) addresses the issue of selecting a cluster-forming threshold for cluster-size inference, by avoiding the choice of an explicit, fixed threshold [53, 47]. It however leads to other arbitrary choices such as integration step and two parameters involved in the TFCE statistic. More generally, tests that combine cluster size and voxel intensity have been proposed [41, 21]. To get p-values maps, the TFCE statistic needs to be tested with a permutation test. The FSL software provides an implementation of the method. A Python version was implemented during this thesis in order to easily plug the method with existing permutation testing frameworks available in the Parietal team code base.

2.2 Covariance estimation

2.2.1 Introduction and relevance to neuroimaging

Let us assume that empirical observations are seen as realizations of a p-dimensional random variable $\boldsymbol{X}$. Under the further assumption that $\boldsymbol{X}$ is distributed according to a unimodal distribution, it is possible to define the *location* and the *covariance* of the population described by $\boldsymbol{X}$. These quantities are the first and the second statistical moments of the distribution of $\boldsymbol{X}$ and can be considered as the best first (resp. second) order approximations of it. The location and the covariance matrix of a population convey a lot of information about the statistical structure of the data. In neuroimaging, they turn out to be useful in various applications such as study of the brain connectivity (especially in resting state studies), outlier detection, or estimation of a General Linear Model (GLM). Most of these applications actually need a good estimate of the inverse covariance matrix, called the *precision matrix.*

The limited number of observations available in practice is called a *sample*: $\boldsymbol{X}_n = \{\boldsymbol{x}_1, \dots, \boldsymbol{x}_n\}$, in which each $\boldsymbol{x}_i$ $(i \in \{1, \dots, n\})$ a p-dimensional vector describing one observation. Under the assumption that $\boldsymbol{X}$ follows a multivariate Gaussian distribution, empirical maximum-likelihood (sample) estimates of the location $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$ can be obtained from the well-known formulas:

$$\hat{\boldsymbol{\mu}}_n = \frac{1}{n}\boldsymbol{X}_n^{\mathsf{T}}\mathbf{1},$$
$$\hat{\boldsymbol{\Sigma}}_n = \frac{1}{n}(\boldsymbol{X}_n - \hat{\boldsymbol{\mu}}_n)^{\mathsf{T}}(\boldsymbol{X}_n - \hat{\boldsymbol{\mu}}_n).$$

$(\hat{\boldsymbol{\mu}}_n,\, \hat{\boldsymbol{\Sigma}}_n)$ is indeed the maximum likelihood estimate solution of the following estimation problem:

$$(\hat{\boldsymbol{\mu}}_n, \hat{\boldsymbol{\Sigma}}_n) = \underset{\boldsymbol{\mu},\boldsymbol{\Sigma}}{\operatorname{argmin}}\left(\log|\boldsymbol{\Sigma}| + \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{x}_i - \boldsymbol{\mu})^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_i - \boldsymbol{\mu})\right), \tag{2.3}$$

2.2.2 Regularization

Unfortunately, the eigenvalues of the real population covariance matrix $\boldsymbol{\Sigma}$ are only poorly approximated by those of $\boldsymbol{\Sigma}_n$, especially as the p/n ratio increases. This is likely to occur in neuroimaging as the number of image descriptors (the image voxels, usual p) is often above 40,000, for 2,000 subjects at most, causing the smallest (resp. largest) eigenvalue of $\hat{\boldsymbol{\Sigma}}_n$ to be biased downwards (resp. upwards). We reduce this effect by considering signal averages within pre-defined parcels and thus obtain values of p between 100 and 1000. $\hat{\boldsymbol{\Sigma}}_n$ can still be ill-conditioned yet. We completely correct the artifact by adding a *penalization term* to the problem 2.3:

$$(\hat{\boldsymbol{\mu}}_{n,J}, \hat{\boldsymbol{\Sigma}}_{n,J}) = \underset{\boldsymbol{\mu},\boldsymbol{\Sigma}}{\operatorname{argmin}}\left(\log|\boldsymbol{\Sigma}| + \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{x}_i - \boldsymbol{\mu})^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_i - \boldsymbol{\mu}) + \lambda J(\boldsymbol{\Sigma})\right), \tag{2.4}$$

where J is a convex function, such as a matrix norm and λ is a scalar that controls the amount of penalization. The resulting sample estimate is called a *regularized* estimate. This avoids large differences between the smallest and the largest eigenvalue of the covariance matrix estimate, but introduces some bias in the estimation. The main challenge associated with covariance matrix regularization is thus to choose the right amount of regularization (i.e. the right λ) to obtain the best compromise between *(i)* a good estimation of the covariance matrix eigenvalues (especially useful when the goal is in fact the estimation of the precision matrix via inversion of the covariance estimate), and *(ii)* a limited loss of precision regarding the estimation of the specific structure of the target $\boldsymbol{\Sigma}$ (i.e. approximating $\boldsymbol{\Sigma}$ with a diagonal matrix may yield a good approximation in terms of eigenvalues, but the structure encoded by the off-diagonal coefficients of $\boldsymbol{\Sigma}$ would be lost). The choice of the imposed structure potentially plays an important role. For instance, not all matrices can be well approximated by a diagonal matrix. We limit ourselves to diagonal structure in our applications, as the associated computations may be way less complex and time-consuming.

2.2.2.1 ℓ_2 regularization

ℓ_2 regularization corresponds to taking $J(\boldsymbol{\Sigma}) = \mathrm{tr}(\boldsymbol{\Sigma}^{-1})$ in Equation 2.4:

$$\begin{aligned}(\hat{\boldsymbol{\mu}}_{n,\ell_2}, \hat{\boldsymbol{\Sigma}}_{n,\ell_2}) = \underset{\boldsymbol{\mu},\boldsymbol{\Sigma}}{\mathrm{argmin}} \Bigg(& \log|\boldsymbol{\Sigma}| \\ & + \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{x}_i - \boldsymbol{\mu})^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_i - \boldsymbol{\mu}) + \lambda\mathrm{tr}(\boldsymbol{\Sigma}^{-1})\Bigg),\end{aligned} \tag{2.5}$$

The resulting covariance matrix is biased toward a spherical covariance matrix. This bias correspond to an underlying assumption of isotropy. The main advantage of ℓ_2 regularization is that the solution of the problem 2.5 is explicitly given by: $\hat{\boldsymbol{\Sigma}}_{n,\ell_2} = \frac{\boldsymbol{X}_n^{\mathsf{T}}\boldsymbol{X}_n}{n} + \lambda\mathrm{I}$ and $\hat{\boldsymbol{\mu}}_{n,\ell_2} = \frac{\boldsymbol{X}_n^{\mathsf{T}}\boldsymbol{1}}{n}$. Thus, problem 2.5 is fast to solve for a fixed λ and available computing time/facilities can be instead used to perform parameter selection (i.e. to choose the "optimal" λ). Much of the effort is therefore spent on choosing the λ that yields the most sensitivity in the subsequent analysis [19].

2.2.2.2 ℓ_1 regularization

The penalized-likelihood problem corresponding to ℓ_1 regularization is the following:

$$(\hat{\boldsymbol{\mu}}_{n,\ell_2}, \hat{\boldsymbol{\Sigma}}_{n,\ell_2}) = \underset{\boldsymbol{\mu},\boldsymbol{\Sigma}}{\mathrm{argmin}} \left(\log|\boldsymbol{\Sigma}| + \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{x}_i - \boldsymbol{\mu})^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_i - \boldsymbol{\mu}) + \lambda\|\boldsymbol{\Sigma}\|_{\mathrm{off}}\right), \tag{2.6}$$

where the ℓ_1 penalty $\|\boldsymbol{A}\|_{\mathrm{off}} = \sum_{i\neq j}|a_{i\,j}|$ corresponds to the ℓ_1 norm of the off-diagonal coefficients of the matrix $\boldsymbol{A}$ (note that this is not a matrix norm). The solution of this problem is known to have a sparse inverse [54]. This sparsity property is useful for interpretation of the solution in terms of graphical models. For instance in the functional neuroimaging context, not all brain regions are statistically related to each other [55]. The choice of the regularization

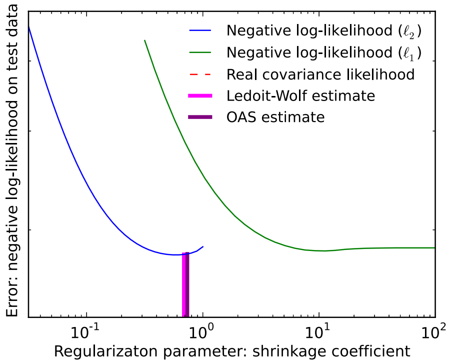


Figure 2.4: **Regularization parameter selection and associated negative log-likelihood.** Various choices for the ℓ_1 and ℓ_2 regularization parameter are considered. We observe that the negative log-likelihood reaches optimal values that do not correspond between the two different regularizations. The values yielded by Ledoit-Wolf and OAS formulas are reported. They are close to the ℓ_2 regularization optimal shrinkage.

parameter λ is particularly important, as the estimate is very sensitive to this value. When $\lambda \to \infty$ this converges to a diagonal matrix. Figure 2.4 shows the variations of the negative log-likelihood according to λ.

2.2.2.3 Ledoit-Wolf estimation

Ledoit & Wolf [27] employ convex shrinkage in order to correct the eigenvalues discrepancies between the empirical covariance matrix ($\hat{\mathbf{\Sigma}}_n$) and the population covariance matrix ($\mathbf{\Sigma}$). Thus, they build a new covariance sample estimate from a convex combination between the sample empirical covariance matrix $\hat{\mathbf{\Sigma}}_n$ and a structured estimator $\boldsymbol{M}$. A classical choice is $\boldsymbol{M} = \frac{\mathrm{Tr}(\hat{\mathbf{\Sigma}}_n)}{p}\mathrm{Id}$, and we denote $\hat{\mathbf{\Sigma}}_{n,M,\rho}$ the estimate given by $(1-\rho)\hat{\mathbf{\Sigma}}_n + \rho\boldsymbol{M}$, where $0 < \rho < 1$. Ledoit & Wolf consider as a criterion the Frobenius norm, $\|\boldsymbol{A}\|_F^2 = \mathrm{Tr}(\boldsymbol{A}^\mathsf{T}\boldsymbol{A})/p$ so their *oracle solution* for the choice of ρ is:

$$\rho* = \underset{\rho}{\arg\min}\, \mathbb{E}\left\{\|\hat{\mathbf{\Sigma}}_{n,M,\rho} - \mathbf{\Sigma}\|_F^2\right\}.$$

In [27], a closed-form solution is proposed as an approximation of $\rho*$:

$$\hat{\rho}_{\mathrm{LW}} = \min\left\{\frac{\sum_{i=1}^n \|x_i x_i^\mathsf{T} - \hat{\mathbf{\Sigma}}_n\|_F^2}{n^2\left[\mathrm{Tr}(\hat{\mathbf{\Sigma}}_n^2) - \frac{\mathrm{Tr}^2(\hat{\mathbf{\Sigma}}_n)}{p}\right]}, 1\right\},$$

yielding the corresponding covariance estimate:

$$\hat{\mathbf{\Sigma}}_{n,M,\mathrm{LW}} = \hat{\mathbf{\Sigma}}_{n,M,\rho_{\mathrm{LW}}}$$

Equivalent results were derived for various choices of $\boldsymbol{M}$ and a better approximation of ρ exists under the assumption that data are Gaussian-distributed. The work of [4] refines the latter result with the *Oracle Approximating Shrinkage (OAS)* estimate:

$$\hat{\rho}_{\mathrm{OAS}} = \min\left\{\frac{\mathrm{Tr}(\hat{\mathbf{\Sigma}}_n^2) + \mathrm{Tr}^2(\hat{\mathbf{\Sigma}}_n)}{(n+1)[\mathrm{Tr}(\hat{\mathbf{\Sigma}}_n^2) - \frac{\mathrm{Tr}^2(\hat{\mathbf{\Sigma}}_n)}{p}]}, 1\right\},$$

but $\hat{\mathbf{\Sigma}}_{n,M,\mathrm{LW}}$ is more used in practice because the Gaussian-distributed data assumption is hard to verify.

Ledoit & Wolf investigated non-linear approximation for high-dimensional covariance matrices [28, 29], but the resulting estimates require a much larger computation time and implementation efforts that are not justified in practical applications, which limits the usefulness of this approach.

2.2.2.4 Parameter selection with cross validation

The choice of the regularization parameter λ is crucial as it represents a trade-off between bias and variance. The use of a biased estimate in the context of statistical inference yields poorly sensitive results while too much variance potentially makes the statistical algorithms unstable (e.g. the inversion of a rank-deficient covariance matrix fails). There is of course no general solution to that problem and the optimal choice actually depends on the data and the targeted application. Empirical methods based on dataset splitting are therefore useful. Most of them consist in estimating a parameter on part of the available data, while the likelihood of the learned parameter is then computed on another(others) part(s) of the data. These are *cross-validation* methods. More specifically, *k-fold cross validation* starts with splitting the whole dataset into k disjoint parts of (almost) equivalent size. We then define a *training set* as the whole data set from which one of the splits (the *test set*) is taken out. A range of potential values for the λ parameter are considered, each one yielding a covariance estimate (we omit the role of location estimate for the clarity of the explanation) that can be used to compute the likelihood of the test data. Repeating the same procedure again for each of the k training sets possibilities finally results in k test set likelihood values associated with each considered value of λ. Those can be averaged according to the λ value and the λ associated with the largest average likelihood is take as the k-fold cross validation estimate of the λ parameter. For large enough values of k, this estimate is stable enough and does not require bootstrapping (see Section 2.3.1). The special case $k = n - 1$ is called *leave-one-out (cross validation)*. The most used schemes are 3-, 5- and 10-fold cross validation. To date, cross validation is the only reliable method in practice to estimate the ℓ_1 regularization parameter (see Scikit-learn documentation and examples[3]).

2.2.3 Robust estimation

Covariance estimation is only relevant in the case of unimodal data, and the more specific assumption that data are Gaussian distributed is often made. Slight departures from Gaussianity weakly affect most of the estimators presented above, but the situation becomes more critical when the hypothesis is strongly violated, for instance when strong outliers are present or when there is a strong asymmetry in the data distribution. The first assumption of unimodal data is more important because its violation strongly affects covariance estimation, as illustrated in Figure 2.5. The *finite-sample breakdown value* $\epsilon^*_n(\hat{\boldsymbol{S}}, \boldsymbol{X}_n)$ of an estimator $\hat{\boldsymbol{S}}$ at the dataset $\boldsymbol{X}_n$ is the smallest amount of contamination that can have an arbitrarily large effect on $\hat{\boldsymbol{S}}$ [8, 25]:

$$\epsilon^*_n(\hat{\boldsymbol{S}}, \boldsymbol{X}_n) \equiv \min_m \left\{ \frac{m}{n}; \sup_{\tilde{\boldsymbol{X}}_{n,m}} \|\hat{\boldsymbol{S}}(\tilde{\boldsymbol{X}}_{n,m}) - \hat{\boldsymbol{S}}(\boldsymbol{X}_n)\| = \infty \right\},$$

[3] http://scikit-learn.org.

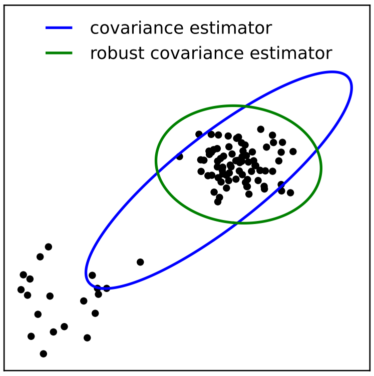


Figure 2.5: **Robust vs. non-robust covariance estimation.** If several modes are present in the data, non-robust covariance estimation breaks down. Here, the empirical sample covariance estimate (in blue) makes no sense, while the robust estimate (in green) correctly captures the shape of the main mode.

where $\widetilde{\boldsymbol{X}}_{n,m}$ is a dataset obtained by replacing m observations of $\boldsymbol{X}_n$ by arbitrary points. Then, the *breakdown point* of the estimator $\hat{\boldsymbol{S}}$ at the dataset $\boldsymbol{X}_n$ is defined as $\lim_{n \to +\infty} \epsilon_n^*(\hat{\boldsymbol{S}}, \boldsymbol{X}_n)$.
Another useful tool in robust statistics is the *influence function* of an estimator. It measures the *stability* of an estimator, i.e. how well the estimator preserves its properties (e.g. unbiasedness, consistency) under small deviation from the model assumptions. In order to define the influence function, we need to rewrite any covariance estimator $\hat{\boldsymbol{S}}$ at the dataset $\boldsymbol{X}_n$ as a function of the sample empirical distribution function, $F_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{x_i<x\}}$ and we take $F = \lim_{n\to\infty} F_n$. Finally, δ_x is the pointmass 1 at x. The influence function $IF(x, \boldsymbol{X}_n, \hat{\boldsymbol{S}})$ of an estimator $\hat{\boldsymbol{S}}$ at the dataset $\boldsymbol{X}_n = \{x_1, \dots, x_n\}$ and the observation $x \in \mathbb{R}$ is then [24]:

$$IF(x, \boldsymbol{X}_n, \hat{\boldsymbol{S}}) = \lim_{s\to 0} \frac{\hat{\boldsymbol{S}}((1-s)F + s\delta_x) - \hat{\boldsymbol{S}}(F)}{s}$$

2.2.3.1 Projection pursuit

Robust estimation of covariance can be done by first removing spurious (group of) observations with *ad hoc* techniques, and then using a non-robust estimate. Projection pursuit methods are a family of methods that project the data into some *well chosen* subspaces in which the data statistical structure appears more clearly. They are simple to understand and they rapidly provide interpretable results. Thus, projection pursuit is more a diagnosis technique than an estimation procedure and we only included them in our survey about robust procedures because they are fairly used in practice [57].

The simpler example of projection is the so-called *random projections* procedure. For example, it may be easier to separate two clusters of observations when the data are projected on a line [31]. Performing several random projections in order to have different *views* on the data may be a efficient way of rapidly investigating the basic structure of the data (and discard outliers). The latter can be explored with more clever projections. For example, Filzmoser et al. [11] performs outlier detection and covariance estimation by considering only $2p$ directions of a p-dimensional space. The choice of the directions is based on maximization of the data kurtosis. The most famous dimension reduction technique that can be used to select projection spaces is the *Principal Component Analysis (PCA)*. PCA aims at finding the directions (the *principal components* that maximize the variance of the data, when those are projected on it. Technically, PCA consists in finding the eigenvalues and eigenvectors of the data empirical covariance matrix $\boldsymbol{X}^\top\boldsymbol{X}$, but this estimate is typically poor in the presence of outliers. Covariance estimation and PCA are thus clearly related. Some estimation techniques as the *repeated median* [51] build robust location and/or covariance estimates by considering the p features describing the data one by one, but they ignore the potential correlations between the features and therefore may provide biased estimates.

2.2.3.2 M-estimators

Assuming centered data, a M-estimate of covariance [32] is defined as the solution $\hat{\boldsymbol{S}}_n$ of:

$$\boldsymbol{S} = \frac{1}{n}\sum_{i=1}^{n}\rho(x_i^\top \boldsymbol{S}^{-1}x_i)x_ix_i^\top,$$

where *(i)* $\rho : s \mapsto u(s)$ is non negative, non increasing, and continuous for $s \geq 0$; *(ii)* ρ' is bounded by a scalar K; *(iii)* ρ' is non decreasing, and strictly increasing in the interval where $\rho'(s) < K$. M-estimators include maximum likelihood estimation as a special case, but their interest is to use a function ρ that dampens the influence of observations deviating from the population pattern (assuming once again unimodal data).

Interesting results about M-estimates are given by Huber [24]. The most important seems to be that their influence function is proportional to ρ'. Thus, the choice of ρ gives control on the stability of the estimate. The breakdown point of M-estimates depends on the choice of ρ as well, but does not reach 0.5 (its maximal theoretical value). The main limitation regarding the use of M-estimators of covariance is that there is no converging algorithm to compute them. According to Huber, existing algorithms give good practical solutions[24] but they remain slow since they alternatively estimate the location (that is not zero in practice) and covariance of the data. We consider M-estimates in the context of regression in Chapter 3.

2.2.3.3 MCD

The state-of-the-art high breakdown point robust covariance estimator for multidimensional Gaussian data is Rousseeuw's Minimum Covariance Determinant (MCD) estimator [45]. Given a dataset with n p-dimensional observations, $\boldsymbol{X} \in \mathbb{R}^{n\times p}$, MCD aims at finding h observations considered as inliers, by minimizing the determinant of their scatter matrix. We refer to these observations as the *support* of the MCD.
The core procedure commonly used to compute the MCD estimate of the covariance of a population is given in Algorithm 1. It consists of alternatively choosing a subset $\boldsymbol{X}_H$ of h observations to minimize a Mahalanobis distance, and updating the covariance matrix $\hat{\boldsymbol{\Sigma}}_H$ used to compute the Mahalanobis distance. $|\hat{\boldsymbol{\Sigma}}_H|$ decreases at each update of $\boldsymbol{X}_H$. Standard algorithms such as the Fast-MCD algorithm [46] perform this simple procedure several times from

Algorithm 1 MCD estimation algorithm

```
1. Select $h$ observations (call the corresponding dataset $\boldsymbol{X}_H$);
2. Compute the empirical covariance $\hat{\boldsymbol{\Sigma}}_H$ and mean $\hat{\boldsymbol{\mu}}_H$;
3. Compute the Mahalanobis distances $d^2_{\boldsymbol{\mu}|H,\boldsymbol{\Sigma}|H}(\boldsymbol{x}_i)$, $i = 1..n$;
4. Select the $h$ observations having the smallest Mahalanobis distance;
5. Update $\boldsymbol{X}_H$ and repeat steps 2 to 5 until $|\hat{\boldsymbol{\Sigma}}_H|$ no longer decreases.
```

different initial subsets $\boldsymbol{X}_H$ and retain only the solution with the minimal determinant. The MCD can be understood as an alternated optimization of the following problem:

$$(\hat{H}, \hat{\boldsymbol{\mu}}_{\mathrm{h}}, \hat{\boldsymbol{\Sigma}}_{\mathrm{h}}) = \underset{\boldsymbol{\mu},\boldsymbol{\Sigma},H}{\operatorname{argmin}} \left(\log |\boldsymbol{\Sigma}| + \frac{1}{h} \sum_{i \in H} (\boldsymbol{x}_i - \boldsymbol{\mu})^{\mathsf{T}} \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}_i - \boldsymbol{\mu}) \right). \tag{2.7}$$

2.3 Non-parametric statistics

This section gives an overview of several well-known non-parametric algorithms that can be used for unsupervised tasks. We first present sampling schemes that we often use to build subsets of observations that have a statistical structure close than that of the original sample. We need such subsets when performing for instance Monte-Carlo simulations or random projections. We then describe various algorithms suited to the study of multimodal datasets for which the number of modes is often unknown and difficult to estimate, especially under high-dimensional settings. They can sometimes be used to extract Gaussian components from complex population structures.

2.3.1 Sampling schemes

Monte-Carlo simulations, permutation testing and the *bootstrap* method [10] consist in generating several artificial datasets by randomly drawing n observation with replacement from a dataset $\boldsymbol{X}_n$. The estimation of a statistic can be performed on each bootstrap dataset so as to obtain a distribution of the estimate. The mean (or the median) of the distribution is taken as the bootstrap covariance estimate of the dataset, and is therefore stable with regard to slight changes in $\boldsymbol{X}_n$ (see the definition of stability in section 2.2.3). *Bagging* [3] is very similar to bootstrap, with the only difference that the artificial samples may contain less than n observations. This results in a loss of efficiency for the estimates but their computation is faster. When the Gaussian-distributed data assumption does not hold, standard statistical procedures may be inaccurate. In particular, the sensitivity of most testing procedures is controlled under the Normal assumption. *Boosting* [48] can be employed in that context: Several testing procedures are used and a final consensus (or vote) is made to come to a final decision. By combining several weakly performing procedures, it is thus possible to obtain a powerful one.

It is important to note that robustness and stability are easily confused. Robustness implies a certain amount of stability (see the definition of the influence function in section 2.2.3) but stable does not mean robust. Bootstrap and

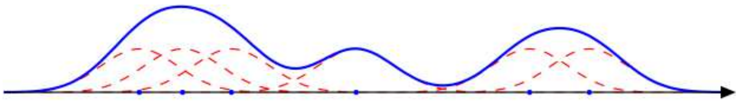


Figure 2.6: **Kernel density estimation with the Parzen-Rozenblatt estimator.** Each blue point on the axis has a smooth contribution (red dashed lines). The sum of all the contribution yields a smooth density function (in blue). The choice of the kernel defines the shape of the contributions, while the bandwidth parameter controls for their width.

ensemble methods provides stable estimates because they avoid to choose an estimate that deviates too much from the average estimate obtained for slightly different datasets. They are not robust *per se* because a gross outlier could completely shift the bootstrap distribution of an estimate, and, in turns, result in a bad bootstrap estimate.

2.3.2 Density estimation

Let $\boldsymbol{X}_n = \{x_1, \dots, x_n\}$ be a sample of n observations. We make the assumption that those observations are realizations of a random variable $\boldsymbol{X}$. *Density estimation* relates to the estimation of the distribution of $\boldsymbol{X}$, usually from the observed sample $\boldsymbol{X}_n$. A simple example of density estimation is the *cumulative density function* of $\boldsymbol{X}$: $F_n(x) = \sum_{i=1}^{n} \mathbb{1}_{\{x_i < x\}}$. In order to estimate the density function D_n and not its cumulative, we can put

$$D_n(x) = \sum_{i=1}^{n} \frac{1}{2} \mathbb{1}_{\left\{\frac{|x_i - x|}{h} < 1\right\}}, \tag{2.8}$$

where h is a parameter (the *bandwidth*) that may be chosen by cross-validation so as to offer a good compromise between a noisy (i.e. too much variance) and a flat (i.e. too much bias) estimation. Figure 2.6 provides an illustrative example of Parzen-Rozenblatt density estimation. Equation 2.8 can be generalized to the multidimensional case by rewriting it with a *kernel* K, which is an even, non-negative function verifying $\int_{+\infty}^{-\infty} K(u)du = 1$:

$$D_n(x) = \sum_{i=1}^{n} K\left(\frac{x_i - x}{h_i}\right). \tag{2.9}$$

According to the kernel K, the estimation may also be smoother than the usual histogram. The parameters of the kernel are once again chosen by cross-validation in most cases. Equation 2.9 correspond to *Parzen-Rosenblatt window estimation* [38, 44], where the kernel intuitively defines a neighborhood around each observation. The most famous kernel is the Gaussian kernel: $K(u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}u^2}$. If d-dimensional Gaussian distributed data are considered, the optimal bandwidth for the Gaussian kernel has a theoretical expression [52]: $\boldsymbol{h} = \left(\frac{4\hat{\boldsymbol{\sigma}}^{(d+4)}}{(d+2)n}\right)^{\frac{1}{d+4}}$, where $\hat{\boldsymbol{\sigma}}$ is the coordinate-wise standard deviation of the n samples. This formula can be used as a relatively good approximation of h when the approximated distribution is "as smooth as a Gaussian" and can serve as a basis for multivariate estimation as well.

Algorithm 2 Expectation-Maximization algorithm (for Gaussian mixture)

```
Define: $k$: number of components in the model;
  $\boldsymbol{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ the set of $n$ $p$-dimensional observations;
  $\boldsymbol{Z} = \{z_1, \ldots, z_n\}$ the vector of latent variables that encode the $k$ components,
  such as $\boldsymbol{x}_i|(z_i = j) \sim \mathcal{N}_p(\boldsymbol{\mu}_j, \boldsymbol{\sigma}_j)\,\forall(i,j) \in \{1,\ldots,n\} \times \{1,\ldots,k\}$;
  $\boldsymbol{\tau} = \{\tau_1, \ldots, \tau_k\}$ represents the weights of the components, and $\sum_{j=1}^{k} \tau_j = 1$.
  $L(\boldsymbol{\theta}; \boldsymbol{X}, \boldsymbol{Z})$ is the likelihood function, where $\boldsymbol{\theta} = (\boldsymbol{\tau}, \boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_k, \boldsymbol{\sigma}_1, \ldots, \boldsymbol{\sigma}_k)$
  is the quantity to be estimated by the algorithm.
Init: $\epsilon = 10^{-5}$, $\boldsymbol{\theta}^{(0)}$ = random values
  repeat
    $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) = \mathrm{E}[\log L(\boldsymbol{\theta}; \boldsymbol{X}, \boldsymbol{Z})]$        (Expectation step)
    $\boldsymbol{\theta}^{(t+1)} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\, Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$          (Maximization step)
    $t \leftarrow t + 1$
  until $|L(\boldsymbol{\theta}^{(t)}; \boldsymbol{X}, \boldsymbol{Z}) - L(\boldsymbol{\theta}^{(t-1)}; \boldsymbol{X}, \boldsymbol{Z})| < \epsilon$
  $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta}^{(t)}$
```

2.3.3 Gaussian Mixture Models

A special case of non-Gaussian distributed dataset is a dataset that combines several Normal populations (called *components*) having different locations and/or covariance parameters. Note that any distribution can be approximated by combining a finite number of Gaussian distributions with different locations and/or covariances parameters. The associated unsupervised learning challenge is to estimate the number of components, their parameters and their weights (in the case of unbalanced classes). For a fixed number of components, an *Expectation-Maximization (EM)* algorithm can be used for adjusting the parameters and weights of the components. Algorithm 2 details the EM algorithm[4]. Estimation of the number of classes is typically performed by cross-validation or by using the BIC criterion [50]. Estimation of a Gaussian Mixture Model parameters can be challenging in high-dimension because the likelihood function requires the inversion of the covariance matrix associated with each component. If the number of observations within one component is too small with respect to the number of data descriptors, the covariance matrix may be ill-conditioned or non-invertible. We refer to section 2.2 for a more complete description of the resulting computation issues.

