UNIVERSITY OF PARIS SUD ORSAY

Selection of a Model of Cerebral Activity for fMRI Group Data Analysis

by

Merlin Keller

A thesis submitted in partial fulfillment for the degree of Doctor of Philosophy

in the

Faculty of Science

Department of Statistics

2010

"As followers of natural science we know nothing of any relation between thoughts and the brain, except as a gross correlation in time and space."

Sir Charles Scott Sherrington

Résumé

L'imagerie par résonance magnétique fonctionnelle (IRMf) permet d'acquérir des images tridimensionnelles de l'activité cérébrale d'un sujet soumis à une séquence de stimulations sensorielles. L'analyse statistique des ces données permet de détecter les aires cérébrales actives en réponse aux différentes stimulations.

Lorsque plusieurs sujets ont été recrutés pour une expérience, l'analyse de groupe consiste à généraliser les résultats individuels à la population d'intérêt dont sont issus les sujets. La variabilité morphologique du cerveau humain rend cependant la comparaison des images acquises sur les différents sujets problématique

L'approche usuelle pour contrer cette difficulté consiste à recaler les sujets dans un référentiel commun, puis de comparer les cerveau séparément en chaque point de ce référentiel. Cette étape de recalage n'étant jamais parfaite, il en résulte une incertitude sur la localisation spatiale de chaque sujet.

Nous proposons dans un premier temps d'étendre le modèle classique d'analyse de groupe afin de prendre en compte cette incertitude spatiale. Dans un deuxième temps, nous développons à partir de ce modèle une nouvelle approche de détection d'aires cérébrales actives, basée sur des régions d'intérêt prédéfinies plutôt que sur les procédures de seuillage couramment utilisées.

UNIVERSITY OF PARIS SUD ORSAY

Abstract

Faculty of Science
Department of Statistics

Doctor of Philosophy

by Merlin Keller

This thesis is dedicated to the statistical analysis of multi-subject fMRI data, with the purpose of identifying bain structures involved in certain cognitive or sensori-motor tasks, in a reproducible way across subjects. To overcome certain limitations of standard voxel-based testing methods, as implemented in the Statistical Parametric Mapping (SPM) software, we introduce a Bayesian model selection approach to this problem, meaning that the most probable model of cerebral activity given the data is selected from a pre-defined collection of possible models.

Based on a parcellation of the brain volume into functionally homogeneous regions, each model corresponds to a partition of the regions into those involved in the task under study and those inactive. This allows to incorporate prior information, and avoids the dependence of the SPM-like approach on an arbitrary threshold, called the cluster-forming threshold, to define active regions. By controlling a Bayesian risk, our approach balances false positive and false negative risk control. Furthermore, it is based on a generative model that accounts for the spatial uncertainty on the localization of individual effects, due to spatial normalization errors.

On both simulated and real fMRI datasets, we show that this new paradigm corrects several biases of the SPM-like approach, which either swells or misses the different active regions, depending on the choice of a cluster-forming threshold.

Acknowledgements

Mes remerciements vont en premier lieu à Alexis Roche pour la disponibilité et la patience dont il a fait preuve dans son encadrement tout au long de ma thèse, et pour sa capacité à me faire reprendre de la hauteur vis-à-vis de mon sujet, les nombreuses fois où je me perdais dans des détails. Merci également de m'avoir fait découvrir le monde merveilleux de Python et du développement collaboratif; longue vie à NiPy!

De la même façon, je remercie Marc Lavielle d'avoir accepté de diriger mon travail, avec une rigueur et un soucis de clarté qui ont été des guides précieux dans l'élaboration de ma démarche. Merci également pour les innombrables solutions apportées aux divers problèmes méthodologiques que j'ai rencontrés, et qui m'ont permis d'approfondir ma connaissance des algorithmes d'optimisation stochastique.

Merci à Elisabeth Gassiat d'avoir accepté de présider mon jury de thèse. Un grand merci à Mark Woolrich pour sa relecture détaillée du manuscript, ses remarques pertinentes, ainsi que pour avoir affronté la neige et les pannes de train pour traverser la Manche et venir participer à ma soutenance!

Mon séjour au laboratoire de neuroimagerie assisté par ordinateur (LNAO) fut l'occasion de nouer de nombreuses collaborations, sans lesquelles ce travail n'aurait pas vu le jour. Ma gratitude va en premier lieu à Jean-François Mangin pour son accueil au sein de son équipe, et son soucis d'assurer les meilleures conditions de travail possibles aux membres de son équipe.

Merci à Sébastien, mon prédécesseur, dont le travail a constitué le socle sur lequel je me suis appuyé dans mes recherches. Merci également pour son soutien lors de mon stage de master, et grâce auquel j'ai pu faire mes premières armes dans l'utilisation de SPM et de la fameuse Distance Toolbox!

Merci à Bertrand Thirion, pour l'intérêt dont il a fait preuve vis-à-vis de mon travail, et dont le point de vue toujours pertinent fut une rélle source d'inspiration. Merci aussi pour les barbecues et autres raclettes si conviviaux qui ont ponctués ces trois années!

Merci à Philippe Ciuciu, grand Bayésien et marathonien devant l'Eternel, pour son abondante expertise sur les techniques d'échantillonnage stochastique, ainsi que pour la découverte en course à pied des plus beaux recoins du plateau de Saclay...

Ma gratitude va également à Edouard Duchesnay, par qui toute cette aventure a commencée, qui m'a initié aux mystères de la neuroimagerie, et donné envie d'en connaître plus. Merci à Jean-Baptiste Poline pour son soutien constant et le point de vue toujours pertinent et constructif qu'il a apporté sur mon travail. Merci également aux héros du

service informatique, Dimitri Papadopoulos et Pascal Stokowski, sur qui reposaient la lourde charge d'assurer la continuité d'un réseau informatique capricieux...

Je tiens également à remercier la fine équipe des thésards, post-docs et assimilés dont j'ai eu le plaisir de faire partie, pour l'esprit de stimulation et d'entraide mutuelle, ainsi que pour les nombreuses sorties et soirées qui permettaient de décompresser et repartir de plus belle, bravo en particulier aux remarquables organisatrices que furent Cécilia et Valdis!

Une dédicace spéciale aux 'geeks' de la bande, Thomas, Alan, Benjamin, Gaël, Matthieu et les autres, pour les innombrables heures passées à combler patiemment mes lacunes en informatique! Merci aussi à Dominique, Denis et Yann d'avoir si souvent répondu à mes appels désespérés lorsque je n'arrivais pas à utiliser les outils maison (suite le plus souvent à l'oubli d'une virgule dans une ligne de commande!)

Reprendre des études après cinq années d'interruption ne s'est pas fait sans difficultés, et je suis énormément redevable à toute l'équipe du master de statistiques d'Orsay, qui m'a permis d'effectuer cette reprise dans les meilleures conditions.

Un grand merci tout d'abord à Pascal Massart, qui m'a ouvert les portes du master, puis m'a si bien conseillé tout au long de cette transition. Merci aux enseignants pour leur clarté et leur pédagogie, et plus particulièrement à Liliane Bel et Gilles Celeux dont la bienveillante attention m'a permis de prendre confiance dans mes capacités à faire de la recherche.

Enfin, mes pensées vont à ma famille dont le soutien sans faille a été indispensable à l'accomplissement de ce travail, et vers qui j'ai toujours pu me tourner dans les inévitables moments de doutes qui accompagnent tout chercheur, particulièrement à ses débuts! Merci tout spécialement à Astrig qui fut en première ligne durant ces années, et qui a notamment accepté que j'emmène mon ordinateur portable en vacances pour rattraper le travail en retard...

Je conclue en dédiant cette thèse à mon neveu Stanley, dont la curiosité débordante, l'imagination foisonnante et l'envie toujours renouvelée de se dépasser ne manquent jamais de me surprendre, et laissent augurer à mon avis d'une brilliante carrière d'inventeur génial!

Contents

Abbreviations

AAL	**A**utomated **A**tlas **L**abels
BOLD	**B**lood **O**xygen **L**evel **D**ependent
CSA	**C**ortical **S**ulci **A**tlas
fMRI	**f**unctional **M**agnetic **R**esonance **I**maging
FDR	**F**alse **D**iscovery **R**ate
FFX	**F**ixed **FX** *(effect)*
FPR	**F**alse **P**ositive **R**ate
FSL	**F**MRIB **S**oftware **L**ibrary
FWER	**F**amily-**W**ise **E**rror **R**ate
GLM	**G**eneral **L**inear **M**odel
HRF	**H**aemodynamic **R**esponse **F**unction
MCMC	**M**onte **C**arlo **M**arkov **C**hain
MCP	**M**ultiple **C**omparison **P**rocedure
MFX	**M**ixed **FX** *(effect)*
MH	**M**etropolis–**H**astings
NRL	**N**eural **R**esponse **L**evel
ROI	**R**egion **O**f **I**nterest
RFX	**R**andom **FX** *(effect)*
SAEM	**S**tochastic **A**veraging **E**xpectation **M**aximization
SPM	**S**tatistical **P**arametric **M**apping

Notations

$\mathcal{V} = (\mathbf{v}_1, \ldots, \mathbf{v}_d) \subset \mathbb{N}^3$	Search volume (voxel grid)
$\mathcal{V} = \mathcal{V}_1 \cup \ldots \cup \mathcal{V}_N$	Partition into N regions of interest
$\ell \in \{1, \ldots, N\}^d$	Region labels ($\forall k$, $\mathbf{v}_k \in \mathcal{V}_{\ell_k}$)
$\mathbf{y}_i = (y_{i,1}, \ldots, y_{i,d})$	Subject i estimated effects map (observations), $1 \leq i \leq n$. We also use the notation $\mathbf{y}_i(\mathbf{v}_k) := y_{i,k}$
$\mathbf{s}_i^2 = (s_{i,1}^2, \ldots, s_{i,d}^2)$	Subject i estimation variance map (known)
$\mathbf{x}_i = (x_{i,1}, \ldots, x_{i,d})$	Subject i effects map (latent variable)
$\mathbf{u}_i = (\mathbf{u}_{i,1}, \ldots, \mathbf{u}_{i,d})$	Subject i registration errors
φ_i	Voxel k is displaced to voxel $\varphi_i(k)$ for subject i
$\mathbf{w}_i = (\mathbf{w}_{i,1}, \ldots, \mathbf{w}_{i,B})$	Subject i elementary displacements
$\mathbf{v}_{k_1}, \ldots, \mathbf{v}_{k_B}$	Deformation field control points
$\mathcal{K}(\cdot, \cdot)$	Interpolation kernel
ω	Deformation field regularity
σ_S^2	Elementary displacements variance
$\boldsymbol{\mu} = (\mu_1, \ldots, \mu_d)$	Mean group effect map
$\boldsymbol{\eta} = (\eta_1, \ldots, \eta_N)$	Regional means
$\boldsymbol{\nu}^2 = (\nu_1^2, \ldots, \nu_N^2)$	Regional variances
$\boldsymbol{\gamma} \in \{0, 1\}^N$	Indicator variable ($\gamma_j = 1$ iff $\eta_j \neq 0$)
$\mathbf{p} = (p_1, \ldots, p_N)$	Prior probabilities
$\boldsymbol{\sigma}^2 = (\sigma_1^2, \ldots, \sigma_N^2)$	Between-subject variance (one per region)
α, β, λ	Hyperparemeters of the Normal-Inverse Gamma prior

Chapter 1

Introduction

1.1 Context and objectives

This thesis is dedicated to the analysis of multi-subject fMRI data. Functional magnetic resonance imaging (fMRI) is a modality of *in vivo* brain imaging that allows to measure the variations of cerebral blood oxygen levels induced by the neural activity of a subject lying inside a MRI scanner and submitted to a series of stimuli (see Figure 1.1). One of the main goals of fMRI group studies, through the analysis of the data acquired on a cohort of subjects, is to identify brain structures involved in certain cognitive or sensori-motor tasks, in a reproducible way across subjects.

Many sources of variability are present in fMRI datasets, making statistical methods necessary to perform such analyses. These include, but are not limited to, the movements of the subject inside the machine, his or her level of attention during the experiment, various physiological signals, such as heartbeats and breathings, etc. Furthermore, the BOLD response itself varies across the different runs of a single experiment, due to habituation and attentional effects. For all these reasons, the measures acquired on each subject are uncertain, and variable should the experiment be reproduced with the same subject. Additionally, each subject reacts differently when exposed to the same stimuli, so a certain variability in the brain activation pattern is also expected across the subjects.

Accounting for these two types of variability, classically referred to as *within*-subject and *between*-subject, is a well-known problem in the literature on statistical inference

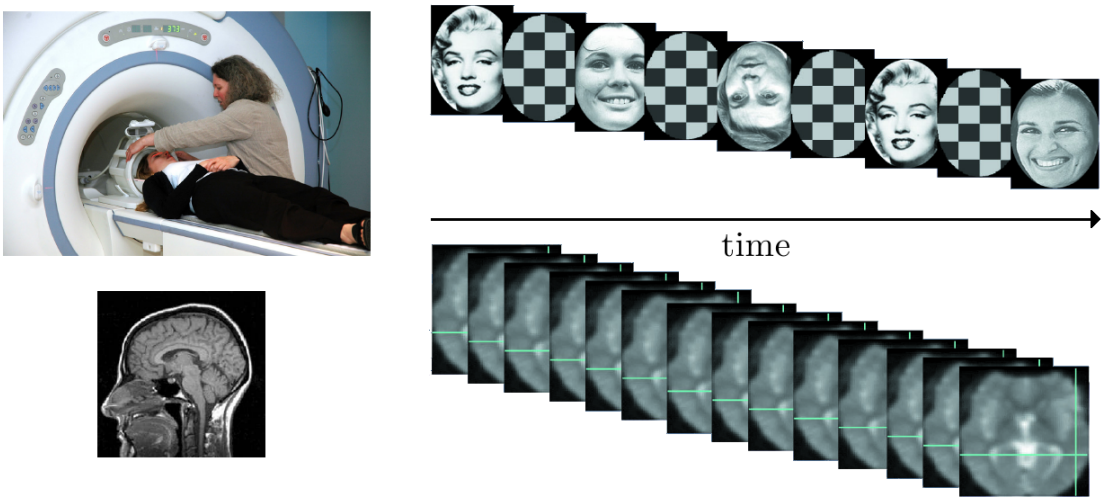


FIGURE 1.1: A typical fMRI experiment. The subject, lying in the MRI scanner (upper left) is submitted to a series of stimuli (upper right), visual, auditory or other, while a series of 3D images of his/her brain activity is acquired (lower right). A structural image of the subject's brain (lower left) is also acquired, using the same scanner.

for biological data [Mcculloch and Searle, 2001]. Another challenge, specific to medical imaging, and in particular to neuroimaging, is the important variability of the human brain anatomy, which makes the comparison of brain images from different subjects far from straightforward. To address this, individual images are most often normalized to a common brain template using nonrigid registration. However, registration is prone to errors (even assuming the existence of point-to-point correspondences between different brains), hence it does not seem reasonable to assume that homologous points are aligned across subjects.

As a consequence of all these fluctuations, the activation pattern inferred from any given fMRI dataset can only be determined with a limited amount of confidence. Ideally, the statistical procedure used to analyse the data should therefore account for the different sources of variability, and assess the uncertainty they induce on the estimated activation pattern.

The purpose of this work is to propose such a statistical analysis framework. This is still an open problem, though the development of appropriate methods for fMRI group inference has been an active field of research for at least two decades. The most widely used approach to address these questions is the co-called mass-univariate, or voxel-based detection, popularized by [Friston, 1997]. It starts with normalizing individual images onto a common brain template, as mentioned above. Next, a t-statistic is computed

in each voxel to locally assess mean group effects. The candidate regions are then defined as the connected components (clusters) of the resulting statistical map above an arbitrary threshold called the cluster-forming threshold. Only the clusters whose sizes exceed a critical value are reported. This critical size acts as a second threshold, and is generally tuned to control the probability of one or more clusters being detected by chance. For scientific reporting purposes, the detected clusters may finally be related to known anatomical regions based on expert knowledge, or using a digital brain atlas such as the Automated Atlas Label (AAL) [Tzourio-Mazoyer et al., 2002].

Though simple and widely applicable, this approach suffers from several shortcomings. One of these is the dependence on an arbitrary cluster-forming threshold. High values of this threshold may result in missing active regions, and low values in merging functionally distinct regions, yielding poor localization power, due to the fact that active voxels cannot be localized within each detected cluster [Hayasaka and Nichols, 2003]. Furthermore, the absence of activations outside the detected clusters cannot be assessed. Consequently, there is no guarantee that the whole functional network can be recovered. Finally, due to unavoidable intersubject registration errors, the observed activations are not well-localized, and possibly displaced across distinct functional regions, which may result in blurring the group activation map, and create non-handled false positives.

To date, these limitations of the mass-univariate approach have been tackled separately. This thesis introduces a new approach for fMRI group data analysis that addresses them jointly.

1.2 Organization and main contributions

The remaining of this thesis is organized in six chapters, which we now summarize:

- In Chapter 2, we give a review of current approaches for fMRI group data analysis, with a focus on mass-univariate detection. This was popularized by the Statistical Parametric Mapping (SPM) software package, and we refer to it as SPM-like in the following. Our goal here is not to give a complete list of existing methods, but rather summarize the main directions of research that have been followed up to

now. This chapter also contains several contributions we have made to the SPM-like approach. [Keller et al., 2007, Keller and Roche, 2008] present an implementation of the Gaussian two-level group model using the maximum likelihood ratio test, a nonparametric generalization of which is developed in [Roche et al., 2007]. We have also participated to another paper [Thirion et al., 2006b] describing a high-level group analysis approach which models the spatial variability of the activation patterns, and which we have included in this review.

- In Chapter 3, we propose to overcome the dependence of the SPM-like approach on an arbitrary detection threshold, and its exclusive control over false positives, by using the random threshold approach developed in [Lavielle and Ludeña, 2007]. One of the key ideas of this chapter, which we re-exploit in the following, is to re-interpret the task of detecting activations, traditionally seen from a multiple testing point of view, as a model selection problem. This work has been submitted to the Canadian Journal of Statistics.

- Chapter 4 proposes a detection method that relaxes the assumption of perfect match. To do this, we extend the classical mass-univariate model by incorporating a set of latent variables representing registration errors, and model them as random deformation fields. Our main results consist in demonstrating the stretching effect of the group estimated activation pattern due to neglecting spatial uncertainty, and its compensation through our approach. A first version of this work was published in [Keller et al., 2008].

- Chapter 5 introduces a new paradigm for fMRI group data analysis that addresses jointly some key limitations of the SPM-like approach, and was published in [Keller et al., 2009]. Based on a Bayesian model selection framework, regions involved in the task under study are selected according to the posterior probabilities of a nonzero mean activation, given a pre-defined parcellation of the search volume into functionally homogeneous regions. Thus our approach is threshold-free, while allowing to incorporate prior information, provided that the parcellation is sensible. By controlling a Bayesian risk, our approach balances false positive and false negative risks, with weights than can be tuned depending on the application domain. Importantly, it is based on the same spatial uncertainty model as in Chapter 4, and thus accounts for the mis-alignment of individual images, due to

inevitable registration errors. Hence for each subject, the membership of a voxel to a given region is probabilistic rather than deterministic.

Results on simulated data show that neglecting spatial uncertainty may result in a bias toward false positives when selecting active regions. This is a consequence of the stretching effect evidenced in Chapter 4. We also show that this bias can be compensated when modeling spatial uncertainty.

- In Chapter 6 we validate our model selection approach on a real fMRI dataset, by successfully recovering the whole network of regions involved in basic number and language processing, in accordance with previous works. The purpose of this chapter is also to illustrate the limitations of the SPM-like approach, which we apply to the same dataset, and the benefits of modeling spatial uncertainty.

- We conclude in Chapter 7 by a summary of the main results and discuss perspectives for future work.

  Finally, all the methods developed throughout this work where implemented in Python, and are now part of the NIPY open-source neuroimaging analysis package http://neuroimaging.scipy.org/site/index.html. Methods concerning the SPM-like approach have also been integrated into the fMRI toolbox of the BrainVisa software, developed at CEA/Neurospin http://www.brainvisa.info/index.html.

Chapter 2

fMRI group data analysis: the 'SPM-like' approach

2.1 Introduction

The main goal of this chapter is to give a detailed account of the current approaches to fMRI group data analysis, with a focus on the approach primarily developed in [Friston, 1997] and popularized by the Statistical Parametric Mapping (SPM) software, which we will hence refer to in the following as the 'SPM-like' approach. It is the most widely used method to date, and has served both as the basis for our research, and the reference to validate our methods.

This approach comprises two main steps. First, each subject's data is processed separately, resulting in a series of summary statistics, such as a statistical map of the subject's brain activity in response to any given contrast of experimental conditions. This step is described in Section 2.2, along with some basic facts on the nature of fMRI data.

Next, the subjects' summary statistics are used as input data for group level analysis. In Section 2.3 we describe how this is accomplished in the 'SPM-like' approach, and discuss its advantages and limitations.

Many alternatives have been proposed to overcome the limitations of the 'SPM-like' approach. We give an overview of the current literature on this subject in Section 2.6,

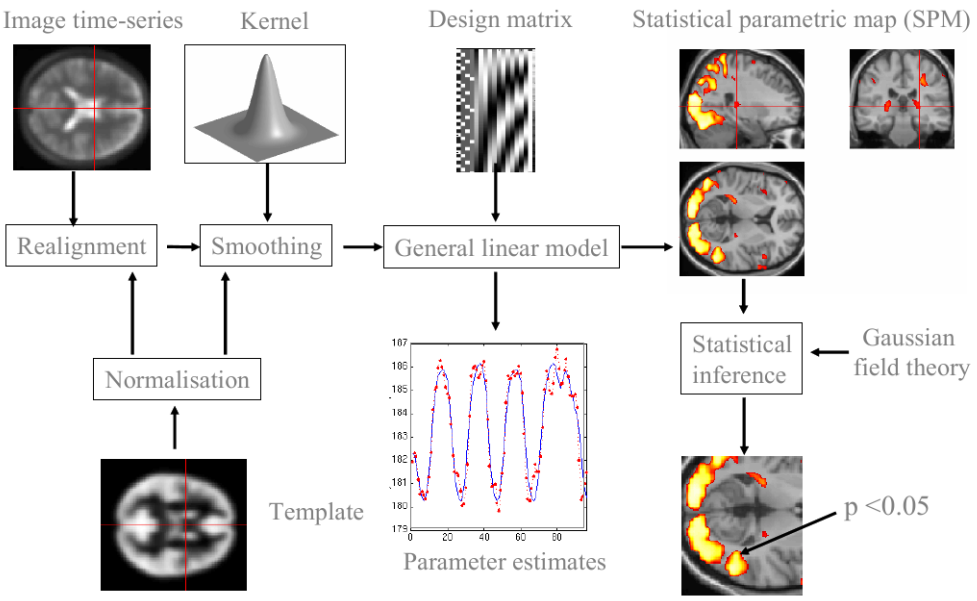


FIGURE 2.1: Pipeline of single-subject data analysis using the SPM-like approach, from [Friston et al., 1995]. The data is realigned, normalized to a brain template representing a standard space, and smoothed. These pre-processings are described in Sections 2.2.1 and 2.3.1. Next, the time-series are modeled separately in each voxel using the general linear model, whose parameters represent the amplitude of the BOLD response evoked by each type of stimuli (see Section 2.2.3). Active brain areas are detected by thresholding a map of test statistics, as explained in Section 2.2.4.

then in Section 2.7 discuss the directions we have decided to follow in our work, and how these relate to the existing approaches.

2.2 Single-Subject Data Analysis

2.2.1 The blood oxygen level dependent effect

Since the early nineties, functional magnetic resonance imaging (fMRI) has become one of the principal tools to investigate the functional organization of the human brain. This popularity is due to both the relatively good spatial resolution and low invasiveness of fMRI compared to other functional imaging modalities. For instance, positron-emission tomography (PET) involves exposure to ionizing radiation; electroencephalography (EEG) or magnetoencephalography (MEG) have significantly less spatial resolution, and involve an inverse-problem for source reconstruction which has no unique solution.

Using fMRI, cerebral activity is measured indirectly, through its impact on the vascular network. More precisely, neural activity induces a local increase in blood oxygenation, known as the blood oxygen level dependent (BOLD) effect [Ogawa et al., 1990] (see Figure 2.2). Because of the magnetic properties of the oxyhemoglobin molecule (oxyHb), this effect can be measured by the scanner. One may note that this inherently limits the precision of this technique in locating activations, since the surge in blood oxygenation does not necessarily take place at the exact place of the neural activity. The link between these two phenomenons is still an active research area.

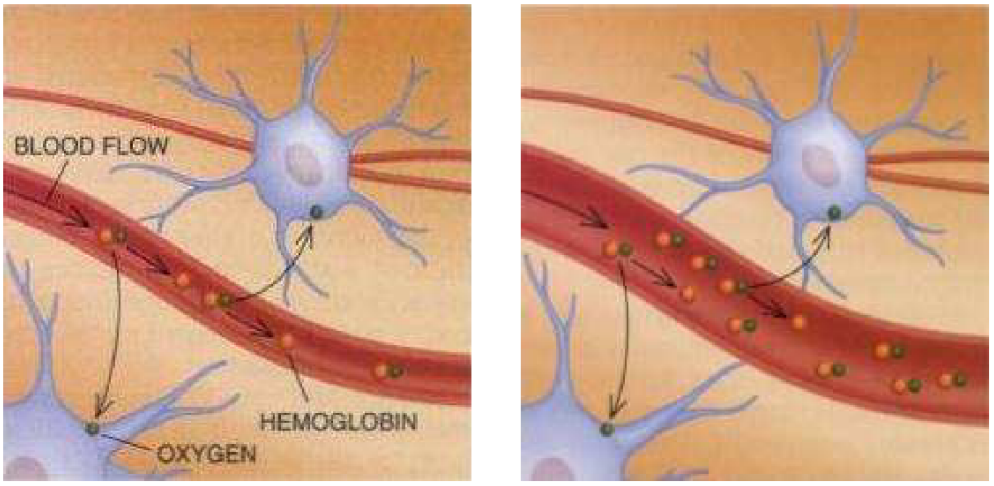


FIGURE 2.2: Illustration of the BOLD effect, from [Raichle, 1994]. (left) at rest (right) during an activation. Each neural activation induces a local increase in blood flow, that is substantially higher than the increase in oxygen consumption of the nearby neurons. This results in a fall in the concentration in *deoxyhemoglobin* (in orange), and a surge in the fMRI signal.

2.2.2 Data acquisition and preprocessings

Most fMRI paradigms are designed to identify brain areas involved in certain cognitive or sensori-motor tasks. They consist in a series of stimuli (visual, auditory, etc.) presented to the subject lying inside the scanner. The primary motive of this thesis lies in the analysis of data arising from such experiments. An alternative which has received much attention lately consists in scanning the subject in absence of external stimuli. The aim of studying such 'resting state' data is to identify latent attentional networks, *e.g.* using multivariate exploratory techniques [Beckmann and Smith, 2004, Perlbarg et al., 2008].

In all cases, the data resulting from a fMRI experiment consists in a sequence of three-dimensional (3D) brain images, with a typical spatial resolution of $3mm^3$ on a standard

3 Tesla (3T) scanner, acquired at the rate of one volume every 2 to 3 seconds. An anatomical image is also acquired using the same scanner, at a higher spatial resolution of about $1mm^2$ planar resolution (inter-slice resolution may be lower).

Before undergoing group analysis, this data is submitted to a number of pre-processing steps. These are briefly summarized hereafter (see [Friston, 1997] for further details). Pre-processing usually starts with the temporal realignment, also termed *slice-timing*, of the successive scans to ensure that all time series inside the volume are sampled at the same time-points, even though the different slices are acquired at slightly different moments. Next, the scans are realigned in order to compensate for subject's motion inside the scanner.

After temporal and spatial realignment, the data is rigidly registered to the associated anatomical image of the same subject, and warped using a nonrigid transformation resulting from the registration of the anatomical image to a brain template, which represents a standard stereotactic space such as the Talairach space [Talairach and Tournoux, 1988]. This so-called spatial normalization step has a key impact on group data analysis, as discussed in Section 2.3.1. Note that the different spatial transformations resulting from motion correction, functional/anatomical registration, and brain warping, are usually composed in order to prevent the data from being resampled more than once after slice timing correction.

Finally, the data is spatially smoothed, in order to enhance the signal-to-noise (SNR) ratio, most often using a Gaussian kernel. The spread of this kernel, measured by its full-width at half maximum (FWHM), usually varies between 5 and 8 mm [Friston, 1997].

2.2.3 GLM analysis of time series

The pre-processed data of a single subject is most often analysed separately at each *voxel* using regression techniques. This approach, often termed *massively univariate* because voxels are processed independently from one another, has been developed primarily in [Friston, 1997, Worsley et al., 2002], and is available in many software packages, such as Statistical Parametric Mapping (SPM) or the FMRIB Software Library (FSL).

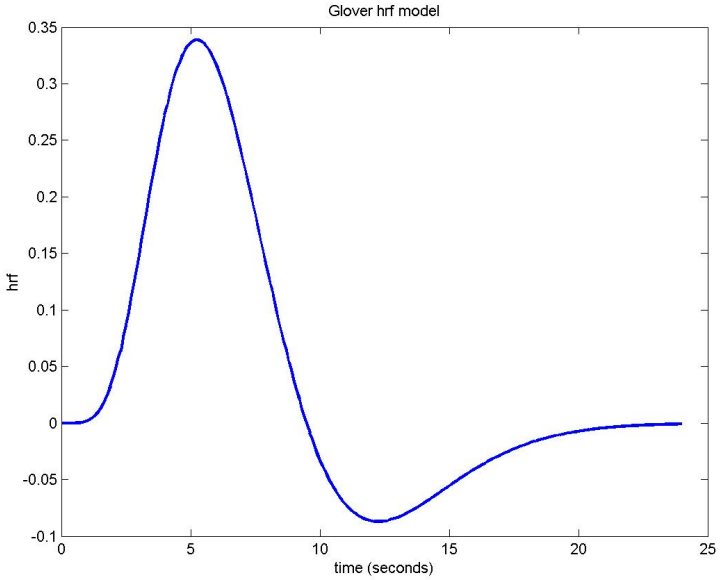


FIGURE 2.3: **Glover haemodynamic response function (HRF).**

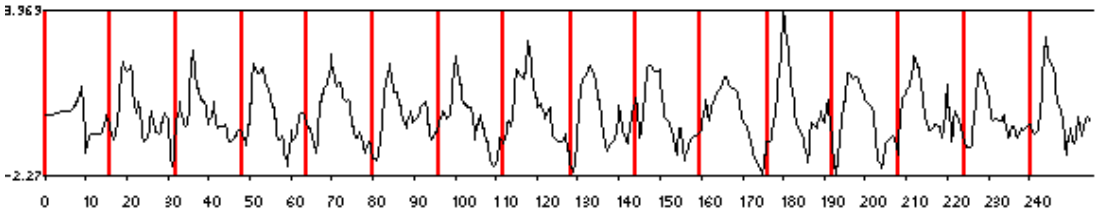


FIGURE 2.4: Example of a fMRI time-series observed in a certain voxel. Occurrence times (onsets) of the successive stimuli appear in red (time is measured in seconds).