Once a mixture model has been estimated, it can be used for various purposes such as density estimation or statistical inference. Gaussian mixture models are also useful to generate synthetic data that have a complex statistical structure. This techniques has been used several times in this thesis. For instance, outliers can easily be simulated by contaminating Gaussian distributed data with observations drawn from a wider Gaussian distribution. Another standard example is the generation of sub-Gaussian data, a problem for which no simple distribution is adapted whereas Student or Laplace distributions are super-Gaussian (i.e. "peaked" data). Figure 2.7 gives an illustration of the differences between "peaked" or "wide" distributions.

[4] source: http://en.wikipedia.org/wiki/Expectation-maximization_algorithm.

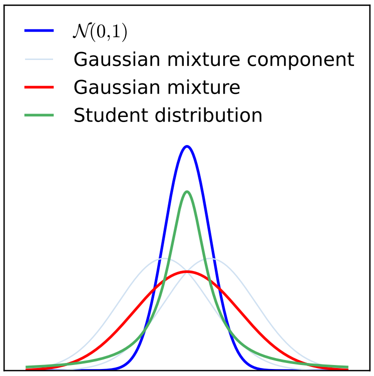


Figure 2.7: **Distributions with various widths.** The standard Normal distribution with zero location and unit variance is the most common distribution (in blue), but it is sometimes useful to consider wider or narrower distribution in simulations for real-data application. Here, we obtain a wide distribution (in red) by a mixture of two Gaussians with different location parameters (in shaded blue). A Student distribution (in green) is more "peaked" than a Gaussian.

2.3.4 Support Vector Machines

Support Vector Machines (SVMs) are mainly classification tools. The well-known linear SVM algorithm was originally designed to estimate the optimal separating hyperplane between two linearly separable classes, that is the hyperplane that can classify new data with the smaller error rate [2]. The simplest formulation of the SVM is:

$$\min_{\boldsymbol{w},\boldsymbol{b}} \frac{1}{2}\|\boldsymbol{w}\|^2$$
$$\text{s.t.} \quad y_i(\boldsymbol{w}\boldsymbol{x}_i - \boldsymbol{b}) \geq 1 \quad \forall i \in \{1, \ldots, n\},$$

where y_i is the label of the observation i (-1 or 1 in this illustrative two-classes example). $\boldsymbol{w}$ and $\boldsymbol{b}$ are the slope and the intercept of the separating hyperplane. SVMs have been extended to the case of non-separable classes (*soft-margin* SVMs) [5], with parameters $(\xi_i, i \in \{1, \ldots, n\})$ that control the trade-off between the classification error of the training sample and the classification error that would be observed on new samples (from the same distribution):

$$\min_{\boldsymbol{w},\boldsymbol{\xi},\boldsymbol{b}} \left\{ \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{i=1}^{n} \xi_i \right\}$$
$$\text{s.t.} \quad y_i(\boldsymbol{w}\boldsymbol{x}_i - \boldsymbol{b}) - \xi_i \geq 1, \quad \xi_i \geq 0 \quad \forall i \in \{1, \ldots, n\},$$

A very useful property of SVMs is that they can also handle the case of non-linear data (as in Figure 2.8), multiple classes and high-dimensional feature space. A lot of applications of SVMs go beyond the standard classification task and there are been various attempts to perform inference from SVMs. An interesting example is the One-Class SVM [49] that was inspired from regular SVMs but which is restricted to the robust estimation of one single class contours (or *frontier*): application to (semi-supervised) outlier detection have been performed in many research fields [7, 6, 58, 62]. *Support Vector Regression (SVR)* is a regression algorithm based on SVMs [9].

References

[1] Anderson, M.J., Robinson, J.: Permutation tests for linear models. Australian & New Zealand Journal of Statistics 43(1), 75–88 (2001)

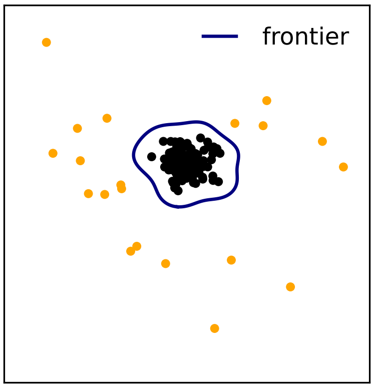


Figure 2.8: **One-Class SVM with a non-linear kernel.** Non-linear classification can be performed by first projecting the data into a higher-dimensional space where the different classes are linearly separable. Then, the frontier can be represented back in the original space as a non-linear curve, here in blue.

[2] Boser, B.E., Guyon, I.M., Vapnik, V.N.: A training algorithm for optimal margin classifiers. In: Proceedings of the fifth annual workshop on Computational learning theory. pp. 144–152. ACM (1992)

[3] Breiman, L.: Bagging predictors. Machine learning 24(2), 123–140 (1996)

[4] Chen, Y., Wiesel, A., Eldar, Y., Hero, A.: Shrinkage algorithms for MMSE covariance estimation. Signal Processing, IEEE Transactions on 58(10), 5016–5029 (oct 2010)

[5] Cortes, C., Vapnik, V.: Support vector machine. Machine learning 20(3), 273–297 (1995)

[6] Cui, W., Yan, X.: Adaptive weighted least square support vector machine regression integrated with outlier detection and its application in QSAR. Chemometrics and Intelligent Laboratory Systems 98(2), 130–135 (2009)

[7] Davy, M., Desobry, F., Gretton, A., Doncarli, C.: An online support vector machine for abnormal events detection. Signal processing 86(8), 2009–2025 (2006)

[8] Donoho, D.L., Huber, P.J.: The notion of breakdown point. A Festschrift for Erich L. Lehmann pp. 157–184 (1983)

[9] Drucker, H., Burges, C.J., Kaufman, L., Smola, A., Vapnik, V.: Support vector regression machines. Advances in neural information processing systems pp. 155–161 (1997)

[10] Efron, B.: Bootstrap methods: another look at the jackknife. The annals of Statistics pp. 1–26 (1979)

[11] Filzmoser, P., Maronna, R., Werner, M.: Outlier identification in high dimensions. Computational Statistics & Data Analysis 52(3), 1694 – 1711 (2008)

[12] Flandin, G., Penny, W.D.: Bayesian fMRI data analysis with sparse spatial basis function priors. Neuroimage 34(3), 1108–1125 (Feb 2007)

[13] Forman, S.D., Cohen, J.D., Fitzgerald, M., Eddy, W.F., Mintun, M.A., Noll, D.C.: Improved assessment of significant activation in functional magnetic resonance imaging (fMRI): use of a cluster-size threshold. Magn Reson Med 33(5), 636–647 (May 1995)

[14] Frackowiak, R., Friston, K., Frith, C., Dolan, R., Price, C., Zeki, S., Ashburner, J., Penny, W.: Human Brain Function. Academic Press, 2nd edn. (2003)

[15] Freedman, D., Lane, D.: A nonstochastic interpretation of reported significance levels. Journal of Business & Economic Statistics 1(4), 292–298 (1983)

[16] Friston, K.J., Holmes, A., Poline, J.B., Price, C.J., Frith, C.D.: Detecting activations in PET and fMRI: levels of inference and power. Neuroimage 4(3 Pt 1), 223–235 (Dec 1996)

[17] Friston, K.J., Penny, W.: Posterior probability maps and SPMs. Neuroimage 19(3), 1240–1249 (Jul 2003)

[18] Friston, K.J., Worsley, K.J., Frackowiak, R.S.J., Mazziotta, J.C., Evans, A.C.: Assessing the significance of focal activations using their spatial extent. Hum. Brain Mapp. 1, 210–220 (1993)

[19] Golub, G.H., Heath, M., Wahba, G.: Generalized cross-validation as a method for choosing a good ridge parameter. Technometrics 21(2), 215–223 (1979)

[20] Hanley, J.A., McNeil, B.J.: The meaning and use of the area under a receiver operating (ROC) curve characteristic. Radiology 143(1), 29–36 (1982)

[21] Hayasaka, S., Nichols, T.E.: Combining voxel intensity and cluster extent with permutation test framework. Neuroimage 23(1), 54–63 (Sep 2004)

[22] Hayasaka, S., Phan, K.L., Liberzon, I., Worsley, K.J., Nichols, T.E.: Nonstationary cluster-size inference with random field and permutation methods. Neuroimage 22(2), 676–687 (Jun 2004)

[23] Holmes, A.P., Blair, R.C., Watson, J.D., Ford, I.: Nonparametric analysis of statistic images from functional mapping experiments. J Cereb Blood Flow Metab 16(1), 7–22 (Jan 1996)

[24] Huber, P.J.: Robust Statistics, chap. 7, p. 149. John Wiley & Sons, Inc. (2005)

[25] Hubert, M., Rousseeuw, P.J., Van Aelst, S.: High-breakdown robust multivariate methods. Statistical Science pp. 92–119 (2008)

[26] Keller, M., Lavielle, M., Perrot, M., Roche, A.: Anatomically informed bayesian model selection for fMRI group data analysis. Med Image Comput Comput Assist Interv 12(Pt 2), 450–457 (2009)

[27] Ledoit, O., Wolf, M.: A well-conditioned estimator for large-dimensional covariance matrices. Journal of Multivariate Analysis 88(2), 365–411 (2004)

[28] Ledoit, O., Wolf, M.: Nonlinear shrinkage estimation of large-dimensional covariance matrices. The Annals of Statistics 40(2), 1024–1060 (2012)

[29] Ledoit, O., Wolf, M.: Spectrum estimation: a unified framework for covariance matrix estimation and PCA in large dimensions. Available at SSRN 2198287 (2013)

[30] Lehmann, E.E.L., Romano, J.P.: Testing statistical hypotheses. Springer Science+ Business Media (2005)

[31] Lopes, M.E., Jacob, L.J., Wainwright, M.J.: A more powerful two-sample test in high dimensions using random projection. arXiv preprint arXiv:1108.2401 (2011)

[32] Maronna, R.A.: Robust M-estimators of multivariate location and scatter. The annals of statistics pp. 51–67 (1976)

[33] Metropolis, N., Ulam, S.: The Monte Carlo method. Journal of the American statistical association 44(247), 335–341 (1949)

[34] Moorhead, T.W.J., Job, D.E., Spencer, M.D., Whalley, H.C., Johnstone, E.C., Lawrie, S.M.: Empirical comparison of maximal voxel and non-isotropic adjusted cluster extent results in a voxel-based morphometry study of comorbid learning disability with schizophrenia. Neuroimage 28(3), 544–552 (Nov 2005)

[35] Nichols, T.E., Holmes, A.P.: Nonparametric permutation tests for functional neuroimaging: a primer with examples. Hum Brain Mapp 15(1), 1–25 (Jan 2002)

[36] Nieto-Castanon, A., Ghosh, S.S., Tourville, J.A., Guenther, F.H.: Region of interest based analysis of functional imaging data. Neuroimage 19(4), 1303–1316 (Aug 2003)

[37] Ou, W., Wells, W.M., Golland, P.: Combining spatial priors and anatomical information for fMRI detection. Med Image Anal 14(3), 318–331 (Jun 2010)

[38] Parzen, E.: On estimation of a probability density function and mode. The annals of mathematical statistics 33(3), 1065–1076 (1962)

[39] Petersson, K.M., Nichols, T.E., Poline, J.B., Holmes, A.P.: Statistical limitations in functional neuroimaging. II. signal detection and statistical inference. Philos Trans R Soc Lond B Biol Sci 354(1387), 1261–1281 (Jul 1999)

[40] Poline, J.B., Mazoyer, B.M.: Analysis of individual positron emission tomography activation maps by detection of high signal-to-noise-ratio pixel clusters. J Cereb Blood Flow Metab 13(3), 425–437 (May 1993)

[41] Poline, J.B., Worsley, K.J., Evans, A.C., Friston, K.J.: Combining spatial extent and peak intensity to test for activations in functional imaging. Neuroimage 5(2), 83–96 (Feb 1997)

[42] Raghavan, V., Bollmann, P., Jung, G.S.: A critical investigation of recall and precision as measures of retrieval system performance. ACM Transactions on Information Systems (TOIS) 7(3), 205–229 (1989)

[43] Roland, P.E., Levin, B., Kawashima, R., Åkerman, S.: Three-dimensional analysis of clustered voxels in 15-o-butanol brain activation images. Hum. Brain Mapp. 1(1), 3–19 (1993)

[44] Rosenblatt, M.: Remarks on some nonparametric estimates of a density function. The Annals of Mathematical Statistics pp. 832–837 (1956)

[45] Rousseeuw, P.J.: Least median of squares regression. J. Am Stat Ass 79, 871–880 (1984)

[46] Rousseeuw, P.J., Van Driessen, K.: A fast algorithm for the minimum covariance determinant estimator. Technometrics 41(3), 212–223 (1999)

[47] Salimi-Khorshidi, G., Smith, S.M., Nichols, T.E.: Adjusting the effect of non-stationarity in cluster-based and TFCE inference. Neuroimage 54(3), 2006–2019 (Feb 2011)

[48] Schapire, R.E.: The strength of weak learnability. Machine learning 5(2), 197–227 (1990)

[49] Schölkopf, B., Platt, J.C., Shawe-Taylor, J.C., Smola, A.J., Williamson, R.C.: Estimating the support of a high-dimensional distribution. Neural Comput. 13, 1443–1471 (July 2001)

[50] Schwarz, G.: Estimating the dimension of a model. The annals of statistics 6(2), 461–464 (1978)

[51] Siegel, A.F.: Robust regression using repeated medians. Biometrika 69(1), 242–244 (1982)

[52] Silverman, B.W.: Density estimation for statistics and data analysis, vol. 26. CRC press (1986)

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

[54] Tibshirani, R.: Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society, Series B 58, 267–288 (1994)

[55] Varoquaux, G., Sadaghiani, S., Pinel, P., Kleinschmidt, A., Poline, J., Thirion, B.: A group model for stable multi-subject ICA on fMRI datasets. NeuroImage 51(1), 288–299 (2010)

[56] Ville, D.V.D., Blu, T., Unser, M.: Integrated wavelet processing and spatial statistical testing of fMRI data. Neuroimage 23(4), 1472–1485 (Dec 2004)

[57] Viviani, R., Grön, G., Spitzer, M.: Functional principal component analysis of fMRI data. Human Brain Mapping 24(2), 109–129 (2005)

[58] Widodo, A., Yang, B.S.: Support vector machine in machine condition monitoring and fault diagnosis. Mechanical Systems and Signal Processing 21(6), 2560–2574 (2007)

[59] Worsley, K.J., Evans, A.C., Marrett, S., Neelin, P.: A three-dimensional statistical analysis for CBF activation studies in human brain. J Cereb Blood Flow Metab 12(6), 900–918 (Nov 1992)

[60] Worsley, K.J., Marrett, S., Neelin, P., Evans, A.C.: Searching scale space for activation in PET images. Hum Brain Mapp 4(1), 74–90 (1996)

[61] Worsley, K.J., Marrett, S., Neelin, P., Vandal, A.C., Friston, K.J., Evans, A.C.: A unified statistical approach for determining significant signals in images of cerebral activation. Hum Brain Mapp 4(1), 58–73 (1996)

[62] Zhang, Y., Meratnia, N., Havinga, P.: Adaptive and online one-class support vector machine-based outlier detection techniques for wireless sensor networks. In: Advanced Information Networking and Applications Workshops, 2009. WAINA'09. International Conference on. pp. 990–995. IEEE (2009)

Chapter 3
The statistical structure of neuroimaging datasets: deviation from Normality and outlier detection

Contents

Contributions

[1] Fritsch, V., Varoquaux, G., Poline, J.B., Thirion, B.: Non-parametric density modeling and outlier-detection in medical imaging datasets. In: Machine Learning in Medical Imaging, vol. 7588, pp. 210–217. Springer Berlin Heidelberg (2012)

[2] Fritsch, V., Varoquaux, G., Thyreau, B., Poline, J.B., Thirion, B.: Detecting outlying subjects in high-dimensional neuroimaging datasets with regularized minimum covariance determinant. In: Medical Image Computing and Computer-Assisted Intervention – MICCAI 2011, vol. 6893, pp. 264–271. Springer Berlin Heidelberg (2011)

[3] Fritsch, V., Varoquaux, G., Thyreau, B., Poline, J.B., Thirion, B.: Detecting outliers in high-dimensional neuroimaging datasets with robust covariance estimators. Medical Image Analysis 16(7), 1359 – 1370 (2012)

3.1 Outliers in neuroimaging

3.1.1 Provenance and influence

Medical image acquisitions are prone to a wide variety of errors such as scanner instabilities, acquisition artifacts, or issues in the underlying bio-medical experimental protocol. In addition, due to the high variability observed in populations of interest, these datasets may also contain uncommon, yet technically correct, observations. In both cases, images deviating from normality are called *outliers*. Outliers may be numerous, especially in neuroimaging, where the between-subjects variability of anatomical and functional features is very high and images can have a low signal-to-noise ratio. The inclusion of overly noisy or aberrant images in medical datasets typically results in additional analysis and interpretation challenges. In particular, outliers have been show to have a dramatic influence in standard statistical procedures such as Ordinary Least Squares regression [19, 44], clustering [9, 15], manifold learning [57] or neuroimaging group analyzes [25, 32]. Figures 3.1 and 3.2 illustrate this on an example dataset on the which a linear regression and a principal component analysis are performed. In both cases, there is a clear discrepancy between the results obtained on a clean dataset and the results obtained on a contaminated dataset.

Figure 3.3 shows the results of a one-sample F-test performed on a dataset including 100 inliers subjects and 20 strong outliers. The same analysis was

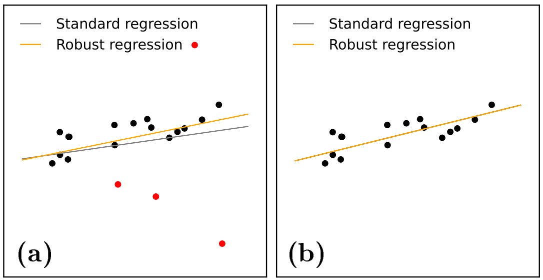


Figure 3.1: **Influence of outliers in a regression task. (a)** Contaminated dataset. **(b)** Uncontaminated dataset. Outliers yield huge residuals that standard regression tries to minimize so as to diminish the sum of squared residuals criterion. Robust regression can accommodate the presence of outliers and is almost equivalent to standard regression when no outlier is present in the dataset.

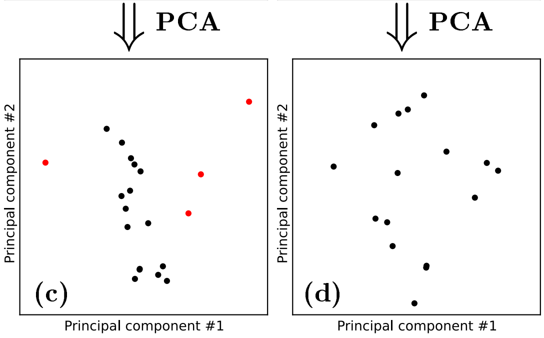


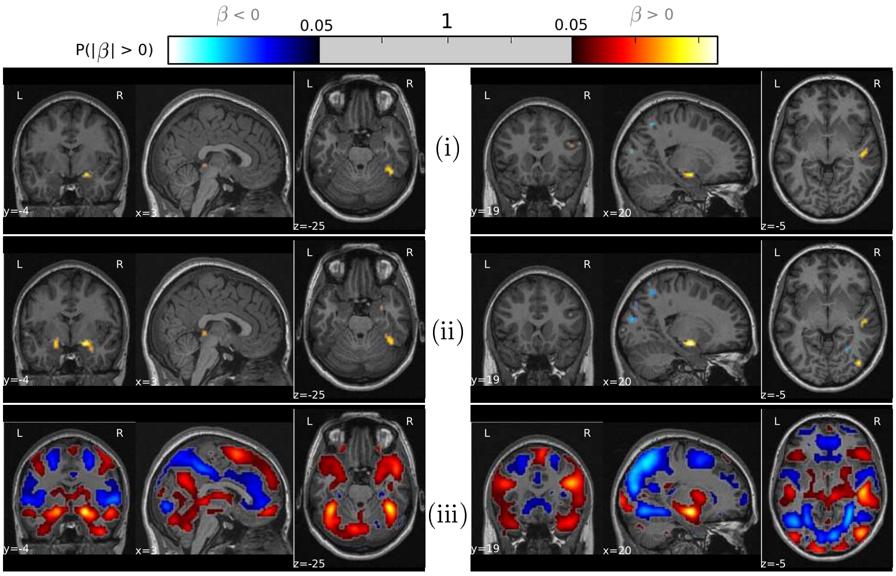
Figure 3.2: **Influence of outliers on a Principal Component Analysis. (c)** Contaminated dataset. **(d)** Uncontaminated dataset. The two plots does not have the same shape. In the first one, one component seems to explain outliers. Thus, the statistical structure of the data is hidden by the presence of outliers.

also performed on the same dataset after outliers removal. Results of both analyzes were compared to a group analysis performed on 1414 inliers subjects[1]. Activation in the left Globus Pallidus was missed in the contaminated set, but was detected after outlier removal. Also, activation in the right occipital cortex was only found from the latter dataset. Although it was obtained from less subjects (resulting in a statistical power loss), the group activation pattern for the "cleaned" group better reflects the activity pattern of the whole dataset, showing a stronger effect in every activated regions than the group map obtained from the contaminated set.

3.1.2 Common practices in neuroimaging

An intrinsic difficulty of outlier detection in medical imaging lies in the lack of formal definition for abnormal data; in particular, no generative model for outliers might be sufficient to model the variety of situations where such data are observed in practice. Moreover, in high-dimensional settings, i.e. when the number of observations is less than five times the number of data descriptors (or *features*) [16], the problem of outlier detection is ill-posed since it becomes very difficult to characterize deviations from normality. From a practical perspective, manual outlier detection is impossible in such a situation. Current methods dealing with outliers in a high-dimensional context are essentially univariate methods, i.e. they consider different dimensions one by one [39, 58]. These methods may fail to tag as outliers observations that are deviant with respect to a combination of several of their characteristics, but for which each descriptor considered individually does not reveal deviation from normality. Medical imaging data, and in particular neuroimaging data, are high dimensional, the underlying dimension being the number of degrees of freedom in their variance,

[1] The details about how we could tag subjects as in- or outliers can be found throughout this chapter.



(x = 3mm, y = $-$4mm, z = $-$25mm) cut. (x = 20mm, y = 19mm, z = $-$5mm) cut.

Figure 3.3: **Illustration of the benefit of removing outliers.** Group activity map (two-sided test for a null intercept hypothesis $\beta = 0$, rejected at $P < 0.05$ level, family-wise corrected) for the angry faces viewing task of the Imagen database (see section 1.4 in chapter 1) on *(i)* a reduced dataset containing 100 inlier subjects and the 20 strongest outlier subjects, *(ii)* the same dataset with outliers removed according to RMCD-ℓ_2 method, *(iii)* the full dataset with outliers removed according to RMCD-ℓ_2 method. The results of the second row, obtained after removal of the outliers, are closer to the full dataset group analysis than the results of the first row. This illustrates the adverse consequences of including outliers in group-level inference.

which can be of the order of the number of image voxels. This is typically much larger than the number of available samples, although parcel-level representations (thanks to parcellations and local signal averaging) reduce this issue and are therefore systematically used in our work. In functional MRI studies, neuroscientists often screen the data manually (see e.g. [40]), because of the lack of an adapted outlier detection framework. The criteria for discarding data are not always quantitatively defined. For instance, images may be discarded if, upon visual inspection, they do not reflect the expected brain activation pattern (e.g. in a so called contrast map). Such a process is tedious and unreliable, but most importantly it makes the statistical analysis of the group data invalid for that pattern –as it implies that the variance of this pattern will be underestimated. While the robust statistics literature generally considers that problems with a number of dimensions comparable to the number of observations cannot be addressed in model-based approaches, we investigate whether outlier detection is still possible in that setting. We consider both parametric and non-parametric approaches and discuss their pros and cons regarding practical application in neuroimaging. However, we insist on the fact that outlier detection cannot be substituted to the use of robust statistical inference in neuroimaging. Experiments in chapter 5 demonstrate that the combination of both techniques improves sensitivity as *(i)* a perfect outlier detection cannot be attained, especially under high-dimensional settings ; *(ii)* there are other sources for deviation from the model assumptions than the mere presence of gross outliers.

3.2 Covariance-based outlier detection

This section deals with covariance-based outlier detection methods, in which the covariance matrix of the population is estimated and then used to compute an outlier score for each observation. Using a covariance estimate relies on the assumption that regular observations, called the *inliers*, are Gaussian distributed, and that outliers are characterized by some distance to the standard model.

3.2.1 State-of-the-art

Assuming a high-dimensional Gaussian model, an observation $\boldsymbol{x}_i \in \mathbb{R}^p$ within a set $\boldsymbol{X}$ can be characterized as outlier whenever it has a large Mahalanobis distance to the mean of the data distribution, defined as $d^2_{\boldsymbol{\mu},\boldsymbol{\Sigma}}(\boldsymbol{x}_i) = (\boldsymbol{x}_i - \boldsymbol{\mu})^\mathsf{T}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_i - \boldsymbol{\mu})$, $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ being respectively the dataset location and covariance. Crucially, robust estimators of location and covariance must be used to compute these distances [8, 38].

3.2.1.1 Minimum Covariance Determinant

The state-of-the-art robust covariance estimator for multidimensional Gaussian data is Rousseeuw's Minimum Covariance Determinant (MCD) estimator [43], presented in chapter 2. Given a dataset with n p-dimensional observations, $\boldsymbol{X} \in \mathbb{R}^{n\times p}$, MCD aims at finding h observations considered as inliers, by minimizing the determinant of their scatter matrix. We refer to these observations as the *support* of the MCD. We recall the alternated optimization problem associated with the MCD:

$$\begin{aligned}(\hat{H}, \hat{\boldsymbol{\mu}}_\mathrm{h}, \hat{\boldsymbol{\Sigma}}_\mathrm{h}) = \operatorname*{argmin}_{\boldsymbol{\mu},\boldsymbol{\Sigma},H} \Bigg(& \log|\boldsymbol{\Sigma}| \\ & + \frac{1}{h}\sum_{i\in H}(\boldsymbol{x}_i - \boldsymbol{\mu})^\mathsf{T}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_i - \boldsymbol{\mu})\Bigg),\end{aligned} \tag{3.1}$$

The limitations of the MCD come from the fact that the scatter matrix must be full rank, as it is used to define a Mahalanobis distance. As a consequence, h must be greater than $h_\mathrm{min} = \frac{n+p+1}{2}$: the MCD cannot learn the inlier distribution if there are less than h_min inliers. In high-dimensional settings, as $\frac{p}{n}$ becomes large, h_min increases and outliers are potentially included in the covariance estimation if there are more than $\frac{n-p-1}{2}$ of them. When $p = n-1$, the MCD estimator is equivalent to the unbiased maximum likelihood estimator, which is not robust. Finally, if $p \geq n$, the MCD estimator is not defined. In practice, the MCD is not recommended when $\frac{p}{n} > 0.2$. To address these issues we propose to use half of the observations in the support ($h = \frac{n}{2}$) and compensate the shortage of data for covariance estimation with regularization, referred to as Regularized MCD in the remainder of the text.

Gaussian Mixture Models (GMM, see section 2.3.3) suffer from the same limitations than the MCD in high-dimension as the algorithm that is used to fit GMM require the estimation and the inversion of one covariance per component. Regularized versions of GMM exist [59, 52] but the associated algorithms require to estimate as many regularization parameters than there are components in the model, which increases a lot the complexity of the procedure. To our knowledge, no practical solution has been developed yet for this promising tool.

3.2.1.2 Distribution of robust Mahalanobis distance

A crucial part of covariance-based outlier detection is the derivation of a threshold on the Mahalanobis distances that helps performing a statistically controlled decision at the τ type I error maximum level. For any random variable $X \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, it is a well known result that $d^2_{\boldsymbol{\mu},\boldsymbol{\Sigma}}(\boldsymbol{X}) \sim \chi^2_p$. Similar result exists for the distribution of $d^2_{\hat{\boldsymbol{\mu}},\hat{\boldsymbol{\Sigma}}}(\boldsymbol{X})$, and [18] derived a theoretical formula approaching the distribution of the MCD-based Mahalanobis distances for the observations that were not part of the MCD's support (the one within are distributed according to [18]). But since the latter approximation only holds for large sample sizes, performing Monte-Carlo simulations remains the reference method to assess the distribution of $d^2_{\hat{\boldsymbol{\mu}}_h,\hat{\boldsymbol{\Sigma}}_h}(\boldsymbol{X})$: Considering a $n \times p$ dataset on which outlier detection has to be performed, the MCD covariance estimate $\hat{\boldsymbol{\Sigma}}_h$ can be used to generate Gaussian distributed data from which a new $\hat{\boldsymbol{\Sigma}}_h$ can be estimated, together with the distribution of the ensuing Mahalanobis distances. Repeating this scheme several times, we obtain a tabulation of the MCD Mahalanobis distance distribution function under the current setting. We verified that this procedure was more accurate than the small-sample correction proposed by Pison et al. [41].

3.2.2 Regularized Minimum Covariance Determinant

We first investigate outlier detection with estimators resulting from a penalized version of the likelihood in Equation 3.1. This corresponds to replacing the step 2 of the MCD Algorithm 1 by a penalized maximum-likelihood estimate of the covariance matrix.

3.2.2.1 RMCD-ℓ_2

We consider ℓ_2 *regularization* (or *ridge regularization*): let $\lambda \in \mathbb{R}^+$ be the amount of regularization, and $\hat{\boldsymbol{\Sigma}}_{\mathrm{r}}|H$ the covariance estimate of a $n \times p$ dataset $\boldsymbol{X}_H$ that maximizes the penalized negative log-likelihood:

$$\begin{aligned}(\hat{\boldsymbol{\mu}}_{\mathrm{r}}, \hat{\boldsymbol{\Sigma}}_{\mathrm{r}}|H) = \operatorname*{argmin}_{\boldsymbol{\mu},\boldsymbol{\Sigma}} \Bigg(& \log|\boldsymbol{\Sigma}| + \lambda \operatorname{Tr} \boldsymbol{\Sigma}^{-1} \\ & + \frac{1}{h} \sum_{i \in H} (\boldsymbol{x}_i - \boldsymbol{\mu})^\mathsf{T} \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}_i - \boldsymbol{\mu}) \Bigg),\end{aligned} \tag{3.2}$$

yielding $\hat{\boldsymbol{\Sigma}}_{\mathrm{r}}|H = \frac{\boldsymbol{X}_H^\mathsf{T}\boldsymbol{X}_H}{h} + \lambda\,\mathrm{I}$ and $\hat{\boldsymbol{\mu}}_r|H = \frac{\boldsymbol{X}_H^\mathsf{T}\mathbf{1}}{h}$. We denote the corresponding estimator RMCD-ℓ_2. The covariance estimate is biased toward a spherical covariance matrix. This bias corresponds to an underlying assumption of isotropy. If the inlier distribution strongly violates this prior, the bias may introduce outliers in the estimator's support.