BOLD response modeling

Let us denote $\mathbf{Y}_k = (\mathrm{Y}_{k,1}, \ldots, \mathrm{Y}_{k,T})$, the time-series observed at voxel k, for $k = 1, \ldots, d$, where d denotes the number of voxels within the search volume. This volume may be defined as the intersection of individual brain regions extracted from the functional images, or a unique region delineated in the template space.

The time-series, illustrated in Figure 2.4, reflects the BOLD response at voxel k. The response to a series of stimulations of the same kind is traditionally modeled as a stationary linear filter with finite impulse response called the HRF. This leads to a general linear model (GLM):

$$\mathbf{Y}_k \quad = \quad \mathbf{X}\boldsymbol{\beta}_k + \boldsymbol{\varepsilon}_k. \tag{2.1}$$

Here $\mathbf{X}$ is the $T \times m$ *design matrix* containing the m regressors (one per column), and $\boldsymbol{\beta}_k$ the corresponding parameters of interest. Regressors include the constant vector $\mathbf{1}_T$, modeling a baseline activation, along with vectors modeling the BOLD responses to the different stimulations, also termed *neural response levels* (NRLs). These are defined for each given stimulus type as the (discretized) convolution of the HRF $\boldsymbol{h}$ by a function $\mathbf{S}$ representing the external stimulation. For event-related fMRI designs, $\mathbf{S}$ is a sum of Dirac masses specifying the occurrence times, or *onsets* of the successive stimuli, whereas for block-related fMRI designs, $\mathbf{S}$ is typically a sum of boxcar functions, specifying the beginning and the end of each stimulation block.

Additional regressors may be added to model confounds, such as a low frequency drift due to physiological signals and scanner-related artefacts. This drift can be decomposed on a family of functions such as periodical (e.g. sine/cosine) functions or polynomials [Friston, 1997].

Noise modeling

In (2.1), $\boldsymbol{\varepsilon}_k \in \mathbb{R}^d$ models the estimation errors, and is usually defined as a zero-mean Gaussian random variable. If its covariance matrix is spherical, *i.e.* if the estimation errors are assumed to be *i.i.d.*, with common variance σ_k^2, then the least square estimate and the maximum likelihood estimate of $\boldsymbol{\beta}_k$ coincide. This common estimator is the best linear unbiased estimate (BLUE) of the model parameters, and is found by solving the linear system $\mathbf{X}'\mathbf{Y}_k = \mathbf{X}'\mathbf{X}\boldsymbol{\beta}_k$. The variance σ_k^2 is estimated in this case by the residual mean squares $\frac{1}{T-r} \parallel \mathbf{Y}_k - \mathbf{X}\boldsymbol{\beta}_k \parallel^2$, where r is the rank of the design matrix $\mathbf{X}$.

However, fMRI time-series are known to be temporally correlated [Boynton et al., 1996, Zarahn et al., 1997]. Noting $\mathbf{V}_k$ the covariance matrix of the noise process $\boldsymbol{\varepsilon}_k$, the BLUE of the model parameters is now obtained by solving the following linear system:

$$\mathbf{X}'\mathbf{V}_k^{-1}\mathbf{Y}_k \quad = \quad \mathbf{X}'\mathbf{V}_k^{-1}\mathbf{X}\boldsymbol{\beta}_k, \tag{2.2}$$

in the case where $\mathbf{V}_k$ is known. In general though, it is unknown and must also be estimated. As noted in [Donnet et al., 2006], this is an ill-posed problem since there are more parameters to be determined than available observations, so constraints must be imposed on $\mathbf{V}_k$.

Several approaches have been adopted. [Friston, 1997, Worsley et al., 2002] model $\boldsymbol{\varepsilon}_k$ as first-order auto-regressive (AR(1)) process, so that at each time-point $t = 1, \ldots, T$:

$$\varepsilon_{k,t} \quad = \quad \rho_k \varepsilon_{k,(t-1)} + \xi_{k,t}, \tag{2.3}$$

where ρ_k is a spatially varying autocorrelation parameter, and $\boldsymbol{\xi}_k$ is a Gaussian white noise, with a spatially varying variance σ_k^2. In [Friston, 1997], a simpler version of this model is used, where ρ_k is assumed uniform across the search volume. In both cases, the variance parameters are estimated by restricted maximum likelihood (ReML), and the estimated covariance matrix $\hat{\mathbf{V}}_k$ is substituted to the unknown true one in (2.2). The precision of the resulting 'plug-in' estimate of model parameters depends crucially on the accuracy of the covariance estimate $\hat{\mathbf{V}}_k$. The study in [Friston and Buechel, 2000] suggests that the use of AR(1) with a global auto-correlation parameter may result in biased parameter estimates. Temporal smoothing is then suggested to improve the estimation.

More sophisticated techniques to estimate $\mathbf{V}_k$ are compared in [Woolrich et al., 2001], including a Tukey taper combined with nonlinear spatial smoothing, which performs optimally among the tested strategies. Also, distinct auto-correlation parameters are estimated in each voxel, an option which is not considered in [Friston and Buechel, 2000], and may in part explain the observed bias.

An alternative to the above plug-in estimation is to estimate the covariance and regression parameters jointly, by maximizing the full likelihood of the model. This is the approach developed in [Roche et al., 2004], where an AR(1) model is fit separately to the time-series acquired in each voxel. This is done iteratively, using an extension of the Kalman filter. This strategy further enables an 'online', or real-time, estimation of the model, the parameter estimates being updated as new scans are acquired.

Generalization to higher-level auto-regressive models (AR(p), with $p \geq 1$) is considered in [Penny et al., 2003], where a variational Bayes (VB) approximation is used to jointly estimate all model parameters. A very interesting feature of this approach is that it allows to select the order p of the auto-regressive model, separately in each voxel. This is done by maximizing the free energy $\mathcal{F}(p)$, that is, the lower bound on the marginal likelihood central to the VB approach, easily available from the output of the VB algorithm. On an application to a particular real fMRI dataset, histograms of the optimal

values for the AR model order p show that it never exceeds $p = 3$. In fact, in most voxels it is equal to either 0 or 1. This provides a rough justification for the models described above, which all assume $p = 1$.

Limits and alternatives

We conclude this review on single-subject data modeling by noting that the formulation in (2.1) relies on the following key assumptions:

- **Fixed HRF.** Several canonical shapes have been proposed for the HRF, such as the Glover HRF [Glover, 1999], used in SPM, defined as the difference of two Gamma functions (see Figure 2.3). This simplifying assumption may lead to a poor model fit in regions where the true HRF is significantly different from the one used in the estimation.

- **Linearity.** The effects of the different stimuli are assumed to contribute additively to the overall effect evoked by the experimental paradigm, thus justifying the use of a linear model. This assumption may hold for an event-related paradigm if the inter-stimulus gap is long enough [Boynton et al., 1996, Glover, 1999]. Otherwise, non-linear effects may occur.

- **Time-Invariance.** For a given stimulus type, the amplitude of the BOLD response is assumed constant across the successive stimulations, justifying the use of a single regressor for each stimulus type. This neglects so-called habituation effects, such as repetition-suppression, *i.e.* a decrease in the BOLD response observed when the same stimulus is presented repeatedly. Modeling such phenomenons using the standard GLM requires to incorporate additional regressors, hence reducing the degrees of freedom. Thus parametric modulation may be preferable, as discussed below.

The modeling and analysis of brain haemodynamics is still a field of active research. To cite only a few examples, [Friston et al., 2000] suggested modeling non linearities in the BOLD response through Volterra series while remaining in the GLM framework, by adding additional regressors, defined as temporal derivatives of the canonical HRF. To overcome the limitations inherent to the use of a canonical HRF, [Penny et al., 2003,

Woolrich et al., 2004b] modeled the HRF as a linear combination of a pre-defined set of functions. In [Makni et al., 2008], a joint detection-estimation (JDE) framework is developed to estimate at the same time the shape $\boldsymbol{h}$ of the HRF, in a nonparametric way, and the NRLs $\boldsymbol{\beta}$, in a Bayesian setting. In [Donnet, 2006] an alternative approach to the statistical modeling of fMRI time series is investigated, based on the physiological model of the blood flow / oxygen consumption coupling proposed in [Buxton and Frank, 1997].

Another limitation, inherent to the mass univariate model, is that it ignores the spatial structure of the data. Several authors, including [Gössl et al., 2001, Penny et al., 2007, Smith and Fahrmeir, 2007, Makni et al., 2008], have proposed to model the NRLs, according to a spatial mixture model, with three classes, corresponding to null, activated and deactivated (inhibited) voxels. The labels identifying the state of each voxel are further modeled jointly as a hidden Markov random field, which tends to group active voxels in clusters. This has a regularizing effect, since isolated voxels, assimilated to noise, are less likely to be detected. Such spatial mixture models have also been proposed to model directly the statistical maps resulting from the GLM analysis, as a means to compute a detection threshold (see Section 2.6.1 for a discussion of these methods).

2.2.4 Statistical map of brain activity

Based on the model (2.1), an effect of interest, such as a difference between stimuli, can be specified by $\mathbf{c}\boldsymbol{\beta}_k$, where $\mathbf{c}$ is a row vector of m contrasts [Worsley et al., 2002]. Detecting a nonzero effect is then equivalent to testing the null hypothesis that the effect is zero, $\mathcal{H}_{0,k} : \mathbf{c}\boldsymbol{\beta}_k = 0$. Assuming the data has been pre-whitened as described in the previous section, so that its covariance structure is approximately spherical, it follows from the Neyman-Pearson lemma that the optimal statistic for testing $\mathcal{H}_{0,k}$ is Student's t-statistic:

$$T_k = \frac{\mathbf{c}\hat{\boldsymbol{\beta}}_k}{\hat{\sigma}_k \sqrt{\mathbf{c}(\tilde{\mathbf{X}}_k' \tilde{\mathbf{X}}_k)^{-1}\mathbf{c}'}}, \tag{2.4}$$

where $\tilde{\mathbf{X}}_k$ is the pre-whitened design matrix given by $\mathbf{V}_k^{-\frac{1}{2}}\mathbf{X}_k$, given the covariance matrix $\mathbf{V}_k\sigma_k^2$ of the data $\mathbf{Y}_k$. If $\mathbf{V}_k$ is known, then under $\mathcal{H}_{0,k}$, T_k follows a Student

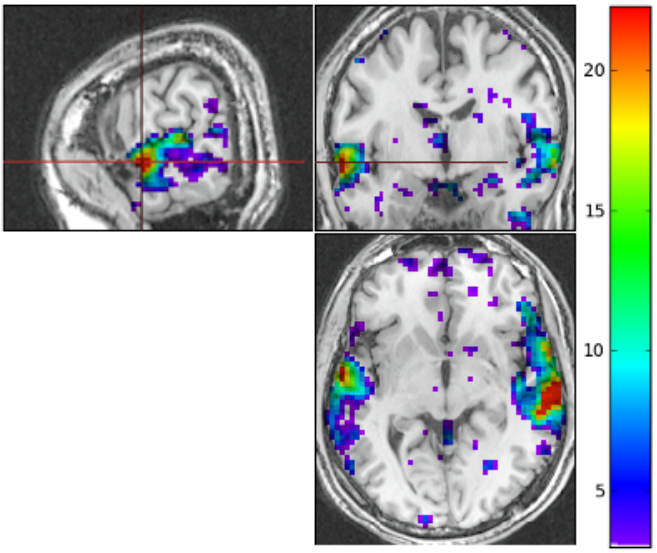


FIGURE 2.5: Individual t-score map for the 'audio-video' contrast, from the Localizer dataset [Pinel et al., 2007], with the the subject's anatomical image in the background. The map is thresholded at 10^{-3}, uncorrected, meaning that the presence of an activation is tested independently in each voxel at level 10^{-3}, without accounting for multiple comparisons. Activations are clearly seen in the bilateral temporal regions; many isolated voxels are also detected, which may be false positives.

distribution with $T-(m+1)$ degrees of freedom (df). Consequently, $\mathcal{H}_{0,k}$ is rejected at level α if $T_k > t_\alpha$, where t_α is the $(1-\alpha)$-th quantile of the Student distribution with $T-(m+1)$ df. In practice $\mathbf{V}_k$ is estimated from the data, as explained in Section 2.2.3, so the test level is not exactly α.

Computing T_k for all voxels $k = 1, \ldots, d$ results in a statistical map, also termed *activation map*, reflecting the subject's activation pattern for the given contrast $\mathbf{c}$. Detecting activated regions in this setting is equivalent to testing simultaneously all the null hypotheses $\mathcal{H}_{0,k}$, the number of which is the number of voxels in the search volume and is typically of the order of 100 000.

To do so, one must address the multiple comparison problem. For instance, the naive procedure that consists in rejecting $\mathcal{H}_{0,k}$ for all voxels k such that $T_k > t_\alpha$, as described above, choosing for instance $\alpha = 0.001$, ensures that the probability of a false detection in each voxel is at most 0.001. However, this means that, in the worst-case scenario where all voxels are in fact inactive, an average of $\alpha \times 100\,000 = 100$ false positives are detected throughout the volume. This issue is illustrated in Figure 2.5.

The same multiple testing problem arises when detecting group activation pattern using the 'SPM-like' approach, as we will see in Section 2.3.3. Thus, we postpone this issue

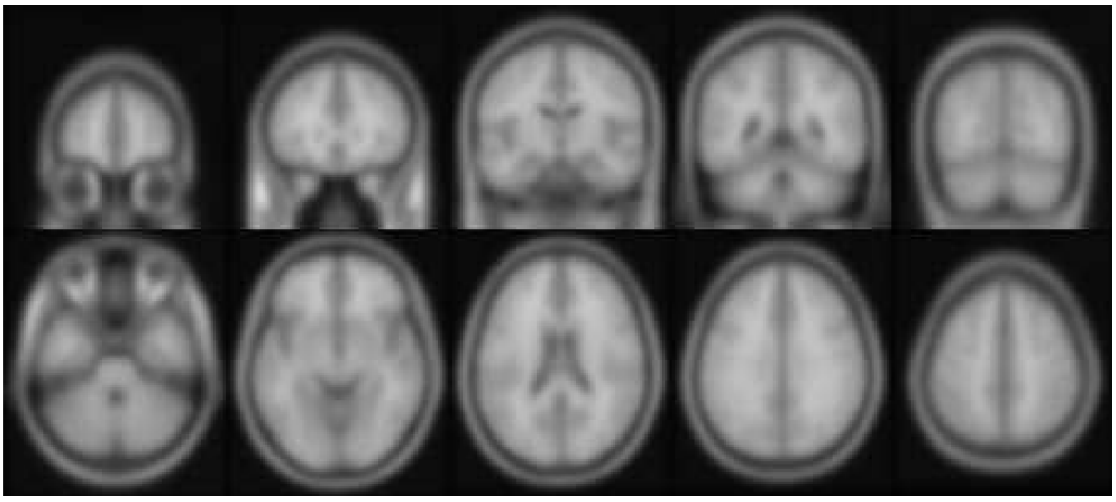


FIGURE 2.6: coronal and axial slices of the MNI template, from [Flandin, 2004]. This is obtained as the average of 152 individual anatomical images, registered to an earlier version of the template (MIN305) using affine transformations.

to Section 2.4, where we adress it jointly for individual and group data analysis.

2.3 Group analysis: The mass univariate approach

In a typical fMRI study, several subjects are recruited from a population of interest and scanned while submitted to the same series of stimuli. Activation maps associated with a given contrast are obtained for each subject, as described in the previous section, and used as input data for inference at the between-subject level, where the goal is to evidence a general brain activity pattern.

In the following, we describe how this is performed in the 'SPM-like' approach, starting with the spatial normalisation step in Section 2.3.1, which aims to match each individual image to a brain template. The activation maps are then compared on a voxelwise basis, as described in Section 2.3.2, resulting in the computation of a map of statistics, detailed in Section 2.3.3, to test in each voxel the presence of a positive mean activation across subjects.

2.3.1 Spatial normalisation and smoothing

The high morphological variability of the human brain [Brett et al., 2002], makes the comparison of cerebral images across subjects problematic.

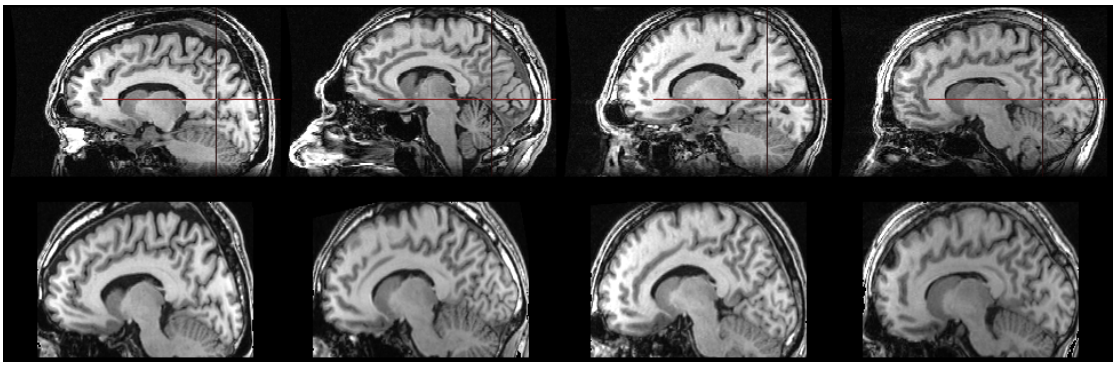


FIGURE 2.7: Illustration of spatial normalisation on a sagittal slice: (top) images, before normalisation; (bottom) after normalisation, using affine transformations. Anatomical differences across subjects are seen to persist, especially in the sulco-gyral geometry.

A traditional way to compensate for this variability, as mentioned in Section 2.2.1, is to register, or normalize, the individual images of all subjects to a common brain template [Ashburner and Friston, 1999], such as the widely used Montreal Neurological Institute (MNI) template (see Figure 2.6). This is usually done by minimizing a measure of discrepancy between image intensities over a suitable class of spatial transformations (see Chapter 4 for a brief review on registration methods). Comparative studies of several normalization methods can be found in [Hellier et al., 2003, Klein et al., 2009].

Any location in the brain can then be marked in a standard coordinate system, such as the one developed by [Talairach and Tournoux, 1988]. However, registration is prone to errors (even assuming the existence of point-to-point correspondences between different brains), hence it does not seem reasonable to assume that homologous points are exactly aligned across subjects.

This fact is often used as a motivation to justify a preliminary linear spatial smoothing of the data, with a typical FWHM of 8 to 12 mm, as a way to increase the overlap of functionally homologous regions over subjects. This smoothing is sometimes applied to the individual activation maps, resulting from the individual data processing, in addition to the first smoothing step used on the raw fMRI time-series to enhance SNR (see section 2.2.1).

The 'SPM-like' approach then compares the individual images on a voxelwise basis, thus making an implicit assumption that each subject is in perfect match with the template. Among the consequences of that assumption, one may anticipate a stretching effect on group activity patterns due to the "jitter" induced by inaccurate registration. This effect

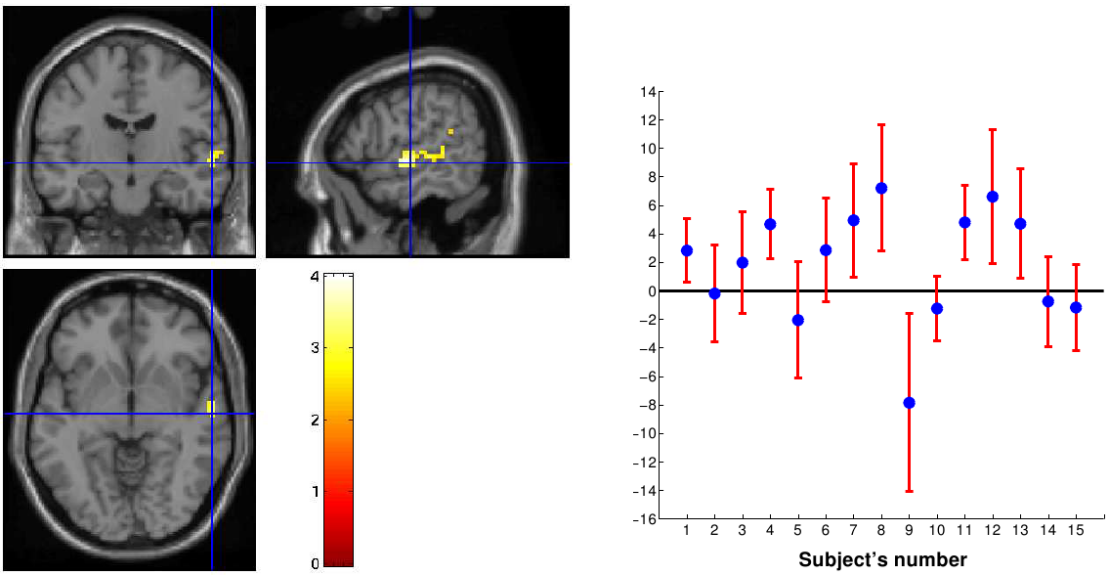


FIGURE 2.8: Example of fMRI group data in one voxel, during a language processing task, from [Mériaux et al., 2006]. Left, activations detected using the one-sided mixed-effect statistic, thresholded using a permutation test (see Section 2.3.3), with cross-hair at $(60; 15; 6)$ Talairach coordinates in mm. Right, plot of the estimated effects in the same voxel and associated 70% confidence intervals

can only be reinforced by the above-mentioned preliminary smoothing step. Evidence for this stretching is provided in [Keller et al., 2008], and will be one of the important results of Chapter 4.

2.3.2 Between-subject modeling

Let $\mathbf{x}_i = (x_{i,1}, \ldots, x_{i,d})$ denote the map of BOLD effects in response to a certain contrast of experimental conditions, for subject $i = 1 \ldots, n$. As seen in Section 2.2.4, a noisy estimate of $\mathbf{x}_i$, $\mathbf{y}_i = (y_{i,1}, \ldots, y_{i,d})$ is available from the analysis of the subject's scans, along with an image of estimation variances $\mathbf{S}_i^2 = (s_{i,1}^2, \ldots, s_{i,d}^2)$.

More precisely, for each subject i, following the notations introduced in Section 2.2.4, we define in each voxel k :

$$x_{i,k} := \mathbf{c}\boldsymbol{\beta}_k; \quad y_{i,k} := \mathbf{c}\hat{\boldsymbol{\beta}}_k; \quad s_{i,k}^2 := \hat{\sigma}_k \sqrt{\mathbf{c}(\tilde{\mathbf{X}}_k' \tilde{\mathbf{X}}_k)^{-1}\mathbf{c}'},$$

for a certain row contrast vector $\mathbf{c}$.

Under sufficient degrees of freedom at the within-subject level, it is reasonable to consider $y_{i,k}$ as being normally distributed around $x_{i,k}$ with standard deviate $s_{i,k}$ considered fixed

[Worsley et al., 2002]. To address questions regarding the variability of the effect in a population, the unobserved effects $x_{1,k}, \ldots, x_{n,k}$ are further modeled as independent random variables drawn from an unknown distribution which characterizes the across-subject variability of BOLD responses. When this distribution is assumed Gaussian with unknown mean and variance (μ_k, σ_k^2), we obtain the same hierarchical model as in [Worsley et al., 2002, Beckmann et al., 2003a, Mériaux et al., 2006]:

- First level (within-subject):

$$y_{i,k} = x_{i,k} + \varepsilon_{i,k}; \quad \varepsilon_{i,k} \overset{ind.}{\sim} \mathcal{N}(0, s_{i,k}^2), \tag{2.5}$$

- Second level (between-subject):

$$x_{i,k} = \mu_k + \eta_{i,k}; \quad \eta_{i,k} \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma_k^2), \tag{2.6}$$

where the independence sampling assumptions at both levels imply that in each voxel k, the pairs $(x_{1,k}, y_{1,k}), \ldots, (x_{n,k}, y_{n,k})$ are mutually independent conditionally on the population parameters (μ_k, σ_k^2). By integrating out the hidden variables $x_{i,k}$, we see that the observed effects are drawn independently but, in general, non-identically from Gaussian distributions:

$$y_{i,k} = x_{i,k} + \xi_{i,k}; \quad \xi_{i,k} \overset{ind.}{\sim} \mathcal{N}(0, s_{i,k}^2 + \sigma_k^2). \tag{2.7}$$

That is to say, the observations are generally heteroscedastic unless all first-level deviations $s_{i,k}$ are equal. In this special case, the model boils down to the simple sampling model in [Friston et al., 1995], that is computationally attractive but may lack robustness against noisy observations .

We conclude this description of the standard model for fMRI group data with the following remarks:

- As for individual subject data (see Section 2.2.3), the group data is modeled separately in each voxel, in a 'massively univariate' fashion. Consequently, possible

correlations between neighboring voxels are ignored at this stage. They will be accounted for in the multiple comparison step, discussed in Section 2.4.

- The within-subject model (2.5) is also referred to as a *fixed-effect* (FFX) model, since it specifies the unobserved effects $x_{1,k}, \ldots, x_{n,k}$ as the (fixed) parameters of interest. Inference in this model is therefore limited to the cohort of subject scanned during the experience. Likewise, the between-subject model (2.6) is referred to as a *random-effect* (RFX) model, since it considers the effects $x_{1,k}, \ldots, x_{n,k}$ as random variables. Finally, the hierarchical model specified by (2.5) and (2.6) is also called a *mixed-effect* (MFX), or *mixed*, model, since it 'mixes' the FFX and RFX models.

- As noted in [Worsley et al., 2002], besides simply inferring on the mean population effect μ_k, we may also wish to compare two or more populations, and more generally regress the subjects' effects $x_{i,k}$ on a certain set of regressors $\boldsymbol{z}_{i,k}$. To this end, the mean effect μ_k in (2.6) may be replaced by a linear term $\boldsymbol{z}'_{i,k}\,\boldsymbol{\mu}_k$ in the regressor variables. Though in the following we will focus on the one-sample setting for the sake of simplicity, the methods exposed here are fully adaptable to the general regression setting, except when mentioned otherwise.

Several extensions to the standard hierarchical model specified by (2.5) and (2.6) have been proposed. In [Woolrich et al., 2004a] for instance, the uncertainty on first-level standard deviates $s_{i,k}$, is accounted for. More specifically, the general linear model (2.1) used to estimate the subject-specific parameters is directly used at the within-subject level, rather than the Gaussian approximation (2.5).

Based on this more realistic model, a Bayesian approach is developed wherein the posterior distribution of the group mean effect, $p(\mu_k|y_{1,k}, \ldots, y_{n,k})$, is estimated, either using MCMC techniques, or a faster approximation, assimilating it to a noncentral t-distribution.

Such an approach may seem computationally intensive at first, since the complete data of all subjects must be analyzed together. However, it is shown in [Woolrich et al., 2004a] that the posterior distribution of μ_k depends in fact only on certain summary statistics of the individual datasets. These can be computed beforehand, hereby greatly reducing the computational complexity of this approach.

Some authors have further proposed to relax the Gaussian assumption at the between-subject level. In [Roche et al., 2007], the distribution of the $x_{i,k}$'s is assumed totally unknown, and a nonparametric maximum likelihood estimation (NMLE) method is developed. In [Woolrich, 2008], a mixture model consisting of two Gaussian distributions is used to describe population heterogeneity, one class corresponding to possible outlier subjects.

2.3.3 Test of a nonzero group effect

Based on the hierarchical model introduced in the previous section, our final goal is to test in each voxel k the presence of a nonzero mean effect, *i.e.*, test $\mathcal{H}_{0,k} : \mu_k = 0$ versus $\mathcal{H}_{1,k} : \mu_k \neq 0$. Besides this *two-sided* null hypothesis, one may also wish to test the presence of a positive effect (corresponding to an activation), that is, test the *one-sided* null hypothesis $\mathcal{H}_{0,k} : \mu_k \leq 0$ versus $\mathcal{H}_{1,k} : \mu_k > 0$. This test involves the two following steps:

- Definition of a test statistic $T_k = T_k(y_{1,k}, \ldots, y_{n,k})$
- Statistical calibration of T_k, *i.e.* computing the threshold u such that $P(T_k \geq u|\mathcal{H}_{0,k}) \leq \alpha$, for any required level α. The level specifies the probability of falsely rejecting $\mathcal{H}_{0,k}$, *i.e.,* the probability of a type I error, or false positive.

Choice of a test statistic

In the simple RFX model, *i.e.*, when the first-level standard deviates $s_{i,k}$ are implicitly assumed constant, as in [Friston et al., 1995], a natural choice is the standard t-statistic

$$T_k = \frac{\bar{\mathbf{y}}_k}{std(\mathbf{y}_k)/\sqrt{n-1}},$$

where $\bar{\mathbf{y}}_k = \frac{1}{n}\sum_{i=1}^{n} y_{i,k}$ and $std(\mathbf{y}_k)^2 = \frac{1}{n}\sum_{i=1}^{n}(y_{i,k} - \bar{\mathbf{y}}_k)^2$. This yields the uniformly more powerful (UMP) test at any level α, both for the one-sided and the two-sided test [Lehmann, 1986].