3.2.2.2 RMCD-ℓ_1

We build another regularized version of the MCD using the ℓ_1 penalty $\|\boldsymbol{A}\|_{\text{off}} = \sum_{i \neq j} |a_{i\,j}|$ that corresponds to the ℓ_1 norm of the off-diagonal coefficients of the matrix $\boldsymbol{A}$ (note that this is not a matrix norm) in the expression of the penalized negative log-likelihood at step 2 of Algorithm 1:

$$\begin{aligned}(\hat{\boldsymbol{\mu}}_{\ell_1}, \hat{\boldsymbol{\Sigma}}_{\ell_1}|H) = \underset{\boldsymbol{\mu},\boldsymbol{\Sigma}}{\operatorname{argmin}}\Bigg(& \log|\boldsymbol{\Sigma}| + \lambda\,\|\Sigma^{-1}\|_{\text{off}} \\ & + \frac{1}{h}\sum_{i \in H}(\boldsymbol{x}_i - \boldsymbol{\mu})^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_i - \boldsymbol{\mu})\Bigg).\end{aligned} \tag{3.3}$$

We denote the corresponding estimator RMCD-ℓ_1. The solution of the problem 3.3 is known to have a sparse inverse [53]. This sparsity property is useful for interpretation of the solution in terms of graphical models. For instance in the functional neuroimaging context, not all brain regions are statistically related to each other [55]. Since no closed form solution exists for the problem (3.3), we use the GLasso algorithm [14], implemented in the scikit-learn package [36].

3.2.2.3 Setting the regularization parameter (λ)

For RMCD-ℓ_2 and RMCD-ℓ_1, the λ parameter has to be chosen carefully to obtain the right trade-off between ensuring the invertibility of the estimated covariance matrix and not introducing too much bias in the estimator. If $\lambda = 0$ we recover the MCD estimator and its limitations. On the contrary, if λ is very large, the data structure is not taken into account since the distance becomes then the Euclidean distance to the data mean. We report here three strategies that we investigated to set the shrinkage parameter:

i) The first strategy is based on *likelihood* maximization under the Gaussian distribution model for the inliers. Starting with an initial guess for $\lambda = \frac{1}{n\,p}\text{Tr}(\hat{\boldsymbol{\Sigma}})$ where $\hat{\boldsymbol{\Sigma}}$ is the unbiased empirical covariance matrix of the whole dataset, we isolate an uncontaminated set of $\frac{n}{2}$ observations that correspond to the RMCD's support. Let $\lambda = \frac{\delta}{n\,p}\text{Tr}(\hat{\boldsymbol{\Sigma}}_{\text{pure}})$, where $\hat{\boldsymbol{\Sigma}}_{\text{pure}}$ is the empirical covariance matrix of the uncontaminated dataset. We choose δ so that it maximizes the ten-fold cross-validated log-likelihood of the uncontaminated dataset. Since we use cross-validation, we refer to the ℓ_2-regularized version of the MCD by RMCD-$\ell_{2(\text{cv})}$. We also used this strategy for the choice of the RMCD-ℓ_1 shrinkage parameter, since the subsequent strategies are not adapted to the ℓ_1 case.

The two other strategies are based on convex shrinkage, where the estimated covariance matrix can be expressed as $(1-\alpha)\hat{\boldsymbol{\Sigma}} + \frac{\alpha}{p}\text{Tr}(\hat{\boldsymbol{\Sigma}})\boldsymbol{I}$. *ii)* O. Ledoit and M. Wolf [31] derived a closed formula for the shrinkage coefficient α that gives the optimal solution in terms of Mean Squared Error (MSE) between the real covariance matrix to be estimated and the shrunk covariance matrix (see section 2.2.2.3). *iii)* In a recent work, Chen et al. [7] derived another closed formula that gives a smaller MSE than Ledoit-Wolf formula under the assumption that the data are Gaussian distributed. They called it the *Oracle Approximating Shrinkage estimator (OAS)* (see section 2.2.2.3). We adapt these results to set the regularization parameter of our MCD ℓ_2-regularized version by taking $\lambda = \frac{\alpha^\star}{p\,(1-\alpha^\star)}\text{Tr}(\hat{\boldsymbol{\Sigma}})$ for $\alpha^\star$ obtained by Ledoit-Wolf and OAS formulas

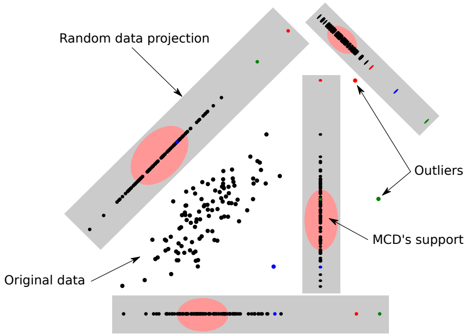


Figure 3.4: **Robust covariance estimation with random projections.** Illustration showing the RMCD-RP estimation framework. Initial data are projected on random subspaces of (fixed) lower dimension, so that a standard Minimum Covariance Determinant estimator can be used in each projected dataset. The geometric mean of the p-values obtained in the projected subspaces gives a final decision score.

applied to the uncontaminated set, respectively yielding estimators that we refer to as RMCD-$\ell_{2(\mathrm{lw})}$ and RMCD-$\ell_{2(\mathrm{oas})}$ estimators.

Outlier detection with RMCD-$\ell_{2(\mathrm{cv})}$ and RMCD-$\ell_{2(\mathrm{oas})}$ systematically yield an accuracy lower than or equal to RMCD-$\ell_{2(\mathrm{lw})}$. This is explained by the additional hypothesis required by OAS and cross-validation with respect to Ledoit-Wolf approach, and by the suboptimal cross-validation scheme. This finding suggests that the cross-validated likelihood may not be optimal as a criterion for choosing the RMCD-ℓ_1's shrinkage parameter and that we do not know how to set this parameter in practice. In the sequel, we restrict ourselves to using RMCD-$\ell_{2(\mathrm{lw})}$ that we refer to as RMCD-ℓ_2.

3.2.2.4 RMCD-RP

Another way to regularize the MCD estimator in a high-dimensional context is to run it on datasets of reduced dimensionality via random projections as illustrated in Figure 3.4. This dimensionality reduction is done by projecting to a randomly selected subspace of dimension $k < p$. Outlier detection can be performed with the MCD on the projected data if k/n ratio is small enough. Since the choice of the projection subspace is crucial for detection accuracy, the procedure has to be repeated several times in order not to miss the most discriminating subspaces. In our experiments, the results of the detections were averaged using the geometric mean of the p-values obtained in the different projections.

Setting the subspace dimension The choice of the dimension k of the projection subspace is crucial. A too small value of k results in a large loss of information during the projection step and thus raises the issues encountered with the univariate method. On the other hand, for large values of k, the geometry is preserved but the method might suffer the same issues as the MCD, even though the dimensionality reduction should make RMCD-RP more robust. We performed several outlier detection experiments with various choices for the value of k between $p/10$ and p. Our observation was that taking $k = p/5$ was a good trade-off (see Figure 3.5). This choice furthermore ensures that the RMCD-RP-based outlier detection method will be applicable for p/n ratios up to 1, since the underlying MCD-based outlier detections take place in a context where the MCD is computationally stable ($k/n < 0.2$).

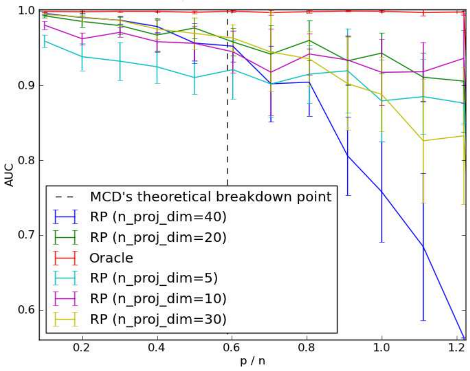


Figure 3.5: **Influence of the projection subspace dimension on RMCD-RP's accuracy (as measured by AUC curves, see section 3.4.2.1).** Outlier detections performed with various values for the projection subspaces dimension in RMCD-RP. Details about the experimental settings are discussed in section 3.4.2.1. Variance outliers ($p = 100$, condition number $= 100$, $a = 1.15$, $\gamma = 20\%$). $k = p/5$ seems a reasonable choice (corresponding here to $k = 10$) for a various set of experimental settings.

Setting the number of projections While too many random projections is computationally costly for a limited gain, too few projections may miss a good *angle* of the dataset. Outlier detection experiments convinced us that a number of projections equal to the number of dimensions is enough to explore the whole working space while being computationally tractable: further increase of this parameter does not improve the performance of the RMCD-RP method.

3.2.2.5 Statistical decision from Mahalanobis distances

Monte-Carlo simulations can be applied to assess RMCD-ℓ_2's Mahalanobis distances distribution, in the same fashion as discussed in subsubsection 3.2.1.2. We adapted it to RMCD-RP in the following manner:

1. We tabulate the distribution F_{X_k} of the MCD-based Mahalanobis distance under $n \times k$ settings (k is the dimension of the projection subspaces);

2. We take τ/p as the new accepted error level as the number of random projections is equal to p;

3. Taking $d^* = F_{X_k}^{-1}(1 - \tau/p)$, define every observations with Mahalanobis distance greater than d^* in at least one subspace as outlier.

Despite the approximation made at step 2 of the previous procedure, Figure 3.6 shows the proportion of type I errors made by the RMCD-RP for a desired theoretical value of $\tau = 0.05$ under various p/n settings. The final decision is a bit conservative but still relevant.

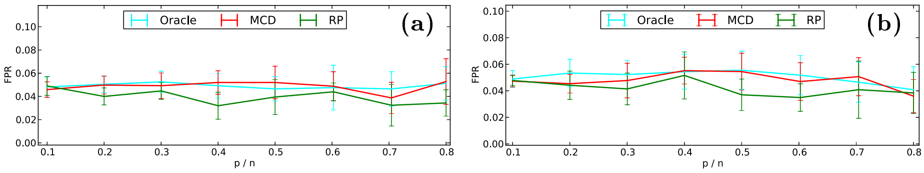


Figure 3.6: **Statistical control of outlier detection from the Mahalanobis distances of robust covariance estimators.** Proportion of detected outliers on a clean Gaussian distributed dataset at $P < 0.05$ uncorrected. **(a)** $\kappa(\mathbf{\Sigma}) = 1$. **(b)** $\kappa(\mathbf{\Sigma}) = 1000$. Type I error rate of RMCD-ℓ_2 and RMCD-RP is close to the nominal value of 0.05 uncorrected chosen in this example.

3.3 Non-parametric outlier detection

Medical imaging data is not necessarily well described by a Gaussian distribution. Thus it might be profitable to seek decision rules not based on Mahalanobis distances to screen deviant data. So far, applications of density-based outlier detection methods in neuroimaging have been mostly restricted to the detection of pathological observations, such as patients amongst healthy individuals or tumors detection. These applications involve example cases or at least some input from the practitioner and therefore fall into the category of (semi-)supervised problems. Here, we consider an unsupervised task because we have no prior knowledge on the form nor the number of outliers.

3.3.1 Density-based outlier detection

Assuming that we have a density model associated with the observations space, outliers can be defined as observations lying in low density region. This definition is the ground for *density-based outlier detection.* Density models are generally defined from distances to the neighboring observations [3, 6, 35, 12]: The (average) distance of an observation to its nearest neighbors is converted into a score, according to the which it can be decided whether the observation is an outlier or not. Alternative methods use angles [30] or rules [13] to compute outlier scores. *Distance-based outlier (DB-outlier)* [26] and *local outlier factor (LOF)* [6] are the most used non-parametric outlier detection methods; their scores are based on the distances to the k-nearest neighbors. These methods have been declined in many variants [2, 4, 27, 37, 42, 21, 22, 50, 51]. They have been merged in the *local distance-based outlier detection approach (LDOF)* [60], with variants for high-dimensional data [56, 28]. The weakness of density-based outlier detection algorithms is that their associated scores are not interpretable and cannot be readily converted into decision statistics [29]. Most of the methods require arbitrary choices and involve transformations of the k-nearest neighbors distance such that the resulting score is not interpretable anymore. In our work, we perform density-based outlier detection with standard machine learning algorithms that behave well in high dimension. These algorithms estimate density models instead of mere scores. One can then directly compute new scores for novel observations (*novelty detection*) and transform the model with algebraic manipulations (e.g. one can filter the density model with trace norm penalization in order to identify outliers, see section 3.3.3).

3.3.2 One-Class SVM

We first consider *unsupervised* outlier detection with the One-Class SVM (see section 2.3.4). This choice was originally motivated by the fact that other robust, high-dimensional, non-parametric tools such as Robust PCA [20] or Local Component Analysis [45] had not yet been considered in practical applications. The One-Class SVM novelty detection algorithm [46] is a clustering algorithm that relies on a thresholded Parzen windows density estimator (see section 2.3.4) to define a frontier around a population from a set of representative observations. It is not limited by any prior shape of the separation between in- and outlying observations. Mourao-Miranda et al. [33] showed that the One-Class SVM can be trained to detect patients, provided an initial cohort of healthy subjects is available. Let us remind that the One-Class SVM algorithm solves the following quadratic program:

$$\min_{w \in F, \boldsymbol{\xi} \in \mathbb{R}^n, \rho \in \mathbb{R}} \frac{1}{2}\|w\|^2 + \frac{1}{\nu\, n}\sum_i \xi_i - \rho \tag{3.4}$$

$$\text{subject to} \quad (w \cdot \Phi(\boldsymbol{x}_i)) \geq \rho - \xi_i, \quad \xi_i \geq 0 \tag{3.5}$$

where Φ is a feature map $\mathbb{R}^p \to F$ verifying $K(\boldsymbol{x}, \boldsymbol{y}) = \Phi(\boldsymbol{x}) \cdot \Phi(\boldsymbol{y})$ for any observations $\boldsymbol{x}$, $\boldsymbol{y} \in \mathbb{R}^p$ and a given kernel K. The One-Class SVM algorithm needs to be trained on observations that are representative of the whole population so that the learned frontier generalizes well to new subjects. This corresponds to a *supervised* setting.

Outlier detection is an *unsupervised problem*. The important parameter of the One-Class SVM is the margin parameter ν, which is both an upper bound on the proportion of observations that lie outside the frontier learned by the algorithm and a lower bound on the number of support vectors of the model [46]. In our experiments on simulated data, we set ν to the amount of contamination (i.e. the proportion of outliers in the dataset). Note that this choice favors the One-Class SVM compared to methods that ignore the ratio of outliers. For real data experiments, we set $\nu = 0.5$ as we work with at most 50% contamination. We use a *Radial Basis Function* (RBF) kernel and select its inverse bandwidth σ with an heuristic inspired by [48]: $\sigma = \frac{0.01}{\Delta}$, where Δ is the 10^{th} percentile of the pairwise distances histogram of the observations. We verified that this heuristic is close to the optimum parameter on simulations, although the results are not very sensitive to mild variations of σ around this value. We use the distance to the frontier as an outlier score on which we need to set a manual threshold in order to take a decision.

3.3.3 Local Component Analysis (LCA)

3.3.3.1 Description

Local Component Analysis is another extension of Parzen windows density estimation where the isotropic assumption inherent in most kernels is relaxed to anisotropic covariance parameters. The θ parameter of the kernel hence becomes the local data covariance matrix $\boldsymbol{\Sigma}$, which we estimate using a leave-one-out cross-validation scheme as in [45]:

$$\boldsymbol{\Sigma}^* = \underset{\boldsymbol{\Sigma}}{\operatorname{argmin}} \left[-\sum_{i=1}^{n} \log \left(\frac{|\boldsymbol{\Sigma}|^{-\frac{1}{2}} (2\pi)^{-\frac{p}{2}}}{n-1} \sum_{j \neq i} \exp\left(-\frac{1}{2} d^2_{\boldsymbol{x}_j, \boldsymbol{\Sigma}}(\boldsymbol{x}_i) \right) \right) \right]. \tag{3.6}$$

3.3.3.2 Setting LCA's regularization parameter

In Leroux et al.'s paper [45], an internal regularization term is used to ensure LCA computation stability. The proposed default value is set to $\lambda = 10^{-4}$, without further discussion about the influence nor the choice of this parameter. In our work, we choose λ so that it models properly the central mode of the data. Since outliers may have a large influence on the observed variance of the dataset along some dimensions, we use a robust heuristic: we select the 50% most

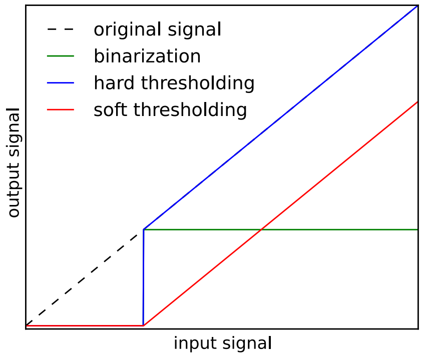


Figure 3.7: **Thresholding functions.** Binarization is used in chapter 4 to compute the RPBI statistic. We use soft thresholding for non-parametric outlier detection, yielding the Δ function. Soft thresholding is the only convex transformation amongst the three presented in the figure. This property yields more stability to the results.

concentrated observations according to a Parzen windows density estimation, compute the Ledoit-Wolf [31] coefficient shrinkage α from this subsample, and set $\lambda = \frac{\alpha}{1-\alpha}$. Leroux et al. mentioned outlier detection as a potential application of LCA, but no further investigation has been performed to our knowledge.

3.3.3.3 Building an interactive outlier detection framework.

We propose an efficient procedure to summarize the necessary information about the data structure so that the practitioner can find how many observations to discard: Within the LCA computation, proximity measures of each observation from another are computed as $k_{ij} = \exp\left(-\frac{1}{2}(x_i - x_j)^\top \boldsymbol{\Sigma}^{*-1}(x_i - x_j)\right)$, thus providing a kernel-based representation of the data as a symmetric positive definite matrix $\boldsymbol{K} = (k_{ij})_{i,j\in[1..n]^2}$, that summarizes the whole data set structure. Let $\boldsymbol{K} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{U}^\top$, where $\boldsymbol{U}$ and $\boldsymbol{D}$ are the matrix of the eigenvectors and diagonal matrix of eigenvalues $(\sigma_i, i \in [1..n])$ of $\boldsymbol{K}$. Let $\boldsymbol{D}_\delta$ be the diagonal matrix obtained by shrinking the elements of $\boldsymbol{D}$ by a factor $\delta \geq 0$ like in [57]: $\boldsymbol{D}_\delta(i,i) = 1 - \frac{\delta}{\sigma_i}$ if $\sigma_i > \delta$, $\boldsymbol{D}_\delta(i,i) = 0$ otherwise. This transformation is called a *soft thresholding* (see Figure 3.7). D_δ yields a shrunk density estimate at each observation $g_\delta(x_i) = \boldsymbol{e}_i^\top \boldsymbol{U}\boldsymbol{D}_\delta\boldsymbol{U}^\top \boldsymbol{e}$, where $\boldsymbol{e}_i$ is a vector whose entries are 0 except its i-th element which is 1 and $\boldsymbol{e}$ is a vector of ones; note that the normalization constant is omitted as it plays no role in our analysis. We finally define $\Delta(x_i) = \min_\delta\{\delta : g_\delta(x_i) < 0.5\}$, which associates each observation with the minimal shrinkage value δ that –almost– cancels it. Δ can be further used to identify different levels of homogeneity amongst the data. Typically, outliers would correspond to a group of observations that vanish with the smallest values of δ, whereas larger δ also trim off regular observations as in Figure 3.14.

We define the *disappearance function*, a ranked version of Δ, as: $\Delta_{\text{rank}}(i) = \Delta(x_i)_{i:n}$, where $\Delta(x_i)_{i:n}$ is the i-th order value of $\Delta(x_i)$. Working with simulated datasets and various values of p, p/n ratios and contamination amount γ, we observe that the first knee in the plot of Δ_{rank} provides a reliable estimation of the number of outliers in the dataset, while no such estimation can be made from the LCA's ranked density function $g_{\text{rank}}(i) = g(x_i)_{i:n}$, where $g(x_i) = \frac{1}{Z}\boldsymbol{e}_i^\top \boldsymbol{K}\boldsymbol{e}$ and Z is a normalization factor. As we will show in section 3.4, Δ_{rank} better characterizes data structure than g_{rank}.

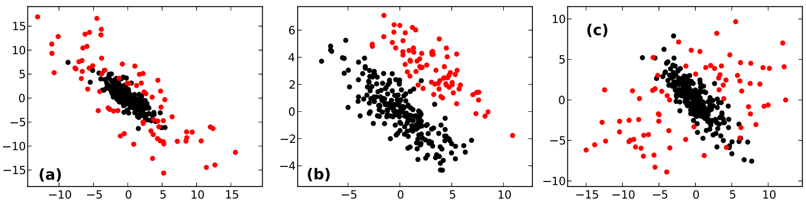


Figure 3.8: **Simulated data.** Three different ways to generate multivariate outliers for Gaussian data ($p = 2$). **(a) Variance outliers** ($a = 3$). **(b) Multi-modal outliers** ($b = 3$). **(c) Multivariate outliers** ($c = 5$). The contamination rate is 25%.

3.4 Applying outlier detection to neuroimaging

3.4.1 Synthetic data

3.4.1.1 Outlier models

Normal population hypothesis. We first stick to the standard hypothesis that neuroimaging data are Gaussian distributed. Since the methods we consider are location invariant, we can make the assumption that the inliers are centered ($\boldsymbol{\mu} = \mathbf{0}$) without loss of generality. Let $\boldsymbol{\Sigma}$ be the covariance matrix for the inliers. Let $\boldsymbol{\mu}_q$ and $\boldsymbol{\Sigma}_q$ be the location and covariance matrix for the outliers. We simulate three outliers types using mixture models (see Figure 3.8):
Variance outliers are obtained by setting $\boldsymbol{\Sigma}_q = a\boldsymbol{\Sigma}$, $a > 1$ and $\boldsymbol{\mu}_q = \mathbf{0}$. This situation models signal normalization issues or aberrant data, where the amount of variance in outlier observations is abnormally large.
Multimodal outliers are obtained by setting $\boldsymbol{\Sigma}_q = \boldsymbol{\Sigma}$ and $\boldsymbol{\mu}_q = b\mathbf{1}$. This simulates the study of an heterogeneous population. Thus, we do not have *outliers* strictly speaking.
Multivariate outliers are obtained by setting $\boldsymbol{\mu}_q = \mathbf{0}$, $\boldsymbol{\Sigma}_q = \boldsymbol{\Sigma} + c\,\sigma_{\max}(\boldsymbol{\Sigma})\,\boldsymbol{a}\,\boldsymbol{a}^{\mathsf{T}}$ where $\boldsymbol{a}$ is a p-dimensional vector drawn from a $\mathcal{N}(\mathbf{0}, \boldsymbol{I})$ distribution. This model simulates outliers as sets of points having potentially abnormally high values in some random direction.
In each case, we relied on the theoretical result $d^2_{\boldsymbol{\mu},\boldsymbol{\Sigma}}(\boldsymbol{X}) \sim \chi^2_p$ to generate the outlier observations in such a way that with a probability of 99%, they do not fall in the inliers support. This was done to ensure that we can distinguish between in- and outliers if we know the real covariance matrix of the former.

Deviation from Gaussian distribution. Real-world data, and in particular, medical imaging data, are often not Gaussian distributed [11, 24, 54]. Yet, in absence of a better model, assuming that the observations are Gaussian distributed is a very popular choice in many fields of applied statistics and within the neuroimaging community, as it amounts to reducing data models to the specification of location and covariance parameters. In order to address deviations from normality, we simulate neuroimaging real data as data coming from a mixture of m Gaussian distributions, the modes of which are randomly drawn from a $\mathcal{N}(\mathbf{0}, \frac{1}{\beta^2}\boldsymbol{I})$ distribution. The β parameter controls the expected distance between the modes. Each component of the model is affected by a given number of variance outliers ($a = 1.15$, see section 3.4.1.1). We choose the β parameter in such a way that the different components overlap. We also consider Student

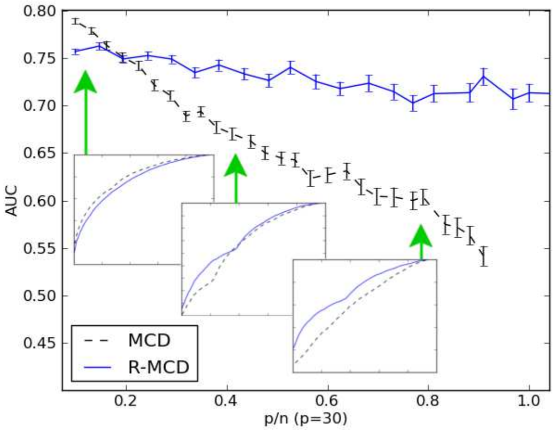


Figure 3.9: **Success metrics – illustrative example.** Simplified AUC curve that illustrates the definition of AUC plots. For a given outlier detection method, 100 outlier detections are performed on samples of the same size (n) and as many ROC curves are built. An average ROC curve is then computed, as well as its Area Under Curve (AUC). This is repeated for various values of n, yielding an AUC value per p/n ratio. Those are presented in a final plot showing the performance of various outlier detection methods. The number of features (p) and the amount of contamination (γ) are fixed during the experiment.

distributed data as a non-Gaussian data distribution (the strength of the deviation is again controlled with a distribution parameter). In both cases (mixture of Gaussian distributions or Student distributed data), we generated outliers so that they do not lie within the 99% support of the inliers. To quantify the deviation from Gaussianity of our simulated dataset, we look at the distribution of the p-values of a thousand normality tests (Shapiro test [49]) performed on random one-dimensional projections of the data, and consider how frequently these p-values are below .05.

3.4.1.2 Success metrics.

For a given outlier model and a fixed p/n ratio, we call an *experiment* 100 outlier detection runs, using a predetermined outlier detection method. We average the results of these runs to build a unique ROC curve [61] per method, and the Area Under the Curve (AUC) [17] is computed. AUC values obtained for various p/n ratios provide a measure of each method's accuracy for outlier detection. Figure 3.9 is an illustrative example showing how we construct the curves that we present in the sequel.

3.4.1.3 Results

All our results are given for a number of features p equal to 100, similar to the real setting ($p = 113$, see section 3.4.2). They hold for greater or lower dimensions (data not shown), although small dimensions are of no interest and computation time becomes a burden for very high dimensions. When reporting results, we denote by *oracle* the best possible decision, knowing the generative model for inliers and outliers.

Variance outliers. As illustrated in Figure 3.10, we observe a significant drop of the MCD accuracy as p/n increases. The MCD ℓ_1- and ℓ_2-regularized versions always give an accuracy above 0.9, RMCD-ℓ_1 performing a bit better. RMCD-RP does not perform well. RMCD-ℓ_1 and RMCD-ℓ_2 performance show that the regularization parameter selection is adapted to our problem, i.e. that we do not introduce too much bias by regularizing the covariance estimate. Indeed,

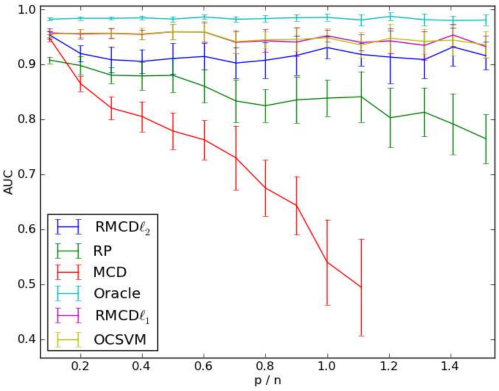


Figure 3.10: **Synthetic data, variance outliers.** AUC for various outlier detection methods in the case of variance outliers ($p = 100$, $\kappa(\mathbf{\Sigma}) = 100$, $a = 1.15$, $\gamma = 40\%$). ℓ_1- and ℓ_2-regularized versions of the MCD outperform by far the standard MCD, benefiting from the isotropic distribution of the outliers. RMCD-RP's accuracy slightly decrease with p/n, which makes it not suitable for our problem. OCSVM$_w$ also gives good performance.

both methods achieve almost perfect outlier detection performance for all values of the covariance matrix condition number.

On both Gaussian and Student distributed data, the accuracy of outlier detection with LCA dominates the accuracy of the other methods, although all density-based methods perform well with an AUC above 0.9 (not shown).

Multimodal outliers. When dealing with multimodal outliers, we observe the expected drop of the MCD accuracy. This demonstrates empirically MCD's theoretical limitations. All the regularized versions of the MCD estimator yielded a perfect outlier detection accuracy, even for $p/n > 1$. Shortening the distance between the modes only impacted the performance of the RMCD-RP-based method, especially when the amount of contamination was high, as shown in Table 3.2: When projecting to a k-dimensional subspace, the expected distance between two observations decreases by a factor $\sqrt{k/n}$ [23] so there is a weaker chance to randomly draw a subspace which preserves the separability between the two modes. Finally, the One-Class SVM is not adapted to this outliers model because it considers every densely populated region as composed of inliers. In the presence of several clusters, the One-Class SVM and any density-based method would only detect outliers as abnormal subjects *with respect to their closest cluster.* This does not correspond to our assumptions of a single main cluster containing inliers, which is anyway debatable as regards neuroimaging data.

p/n	0.1	0.5	0.8	1.0
LCA	**0.99** ±0.0017	**0.99** ±0.0054	**0.99** ±0.0080	**0.98** ±0.0078
Parzen	0.98 ±0.0043	0.98 ±0.0103	0.98 ±0.0091	0.96 ±0.0094
Parzen$_\mathrm{w}$	**0.99** ±0.0022	0.97 ±0.0055	0.97 ±0.0082	0.97 ±0.0095
One-Class SVM	**0.99** ±0.0037	0.91 ±0.0296	0.77 ±0.0795	0.64 ±0.0593
One-Class SVM$_\mathrm{w}$	**0.99** ±0.0022	0.96 ±0.0061	0.97 ±0.0095	0.97 ±0.0104
RMCD	**0.99** ±0.0023	0.95 ±0.0055	0.97 ±0.0083	0.97 ±0.0090

Table 3.1: **Density-based outlier detection methods and variance outliers.** AUC values of the different outlier detection methods confronted with variance outliers (Gaussian distributed data, $p = 100$, $\gamma = 0.4$, $\kappa(\mathbf{\Sigma}) = 1000$, $\alpha = 1.15$).