In the general MFX model, there is no optimal choice of a test statistic in terms of power. In [Worsley et al., 2002], σ_k^2 is estimated by restricted maximum likelihood (ReML), using an expectation-maximization (EM) algorithm [Dempster et al., 1977]. This is equivalent to an iterative reweighted least-square procedure to estimate μ_k, using the weight $(s_{i,k}^2 + \hat{\sigma}_k^2)^{-1}$ for observation $y_{i,k}$. In analogy to the RFX case, the following approximate t-statistic is then used:

$$\tilde{T}_k = \frac{\hat{\mu}_k}{\hat{\sigma}_k / \sqrt{n}}.$$

A similar approximate t-statistic is used in [Woolrich et al., 2004a], where (μ_k, σ_k) are estimated in a Bayesian setting by their posterior mean, having marginalized out all other hidden variables using a Monte-Carlo Markov-Chain (MCMC) sampling algorithm.

Another, more systematic, approach, advocated in [Mériaux et al., 2006, Keller and Roche, 2008], is to use the maximum likelihood ratio (MLR) for the two-sided test:

$$R_k = \frac{\sup_{\mu_k = 0,\ \sigma_k^2 \in \mathbb{R}_+^*} \mathcal{L}_k(\mu_k, \sigma_k^2)}{\sup_{\mu_k \neq 0,\ \sigma_k^2 \in \mathbb{R}_+^*} \mathcal{L}_k(\mu_k, \sigma_k^2)},$$

where $\mathcal{L}_k(\mu_k, \sigma_k^2)$ is the likelihood of the model (2.7). For the one-sided test, the following sign modulation is used:

$$\tilde{T}_k = sign(\hat{\mu}_k)\sqrt{R_k}.$$

The maximum likelihood estimates $(\hat{\mu}_k, \hat{\sigma}_k)$ can be computed using an EM algorithm. In [Roche et al., 2007], this approach is extended to the nonparametric setting, and it is shown that the nonparametric maximum likelihood estimate of the between-subject distribution is a combination of at most n Dirac masses.

Statistical calibration

Voxelwise statistical calibration is straightforward in the simple RFX model in [Friston et al., 1995], using the fact that, under $\mathcal{H}_{0,k}$, T_k follows a Student distribution with $n-1$ degrees of freedom. The two-sided null hypothesis $\mathcal{H}_{0,k} : \mu_k = 0$ is then rejected if $|T_k| > t_{\alpha/2}$, where $t_{\alpha/2}$ is the $(1-\alpha/2)$-th quantile of the t_{n-1} distribution, and the one-sided null hypothesis $\mathcal{H}_{0,k} : \mu_k \leq 0$ is rejected if $T_k > t_\alpha$. The Student distribution is also used as a substitute for the unknown null distribution of the approximate t-statistics in

[Worsley et al., 2002] and [Woolrich et al., 2004a] (see the previous section), using the same degrees of freedom as the standard t-statistic. Use of the Student distribution is valid in both cases if the data distribution is Gaussian, or for large sample sizes ($n \to \infty$), owing to the central limit theorem.

An alternative solution to this calibration problem is to use permutation tests [Good, 2005]. They allow exact control on the type I error, under mild assumptions on the sampling distribution, and for any choice of a test statistic. This method of calibration was introduced in the neuroimaging literature by [Holmes et al., 1996]. It consists in sampling the null distribution of the test statistic by permuting the data, under certain exchangeability hypotheses. Having sampled N values, the threshold of the test is then simply equal to the $[N\alpha]$-th largest sampled value.

The principal limitations of permutation tests are the heavy computations they require, and also their limited applicability. For instance, the universally exact control on false positives does not extend to the test of partial correlations in a multiple regression model. Approximate permutation testing procedures have been proposed in this case [Anderson and Legendre, 1999, Cade, 2005]. They have been found empirically to provide a more precise control over false positives than standard parametric tests, in certain cases where the assumptions underlying the latter are not verified. However, no general result exists to support these observations.

2.4 Multiple comparisons

A multiple comparison problem arises when testing several voxels simultaneously. This is the same problem encountered in the analysis of individual subject data (see Section 2.2.4). In its simplest form, it consists in finding a detection threshold u for a given statistical map $\mathbf{T} = (T_k)_{1 \leq k \leq d}$ such that the balance between specificity (control over false positives) and sensitivity (control over false negatives) is optimal in a certain sense. Appendix A.1 gives a precise definition of these concepts. Furthermore, the chosen multiple comparison procedure (MCP) must also produce results that are useful in terms of neuroscience. Thus the detected activations must be easily linked to known anatomical structures. Moreover, a stringent control over false positives is usually required to avoid erroneous interpretations.

MCP's in neuroimaging can be divided into two main categories. The first one contains voxel-level thresholding procedures, that are presented in Section 2.4.1. However, due to the strict control on false positives, and the large number of tests performed (up to $\approx$ 100 000 voxels in a whole brain image), these procedures typically lack sensitivity. Another shortcoming is that they detect individual voxels, which are hard to interpret, as they do not constitute relevant biological units.

Another type of MCP's has been developed to overcome these limitations, which we will refer to as cluster-level procedures. First, the statistical map is thresholded as before, but this time at an arbitrary level. Next, connected components, or *clusters* of voxels above this threshold are identified. Finally, the presence of activations is tested within each cluster, using a secondary threshold on the cluster's size. Cluster-level procedures are in general more powerful than their voxel-level counterparts, since they entail much less hypotheses to be tested. A review of these approaches is presented in Section 2.4.2, and we conclude by discussing the pro's and con's of the different methods.

2.4.1 Voxel-level inference

We now address the problem of choosing a detection threshold u for a map of voxelwise test statistics $\mathbf{T} = (T_1, \ldots, T_d)$, controlling a certain error rate. The problem can also be defined as that of finding a threshold c for the map of p-values $(p_1, \ldots, p_d)$. Here $\mathbf{T}$ may stand either for the activation map of an individual subject (see Section 2.2.4), or a group activation map (see Section 2.3.3). In both cases, the null hypotheses considered here, noted $\mathcal{H}_k = 0$ in Appendix A.1, are the voxelwise null hypotheses $\mathcal{H}_{0,k}$.

FWER-controlling procedures

A lot of attention has been given to the control of the FWER, that is, the probability of one or more false positives, in the neuroimaging literature, as a strict control on false positive is necessary in order for the detected activated brain areas to be reliable. We review here some of the main approaches. We refer to Appendix A.1 for a rigorous definition of this and other multiple test error rates.

The Bonferroni procedure. The Bonferroni procedure consists in rejecting every null hypothesis $\mathcal{H}_{0,k}$ whose p-value is smaller than α/d, where α is the desired upper bound on the FWER, and d the number of tested hypotheses (one per voxel). Its justification relies on no other assumption than subset pivotality, necessary to define the p-values (see Appendix A.1). It is a direct application of Boole's inequality, and the fact that each p_k is uniformly distributed under $\mathcal{H}_{0,k}$:

$$\begin{aligned} FWER &= P[V > 0 | H_{\mathcal{M}_0}] \\ &= P[\cup_{k \in \mathcal{M}_0} \{p_k < \alpha/d\} | \mathcal{H}_{\mathcal{M}_0}] \\ &\leq \sum_{k \in \mathcal{M}_0} P[p_k < \alpha/d | \mathcal{H}_{0,k}] \\ &= d_0 \times \alpha/d \leq \alpha. \end{aligned} \tag{2.8}$$

In the above derivation, $d_0 = \sharp\mathcal{M}_0$ is the number of true null hypotheses. It can be shown that, under independence of the data across voxels, the Bonferroni procedure is near optimal [Ge et al., 2003]. This means that there is no procedure significantly more powerful having the same level of FWER strong control.

The maxT Procedure. In the case of mutually dependent tests, the more general maxT procedure can be applied. It relies on the fact that, under the global null $\mathcal{H}_{\mathcal{M}}$, the probability of one or more false detections can be controlled knowing the null distribution of the maximal statistic:

$$\begin{aligned} P[V > 0 | H_{\mathcal{M}}] &= P[\exists k \in \mathcal{M}, T_k > u | \mathcal{H}_{\mathcal{M}}] \\ &= P\left[\max_{k \in \mathcal{M}} T_k > u | \mathcal{H}_{\mathcal{M}}\right]. \end{aligned} \tag{2.9}$$

Thus, to control $P[V > 0|\mathcal{H}_{\mathcal{M}}]$ at a given level α, the threshold u must be equal to the $(1-\alpha)$-th quantile of the distribution of the maximal statistic $\max_{k \in \mathcal{M}} T_k$ under the global null $\mathcal{H}_{\mathcal{M}}$. Note that this method gives weak control on the FWER. Strong

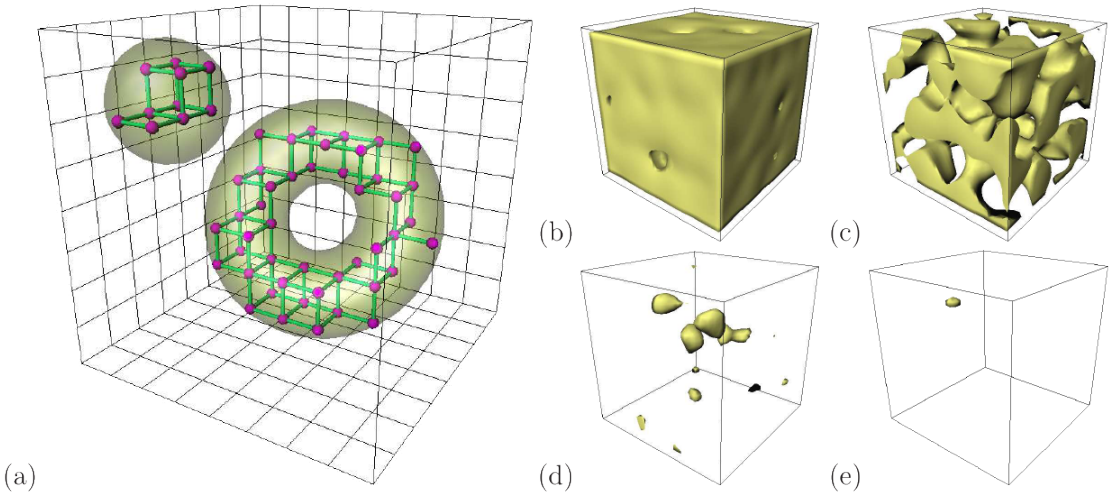


FIGURE 2.9: Excursion sets of $3D$ random fields, from [Taylor and Worsley, 2007]. (a) Ball and torus. The corresponding Euler characteristic is: $\chi = 2 - 1 + 0 = 1$. (b) Isotropic Gaussian field, with zero mean and variance one, above a threshold $t = -2$, corresponding to $\chi = 6$ (unseen hollows contribute +1 each) (c) $t = 0$, $\chi = -6$ (handles dominate) (d) $t = 2$, $\chi = 14$ (handles disappear) (e) $t = 3$, $\chi = 1$.

control follows under subset pivotality [Westfall and Young, 1993]. The justification, elementary, is given in Appendix A.

Computing Tail Probabilities. The main difficulty in applying the maxT principle is the computation of the tail probabilities $P\left[\max_{k\in\mathcal{M}} T_k > u|\mathcal{H}_\mathcal{M}\right]$. These depend strongly on the definition of the test statistics T_k, and on the distribution of the underlying data. Hence, many different approaches have been proposed, that we briefly review here. A detailed study can be found in [Nichols and Hayasaka, 2003].

Parametric approximations were introduced in [Worsley, 1994]. They essentially equate the tail probability $P\left[\max_{k\in\mathcal{M}} T_k > u|\mathcal{H}_\mathcal{M}\right]$ with the expectation of the Euler characteristic χ of the *excursion set* $\{k \in \mathcal{M}, T_k > u\}$. For a 3D excursion set, χ is essentially the number of connected components, minus the number of 'handles', plus the number of 'hollows' (see Figure 2.9). For a high threshold u, the excursion set is expected to be either empty, or composed of a single cluster with no holes, hence the tail probability is approximately equal to $E[\chi|\mathcal{M}]$. This last expectation can be estimated using inequalities from random field theory (RFT), that are based on an estimation of the smoothness of the spatial map $\mathbf{T}$. In practice, these approximations yield satisfying results for smooth maps, but are overly conservative for non smooth maps, where the Bonferroni correction is optimal [Nichols and Hayasaka, 2003]. An approach to bridge

this gap is developed in [Worsley, 2005, Taylor et al., 2007, Taylor and Worsley, 2007], using improved Bonferroni-type inequalities based on the discrete local maxima (DLM) of the statistical map. These lead to bounds on the tail probability evaluated using RFT-type approximations, that are shown to be near optimal at all smoothness levels.

The main drawback of RFT-based approximations is that they rely on heavy parametric assumptions, which are hard to verify in practice, such as the stationarity, Gaussianity and smoothness of the statistical map. They are also restricted to a certain class of test statistics, such as t or F-statistics. Finally, their domain of validity is hard to determine, since they assume a 'high' threshold u, but it is not clear what this means in practice.

In the context of group data analysis, it is possible to avoid these issues by using a permutation test to tabulate the null distribution of $\max_{k\in\mathcal{M}} T_k$ [Holmes et al., 1996]. When applicable, this approach has many advantages: It is valid under minimal assumptions concerning the distribution of the data; it can be applied to any choice of the test statistic T_k; finally it is near optimal in terms of statistical power [Nichols and Hayasaka, 2003]. Despite of all these virtues, permutation testing has not yet replaced RFT techniques as a standard for fMRI group data analysis, though both are available in SPM and FSL, the most used software packages to date for the analysis of fMRI data. This may in part be explained by the important computation time they require, and also by the limited range of questions they allow to answer. Indeed, as mentioned in Section 2.3.3, their exists no generally exact permutation test of a partial correlation in a multiple regression model.

FDR-controlling procedures

In spite of the abundant literature devoted to voxelwise FWER-controlling procedures, their application remains limited by their lack of power. This is due both to the strict control they impose on false positives, and to the large number (up to a hundred thousands) of voxels, and therefore of tested hypotheses, present in brain activation maps. More recently, there have been some attempts to control the less stringent FDR criterion instead in order to gain statistical power. This is illustrated in Figure 2.10, where different error rate controls are compared for a same activation map.

The most famous procedure for controlling the FDR is the Benjamini-Hochberg procedure [Benjamini and Hochberg, 1995]. It consists in sorting the p-values in increasing order: $p_{(1)} \leq \ldots \leq p_{(d)}$. They are then compared to the linear series $\left(\frac{k}{d}\alpha\right)_{1\leq k\leq d}$, where α is the desired FDR level. Next the following index is computed: $k^* = \max\{1 \leq k \leq d,\ p_{(k)} \leq \frac{k}{d}\alpha\}$. If for all k, $p_{(k)} > \frac{k}{d}\alpha$, then no null hypothesis is rejected. Otherwise, all null hypotheses $\mathcal{H}_{0,k}$ for $k = 1, \ldots, k^*$ are rejected. [Benjamini and Hochberg, 1995] showed that, under independence of the p-values, this procedure had strong control over the FDR, *i.e.*,

$$FDR = E\left[\frac{V}{R}1_{R>0}\right] \leq \frac{d_0}{d}\alpha \leq \alpha.$$

This proof, along with the simplicity of the algorithm, greatly contributed to popularize this approach.

Several works have been concerned with the use of FDR for neuroimaging data, such as [Pacifico et al., 2004], which proposes an extension of of the BH approach to random fields, based on an estimation of the field's smoothness similar to the one developed in [Worsley, 1994]. In [Nichols and Hayasaka, 2003], was expressed the belief that FDR would soon replace FWER as a standard type I error rate in fMRI data analysis. However, relatively few applications have been published to date based on the FDR, though the BH procedure has been included in the reference SPM software. To this we see two possible explanations. The first one is that the BH procedure can be unstable with respect to the p-values, which is a problem when analyzing fMRI datasets, known to be very noisy. Hence, while allowing better statistical power, FDR-controlling procedures are also liable to produce more false positives, without being able to discriminate them from true positives, as illustrated in Figure 2.10. Another one is that, as mentioned earlier, voxels do not constitute relevant units from a cognitive point of view. In a broader sense, this may also be the main reason why voxel-level inference methods are not mainstream in neuroimaging papers concerned with the applications.

2.4.2 Cluster-level inference

We now turn to cluster-level multiple testing, which is the most widely used approach to detect activated regions given fMRI activation maps. Using the same notations as above, the statistical map $(T_1, \ldots, T_d)$ is first thresholded at a certain height u. However, in this context u does not serve to test the voxelwise hypotheses $\mathcal{H}_{0,k}$. Rather, it is used

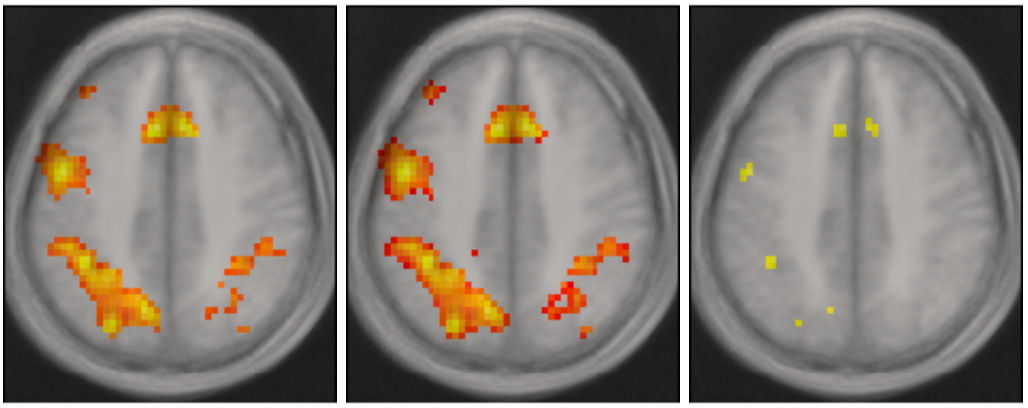


FIGURE 2.10: Voxel-level error rate controls compared on the same activation map. *Left:* FPR controlled at 0.001, using voxelwise tests. *Middle:* FDR controlled at 0.05 using the BH procedure. *Right:* FWER controlled at 0.05 using the Bonferroni procedure. The FDR control enables to detect substantially more voxels than the FWER control at the same level. However, 5% of voxels detect by this approach, *i.e.* 118 in this example, are expected to be false positives. In contrast, less than 60 false positives are detected on the average using a simple voxelwise test (left), for a qualitatively similar result.

to identify connected components, or *clusters*, from the excursion set $\{k \in \mathcal{M}, T_k > u\}$. Thus it is referred to as a *cluster-defining threshold*, and is a fixed, user-specified quantity. The choice of u is arbitrary; a popular heuristic is to tune it in order to control the voxelwise FPR at a certain level α, that is, allow a proportion α of null voxels to be retained at this stage.

Next, the presence of an activation is tested within all detected clusters $C_1, \ldots, C_M$. More precisely, for each cluster C_i, the null hypothesis: $\mathcal{H}_{\mathcal{C}_i} = \bigcap_{k \in C_i} \mathcal{H}_{0,k}$ is tested, against the alternative $\bar{\mathcal{H}}_{\mathcal{C}_i} = \bigcup_{k \in C_i} \mathcal{H}_{1,k}$. The cluster null hypothesis is equivalent to stating that C_i is enclosed in the null set $\mathcal{M}_0$ defined in Section A.1: $\mathcal{H}_{\mathcal{C}_i} = \{C_i \subseteq \mathcal{M}_0\}$. The alternative hypothesis $\bar{\mathcal{H}}_{\mathcal{C}_i}$ states that at least one voxel in C_i is activated.

The decision statistic used to test each cluster-level hypothesis is generally the cluster size $\sharp C_i$, though many other choices are available, as discussed in Section 2.6.1. Hence, the null hypotheses $\mathcal{H}_{\mathcal{C}_i}$ are rejected for all clusters whose sizes exceed a critical value N, tuned to control a certain error rate, that is, for clusters that would be 'unusually large' in absence of any activation.

As in voxel-level inference, cluster-level multiple comparison procedures are usually required to have strong control over the FWER, which in this case is the probability of

detecting one ore more clusters by mistake:

$$\begin{aligned} FWER &= P[V > 0|\mathcal{H}_{\mathcal{M}_0}] \\ &= P[\exists C_i \subseteq \mathcal{M}_0, \sharp C_i > N|\mathcal{H}_{\mathcal{M}_0}]. \end{aligned}$$

Following the maxT principle (see Section 2.4.1), this is usually done by controlling the tail probability of the maximum cluster size, under the global null $\mathcal{H}_{\mathcal{M}}$, since the true subset of null hypotheses is unknown. This means tuning N so that

$$P\left[\max_{C_i} \sharp C_i > N|\mathcal{H}_{\mathcal{M}}\right] \leq \alpha. \tag{2.10}$$

This procedure only provides weak control on the FWER; strong control follows under subset pivotality [Holmes et al., 1996]. The proof of this result is given in Appendix A.

Computing the tail probabilities of null distribution of the the maximum cluster size $\max_{C_i} \sharp C_i$ is the principal difficulty of this approach. The principal methods for doing so are reviewed in [Hayasaka and Nichols, 2003]. As previously, parametric approximations based on RFT theory are available [Worsley et al., 2002], as well as exact calibration based on permutation tests, if $\mathbf{T}$ is a group activation map. These have the same advantages and drawbacks as in the voxel-level setting (see Section 2.4.1). To summarize, though easy and fast to implement, the RFT approach relies on heavy parametric assumptions, and may be overly conservative is the statistical map is not smooth; the permutation testing approach is exact under very mild nonparametric assumption, and is always near optimal [Hayasaka and Nichols, 2003]. Its only drawback is that it is computationally intensive.

Cluster-level inference as described above has several key advantages over voxel-level inference. First, the number M of clusters is much smaller than the number d of voxels. Thus, testing cluster-level hypotheses greatly reduces the multiple comparison problem; this explains why cluster detection is in general more powerful than voxel detection. Another advantage is that results are reported in terms of regions rather than voxels, and are therefore easier to interpret from a cognitive point of view. The detected clusters may be related to known anatomical regions based on expert knowledge, or using a digital brain atlas such as the Automated Atlas Label (AAL) [Tzourio-Mazoyer et al., 2002].

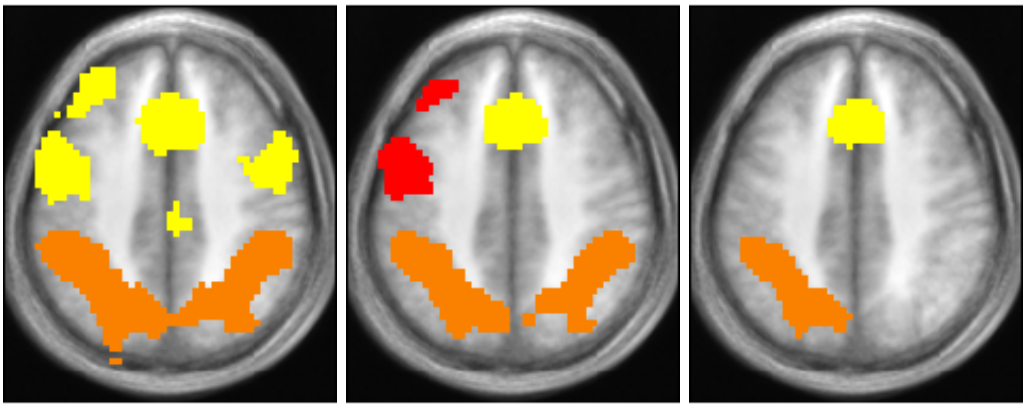


FIGURE 2.11: Clusters detected at different cluster-forming thresholds. Axial slices $z = 37$mm in Talairach, represented with the subjects' mean anatomical image in the background). From left to right, the threshold is tuned to control the false positive rate (FPR) respectively at 10^{-2}, 10^{-3} and 10^{-4} uncorrected. Each cluster surviving the FWER controlling-threshold at 5% is represented with a specific color, showing how distinct functional regions are merged. In particular, left and right hemispheres are not segmented.

The principal disadvantage of cluster-level inference wish respect to voxel-level inference is its dependence on the cluster-defining threshold, which ultimately defines the detected regions. The fact that a suprathreshold cluster is found significant by the cluster size test only implies that it contains *some* active voxels [Hayasaka and Nichols, 2003]. Low values of the cluster-forming threshold may result in merging functionally distinct regions, thus yielding poor localization power, while high values may result in missing active regions. This is illustrated in Figure 2.11.

2.5 Limits of the SPM-like approach

The 'SPM-like' approach has been summarized in Sections 2.2, 2.3 and 2.4. In the following, we will assume that activations are detected at the cluster-level, as described in Section 2.4.2. Though simple, and widely applicable, this approach suffers from several important limitations, which we summarize here:

Arbitrary cluster-forming threshold. As noted in Section 2.4.2, clusters are defined for a certain cluster-defining threshold. The choice of this threshold is arbitrary, even though it is crucial for the resulting inference.

Exclusive control of false positives. Error rates are controlled by maximizing the statistical power of the tests while limiting a certain type I error rate at a given level α (see Section A.1). Consequently, there is no direct control on the amount of type II errors, or false negatives, meaning that the absence of activations outside the detected clusters cannot be assessed, and that there is no guarantee that the whole functional network can been recovered. This partly explains the poor reproducibility of group analyses across datasets [Thirion et al., 2007a].

Assumption of perfect match between individual brains. Due to unavoidable inter-subject registration errors (see Section 2.3.1), the observed activations are not well-localized, and possibly displaced across distinct functional regions, which may result in blurring the group activation map and creating unhandled false positives [Keller et al., 2008].

2.6 Alternative approaches

Many approaches have been developed to overcome the limitations identified in the previous section, which we review now. Research on this subject has been conducted in different directions, which we have chosen to identify as follows:

To start with, many efforts have been devoted to develop new thresholding techniques to address the multiple comparison problem, both at the voxel and cluster level, and are discussed in Section 2.6.1. Their goal is to overcome the limitations of the standard MCPs, such as the exclusive control over false positives, or the dependence on a cluster-forming threshold.

Next, to define potentially active regions, as a natural alternative to suprathreshold clusters one may use pre-defined regions of interest (ROIs), related to the question to be answered by the fMRI experiment. For instance, ROIs may be defined as anatomical structures that one expects to be functionally involved in a certain target task. This strategy avoids the problematic choice of a cluster-forming threshold, and provides a way to incorporate prior informations on the regions involved in the task at hand. However, the use of such fixed regions poses several challenges, which we describe in Section 2.6.2.

An increasingly popular approach, termed surface-based analysis, aims at improving the localization of each subject's activity. It consists in projecting for each subject the fMRI data on the cortical surface, segmented from the anatomical images. Spatial normalisation is also performed on these surfaces rather than on the whole brain volume. We describe several procedures for surface-based analysis, and discuss their pro's and con's in Section 2.6.3.

Finally, so-called *feature-based* approaches represent a very fruitful line of research. They consist in extracting from the individual data certain high-level features, such as local maxima of an activation map, and comparing them across subjects. By performing the group analysis at a higher level than the voxel-level, such approaches naturally reduce the multiple comparison problem, and also provide a way to deal with spatial uncertainty. Several methods are presented in Section 2.6.4.

2.6.1 Thresholding techniques

We now review a few papers which are representative of the directions which have been explored to ameliorate standard voxel and cluster-level thresholding techniques.

Joint control over false positive and false negative risks

Voxel-level Bayesian inference Voxel-level inference methods have been proposed that control both false positive and false negative rates. In [Friston et al., 2002b, Friston et al., 2002a], the usual voxelwise test is extended using Bayesian inference. Thus, the voxelwise p-values p_k are replaced by the posterior probabilities q_k that the mean group effect μ_k is larger than a certain baseline b. The method implemented in [Friston et al., 2002b] has several drawbacks however. First, the meaning of the baseline b is unclear, and so is its choice. Second, the Bayesian inference proposed in [Friston et al., 2002b] is solely aimed at voxel detection. In particular, it provides no equivalent to the cluster-level test (see Section 2.4.2) which is the current standard in fMRI data analysis.

Third, the multiple comparison problem is not addressed. This seems due to a confusion between Bayesian inference and multiple testing. Indeed, it is stated in the abstract that in contrast to 'conventional SPMs' (*i.e.*, the p-value maps $(p_1, \ldots, p_d)$), the posterior probability maps (PPMs) $(q_1, \ldots, q_d)$ 'are not confounded by the multiple comparison

problem'. However, both SPMs and PPMs address the same problem of selecting active voxels, while jointly limiting the number of false positives and false negatives. The fact that these are measured by frequentist rates (in the case of SPMs) or Bayesian risks (in the case of PPMs) does not change the fact that increasing the number of tested hypotheses mechanically increases the risk of detecting activations by chance. Thus, some form of correction for multiple comparisons must be applied in both cases.

These issues may explain why this approach is scarcely used in practice.

Mixture modeling An alternative proposed in [Beckmann et al., 2003b] is to fit a spatial mixture model to the voxel-wise test statistics, where null, activated and deactivated (inhibited) regions are modeled separately. A Gaussian distribution is used for null, or inactivated, voxels, a Gamma distribution for activated voxels, and a negative Gamma distribution for deactivated voxels. The same model has also been proposed in the context of individual subject fMRI data analysis, to model neural response levels (see Section 2.2.3). A threshold can then be derived, which minimizes a certain classification risk, such as the binary risk, associated to the 0-1 loss function, resulting in a 'naive Bayes' classifier.

This approach is revisited in [Woolrich et al., 2005, Woolrich and Behrens, 2006], where the status of each voxel is further modeled according to a Markov random field. This allows to account for the spatial structure of fMRI activation patterns, where active voxels tend to be grouped in clusters rather than isolated. The same model has also been used for the modeling of fMRI time-series, as mentioned in Section 2.2.3. Active voxels are detected using Bayesian inference, by computing a map of posterior probabilities that each voxel is active, and selecting the most probable state for each voxel, as in [Beckmann et al., 2003b].

Threshold-free cluster enhancement

In the context of cluster-level inference, dependence on the cluster-forming threshold is addressed in [Smith and Nichols, 2009], where an original approach is developed, termed threshold-free cluster enhancement (TFCE). It consists in applying a filter to the statistical map, that has the effect of enhancing the amplitude of extended signals, while

reducing less extended ones. This reduces the dependence on the cluster-defining threshold, while at the same time increasing SNR.

It must be noted though that the enhancement filter is defined by two parameters, noted E and H, which need to be tuned by the user, in exchange to the lessened dependence on the cluster-forming threshold. The choice of E and H is however extensively studied in [Smith and Nichols, 2009], and the robustness of the approach with respect to this tuning is demonstrated in practice over a wide range of datasets.

When used in association with a multiple testing procedure, another potential shortcoming of TFCE is that the resulting statistical map does not verify subset pivotality (see Appendix A.1 for a precise definition). In other terms, the value of the enhanced statistic in each voxel is not independent from the statistics in other voxels. As a consequence, only weak control is guaranteed for the resulting procedure, but not strong control (see Appendix A for formal definitions of weak and strong control, and justifications of these assertions). This means that the absence of bias due to TFCE, though supported by simulations, has yet to be demonstrated mathematically.

Alternative choices of cluster-level summary statistic

Finally, the choice of a decision statistic for testing cluster-level hypotheses remains a point of debate. The widespread use of the cluster size has often been criticized, as it naturally favors spatially extended signals, and may result in missing small activations, irrespective of their intensity. Thus, several approaches have been developed to combine cluster statistics. In [Poline et al., 1997] the cluster size and cluster peak (maximum statistic value) are combined through multivariate rejection regions. In [Hayasaka and Nichols, 2004], a wider panel of possible cluster summary statistics are compared on several datasets, using combining functions. Though promising, these approaches have not been shown to produce significantly different results from the conventional cluster size test. The core problem seems to be that each cluster summary statistic defines a test that is optimally powerful for a certain type of signal (extended, peaked, etc.). Combining different statistics does not produce a more powerful test, but rather changes the type of signal that is detected.

To date, there is to our knowledge no cluster-level inference approach that controls false negative clusters, *i.e.*, controls the risk of missing an activated cluster.

2.6.2 ROI-Based Analysis

ROIs in fMRI single-subject data analysis

The use of regions of Interest (ROIs) to segment the brain volume is common in single-subject fMRI data analysis [Poldrack, 2007]. They are often defined as anatomical structures, extracted from the T1-weighted image of the subject under study, or as clusters detected as active from a previous fMRI experiment. fMRI paradigms specifically designed to define ROIs are known as *localizers*. They usually consist of simple stimuli, meant to elicit a response from a known functional region, in order to identify its location for the subject under study.

In both cases, ROIs are an attractive alternative to suprathreshold clusters for the definition of candidate activated regions; they avoid the troublesome choice of cluster-defining threshold, and provide a way to include prior information on the regions involved in the investigated task. When defined anatomically, they also provide a statistically sound way to assess the link between brain structure and function, which is one of the fundamental questions fMRI studies aspire to answer.

ROIs in fMRI group data analysis

In spite of these advantages, ROIs are scarcely used as a basis for inference in the context of fMRI group data analysis. Rather, clusters detected as active are empirically compared to anatomical ROIs, usually in the form of an atlas of brain regions, such as the single subject Automated Atlas Labels (AAL) [Tzourio-Mazoyer et al., 2002], in a post-treatment step, as illustrated in Figure 2.12. This raises several questions, because active voxels cannot be localized within an active cluster (see Section 2.4.2). Thus, when a cluster extends over several atlas regions, there is no guarantee that any of these regions contain any active voxels. Moreover, there is no single way of performing this comparison, so several heuristics are traditionally used, with potentially different answers.

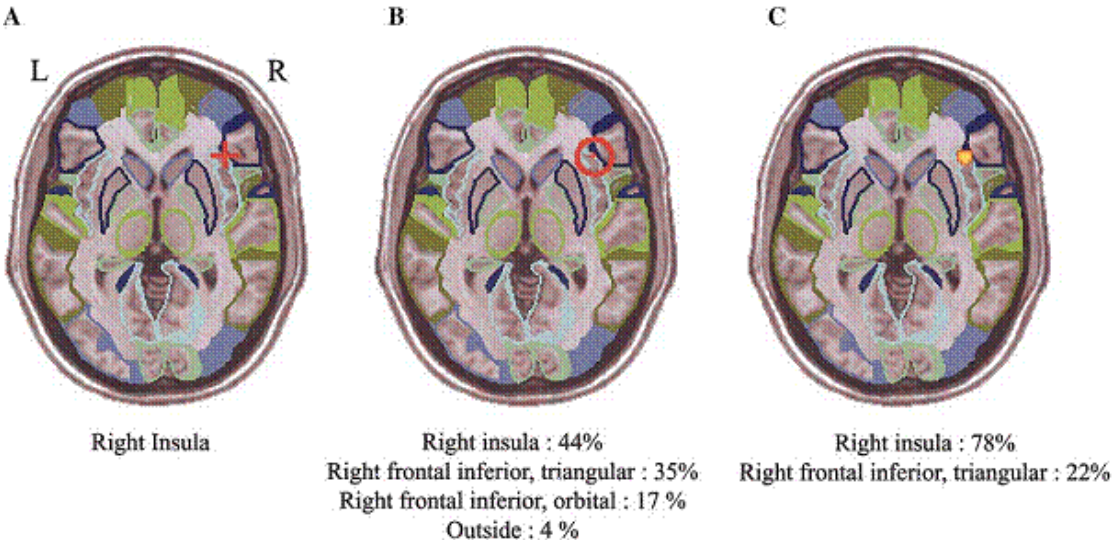


FIGURE 2.12: Three procedures to perform the automated anatomical labeling of functional activations, proposed in [Tzourio-Mazoyer et al., 2002] (the outline of the AAL parcellation is overlaid on the z=5.0 mm axial slice of the single-subject anatomical image used to define it): (A) Local maxima labeling: the red cross indicates the location of the local maximum. (B) Extended local maxima labeling: percentage of overlap between a 10-mm sphere radius centered on the local maximum and the different parcels. (C) Cluster labeling: percentage of overlap between the activation cluster and the different parcels.

In fact, it seems rather surprising to us that, after all the efforts devoted to develop statistically well-defined activation detection procedures, the crucial step wherein these activations are identified with anatomical structures is performed using post-hoc methods, producing results which cannot be validated or refuted on an objective basis.

A notable exception is found in [Bowman et al., 2008], which uses AAL to divide the brain volume in regions assumed to represent distinct functional units. Based on this parcellation, a Bayesian hierarchical approach is developed to identify regions whose functional responses are statistically correlated, a phenomenon known as *functional connectivity*. Thus, parameters at the regional level are modeled as a Gaussian vector with arbitrary covariance structure. By estimating the covariance matrix, functional connectivity may be assessed between any couple of ROIs, for the specific task investigated. Moreover, this framework could also be used to test regionwise hypotheses, as in the SPM-like approach, even though this is not done in [Bowman et al., 2008].

However, the use of ROIs for fMRI group analysis poses a series of challenges, which may explain why they have not yet been adopted as a standard tool. As noted in [Poldrack, 2007], a major issue is the important variability of anatomical structures accross subjects, only crudely compensated by the normalisation step (see Section 2.3.1).

Thus, anatomical ROIs are hard to define at the group level, and individual images are never aligned with any given set of ROIs. This may result in false positives, since activations are likely to be displaced accross regions.

This point of view is developed in [Nieto-Castanon et al., 2003]. It is shown that the overlap of individual anatomical ROIs across subjects vanishes quickly as the group size increases (see 2.13). This observation is used to motivate an alternative analysis of multi-subject neuroimaging data, based on individual ROIs rather than an anatomical atlas. Briefly summarized, the functional data is analyzed separately for each subject-specific ROI. For each ROI, the corresponding summary statistics are compared across subjects. Note that this approach does not require inter-subject registration, since each subject's functional data is analyzed in reference to its own anatomy.

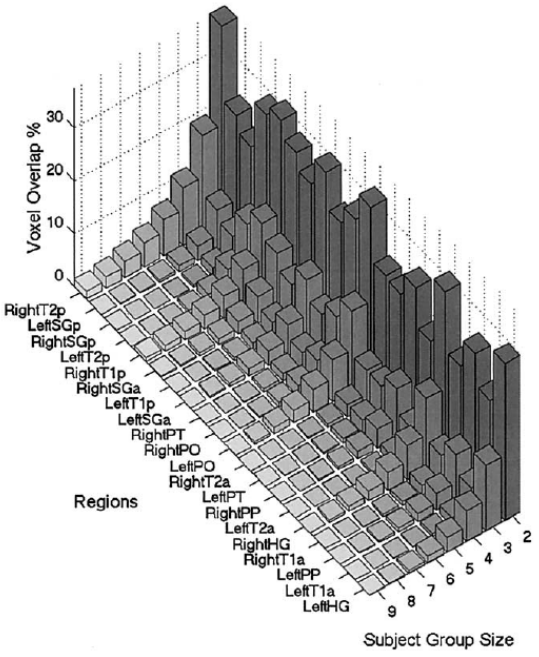


FIGURE 2.13: Mean overlap of individual anatomical ROIs across different subject group sizes in [Nieto-Castanon et al., 2003].

However, as discussed in [Poldrack, 2007], measuring the functional signal of a single subject within each given ROI is a challenging task. For each subject, there may be only a small proportion of voxels activated inside a given region, so that averaging the signal, as seems to be done in [Nieto-Castanon et al., 2003], may swamp the signal with the noise from the many inactive voxels. In contrast, standard mass-univariate methods gain statistical power by averaging the data across subjects in each voxel. Thus, they

may be able to recover activations which would otherwise be lost at the single-subject level.

Another issue in terms of group modeling is that individual anatomical structures may not be present in all subjects. This must be taken into account if one wants to perform a RFX analysis, *i.e.* generalize the findings on the cohort of subjects under study to their parent population.

These shortcomings may explain why this methodology has only been demonstrated so far in a fixed-effect (FFX) setting. Though [Nieto-Castanon et al., 2003] claim that random-effect (RFX) analyses would yield 'similar results', there may be both a sensitivity and a modeling issue in the generalization of this approach.

2.6.3 Surface-based analysis

Motivation

The SPM-like approach searches for BOLD activations throughout the whole brain volume. However, neurons are mainly concentrated near the cortical surface (though they are also present in subcortical structures). Moreover, there is evidence supporting the assertion that cortical landmarks, such as sulci and gyri, correspond to boundaries between functionally distinct regions [Brett et al., 2002].

Consequently, volume-based detection may be sub-optimally sensitive because statistical power is wasted by searching for activations in the wrong places, such as the white matter. Also, volume-based registration may fail to match cortical structures, which may in part explain the bad overlap between homologous functional regions. Moreover, distinct functional areas which are well separated along the cortical surface may be close in terms of the euclidean distance, due to the highly convoluted topology of the cortical folds. Hence, They may be hard to separate in a fMRI contrast map, whose lower spatial resolution (3mm^2 against up to 1mm^2) cannot account for such fine details.

Based on these observations, the analysis of functional neuroimaging data based on the cortical surface rather than the entire brain volume has received an increasing attention over the last decade. The goal of such surface-based techniques is to obtain a more precise localization of individual functional areas, and improve their matching across subjects,

based on the assumption that functional regions are better identified on the cortical surface by the associated anatomical landmarks than in the 3D Talairach coordinate system.

Surface-based registration

A method for surface-based registration is presented in [Fischl et al., 1999]. For each individual, the procedure starts by reconstructing the cortical surface from a structural MRI image. This surface is then inflated, and transformed into a sphere using a registration algorithm that minimizes metric distortions, as illustrated in Figure 2.14. In this fashion, the folding pattern of the subject's cortex is mapped using a polar coordinate system, so that to each point corresponds a measure C of curvature, with negative and positive values indicating gyral and sulcal regions respectively. This folding pattern is then aligned with a canonical, average folding pattern (Figure 2.14, to the right).

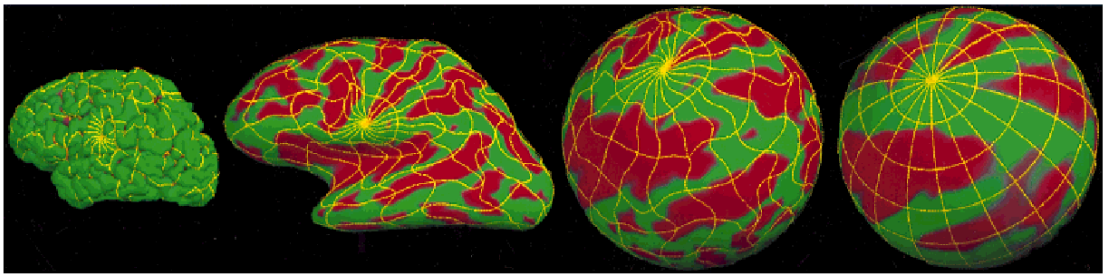


FIGURE 2.14: Unfolding of the cortical surface onto a sphere (in [Fischl et al., 1999]).

The advantages of this approach with respect to volume-based registration are demonstrated in [Fischl et al., 1999]: The overlap of cortical structures, such as the central sulcus, are greatly improved, and sensitivity in the detection of fMRI activation is observed. However, no clue is provided regarding how the fMRI data is projected on the cortical surface, though this step is bound to be a challenge in this type of analysis.

A similar approach is developed in [Essen, 2005], involving a flattened representation of the cortical surface, which is mapped onto a sphere and matched to an atlas of the cortical folding pattern. However, a more sophisticated atlas is constructed here, obtained by matching anatomical landmarks extracted from each individual's cortex. As in [Fischl et al., 1999], this approach significantly improves the alignment of cortical structues (see Figure 2.15). The issue of mapping fMRI data to the cortical surface is addressed by simply assigning to each node of the mesh representing the cortex the voxel

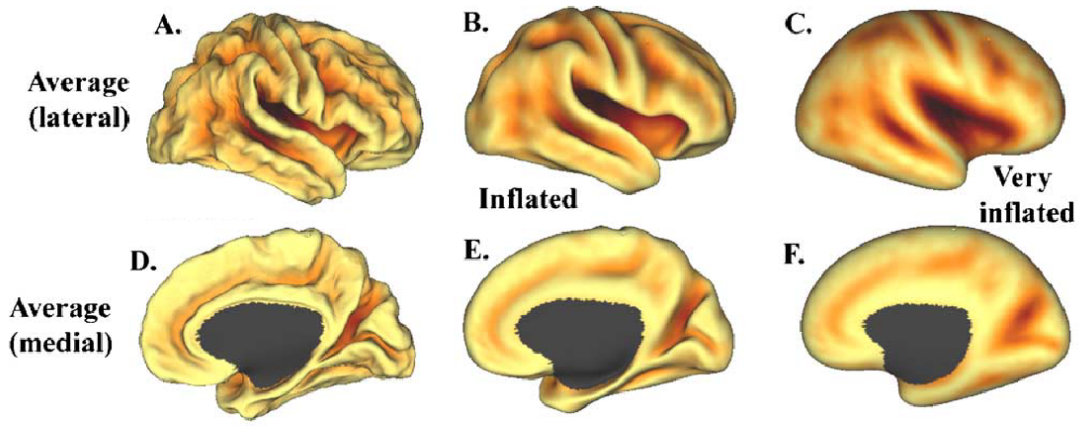


FIGURE 2.15: Average cortical surfaces of 12 subjects, registered to the atlas in [Essen, 2005]. A – C. Lateral views of the average surface, both with and without inflation. D – F. Similar to A – C, but showing the medial views. The average surface preserves the main features of the sulcal anatomy, showing a clear improvement over volume-based registration using affine transformations (see Section 2.3.1).

containing it, though [Essen, 2005] mentions 7 (at least) other algorithms that could be used instead.

Projection of fMRI data onto the cortical surface

Mapping the fMRI data onto the cortical surface thus seems to be one of the main difficulties of surface-based analysis. Indeed, because of the low resolution of fMRI data, it is impossible to assign each voxel to a particular point of the cortical surface. In practice, no reference projection method exists, and the choice of a particular strategy may influence the results. A recent development on this subject can be found in [Operto et al., 2008a], wherein fMRI data is projected onto the cortical surface, using interpolation kernels informed by the subject's anatomy. This approach is shown to be more robust to registration errors than other, less sophisticated approaches. However, the authors agree that matching of the fMRI data volume with the cortical ribbon remains largely an open issue, due mainly to the low resolution of current functional images, and the ensuing partial volume effects.

In conclusion, surface-based analysis of fMRI data seems a promising alternative to volume-based approaches, by providing a better match between the cortical structures of different subjects, and consequently increasing the overlap of homologous functional regions. However, this increased anatomical precision seems to come at the cost of a degraded match between functional and anatomical data. Thus, in both cases the

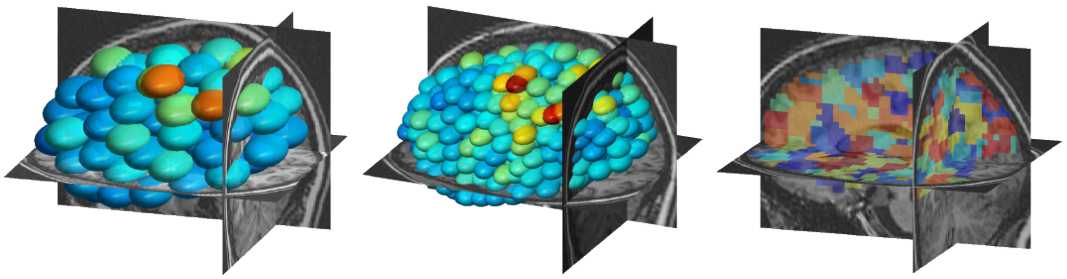


FIGURE 2.16: Examples of anatomo-functional parcellations for a grasping task in [Flandin, 2004]. From left to right: mixture model with 100 Gaussian components, 500 components, corresponding maximum *a posteriori* parcellation.

uncertainty on the localization of individual activations persists, and should be taken into account when conducting inference at the group level.

2.6.4 Feature-based approaches

As mentioned in Section 2.6.3, the SPM-like approach detects activations throughout the whole brain volume, by comparing the individual images in a voxelwise fashion. The motivation behind surface-based analysis is that voxels are not relevant entities from a physiological point of view.

This criticism of voxel-based comparisons leads to a broader class of methods, which we refer to as *feature-based*. The principle they rely on consists in extracting high-level features from the individual data, and matching them across subjects. By choosing relevant features that summarize the information contained in the subject's data, these approaches may be expected to provide results that are easier to interpret, and gain statistical power by reducing the number of multiple comparison. For instance, the ROI-based analysis of group fMRI data developed in [Nieto-Castanon et al., 2003] is feature-based, the features being the summary statistics measuring the BOLD response for each subject-specific ROI.

The following examples have been chosen to illustrate the diversity of feature-based approaches. These methods are best classified according to the choice of a feature set.

Functionally homogeneous parcels

The idea of parcelling the brain volume into parcels that are homogeneous both anatomically and functionally is investigated in [Flandin, 2004]. A first approach is developed, based on the anatomical data of a single subject, to compute a Voronoi tesselation of the search volume into K connected components. These tesselations are then used to reduce the dimensionality of the subject's fMRI data, by averaging the time courses (see Section 2.2.3) within each parcel. The K average time-courses are then used as input for the SPM-like approach, instead of the d original time-courses.

In particular, the presence of activation is tested at the parcel rather than the voxel level, using the methodology presented in Section 2.4.1. As a consequence, the multiple comparison problem is reduced, and the greater sensitivity of this approach is demonstrated compared to the standard voxelwise detection procedure, even though the data is spatially smoothed in the last case, another way of reducing the data's resolution.

Going one step further, [Flandin, 2004] then generalizes this approach by introducing a parcellation into clusters containing voxels with similar BOLD responses, based on both the anatomical and functional MRI data of the subject under study. Thus, in addition to reducing the data's dimension, the proposed procedure also defines units that are relevant in terms of neuroscience, since they may be interpreted as functionally homogeneous areas.

Finally, this approach is extended to the context of fMRI group data analysis in the following way. The individual images are first normalized to a given template, to ensure proximity between homologous functional areas across subjects (perfect alignement of the images is not assumed however). Then, the fMRI time-courses of the different subjects are concatenated and analyzed as if they belonged to a single subject, using the anatomo-functional parcellation described above. This creates group-level clusters regrouping voxels with similar fMRI time-courses. A nice feature of this approach is that it implicitly defines clusters at the subject level that are automatically matched, clusters from two distinct subjects being homologous if they belong to the same group-level clusters.

This approach provides an elegant way to deal with inter-subject anatomical variability, but has several limitations. Indeed, because it includes no inter-subject model,

this approach can only characterize the group of subjects under study, *i.e.* perform a fixed-effect analysis. Additionally, single-subject clusters are not necessarily spatially connected, making them more difficult to interpret as functional units. Moreover, there is a risk that clusters defined according to some measure of fMRI time-course similarity may be influenced by the main effect studied in the experiment, which in general will dominate the fMRI signal. Hence, such a clustering may not work well when analyzing more subtle effects (such as interactions between different experimental factors).

These shortcomings are addressed in [Thirion et al., 2006c], which revisits the approach in [Flandin, 2004] by adding a group level to the model defining the subject parcels. According to this model, subject parcels are no longer forced to be subsets of group-level parcels, or *cliques*. The relaxation of this constraint allows the definition of spatially connected parcels, such that all subjects are represented in each clique. Furthermore, the data is clustered according to the vector of estimated effects rather than the raw fMRI time-courses, which means that it is no more dominated by the main effect.

As in [Flandin, 2004], presence of activity is tested for each clique C, by averaging for each subject i the map of estimated effects over the subject-cluster homologous to C and performing a t-test on the resulting values. Such an approach is shown to outperform conventional voxelwise tests. There is some concern however about the control this approach offers on false positives. Indeed, because subject clusters and cliques are estimated from the whole dataset, they cannot be considered spatially independent. Thus, there is no guarantee that the vector of clique-level t-statistics verifies the subset pivotality, hence that this procedure has strong control over the false positive risk considered.

In a recent development based on similar ideas [Tucholka et al., 2008a], the same type of multi-subject anatomo-functional parcellations is derived, combined to a surface-based approach (see Section 2.6.3). More precisely, A first segmentation of the cortical surface into gyri defines distinct anatomical regions. Each region is then segmented into a certain number of cliques, represented for each subject by a parcellation into an equal number of clusters. The total number of cliques is optimized using a cross-validation criterion, providing an insight into the number of functionally distinct regions involved in the fMRI paradigm under study.

Spatial mixture modeling

Following the same idea of grouping voxels with similar responses that is at the core of clustering approaches, several mixture models have been proposed for the statistical analysis of fMRI data, with the specific goal of detecting activated regions. This idea was first pursued in [Vaever Hartvig, 2000], which considered representing the BOLD activation pattern as a superposition of Gaussian shaped 'bumps', in addition to a uniform background level. The number of bumps was determined automatically, using a stochastic geometry model. More recently, [Penny and Friston, 2003] defined a mixture model to represent the map of BOLD effects, each component corresponding to an activated cluster, modeled using a similar Gaussian shaped response surface, plus an additive noise. This idea was further extended in [Kim and Smyth, 2006], which used an infinite Bayesian mixture model, based on a Dirichlet process prior, to deal with the problem of selecting the number of components.

The above works are all concerned with single-subject fMRI data analysis, and use a parametric surface response to model activation patterns. In a recent work, [Xu et al., 2009] uses a similar Bayesian spatial mixture modeling approach for fMRI group data analysis, using different levels of hierarchy to model individual activation centers and population-level activation centers. The spatial variability of individual centers around the population centers is explicitly modeled, allowing to account for mis-registrations. An original feature of this work is that it adopts a Bayesian model averaging point of view, integrating out the number of mixture components, rather then estimating it and characterizing each cluster. Model averaging is implemented through a reversible-jump MCMC algorithm.

Critical points of the individual activation maps

While in the previous approaches the individual features are defined, and possibly estimated, as latent variables of a certain model describing the data, it is also possible to directly extract them as *e.g.* salient features of the activation maps of the individual subjects, and build a group model from these. This is exactly what is done in [Thirion et al., 2007b], where features are defined as the critical points of the individual activation maps, such as local maxima and minima, and the associated level sets. These

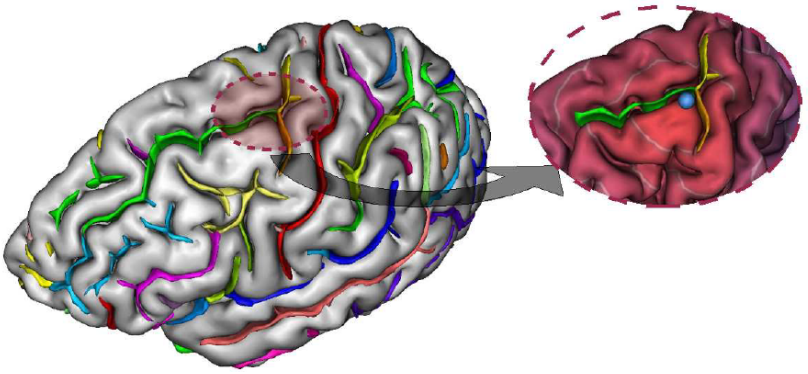


FIGURE 2.17: Within labelled sulci framework, the activation landmark of computation (blue ball on the picture) in the prefrontal lobe is localized near the intersection of the three sulci: superior precentral, marginal precentral and superior frontal (from [Tucholka et al., 2008b]).

are used to represent individual activation centers, and are modeled using a Bayesian spatial mixture model similar to those described above. In particular, a Dirichlet process prior is used to define an infinite mixture model to deal with the estimation of the number of classes, as in [Kim and Smyth, 2006]. Each class defines a separate group-level activity cluster, and the set of such clusters provides a description of the activation pattern at the population level. These group-level clusters are shown to be more reproducible across datasets than standard clusters obtained with the standard SPM-like analysis.

Along similar lines, [Tucholka et al., 2008b] propose an original approach to localize the local maxima extracted from the individual activation maps of a group of subjects with known anatomical landmarks, using triangulation techniques. The local maxima are first matched across subjects, and viewed as functional landmarks. Then, for each subject, the distance of each landmark to three user-chosen neighbouring sulci (see Figure 2.17) is computed, then averaged over subjects. Given a new subject, the position of the homologous functional landmark is then predicted by triangulation, based on these average distances. This is compared with the position predicted by the average coordinates in the standard space, and shown to be much more accurate. This nice result shows that brain anatomy and function are linked, and suggests that the standard coordinate system may not be optimal to localize functional areas.

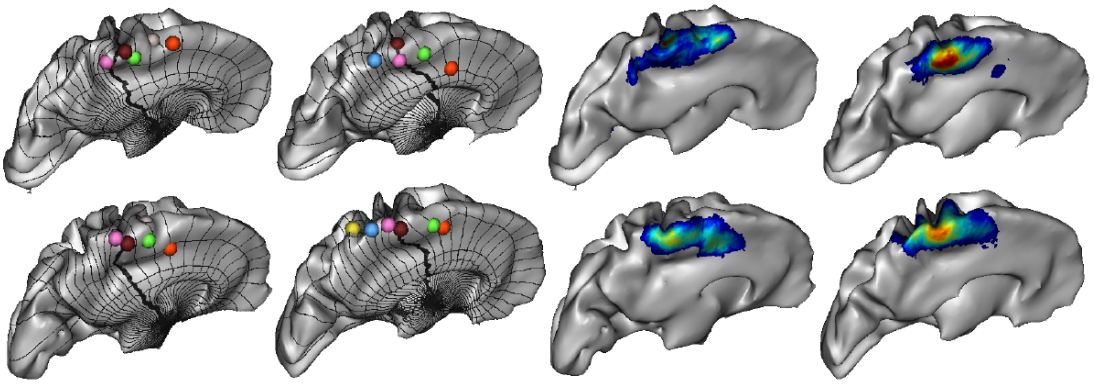


FIGURE 2.18: Reproducible activations from a right-motor contrast for 4 subjects, in [Operto et al., 2008b]. (left) labeled after maximization (right) thresholded individual statistical t-maps ($p < 0.05$ corrected).

Scale-space blobs

We conclude this review of feature-based approaches with the fMRI group data analysis method proposed initially in [Coulon et al., 2001] for the analysis of 3D activation maps and revisited in [Operto et al., 2008b] in the setting of surface-based analysis. In the latter, the individual fMRI data are first projected on the cortical surface, as explained and for the reasons exposed in Section 2.6.3. The critical points are then extracted from the individual t-maps as in [Thirion et al., 2007b, Tucholka et al., 2008b]. The key point of this approach however is the scale-space representation of these features it then constructs. This adds to the spatial structure linking the critical points another level of hierarchy, obtained by considering different levels of spatial smoothing. This description allows to account for patterns that appear at different levels of resolution, based on the idea that objects of interest happen at many different scales.

This complex structure is coded for each subject by a graph, whose nodes are referred to as *scale space blobs*, and represent the features of interest. Group analysis is performed by embedding these graphs in an encompassing graph, and matching the blobs across subjects. The optimal matching is found by optimizing an objective function which accounts for the proximity of each pair of matched blobs, and the likelihood that these blobs correspond to an active brain area. This is performed in a Bayesian framework, the field of labels coding matches between blobs being modeled as hidden Markov field, that is estimated by *a posteriori* maximization.

The final representation of the functional reponse of the group of subjects consists in a list of labels representing activations that are reproducible across subjects, and for each subject a set of scale-space blobs representing instances of these activations, as illustrated in Figure 2.18.

As in [Flandin, 2004] and [Thirion et al., 2006a], this creates a 'structural' description of the group of subjects, with informations that may be exploited both at the group and at the subject level. Results are however not generalizable to the general population, so that this approach can be categorized as performing a fixed effect analysis.

2.7 Conclusion

The review we have just presented, though far from exhaustive, suggests the wide variety of approaches that have been developed for the statistical analysis of fMRI data during the last decades, to address the limitations of the 'SPM-like' approach identified in Section 2.5: dependence on an arbitrary cluster-defining threshold, exclusive control of false positives, and the assumption of a perfect match between the different brains. However, to date these issues have been addresses separately.