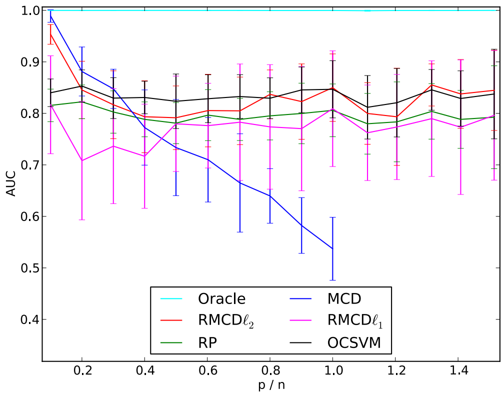


Figure 3.11: **Synthetic data, multimodal outliers.** AUC for various outlier detection methods in the case of multivariate outliers ($p = 100$, $\kappa(\boldsymbol{\Sigma}) = 50$, $c = 20$, $\gamma = 30\%$). While MCD's accuracy drops, the regularized versions of the MCD almost give the same detection accuracy for each p/n ratio. RP and RMCD-ℓ_1 have a lower AUC than RMCD-ℓ_2.

Multivariate outliers. Provided the outliers are strong enough (i.e. $c \geq 10$), the MCD estimator is well adapted to the case of multivariate outliers for $p/n < 0.2$, since its AUC is almost always above 0.9. Yet, the latter drops as the p/n ratio increases. Since RMCD-ℓ_1 and -ℓ_2 have stable performance, they outperform the MCD for large p/n values. In-between, depending on the condition number of the inliers covariance matrix and on the amount of contamination, the relative performance may vary in favor of one method or another. Figure 3.11 gives a general picture of the results obtained with the different methods confronted with multivariate outliers. For $c < 10$, none of the methods can distinguish between in- and outliers and the AUC of each method increases with c, the strongest outliers being detected first. Even though the regularization parameter selection was adapted in the case of variance outliers, RMCD-ℓ_1 and RMCD-ℓ_2 confuse in- and outliers when confronted with multivariate outliers, because of the difficulty to choose an adapted regularization parameter in that case: the *most concentrated set of observations* depends on a prior knowledge about the shape of the global data set. The (R)MCD support is thus difficult to define, and so is the (R)MCD.

Covariance matrix condition number. Because regularized estimators of covariance are biased toward a spherical covariance model, we evaluate the methods performance for inliers covariance matrix having a *condition number* $\kappa(\boldsymbol{\Sigma})$ comprised between 1 and 10,000. With multivariate and variance outliers, we observe an improved accuracy for RMCD-ℓ_1, RMCD-ℓ_2 and density-based methods when the condition number is small. MCD is not affected by this parameter. Unlike RMCD-ℓ_1, RMCD-ℓ_2 and RMCD-RP, OCSVM and Parzen methods

p/n	0.1	0.4	0.6	0.8	1.
MCD	**1.** ±0.008	**1.** ±0.065	0.8 ±0.052	0.65 ±0.058	0.55 ±0.067
RMCD-RP	**1.** ±0.008	0.98 ±0.035	0.95 ±0.031	0.90 ±0.056	0.8 ±0.057
One-Class SVM	0.76 ±0.009	0.76 ±0.020	0.76 ±0.016	0.75 ±0.025	0.76 ±0.028
RMCD-ℓ_1 / RMCD-ℓ_2	**1.** ±0.	**1.** ±0.	**1.** ±0.	**1.** ±0.	**1.** ±0.

Table 3.2: **Synthetic data, multimodal outliers.** AUC for MCD and RMCD-RP confronted with multimodal outliers ($p = 100$, $b = 3$, $\kappa(\boldsymbol{\Sigma}) = 10$, $\gamma = 30\%$). MCD breaks down for $p/n > 0.4$, which is the theoretical breakdown point. RMCD-RP's AUC stays above 0.8, which indicates good performance although it decreases when p/n increases. Other regularized methods achieve perfect outlier detection (AUC$=1$) and the One-Class SVM's AUC remains constant at a low level.

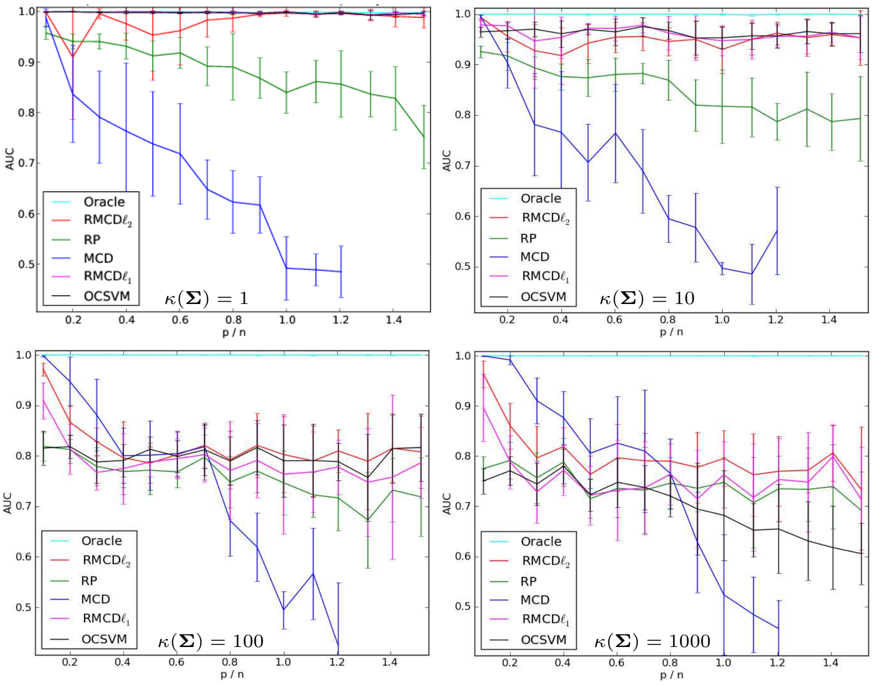


Figure 3.12: **Synthetic data, multivariate outliers.** AUC for various outlier detection methods in the case of multivariate outliers ($p = 100$, $\kappa(\mathbf{\Sigma}) = \{1, 10, 100, 1000\}$, $c = 10$, $\gamma = 20\%$). A small condition number give advantage to the RMCD-ℓ_1 and -ℓ_2 methods, as well as the OCSVM. For $\kappa(\mathbf{\Sigma}) > 100$, all RMCD approaches perform similarly.

break down when $\kappa(\mathbf{\Sigma}) \geq 1000$ (see Table 3.1). In the latter case, one has to whiten the data previously to using One-Class SVM and Parzen. The main advantage of LCA is that such a transformation is part of the algorithm.
The inliers covariance matrix condition number has the most influence when multivariate outliers are considered. In that case, all methods have poor performance. This phenomenon is depicted in Figure 3.12.

Contamination rate. Outlier detection accuracy remains similar for each method and amount of contamination, except for the RMCD-RP method that is very sensitive to the number of outliers when these are of the multimodal type (see Table 3.3).

Sparsity coefficient. As the ℓ_1 regularization is known to benefit from the *sparsity* of the original inliers precision matrix, we also look at this parameter's influence. Sparsity of the precision matrix does not have a strong influence on

p/n	0.1	0.4	0.6	0.8	1.
$\gamma = 20\%$	**1.** ±0	**1.** ±0.005	**1.** ±0.008	**0.99** ±0.019	**0.98** ±0.027
$\gamma = 30\%$	**1.** ±0	0.98 ±0.023	0.95 ±0.051	0.90 ±0.104	0.8 ±0.187
$\gamma = 40\%$	0.65 ±0.198	0.6 ±0.164	0.59 ±0.075	0.55 ±0.063	0.58 ±0.084

Table 3.3: **Influence of the amount of contamination on outlier detection.**We observe a drop of the RMCD-RP-based outlier detection method AUC with the amount of contamination γ. Multimodal outliers ($p = 100$, $b = 3$, $\kappa(\mathbf{\Sigma}) = 10$).

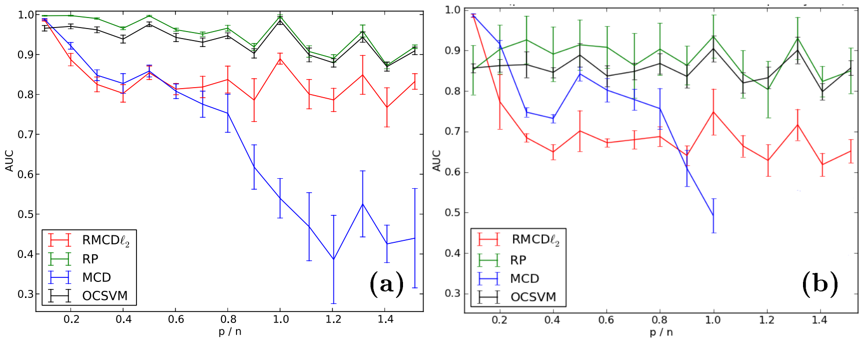


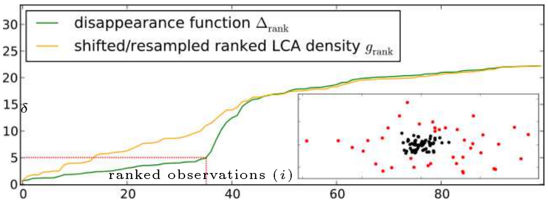
Figure 3.13: **Outlier detection in the non-Gaussian inliers case.** AUC curves of the methods on datasets generated by a mixture of Gaussian distributions. Observations were equally distributed between the components. $\gamma = 0.4$. **(a) Mild deviation from normality.** $m = 4$, $\beta = 1.1$. RMCD-ℓ_2 AUC is stable for large p/n ratios but is roughly 0.1 below RMCD-RP and One-Class SVM AUC. RMCD-RP has the best accuracy. **(b) Strong deviation from normality.** $m = 4$, $\beta = 0.7$. The different modes are observable in two- or one-dimensional projections of the data. RMCD-ℓ_2's performance is poor compared to OCSVM and RMCD-RP.

the methods AUC: only the RMCD-ℓ_1 has a slightly improved AUC when the inverse covariance is very sparse. Yet, RMCD-ℓ_1 is not more accurate than RMCD-ℓ_2, so we did not report the results for RMCD-ℓ_1 in the sequel.

Non-Gaussian models. Under deviations from normality, RMCD-RP and One-Class SVM outperform RMCD-ℓ_2, as shown in Figure 3.13. RMCD-ℓ_1 results are not reported since RMCD-ℓ_2 always yields better performance. All methods but MCD have similar and stable performance for $p/n > 0.4$. Interestingly, all methods have an AUC close to 1 for $p/n < 0.1$, which justifies the use of MCD on the complete database to build a reference labeling in our real-data experiments. A stronger deviation from normality yields poorer performance as well as a larger variability of the outlier detection accuracy. RMCD-RP remains the best method for detecting outliers with an AUC above 0.85. MCD and RMCD-ℓ_2 still achieve almost perfect outlier detection for $p/n < 0.1$ with an AUC close to 1. RMCD-RP performance is explained by the fact that in high-dimension, the distribution of randomly projected observations is closer to normal than the original data [10]. Therefore, applying the MCD on projected data yields a more accurate detection since the outlier detection threshold can be set exactly.

Experiments on synthetic data give an overview of the relative performance of the various outlier detection methods that we consider in this work. The accuracy of density-based methods is generally higher than that of covariance-based methods, but RMCD-RP behaves well even when the model assumptions are not met or when the data dimensionality is high. RMCD-RP comes with a statistical control on outlier discarding, which makes it a good compromise overall. Its only weakness is that it is sensitive to the amount of contamination, unlike the other methods.

Interactive outlier detection procedure We verified with extensive simulations that the first knee of the disappearance function directly provides an estimate of the number of outliers. Figure 3.14 illustrates this statement. This



Gaussian-distributed data, $p = 2$, $n = 100$, $\alpha = 4$, $\gamma = 0.4$

Figure 3.14: **Functions summarizing the data structure from their density.** Difference between outliers (red dots) density and inliers (black dots) density only appears in the disappearance function. Choosing $\delta \simeq 5$ yields an outlier detection corresponding to estimating $\hat{\gamma} \simeq 35\%$ for a real value of $\gamma = 40\%$.

result holds for various p, p/n and α values, even though the decision showed to be a bit conservative. This behavior is yet required to guarantee a low false detections rate on heavy tailed distributions such as the Student distribution. Figure 3.19 shows that our procedure does not encourage discarding observations when applied to pure Student distributed data.

3.4.2 Real data

3.4.2.1 Data and validation procedure

Functional data. We work with four different types of contrasts images from the Imagen database (see section 1.4) that show brain regions implied in simple cognitive tasks:

- an auditory task as opposed to a visual task;
- a left motor task as opposed to a right motor task;
- a computation task as opposed to a sentences reading task;
- an angry faces viewing task.

For outlier detection, we extracted 113 features by computing on each contrast image the average activation intensity value from 113 regions of interest. These regions were given by the Harvard-Oxford cortical and sub-cortical structural atlases[2]. We removed the regions covering more than 1% of the whole brain volume, because the mean signal within such large regions does not summarize the functional signal well. We removed the effect of gender, handedness and acquisition center by using a robust regression based on M-estimators [19], using the scikit.statsmodels Python package [47] implementation. We then performed an initial outlier detection at a p-value $P < 0.1$ family-wise corrected, including all subjects ($n > 1500$). With such a small $\frac{p}{n}$ value, a statistically controlled outlier detection can be achieved using the MCD estimate. The outlier list obtained from this first outlier detection was then held as a reference labeling for further outlier detection experiments performed on reduced sample sizes, using MCD and all the Regularized MCD estimators. Note that for very small samples, we could not use the MCD-based outlier detection method. The outlier lists were compared to the reference labeling and ROC curves were constructed. For each sample size, we repeated the detection 10 times with 10 different, randomly selected samples.

[2] http://www.cma.mgh.harvard.edu/fsl_atlas.html

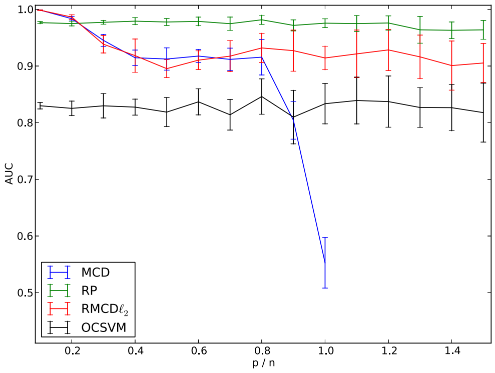


Figure 3.15: **Outlier detection on real fMRI data.** Results on functional MRI data after removal of the effect of gender, handedness and acquisition center. AUC curve illustrating the ability of each method to find back a reference labeling from randomly selected sub-samples corresponding to various p/n ratios. Reference labeling was constructed with the MCD from $n = 1995$ observations ($p = 113$).

Voxel-Based Morphometry (VBM). We use the gray matter probability maps available in the Imagen database. We use 120 regions of interest defined as 4mm-radius balls centered around locations of highly variable gray matter probability value trough subjects: we used the watershed algorithm [34] to segment the voxel-wise variability map into homogeneous regions, and the signal peak locations of the 120 regions of highest mean signal were retained as regions of interest. We limited the number of regions to 120 in order to keep an accurate statistical control of outlier detection with the full dataset. However, the choice and the size of the regions as well as the different type of data used in this second experiment should demonstrate how well regularized covariance-based outlier detection methods generalize to different contexts encountered in medical imaging. For the sake of completeness, we also tried outlier detection using the Harvard-Oxford atlas regions of interest on the gray matter probability maps.

3.4.2.2 Results

Figure 3.15 shows the outlier detection performance obtained on a dataset constructed from an fMRI contrast reflecting the brain activity related to angry faces viewing. The RMCD-RP method's curve dominates the others methods' curves for $p/n > 0.2$. RMCD-ℓ_2's accuracy is always above 0.9 while MCD-based outlier detection breaks down when p/n becomes large. Results obtained without removing the effect of gender, handedness and acquisition center are similar to our first results, although the difference between RMCD-RP and RMCD-ℓ_2 is a bit larger (not shown). Results obtained with others functional contrasts are similar to those of Figure 3.15. This suggests that the general structure of observations distribution does not depend on the contrast.

Regarding Density-based outlier detection, Figure 3.16 shows the accuracy of the different methods on a real neuroimaging dataset, using a contrast related to the perception of angry versus neutral faces. All methods perform well with an AUC above 0.8. Yet, LCA achieves the highest accuracy, which remains above 0.95 for all p/n ratios. Whitening the data prior to outlier detection with One-Class SVM or Parzen density estimation is relevant since it increases the accuracy of the latter methods by roughly 0.1. Similar results were obtained in five other functional contrasts.

Figure 3.17 shows activity maps (thresholded at $P < 0.01$ family-wise corrected) of out- and inliers subjects in a plot of the first two components of a Principal Components Analysis performed on the full, outlier-free data set.

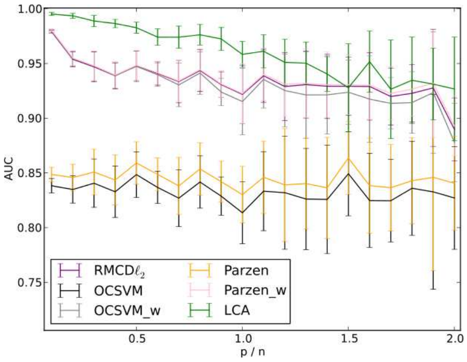


Figure 3.16: **Outlier detection accuracy of non-parametric density estimation algorithms, represented by their AUC (real data).** LCA outperforms both Parzen density estimation and One-Class SVM, even applied on whitened data. RMCD parametric method has the same accuracy than the latter. LCA seems to be sensitive to the p/n ratio as its performance decreases with this ratio.

Outlier observations were projected to the same low-dimensional space. Outliers found by RMCD-ℓ_2-based method stand far from the central cluster, which illustrates the accuracy of the method. State-of-the-art MCD finds only three outliers. It is clear from the figure that some observations should be tagged as abnormal because the global activation pattern deviates from the standard ones (e.g. too much activity for subjects *(a)*, *(b)* and *(c)*). Yet manual screening may not be sufficient to detect some subtleties in the pattern differences. For instance, the dissimilarity between subjects *(d)* and *(e)* (both were yet in the RMCD-RP support) is not apparent in the low-dimensional projection. Note that some outliers seem to fall amongst inliers due to an artifact of projection since the original data lie in a 100-dimensional space. Indeed, only 70% of the variance is fit by the first two components.

Voxel-Based Morphometry Figure 3.18 gives the outlier detection accuracy of the RMCD-ℓ_2, RMCD-RP, MCD and OCSVM methods on gray matter probability maps. Despite the use of a different imaging modality and ROI selection procedure, the relative performance of the methods is very similar to the performance obtained in our experiment with functional data. The number of outliers is much smaller in the reference labeling ($\simeq$ 3%). The MCD drops faster and breaks down for $p/n > 0.5$. The variability of all methods but RMCD-RP is much larger, which may be related to the deviation from the Gaussian distribution hypothesis that can be observed in the PCA plot given in Figure 3.18.

Using the Harvard-Oxford atlas's regions of interest mean signal as a descriptive feature of the gray matter images, we obtained similar results, confirming the RMCD-RP's more accurate performance for outlier detection on real datasets (not shown).

Interactive outlier detection procedure Figure 3.20 gives the spectrum of real neuroimaging datasets as obtained from LCA-learned density transformations. Knees can be easily identified in this curve, indicating that two or more relevant groups of observations are present. This observation rank property could not be inferred from the standard decision function. It is noticeable that many observations (about half of the dataset) seem to be suggested as outliers, while looking at a standard bidimensional PCA plot (not shown) would have suggested a much lower number.

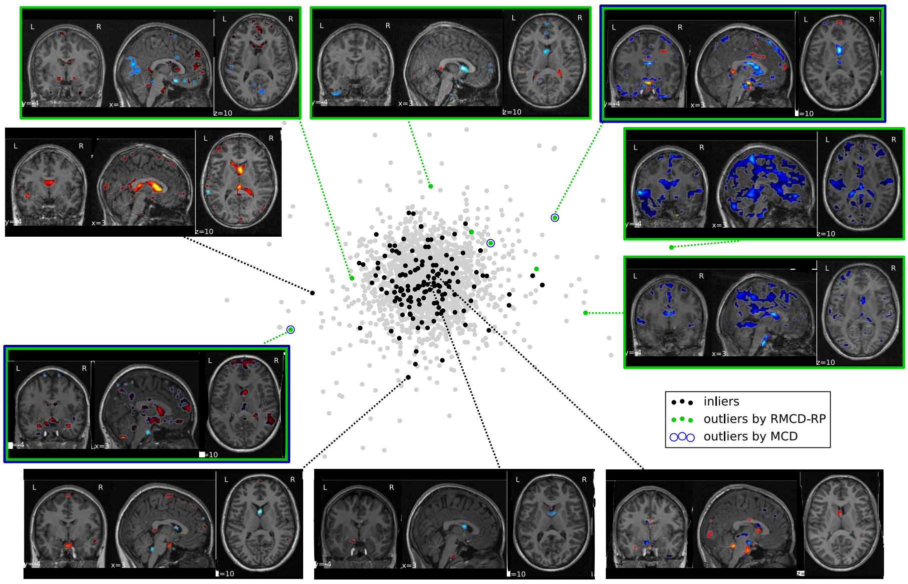


Figure 3.17: **Neuroimaging data projection on the space spanned by the two principal components of the full, cleaned dataset.** Observations tagged as outliers by the RMCD-RP method are indeed outliers at least along the two first PCA components. MCD-based outlier detection method only finds three outliers and misses strong ones. This figure illustrates the difficulty of manual outlier detection: the deviation from normality can result in unusual patterns that are not easily compared to the others.

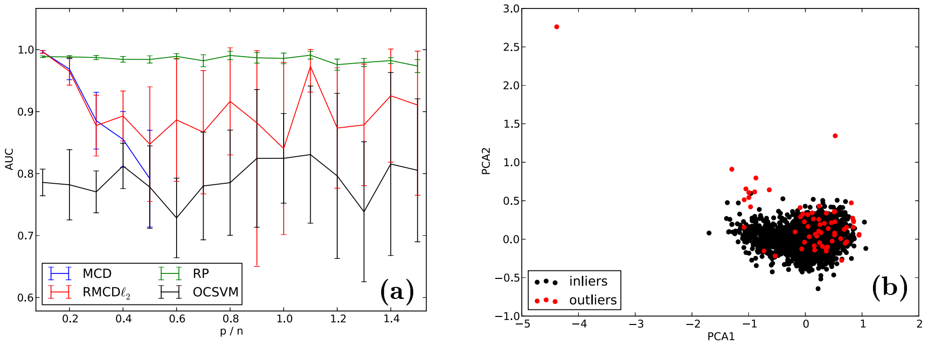


Figure 3.18: **Outlier detection on voxel-based morphometry data.** Outlier detection accuracy of the RMCD-ℓ_2, RMCD-RP, MCD and OCSVM methods on gray matter probability maps and representation of the corresponding dataset. **(a)** The relative performance is very similar to the performance obtained with functional data, although MCD drops faster. RMCD-RP still outperforms with an AUC above 0.95. **(b)** Projection of the dataset according to the first two components of a PCA decomposition. Outliers (in red) and inliers (in black) of the reference labeling are represented.

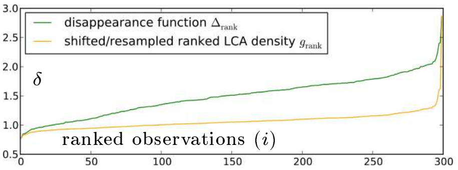


Figure 3.19: **"Spectrum" of Student-distributed data.** Data statistical structure investigation in an uncontaminated Student-distributed data. No hard decision seems to be suggested. $p = 100$, $n = 300$, $\gamma = 0$.

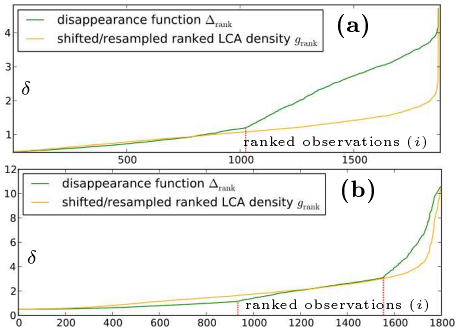


Figure 3.20: **Dataset structure spectrum obtained by density analysis on real neuroimaging datasets.** **(a)** Viewing Angry faces - viewing neutral faces. A slope breakdown is observed at $i \simeq 1000$, suggesting that half the observations should be removed to obtain an heterogeneous set. **(b)** Rewarding task. The procedure suggests that three observations scales are present in the data. The first one may be composed of outliers. In both cases, g_{rank} (see 3.3.3.3) does not reveal any structure.

3.5 Discussion

3.5.1 RMCD-RP is robust to non-Gaussian distributions

As most outlier-detection procedures, the RMCD-RP's accuracy slightly drops as p/n increases. Yet, except for extreme cases such as *multivariate outliers and large condition number* or *multimodal outliers and large amount of contamination*, the method's AUC is higher than 0.8, which makes it attractive in practice. Importantly, RMCD-RP was shown to have the best accuracy for non-Gaussian distributed data sets (see 3.4.1.3) under mild or strong deviation from normality. While the performance of RMCD-ℓ_2 breaks with stronger deviations from normality, RMCD-RP performances dominates with a gain in AUC of 0.2 or more in non-Gaussian settings. In medical imaging settings, RMCD-RP can be considered as useful, due to its robustness to deviations from normality. A procedure for the explicit control of false detections with RMCD-RP is presented in section 3.2.2.5. We showed in section 3.4.1.3 that using regularized versions of the MCD was relevant to detect outliers in practice, as the neuroimaging datasets that we used appeared to be non-Gaussian distributed. The RMCD-RP estimator is particularly adapted to that context (see Figure 3.13 and Figure 3.18) since the actual outlier detection is made on projected subspaces that appear *more Gaussian* than in the native space. This is a straightforward consequence of the Central Limit Theorem. Even on small datasets ($p/n > 0.2$), RMCD-RP outlier detection method can detect outliers that would not be detected by hand.

3.5.2 LCA is a powerful density-based method

Outlier detection with Local Component Analysis (LCA) achieves higher accuracy than state-of-the-art methods, including covariance-based ones. Only RMCD-RP offers comparable performance. This was shown for various p/n settings on both Gaussian and Student distributed data contaminated with up to 40% outliers. Real data experiments showed that LCA accuracy is generally above 0.9, although it seems to slightly decrease in high-dimension. Our

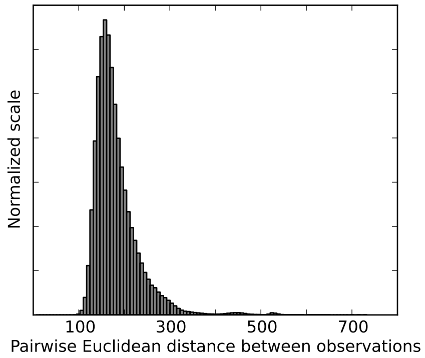


Figure 3.21: **Pairwise Euclidean (L_2) distance between observations.** The shape of the distribution does not change with L_k ($k \leq 1$) distances, indicating that the observations of [1] are not worth to consider in neuroimaging. Therefore, non-parametric outlier detection algorithms can trustworthy rely on observations pairwise Euclidean distances (we confirmed this statement by experiments for which we do not show the results).

choice of the LCA regularization parameter seemed to be optimal in that regard and our experiments demonstrated that LCA should be preferred to other non-parametric methods. Because it uses an internal cross-validation scheme, LCA adapts to the data local structure and comes with a natural kernel-based representation of the data, on which we apply a soft thresholding operation. Formally, we apply a trace norm penalization to capture the information carried in the kernel matrix which reveals important features on the structure of the data [57]. As this penalization is the convex relaxation of principal components analysis-based truncation of the kernel matrix, it results in a stable criterion. We used it to characterize the difference between outliers and inliers in a robust procedure. This is meant to provide practitioners a faithful representation of possible inhomogeneities in the population under study. We verified on several simulated and real functional neuroimaging datasets that this heuristic to chose the regularization of LCA does not yield spurious outlier detections. An attractive generalization of the LCA approach for high-dimensional settings is a mixed model, in which some dimensions are simply modeled as a Gaussian, while others are modeled through equation (3.6) [45].

3.5.3 L_k distances

The classical euclidean distance (i.e. the L_2 distance) may not be optimal to describe high-dimensional datasets as the distribution of the pairwise distances between the observations becomes tighter as the number of features p gets large [5]. Aggarwal et al. [1] showed that L_k norms with $k < 1$ improve the contrast between the pairwise distances, thus improving the accuracy of the algorithms that rely on it. Figure 3.21 shows the histogram of the pairwise L_2 distances. On fMRI and structural data, we do not observe any change of the distribution when considering L_k distances with $k < 1$. We do neither observe any change in the accuracy of outlier detection with Parzen density estimator (not shown).

3.5.4 Statistical control of outlier detection

Non-parametric methods, in particular LCA, have good outlier detection accuracy and do not require a lot of computation time. They furthermore do not rely on distributional assumptions, but at the expense of explicit statistical control. Covariance-based outlier detection methods can be used instead so as to obtain a statistical control on outlier detection. This feature is of broad interest when considering studies replications since no *ad hoc* choice on the proportion of subjects to discard is needed. Depending on the application and the number of available data, the practitioner may use RMCD-RP or LCA to perform outlier detection. Would LCA be chosen, our interactive outlier detection procedure helps setting up an interpretable outlier detection threshold that depends on the data statistical structure.

References

[1] Aggarwal, C.C., Hinneburg, A., Keim, D.A.: On the surprising behavior of distance metrics in high dimensional space. Springer (2001)

[2] Angiulli, F., Pizzuti, C.: Fast outlier detection in high dimensional spaces. In: Principles of Data Mining and Knowledge Discovery, pp. 15–27. Springer (2002)

[3] Arning, A., Agrawal, R., Raghavan, P.: A linear method for deviation detection in large databases. In: KDD. pp. 164–169 (1996)

[4] Bay, S.D., Schwabacher, M.: Mining distance-based outliers in near linear time with randomization and a simple pruning rule. In: Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 29–38. ACM (2003)

[5] Beyer, K., Goldstein, J., Ramakrishnan, R., Shaft, U.: When is "nearest neighbor" meaningful? In: Database Theory–ICDT'99, pp. 217–235. Springer (1999)

[6] Breunig, M.M., Kriegel, H.P., Ng, R.T., Sander, J.: LOF: identifying density-based local outliers. In: ACM Sigmod Record. pp. 93–104. ACM (2000)

[7] Chen, Y., Wiesel, A., Eldar, Y., Hero, A.: Shrinkage algorithms for MMSE covariance estimation. Signal Processing, IEEE Transactions on 58(10), 5016–5029 (oct 2010)

[8] Daszykowski, M., Kaczmarek, K., Heyden, Y.V., Walczak, B.: Robust statistics in data analysis – A review: Basic concepts. Chemometrics and Intelligent Laboratory Systems 85(2), 203–219 (2007)

[9] Dave, R., Krishnapuram, R.: Robust clustering methods: A unified view. Fuzzy Systems, IEEE Transactions on 5(2), 270–293 (may 1997)

[10] Diaconis, P., Freedman, D.: Asymptotics of graphical projections. The Annals of Statistics 12(3), 793–815 (1984)

[11] Falangola, M., Jensen, J., Babb, J., Hu, C., Castellanos, F., Di Martino, A., Ferris, S., Helpern, J.: Age-related non-Gaussian diffusion patterns in the prefrontal brain. Journal of Magnetic Resonance Imaging 28(6), 1345–1350 (2008)

[12] Fan, H., Zaïane, O.R., Foss, A., Wu, J.: A nonparametric outlier detection for effectively discovering top-n outliers from engineering data. In: Advances in Knowledge Discovery and Data Mining, pp. 557–566. Springer (2006)

[13] Fawcett, T., Provost, F.: Activity monitoring: Noticing interesting changes in behavior. In: Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 53–62. ACM (1999)

[14] Friedman, J., Hastie, T., Tibshirani, R.: Sparse inverse covariance estimation with the lasso. ArXiv e-prints (Aug 2007)

[15] Garcia-Escudero, L., Gordaliza, A.: Robustness properties of K-Means and trimmed K-Means. Journal of the American Statistical Association 94(447), 956–969 (september 1999)

[16] Hamilton, W.C.: The revolution in crystallography. Science 169, 133–141 (jul 1970)

[17] Hanley, J.A., McNeil, B.J.: The meaning and use of the area under a receiver operating (ROC) curve characteristic. Radiology 143(1), 29–36 (1982)

[18] Hardin, J., Rocke, D.M.: The distribution of robust distances. Journal of Computational and Graphical Statistics 14(4), 928–946 (2005)

[19] Huber, P.J.: Robust Statistics, chap. 7, p. 149. John Wiley & Sons, Inc. (2005)

[20] Hubert, M., Engelen, S.: Robust PCA and classification in biosciences. Bioinformatics 20(11), 1728–1736 (Jul 2004)

[21] Jin, W., Tung, A.K., Han, J.: Mining top-n local outliers in large databases. In: Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 293–298. ACM (2001)

[22] Jin, W., Tung, A.K., Han, J., Wang, W.: Ranking outliers using symmetric neighborhood relationship. In: Advances in Knowledge Discovery and Data Mining, pp. 577–593. Springer (2006)

[23] Johnson, W., Lindenstrauss, J., Schechtman, G.: Extensions of Lipschitz maps into Banach spaces. Israel Journal of Mathematics 54, 129–138 (1986)

[24] Joshi, S., Bowman, I., Toga, A., Van Horn, J.: Brain pattern analysis of cortical valued distributions. Proc IEEE Int Symp Biomed Imaging pp. 1117–1120 (2011)

[25] Kherif, F., Flandin, G., Ciuciu, P., Benali, H., Simon, O., Poline, J.B.: Model based spatial and temporal similarity measures between series of functional magnetic resonance images. Med Image Comput Comput Assist Interv pp. 509–516 (2002)

[26] Knorr, E.M., Ng, R.T.: A unified notion of outliers: Properties and computation. In: KDD. pp. 219–222 (1997)

[27] Kollios, G., Gunopulos, D., Koudas, N., Berchtold, S.: Efficient biased sampling for approximate clustering and outlier detection in large data sets. Knowledge and data engineering, ieee transactions on 15(5), 1170–1187 (2003)

[28] Kriegel, H.P., Kröger, P., Schubert, E., Zimek, A.: Outlier detection in axis-parallel subspaces of high dimensional data. In: Advances in Knowledge Discovery and Data Mining, pp. 831–838. Springer (2009)

[29] Kriegel, H.P., Kröger, P., Schubert, E., Zimek, A.: Interpreting and unifying outlier scores. In: SDM. pp. 13–24 (2011)

[30] Kriegel, H.P., Zimek, A., et al.: Angle-based outlier detection in high-dimensional data. In: Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 444–452. ACM (2008)

[31] Ledoit, O., Wolf, M.: A well-conditioned estimator for large-dimensional covariance matrices. Journal of Multivariate Analysis 88(2), 365–411 (2004)

[32] Mériaux, S., Roche, A., Thirion, B., Dehaene-Lambertz, G.: Robust statistics for nonparametric group analysis in fMRI. In: Biomedical Imaging: Nano to Macro, 2006. 3rd IEEE International Symposium on. pp. 936–939 (april 2006)

[33] Mourão-Miranda, J., Hardoon, D.R., Hahn, T., Marquand, A.F., Williams, S.C., Shawe-Taylor, J., Brammer, M.: Patient classification as an outlier detection problem: An application of the one-class support vector machine. NeuroImage 58(3), 793–804 (2011)

[34] Najman, L., Schmitt, M.: Watershed of a continuous function. Signal Processing 38(1), 99–112 (1994)

[35] Papadimitriou, S., Kitagawa, H., Gibbons, P.B., Faloutsos, C.: Loci: Fast outlier detection using the local correlation integral. In: Data Engineering, 2003. Proceedings. 19th International Conference on. pp. 315–326. IEEE (2003)

[36] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, É.: Scikit-learn: Machine learning in Python. Journal of Machine Learning Research (Oct 2011)

[37] Pei, Y., Zaiane, O.R., Gao, Y.: An efficient reference-based approach to outlier detection in large datasets. In: Data Mining, 2006. ICDM'06. Sixth International Conference on. pp. 478–487. IEEE (2006)

[38] Peña, D., Prieto, F.J.: Multivariate outlier detection and robust covariance matrix estimation. Technometrics 43, 286–310 (2001)

[39] Penny, W.D., Kilner, J., Blankenburg, F.: Robust bayesian general linear models. Neuroimage 36, 661–671 (2007)

[40] Pinel, P., Dehaene, S., Rivière, D., LeBihan, D.: Modulation of parietal activation by semantic distance in a number comparison task. NeuroImage 14(5), 1013–1026 (2001)

[41] Pison, G., Van Aelst, S., Willems, G.: Small sample corrections for LTS and MCD. Metrika 55(1-2), 111–123 (2002)

[42] Ramaswamy, S., Rastogi, R., Shim, K.: Efficient algorithms for mining outliers from large data sets. In: ACM SIGMOD Record. pp. 427–438. ACM (2000)

[43] Rousseeuw, P.J.: Least median of squares regression. J. Am Stat Ass 79, 871–880 (1984)

[44] Rousseeuw, P.J., Leroy, A.M.: Robust Regression and Outlier Detection, chap. 1, pp. 4–5. John Wiley & Sons, Inc. (2005)

[45] Roux, N.L., Bach, F.: Local component analysis. CoRR abs/1109.0093 (2011)

[46] Schölkopf, B., Platt, J.C., Shawe-Taylor, J.C., Smola, A.J., Williamson, R.C.: Estimating the support of a high-dimensional distribution. Neural Comput. 13, 1443–1471 (July 2001)

[47] Seabold, S., Perktold, J.: Statsmodels: Econometric and statistical modeling with Python. In: van der Walt, S., Millman, J. (eds.) Proceedings of the 9th Python in Science Conference. pp. 57–61 (2010)

[48] Segata, N., Blanzieri, E.: Fast and scalable local kernel machines. J. Mach Learn Res 11, 1883–1926 (2009)

[49] Shapiro, S.S., Wilk, M.B.: An analysis of variance test for Normality (complete samples). Biometrika 52(3/4), 591–611 (Dec 1965)

[50] Sun, P., Chawla, S.: On local spatial outliers. In: Data Mining, 2004. ICDM'04. Fourth IEEE International Conference on. pp. 209–216. IEEE (2004)

[51] Tang, J., Chen, Z., Fu, A.W.C., Cheung, D.W.: Enhancing effectiveness of outlier detections for low density patterns. In: Advances in Knowledge Discovery and Data Mining, pp. 535–548. Springer (2002)

[52] Tao, T., Zhai, C.: Regularized estimation of mixture models for robust pseudo-relevance feedback. In: Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval. pp. 162–169. ACM (2006)

[53] Tibshirani, R.: Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society, Series B 58, 267–288 (1994)

[54] Upadhyaya, A., Rieu, J., Glazier, J., Sawada, Y.: Anomalous diffusion and non-Gaussian velocity distribution of hydra cells in cellular aggregates. Physica A: Statistical Mechanics and its Applications 293(3–4), 549–558 (2001)

[55] Varoquaux, G., Sadaghiani, S., Pinel, P., Kleinschmidt, A., Poline, J., Thirion, B.: A group model for stable multi-subject ICA on fMRI datasets. NeuroImage 51(1), 288–299 (2010)

[56] de Vries, T., Chawla, S., Houle, M.E.: Finding local anomalies in very high dimensional space. In: Data Mining (ICDM), 2010 IEEE 10th International Conference on. pp. 128–137. IEEE (2010)

[57] Wang, J., Saligrama, V., Castañón, D.A.: Structural similarity and distance in learning. ArXiv e-prints (oct 2011)

[58] Woolrich, M.: Robust group analysis using outlier inference. Neuroimage 41, 286–301 (2008)

[59] Woolrich, M.W., Behrens, T.E.J., Beckmann, C.F., Smith, S.M.: Mixture models with adaptive spatial regularization for segmentation with an application to fMRI data. Medical Imaging, IEEE Transactions on 24(1), 1–11 (2005)

[60] Zhang, K., Hutter, M., Jin, H.: A new local distance-based outlier detection approach for scattered real-world data. In: Advances in Knowledge Discovery and Data Mining, pp. 813–822. Springer (2009)

[61] Zweig, M., Campbell, G.: Receiver-operating characteristic (ROC) plots: A fundamental evaluation tool in clinical medicine. Clin Chem 39(4), 561–577 (1993)

Chapter 4
Enhancing the Reproducibility of Group Analysis with Randomized Brain Parcellations

Contents

Contributions

[1] Da Mota, B., Fritsch, V., Varoquaux, G., Frouin, V., Poline, J.B., Banaschewski, T., Barker, G.J., Bokde, A.L., Bromberg, U., Conrod, P., Gallinat, J., Garavan, H., Martinot, J.L., Nees, F., Paus, T., Pausova, Z., Rietschel, M., Smolka, M.N., Ströhle, A., the IMAGEN consortium, Thirion, B.: Randomized parcellation based inference. Neuroimage submitted manuscript (2013)

[2] Da Mota, B., Fritsch, V., Varoquaux, G., Frouin, V., Poline, J.B., Thirion, B.: Enhancing the Reproducibility of Group Analysis with Randomized Brain Parcellations. In: MICCAI - 16th International Conference on Medical Image Computing and Computer Assisted Intervention - 2013 (Oct 2013)

Neuroimaging group analysis are used to relate inter-subject signal differences observed in brain imaging with behavioral or genetic variables and to assess risks factors of brain diseases. The lack of stability and of sensitivity of current voxel-based analysis schemes may however lead to non-reproducible results. We introduce a new approach to overcome the limitations of standard methods, in which active voxels are detected according to a consensus on several random parcellations of the brain images, while a permutation test controls the false positive risk. Both on synthetic and real data, this approach shows higher sensitivity, better accuracy and higher reproducibility than state-of-the-art methods. In a neuroimaging-genetic application, we find that it succeeds in detecting a significant association between a genetic variant next to the *COMT* gene and the BOLD signal in the left thalamus for a functional Magnetic Resonance Imaging contrast associated with incorrect responses of the subjects from a *Stop Signal Task* protocol.

4.1 Brain parcellations

4.1.1 Spatial models for group analysis in neuroimaging

Spatial models try to overcome the lack of correspondence between individual images at the voxel level. The most straightforward and widely used technique consists in smoothing the data to increase the overlap between subject-specific activated regions [34]. In the literature, several approaches propose more elaborate techniques to model the noise in neuroimaging, like Markov Random Fields [19], wavelets decomposition [31], spatial decomposition or topographic methods [8, 6] and anatomically informed models [14]. These techniques are not widely used probably because they are computationally costly and not always well-suited for analysis of a group of subjects. A popular approach consists in working with subject-specific Regions of Interest (ROIs), that can be defined in a way that accommodates inter-subject variability [18]. The main limitation of such an approach [2] is that there is no widely accepted standard for partitioning the brain, especially for the neocortex. Data-driven parcellation was proposed by Thirion et al. [27] to overcome this limitation: they improve the sensitivity of random effect analysis by considering adaptative parcels that vary across subjects to better fit the fMRI signal.

4.1.1.1 The randomized parcellation approach

The parcellation model [27] has several advantages: *(i)* it is a simple and easily interpretable method, *(ii)* by reducing the number of descriptors, it alleviates the multiple comparisons problem, and *(iii)* the choice of the parcellation algorithm can lead to parcels adapted to the local smoothness. But parcellations, when considered as spatial functions, highly depend on the data used to construct them and the choice of the number of parcels. In general, a parcellation defined in a given context might not be a good descriptor in a slightly different context, or may generalize poorly to new subjects. This implies a lack of reproducibility of the results across subgroups, as illustrated latter in Figure 4.7. The weakness of this approach is the large impact of a parcellation scheme that cannot be optimized easily for the sake of statistical inference; it may thus fail to detect effects in poorly segmented regions. We propose to solve this issue by using several randomized parcellations [30, 3] generated using resampling methods (bootstrap) and average the corresponding statistical decisions. Replacing an estimator such as parcel-level inference by a mean of bootstrap estimates is known to *stabilize* it; a fortunate consequence is that the *reproducibility* of the results (across subgroups of subjects) is improved. Formally, this can be understood as handling the parcellation as a hidden variable that needs to be integrated out in order to obtain the posterior distribution of statistics values. The final decision is taken with regard to the stability of the detection of a voxel [16, 1] across parcellations, compared to the null hypothesis distribution obtained by a permutation test.

4.1.1.2 A multivariate problem : the detection of outliers

The benefits of the randomized parcellation approach can also be observed in multivariate analysis procedures, such as predictive modeling [30] or outliers detection. In this work, we focus on the latter: neuroimaging datasets often contain atypical observations; such *outliers* can result from acquisition-related issues [12], bad image processing [35], or they can merely be extreme examples of the high variability observed in the population. Because of the high dimensionality of neuroimaging data, screening the data is very time consuming, and becomes prohibitive with large cohort studies. Covariance-based outlier detection methods have been proposed to perform statistically-controlled inclusion of subjects in neuroimaging studies [9] (see chapter 3) and yield a good detection accuracy. These methods rely on prior reduction of the data dimension which is obtained by taking signal averages within predefined brain parcels. As a consequence, the results depend on a fixed brain parcellation and are unstable. Randomization might thus improve the procedure.

4.1.2 Construction of variable parcellations

An important prerequisite for our approach is to generate several parcellations that are different enough from each other to guarantee that the analysis conducted with each of these parcellations represents correctly the set of regions that display some activation for the effect considered. One way to achieve this is to take bootstrap samples of subjects and apply Ward's clustering algorithm to their contrast maps in order to build brain parcellations that best summarize the data subsamples, i.e. so that the parcel-level mean signal summarizes the signal within each parcel, in each subject. If enough subjects are used, all the parcellations offer a good representation of the whole dataset. It is important that the bootstrap scheme generates parcellations with enough entropy [30].

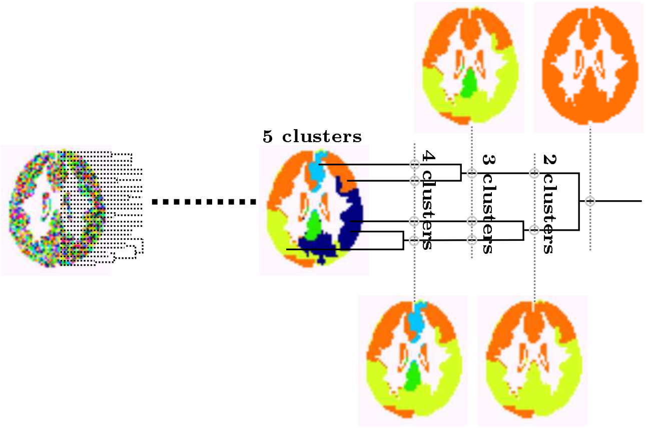


Figure 4.1: **Hierarchical clustering.** At each step, two clusters are merged so as to minimize the variance within the resulting cluster. The binary tree that represents the consecutive merges has to be cut according the a given number of desired clusters. In this example, the clustering algorithm was applied to only one subject's image.

4.1.2.1 Ward's clustering algorithm

Ward's clustering algorithm [33] is a particular case of hierarchical agglomerative clustering algorithm [13]: Starting with as many (singletons) clusters as observations, two clusters are merged at each step of the algorithm, until only one cluster containing all the observations remains. The history of the merges is kept in the form of a *tree* that has to be *cut* according to an expected number of clusters. Various criteria exist to decide which clusters should be merged at each step, yielding as many variants of the algorithm. Ward's clustering is one of those, in which the criterion for merging two clusters is that the intra-cluster variance of the resulting cluster should be minimal (as compared to the intra-cluster variance of other clusters resulting from potential merges). The main advantage associated with Ward's criterion is that variance minimization ensures that the mean is a good representation of the data.

Formally, let $\boldsymbol{Y} = \{\boldsymbol{y}_1, \dots, \boldsymbol{y}_p\} \in \mathbb{R}^{n \times p}$ be a set of n fMRI volumes described by p voxels each. For two clusters of voxels c and c', we define the distance:

$$\Delta(c, c') = \frac{|c||c'|}{|c| + |c'|} \| \langle \boldsymbol{Y} \rangle_c - \langle \boldsymbol{Y} \rangle_{c'} \|_2^2, \tag{4.1}$$

where $\langle \boldsymbol{Y} \rangle_c = \frac{1}{|c|} \sum_{j \in c} \boldsymbol{y}^j$. For each partition $C = \{c_1, \dots, c_k\}$ of the set of voxels $\boldsymbol{Y}$ (i.e. $\cup_{c \in C} = Y$ and $c_i \cap c_j = \emptyset \forall (c_i, c_j) \in C^2$), we note C^* the set of all pairs of clusters that share at least one neighboring voxel. Ward's clustering algorithm starts with an initial partition of p clusters $C = \{\{y_1\}, \dots, \{y_p\}\}$ that correspond to one singleton cluster per voxel. At each iteration, we merge the two clusters c_i and c_j of C^* that minimize the distance Δ:

$$(c_i, c_j) = \underset{(c, c') \in C^*}{\operatorname{argmin}} \Delta(c, c'). \tag{4.2}$$

The spatial constraint comes from the fact that we restrict the solution of the minimization criterion to C^*. When constructing a K-parcellations, the algorithm stops when $\operatorname{card}(C) = K$. In section 4.2.2.2, we use various Ward's clustering schemes that simply correspond to different choices for $\boldsymbol{Y}$.

We rely on the implementation of Ward's clustering that is available in the *Scikit-learn* Python package [20]. This implementation is a structured version of the algorithm that takes into account the topological structure between the samples (the *connectivity*). In neuroimaging, the connectivity is typically defined from the data mask. Figure 4.1 illustrates the application of Ward's clustering applied to fMRI images.

4.1.2.2 Influence of the sample size

The variability and quality of the parcellations are directly related to the variability and number of the images on which they are built. If few subjects are used, the benefit of taking the data structure into account is hindered. Conversely, if all the subjects are used, the parcellations may be very similar, due to the deterministic clustering algorithm. Obtaining variable parcellations is crucial (see next section), but the shape of the parcels is expected to match the anatomical organization of the brain. Thus, the number of subjects involved in the construction of the parcellations acts as a trading parameter between two crucial characteristics of the set of parcellations. Drawing a fixed number of subjects *with replacement* (i.e. *bagging*, see chapter 2) reduces the impact of that parameter, and alleviate the reliance on it.

Table 4.1 reports the evolution of both the variability of the parcellations and their averaged parcels spatial spread with the number of subjects. The variability of the parcellations is measured for each pair of parcellation as their adjusted mutual information [32]. The spatial spread of a parcel is measured as the mean of the variance of the parcel's voxels coordinates along each axis. In a first case, we select the subjects without replacement, leading to the extreme case where all the parcellations are the same (for parcellations built from the whole set of available images). In a second case, we draw the subjects with replacement to be sure that we obtain variable parcellations whatever the number of subject drawn. One should however include enough subjects in order to have population-representative parcels. When all the subjects are used, bag-

Subsample size (over 20)	1	5		10		15		20	
replacement	-	✗	✓	✗	✓	✗	✓	✗	✓
Adjusted mutual information	0.40	0.55	0.54	0.64	0.60	0.69	**0.62**	1	**0.65**
	(± 0.09)	(± 0.02)	(± 0.03)	(± 0.02)	(± 0.02)	(± 0.01)	(± 0.03)	(0)	(± 0.01)
Spatial spread (mm)	120.5	115.7	118.4	114.3	114.9	114.3	114.0	113.7	116.4
	(± 11.6)	(± 5.0)	(± 6.1)	(± 2.2)	(± 2.4)	(± 2.6)	(± 2.7)	(0)	(± 3.1)

Table 4.1: **Variability of the parcellation and spatial extent of the parcels according to the number of subjects involved in the construction of the parcellations.** Drawing the subjects with replacement helps obtaining variable parcels while ensuring a good amount of spatial regularization. Typically, one can perform a bootstrap subsampling of the subjects (i.e. drawing n subjects amongst n with replacement) as a default choice.

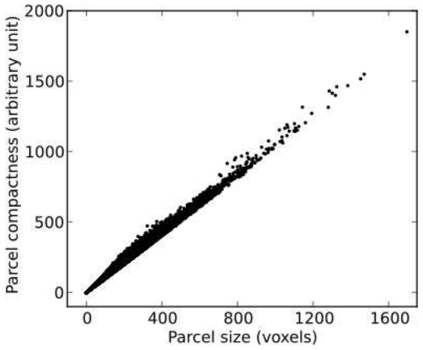


Figure 4.2: **Size and compactness of the parcels obtained with Ward's clustering algorithm on fMRI contrast maps.** For each parcel, the compactness is measured as a the difference between a mask of the parcel and its 1-eroded image). One can observe a great variability in parcel size/compactness, which reflects the structure of the individual fMRI contrast maps.

ging becomes *bootstrapping* (i.e. randomly drawing n subjects amongst n with replacement, see section 2.3.1): We advocate the use of this sampling procedure to build randomized parcellations.

4.1.2.3 Setting the number of parcels

We determined empirically that using 1000 parcels is a good trade-off between accurate parcellations and dimension reduction. This choice leads to using an average of 50 voxels per parcel, which seems relevant to describe the activation clusters given the typical inter-subjects variability. Note that this number of parcels is far from standard brain atlases with, at best, two hundred ROIs, suggesting that current atlases are not well-suited for such studies (see e.g. [29]). Figure 4.2 shows the size and compactness of the parcels for parcellations of 100 parcels obtained with Ward's clustering algorithm on real fMRI data.

4.2 Randomized Parcellation Based Inference

4.2.1 Method description

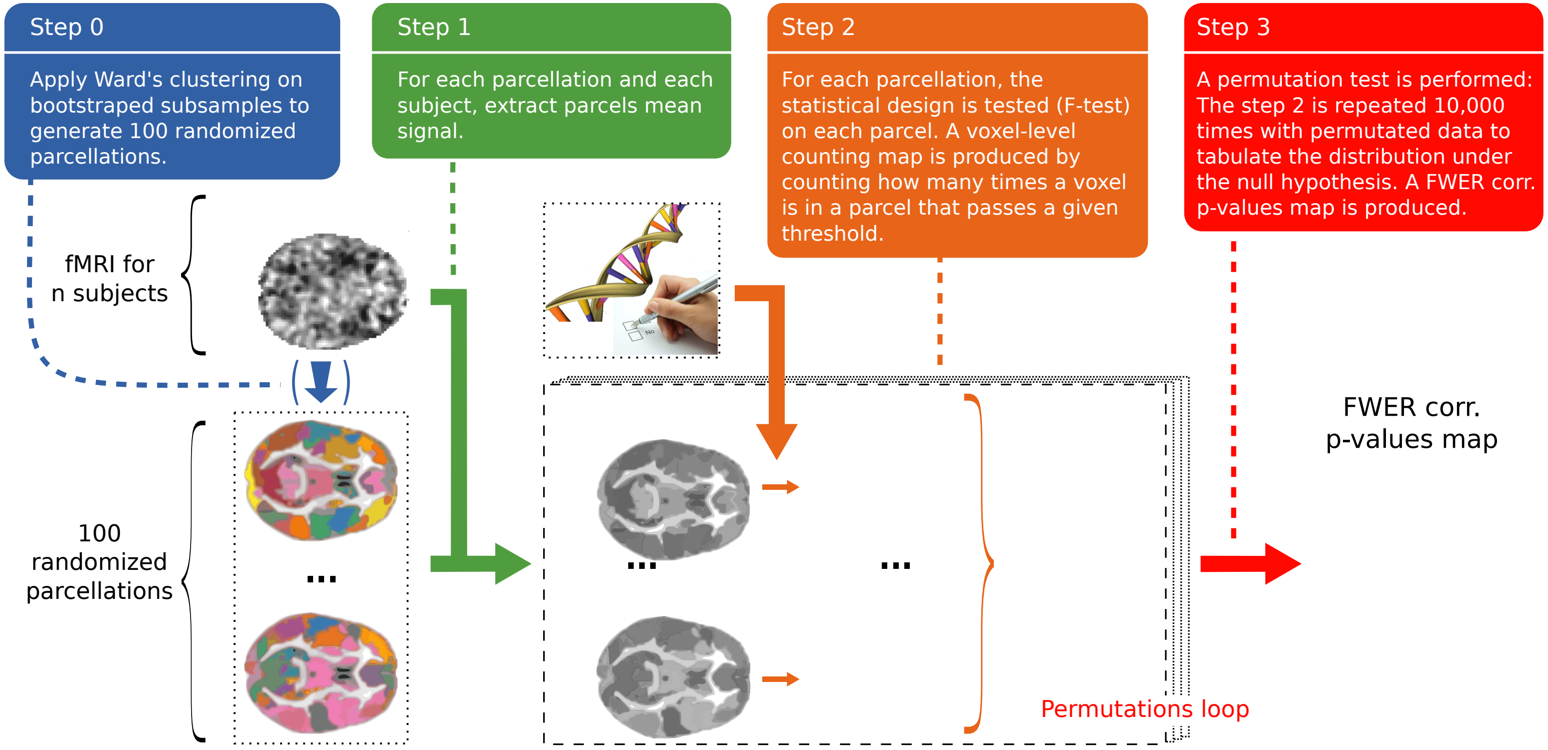
Randomized parcellation based inference (RPBI) performs several standard analyzes based on different parcellations and aggregates the corresponding statistical decisions. Let $\mathcal{P}$ be a finite set of parcellations, and V be the set of voxels under consideration. Given a voxel v and a parcellation P, the parcel-based thresholding function θ_t is defined as:

$$\theta_t(v, P) = \begin{cases} 1 \text{ if } F(\Phi_P(v)) > t \\ 0 \text{ otherwise} \end{cases} \tag{4.3}$$

where $\Phi_P : V \to P$ is a mapping function that associates each voxel with a parcel from the parcellation P ($\forall v \in P^{(i)}$, $\Phi_P(v) = P^{(i)}$). For a predefined test, F returns the F-statistic associated with the average signal of a given parcel. Finally, the aggregating statistic at a voxel v is given by the counting function C_t:

$$C_t(v, \mathcal{P}) = \sum_{P \in \mathcal{P}} \theta_t(v, P). \tag{4.4}$$

$C_t(v, \mathcal{P})$ represents the number of times the voxel v was part of a parcel associated with a statistical value larger than t across the folds of the analysis conducted on the set $\mathcal{P}$ of parcellations. We set the parameter t to ensure a



Step 0

Apply Ward's clustering on bootstraped subsamples to generate 100 randomized parcellations.

Step 1

For each parcellation and each subject, extract parcels mean signal.

Step 2

For each parcellation, the statistical design is tested (F-test) on each parcel. A voxel-level counting map is produced by counting how many times a voxel is in a parcel that passes a given threshold.

Step 3

A permutation test is performed: The step 2 is repeated 10,000 times with permutated data to tabulate the distribution under the null hypothesis. A FWER corr. p-values map is produced.

Figure 4.3: Overview of the randomized parcellation based inference framework on an example with few parcels. The variability of the parcels definition is used to obtain voxel-level statistics.