The goal of the present work is to propose a new procedure for fMRI group data analysis, addressing them jointly. To position ourselves with respect to the existing methods, we start by discussing their similarities and divergences. We have found three key points that summarize this comparison:

- First, all the described approaches agree on one point, namely that group inference should be performed at a higher level than the single voxel. To quote [Operto et al., 2008b]: "The voxels are only the acquisition space, and have never had any anatomical meaning other than the simple localization provided by spatial normalization."

  Thus, the group activation pattern is always described in terms of higher-level features, such as suprathreshold clusters, ROIs, group activation centers, parcels or multiscale blobs.

- Secondly, most of the methods presented here rely on preliminary spatial normalization to align subjects, using either volume or surface-based registration.

A notable exception is [Nieto-Castanon et al., 2003], which relies solely on the anatomical structures extracted in each subject to compare them. It is generally accepted that perfect alignment of the anatomies is never attained in practice (see Section 2.3.1). In spite of this, to date feature-based approaches have been the only ones to account for the resulting spatial variability of the activation patterns, through explicit modeling.

- Finally, apart from the ROI analysis framework presented in Section 2.6.2, none of the proposed approaches provide an assessment of the link between detected activations and anatomical brain structures, though this question is at the core of most fMRI studies. An exception is [Tucholka et al., 2008b], which shows that activation foci are better localized with respect to neighbouring sulci than than in the standard coordinate system (see Section 2.6.4).

Based on our review and these final remarks, it seems to us that the ROI framework has a certain number of key advantages, in relation to the issues we want to address. It is a simple and flexible tool that allows to define regions of potential activity at the group level based on expert knowledge rather than an arbitrary threshold. Furthermore, joint control on false positive and false negative risks is not only possible, but conceptually simple in this setting, by adopting a Bayesian framework. Moreover, as shown in [Bowman et al., 2008], it also provides a means of investigating the functional connectivity of distant brain regions.

Feature-based approaches also constitute an appealing starting point for our work, as they have demonstrated their ability to account for imperfect matches between subjects. However, we feel that using a high-level description of the single subjects estimated effects maps would not necessarily help us attain our goal. The main reason for that is that we are not interested in describing each subject, but rather the functional network involved in the task at hand, in a reproducible way across subjects. Furthermore, we have no idea how to define these features, since there is an infinity of possible choices, that may all be at the expense of discarding valuable information. Thus it appears to us that adopting a feature-based approach in our case would only add un-necessary complexity.

Thus, we propose to extend the current state-of-the-art ROI analysis methodology, by relaxing the assumption that the individual images are in perfect match on a voxelwise

basis. This can be done by explicitly modeling spatial variability, as is common in feature-based procedures, but this time at the voxel-level.

Chapter 3

Random thresholds for linear model selection, revisited. Application to fMRI data analysis.

Abstract

In [Lavielle and Ludeña, 2007], a random thresholding method is introduced to select the significant, or non null, mean terms among a collection of independent random variables, and applied to the problem of recovering the significant coefficients in non ordered model selection. We introduce a simple modification which removes the dependency of the proposed estimator on a window parameter while maintaining its asymptotic properties. A simulation study suggests that the modified estimator performs better at low signal to noise ratios, where the original one is unstable with respect to the window parameter. An application of the method to the problem of activation detection on functional magnetic resonance imaging (fMRI) data is discussed.

3.1 Introduction

A popular approach to activation detection in fMRI data analysis consists in thresholding a statistical map of brain activity [Friston, 1997].

The choice of a detection threshold is usually addressed by controlling at a user-fixed level a certain type I error rate, such as the family wise error rate (FWER) [Nichols and Hayasaka, 2003] or the false discovery rate (FDR) [Pacifico et al., 2004, Benjamini and Hochberg, 1995]. It can be argued however that the choice of a level, which ultimately defines the detected regions, is arbitrary, as their is no safe guideline to what an 'optimal' level of false detections should be.

We consider here as an alternative the random threshold method recently developed in [Lavielle and Ludeña, 2007], which consists in estimating the number of non null mean terms among a collection of independent random variables. It is based on a random centering of the partial sums of the ordered observations. Consistency of the proposed estimator can be shown, using L-statistics techniques.

Though requiring no prior tuning of a type I error rate, this method still depends on a user-chosen window width, and has so far been demonstrated on high signal to noise (SNR) simulated data only, whereas fMRI data is known to be very noisy. This paper describes a simple modification of the procedure which makes it totally unsupervised; the modification simply consists in replacing the fixed window width by a varying one.

The article is organized as follows: In Section 3.2, we review the random threshold method in [Lavielle and Ludeña, 2007], and introduce the varying window extension. In Section 3.3, a simulation study compares the random threshold procedure to the standard type I error rate control methods described above. Application to fMRI data analysis is discussed in Section 3.4.

3.2 Method

Assume we observe $y_i = \mu_i + \epsilon_i$, where variables ϵ_i are centered, independent and identically distributed with common cumulative distribution function (cdf) F_ϵ.

3.2.1 Original random threshold procedure

Testing the presence of significant coefficients

We start by recalling the procedure for testing if all the μ_i's are null or not, given the observations y_i; $1 \leq i \leq n$, that is, testing the null hypothesis $\mathbf{H_0} : \mu_i \equiv 0$ for $i = 1, \ldots, n$ against the alternative $\mathbf{H_1} : \exists I \subseteq \{1, \ldots, n\} \mid \forall i \in I, \mu_i > 0$. The procedure is defined as follows :

i) Order the observations $|y_{(1)}| \geq |y_{(2)}| \geq \ldots \geq |y_{(n)}|$.

ii) For $i = 1, \ldots, n$, let $X_{(i)} = -\log(1 - F_{|\epsilon|}(|y_{(i)}|))$.

iii) Let $T_j = \sum_{i=1}^{n} X_{(i)}$ and $Q_j = \mathbb{E}_{\mathbf{H_0}}(T_j|T_n)$.

iv) Define the test statistic $D_n = \max_j |T_j - Q_j|/\sqrt{n}$. The null hypothesis is rejected if $D_n > d_\alpha$, where d_α ensures that the level of the test is at most α.

The conditional expectation $\mathbb{E}_{\mathbf{H_0}}(T_j|T_n)$ is easily computed due to the fact that under the null hypothesis, $(X_{(i)})_{1 \leq i \leq n}$ is an ordered sequence of exponential random variables, and to the following result, whose proof is omitted here:

Proposition 3.1. *Assume $X_{(1)}, \ldots, X_{(n)}$ is an ordered sequence of Exp(1) random variables, with $X_{(1)} \geq \ldots \geq X_{(n)}$. For any $1 \leq j \leq n$, let $T_j = \sum_{i=1}^{n} X_{(i)}$. Then, for any $1 \leq j \leq K \leq n$:*

$$\begin{aligned}
\mathbb{E}_{\mathbf{H_0}}(X_{(i)}) &= \sum_{\ell=i}^{n} \frac{1}{\ell} \\
\mathbb{E}_{\mathbf{H_0}}(T_j) &= j + j \sum_{\ell=j}^{n} \frac{1}{\ell} \\
\mathbb{E}_{\mathbf{H_0}}(T_j|T_n) &= \frac{\mathbb{E}_{\mathbf{H_0}}(T_j)}{\mathbb{E}_{\mathbf{H_0}}(T_n)} T_n.
\end{aligned}$$

Selecting the significant coefficients

Upon rejection of the null hypothesis, the following task consists in selecting the significant coefficients. The procedure for doing so can be interpreted in a data-dependent

'multiple hypothesis testing' setting, as described hereafter. Consider the null hypothesis $\mathbf{H_0}$ as defined in Section 3.2.1, and the set of alternative hypotheses:

$$\mathbf{H_1(k)} : \text{for any } i \leq k,\ \mu_{(i)} > 0, \text{ and } \mu_{(k+1)} = \ldots = \mu_{(n)} = 0.$$

Denote $\mathbb{E}_k$ the expectation under $\mathbf{H_1(k)}$ (instead of $\mathbb{E}_{\mathbf{H_1(k)}}$). The procedure first computes the $X_{(i)}$'s using the same steps i) and ii) as in Section 3.2.1, then adds the following steps:

iii) Let K_n be some positive integer. For $1 \leq k \leq n - K_n$ and $1 \leq j \leq K_n$, compute:

$$\begin{aligned} T_{k,j} &= \sum_{i=k+1}^{k+j} X_{(i)} \\ Q_{k,j} &= \mathbb{E}_k(T_{k,j}|T_{k,K_n}) \\ \eta_k &= \max_{1\leq j\leq K_n} |T_{k,j} - Q_{k,j}|/\sqrt{n}. \end{aligned}$$

iv) Let $\hat{k}_n = \text{argmin}_{1\leq k\leq K_n}\eta_k$.

As in the preceding section, $Q_{k,j}$ can easily be computed using Proposition 3.1. η_k can also be defined from the centered partial sums $(T_{k,j} - Q_{k,j})$ using the ℓ_p norm for $1 \leq p < \infty$ rather than the ℓ_∞ norm, by setting $\eta_k = n^{-p/2-1}\sum_{j=1}^{K_n} |T_{k,j} - Q_{k,j}|^p$.

Unknown distribution extension

The above procedure can be extended to the case where the distribution F_ϵ of the ϵ_i's is a parametric distribution $F_\epsilon(\cdot\,;\theta^\star)$, but where $\theta^\star$ is unknown.

For $0 \leq k \leq n-1$, let $\hat{\theta}_k = \hat{\theta}(y_{k+1}, \ldots, y_n)$ be an estimator of θ. Let $F_{|\epsilon|}(\cdot\,;\theta^\star)$ be the distribution of the $|\epsilon_i|$'s. For any $\theta \in \Theta$, let $X_i(\theta) = -\log\left(1 - F_{|\epsilon|}(|y_{(i)}\,;\theta|)\right)$ and $T_{k,j}(\theta) = \sum_{i=k+1}^{k+j} X_{(i)}(\theta)$. Then the following procedure is defined :

i) Let $K_n \leq [(1-b)n]$ be some positive integer. For $[a\,n] \leq k \leq n - K_n$:

1. let $\hat{\theta}_k = \hat{\theta}(y_{k+1}, \ldots, y_n)$,
2. for $i = 1, \ldots, n$, let $X_{(i)}(\hat{\theta}_k) = -\log\left(1 - F_{|\epsilon|}(|y_{(i)}|\,;\hat{\theta}_k)\right)$,

3. for $1 \leq j \leq K_n$, compute

$$\begin{aligned} T_{k,j}(\hat{\theta}_k) &= \sum_{i=k+1}^{k+j} X_{(i)}(\hat{\theta}_k) \\ Q_{k,j}(\hat{\theta}_k) &= \mathbb{E}_k(T_{k,j}(\hat{\theta}_k)|T_{k,K_n}(\hat{\theta}_k)) \\ \eta_k(\hat{\theta}_k) &= \max_{1\leq j\leq K_n} |T_{k,j}(\hat{\theta}_k) - Q_{k,j}(\hat{\theta}_k)|/\sqrt{n}. \end{aligned}$$

ii) Let $\hat{k}_n = \text{argmin}_{a\,n\leq k\leq b\,n}\eta_k(\hat{\theta}_k)$.

In the applications presented in Section 3.3 and Section 3.4, we consider Gaussian noise: $F_\epsilon(\cdot\,;\sigma^2) = \mathcal{N}(\cdot\,;0,\sigma^2)$ and estimate $\theta = \sigma^2$ by the usual mean squares estimator: $\hat{\theta}_k = \frac{1}{n-k}\sum_{i=k+1}^{n} y_{(i)}^2$.

3.2.2 Varying window extension

The procedures defined in the preceding sections depend on a parameter K_n which can be interpreted as a window width, since η_k is a function of $X_{(k+1)},\ldots,X_{(k+K_n)}$. Tuning this parameter may be difficult, as seen in Section 3.3. This issue can be avoided by re-defining η_k as a function of $X_{(k+1)},\ldots,X_{(n)}$, thus replacing the fixed width K_n by a varying width $n-k$, which requires no prior tuning. We define the following procedure:

iii) Let κ_n be a lower bound on the number of null coefficients. For $1 \leq k \leq n-\kappa_n$ and $1 \leq j \leq n-k$, compute:

$$\begin{aligned} T_{k,j} &= \sum_{i=k+1}^{k+j} X_{(i)} \\ Q_{k,j} &= \mathbb{E}_k(T_{k,j}|T_{k,n-k}) \\ \eta_k &= \max_{1\leq j\leq n-k} |T_{k,j} - Q_{k,j}|/\sqrt{n-k}. \end{aligned} \tag{3.1}$$

iv) Let $\hat{k}_n = \text{argmin}_{1\leq k\leq n-\kappa_n}\eta_k$.

In other terms, η_k would be strictly equal to the test statistic D_n defined in Section 3.2.1, if the sequence $(X_{(i)})_{1\leq i\leq n}$ where replaced by the subsequence $(X_{(i)})_{k+1\leq i\leq n}$, *i.e.*, the

null set of variables under $\mathbf{H_1(k)}$. As in the previous section, ℓ_p norms can be used instead of the ℓ_∞ norm by setting $\eta_k = (n-k)^{-p/2-1} \sum_{j=1}^{n-k} |T_{k,j} - Q_{k,j}|^p$.

Notice that $\hat{k}_n$ is independent of κ_n, as long as η reaches its global minimum on $\{1, \ldots, n-\kappa_n\}$. It is also immediate that the varying window extension, which we present here for the procedure in Section 3.2.1, can be applied to the unknown distribution extension in Section 3.2.1.

Asymptotic properties

The estimator of the number of significant coefficients presented in Section 3.2.1 is consistent. This is the main result in [Lavielle and Ludeña, 2007], and it can be extended to the varying window setting. We start by recalling the following asymptotic framework:

AF1 There exists $t^\star \in (0,1)$ and a subset $I_{k_n^\star}$ of $\{1, \ldots, n\}$, with $k_n^\star = [t^\star n]$ and $|I_{k_n^\star}| = k_n^\star$, such that $\mu_i \neq 0$ if $i \in I_{k_n^\star}$. For all other index, $\mu_i = 0$.

AF2 For any $i \in I_{k_n^\star}$, $|\mu_i| \geq \alpha_n$, for a certain sequence (α_n) with $\alpha_n \to \infty$ (see [Lavielle and Ludeña, 2007] for further details).

AF3 $\kappa_n / n \to c$ such that $0 < c < 1 - t^\star$.

We then have the following result:

Theorem 3.2. *Let $\hat{k}_n$ stand for the estimator defined in Section 3.2.2. Under assumptions* **AF1**, **AF2**, **AF3**, *and appropriate Von-Mises type conditions on the cdf F_ϵ of the errors (see [Lavielle and Ludeña, 2007] for further details), $\hat{k}_n$ is consistent in the sense that:*

$$P\left(\left| \frac{\hat{k}_n}{n} - t^\star \right| > u_n \right) \to 0, \tag{3.2}$$

for any positive decreasing sequence (u_n) such that $\sqrt{n} u_n \to \infty$.

This result can be refined by deriving an upper bound, which we do not detail here, on the convergence rate of the probability in Equation (3.2), for a particular choice of sequence (u_n). The proof of this Theorem is given in Appendix B.

Consistency also holds in the unknown distribution case (see Section 3.2.1). Consider the following assumptions on the cdf $F_{|\epsilon|}$ of the $|\epsilon_i|$'s:

F1. $F_{|\epsilon|}$ is two times differentiable as a function of θ with *a.e.* strictly positive derivative at $\theta = \theta^\star$.

F2. $\theta^\star$ belongs to some compact set Θ and there exists under $H_{k_n^\star}$ a consistent estimator $\hat{\theta_{k_n^\star}} = \hat{\theta}(y_{k_n^\star+1}, \ldots, y_n)$ of $\theta^\star$.

F3. There exists (a, b) such that $0 < a < t^\star < b < 1$ and a Lipschitz continuous function $\tilde{\theta}$ defined on $[a, b]$ such that, under $H_{k_n^\star}$, $(\hat{\theta}_{[tn]})$ converges in probability uniformly on $[a, b]$ to $\tilde{\theta}(t)$.

Then we have the following result:

Theorem 3.3. *Let $\hat{k}_n$ stand for the estimator defined in Section 3.2.2. Assume* **F1**, **F2** *and* **F3**. *Let (u_n) be any positive decreasing sequence such that $\sqrt{n}u_n \to \infty$. Under the asymptotic framework defined by* **AF1**, **AF2**, **AF3**,

$$P_{H_1(k_n^\star)}\left(\left|\frac{\hat{k}_n}{n} - t^\star\right| > u_n\right) \to 0.$$

The proof of this result in the varying window setting follows the same lines as that of Theorem 3.2, and is not detailed here.

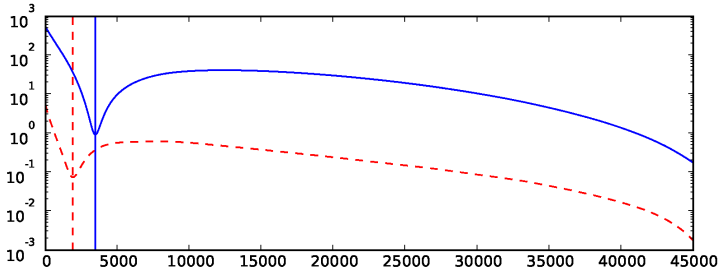


FIGURE 3.1: **The sequence η_k for $0 \leq \mu_i \leq 4$**
Dashed line: fixed window width $K_n = 5\,000$; *Solid line:* varying window width.

TABLE 3.1: **Number of misclassified voxels (false positives and false negatives) for different thresholding methods.**

	$K_n = 15\,000$	$K_n = 35\,000$	var. window	GGM
$0 \leq \mu_i \leq 4$	7477 ± 193	7148 ± 228	7120 ± 237	11016 ± 666
$1 \leq \mu_i \leq 5$	5079 ± 243	4736 ± 250	4706 ± 255	8306 ± 1862
$2 \leq \mu_i \leq 6$	2561 ± 185	2385 ± 182	2381 ± 184	1901 ± 125
$3 \leq \mu_i \leq 7$	930 ± 118	892 ± 120	891 ± 120	766 ± 102
$4 \leq \mu_i \leq 8$	297 ± 55	298 ± 68	297 ± 67	262 ± 52

Results are obtained over 100 replications, and given in the form: mean ± std. deviate.

3.3 Simulations

3.3.1 Experiment summary

In this experiment, we have repeatedly simulated $n = 50\,000$ Gaussian random variables with μ_i distributed uniformly on $[a, b]$ for $1 \leq i \leq 10\,000$, and $\mu_i = 0$ for $10\,001 \leq i \leq 50\,000$. (ϵ_i) is a sample from the $\mathcal{N}(0, 1)$ distribution.

(a, b) was chosen successively equal to $(0, 4)$, $(1, 5)$, $(2, 6)$, $(3, 7)$, $(4, 8)$. 100 datasets where simulated for each of these values. On each simulated dataset, we then used the procedure described in Section 3.2.1 (the unknown distribution extension), assuming a Gaussian cdf for $F_{|\epsilon|}$ with unknown variance, and compared the results for different window widths $K_n = 15\,000,\ 35\,000$. We also used the varying window extension, as described in Section 3.2.2, applied to the unknown variance setting, and the Gamma-Gaussian mixture (GGM) modeling procedure in [Beckmann et al., 2003b].

Our choice of simulation parameters was guided by the fact that 50 000 is roughly the number of voxels in a whole-brain fMRI activation map. 8 can be seen as an upper bound on the maximal signal intensity, since in the real data we analyzed (see Section 3.4), the maximum z-score was equal to 6.82 for the individual subject activation map, and 5.08 for the between-subject activation map.

3.3.2 Results and discussion

We compared the different approaches through the error rates they achieved on the different datasets. We considered the binary risk, *i.e.*, the number of misclassified voxels,

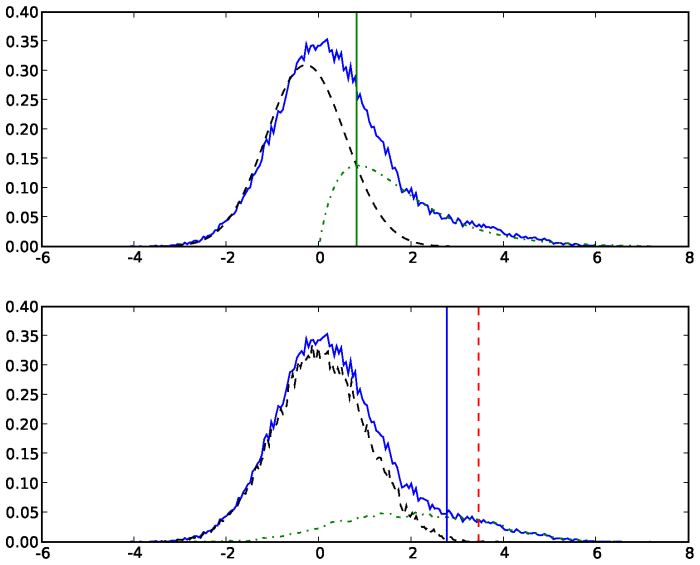


FIGURE 3.2: **Different thresholding strategies for** $0 \leq \mu_i \leq 4$.
Top: Gamma-Gaussian mixture fit (solid curve: data, dashes: Gaussian, dash-dots: Gamma). *Bottom*: Random thresholds with $K_n = 5\,000$ (dashed line) and varying K_n (solid line). Data corresponds to the solid curve, null data to the dashes and non null data to dash-dots.

TABLE 3.2: **False positive rates for different thresholding methods.**

	$K_n = 15\,000$	$K_n = 35\,000$	var. window	GGM
$0 \leq \mu_i \leq 4$	0.001 ± 0.001	0.002 ± 0.001	0.002 ± 0.001	0.219 ± 0.019
$1 \leq \mu_i \leq 5$	0.001 ± 0.001	0.002 ± 0.001	0.002 ± 0.001	0.182 ± 0.062
$2 \leq \mu_i \leq 6$	0.002 ± 0.001	0.003 ± 0.001	0.003 ± 0.001	0.009 ± 0.003
$3 \leq \mu_i \leq 7$	0.002 ± 0.001	0.002 ± 0.001	0.002 ± 0.001	0.004 ± 0.001
$4 \leq \mu_i \leq 8$	0.001 ± 0.001	0.001 ± 0.001	0.001 ± 0.001	0.002 ± 0.001

Results are obtained over 100 replications, and given in the form: mean $\pm$ std. deviate.

as a measure of how well the null and non null sets where separated (see Table 3.1). We also computed the false positive rate (FPR) (see Table 3.2).

As could be expected, differences between methods were most observed when the data was simulated with a low signal-to-noise ratio (SNR), (first and second line of each table). In this case, the Gamma-Gaussian mixture (GGM) model is clearly misspecified since it cannot account for negative observations in the non-null set, as illustrated in Figure 3.2. This leads to risk and false positive rate (FPR) values much higher than those obtained by the other thresholding methods.

For higher SNRs, the different approaches gave similar results, even though the GGM model yielded slightly lower risks than the different random thresholds, at the cost of higher FPR values.

On the whole, the different random thresholding procedures gave very similar results in all cases, and yielded mean FPR values that were systematically below 0.003, and also much lower than those found by GGM fit.

In conclusion, this experiment suggests that the random thresholding approach provides a very good control on false positives, and yields reasonable risk values at all SNRs, with a slight advantage for the varying window extension. In contrast, the GGM fit appears to yield more false positives, and performs poorly at low SNRs.

3.4 fMRI data

We used an event-related fMRI protocol involving a cohort of 37 right-handed subjects. The participants were presented with a series of stimuli or were engaged in tasks such as passive viewing of horizontal or vertical checkerboards, left or right click after audio or video instruction, computation (subtraction) after video or audio instruction, sentence listening and reading. Events occurred randomly in time (mean inter stimulus interval: 3s), with ten occurrences per event type, and ten event types in total.

The subjects gave informed consent and the protocol was approved by the local ethics committee. Functional images were acquired on a General Electric Signa 1.5T scanner using an Echo Planar Imaging sequence (time of repetition = 2400 ms, time to echo = 60 ms, matrix size = 64×64, field of view = 24 cm^2). Each volume consisted of 34 64×64 3 mm-thick axial contiguous slices. A session comprised 130 scans. Anatomical T1 weighted images were acquired on the same scanner, with a spatial resolution of $1 \times 1 \times 1.2$ mm^3. Finally, the cognitive performance of the subjects was checked using a battery of syntactic and computation tasks.

First-level analyses were conducted using SPM5 http://www.fil.ion.ucl.ac.uk. Data were submitted successively to motion correction, slice timing and normalization to the MNI template. For each subject, BOLD contrast images were obtained from a fixed-effect

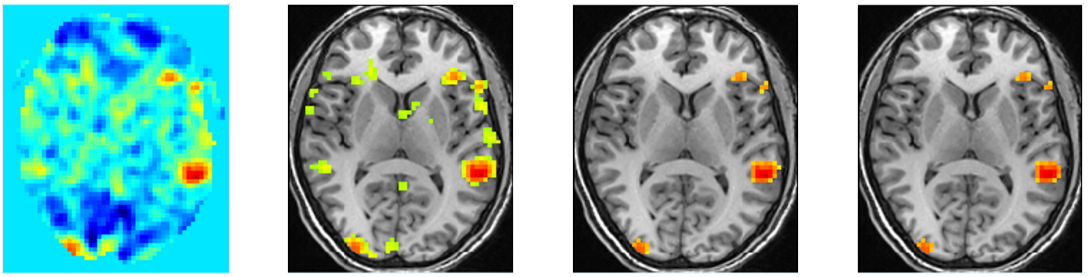


FIGURE 3.3: **Axial slice from a z-score map for the "sentence–checkerboard" contrast.**
From left to right: Unthresholded, thresholded by GGM model fit, varying-window and fixed-window ($K_n = 15\,000$) random thresholding. Detected activations are superimposed on the subject's anatomical image.

analysis on all sessions. Group analyses were restricted to the intersection of all subjects' whole-brain masks, comprising 43 367 voxels.

We considered the t-score maps computed for different contrasts of experimental conditions. These were first converted to z-score maps, to obtain approximatively Gaussian statistics in inactivated voxels. Using these maps as input data, we then compared, as in the simulation study, the detection thresholds obtained by Gamma-Gaussian mixture modeling (GGM), fixed-window random thresholding and the varying-window extension using in the last two cases the unknown variance extension of the method (see Section 3.2.1). For simplicity, we only present here the results obtained for a fixed window equal to $K_n = 15\,000$.

3.4.1 Individual subject activation map

Our first illustration concerns the activation map of a single subject, for the "sentence–checkerboard" contrast. This contrast subtracts the effect of viewing horizontal and vertical checkerboards from that of reading video instructions, thus allowing to detect brain regions specifically implicated in the reading task.

Figure 3.3, left, shows an axial slice from the z-score map before thresholding. Activations are clearly seen in Wernicke's and Broca's areas (right and upper right), which are known to be involved in language processing (see [Price, 2000], for instance). The detection threshold found by GGM fit for the z-score map (2.03) is much lower than those found by the random threshold procedure, both with a varying window (3.19) and

a fixed window (3.33). We note that these thresholds follow the same order found in the simulation study.

The random thresholds with fixed and variable windows yield very similar activation maps in this case, which seem to capture the activated regions seen in the raw map. In contrast, the much lower threshold found by mixture modeling detects several smaller clusters, some of which may be false positives.

3.4.2 Group activation map

In this second example, we consider a group activation map, specifically a map of t-statistics computed from the individual contrast maps of 15 subjects, thus enabling to infer regions of positive mean effects in the parent population. Our choice of limiting the number of subjects, rather than using the whole cohort, was driven by the fact that many fMRI studies are conducted on groups of less then 20 subjects. The remaining subjects were used to assess the variability of the thresholds found by the different approaches with respect to the choice of the subgroup, as described in Section 3.4.3.

We report results from the "calculation–sentences" contrast, which subtracts activations due to reading or hearing instructions from the overall activations detected during the mental calculation tasks. This contrast may thus reveal regions that are specifically involved in the processing of numbers.

Figure 3.3, left, shows an axial slice from the activation map before thresholding, with clear activations in the bilateral anterior cingulate (upper middle), bilateral parietal (lower left and right) and right precentral (upper right) regions, all known to be involved in number processing, as explained in [Pinel et al., 2007].

Though sorted in the same order as previously, the varying window random threshold (2.49) is now roughly at equal distances from the threshold found by GGM modeling (1.79) and the fixed window random threshold (3.06).

The three methods detected activations in the regions described above, though the fixed window random threshold seemed to miss some activations, and the GGM approach further detected smaller clusters, some of which may be false positives.

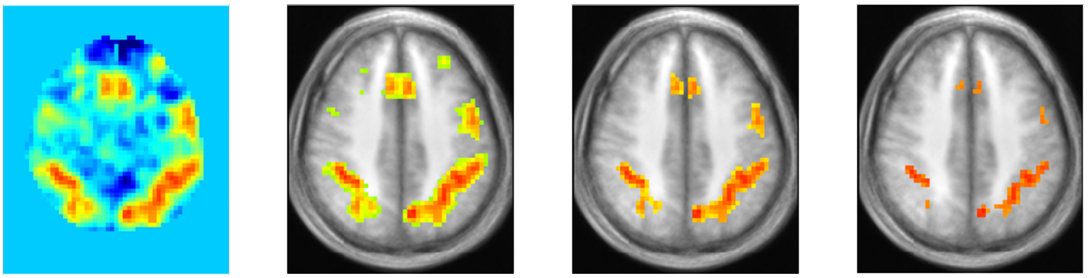


FIGURE 3.4: **Axial slice from the group activation z-score map for the "calculation–sentence" contrast.**
From left to right: Unthresholded, thresholded by GGM model fit, varying-window and fixed-window ($K_n = 15\,000$) random thresholding. Detected activations are superimposed on the mean anatomical image of all subjects.

3.4.3 Reproducibility study

In both experiments described above, it seemed that the varying window threshold found a good compromise between the GGM fit, which selected possibly inactive voxels, and the fixed window random threshold, which seemed to be overly conservative. However, we cannot conclude from these observations alone that the varying window threshold is 'better' than the other thresholds.