Bonferroni-corrected control at $p < 0.1$ [1] in each of the parcel-level analyzes. In practice, the results are weakly sensitive to mild variations of t. In order to assess the significance of the counting statistic at each voxel, we perform a permutation test, i.e. we tabulate the distribution of $C_t(v, \mathcal{P})$ under the null hypothesis that there is no significant correlation between the voxels' mean signal and the target variable. As a result, we get a voxel-wise p-values map similar to a standard group analysis map (see Figure 4.3). We obtain family-wise error control by tabulating the maximal value across voxels in the permutation procedure. The θ_t function of Equation 4.3 can be replaced by any function that is convex with respect to t. In particular, the natural choice $\theta_t(v, P) = F(\Phi_P(v))$ yields similar results (not shown in the paper) but its computation requires much more memory since the $v \to \theta_t(v, P)$ mapping and bootstrap averages are no longer sparse. We discuss the pivotality properties of the statistic in the discussion section of this chapter (4.3.4).

Spatial models try to address the problem of imperfect voxel-to-voxel correspondence after co-registration of the subjects in the same reference space. Our approach is clearly related to anisotropic smoothing [24], in the sense that obtained parcels are not spherical and in the aggregation of the signals of voxels in a given parcel, certain directions are preferred. Unlike smoothing or spatial modeling applied as a preprocessing, our statistical inference embeds the spatial modeling in the analysis and decreases the number of tests and their dependencies. In addition to the expected increase of sensitivity, the randomization of the parcellations ensures a better reproducibility of the results, unlike inference on one fixed parcellation.

4.2.2 Comparison with state-of-the-art methods

4.2.2.1 Simulations

Description. We simulate fMRI contrast images as volumes of shape $40 \times 40 \times 40$ voxels. Each contrast image contains a simulated $4 \times 4 \times 4$ activation patch at a given location, with a spatial jitter following a three-dimensional $\mathcal{N}(\mathbf{0}, \boldsymbol{I}_3)$ distribution (coordinates of the jitter are rounded to the nearest integers). The strength of the activation is set so that the signal to noise ratio (SNR) peaks at 2 in the most associated voxel. The background noise is drawn from a $\mathcal{N}(0, 1)$ distribution, Gaussian-smoothed at σ_{noise} isotropic and normalized by its global empirical standard deviation. After superimposing noise and signal images, we optionally smooth at $\sigma_{\text{post}} = 2.12$ voxels isotropic, corresponding to a 5 voxels Full Width at Half Maximum (FWHM). Voxels with a probability above 0.1 to be active in a large sample test are considered as part of the ground truth. Ten subsamples of 20 images are drawn to perform analyzes. Each time, RPBI was conducted with one hundred 1000-parcellations built from a bootstrapped selection of the 20 images involved.

For each of the 10 groups, we expect to obtain a p-values map that shows a significant effect at the mean location of generated artificial activations in the contrast images.

We investigate the ability of four methods to actually recover the region of activation:

[1] We determine this value empirically to obtain a well-behaved null distribution of the counting statistic. With 1 target and 1,000 parcels, it corresponds to a raw p-value $< 10^{-4}$.

(i) voxel-level group analysis, which is the standard method in neuroimaging;

(ii) cluster-size group analysis – introduced in chapter 2, section 2.1.3.2, which is know to be more sensitive than voxel-intensity group analysis [17, 7, 21];

(iii) threshold-free cluster enhancement (TFCE) [23] – introduced in chapter 2, section 2.1.3.3;

(iv) RPBI, which is our contribution.

We control the specificity of each procedure by permutation testing. In order to ensure an accurate type 1 error control, we generate 400 sets of 20 images with no activation (i.e. the images are only noise with $\sigma_{\text{noise}} = 1$, and SNR $= 0$). We evaluate the false positive rate at voxel level for RPBI.

Results. Table 4.2 gives the number of times a significant effect was reported according to the different methods. RPBI always achieves more or as many detections as the other approaches. Since the specificity of the detections is controlled for all the methods at 5%, corrected for multiple comparisons, the results indicates that RPBI is more sensitive. Voxel-intensity group analysis is the only method that benefits from a posteriori smoothing, while spatial methods (cluster-level, TFCE, RPBI) lose accuracy when the images are smoothed. This is in agreement with the theory and the results of [34]. Figure 4.4 shows that detections made by spatial methods (cluster-size group analysis, TFCE and RPBI) does not come with wrongly reported effects in voxels close to the actual effect location. That would be the case for a method that simply extends a recovered effect to the neighboring voxels and would wrongly be thought to be more sensitive because it points out more voxels. RPBI offers the best precision-recall compromise as its precision-recall curve dominates in Figure 4.4. The shape of the curves for the cluster-size method illustrates the problem of the cluster-forming threshold: most voxels do not pass the threshold and then were discarded by the method, leading to a precision equal to zero. The cluster-forming threshold directly acts on the recovery capability of the method, but lowering the threshold does not increase the sensitivity of this approach in general. By integrating over multiple thresholds, the TFCE partially addresses this issue. When no signal is put in the data (SNR $= 0$), RPBI reports an activation 37 times over 400 at $P < 0.1$ FWER corrected, 20 times at $P < 0.05$ FWER corrected, and 4 times at $P < 0.01$ FWER corrected. In all cases, it corresponds to the nominal type I error rate.

4.2.2.2 Real data

Random effects analysis. In this experiment, we work with an *[angry faces - control]* fMRI contrast. 1567 subjects were available after removal of the subjects with too many missing data and/or bad/missing covariables. After standard preprocessing of the images, including registration of the subjects onto the same template, we test each voxel for a zero mean across the 1567 subjects with an OLS regression, including handedness and gender as covariables, yielding a reference voxel-wise p-values map. We threshold this map in order to keep 5% of the most active voxels (corresponding to $-\log_{10} P > 77.5$), and we consider it as the ground truth. Since we use a voxel based threshold, the ground truth may be biased to voxel-level statistics (thus disadvantaging our method).

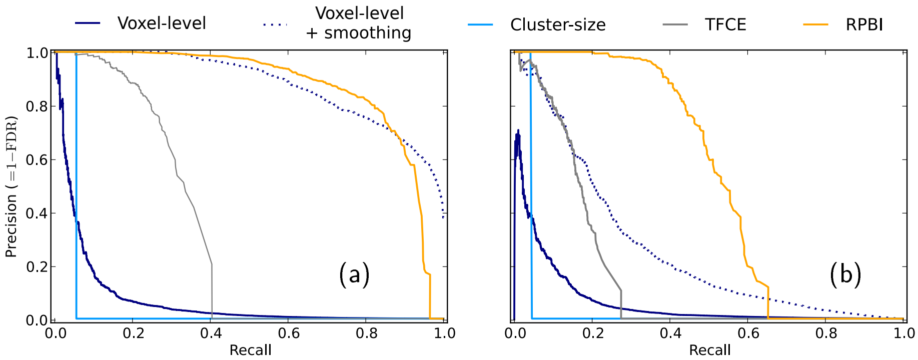


Figure 4.4: **Simulated data.** Precision-recall curves for various analysis methods across 10 random subsamples containing 20 subjects. SNR = 2 and noise spatial smoothness: (a) $\sigma_{\text{noise}} = 0$, (b) $\sigma_{\text{noise}} = 1$. The curves are obtained by thresholding the statistics brain maps at various levels, yielding as many precision-recall points.

Our objective is to retrieve the population's reference activity pattern on subsamples of 20 randomly drawn subjects and compare the performance of several methods in this problem. Because of the reduced number of subjects used, we cannot expect to retrieve the same activation map as in the full-sample analysis due to a loss in statistical power. We therefore measure the sensitivity and we build precision-recall curves to assess the performance of the methods. We perform our experiment on 10 different subsamples and we use the same analysis methods as the previous experiment. We propose to observe the behavior of our method with the use of parcellations of different kinds. We perform analysis of the 10 different subsamples with the following parcellation schemes:

(i) RPBI (sh. parcels) with parcellations built on bootstrapped subsamples of 150 images amongst the 1567 images corresponding to the fMRI contrast under study;

(ii) RPBI (alt. parcels) with shared parcellations built on images corresponding to another, independent fMRI contrast;

(iii) RPBI (rand. parcels) with shared parcellations built on smooth Gaussian noise (FWHM = 2 voxels);

We also assess the stability of all these methods by counting how many times each voxel was associated with a significant effect across subgroups. We present

σ_{noise}	0		1		2	
post-smoothing	✗	✓	✗	✓	✗	✓
Voxel-intensity	3	**10**	3	4	3	**2**
Cluster-size	9	**10**	6	3	1	0
TFCE	**10**	**10**	5	2	2	1
RPBI	**10**	**10**	**7**	4	**4**	**2**

Table 4.2: **Simulated data.** Frequency (over 10 simulations) of significant effect reporting according to the analysis methods (peak SNR = 2; significance threshold at $P < 0.05$ corrected). For several values of σ_{noise} used to smooth the white noise, the left column report detection without post-smoothing and right column with standard post-smoothing with FWHM = 5 voxels. RPBI reports more frequently some effects, which shows that the method is more sensitive. Generally, a posteriori smoothing undermines the spatial methods' performance.

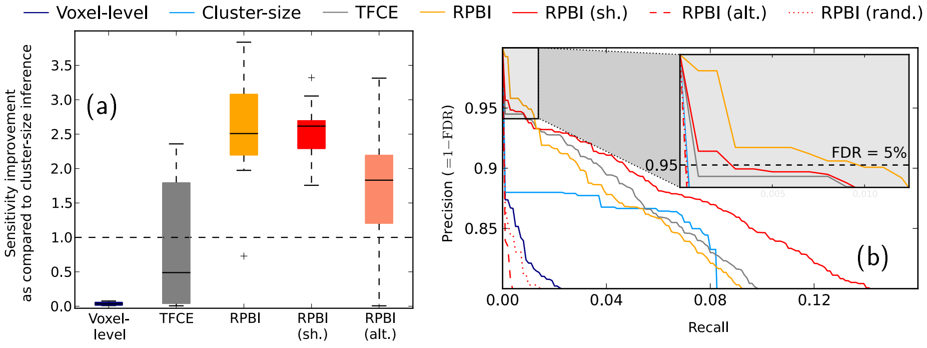


Figure 4.5: **Real fMRI data.** Evaluation of the performances for various analysis methods across 10 random subsamples containing 20 subjects. (a) Sensitivity improvement relative to cluster-size under control of the specificity at 5% FWER. (b) Precision-recall curves built with a pseudo ground truth where 5% of the most active voxels across 1567 subjects are kept. RPBI achieves the best sensitivity improvement and shows a good recall rate at FDR=5%.

the inverted cumulative normalized histogram of that count for each method, restricting our attention to the voxels that were reported at least once. A method is considered to be more stable than another if the same voxels appear more often, that is if its histogram shows many high values.

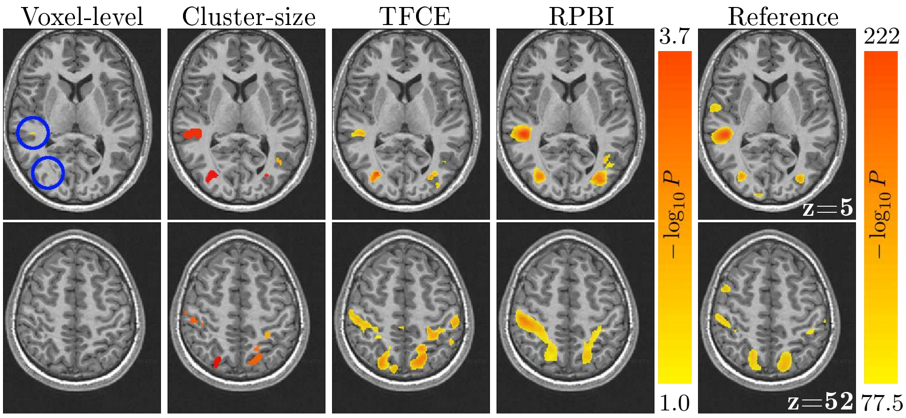
Neuroimaging-genetic study. The aim of this experiment is to show that RPBI has the potential to uncover new relationships between neuroimaging and genetics. We consider an fMRI contrast corresponding to events where subjects make motor response errors (*[go wrong]* fMRI contrast from a Stop Task Signal) and its associations with *Single-Nucleotide Polymorphisms (SNPs)* in the *COMT* gene. This gene codes for the Catechol-O-methyltransferase, an enzyme that catalyzes transfer of neurotransmitters like dopamine, epinephrine and norepinephrine, making it one of the most studied genes in relation to brain. Subjects with too many missing voxels in the brain mask or with bad task performance were discarded. Regarding genetic variants, we keep only 27 SNPs in the *COMT* gene (±20kb) that pass *plink* standard parameters (Minor Allele Frequency < 0.05, Hardy-Weinberg Equilibrium $P < 0.001$, missing rate per SNP < 0.05). Age, sex, handedness and acquisition center were included in the model as confounding variables. Remaining missing data were replaced by the median over the subjects for the corresponding variables. After those filtering steps, our experiment involves 1,372 subjects.
For each of the 27 SNPs, we perform a massively univariate voxel-wise analysis with the algorithm presented in [4], including cluster-size analysis [11], and RPBI through 100 different Ward's 1000-parcellations; 10,000 permutations were performed to assess statistical significance with a good degree of confidence.

Results. Figure 4.5a shows the sensitivity improvement relative to cluster-size for various analysis methods under control for false detections at 5% FWER. Cluster-size was taken as the reference because it is the method that yields the most sensitivity amongst state-of-the-art methods to which we compare RPBI to. RPBI achieves the best sensitivity improvement, and RPBI with shared, alternative or random parcels are always more sensitive than TFCE. Voxel-level group analysis yields poor performance while cluster-size analysis is comparable to TFCE. These gains in sensitivity should be linked with a measure of accuracy (see section 4.2.2.2). In our experiments, to estimate a method's accuracy, we construct precision-recall curves by reporting the proportion of true positives in the detections (precision $= 1 -$ False Discovery Rate, FDR, see section 2.1.1.2) for different levels of recovery of the ground truth (the recall is the proportion of true positives among detections). The curves have to be read vertically: at a fixed level of precision, the best method is the one with the highest recall (ie. less false negatives). In practice, it is standard to choose a FDR at 5%. Precision-recall analysis constitute an alternative to ROC analysis that is better suited to unbalanced classes [5].

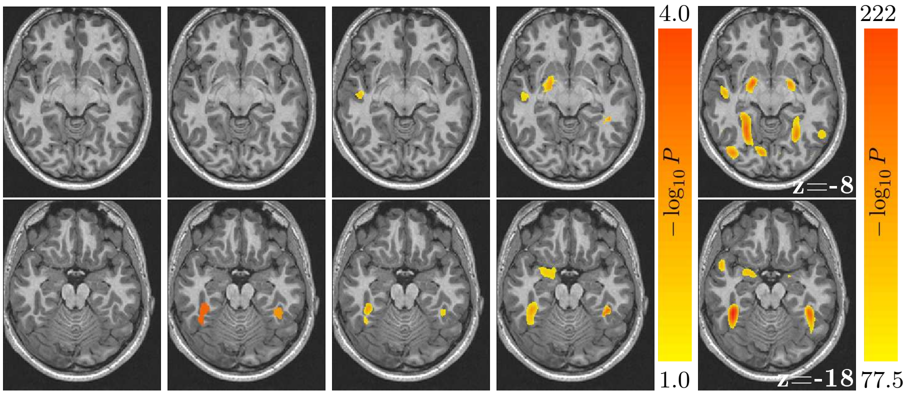
Figure 4.5b shows the precision-recall curves associated with the performance of the methods under comparison. For acceptable levels of precision, RPBI outperforms other methods when we use parcellations that have been built on the contrast under study. RPBI with alternative or random parcels yields poor recovery although these approaches are based on the randomized parcellation scheme. This demonstrates that the sensitivity is not a sufficient criterion and that the choice of parcellations plays an important role in the success of RPBI. Unlike simulations, real data may contain outliers, which reduce the effectiveness of all the presented methods. One benefit of RPBI with shared parcels is that the impact of a bad observation is lowered, because the fitness of the parcellations no longer depends on the analyzed data but similar data in greater quantity. This requires other data from a similar protocol, but Figure 4.5b shows that this approach outperforms other methods in terms of recall.

The lack of reproducibility of group studies is a well-known issue [26, 28], but it can be improved if better models are used. RPBI has better reproducibility than the other methods, as shown in Figure 4.7. The histogram of the RPBI method dominates, which means that significant effects were reported more often at the same location (i.e. the same voxel) across subgroups when using RPBI than when using the other methods. For RPBI with shared parcels, it is even more pronounced and this is explained by the fact that parcellations are shared across subgroups, which is another advantage to this method. In general, the same activation peaks arises from the cluster-size, the TFCE and the RPBI maps (see Figure 4.6). The TFCE slightly improves the results of cluster-size and provides voxel-level information. As is can be seen in Figure 4.6, the map returned by RPBI better matches the patterns of the reference map and is less scattered. Voxel-based group analysis clearly fails to detect some of the activation peaks.

The SNP rs917478 yields the strongest correlation with the phenotypes and lies in an intronic region of *ARVCF*. The number of subjects in each genotype group is balanced: 523 homozygous with major allele, 663 heterozygous and 186 homozygous with minor allele. For RPBI, 31 voxels (resp. 81) are significantly associated with that SNP at $P < 0.05$ corrected (resp. $P < 0.1$) in the left thalamus, a region involved in sensory-motor cognitive tasks. The association peak has a p-value of 0.016 FWER corrected. Cluster-size inference finds this effect but with a higher p-value ($P = 0.046$). Voxel-based inference does not find any significant effect. A significant association for rs917479 is only reported



(a) subgroup no. 1



(b) subgroup no. 2

Figure 4.6: Negative logp-value associated with a non-zero intercept test with confounds (handedness, site, gender). The subgroups maps are thresholded at $-\log_{10} P > 1$ corrected and the reference map at $-\log_{10} P > 77.5$ (i.e. 5% of the most active voxels). Small activation clusters are surrounded with a blue circle in order to make them visible.

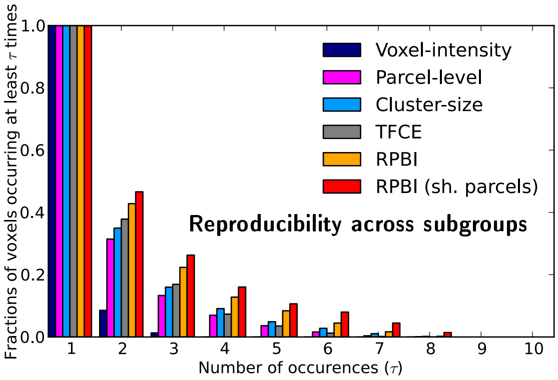


Figure 4.7: **Real fMRI data.** Inverse cumulative histograms of the relative number of voxels that were reported as significant several times through the 10 subsamples ($P < 0.05$ FWER corrected).

by RPBI; The Figure 4.8 shows that this SNP is in high linkage disequilibrium (LD) with rs917478 ($D' = 0.98$ and $R^2 = 0.96$). Those SNPs are also in LD with rs9306235 and rs9332377 in *COMT*, the targeted gene for this study. Figure 4.8 shows the thresholded p-values maps obtained with RPBI with rs917478. The *ARVCF* gene has already been found to be associated with intermediate brain phenotypes and neurocognitive error tests in a study about schizophrenia [22]. We applied our method on this gene, for which we have 33 SNPs, and did not find any effect except from rs917478 and SNPs in LD with it.

4.2.2.3 Outlier detection

We finally apply the concept of randomized parcellations to outlier detection. We work with a cohort of 1886 fMRI contrast images. In a first step, we randomly select 300 subjects and summarize the dataset by computing a 500-parcellation (obtained by Ward's) and averaging signal over each parcel. We perform a reference outlier detection on this dataset with a regularized version of a robust covariance estimator *RMCD-RP* [9]. This outlier detection algorithm consists in fitting robust covariance estimators to random data projections. For the outliers detection we use the average of the Mahalanobis distances of the observations to the population mean in every projection subspace. In a second step, we perform outlier detections with RMCD-RP on random subsamples : We randomly draw a subsample of n subjects and perform 100 outlier detections with RMCD-RP on 100 different p-dimensional representations of the data defined by 100 Ward's p-parcellations built on 300 bootstrapped subjects from the whole cohort. Following the model of RPBI, we report how many times each subject was reported as an outlier through these 100 outlier detections and we use that number as an outlier score. We hence construct two Receiver Operating Characteristic (ROC) curves [10]: one for randomized parcellations-based (RPB) outlier detection and the other as the average ROC curve of the 100 inner outlier detections used to obtain the RPB outlier detection. Finally, we report the rate of correct detections when 5% of false detections are accepted, to control the sensitivity of this test when wrongly rejecting few non-outlier data. These statistics make it possible to easily measure the accuracy improvement of RPB outlier detection across several experiments performed with different subsamples of n subjects (keeping the same reference decision obtained at the first step). In our experiment, we choose to work with $p = 100$ and $n = \{80, 100, 200, 300, 400\}$, yielding p/n configurations that correspond to various problem difficulty. For a fixed (n, p) couple, we run the experiment on 50 different subsamples and we present the rate of correct detections in a box-plot.

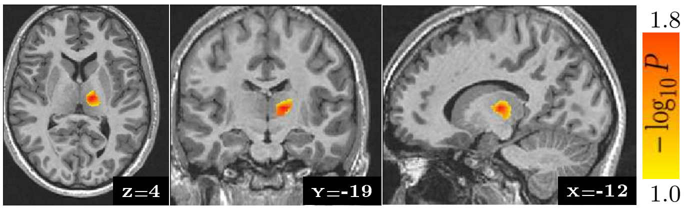


Figure 4.8: **Association study between 27 SNPs from the *COMT* gene and fMRI contrast phenotypes.** Corrected p-values map (thresholded at $P < 0.1$) obtained with RPBI for rs917478, the SNP with the strongest reported effect.

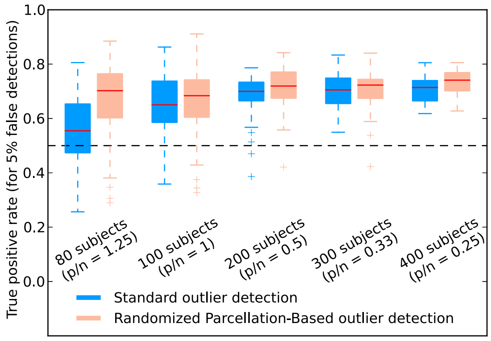


Figure 4.9: **Outliers detection with randomized parcellations.** Proportion of observations correctly tagged as outliers when 5% errors are accepted. Results are represented as boxes according to the number of subjects present in the subsamples in which we seek for outliers. Chance level is given by the dashed black line. RPB outlier detection always outperforms standard outlier detection, although the difference between both is small and may not worth the implementation and computation costs. It is larger in the case where there are more features than subjects.

Results. Figure 4.9 illustrates the accuracy of RPB outlier detection as compared to standard outlier detection performed on data issued from a single parcellation. We present the rate of correct detections when 5% false detections is accepted. Since the experiment is conducted on 50 subsamples of n subjects, we present the results for various values of n ($n \in \{80, 100, 200, 300, 400\}$) with box-plots. For a large number of subjects (low-dimensional settings: $n < p$) RPB outlier detection performs slightly better than standard outlier detection, while in high-dimensional settings ($p > n$) it clearly outperforms the classical approach. Relative results are the same when allowing for any proportion of false detection comprised between 0% and 10%.

4.3 Discussion

4.3.1 Influence of smoothing and images properties

Our first experiment shows that RPBI performance drops when additional smoothing of the images if performed. Unlike voxel-intensity analysis, cluster-size analysis, TFCE and RPBI, which are spatial methods, suffer from data smoothing. In the presence of smooth noise, this experiment also shows that RPBI outperforms other methods. Our experiment on real data shows that RPBI can recover activations clusters of various size and shape, as can be seen on the effects maps reported in Figure 4.6. Yet, the use of parcels clearly helps focusing on activations with a spatial extent of the order of the average parcel size. Cluster-size group analysis also focuses more easily on some activations with a given size, according to internal parameters such as the cluster forming threshold or an optional data smoothing. TFCE is designed to address this issue and clearly enhances the results of the cluster-size inference.

4.3.2 Sensitivity and reproducibility

Usually, the sensitivity of a procedure is compared under a given control for false positives. Under this criterion, RPBI outperforms voxel- intensity, cluster-size analysis and TFCE (Figure 4.5). By aggregating 100 × 1000 measurements, RPBI drastically reduces the multiple comparisons problem and stabilizes parcel-based statistics. Neuroimaging studies are subject to a lack of reproducibility and using the most sensitive procedure does not guarantee to unveil reproducible results [26, 28]. By nature, the randomization of the par-

cellations ensures a good stability of the procedure. Experiments on real data show the gain in terms of reproducibility of RPBI compared to other methods when the subset of subjects changes (Figure 5). RPBI with shared parcels has a better recovery and yields results that are more reproducible across various analysis settings.
Randomized parcellation can be applied to various neuroimaging tasks. However, sensitivity improvement is not straightforward and may depend on problem-specific settings. In particular, our experiment about outlier detection suggests that multivariate statistical algorithms require a more subtle use of randomized parcellation in order to get significant sensitivity improvement.

4.3.3 Computational aspects

The procedure is separated in two distinct steps: *(i)* the generation of the 100 Ward K-parcellations and extraction of the signal means, then *(ii)* the statistical inference. The generation of parcellations is optional (parcellations can be replaced by precomputed ones), but Ward's hierarchical clustering algorithm is fast and this step takes only few minutes on a desktop computer for 100 parcellations. The second step involves a permutation test. Our implementation fits a Massively Univariate Linear Model [25, 4] in an optimized version adapted to permutation testing and our application. As a result, in our experiments with 20 subjects and 10,000 permutations, the statistical inference takes only 1 minutes×cores, i.e. 5 seconds on a 12-core computer. The total computation time thus amounts to a few minutes on a desktop computer and is limited by the construction of the parcellations. Asymptotically, the computation time increases linearly with the number of subjects and the number of variables to test, which is a desirable property to scale to larger problem like neuroimaging-genetic studies.

4.3.4 Pivotality of the counting statistic

An important question is whether the counting statistic introduced in Eq. 4.3 is a valid statistic to detect activated voxels. One essential criterion for this is to check the pivotality, i.e. the convergence –under the null hypothesis– of the statistic distribution toward a law that is invariant under data distribution parameters. In the present case, the main deviation from pivotality could result

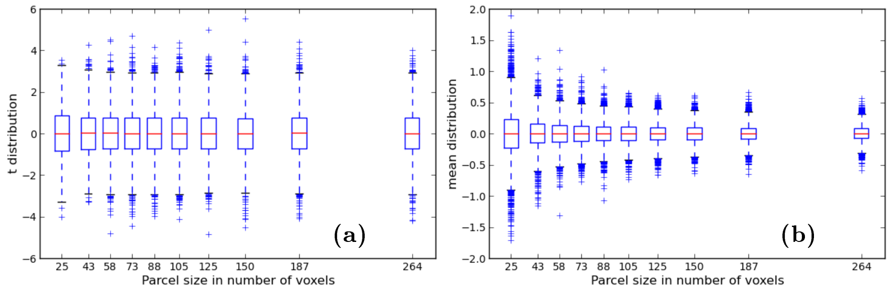


Figure 4.10: **Impact of the parcel size on the distribution of the second-level one-sample t statistic (a) and of the mean value (b).** While there is an obvious effect on the mean, there is no conspicuous effect on the t distribution.

from a distribution of (extreme) statistical values that depends on the parcel size: large parcels would represent fMRI signal averaged over larger domains, and thus would get typically lower values. This is indeed typically the case for the mean statistic (see Figure 4.10, (b)); however, we show for instance that the t statistic used in Section 4.2.2.2 is very weakly influenced by the parcel size: we repeated the experiment described in section 4.2.2.2, i.e. computing the t statistic on parcels obtained by Ward's algorithm, based on 100 random batches of 20 subjects, after permutation by random sign swap. We tabulate the t distribution according to the parcel size by using 10 size bins. The result, shown in Figure 4.10, (a), is that the effect, if any, is not detectable by visual inspection.

To test more precisely the independence on the t distribution with respect to the parcel size, we tested the equality of the mean, median and variance of the size-specific distributions using the One-way (mean), Kruskal (median), Bartlett (variance), Levene (variance) and Fligner (variance) tests as implemented in the SciPy library[2]. All the tests are performed on the 10 bins jointly. We obtain the following p-values: Oneway, $P = 0.36$; Kruskal, $P = 0.27$; Bartlett: $P = 0.95$; Levene: $P = 0.016$; Fligner: $P = 0.06$. This means that there is only a small effect on the variance, as reported by the Levene test, that is more sensitive than Fligner (which is non-parametric) and Bartlett, which assumes Gaussian distributions. However this effect is very small, and has no obvious consequence on the number of peak values of the statistic; in particular, we do not observe monotonic trends with size. Note that the small effect fades out when using larger number of subjects (here, only $n = 20$ subjects per groups were used). Finally, we did not find any significant correlation between the number of detections above a given threshold (using uncorrected p-values of $10^{-2}, 10^{-3}, 10^{-4}$) and the parcel size.

As a conclusion, the effect of parcel size is too small to jeopardize the usefulness of the counting statistic.

[2] http://www.scipy.org/

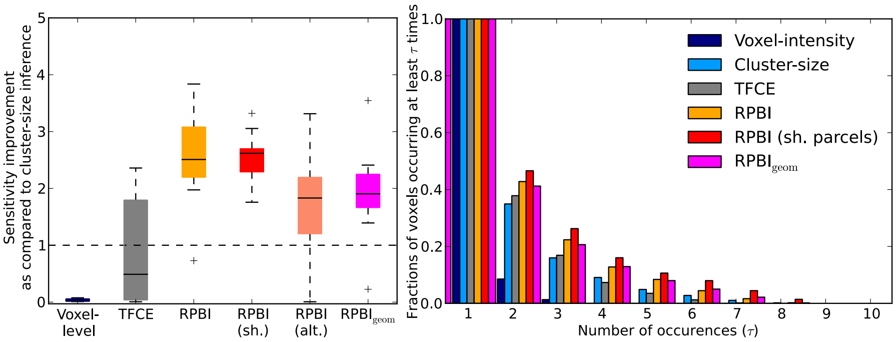


Figure 4.11: **(a)** Sensitivity improvement relative to cluster-size under control of the specificity at 5% FWER. **(b)** Inverse cumulative histograms of the relative number of voxels that were reported as significant several times through the 10 subsamples ($P < 0.05$ FWER corrected), on a *[angry faces - control]* fMRI contrast from the *faces* protocol. RPBI with geometric parcellations yields poor sensitivity and poor reproducibility.

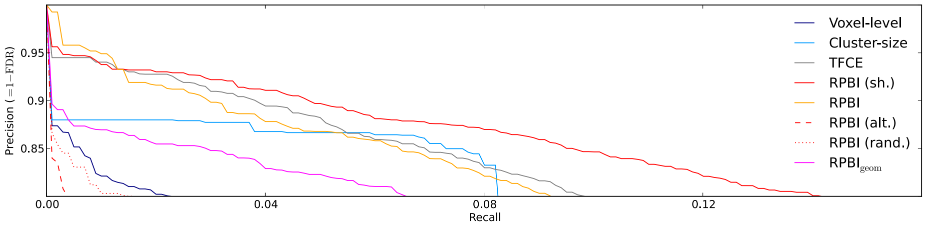


Figure 4.12: **Real fMRI data – RPBI with geometric parcellations.** Evaluation of the performances for various analysis methods across 10 random subsamples containing 20 subjects, on a *[angry faces - control]* fMRI contrast from the *faces* protocol. Precision-recall curves built with a pseudo ground truth where 5% of the most active voxels across 1430 subjects are kept. RPBI with geometric parcellations has poorer performance than RPBI with Ward's clustering parcellations.

4.3.5 Geometric parcellations

We run experiment on real data with parcellations coming from a geometric parcellation approach. We built parcellations of 1000 parcels with the K-means clustering algorithm [15] (using random initializations) on the 3D coordinates of a brain mask voxels. Geometric parcellations yield more regular parcels than those obtained by performing a Ward's clustering algorithm on simulated and real data. Geometric parcellations lose the anisotropic effect of Ward's parcellations. In practical terms, they do not give good results, as compared to RPBI with Ward's clustering.

References

[1] Alexander, D.H., Lange, K.: Stability selection for genome-wide association. Genet Epidemiol 35(7), 722–728 (Nov 2011)

[2] Bohland, J.W., Bokil, H., Allen, C.B., Mitra, P.P.: The brain atlas concordance problem: quantitative comparison of anatomical parcellations. PLoS One 4(9), e7200 (2009)

[3] Bühlmann, P., Rütimann, P., van de Geer, S., Zhang, C.H.: Correlated variables in regression: clustering and sparse estimation. ArXiv e-prints (Sep 2012)

[4] Da Mota, B., Frouin, V., Duchesnay, E., Laguitton, S., Varoquaux, G., Poline, J.B., Thirion, B.: A fast computational framework for genome-wide association studies with neuroimaging data. In: 20th International Conference on Computational Statistics (2012)

[5] Davis, J., Goadrich, M.: The relationship between precision-recall and ROC curves. In: Proceedings of the 23rd international conference on Machine learning. pp. 233–240. ACM (2006)

[6] Flandin, G., Penny, W.D.: Bayesian fMRI data analysis with sparse spatial basis function priors. Neuroimage 34(3), 1108–1125 (Feb 2007)

[7] Friston, K.J., Holmes, A., Poline, J.B., Price, C.J., Frith, C.D.: Detecting activations in PET and fMRI: levels of inference and power. Neuroimage 4(3 Pt 1), 223–235 (Dec 1996)

[8] Friston, K.J., Penny, W.: Posterior probability maps and SPMs. Neuroimage 19(3), 1240–1249 (Jul 2003)

[9] Fritsch, V., Varoquaux, G., Thyreau, B., Poline, J.B., Thirion, B.: Detecting outliers in high-dimensional neuroimaging datasets with robust covariance estimators. Med Image Anal 16(7), 1359 – 1370 (2012)

[10] Hanley, J., McNeil, B.: The meaning and use of the area under a receiver operating (ROC) curve characteristic. Radiology 143(1), 29–36 (1982)

[11] Hayasaka, S., Nichols, T.E.: Validating cluster size inference: random field and permutation methods. Neuroimage 20(4), 2343–2356 (Dec 2003)

[12] Hutton, C., Bork, A., Josephs, O., Deichmann, R., Ashburner, J., Turner, R.: Image distortion correction in fMRI: A quantitative evaluation. Neuroimage 16(1), 217–240 (May 2002)

[13] Johnson, S.C.: Hierarchical clustering schemes. Psychometrika 32(3), 241–254 (1967)

[14] Keller, M., Lavielle, M., Perrot, M., Roche, A.: Anatomically informed bayesian model selection for fMRI group data analysis. Med Image Comput Comput Assist Interv 12(Pt 2), 450–457 (2009)

[15] MacQueen, J., et al.: Some methods for classification and analysis of multivariate observations. In: Proceedings of the fifth Berkeley symposium on mathematical statistics and probability. p. 14. California, USA (1967)

[16] Meinshausen, N., P.Bühlmann: Stability selection. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 72(4), 417–473 (2010)

[17] Moorhead, T.W.J., Job, D.E., Spencer, M.D., Whalley, H.C., Johnstone, E.C., Lawrie, S.M.: Empirical comparison of maximal voxel and non-isotropic adjusted cluster extent results in a voxel-based morphometry study of comorbid learning disability with schizophrenia. Neuroimage 28(3), 544–552 (Nov 2005)

[18] Nieto-Castanon, A., Ghosh, S.S., Tourville, J.A., Guenther, F.H.: Region of interest based analysis of functional imaging data. Neuroimage 19(4), 1303–1316 (Aug 2003)

[19] Ou, W., Wells, W.M., Golland, P.: Combining spatial priors and anatomical information for fMRI detection. Med Image Anal 14(3), 318–331 (Jun 2010)

[20] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al.: Scikit-learn: Machine learning in Python. Journal of Machine Learning Research 12, 2825–2830 (2011)

[21] Poline, J.B., Worsley, K.J., Evans, A.C., Friston, K.J.: Combining spatial extent and peak intensity to test for activations in functional imaging. Neuroimage 5(2), 83–96 (Feb 1997)

[22] Sim, K., Chan, W.Y., Woon, P.S., Low, H.Q., Lim, L., Yang, G.L., Lee, J., Chong, S.A., Sitoh, Y.Y., Chan, Y.H., Liu, J., Tan, E.C., Williams, H., Nowinski, W.L.: ARVCF genetic influences on neurocognitive and neuroanatomical intermediate phenotypes in Chinese patients with schizophrenia. J Clin Psychiatry 73(3), 320–326 (Mar 2012)

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

[24] Solé, A.F., Ngan, S.C., Sapiro, G., Hu, X., Lòpez, A.: Anisotropic 2-D and 3-D averaging of fMRI signals. IEEE Trans Med Imaging 20(2), 86–93 (Feb 2001)

[25] Stein, J.L., Hua, X., Lee, S., Ho, A.J., Leow, A.D., Toga, A.W., Saykin, A.J., Shen, L., Foroud, T., Pankratz, N., et al.: Voxelwise genome-wide association study (vGWAS). Neuroimage 53(3), 1160–1174 (2010)

[26] Strother, S.C., Anderson, J., Hansen, L.K., Kjems, U., Kustra, R., Sidtis, J., Frutiger, S., Muley, S., LaConte, S., Rottenberg, D.: The quantitative evaluation of functional neuroimaging experiments: the NPAIRS data analysis framework. Neuroimage 15(4), 747–771 (Apr 2002)

[27] Thirion, B., Flandin, G., Pinel, P., Roche, A., Ciuciu, P., Poline, J.B.: Dealing with the shortcomings of spatial normalization: multi-subject parcellation of fMRI datasets. Hum Brain Mapp 27(8), 678–693 (Aug 2006)

[28] Thirion, B., Pinel, P., Mériaux, S., Roche, A., Dehaene, S., Poline, J.B.: Analysis of a large fMRI cohort: Statistical and methodological issues for group analyses. Neuroimage 35(1), 105–120 (Mar 2007)

[29] Tucholka, A., Thirion, B., Perrot, M., Pinel, P., Mangin, J.F., Poline, J.B.: Probabilistic anatomo-functional parcellation of the cortex: how many regions? In: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2008, pp. 399–406. Springer (2008)

[30] Varoquaux, G., Gramfort, A., Thirion, B.: Small-sample brain mapping: sparse recovery on spatially correlated designs with randomization and clustering. In: John, L., Joelle, P. (eds.) International Conference on Machine Learning. Edimbourg, United Kingdom (Jun 2012)

[31] Ville, D.V.D., Blu, T., Unser, M.: Integrated wavelet processing and spatial statistical testing of fMRI data. Neuroimage 23(4), 1472–1485 (Dec 2004)

[32] Vinh, N.X., Epps, J., Bailey, J.: Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. The Journal of Machine Learning Research 9999, 2837–2854 (2010)

[33] Ward, J.: Hierarchical grouping to optimize an objective function. Journal of the American Statistical Association 58(301), 236–244 (1963)

[34] Worsley, K.J., Marrett, S., Neelin, P., Evans, A.C.: Searching scale space for activation in PET images. Hum Brain Mapp 4(1), 74–90 (1996)

[35] Wu, D.H., Lewin, J.S., Duerk, J.L.: Inadequacy of motion correction algorithms in functional MRI: Role of susceptibility-induced artifacts. Journal of Magnetic Resonance Imaging 7(2), 365–370 (1997)

Chapter 5

Dealing with multiple sources of variability in neuroimaging with robust regression

Contents

Contributions

[1] Fritsch, V., Da Mota, B., Varoquaux, G., Frouin, V., Loth, E., Poline, J.B., Thirion, B., et al.: Robust group-level inference in neuroimaging genetic studies. In: Pattern Recognition in Neuroimaging (2013)

[2] Fritsch, V., Da Mota, B., Varoquaux, G., Frouin, V., Loth, E., Poline, J.B., Thirion, B., et al.: Robust regression for large-scale neuroimaging studies. in preparation (2014)

The multi-subject brain imaging datasets used in neuroimaging group studies have a complex statistical structure, including local and long range correlations, non-stationarity of the statistical properties and the presence of artifacts. While small-sample size studies can hardly be proved to deviate from standard hypotheses (such as the normality of the residuals) due to the low degrees of freedom of the statistical model, large-scale studies (e.g. on more than 100 subjects) give a different picture and encourage the practitioner to use finer models to perform statistical inference. In this work, we demonstrate the benefits of robust regression as a tool for analyzing large neuroimaging cohorts. Our first contribution is to design an analytic test based on robust parameters estimates; this procedure makes it possible to forgo permutation testing and thus to perform whole brain analysis in a reasonable time. Then we demonstrate that robust regression yields sensitivity improvements in two real data examples on 392 and 1502 subjects. We finally show that robust regression can be combined with more complex analysis techniques to improve whole-brain tests sensitivity.

5.1 Robust regression in neuroimaging

5.1.1 Linear model and statistical inference

We consider the linear model defined in chapter 1:

$$\boldsymbol{B} = \boldsymbol{X}_2\boldsymbol{\gamma} + \boldsymbol{\epsilon}_2, \tag{5.1}$$

where $\boldsymbol{B}$ (the *target feature*) is a set of n samples of an imaging feature (corresponding to n subjects typically), $\boldsymbol{X}_2$ is a $n \times p$ matrix of p variates describing the same n samples, and $\boldsymbol{\epsilon}_2$ is some noise. Some columns of $\boldsymbol{X}_2$ may correspond to variates of no interest, also called *confounds*, while the other columns correspond to explanatory variates, i.e. variates for the which we want to perform a statistical test in order to know if they have an influence on the target variate $\boldsymbol{B}$. The purpose of linear regression is to estimate the unknown coefficients $\boldsymbol{\gamma}$ and this is generally done using *Ordinary Least Squares (OLS) regression*, which intends to minimize the sum of the squared residuals of the fitted model:

$$\hat{\gamma}_{\text{OLS}} = \underset{\boldsymbol{\gamma}}{\operatorname{argmin}} \|\boldsymbol{B} - \boldsymbol{X}_2\boldsymbol{\gamma}\|^2. \tag{5.2}$$

In neuroimaging, the most famous analysis method is a *voxel-wise inference*, that consists in fitting the model (5.1) independently at each voxel (assuming that the images involved in the study have been registered prior to analysis, an assumption that we systematically rely on in this work) by using (5.2). To reduce the number of statistical tests as well as the impact of registration mismatch, analysis methods based on regions of interest have been developed: Signal averages within predefined regions-of-interest are taken as descriptors instead of the voxel-wise signal. If the regions of interest are defined by a brain parcellation, we call the corresponding analysis a *parcel-based analysis*. The model (5.1) is fit at the parcel level or at the voxel level in the same way, i.e. using (5.2). The coefficients $\hat{\gamma}_{\text{OLS}}$ are tested for non-zero significance and one p-value per descriptor is computed, yielding a statistical parametric *map*. Such maps show what regions of the brain the target variates are associated with.

Statistics ensure that $\hat{\gamma}_{\text{OLS}}$ is the maximum likelihood estimate of γ for Gaussian-distributed noise (ϵ_2), but this assumption does not hold for neuroimaging data (see section 1.2.4.3). Corrupted data and observations that deviate from the population-representative pattern potentially introduce some bias in the parameter estimate. We call these unusual observations *outliers*. Formally, outliers have large residual values that contribute to increasing the OLS criterion presented in equation 5.2, resulting in a poor estimation of γ. But as we mentioned in chapter 1, outliers alone are not responsible for deviations from the models used in practice. In fact, it is very likely that high-dimensional neuroimaging data have a complex structure that is unknown, and that the models try to approximate. That said, we consider that *presence of outliers* and *mere deviation from the model* are two phenomena that are hardly distinguishable in practice. We therefore consider that valid and reproducible statistical inference can be performed in the presence of any kind of deviation thanks to robust statistical procedures. This chapter demonstrates this statement and shows that robust procedures yield more sensitivity in the analysis.

5.1.2 Large cohorts and the need for robust tools

If standard hypotheses about the statistical structure of the data cannot be easily rejected and do not require to be fixed when 10 to 20 subjects are included in a neuroimaging experiment, significant departure may be observable when more than 100 subjects are considered. Consequently, we can expect a much better model of the data, hence some gains in sensitivity if we can use a model that relaxes these common hypotheses. This use case is gaining some importance cohorts that are now emerging (ADNI [10], IMAGEN [20], Human Connectome [22] cohorts). This implies getting rid of the standard, caricatural assumptions like Gaussian-distributed data, or homoscedastic noise (see section 1.2.4.3. One precursive step towards that direction was proposed by Wager [23]: Going back to the simplest analysis scheme, the massively univariate voxel-wise inference, Wager suggested to replace the standard ordinary least squares regression by a robust regression (Huber regression [9]), which has the advantage of *(i)* relying on weak structural assumptions (symmetric, unimodal data) and *(ii)* being robust to outliers. Wager's work successfully showed sensitivity improvements for both inter- and intra-subject analyzes, as well as stability regarding the presence of outliers. But this work was limited to the consideration of small groups of subjects (< 20) and only the "outlier-resistant" property of the method seems to have had an impact on the community [17, 15, 14, 11, 1].

5.1.3 Robust regression algorithms

5.1.3.1 A short review

Many robust regression settings have been proposed in the statistical literature. *Least Absolute Deviation (LAD)* regression (or ℓ_1 regression) [4] minimizes the sum of the absolute value of the model residuals. ℓ_1 regression estimate is hard to compute in practice and it can break down in the presence of atypical values in the model design matrix (*leverage points*) [19]. Any regression estimate that minimizes the sum of a given function of the residuals without rescaling them by their variance suffer from the same problem [9]. Indeed, the residual value of an observation is proportional to the absolute deviation of the observation to the data mean. As a result, leverage points attract the regression hyperplane. The *repeated median algorithm* [21] is a regression algorithm that targets a high level of outlier resistance, namely up to 50% amount of contamination (a property known as *high-breakdown point*). It is computationally costly and the resulting estimate is not affinely equivariant. The *Least Median of Squares (LMS)* [7] and *Least Trimmed Squares (LTS)* [18] estimates also have a high breakdown point but can only be computed with algorithms that only provide a local optimum. The efficiency in uncontaminated models remains poor. All of the above-mentioned regression estimates are difficult to apply in the context of high-dimensional neuroimaging problems. *Support Vector Regression (SVR)* can use kernel representations of the data to deal with their dimensionality, but lacks a direct conversion of the estimated regression coefficients into p-values. Huber's regression (presented in section 5.1.4) is the most convenient regression criterion to date, as it offers a good compromise between interpretability, computational cost, and robust (including robust to the statistical structure of the data and outlier resistance). We present SVR and LTS in the next sections because we think they are good examples of robust regression algorithms: LTS illustrates the concept of high breakdown point while SVR is an example of robust non-parametric algorithm.

5.1.3.2 Least Trimmed Squares regression

Some alternative regression criteria target a *breakdown point* of 50 %, i.e. there goal is to ensure a correct estimation of the regression hyperplane under amounts of contamination going up to 50 %. While this robustness property is crucial regarding applications such as outlier detection or analysis of a high-dimensional multimodal cohort, the use of high breakdown point regression criteria should be limited to diagnosis purpose and is inefficient. Least Trimmed Squares (LTS) regression [18] is the state-of-the-art high breakdown point robust regression:

$$\sum_{i=1}^{h} (r^2)_{i:n}, \tag{5.3}$$

where $(r^2)_{i:n} = \left((b_i - \sum x_{2_{i\,j}} \gamma_j)^2\right)_{i:n}$ is the i^{th} ordered squared residual and $1 \leq h \leq \frac{n}{2}$ sets up the breakdown point of the corresponding regression estimate. The p-values associated with the statistical test on the estimated model coefficients are obtained by a permutation test, which is another drawback of the method.

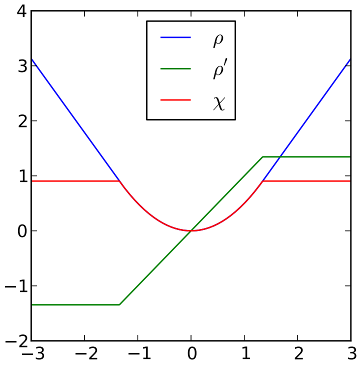


Figure 5.1: **Huber's ρ function.** ρ is a parabola that is continued by a line on its both ends. The square part of the function smoothly downweights the influence of the observations as they depart from the regression hyperplane, yielding a fit (and a testing framework) that accommodates to the data distribution. The linear part prevents the model to be dragged by strong outliers as their influence is bounded (see the flat part of ρ'). $\chi : x \mapsto x\rho'(x) - \rho(x)$, is useful in algorithm 3.

5.1.3.3 Support Vector Regression

Support Vector Regression (SVR) is a regression algorithm that is derived from the *Support Vector Machine (SVM)* classification algorithm already introduced in chapter 2, section 2.3.4. It can be written as a convex minimization problem and solved in the same fashion than the well-known SVM. Notably, not all the observations are used to define the regression hyperplane, which makes the algorithm robust to deviant observations. However, to the best of our knowledge, no statistical test on the estimated model coefficients of SVR regression has been derived so far, hence we need to resort to costly permutation tests.

5.1.4 Huber's robust regression

5.1.4.1 Formulation

As the influence of outliers is accentuated by the square function in equation 5.2, Huber [9] proposed to replace it by a function ρ that dampens the influence of the outliers:

$$\hat{\gamma}_{\mathrm{RLM}} = \underset{\gamma}{\operatorname{argmin}} \sum_{i=1}^{n} \rho\left(\frac{b_i - \sum_{j=1}^{p} x_{2_{i\,j}}\gamma_j}{\sigma}\right). \tag{5.4}$$

The dampening is non-linear and thus offers an interesting trade-off between statistical efficiency at the Normal model and outlier resistance. σ is the standard deviation of the residuals, and acts as a scaling factor to tune the non linearity induced by ρ. A standard choice for ρ is

$$\rho(x) = \begin{cases} \frac{1}{2}x^2, & \text{if } |x| < c, \\ c|x| - \frac{1}{2}c^2, & \text{if } |x| \geq c, \end{cases} \tag{5.5}$$

with $c = 1.345$ for 95% asymptotic efficiency on the standard normal distribution. In this work, we only use this definition of ρ.

Fitting a linear model with the criterion (5.4) is equivalent to downweighting the observations according to their residual value with respect to the *true model*. Thus, beyond outlier resistance, a robust regression criterion ensures that the fit does not depend on the data in the tails of the distribution. This is also the

Algorithm 3 Iteratively Reweighted Least Squares

```
Require: $\boldsymbol{X}_2, \boldsymbol{b}, \rho$.
Ensure: $\epsilon = 10^{-8}$, $\boldsymbol{W}_{\text{old}} = \infty$, $h(p)$ a normalization factor.
  define function $\chi : x \mapsto x\rho'(x) - \rho(x)$
  $\boldsymbol{\gamma} \leftarrow (\boldsymbol{X}_2^\top \boldsymbol{X}_2)^{-1} \boldsymbol{X}_2^\top \boldsymbol{b}$
  $\sigma^2 \leftarrow \text{Var}[b_i - \sum_{j=1}^p x_{2_{i\,j}} \gamma_j]$
  $\sigma^2 \leftarrow \frac{1}{n\,h(p)} \sum_{i=1}^n \chi\left(\frac{b_i - \sum_{j=1}^p x_{2_{i\,j}} \gamma_j}{\sigma}\right) \sigma^2$
  $W \leftarrow \rho'\left(\frac{b_i - \sum_{j=1}^p x_{2_{i\,j}} \gamma_j}{\sigma}\right)$
  while $\|\boldsymbol{W}_{\text{old}} - \boldsymbol{W}\|_\infty > \epsilon$ do
    $\boldsymbol{W}_{\text{old}} \leftarrow \boldsymbol{W}$
    $\sigma^2 \leftarrow \frac{1}{n\,h(p)} \sum_{i=1}^n \chi\left(\frac{b_i - \sum_{j=1}^p x_{2_{i\,j}} \gamma_j}{\sigma}\right) \sigma^2$ (scale step)
    $\boldsymbol{W} \leftarrow \left(\frac{\rho'(b_i - \sum_{j=1}^p x_{2_{i\,j}} \gamma_j / \sigma)}{b_i - \sum_{j=1}^p x_{2_{i\,j}} \gamma_j / \sigma}\right)_{i \in \{1..n\}}$ (reweighting)
    Let $\hat{\boldsymbol{\tau}}$ be the solution of $\boldsymbol{X}_2^\top \boldsymbol{W} \boldsymbol{X}_2 \hat{\boldsymbol{\tau}} = \boldsymbol{X}_2^\top \boldsymbol{W} \boldsymbol{b}$
    $\boldsymbol{\gamma} \leftarrow \boldsymbol{\gamma} + \hat{\boldsymbol{\tau}}$
  end while
  $\text{cov}(\hat{\boldsymbol{\gamma}}) = K \frac{[1/(n-p)] \sum_{i=1}^n \rho''(b_i - \sum_{j=1}^p x_{2_{i\,j}} \gamma_j)^2}{(1/n) \sum_{i=1}^n \rho''(y_i - \sum_{j=1}^p x_{2_{i\,j}} \gamma_j)} \boldsymbol{W}^{-1}$,
  $K = 1 + \frac{p}{n} \frac{\text{Var}[\psi'(b_i - \sum_{j=1}^p x_{2_{i\,j}} \gamma_j)]}{(\text{E}[\psi'(b_i - \sum_{j=1}^p x_{2_{i\,j}} \gamma_j)])^2}$
```

case for the p-values that are derived from the associated robust test, which is presented in the sequel.

5.1.4.2 Algorithm

In practice, σ is not known and has to be estimated while the model is being fit, yielding a joint estimation challenge (i.e. one has to estimate $\hat{\gamma}$ and $\hat{\sigma}$ at the same time. The *Iteratively Reweighted Least Squares (IRLS)*, presented as Algorithm 3, is often used to solve this problem. One important step to ensure the convergence of the algorithm is the *scale step*, that correspond to the update of $\hat{\sigma}$. In the applied literature, a robust estimate of the residuals' standard deviation is taken so as to update $\hat{\sigma}$ (e.g. in [16, 3]), but Huber raises the point that no theoretical proof of the algorithm convergence has been given in that setting and suggests a more complex update that guarantees the algorithm convergence when a convex weighting function is used [9]. We use a Python implementation of robust regression available in the statsmodels [1] library, which we optimized for our application. The implementation strictly follows Huber's definition of the scale update step.

5.1.4.3 Significance testing

Huber [9] proposed to adapt the standard F-test to robust regression by considering a robust unbiased estimate of cov($\hat{\gamma}$) (given at the end of Algorithm 3). Such an analytic testing procedure is however crucial to us as the IRLS algorithm costs too much to be considered with permutation testing. We dedicate

[1] http://statsmodels.sourceforge.net

the next section to a validation of this testing procedure as it has never been done to our knowledge.

5.2 Validation of the statistical procedure

5.2.1 Methods

5.2.1.1 Generative model

Before applying robust regression (RLM) to real data, we carry out an empirical validation of the testing procedure associated with it, and compare it with standard ordinary least squares regression (OLS). We use the following model to generate n observations $\{y_1, \dots, y_n\}$:

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{a}_q\boldsymbol{\epsilon} + \alpha(\mathbf{1}_n - \boldsymbol{a}_q)\boldsymbol{\epsilon}, \tag{5.6}$$

where $\boldsymbol{X}$ is a random $(n \times r)$ design matrix, $\boldsymbol{\beta}$ is the $(r \times p)$ matrix of the model coefficients, $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathrm{Id}_n)$ models a Gaussian noise, $\boldsymbol{a}_q$ is a n-dimensional vector with coordinates drawn from a Bernoulli distribution $\mathcal{B}(1-q)$, and $\alpha > 1$ is a scalar. q is thus the expected proportion of outliers in the generated dataset, and α is a parameter that controls how strongly the outliers deviate from the regular model. We set α to 5.

5.2.1.2 Control of the type I error rate

To verify that we can control the rate of type I error under the null hypothesis, we set a column of $\boldsymbol{\beta}$ to 0 in model (5.6), say column j. For various values of q, we fit both a standard and a robust linear model to a dataset generated according to the obtained model, respectively yielding matrices of estimated model coefficients $\hat{\boldsymbol{\beta}}_{\mathrm{OLS}}$ and $\hat{\boldsymbol{\beta}}_{\mathrm{RLM}}$. Finally, we test the parameters corresponding to the j-th column of $\hat{\boldsymbol{\beta}}_{\mathrm{OLS}}$ and $\hat{\boldsymbol{\beta}}_{\mathrm{RLM}}$, that are associated with a null true effect. The proportion of null rejection should thus be equal to or less than the nominal test p-value. For instance, providing the testing procedures are unbiased, it should happen in 5% cases that OLS/RLM reports a significant effect at $P < 0.05$ uncorrected. We are also interested in lower thresholds ($< 10^{-3}$) as we intend to work with corrected p-values.

5.2.1.3 Statistical power (type II error rate)

We show that in the presence of outliers, the statistical power of the robust test is higher than that of the statistical power achieved by an F-test subsequent to an OLS fit. The simulation framework is the same than in the previous experiment, except that we do not set any column of $\boldsymbol{\beta}$ to 0, so we perform tests on a variable that is known to have an effect. We construct Receiver Operating Characteristic (ROC) curves [8] for RLM and OLS according to the true/false acceptance/rejection of the null hypothesis (i.e. "no correlation exists between the tested variable(s) and the fMRI signal values").

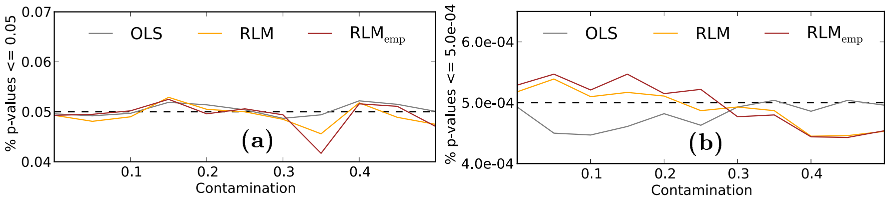


Figure 5.2: **Proportion of type I errors for OLS and RLM (for two different algorithms).** RLM$_{\text{emp}}$ corresponds to robust regression as computed with the IRLS algorithm for which the scale is estimated with a median absolute deviation. **(a)** Nominal rate at $P < 0.05$ uncorrected, estimated on 10,000 independent tests performed under a null hypothesis. **(b)** Nominal rate at $P < 5.0e^{-4}$ uncorrected, estimated on 1,000,000 independent tests performed under a null hypothesis. The experimental design involves 300 observations ($n = 400$), 1 tested variable and 10 confounding variables. The error rate corresponds to the expected nominal rate with all methods.

5.2.2 Results

5.2.2.1 Control of the type I error rate

The control of type I error obtained with the testing procedures associated with OLS and RLM is exact, as shown in Figure 5.2. This result hold for all number of observations involved in the simulation (n). We also obtained the same performance when confounding variables were included, and with multivariate test (i.e. several columns of the design matrix were associated with null coefficients and tested for a joint effect).

5.2.2.2 Control of the type II error rate

The ROC curves presented in Figure 5.3 illustrate the ability of the testing procedures associated with OLS (resp. RLM) to detect a significantly non-null effect in the presence of outliers. The latter potentially mislead OLS while RLM keeps a good sensitivity, as defined by the trade-off between correct and uncorrect null-hypothesis rejections. The curves may drop as more confounding variables are included in the experimental design, but the relative performance of both regression framework is preserved. Conversely, testing several variables at a time increases the performance of the methods simultaneously. We also give ROC curves that show the performance of Least Trimmed Squares (LTS) regression in Figure 5.4. As expected, this algorithm behaves better than RLM only in the case of an extremely strong contamination, namely 50 %, which does not correspond to a realistic case in neuroimaging. Moreover, LTS consistency at the regular model is poor and we will therefore not investigate further this regression algorithm.

5.3 Application to neuroimaging data

5.3.1 Synthetic neuroimaging data

We measure the accuracy improvement yielded by robust regression by first applying it on synthetic neuroimaging data. We use a generative model that makes it possible to compare the analysis results to a ground truth. The parameters of the simulation are chosen to yield data that model real neuroimaging data well.