To compare them on a more objective basis, we studied the variability of the different thresholds with respect to the choice of a subgroup. More precisely, we repeatedly and randomly selected a group of 15 distinct subjects from the cohort of 38, computed the corresponding group activation map for the computation task, and thresholded it by all three methods. We then computed the empirical mean and variance of each sample of threshold values. These provided unbiased estimates for each threshold's mean and variance, with respect to the data's sampling distribution:

Proposition 3.4. *Let $Y_i = (y_{i1}, \ldots, y_{in})$ be the activation map for subject $i \in \{1, \ldots, N\}$, let $J \subseteq \{1, \ldots, N\}$ be any subgroup of subjects, and $T = T(Y, J) = T(\{Y_i\}_{i \in J})$ any statistic computed from the activation maps indexed by J. Define:*

$$\begin{aligned} \hat{\mathbb{E}}(T) &= \frac{1}{N!} \sum_{\pi \in \Pi_n} T(\{Y_{\pi(i)}\}_{i \in J}), \\ \hat{\mathbb{V}}(T) &= \frac{1}{N!} \sum_{\pi \in \Pi_n} (T(\{Y_{\pi(i)}\}_{i \in J}) - \hat{\mathbb{E}}(T))^2, \end{aligned}$$

TABLE 3.3: **Mean and variance estimates of the thresholds found by different methods for the computation task.**

Method	GGM	fix. window	var. window
Mean	1.92	2.96	2.54
Variance	0.76	0.12	0.08

where Π_N *is the set of all permutations on* $\{1, \ldots, N\}$. *If the* Y_i*'s, for* $i \in \{1, \ldots, N\}$, *are independent random variables following a common distribution* $\mathcal{P}$ *on* $\mathbb{R}^n$, *then* $\hat{\mathbb{E}}(T)$ *and* $\hat{\mathbb{V}}(T)$ *are unbiased estimates of the expectation and variance of* T, *respectively.*

Proof. This is an immediate consequence of the first two lemmas from Section 6 in [Strasser and Weber, 1999].

Note that we used Monte-Carlo estimates of the quantities defined in Proposition 3.4 by using 100 random permutations, since performing exhaustive permutations was intractable.

As can be seen in Table 3.3, The GGM yields the most variable threshold, and the varying window is slightly less variable than the fixed window random threshold. Thus the varying window variant yields the most stable threshold among those tested here.

3.5 Discussion

By introducing a simple modification to the random procedure proposed in [Lavielle and Ludeña, 2007], we have obtained an entirely unsupervised procedure for recovering non null mean terms from a collection of independent random observations, based solely on a parametric model of the null terms. Importantly, our modification, which requires no prior tuning, conserves the consistency properties of the original procedure.

Simulation results suggest that the random threshold approach has very good properties in terms of type I error rate control. While the fixed window seems overly conservative, and unstable with respect to the window parameter at low signal to noise ratios, the varying window extension, which avoids instability, also seems to find a better compromise between specificity and sensitivity.

The first results on real fMRI data are encouraging and seem to confirm the good stability achieved by the varying window random threshold, with respect to the other tested strategies.

A more thorough investigation is needed to confirm these preliminary results. Reproducibility of the regions found by random thresholding, rather than just the threshold values, can be assessed by resampling techniques similar to those used here, as described in [Thirion et al., 2007a], and compared to mixture modeling, as well as to false positive control strategies. Finally, studying the asymptotic properties of the random threshold estimator when null mean and non-null mean terms are not well separated is also very important in view of the applications.

Chapter 4

Modeling spatial uncertainty

Abstract

This chapter presents an extension of the mass univariate detection approach for group fMRI data, which relaxes the assumption of perfect match between the effect maps of the different subjects. A set of hidden variables is introduced in the classical mass univariate model, which represent registration errors, and are modeled as random deformation fields. The group mean effect map is estimated in a Bayesian setting by its posterior expectation. This is evaluated numerically, using a Metropolis-within Gibbs algorithm to sample from the posterior density of all hidden variables. We also show the consistency of the posterior density of the model parameters.

Using simulations, we evidence a stretching effect of the estimated activation pattern when the registration errors are unaccounted for, causing neighboring activations to be merged. This stretching effect is substantially reduced when registration errors are modeled. When applied to real fMRI data, our method yields group effect maps under spatial uncertainty that are both smoother and more contrasted than under no spatial uncertainty, an effect that cannot be reproduced by linear isotropic smoothing. These results are obtained in spite of the slow mixing of our posterior sampling algorithm, suggesting some space for improvement.

4.1 Introduction

In this chapter, we relax the assumption of perfect match between individual brains, which is one of the limitations of the SPM-like approach identified in Section 2.6. To this end, we incorporate in the mass univariate model specified by (2.5) and (2.6) a set of hidden variables, representing spatial normalisation errors, which are modeled as multivariate random fields. A first implementation of this idea has been published in [Keller et al., 2008].

In the following, the individual effect maps are seen as warped and noisy versions of the group activation pattern to be estimated. Our goal here is not to solve the registration problem but rather to develop an inference strategy that accounts for registration errors. When performing registration, the spatial transformations are the parameters of interest. In our case, these transformations are viewed as nuisance parameters to be integrated out during the analysis. Also, they are modeled as residual deformation fields, which persist after the actual registration has taken place. This precludes the use of global linear transformations, such as translations or rotations, to model the deformations. Rather, these are seen as local, zero-mean displacements, with no pre-determined form.

There exists an extensive literature on image registration. General reviews on this subject can be found in [Brown, 1992, Zitova and Flusser, 2003]. More specific reviews include [Van den Elsen et al., 1993, Maintz and Viergever, 1998], on the registration of medical images, and [Toga, 1999], on registration in brain imaging. Because we wish to model registration errors, we are interested in a statistical formulation of the registration problem. In [McGillem and Svedlow, 1976], a first step in this direction is taken, where the two images to be aligned are considered as identical up to a deformation and an additive Gaussian noise, justifying the use of the variance as a measure of overlay quality. More generally, in addition to such a dissimilarity measure, the criterion to be optimized by the registration procedure may contain a so-called regularization or penalty term, which imposes constraints on the transformation [Hajnal, 2001]. This penalty term may take many different forms. In the framework of pattern theory [Grenander, 1993], deformations are modeled in a Bayesian setting, in which case the penalty is interpreted as a prior density.

Furthermore, the problem we address here is closely related to template estimation, also referred to as *atlas construction* in the medical image analysis literature [Subsol, 1995, Joshi et al., 2004]. A simple atlas construction method consists in iterating a 'registration step' and an 'averaging step': on each iteration, individual images are registered to a current version of the template, and then averaged in order to update the template. For instance, the widely used MNI anatomical template was constructed in this fashion (see Section 2.3.1). However, there is no guaranty that this scheme yields a statistically consistent estimate of the mean image. More recently, [Allassonnière et al., 2007, Allassonnière et al., 2008, Leporé et al., 2008, Sabuncu et al., 2008] have proposed to construct a Bayesian estimate of the template, given prior distributions on both the deformation fields and the template parameters. Consistency of the template's MAP estimator is shown in [Allassonnière et al., 2007].

This chapter is organized as follows: in Section 4.2, we introduce spatial uncertainty in the mass univariate model specified by (2.5) and (2.6), under the form of unknown spatial deformation fields. The deformation fields are modeled in Section 4.3. From this model, we derive in Section 4.4 a Bayesian estimate of the template $\boldsymbol{\mu}$, given a certain prior distribution on the model parameters, in an approach similar to [Allassonnière et al., 2007]. Section 4.5 illustrates on several simulation studies how techniques that ignore spatial variability may result in blurring and stretching activation patterns, and how this effect can be reduced by our approach. Finally, results on real fMRI data are presented in Section 4.6.

4.2 Observation model

Following the notations introduced in Section 2.3.2, $\mathbf{x}_i = (x_{i,1}, \ldots, x_{i,d})$ is the map of BOLD effects of subject $i = 1, \ldots, n$, in response to a certain contrast of experimental conditions; $\mathbf{y}_i = (y_{i,1}, \ldots, y_{i,d})$ is the noisy estimate of $\mathbf{x}_i$, available from the analysis of the subject's scans (see Section 2.2), and $\mathbf{s}_i^2 = (s_{i,1}^2, \ldots, s_{i,d}^2)$ an image of estimation variances. Recall that, under a sufficient number of acquired scans, the estimation error $\boldsymbol{\varepsilon}_i$ can be assumed Gaussian, so that:

$$\mathbf{y}_i \quad = \quad \mathbf{x}_i + \boldsymbol{\varepsilon}_i; \quad \boldsymbol{\varepsilon}_i \sim \mathcal{N}\left(\mathbf{0}, \mathrm{diag}(\mathbf{s}_i^2)\right), \tag{4.1}$$

where $\mathrm{diag}(\mathbf{s}_i^2)$ denotes the diagonal matrix whose diagonal is given by $(s_{i,1}^2, \ldots, s_{i,d}^2)$. As in [Keller et al., 2008], we extend the mass-univariate Gaussian between-subject model described in Section 2.3.2 by incorporating spatial uncertainty so that at voxel k :

$$\mathbf{x}_i(\mathbf{v}_k) \quad = \quad \boldsymbol{\mu}(\mathbf{v}_k + \mathbf{u}_{i,k}) + \xi_{i,k}; \quad \boldsymbol{\xi}_i \sim \mathcal{N}\left(\mathbf{0}, \sigma^2 \mathbf{I}_d\right). \tag{4.2}$$

Here, we note $\mathbf{x}_i(\mathbf{v}_k) = x_{i,k}$ to emphasize that it is a spatial map; $\boldsymbol{\mu} \in \mathbb{R}^d$ is the map of mean population effects; the vector $\mathbf{u}_{i,k} \in \mathbb{R}^3$ is a hidden variable that models the subject-to-atlas registration error for subject i at voxel k; and the $\xi_{i,k}, 1 \leq i \leq n$, model the between-subject variability of the effect at voxel k. Finally, we define $\boldsymbol{\mu}(\mathbf{v}_k + \mathbf{u}_{i,k})$ by discrete interpolation, as being equal to the mean population effect in the nearest voxel $\mathbf{v}_{k'} = \arg\min_{\mathbf{v}_l} ||\mathbf{v}_l - (\mathbf{v}_k + \mathbf{u}_{i,k})||$. In practice, this is done by rounding each coordinate of $\mathbf{u}_{i,k}$ toward the nearest integer. We will also note in the following φ_i the function that maps each voxel index k to the corresponding displaced voxel index k' for subject i, so that $\boldsymbol{\mu}(\mathbf{v}_k + \mathbf{u}_{i,k}) = \mu_{\varphi_i(k)}$.

(4.2) generalizes the classical mass univariate model (2.6), which corresponds to the special case where the registration errors $\mathbf{u}_{i,k}$ are neglected. Note however that the between-subject variance σ^2 is uniform over the search volume in our formulation, whereas it is voxel dependent in the classical setting. Indeed, we found from practical experience that a voxel dependent variance raised overfitting issues when modeling registration errors, and resulted in degenerate estimates in certain voxels. We used a uniform variance as the simplest form of constraint to avoid this overfitting, at the price of a possible over-simplification. In Section 5.2, we introduce a regionalized version of Equation (2.6) where σ^2 is allowed to vary across functionally distinct regions. More sophisticated methods could be considered, for instance by modeling the variance map σ_k^2 as a smooth random field, and constitute a possible direction for future work.

4.3 Deformation field model

As a standard way of representing nonlinear local deformations [Zitova and Flusser, 2003], we use Gaussian splines to model the displacements $\mathbf{u}_{i,k}$. Specifically, elementary displacements $\mathbf{w}_{i,b}$ are defined in a limited number of fixed control points $\{\mathbf{v}_{k_b}, b = 1, \ldots, B\}$. These are then interpolated to the whole brain image using a radial basis function $\mathcal{K}$,

$$\mathbf{u}_{i,k} = \sum_{b=1}^{B} \mathcal{K}(\mathbf{v}_k, \mathbf{v}_{k_b}) \mathbf{w}_{i,b}, \tag{4.3}$$

where $\mathcal{K}(\mathbf{v}_k, \mathbf{v}_{k_b}) = \exp -\{\|\mathbf{v}_k - \mathbf{v}_{k_b}\|^2 / 2\omega^2\}$, and ω is a user-chosen parameter controlling the displacement field's smoothness.

The $\mathbf{w}_{ib}$'s are modeled as a *i.i.d.* Gaussian variables, with spherical covariance matrix $\sigma_S^2 \mathbf{I}_3$, where σ_S models the standard registration error:

$$\pi(\mathbf{w}_i | \sigma_S^2) \propto \prod_{b=1}^{B} \mathcal{N}(\mathbf{w}_{i,b}; \mathbf{0}, \sigma_S^2 \mathbf{I}_3). \tag{4.4}$$

4.4 Posterior mean estimate

We estimate the population effect map $\boldsymbol{\mu}$ by its posterior mean:

$$\mathbb{E}[\boldsymbol{\mu}|\mathbf{y}] = \int \boldsymbol{\mu} \pi(\boldsymbol{\mu}|\mathbf{y}) d\boldsymbol{\mu},$$

where $\pi(\boldsymbol{\mu}|\mathbf{y}) \propto \pi(\boldsymbol{\mu}) f(\mathbf{y}|\boldsymbol{\mu})$ is the posterior density of $\boldsymbol{\mu}$ relative to a given prior density $\pi(\boldsymbol{\mu})$. To define this prior, $\boldsymbol{\mu}$ is modeled according to:

$$\mu_k = \eta + \chi_k; \quad \chi_k \sim \mathcal{N}(0, \nu^2),$$

where η represents the mean activation across voxels, and ν^2 the variance of the activation pattern. This hierarchical prior can be seen as an instance of the regional model

developed in Chapter 5, in the special case of a single functionally homogeneous region. We then define a prior density on these hyperparameters, as well as on the spatial uncertainty parameter σ_S^2, as explained in Section 5.4.

The integral in the above display cannot be computed analytically, so we resort to MCMC strategies instead to produce a sample $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_G$ from the posterior density $\pi(\boldsymbol{\mu}|\mathbf{y})$, yielding the following Monte-Carlo estimate:

$$\hat{\boldsymbol{\mu}} = G^{-1} \sum_{g=1}^{G} \boldsymbol{\mu}_g.$$

Sampling the posterior density in the model with spatial uncertainty raises some technical issues. The simplest strategy would be to use a Gibbs sampler [Geman and Geman, 1984], to generate a sequence of samples from the joint posterior density of all hidden variables $\mathbf{x}$, $\mathbf{w}$, $\boldsymbol{\mu}$, (η, ν), σ^2, and σ_S^2. This is done by partitioning the variables into blocks, then sampling successively each block conditionally on all others. However, it turns out that the conditional distribution of each elementary displacement $\mathbf{w}_{ib}$ has no closed-form, and cannot be directly sampled. Therefore, we use the more general Metropolis-Hastings (MH) algorithm [Hastings, 1970]. This involves the choice of a proposal density to generate candidate values $\mathbf{w}_{ib}$, which are then accepted with a certain rate. Thus, we use a *Metropolis-within-Gibbs* algorithm [Tierney, 1994] that generalizes the standard Gibbs sampler by the incorporation of MH iterations. Technical details on this algorithm can be found in appendix D, for the more general model introduced in Chapter 5.

As is often the case with Bayesian estimation, the posterior distribution $\pi(\boldsymbol{\mu}|\mathbf{y})$ has very good frequentist properties. In particular, it is *consistent*, *i.e.,* it concentrates on the true value $\boldsymbol{\mu}_0$ of the mean population effect map, assuming the data is indeed distributed under our generative model $f(\mathbf{y}|\boldsymbol{\mu})$. Thus $\hat{\boldsymbol{\mu}}$, as well as any other Bayes estimator based on this posterior distribution, are reasonable choices to estimate $\boldsymbol{\mu}_0$.

Theorem 4.1 (**C**onsistency of the Posterior Distribution)**.** *Let $\boldsymbol{\theta} = (\boldsymbol{\mu}, \sigma^2, \sigma_S^2)$ denote the parameter vector of the model defined by (4.1), (4.2), (4.3), and (4.4), and $\mathbf{y}^{(n)} = \mathbf{y} = (\mathbf{y}_i)_{1 \leq i \leq n}$ the data vector, where n is the number of observed activation maps. Then for all $\varepsilon > 0$:*

$$\mathbb{P}[\| \boldsymbol{\theta} - \boldsymbol{\theta}_0 \| > \varepsilon | \mathbf{y}^{(n)}] \overset{n \to \infty}{\longrightarrow} 0,$$

for any value of the true parameter $\boldsymbol{\theta}_0$, except possibly on a set of π-measure 0.

Proof. This derives from Doob's theorem (see [van der Vaart, 2000], Chapter 10, for instance). Doob's theorem states that in an identifiable parametric model and for any prior distribution π, the sequence of posterior densities is consistent for π-almost every parameter value. The identifiability of our model is demonstrated in Appendix C.

However, Doob's theorem applies only to independent, identically distributed (iid) variables. In contrast, the observations $\mathbf{y}_i$ in (4.1) are defined conditional on the first-level variances $\mathbf{s}_i^2$, which are different for each subject, and so are not identically distributed. This can be arranged through the device introduced in [Mériaux et al., 2006], where the $\mathbf{s}_i^2$ are considered as part of the data vector, and modeled as a sample from an unspecified density $f(\mathbf{s}_i^2)$, independent from all other variables. This has no effect on the posterior density, and under this assumption the observations $(\mathbf{y}_i, \mathbf{s}_i^2)_{1 \leq i \leq n}$ are iid, so that the result holds □

4.5 Simulations

4.5.1 1D simulations

We now illustrate our model of spatial uncertainty on a simplistic one-dimensional (1D) example. Our purpose here is to give some insight on how deformations are modeled, and how they can impact the estimation of the activation pattern $\boldsymbol{\mu}$, which in this case is simply a real-valued function.

Description of the dataset

The voxels, aligned and equally spaced in this example, are represented by the integers $k = 1, \ldots, 50$. We simulated a deformation field according to model (4.3), with a standard displacement of $\sigma_S = 3$ voxels, and a kernel width of $\omega = 6.5$ voxels. We also empirically set the distance between adjacent control points to $2 \times \omega = 13.0$ voxels, with no control points at at less then $2.5 \times \omega = 16.25$ voxels from the voxel set boundaries. In the present case, this resulted in two control points only, as illustrated in Figure 4.1.

This displacement field was used to warp a signal defined as the sum of three Gaussian 'bumps', representing the mean effect map $\boldsymbol{\mu}$, following (2.6), (see Figure 4.2, middle).

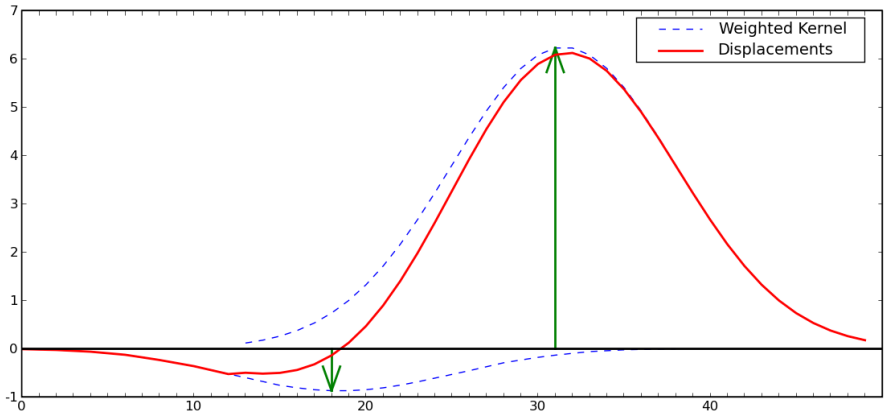


FIGURE 4.1: Illustration of the deformation field model. Elementary displacements (arrows) in pre-specified control points are interpolated to all other points using a radial kernel (dashed lines). The displacement field (solid line) is the sum of the resulting weighted kernels.

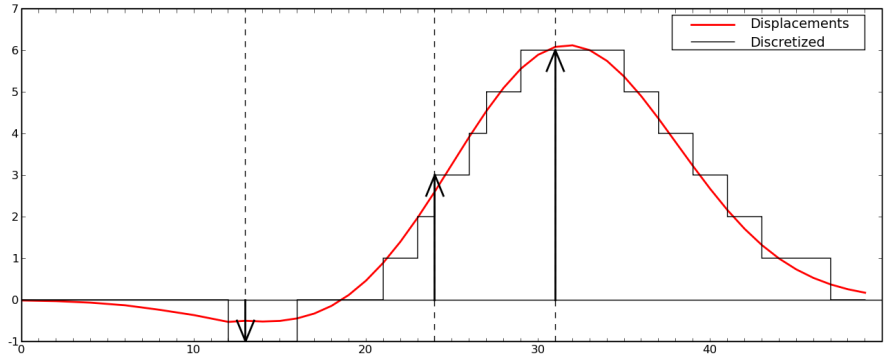
Heteroscedastic noise with variance $\sigma^2 + s_{i,k}^2$ was then added to this warped signal to produce a synthetic observation $\mathbf{y}_i$ (see Figure 4.2), according to (4.1), (4.2). We chose $\sigma = 1.0$, and the individual standard deviates $s_{i,k}$ where generated as independent standard normal variables, multiplied by a noise level of $\varepsilon = 4.0$.

Methods compared

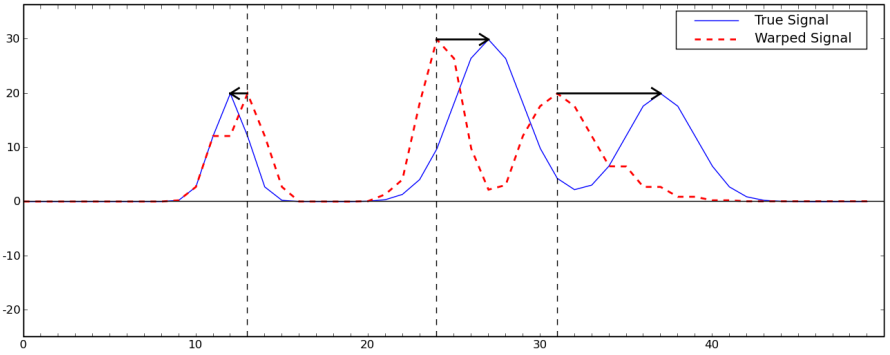
We generated $n = 40$ observations as described above, and tried to recover the original signal $\boldsymbol{\mu}$ from these warped, noisy versions. To do this, we computed the posterior mean $\mathbb{E}[\boldsymbol{\mu}|\mathbf{y}]$, as explained in Section 4.4. We compared the estimate obtained in the full model with that obtained in the model without spatial uncertainty, setting $\sigma_S^2 = 0$. In both cases, the posterior mean was averaged over 100 Gibbs iterations, following 100 'burn-in' iterations which were discarded.

It can be seen in Figure 4.3 that, when sampling the posterior distribution of the full model, the Markov chain quickly reached a stable state, after approximately 100 iterations. However, the resulting posterior mean estimate of the spatial incertitude parameter, $\hat{\sigma}_S = 1.6$, was lower than the true value, $\sigma_S = 3.0$. Conversely, the posterior mean of the between-subject standard deviate, $\hat{\sigma} = 2.0$ was higher than the actual value, $\sigma = 1.0$.

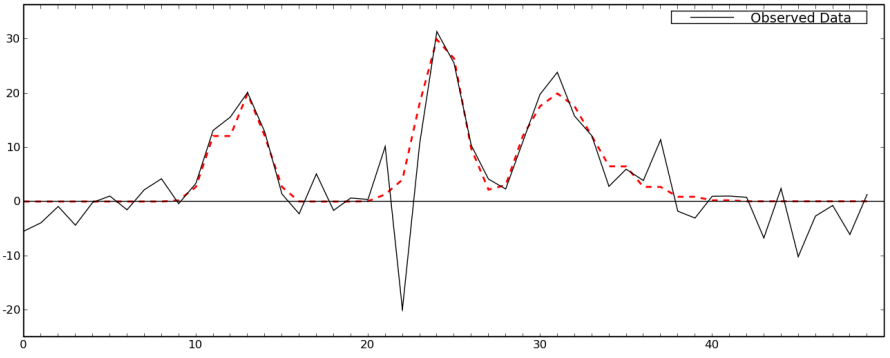
FIGURE 4.2: Illustration of the model with spatial uncertainty on 1D data.



Deformation field (corresponding to $\mathbf{u}_{i,k}$ in (2.6)). For instance, $[u_{i,13}] = -1$, $[u_{i,24}] = +3$,and $[u_{i,31}] = +6$.



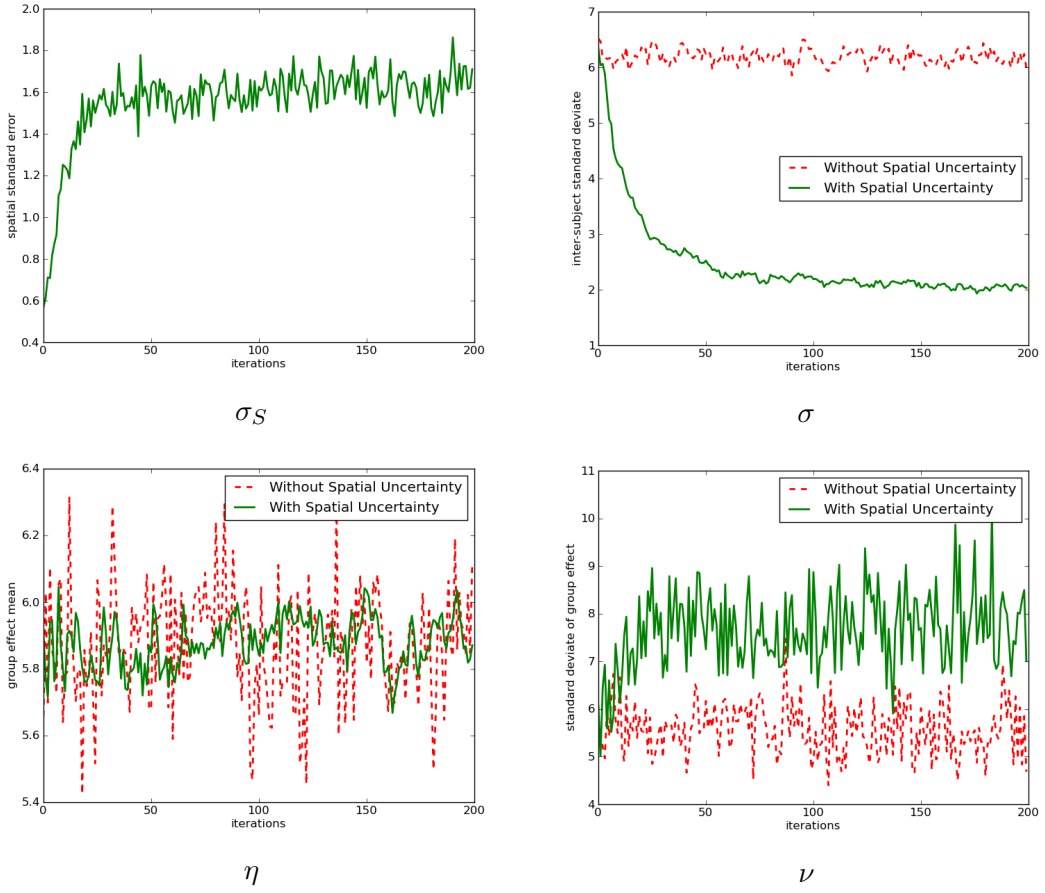
Effect of the deformation field on the signal $\boldsymbol{\mu}$ (solid line). For instance, the warped signal at point $k = 24$ is equal to the true signal at point $24 + [u_{i,24}] = 24 + 3 = 27$.



Data $\mathbf{y}_i$ (solid line), obtained by adding Gaussian noise to the warped signal.

A possible explanation is that the chain was trapped next to a mode of the posterior distribution, away from the global maximum. Escaping from this mode would theoretically happen after a sufficient number of iterations, but there is no upper bound on the CPU time this would require. This illustrates the difficulty of sampling from the joint distribution of the deformation fields, even in this simplistic 1D example with two

FIGURE 4.3: Posterior sampling of different parameters in the 1D model with and without spatial uncertainty



control points per field.

Alternatively, this bias in the variance estimates may also be a consequence of the so-called *shrinkage* (or *regularizing*) effect, *i.e.* a compromise between estimator bias and variance induced by the prior distribution. Specifically, the model on spatial displacements $\mathbf{w}$ (4.4), together with the prior on the spatial standard error σ_S (5.12), favor small displacements, thus preventing unreasonable distortions of the individual images which would artificially increase the likelihood by overfitting the data.

Supporting this explanation, the convergence plot σ (Figure 4.3, upper right, solid line), shows that its sampled values rapidly decrease at first, while the spatial incertitude parameters augments (Figure 4.3, upper left). This suggests that the Markov Chain is able to find a significantly better match between the individual images by spatially deforming them, σ being a measure of their global dissimilarity (in contrast, σ values sampled in the model without spatial uncertainty (Figure 4.3, upper right, dashed line)

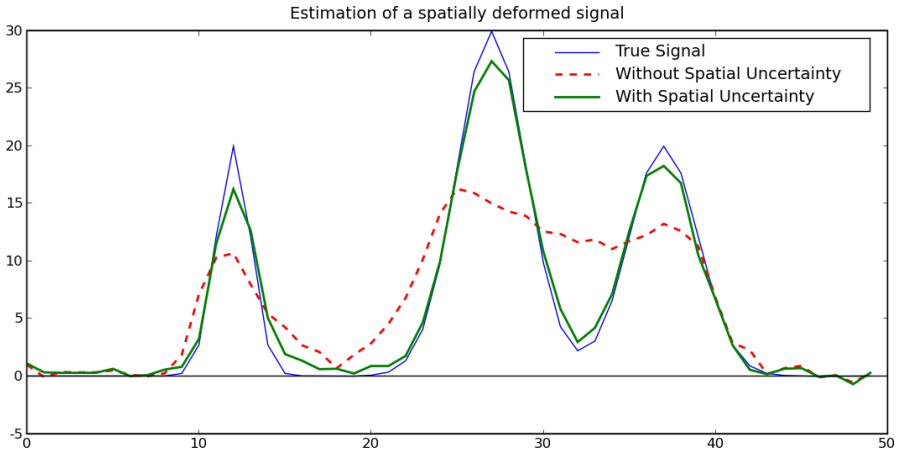


FIGURE 4.4: Posterior estimates of $\boldsymbol{\mu}$ in the 1D-model, with and without modeling spatial uncertainty.

do not vary significantly from their initial value). Then, both σ and σ_S settle in a stable state, presumably because an equilibrium has been met between a good match across the images (low σ) and small, regular deformations (low σ_S).

Additional indicators of the enhanced match between individual images, achieved when modeling spatial uncertainty, is provided by the convergence plot of the group effect map's mean η (Figure 4.3, bottom left, solid line) which is less variable than in the model without spatial uncertainty (dashed line). Also, the resulting map $\hat{\boldsymbol{\mu}}$ is more contrasted, as can be seen from the higher sampled values of its variance ν (Figure 4.3, bottom right).

Finally, we can see from Figure 4.4 that the posterior mean estimate $\hat{\boldsymbol{\mu}}$ under the full model successfully recovered the three different modes present in the original signal. In contrast, as could be expected beforehand, the estimate which neglected the signal's spatial variability behaved poorly. As a result, regions of positive signal were swelled, and the two nearby peaks merged in a single mass.

Sensitivity study

Next, we investigated the reproducibility of the above results, and their sensitivity to various features of the simulated data. We measured the quality of each estimate $\hat{\boldsymbol{\mu}}$ by its mean-square error $MSE(\hat{\boldsymbol{\mu}}) = \frac{1}{d}\sum_k(\hat{\mu}_k - \mu_k)^2$.

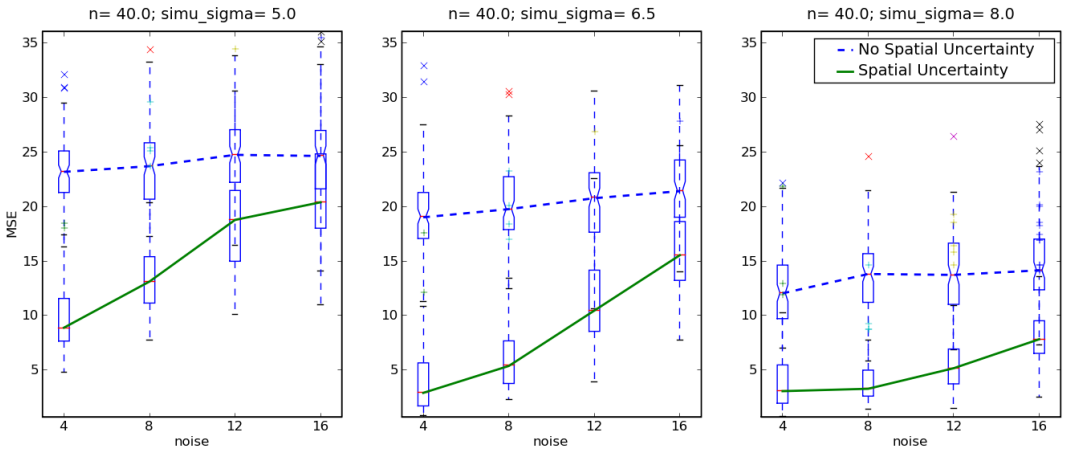


FIGURE 4.5: Sensitivity of the posterior mean estimate to noise and deformation field smoothness on 1D data.

Influence of noise and deformation field smoothness. In a first experiment, we simulated the data for different noise levels ($\varepsilon = 4.0, 8.0, 12.0, 16.0$) and different values of the deformation field smoothness parameter ($\omega = 5.0, 6.5, 8.0$). For each of the 4×3 possible combinations of these quantities, we generated 100 datasets, which we used to estimate $\boldsymbol{\mu}$, using the two methods described above. In particular, the deformation field model used for this posterior estimation assumed a field smoothness of $\omega_{est} = 6.5$, so that it alternatively under or over estimated the true field smoothness. This allowed to investigate the behavior of our sampling scheme when using a mis-specified deformation field model, which is inevitable when analysing real fMRI data.

Results of this experiment are presented in Figure 4.5. It can be seen that both posterior estimates are very sensitive to noise, since the MSE increases with the noise level ε in all cases. Also, the difference in MSE values between the estimates with and without spatial uncertainty is clearly seen at the lowest noise level considered $\varepsilon = 4.0$. However, it tends to disappear at the highest noise level $\varepsilon = 16.0$. Deformation field smoothness, measured here by the parameter ω, also seems to be a critical parameter, with estimations that are better for smoother fields.

Influence of Sample Size. In a second experiment, we studied how the performance of the posterior mean estimate varied when applied to datasets, simulated as described in Section 4.5.1, of increasing sizes $n = 20.0$, 30.0, 40.0, 50.0. As the sample size increases, the posterior distribution of the full model concentrates around the true value of the

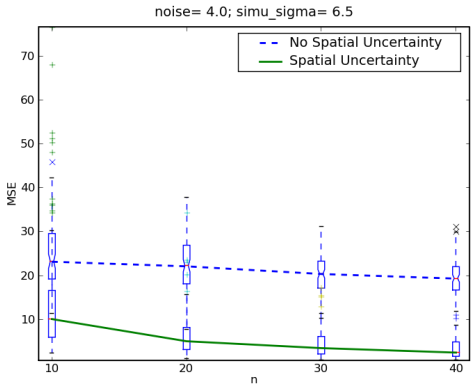


FIGURE 4.6: Sensitivity of the posterior mean to sample size on 1D data.

parameters, according to Theorem 4.1, unlike the posterior distribution of the model without spatial uncertainty, whose asymptotic behavior is unknown. Hence, we expect the difference in MSE between the two corresponding estimates of $\boldsymbol{\mu}$ to increase with sample size. This is precisely what we observe in Figure 4.6.

4.5.2 3D simulations

In this second numerical experiment, we apply our method to 3D simulated data, and show that the observations on simulated 1D data are confirmed, specifically, that standard voxelwise techniques (not accounting for spatial uncertainty) may lead to overestimating the size of positive effected regions. We also show that this undesirable "stretching effect" may be reduced by modeling spatial uncertainty.

Data simulation

A synthetic dataset was generated as follows. We defined a volume of $24 \times 32 \times 32$ voxels, containing two spherical activated regions, with uniform intensity value 5 (the background was set to 0) and a fixed diameter of 7 voxels.

This idealized activation, interpreted as the mean population effect map $\boldsymbol{\mu}$, was slightly smoothed (using a Gaussian kernel with standard deviation 0.5 voxels) to emulate the partial volume effect present in real fMRI data. It was then deformed according to a displacement field $\mathbf{u}$, simulated under the model described in Section 4.3, with one control point in each voxel. The standard displacement was taken equal to $\sigma_S = 2.0$ voxels and the field smoothness parameter was set to $\omega = 4.0$ voxels.

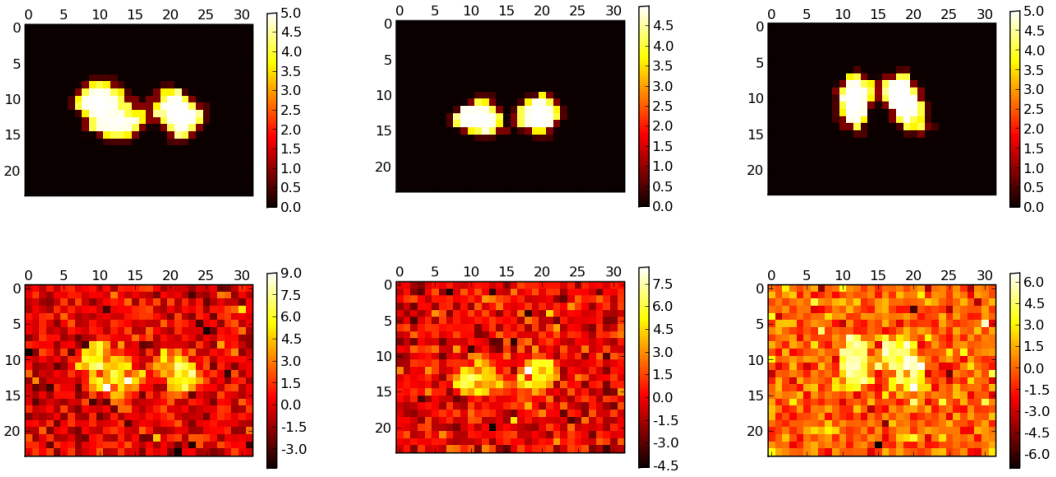


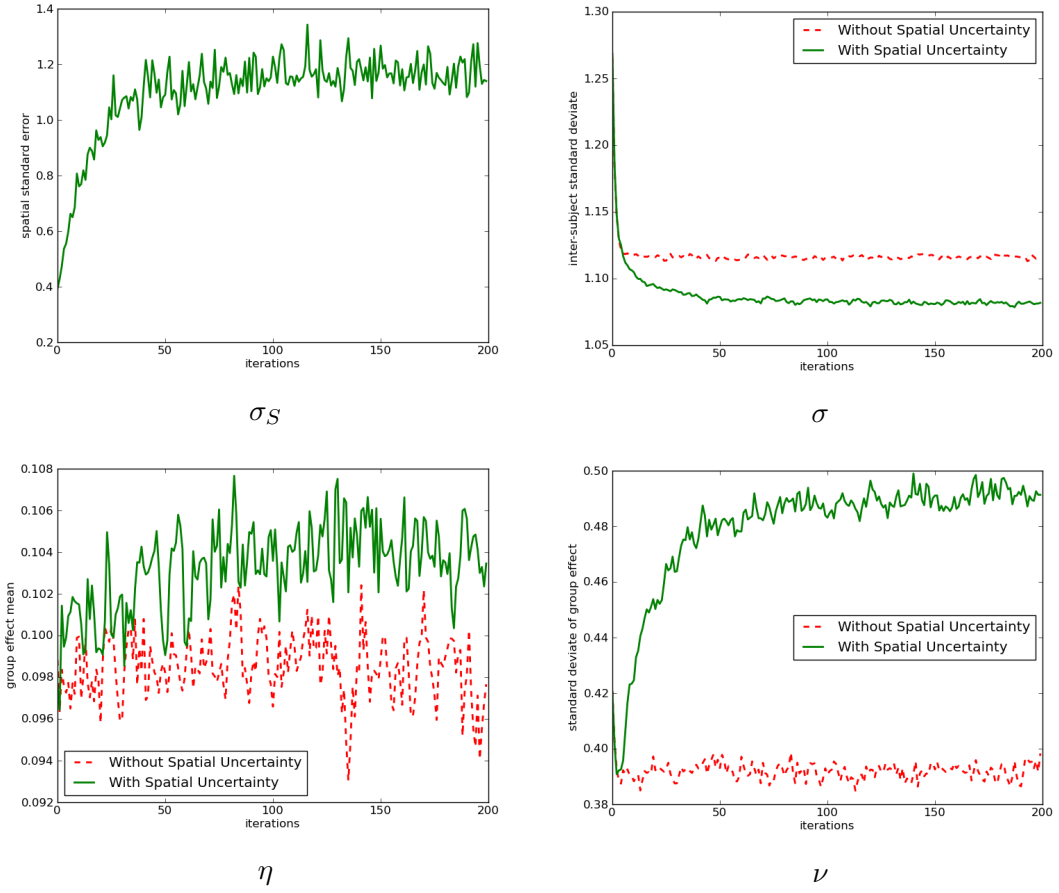
FIGURE 4.7: Examples of simulated data. (top) Deformed signals. (bottom) Data (deformed, noisy signals).

As for 1D simulations, independent heteroscedastic Gaussian noise was then added to each voxel $\mathbf{v}$, the variance of which was taken equal to $1+\mathbf{s}^2(\mathbf{v})$, with $(\mathbf{s}/\varepsilon)^2(\mathbf{v}) \sim \chi^2(1)$, ε being the noise level, set to 1.0 in this example. A total of $n = 40$ pairs $(\mathbf{y}_i, \mathbf{s}_i^2)$ of effect and variance maps were sampled in this fashion, as illustrated in Figure 4.7, and constitute a sample from the hierarchical model in Section 4.2 .

Methods compared

Based on the synthetic observations $\mathbf{y}_1, \ldots, \mathbf{y}_n$, and the hierarchical model in Section 4.2, we estimated the original signal $\boldsymbol{\mu}$ by its posterior mean $\mathbb{E}[\boldsymbol{\mu}|\mathbf{y}]$, as described in Section 4.4. Importantly, the estimation model used to compute $\hat{\boldsymbol{\mu}}$ differed from the one used to simulate the data, in that the deformation fields were defined by two control points, instead of one control point in each voxel. This modification was introduced to verify that the signal could be correctly estimated, even though the model was misspecified, as is unavoidable in real-world datasets. Finally, $\boldsymbol{\mu}$ was estimated in the model without spatial uncertainty, setting $\sigma_S^2 = 0, \mathbf{u} \equiv 0$, allowing to verify the presence of a stretching effect due to unaccounted spatial uncertainty, as was observed on the 1D example (Section 4.5.1). In order to compare objectively the different methods, we again measured the performance of each estimate $\hat{\boldsymbol{\mu}}$ by its mean-square error: $MSE = \frac{1}{d}\sum_k (\hat{\mu}_k - \mu_k)^2$.

FIGURE 4.8: Posterior sampling of different parameters in the 3D model with and without spatial uncertainty



As in the 1D case, we can see in Figure 4.8 that the Markov chain quickly settles in a stable state. A similar srinkage effet is suggested from the estimated spatial standard error $\hat{\sigma}_S = 1.2$ lower than the true value, $\sigma_S = 2.0$. The estimated between subject standard error $\hat{\sigma} = 1.1$, is however closer now to its true value $\sigma = 1.0$. This may be because the 3D data used in this simulations include many more voxels (18 432) than the length of the previous $1D$ data (50 points), and thus allow a much better estimation of certain model parameters (though σ_S remains conservatively biased).

As previously, the posterior estimate of the group effect is significantly more accurate using the full model than its version without spatial uncertainty, as can be seen in Figure 4.9. As expected, the latter (right) is visibly more spread out spatially than the original signal (left), and the two activated regions are merged into a single blurry region. In contrast, the signal estimated in the model with spatial uncertainty (center), correctly recovers the two separate activation spheres, which also appear less blurred.

These impressions are confirmed by the fact that its MSE (0.11) is lower than the one obtained without spatial uncertainty (0.19). $\hat{\sigma} = 2.1$

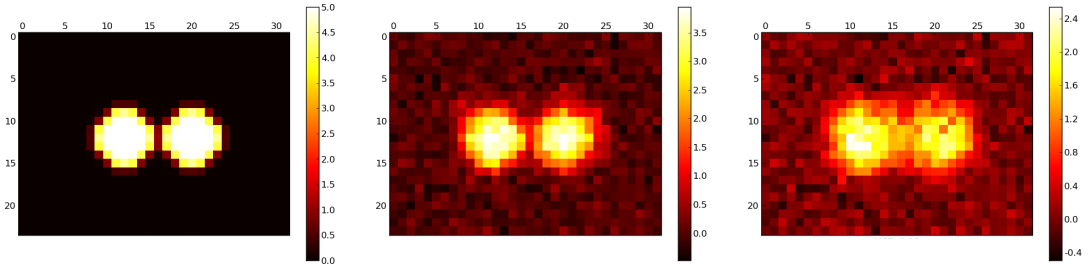


FIGURE 4.9: Posterior estimates of $\boldsymbol{\mu}$ in the 3D-model, with and without spatial uncertainty.

Sensitivity study

Next, we investigated the reproducibility of the above results, and the sensitivity of the posterior estimates with respect to various simulation parameters.

Influence of Noise and Deformation Field Regularity. We studied the influence of both the regularity of the deformations and the amount of observation noise. Thus, the field regularity parameter used to simulate the data, ω_{sim}, was chosen in $[3.0, 4.0, 5.0]$. Also, we generated the observation variance $s_{i,k}$ for subject i at voxel k according to $\frac{s_{i,k}}{\epsilon}^2 \sim \chi^2(1)$, with ϵ, controlling the noise level, varying in $[1.0, 2.0, 3.0, 4.0]$.

For all 3×4 combinations of these parameters, 100 synthetic datasets where generated and used to estimate $\boldsymbol{\mu}$, both with and without spatial uncertainty, as described above. In all cases, the estimation model used two control points to define each deformation field, and a field regularity parameter ω_{est} fixed to 4.0. It was therefore mis-specified with respect to the model used to simulate the data, for which the deformation fields where defined with one control point in each voxel, and varying values of the regularity parameter ω_{sim}.

The results of the sensitivity analysis are presented in Figure 4.10. As in the $1D$ case, both approaches are sensitive to observation noise, and the MSE is seen to decrease when the deformation field smoothness ω increases.

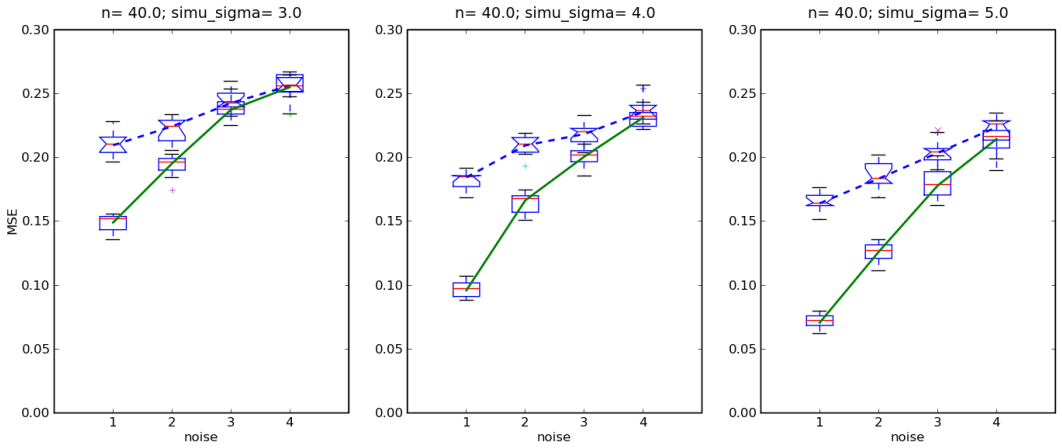


FIGURE 4.10: Sensitivity of posterior estimation to noise and deformation field smoothness. Mean-Square Error (MSE) for the estimation of signal ($\boldsymbol{\mu}$), in the model with spatial uncertainty (solid line) and without (dashed line). The lines represent the median MSE across 10 replications, ploted against the noise level ϵ. Dispersion of MSE values is represented by the boxplots. Each sub-figure corresponds to a different value of the deformation field smoothness parameter ω_{sim}, which increases from left to right.

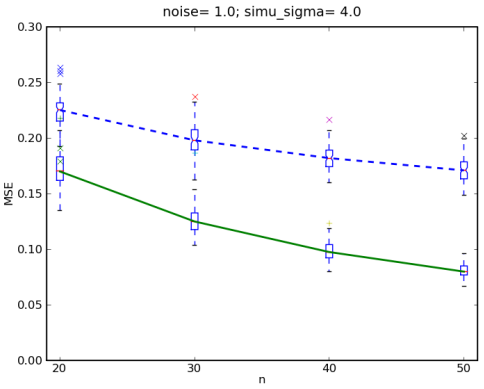


FIGURE 4.11: Sensitivity of posterior estimation to sample size.

Mean-Square Error (MSE) for the estimation of signal ($\boldsymbol{\mu}$), in the model with spatial uncertainty (solid line) and without (dashed line). The lines represent the median MSE across 100 replications, ploted against the number of observed images n. Dispersion of MSE values is represented by the boxplots.

Influence of sample size. Finally, we studied the behavior of our approach when the sample size varied, choosing $n = 20, 30, 40, 50$. (see Figure 4.11). As in the $1D$ case, the drop in MSE due to spatial uncertainty modeling is seen to increase with the data size.

4.5.3 Conclusion

The simulations presented in the above sections 4.5.1 and 4.5.2 show how the modeling of unknown deformations can lead to significant improvements in the estimation of a

spatially warped signal.

The sensitivity to noise of our estimation method is also highlighted in these illustrations. This sensitivity may be over-estimated here due to the slow mixing of our sampling algorithm, which in particular under-estimated the spatial uncertainty parameter. Nevertheless, this suggests that modeling spatial uncertainty is most useful in cases where the data is not too noisy, and motivates the use of a moderate preliminary smoothing step when processing the individual fMRI datasets (see Section 2.2.1). We would however recommend nonlinear smoothing strategies, such as cortical surface-constrained filtering [Andrade et al., 2001] or anisotropic diffusion [Kim et al., 2005], in order to limit the blurring effect inherent to Gaussian and other linear filters.

Finally, our simulations suggest that the field regularity parameter ω_{est} of the estimation model should ideally be not larger than the actual deformation field smoothness to optimally account for spatial uncertainty. It should also not be too small to avoid over-fitting and computational issues (since more control points would be necessary to define less regular fields). However, tuning ω_{est} remains an issue in real-life applications, since we see no way of estimating the actual regularity of the deformations, and their distribution may be very far from that specified in Section 4.3.

4.6 fMRI data

We conclude this chapter on spatial incertitude modeling by an illustration on real fMRI data. We used the Localizer dataset [Pinel et al., 2007], which is described in Section 3.4, restricted to a group of 38 right-handed subjects. Moderate preliminary smoothing by a $5mm$ Gaussian kernel was first used, in order to increase the SNR, as justified in the previous section. Then, each subject's data was pre-processed and analyzed using SPM5, as explained in Section 2.2.

For each contrast of interest $\mathbf{c}$ considered, the subject's BOLD effect maps $\mathbf{y}_i$ and estimation variance maps $\mathbf{s}_i^2$ where calculated, following the notations in Section 2.3.2. These were used as input data to estimate the mean effect map $\boldsymbol{\mu}$, as described in Section 4.4. As in the simulation studies, we estimated $\boldsymbol{\mu}$ both in the model with spatial

uncertainty, and in the model without spatial uncertainty, setting $\sigma_S = 0$. The deformation field smoothness was set to $\omega = 4.5$ voxels, and each deformation field was defined by 60 equally spaced control points.

100 iterations of the Gibbs sampler were used to average the posterior mean, following 100 steps which were discarded, corresponding to a burn-in period. This took approximately 1 hour and 40 minutes, on a PC with 2.8 GHz clock rate.

4.6.1 Number processing task

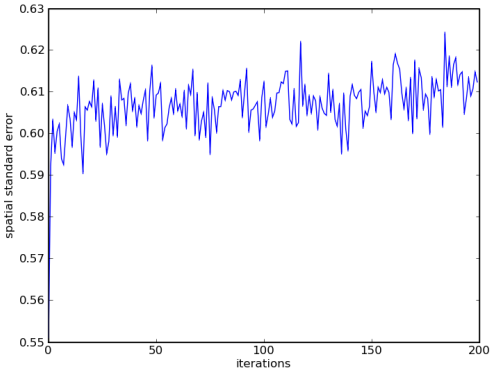


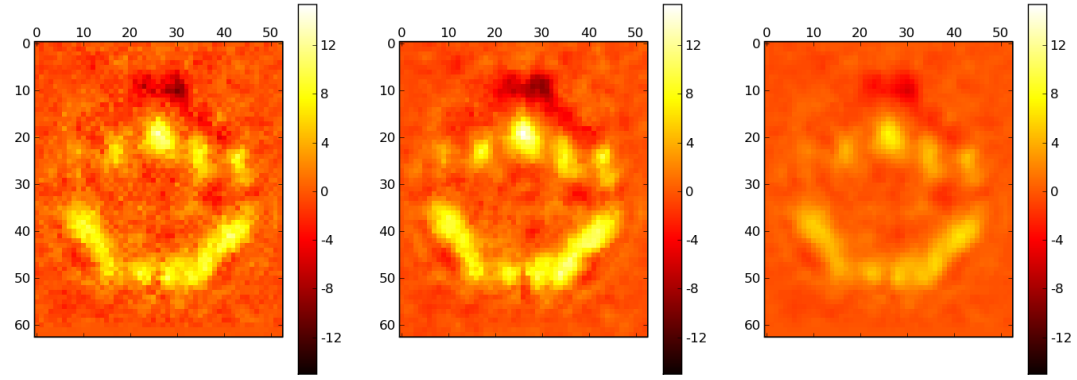
FIGURE 4.12: Posterior sampling of σ_S on real fMRI data.

Behavior of the Markov chain is illustrated in Figure 4.12, by the sampled values of the standard spatial displacement σ_S. At first glance, it seems to settle almost instantly in a stable state, presumably next to a local mode of the posterior distribution.

However, the average value of σ_S increases slightly across the iterations, suggesting that the chain is in fact progressing extremely slowly across the parameter space, and is still far from its stationary distribution. This implies that the Monte-Carlo approximation to the posterior mean, $\hat{\sigma}_S = 0.61$ voxels, corresponding to 4.3mm, is probably lower than the true posterior mean.

Even so, the effect of spatial modeling on the posterior mean estimation is clearly visible in Figure 4.13. The posterior mean effect map under spatial uncertainty (middle) is clearly smoother than the map obtained under no spatial uncertainty (left). It is also slightly more contrasted, with values ranging from -14.9 to 15.7 against -13.8 to 12.7 when not accounting for spatial uncertainty.

FIGURE 4.13: Posterior estimates of $\boldsymbol{\mu}$ for a number processing task.



No Spatial Uncertainty | Spatial Uncertainty | No Spatial Uncertainty + Smoothing

This regularization effect can be explained by the fact that the mean effect map is averaged across a large number of displacements, sampled according to their posterior likelihood, resulting in a smoothing effect. Thus, it is natural to wonder whether spatial modeling does not act merely as a spatial filter, comparable to the linear and isotropic smoothing customarily applied to fMRI data.

We investigated this question by smoothing the posterior mean effect map obtained without spatial uncertainty, using a Gaussian filter with FWHM equal to $4,3mm$, according to the standard spatial displacement estimated in the model with spatial uncertainty. As can be seen in Figure 4.13, right, the result is very different from the spatially uncertain posterior mean effect map, in that the image is smoother, but with less contrast, the values ranging from -12.3 to 11.6.

Thus, the regularization induced by our spatial modeling approach is seen to be highly anisotropic, as it enhances the salient features of the effect map, while reducing the background noise. This effect is reminiscent of other Bayesian hierarchical modeling approaches to fMRI data analysis discussed in Sections 2.2.3 and 2.6.1, such as [Gössl et al., 2001, Woolrich et al., 2005, Woolrich and Behrens, 2006, Smith and Fahrmeir, 2007, Penny et al., 2007, Makni et al., 2008]. In all of these, a similar anisotropic smoothing effect is obtained by explicitly modeling correlations between neighboring voxels through a hidden discrete-valued Markov field. This Markov field describes the *state* of each voxel, *i.e.*, whether it is active, inactive, or de-activated. Yet another approach can be found

in [Harrison et al., 2008], which uses a non-stationary spatial regularisation prior on the effect map, which promotes regularization while preserving activation contours.

However, it is worth noting that all the above methods are concerned with single-subject analysis only, and promote regularization by imposing constraints on the subject's activation map. This is very different from our spatial uncertainty model (4.2), which is defined at the between-subject level. Indeed, the regularization of the group effet map $\boldsymbol{\mu}$ is essentially a by-product of integrating out unknown spatial deformations $\mathbf{u}$ applied to the individual activation maps. The posterior distributions of these deformations are more sharply peaked in regions were the individual images are easily matched (such as at the border of an activation region), than in regions where the matching is unclear (such as the background). This explains how the anisotropic smoothing effect is obtained. Their is little chance that the exact same effect could be achieved by only modeling the group effect map $\boldsymbol{\mu}$, according to one of the approaches cited above, which would ignore correlations between individual images.

4.6.2 Language processing task

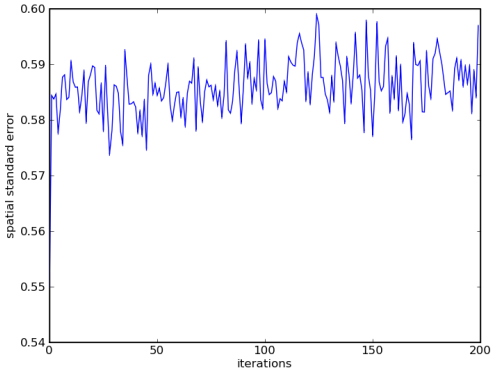


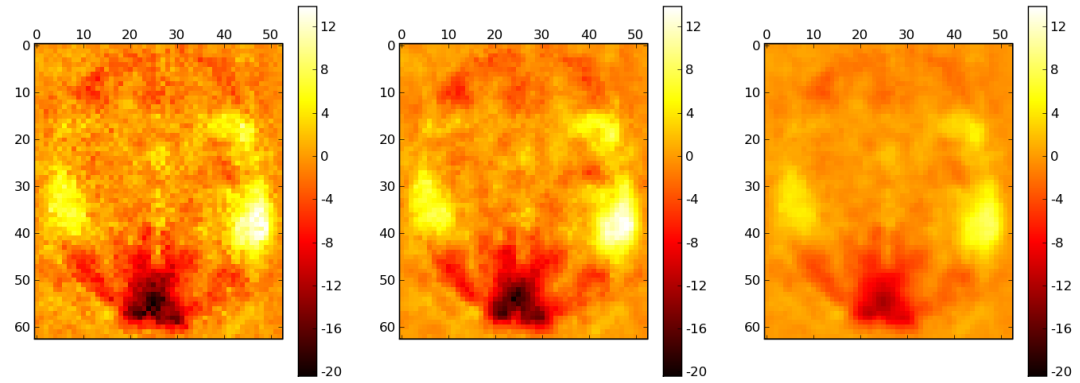
FIGURE 4.14: Posterior sampling of σ_S on real fMRI data.

We obtained similar results when analyzing data from a language processing task. The estimate of the standard spatial displacement $\hat{\sigma}_S = 0.59$, corresponding to a FWHM of $4mm$, was only slightly smaller, and the mixing of the Markov chain is also seen to be very slow from Figure 4.14.