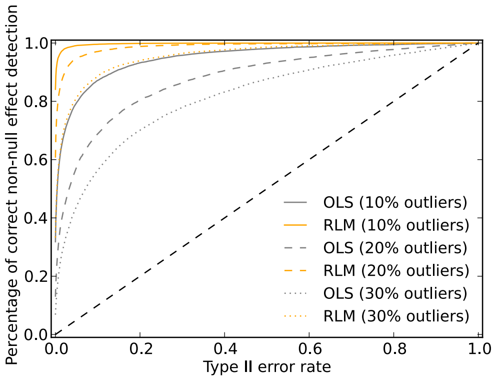


Figure 5.3: **Accuracy of standard and robust regression algorithms under various amounts of contamination.** Robust regression and its associated testing procedure always achieve a better compromise between type I and type II errors.

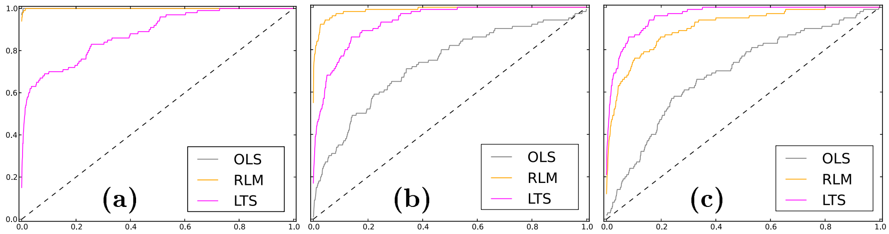


Figure 5.4: **Accuracy of Least Trimmed Squares (LTS) regression as compared to OLS and RLM. (a)** no contamination, **(b)** 30 % contamination, **(c)** 50 % contamination. LTS only outperforms for an amount of contamination that is equal to its breakdown point, i.e. 50 % contamination (c). Such a case is not relevant in practice and one would therefore prefer RLM as a robust tool. LTS consistency at the regular model is poor (a).

Data generation. We simulate fMRI contrast images as volumes of shape $40 \times 40 \times 40$ voxels. Each contrast image contains a simulated $4 \times 4 \times 4$ activation patch at a given location, with a spatial jitter following a three-dimensional $\mathcal{N}(\mathbf{0}, \boldsymbol{I}_3)$ distribution (coordinates of the jitter are rounded to the nearest integers). The strength of the activation is set so that the signal to noise ratio (SNR) peaks at 2 in the most associated voxel. The background noise is drawn from a $\mathcal{N}(0, 1)$ distribution, Gaussian-smoothed at σ_{noise} isotropic and normalized by its global empirical standard deviation. After superimposing noise and signal images, we optionally smooth at $\sigma_{\text{post}} = 2.12$ voxels isotropic, corresponding to a 5 voxels Full Width at Half Maximum (FWHM). Voxels with a probability above 0.1 to be active in a large sample test are considered as part of the ground truth. Ten subsamples (or groups) of 100 images are then generated and analyzed. Each subsample was then contaminated by 15 % outliers, i.e. in each group, we replace 15 observations by images for which the activation is located at a random position in the image. All the others parameters stay the same as for the generation of the 85 % valid observations.

Experiment. We perform two voxel-intensity based analyzes on each subgroups: the first using standard regression (OLS) and the second using robust regression (RLM). For a given type of analysis, the output statistical maps corresponding to the ten groups are pooled together so as to obtain one single statistical map that is thresholded according to a varying threshold. The various threshold levels are compared to the ground truth described above and we summarize the results in a ROC curve. We present these curves for the

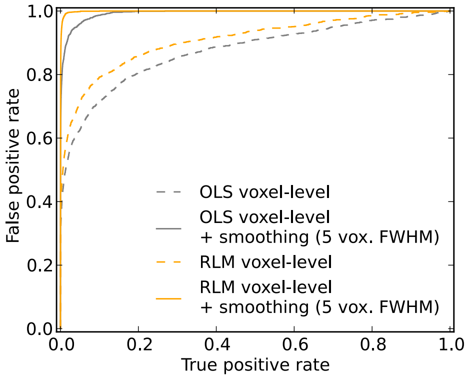


Figure 5.5: **Sensitivity/specificity trade-off for standard (OLS) and robust (RLM) regression regarding voxel-level group analysis in the presence of outliers.** Robust regression achieves more sensitivity than standard regression, because is accommodates to deviation from the model assumptions. This is likely to happen in the context of neuroimaging. The performance of the two algorithms is almost equivalent when no outlier is present (confounded curves, not shown), which confirms the good consistency of robust regression at the regular model.

experiment performed with and without smoothing.

Results. When outliers are present in the data, a deviation from the model assumptions is likely to be observed. Robust regression can accommodate these potential deviations and therefore yields more sensitive results than OLS regression, as shown in Figure 5.5. Robust regression is yet consistent at the regular model because its performance is similar to that of OLS regression (not shown, since the curves are almost confunded). This agrees with the theory. This simulation suggests that robust regression should be systematically used in neuroimaging, because the quality of the data can hardly be checked, especially when large and complex cohorts are considered.

5.3.2 Real data

5.3.2.1 A gene-neuroimaging study

We apply robust regression to a gene-neuroimaging study examining gene × environment (G×E) interaction effects on fMRI BOLD activity to angry faces ([6]) in a large sample of $n = 392$ subjects taken from the multi-centre study IMAGEN [20]. Severe outliers due to motion or deformation artifacts as well as those detected using a multivariate outlier procedure covering the whole brain, were removed. All of the 392 available observations are thus considered as correct upon manual quality check. The example illustrates the common hypothesis that genetic effects on brain function (and behavior) may often only be detected under certain environmental conditions [2]. Consequently, compared to main effects models, tests of the G×E interaction term render the need for sensitive neuroimaging 'endophenotypes' all the more pertinent. The model covariates were genotype, environmental risk (number of stressful life events, SLE), sex, puberty development, study center and handedness. As in many gene-neuroimaging studies we employed an unbalanced design, comparing 65 minor allele carriers of a common *Single Nucleotide Polymorphism (SNP)* in the oxytocin receptor gene (rs2268494) to 327 major-allele homozygous. Our aim is to compare the ability of standard and robust regression to uncover interesting effects at a fixed specificity level, i.e. the sensitivity of both methods. We construct 200 brain parcellations (from 100 to 2000 parcels by increment of 100, and 10 brain parcellations per number of parcels) using Ward's clustering [24] on the contrast images of 300 bootstrapped subjects amongst the available 392.

Each parcellation is used to convert the contrast images of the 392 subjects into neuroimaging features by averaging the voxels signal within each parcel. For each set of features, we conduct two analyzes: one with OLS and the other with RLM. We report the number of significant effects found ($P < 0.1$ Bonferroni corrected) divided by the number of parcels, which gives an estimate of the method sensitivity under a given type I error control. We also perform a voxel-wise Bonferroni-corrected analysis, using respectively standard and robust regression.

5.3.2.2 Embedding robust regression into complex methods

Robust regression can straightforwardly be combined with more advanced analysis methods (TFCE or RPBI, see chapter 2). We demonstrate that such combinations actually yield more sensitivity than the standard, non-robust version of the methods. We applied RPBI to the gene-neuroimaging study described in the previous section. We generated 100 random parcellations with 1000 parcels each, following the description given above. RPBI was performed twice: the first time with a standard regression algorithm (RPBI_{OLS}), the second time with a robust regression algorithm (RPBI_{RLM}).

As another example of the importance of using robust regression, contrast images of 1364 subjects from the Imagen study were regressed against an impulsivity factor variable using Randomized Parcellation Based Inference. Covariables such as handedness, acquisition center, sexe and age were included. The task was a stop-signal task and the contrast corresponds to the *[stop success - stop fail]* condition [12], where the *stop success* (resp. *stop fail*) condition corresponds to the event where the subject managed (resp. failed) to inhibit its response when asked to do so (i.e. not pressing a button when the first intention was to press it but a stop signal occured). As above, 300 bootstrapped subjects were used to build 100 different wards parcellations of 1000 parcels each.

5.3.2.3 Results

Gene-neuroimaging study Figure 5.6 shows that robust regression always reports more significant activations than standard regression, whatever the number of parcels considered to reduce the data dimension. As the proportion of reported significant activations stabilizes as soon as 500 parcels are used, Figure 5.6 moreover suggests the exact number of parcels does not have a strong impact if enough are considered. This justifies to same extent our choice of using parcellation-based analyzes with a fixed number (1000) of parcels.

Regarding voxel-level analyzes performed with OLS (resp. RLM), a significant association was reported in only one (resp. four) voxel(s) located in the right (resp. left and right) ventral striatum. The results are relevant to the study, as the ventral striatum plays a key role in the processing of positive and negative reward signals, including anger expressions [13]. Robust regression improves the original findings and interestingly uncover the symmetric activation as well.

Five brain locations were reported as significantly associated with a non-null effect when applying RPBI_{RLM} to the gene-neuroimaging study of this real data application. Only three of them were reported by RPBI_{OLS}, as shown in

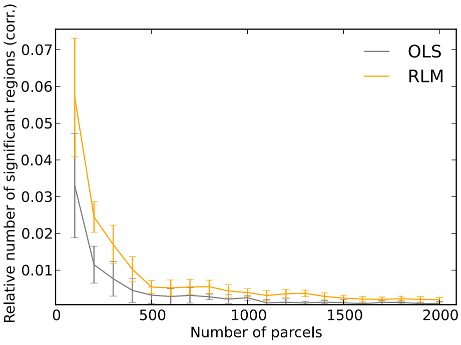


Figure 5.6: **Percentage of parcels significantly associated with a non-null effect of the *SNP* × *Stressful Life Events (SLE)* interaction**, according to standard and robust regression. The latter always uncovers more significant associations.

Figure 5.7. The activation in the left amygdala ($z = 7$ mm slice) is larger and more significant according to $\mathrm{RPBI}_{\mathrm{RLM}}$.

Neuroimaging study with behavioral features As shown in Figure 5.8, a standard regression framework reports a significant effect of the impulsivity factor within the right hypothalamus ($P < 0.1$ Bonferroni corrected) while a robust algorithm does not report anything. Further investigation on the data shows that one subject is an outlier and influences OLS regression. As an illustration, we focused on a single parcellation, and particularly on a parcel which overlaps with the hypothalamus location. Figure 5.9 represents the corresponding data in a scatter plot, and notably one subject having both a very low average signal value and a high impulsivity score. We presented OLS and RLM regression lines within the same scatter plot. The shift created by the outlier observation can be observed for each individual parcellation, in the parcel that matches the right hypothalamus at best. Renewing the experience without the outlier results in the disappearance of the significant effect observed in the hypothalamus with OLS.

5.4 Discussion

Huber's robust regression can advantageously replace Ordinary Least Squares regression in the context of a neuroimaging study. After ensuring that the testing procedure associated with robust regression comes with an exact control on the accepted errors, we showed that resistance to outliers yields a better stability and, in turn, improves the sensitivity of the analyzes. These results are confirmed on real neuroimaging data.

Robust regression and outlier detection Robust statistical tools are mainly used for their outlier-resistance properties. Equivalently, some studies do not use robust tools because the data are previously quality checked and should not contain any outlier. This step is relevant and we advocate performing it in order to remove gross outliers and make the data more homogeneous. Outlier detection algorithms [5] or Least Trimmed Squares regression [18] can be useful for an automated diagnosis. However, the poor performance of LTS stresses the fact that a robust fit does not simply boils down to discarding potential outliers. Deviation from normality is actually a widespread phenomenon, that

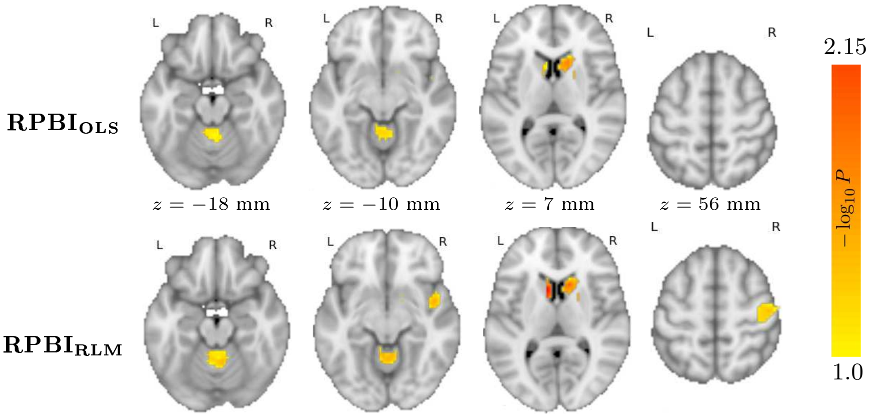


Figure 5.7: **Voxel-level FWER-corrected p-values maps given by RPBI$_{\text{OLS}}$ and RPBI$_{\text{RLM}}$ on our gene-neuroimaging study** (represented as negative $\log_{10}$ p-values). Four brain regions are associated with a significant non-null effect according to the robust version of RPBI, while only two of them are reported by standard RPBI. The significant associations observed in the left and right ventral striatum (third column, $z = 7$ mm) are particularly relevant to the study, as the ventral striatum plays a key role in the processing of positive and negative reward signals, including anger expressions.

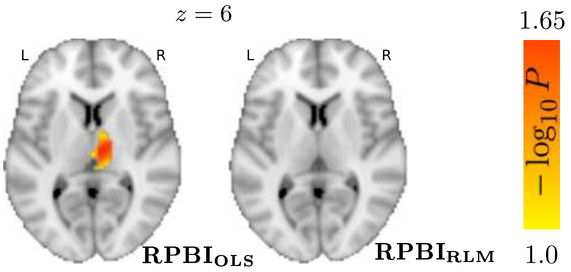


Figure 5.8: **RPBI analysis with standard regression (OLS) and robust regression (RLM).** RPBI$_{\text{OLS}}$ (left) reports a significant effect for a group of voxel within the right hypothalamus while RPBI$_{\text{RLM}}$ (right) does not report any significant effect. The difference is explained by the presence of a gross outlier (details given in Figure 5.9).

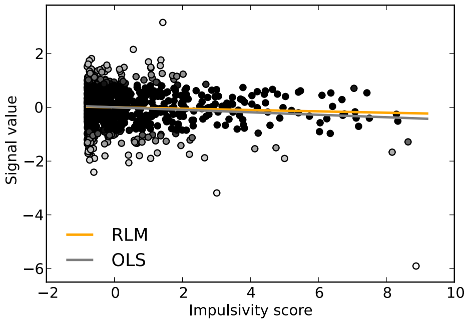


Figure 5.9: **Relationship between the mean signal within a parcel centered at the hypothalamus and the impulsivity factor (1364 subjects).** Covariables effect has been removed from the two values. Regression lines have been drawn on top of the data for standard (OLS) and robust regression (RLM).

is much more general than the contamination by outliers values, and that needs to be taken into account in inferential procedures. More subtle deviations from the model assumptions cannot be systematically detected, and robust tools still turn out to be useful whatever the quality of input data, as our real data experiments demonstrate. Indeed, more sensitivity is achieved by robust regression in our neuroimaging genetic experiment, while we control for the specificity. The experiment with behavioral factors also shows differences between a robust and a non-robust fit, and further investigation reveals the presence of a multivariate outlier that could not be detected with a mere quality check (the impulsivity score of the main outlier is not an extreme value, but its conjunction with the imaging phenotypes makes it an outlier).

Computation Unlike Support Vector Regression and other alternative robust regression algorithms, Huber's robust regression has the advantage that an analytic procedure exist to test the estimated model coefficients. This reduces the running time of the algorithm regarding neuroimaging applications, where the ultimate goal is to find significant associations between experimental variables and brain imaging phenotypes. We optimized the implementation of robust regression so that we can perform voxel-level analyzes of a cohort of hundreds of subjects in a few minutes on a desktop computer. A robust fit is yet 10 to 100 times slower than an OLS fit, which prevents RLM to be routinely used with permutation testing, including in the scope of a more complex statistic such as $\mathrm{RPBI}_{\mathrm{RLM}}$ or robust cluster-size inference. We use a cluster of computers to run the analyzes with $\mathrm{RPBI}_{\mathrm{RLM}}$.

Embedded robust regression Our experiments demonstrate that robust regression can successfully be combined with state-of-the-art neuroimaging analysis method for an improved sensitivity. This approach, albeit more expensive than more traditional inference schemes, is promising as the number of large cohorts and neuroimaging genetic studies are growing. Those datasets have a complex statistical structure and specific statistical procedures are therefore required to address this challenging problem. We focused on Randomized Parcellation Based Inference because it outperforms the others state-of-the-art methods, but robust regression would be embedded in cluster-size inference of TFCE as well.

References

[1] Atlas, L.Y., Bolger, N., Lindquist, M.A., Wager, T.D.: Brain mediators of predictive cue effects on perceived pain. The Journal of Neuroscience 30(39), 12964–12977 (2010)

[2] Caspi, A., Hariri, A.R., Holmes, A., Uher, R., Moffitt, T.E.: Genetic sensitivity to the environment: the case of the serotonin transporter gene and its implications for studying complex diseases and traits. The American journal of psychiatry 167(5), 509 (2010)

[3] Chatterjee, S., Mächler, M.: Robust regression: a weighted least squares approach. Communications in Statistics-Theory and Methods 26(6), 1381–1394 (1997)

[4] Dodge, Y.: on Statistical data analysis based on the L1-norm and related methods. Elsevier Science Inc. (1987)

[5] Fritsch, V., Varoquaux, G., Thyreau, B., Poline, J.B., Thirion, B.: Detecting outliers in high-dimensional neuroimaging datasets with robust covariance estimators. Med Image Anal 16(7), 1359 – 1370 (2012)

[6] Grosbras, M.H., Paus, T.: Brain networks involved in viewing angry hands or faces. Cereb Cortex 16(8), 1087–1096 (Aug 2006)

[7] Hampel, F.R.: Beyond location parameters: Robust concepts and methods. Bulletin of the International statistical Institute 46(1), 375–382 (1975)

[8] Hanley, J.A., McNeil, B.J.: The meaning and use of the area under a receiver operating (ROC) curve characteristic. Radiology 143(1), 29–36 (1982)

[9] Huber, P.J.: Robust Statistics, chap. 7, p. 149. John Wiley & Sons, Inc. (2005)

[10] Jack, C.R., Bernstein, M.A., Fox, N.C., Thompson, P., Alexander, G., Harvey, D., Borowski, B., Britson, P.J., L Whitwell, J., Ward, C., et al.: The Alzheimer's disease neuroimaging initiative (ADNI): MRI methods. Journal of Magnetic Resonance Imaging 27(4), 685–691 (2008)

[11] Kober, H., Mende-Siedlecki, P., Kross, E.F., Weber, J., Mischel, W., Hart, C.L., Ochsner, K.N.: Prefrontal-striatal pathway underlies cognitive regulation of craving. Proceedings of the National Academy of Sciences 107(33), 14811–14816 (2010)

[12] Logan, G.D.: On the ability to inhibit thought and action: A users' guide to the stop signal paradigm. Psychological Review 91(3), 295–327 (1994)

[13] Loth, E., Poline, J.B., Thyreau, B., Jia, T., Tao, C., Lourdusamy, A., Stacey, D., Cattrell, A., Desriviéres, S., Ruggeri, B., Fritsch, V., Banaschewski, T., Barker, G.J., Bokde, A.L., Büchel, C., Carvalho, F.M., Conrod, P.J., Fauth-Buehler, M., Flor, H., Gallinat, J., Garavan, H., Heinz, A., Bruehl, R., Lawrence, C., Mann, K., Martinot, J.L., Nees, F., Paus, T., Pausova, Z., Poustka, L., Rietschel, M., Smolka, M., Struve, M., Feng, J., Schumann, G., the IMAGEN Consortium: Oxytocin receptor genotype modulates ventral striatal activity to social cues and response to stressful life events. Biological Psychiatry in press (2013)

[14] McRae, K., Hughes, B., Chopra, S., Gabrieli, J.D., Gross, J.J., Ochsner, K.N.: The neural bases of distraction and reappraisal. Journal of Cognitive Neuroscience 22(2), 248–262 (2010)

[15] Ochsner, K.N., Hughes, B., Robertson, E.R., Cooper, J.C., Gabrieli, J.D.: Neural systems supporting the control of affective and cognitive conflicts. Journal of Cognitive Neuroscience 21(9), 1841–1854 (2009)

[16] Phillips, G.R., Eyring, E.M.: Comparison of conventional and robust regression in analysis of chemical data. Analytical Chemistry 55(7), 1134–1138 (1983)

[17] Poldrack, R.A.: Region of interest analysis for fMRI. Social cognitive and affective neuroscience 2(1), 67–70 (2007)

[18] Rousseeuw, P.J.: Least median of squares regression. Journal of the American statistical association 79(388), 871–880 (1984)

[19] Rousseeuw, P.J., Leroy, A.M.: Robust regression and outlier detection, vol. 589. Wiley. com (2005)

[20] Schumann, G., Loth, E., Banaschewski, T., Barbot, A., Barker, G., Büchel, C., Conrod, P.J., Dalley, J.W., Flor, H., Gallinat, J., Garavan, H., Heinz, A., Itterman, B., Lathrop, M., Mallik, C., Mann, K., Martinot, J.L., Paus, T., Poline, J.B., Robbins, T.W., Rietschel, M., Reed, L., Smolka, M., Spanagel, R., Speiser, C., Stephens, D.N., Ströhle, A., Struve, M., IMAGEN consortium: The

IMAGEN study: reinforcement-related behaviour in normal brain function and psychopathology. Mol Psychiatry 15(12), 1128–1139 (Dec 2010)

[21] Siegel, A.F.: Robust regression using repeated medians. Biometrika 69(1), 242–244 (1982)

[22] Van Essen, D.C., Ugurbil, K., Auerbach, E., Barch, D., Behrens, T., Bucholz, R., Chang, A., Chen, L., Corbetta, M., Curtiss, S., et al.: The human connectome project: a data acquisition perspective. Neuroimage 62(4), 2222–2231 (2012)

[23] Wager, T.D., Keller, M.C., Lacey, S.C., Jonides, J.: Increased sensitivity in neuroimaging analyses using robust regression. NeuroImage 26, 99 (2005)

[24] Ward, J.: Hierarchical grouping to optimize an objective function. Journal of the American Statistical Association 58(301), 236–244 (1963)

Conclusion

In this thesis, we have investigated how to improve neuroimaging studies with the use of robust statistical methods. Our main aim is to ascertain that the results of statistical inference procedures are *(i)* robust to the presence of abnormal observations; *(ii)* reproducible across studies; *(iii)* do not rely crucially on model assumptions. Although these requirements are somewhat related, they have to be considered in turn independently before global solutions can be proposed. We therefore develop our contributions according to three main directions:

Outlier detection. First, we have considered the automatic detection of abnormal data that potentially mislead the analysis procedures, the so-called *outliers*. In our experiment, we considered n observations that corresponds to the brain images of n subjects, from which we reduce the dimension by considering local signal averages in p predefined parcels covering the whole brain. We adapted a robust covariance estimator, the *Minimum Covariance Determinant (MCD)* to high-dimensional settings ($\frac{p}{n} > 0.2$) by regularizing it. Assuming that the data are Gaussian-distributed, we showed that with such a covariance estimator, it is possible to perform statistically-controlled multivariate outlier detection upon consideration of the Mahalanobis distances of the observations.

Amongst various types of regularization, we observed that *random projections* (yielding the *RMCD-RP* estimator) are most suited to our applications: this approach has the highest outlier detection accuracy when confronted to various types of outliers, it is robust to deviations from the Normal model, it keeps a high accuracy even when $p > n$. However, we found ℓ_2 regularization useful as the corresponding covariance estimate (*RMCD-*ℓ_2) is faster to compute and to set up (e.g. the choice of the regularization amount can be done with a closed form formula). Non-parametric outlier detection procedures such as *One-Class SVM* or *Local Component Analysis (LCA)* also have a high accuracy but they do not come with a statistical control on subjects inclusion. Note that the most powerful of them, LCA, can be used to build a representation of the dataset that provides insight about its statistical structure.

Improving the reproducibility of neuroimaging studies. The number of subjects included in neuroimaging studies is relatively low as compared to the large number of image descriptors (more than 40,000 voxels) and potential external variables (hundreds of behavioral variates, more than 100,000 genetics variants). Thus, the datasets used in neuroimaging are poor representative of the real data structure and with state-of-the-art methods, the results of the same analysis performed on two different samples from the same population may vary a lot. The phenomenon is even worse across cohorts.

We proposed a new approach, *Randomized Parcellation Based Analysis (RPBI)*, to overcome the limitations of standard methods, in which active voxels are detected according to a consensus on several random parcellations of the brain images, while a permutation test controls the false positive risk. Thus, RPBI stabilizes standard parcel-based analysis to obtain reproducible results, which is a form of statistical robustness (namely robustness regarding the choice of the input sample). Data-driven parcellations are obtained with *Ward's clustering algorithm*. Both on synthetic and real data, this approach shows higher sensitivity, better accuracy and higher reproducibility than state-of-the-art methods. These improvements are especially useful for large-scale studies, such as neuroimaging genetic studies, as groups with uneven cell sizes are compared through complex designs.

Robustness to deviations from the model. Due to the complex statistical structure of neuroimaging data, approximations are widely used to model neuroimaging datasets. For instance, a standard assumption is to assume that the data are Gaussian-distributed. We demonstrated that such assumptions are not verified in practice, especially when complex and unbalanced designs are considered (e.g. neuroimaging genetic studies). Also, outlier detection procedures may not be very accurate, which is a potential source for even more deviation of the data distribution from the assumed model. We therefore need to resort to outlier-resistant statistical procedures for data analysis in order to guarantee that the analysis is not driven out by discrepancies that exist between the real data structure and the model considered by the practitioner.

We have considered robust regression for neuroimaging group studies. We emphasized the analysis of large cohorts (more than 50 subjects) as small-sample size problems were already considered in the literature. We performed a validation of the analytic testing procedure associated with robust regression to verify that it provides a correct control on the type I and II error rates. We then showed on two real data experiments that robust regression yields more sensitivity than state-of-the-art non-robust methods. Importantly, multivariate outlier detection procedures were used as preprocessing prior to analysis. This is a proof that robust procedures are still needed for the analysis of quality-checked data.

A global methodology for robust results in neuroimaging

Our last contribution was to show that robust regression can be combined with random parcellation based inference to obtain even more sensitive results. Increased sensitivity is particularly vital for studies examining brain-behavior relationships or gene-neuroimaging studies. We strongly recommend using RPBI in combination with robust regression, therefore, we are currently working on an efficient implementation of this framework. We also advise to perform a preliminary outlier detection. This can be advantageously done with RMCD-RP, with a precise control on subject exclusion, but we need to mention that the exact number of removed subject plays only a limited role in practice. Gross outliers can easily be detected with very conservative thresholds or with non-parametric outlier detection algorithms such as LCA, while a resistance to weak outliers in ensured by the robust procedure used for the analysis.

An interesting observation is that all outlier detection methods (including descriptive tools such as LCA) suggest that fMRI datasets contain more than 30% outliers. This may be due to the fact that fMRI is not a quantitative method, but only relates *activation patterns* that need to be rescaled across subjects. It would therefore be very sensitive to artifacts. A practical solution could be to systematically work on t- or z-maps instead of contrast maps, although this only partially addresses the problem (i.e. the scaling problem is solved, but the sensitivity to artifacts remains).

Future directions

Each of the three point mentioned in the first paragraph of this conclusion can be pushed further.

Outlier detection would benefit from a visualization tool that helps the practitioner interpreting it. We observed that when $\frac{p}{n} > 1$, *diagonal covariance models* have a good accuracy. They are, in turn, more interpretable and their decision function may be easier to represent in a human-readable way.

Random parcellation based analysis may be improved by considering parcellations with variable number of parcels, or by changing the type of the decision function used in the inner parcel-based analyzes. We plan to use *soft thresholding* as it correspond to a convex transformation and would therefore yields results that are more stable regarding the value of the significance threshold t.

Robust regression suffers from the fact that a recursive algorithm (IRLS) is used to fit it. We want to replace the IRLS algorithm by *optimization algorithms* used in machine learning (e.g. Stochastic Gradient Descent [1]) in order to jointly minimize the coefficient of the robust linear model and the corresponding scale parameter. We are mainly inspired by the recent work of the Sierra Inria team that works on robust principal component analysis and its implementation [4].

We also plan to investigate multivariate robust regression. Penalized regression started recently to be used in genome-wide association studies [3, 2], but not in a robust version.

Software distribution

All the algorithms developed in this thesis were implemented in Python. RPBI and robust regression are packaged into the Parietal team code base, and will soon be released as open-source projects. Regarding outlier detection, the code is also available in the Parietal team code base but still need to be cleaned before distribution because the actual implementation was developed for research purpose and is not optimized. Our work on covariance estimation however resulted in the implementation of the *Minimum Covariance Determinant (MCD)* estimator and *Least Trimmed Squares (LTS)* regression. They are respectively available as a part of the Scikit-learn[2] [5] and Statsmodels[3] Python packages.

References

[1] Bottou, L.: Large-scale machine learning with stochastic gradient descent. In: Proceedings of COMPSTAT'2010, pp. 177–186. Springer (2010)

[2] Hibar, D.P., Kohannim, O., Stein, J.L., Chiang, M.C., Thompson, P.M.: Multilocus genetic analysis of brain images. Frontiers in genetics 2 (2011)

[3] Kohannim, O., Hibar, D.P., Stein, J.L., Jahanshad, N., Jack, C.R., Weiner, M.W., Toga, A.W., Thompson, P.M.: Boosting power to detect genetic associations in imaging using multi-locus, genome-wide scans and ridge regression. In: Biomedical Imaging: From Nano to Macro, 2011 IEEE International Symposium on. pp. 1855–1859. IEEE (2011)

[4] Lacoste-Julien, S., Schmidt, M., Bach, F.: A simpler approach to obtaining an O(1/t) convergence rate for the projected stochastic subgradient method (2012)

[5] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, É.: Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research (Oct 2011)

[2] http://scikit-learn.org.

[3] http://statsmodels.sourceforge.net/.