The gain in contrast and regularizing effect due to spatial modeling is also apparent from Figure 4.15, and we noticed the same marginal increase in contrast, with extremal values

of -33.4 and 20.0 against -31.6 and 19.0 This combined effect could not be reproduced by a simple isotropic smoothing with the same FWHM value of $4mm$.

FIGURE 4.15: Posterior estimates of $\boldsymbol{\mu}$ for a language processing task.



No Spatial Uncertainty | Spatial Uncertainty | No Spatial Uncertainty + Smoothing

4.7 Conclusion

We have devised an approach for Bayesian inference on a spatially distorted signal, which can be applied to fMRI data. On simulated data, this approach estimated the original signal from noisy, warped observations with significantly more accuracy than when ignoring spatial uncertainty. The latter resulted in a blurred estimate, and the merging of neighboring signal peaks. The benefits of spatial modeling were also assessed on a large number of simulated datasets, with different values for the noise level, sample size and deformation field smoothness

Applied to real fMRI data, our method yielded smoother and more contrasted posterior mean effect maps when modeling spatial uncertainty. This effect could not be reproduced by naive linear isotropic smoothing. These results are very encouraging, and illustrate the benefits of accounting for the spatial uncertainty present in neuroimaging data through proper statistical modeling rather than isotropic smoothing, as is the common heuristic. The main difficulty we have encountered is the slow mixing rate of the Markov chain under spatial uncertainty, resulting in a conservative estimate of spatial uncertainty. This leaves space for further improvement, through the exploitation of more sophisticated sampling techniques. Thus, the proposal density for the Metropolis

Hastings algorithm used to sample the elementary displacements is clearly sub-optimal; testing alternative proposals appears a promising line of work. Another possible strategy would be to design a *tempering* scheme to 'flatten' the landscape of the posterior distribution [Marin and Robert, 2007], which would allow the Markov chain to move more freely around the parameter space.

Chapter 5

A Bayesian model selection approach to the detection of functional networks

Abstract

In this chapter, we introduce a new paradigm for ROI-based fMRI group data analysis that overcomes certain limitations of the SPM-like approach. Using a Bayesian model selection framework, the functional network associated with a certain cognitive task is selected according to the posterior probabilities of mean regional activations, given a pre-defined parcellation of the brain. Thus our approach is threshold-free, while allowing to incorporate prior information, provided that the parcellation is sensible. Furthermore, by controlling a Bayesian risk, our approach balances false positive and false negative risks. Finally, it is based on the same spatial uncertainty model as in Chapter 4, and thus accounts for the mis-alignment of individual images, due to inevitable registration errors. As a consequence, the assignment of each voxel to a region is random rather than deterministic, and differs from one subject to another.

Results on simulated data show that badly localized effect can cause inactive regions to be detected by mistake. This bias toward false positives is reduced when modeling spatial uncertainties. However, the posterior probabilities estimates in the model with spatial uncertainty are numerically unstable, presumably because of the slow mixing of

the Metropolis-Hastings algorithm used to simulate the displacement fields. We therefore propose a more stable approximate procedure, which consists in fixing the displacement fields to their most probable value *a posteriori*. This does not entirely reduce the bias toward false positives, which is further compensated through an additional penalty on model fit. The final procedure is validated on a simulated dataset.

5.1 Introduction

In the previous chapter, we have dealt with the assumption of perfect match between individual brains. This assumption was relaxed, by incorporating into the mass univariate model specified by (2.5) and (2.6) (see Section 2.3.2) a set of hidden variables, representing spatial normalisation errors. These are modeled as multivariate random fields, as described in Section 4.3.

We now define a more general hierarchical model for the data, based on this modeling of spatial uncertainty, and on a pre-defined parcellation of the brain volume into regions that are assumed functionally homogeneous, as described in Section 5.2. This model is conveniently represented by its directed acyclic graph (DAG) [Jordan, 2003], as illustrated in Figure 5.1, showing its conditional dependence structure.

Based on this model, we define each region as *involved* in the task under study if it contains a nonzero mean effect, and *inactive* otherwise. Thus the unknown functional network we wish to recover is defined as a partition of regions into involved and inactive. Each partition defines a different generative model for the data. Therefore, selecting the right functional network can be seen as a model selection problem. In Section 5.3, we propose to do so in a Bayesian framework, rating each candidate network in terms of its posterior probability, given a prior distribution over all model parameters, defined in Section 5.4.

Because regions are defined beforehand, this method is threshold-free, while taking advantage of the prior knowledge about functionally homogeneous regions, provided that the parcellation is sensible. By controlling a Bayesian risk, our approach balances false positives and false negatives, with relative weights that may be tuned depending on the application.

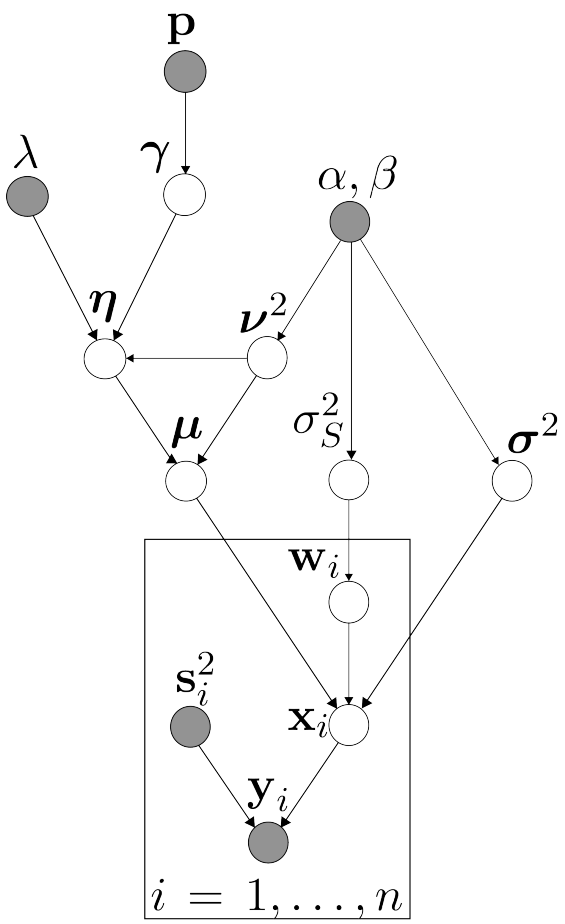


FIGURE 5.1: Directed acyclic graph (DAG) of the full hierarchical model. Gray circles correspond to fixed quantities (hyperparameters and observations), white circles to random quantities (latent variables). The conditional distribution of the subject-specific hidden effects $\mathbf{x}_i$ and their estimation $\mathbf{y}_i$, are described in Section 5.2.2, while the elementary displacements $\mathbf{w}_i$ are modeled in Section 4.3; the distribution of the population mean effect $\boldsymbol{\mu}$, conditional on the regional parameters $(\boldsymbol{\eta}, \boldsymbol{\nu}^2)$, is defined in Section 5.2. Finally, the prior distribution on the model parameters $(\boldsymbol{\eta}, \boldsymbol{\nu}^2, \boldsymbol{\sigma}^2, \sigma_S^2)$ and the indicator variable $\boldsymbol{\gamma}$, is specified in Section 5.4.

The idea of using pre-defined regions of interest (ROIs) for fMRI group data analysis has already been exploited in [Bowman et al., 2008], but under the implicit assumption that individual images are perfectly registered. As shown in Section 4.5, activations estimated by mass univariate detection methods that neglect the spatial uncertainty caused by inevitable registration errors are swelled. This 'stretching' effect may in turn cause inactive regions to be detected by error, a fact illustrated in Section 5.7.

One of the main contributions of our work is to relax this assumption of perfect registration, by combining the spatial uncertainty model presented in Chapter 4 with the regional response model in Section 5.2. Thus for each subject, the membership of a

voxel to a given region is probabilistic rather than deterministic. Because the membership probability accounts for the subject's own functional data, this effectively allows to de-weight the contribution of individual activations to regions they are likely not to belong to, but rather have been projected onto due to spatial normalization errors. In Section 5.7, we show that accounting for this uncertainty indeed makes it possible to reduce the risk of falsely selecting a region as active. In contrast, using a probabilistic atlas, such as found in FSL, would not provide such subject-specific information, but rather inform on the uncertainty in the definition of the region itself.

The main technical difficulty of our approach is the computation of the posterior probability of each functional network in the model under spatial uncertainty, because it involves evaluating complex integrals on high dimensional spaces. Section 5.5 describes how these integrals can be evaluated numerically, using Monte-Carlo Markov Chain techniques. However, these techniques turn out to be too computer intensive and numerically unstable to be used in practice. Instead, in Section 5.8 we introduce an approximation based on posterior modes, which requires much less computer time, and is more stable numerically. However, this approximation is less efficient in compensating the stretching effect due to displaced activations, so that the bias toward false positives, though reduced, is still present. In Section 5.8.3, we compensate for this residual bias by incorporating an additional penalty on model fit, calibrated on simulated data. The final procedure is validated on a simulated dataset in Section 5.8.4, and we conclude by a discussion in Section 5.9.

5.2 Regional response model

Our approach to functional network selection is based on an *a priori* partition of the search volume $\mathcal{V}$ into N disjoint regions of interest $\mathcal{V} = \mathcal{V}_1 \cup \ldots \cup \mathcal{V}_N$, assumed to be homogeneous functional areas. More precisely, we re-define the population mean effects μ_k as spatially independent Gaussian random variables, identically distributed within each region.

Thus, for all region $j = 1, \ldots, N$, and for all voxel k such that $\mathbf{v}_k \in \mathcal{V}_j$:

$$\mu_k \quad = \quad \eta_j + \chi_k; \quad \chi_k \overset{i.i.d.}{\sim} \mathcal{N}(0, \nu_j^2), \tag{5.1}$$

In other terms, μ_k, previously defined as a voxelwise fixed effect (see Section 2.3.2), is now expressed as the sum of a regional fixed effect η_j, representing the average BOLD response in region j, and a voxelwise random effect χ_k representing the variability of the response across voxels.

The same idea of modeling fMRI data based on fixed parcels containing voxels with similar BOLD responses can be found in [Bowman et al., 2008] (see Section 2.6.2). Furthermore, in this work the regional means are modeled jointly as a Gaussian vector, with arbitrary covariance matrix, allowing to measure functional correlations between distant regions, as well as between voxels from a same region. Though modeling these correlations is not our primary goal, and has not been considered here for simplicity, it does constitute a promising direction in which our approach could be extended.

5.2.1 Generative model without spatial uncertainty

We start by assuming that the estimated effect maps $(\mathbf{y}_i)_{1\leq i\leq n}$ are perfectly aligned. Under the assumptions of the regional model (5.1), the within-subject (2.5) and between-subject (2.6) models defined in Section 2.3.2, we specify a generative model for the the data, with the further assumption that the between-subject variance σ_k^2 in (2.6) is now uniform within each region j. As discussed in Section 4.2, this modification will be necessary when modeling spatial uncertainty in the next section, to avoid overfitting the data.

Thus, for all subject $i = 1, \ldots, n$, all region $j = 1, \ldots, N$ and all voxel k such that $\mathbf{v}_k \in \mathcal{V}_j$:

$$y_{i,k} = x_{i,k} + \varepsilon_{i,k}; \quad \varepsilon_{i,k} \overset{ind.}{\sim} \mathcal{N}(0, s_{i,k}^2) \tag{5.2}$$

$$x_{i,k} = \mu_k + \xi_{i,k}; \quad \xi_{i,k} \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma_j^2), \tag{5.3}$$

where we assume the mutual independence of the noise processes $\xi_{i,k}$ and $\varepsilon_{i,k}$.

Using the expression of μ_k in (5.1), we can re-write the above model as:

$$y_{i,k} = \eta_j + \chi_k + \xi_{i,k} + \varepsilon_{i,k}. \tag{5.4}$$

This formulation corresponds to a mixed-effect analysis of variance (MFX-ANOVA) model with a single factor, whose N levels correspond to the different regions. In fact, since the noise processes are assumed independent across voxels, it is simply the aggregation of N independent models, one for each region j.

This generalizes the mass univariate model, defined in Section 2.3.2, which corresponds to the limiting case $N = d$, that is, when regions reduce to single voxels, under the identifiability constraint: $\nu_j^2 \equiv 0$. In this case, the regional mean η_j coincides with the population mean effect μ_k, and the regional between-subject variance σ_j^2 becomes voxel dependent.

5.2.2 Generative model with spatial uncertainty

We now relax the assumption that the individual images are perfectly normalized, so that the between-subject model (2.6) is replaced by its generalization to spatial uncertainty (4.2), introduced in Chapter 4. Spatial normalization errors are defined for each subject i as a spatial deformation field $\mathbf{u}_i$, controlled by a set of hidden variables $\mathbf{w}_i$, as explained in Section 4.3.

As in the previous section, we combine this observation model with the regional model (5.1), and assume the between-subject variance σ^2 to be region-dependent. It turns out that, conditionally on the hidden displacements variables $\mathbf{w}_i$, the data is still distributed according to a MFX-ANOVA model, independently across regions. However, (5.2) must be adapted to account for observations being displaced across regions.

Consequently, for all subject $i = 1, \ldots, n$, all region $j = 1, \ldots, N$, we have for all voxels k such that $\mathbf{v}_k + \mathbf{u}_{i,k} \in \mathcal{V}_j$:

$$\begin{aligned} y_{i,k} &= x_{i,k} + \varepsilon_{i,k}; \quad \varepsilon_{i,k} \overset{ind.}{\sim} \mathcal{N}(0, s_{i,k}^2) \\ x_{i,k} &= \boldsymbol{\mu}(\mathbf{v}_k + \mathbf{u}_{i,k}) + \xi_{i,k}; \quad \xi_{i,k}|\mathbf{w}_i \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma_j^2). \end{aligned} \tag{5.5}$$

Again, using (5.1), this can be re-written as:

$$y_{i,k}|\mathbf{w}_i = \eta_j + \boldsymbol{\chi}(\mathbf{v}_k + \mathbf{u}_{i,k}) + \xi_{i,k} + \varepsilon_{i,k}. \tag{5.6}$$

Thus, the main difference with the case of no spatial uncertainty is that the mean of any given observation $y_{i,k}$ now depends on the region j toward which the voxel k is displaced, conditional on $\mathbf{w}_i$, rather than the region it belongs to. The same observation holds for the conditional variance σ_j^2 of the hidden effects $x_{i,k}$. Finally, because the set of voxels displaced to region j fluctuates with $\mathbf{w}$, the data is no longer independent across regions, after integrating out $\mathbf{w}$.

This model generalizes all the ones introduced in the previous sections and chapters of this work. Indeed, the regional model without spatial uncertainty is the special case $\sigma_S = 0$, while the mass univariate model in Section 2.3.2, is obtained for $\sigma_S = 0$ and $N = d$. Finally, the model introduced in Chapter 4 corresponds to the case of a single region ($N = 1$).

5.3 Bayesian model selection framework

Based on the regional response model, we now propose an answer to the basic question underlying most fMRI paradigms, namely, that of selecting the subset of regions that are involved in the task at hand. More precisely, we define region j as:

- *active* if $\eta_j > 0$;
- *negatively active* if $\eta_j < 0$;
- *inactive* if $\eta_j = 0$,

and define a region j as *involved* if it is either active or negatively active. From a hypothesis testing perspective, we wish to test, for each $j = 1, \ldots, N$, $\mathcal{H}_{0,j}$: '$\eta_j = 0$' versus $\mathcal{H}_{1,j}$: '$\eta_j \neq 0$'.

Alternatively, recovering the subset of regions that are involved in the considered task can be seen as a model selection problem, by introducing the indicator variable $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_N) \in \{0,1\}^N$, such that for all region j: $\gamma_j = 0$ means that $\eta_j = 0$, and $\gamma_j = 1$ means that $\eta_j \neq 0$. Each of the 2^N possible values of the indicator variable $\boldsymbol{\gamma}$ then defines a different generative model $\mathfrak{M}_{\boldsymbol{\gamma}}$ describing the data $\mathbf{y}$, corresponding to a different subset of explanatory variables $\boldsymbol{\eta}_{\boldsymbol{\gamma}} = (\eta_j)_{j,\,\gamma_j=1}$. In the following, we will refer to

$\boldsymbol{\gamma}$ as the *functional network* variable, since it defines the network of regions functionally engaged in the task under study.

According to these notations, our task consists in selecting the 'best' (in a certain sense) model $\mathfrak{M}_{\boldsymbol{\gamma}}$ to describe the data $\mathbf{y}$, *i.e.*, the best segmentation of the regions into involved and inactive. Equivalently, it can be formulated as an estimation problem, with $\boldsymbol{\gamma}$ as the interest parameter. The Bayesian approach to this problem consists in considering the unknown quantities $\boldsymbol{\gamma}$ and $\boldsymbol{\theta} = (\boldsymbol{\eta}, \boldsymbol{\nu}^2, \boldsymbol{\sigma}^2, \sigma_S^2)$ as random variables, and defining a prior density $\pi(\boldsymbol{\theta}|\boldsymbol{\gamma})\pi(\boldsymbol{\gamma})$, representing the prior knowledge on their possible values.

Note that, from a frequentist viewpoint, limiting $\boldsymbol{\theta}$ to the variables cited above makes sense, since these are the unknown, but fixed, quantities, we wish to infer, as opposed to $\mathbf{z} = (\mathbf{x}, \mathbf{w}, \boldsymbol{\mu})$, which are unknown random variables considered as nuisance factors, and marginalized out during the inference. In the Bayesian setting we have adopted on the other hand, this opposition has no real justification, since both $\boldsymbol{\theta}$ and $\mathbf{z}$ are vectors of hidden random variables, whose posterior distribution we wish to determine. However, we will see that these notations are convenient in the context of the model selection algorithm proposed in Section 5.5.1. In particular, the expression of the complete density $f(\mathbf{y}, \mathbf{z}|\boldsymbol{\theta})$ is that of a curved exponential model, allowing the evaluation of the MAP estimates of $\boldsymbol{\theta}$ using an efficient MCMC-SAEM algorithm, as described in Appendix E.

Based on the prior $\pi(\boldsymbol{\theta}|\boldsymbol{\gamma})\pi(\boldsymbol{\gamma})$ and on the likelihood function $f(\mathbf{y}|\boldsymbol{\theta})$ of the full hierarchical model specified by (5.5), (4.3) and (4.4), the posterior distribution of $\boldsymbol{\theta}$ under each model $\mathfrak{M}_{\boldsymbol{\gamma}}$ can be computed according to Bayes' theorem:

$$\pi(\boldsymbol{\theta}|\mathbf{y}, \boldsymbol{\gamma}) = \frac{f(\mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}|\boldsymbol{\gamma})}{\int f(\mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}|\boldsymbol{\gamma})d\boldsymbol{\theta}}, \tag{5.7}$$

where

$$m(\mathbf{y}|\boldsymbol{\gamma}) = \int f(\mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}|\boldsymbol{\gamma})d\boldsymbol{\theta} \tag{5.8}$$

is the *marginal likelihood*, or *evidence*, of model $\mathfrak{M}_{\boldsymbol{\gamma}}$. This quantity is central to Bayesian model selection, since the posterior distribution of the indicator variable can be expressed as:

$$\pi(\boldsymbol{\gamma}|\mathbf{y}) = \frac{m(\mathbf{y}|\boldsymbol{\gamma})\pi(\boldsymbol{\gamma})}{\sum_{\boldsymbol{\gamma}} m(\mathbf{y}|\boldsymbol{\gamma})\pi(\boldsymbol{\gamma})}. \tag{5.9}$$

Assuming that $m(\mathbf{y}|\boldsymbol{\gamma})$ can be computed for all possible values of $\boldsymbol{\gamma}$, the most probable functional network given the data can be selected, according to:

$$\begin{aligned} \hat{\boldsymbol{\gamma}} &= \arg\max_{\boldsymbol{\gamma}} \pi(\boldsymbol{\gamma}|\mathbf{y}), \\ &= \arg\max_{\boldsymbol{\gamma}} m(\mathbf{y}|\boldsymbol{\gamma})\pi(\boldsymbol{\gamma}). \end{aligned}$$

Besides being intuitive, this procedure has a number of attractive features. For instance, because $\boldsymbol{\gamma}$ takes only discrete values, its MAP estimate $\hat{\boldsymbol{\gamma}}$ is also the Bayesian estimator associated to the $0-1$ loss function $\ell(\boldsymbol{\gamma}, \tilde{\boldsymbol{\gamma}}) = \sum_j \mathbf{1}_{\gamma_j \neq \tilde{\gamma}_j}$, *i.e.*, it minimizes over all possible networks $\tilde{\boldsymbol{\gamma}}$ the Bayesian risk:

$$E[\ell(\boldsymbol{\gamma}, \tilde{\boldsymbol{\gamma}})|\mathbf{y}] = \sum_{\boldsymbol{\gamma} \in \{0,1\}^N} \ell(\boldsymbol{\gamma}, \tilde{\boldsymbol{\gamma}})\pi(\boldsymbol{\gamma}|\mathbf{y}). \tag{5.10}$$

From a multiple testing perspective, this means that $\hat{\boldsymbol{\gamma}}$ minimizes the *a posteriori* expected sum of type I (false positive) and type II errors (false negative) errors. Other loss functions could be used, to give different weights to false positive and false negative risks. Thus our Bayesian decision-theoretic framework answers one of the limitations of the SPM-like approach, which only controls type I risks.

5.4 Prior specification

We now address the choice of a prior distribution for the functional network $\boldsymbol{\gamma}$, and the model parameters $\boldsymbol{\theta}$. Under limited information about their possible values, it is natural to use a weak prior. This choice is problematic in generalized linear mixed models (GLMMs), because Jeffreys prior may be intractable, and the standard scale invariant (improper) prior may lead to an improper posterior [Hobert and Casella, 1996]. Instead, we have adopted normal priors on the effects and inverse-Gamma priors for variance

components. These are practical choices, their conditional conjugate form making them amenable to posterior sampling. Our prior is specified as follows:

5.4.1 Regional means

For each region j, we define:

$$\begin{aligned} \eta_j|\nu_j^2, \gamma_j = 0 &\sim \delta(0) \\ \eta_j|\nu_j^2, \gamma_j = 1 &\sim \mathcal{N}(m, \nu_j^2/\lambda), \end{aligned} \tag{5.11}$$

where $\delta(0)$ is the Dirac mass in 0, meaning that in inactive regions, the regional mean vanishes almost surely; in the other regions, the prior mean is set to $m = 0$ so as not to bias the inference towards positive or negative effects and, the scale parameter is set to $\lambda = 10^{-3}$, which may be interpreted as the weight given to the prior mean m with respect to one observation.

5.4.2 Variance components

All variance components share the same prior:

$$\pi(\sigma_S^2, \boldsymbol{\nu}^2, \boldsymbol{\sigma}^2) = \mathcal{IG}(\sigma_S^2; \alpha, \beta) \prod_{j=1}^{N} \left\{ \mathcal{IG}(\nu_j^2; \alpha, \beta) \mathcal{IG}(\sigma_j^2; \alpha, \beta) \right\} \tag{5.12}$$

where $\mathcal{IG}(\alpha, \beta)$ is the Inverse-Gamma distribution with parameters (α, β), and density function

$$\mathcal{IG}(\, z\,; \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} z^{-\alpha-1} \exp\left(\frac{-\beta}{z}\right).$$

Tuning α, and β is a difficult task in absence of prior information. A popular choice is to use a "just" proper prior in absence of prior knowledge, given for instance by $\alpha = \beta = 10^{-3}$ [Spiegelhalter et al., 1996]. This common practice has raised some concerns [Natarajan and McCulloch, 1998], because they can result in poorly behaved posterior sampling schemes, due to the near impropriety of the resulting posterior.

Thus we chose values which reflected the limited amount of knowledge available on the interest parameters, while defining truly proper priors, setting $\alpha = 3, \beta = 20$. This means for instance that in each region j, the population variance σ_j^2 has a prior mean of 10, and a prior variance of 10^2, which seems reasonable from practical experience. The same considerations hold for the regional variances ν_j^2, and the spatial variance parameter σ_S^2.

Several alternative priors are proposed in [Natarajan and Kass, 2000], including a uniform shrinkage prior, and an approximate Jeffreys prior. However, we found the results obtained by the standard priors satisfying enough not to consider these more complex strategies.

5.4.3 Indicator variables

We define independent Bernoulli priors for the indicators variables:

$$\pi(\boldsymbol{\gamma}) \quad = \quad \prod_{j=1}^{N} \mathcal{B}(\gamma_j; 1, p_j), \tag{5.13}$$

meaning that region j has a prior probability of p_j of being activated. In the following, we use the default choice $p_j \equiv 0.5$, but prior information on the state of regions can easily be included at this stage, such as results of previous analyses.

5.5 Evaluating the marginal likelihood

We now address the computation of the marginal likelihood $m(\mathbf{y}|\boldsymbol{\gamma})$, defined by (5.8). As is often the case, this is a difficult task, because it requires integrating the probability density function (pdf) $f(\mathbf{y}|\boldsymbol{\theta})$, defined on a high dimensional space, with respect to the prior density $\pi(\boldsymbol{\theta}|\boldsymbol{\gamma})$, which cannot be done analytically in our case.

Thus, we turn to numerical approximation strategies. The main challenge in evaluating $m(\mathbf{y}|\boldsymbol{\gamma})$ is that the integration must be done with respect to the prior density $\pi(\boldsymbol{\theta}|\boldsymbol{\gamma})$, but the values significantly contributing to the integral are concentrated around the high density points of the posterior $\pi(\boldsymbol{\theta}|\mathbf{y}, \boldsymbol{\gamma})$. We start by briefly reviewing some of the main

strategies that have been developed to deal with this classical problem, and discuss their relevance to the present case.

Importance sampling

A naive way of estimating $m(\mathbf{y}|\boldsymbol{\gamma})$, justified by (5.8), would be to generate a sample $(\boldsymbol{\theta}_g)_{1\leq g\leq G}$ from the prior distribution $\pi(\boldsymbol{\theta}|\boldsymbol{\gamma})$, and deduce the following Monte-Carlo estimate:

$$\hat{m}(\mathbf{y}|\boldsymbol{\gamma}) = G^{-1}\sum_{g=1}^{G} f(\mathbf{y}|\boldsymbol{\theta}_g).$$

As mentioned earlier, the significant values of the likelihood function $f(\mathbf{y}|\boldsymbol{\theta})$ are concentrated on a small region of the support of $\pi(\boldsymbol{\theta}|\boldsymbol{\gamma})$, so most of the sampled values $f(\mathbf{y}|\boldsymbol{\theta}_g)$ are likely to be close to zero. Hence, this estimate, though unbiased, would have such a high variance that it would be useless.

A natural alternative to sampling under the prior, known as *importance sampling*, is to use a proposal distribution $q(\boldsymbol{\theta})$ instead, noting that:

$$m(\mathbf{y}|\boldsymbol{\gamma}) = \int \frac{f(\mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}|\boldsymbol{\gamma})}{q(\boldsymbol{\theta})} q(\boldsymbol{\theta})\boldsymbol{d}\theta.$$

Thus, if $(\boldsymbol{\theta}_g)_{1\leq g\leq G}$ is a sample from $q(\boldsymbol{\theta})$, an unbiased estimate of $m(\mathbf{y}|\boldsymbol{\gamma})$ is given by:

$$\hat{m}_q(\mathbf{y}|\boldsymbol{\gamma}) = G^{-1}\sum_{g=1}^{G} \frac{f(\mathbf{y}|\boldsymbol{\theta}_g)\pi(\boldsymbol{\theta}_g|\boldsymbol{\gamma})}{q(\boldsymbol{\theta}_g)}.$$

The likelihood function $f(\mathbf{y}|\boldsymbol{\theta})$ is unknown in the model with spatial uncertainty, but this method could still be applied, replacing $\boldsymbol{\theta}$ by $(\boldsymbol{\theta}, \mathbf{w})$, since the density conditional on the displacement parameters $\mathbf{w}$ can be computed explicitly (see Appendix F).

Thus, $\hat{m}_q(\mathbf{y}|\boldsymbol{\gamma})$ is still an unbiased estimator of $m(\mathbf{y}|\boldsymbol{\gamma})$. Its variance, which depends on the proposal, is given by:

$$V\hat{m}_q(\mathbf{y}|\boldsymbol{\gamma}) = \frac{1}{G}\int \left(\frac{f(\mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}|\boldsymbol{\gamma})}{q(\boldsymbol{\theta})} - m(\mathbf{y}|\boldsymbol{\gamma})\right)^2 q(\boldsymbol{\theta})d\boldsymbol{\theta}.$$

In particular, the variance tends to zero as $q(\boldsymbol{\theta})$ gets closer to the posterior density $\pi(\boldsymbol{\theta}|\mathbf{y}, \boldsymbol{\gamma})$. Thus, the main difficulty of this approach is to choose $q(\boldsymbol{\theta})$ 'close to' $\pi(\boldsymbol{\theta}|\mathbf{y}, \boldsymbol{\gamma})$

which can indeed be hard when little is known on the shape of the posterior density, as is the case here.

A generalization of importance sampling is bridge sampling [Meng and Wong, 1996], based on the following identity:

$$m(\mathbf{y}|\boldsymbol{\gamma}) = \frac{\int f(\mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}|\boldsymbol{\gamma})h(\boldsymbol{\theta})q(\boldsymbol{\theta})\boldsymbol{d}\theta}{\int q(\boldsymbol{\theta})h(\boldsymbol{\theta})\pi(\boldsymbol{\theta}|\mathbf{y},\boldsymbol{\gamma})d\boldsymbol{\theta}},$$

valid for any choice of functions h and q. Given a sample $(\boldsymbol{\theta}_g)_{1\leq g\leq G}$ from $q(\boldsymbol{\theta})$ as before, and a sample $(\boldsymbol{\theta}_j)_{1\leq j\leq J}$ from the posterior density $\pi(\boldsymbol{\theta}|\mathbf{y},\boldsymbol{\gamma})$ (in our case, replacing $\boldsymbol{\theta}$ by $(\boldsymbol{\theta},\mathbf{w})$, this can be done by by the Metropolis-with Gibbs algorithm in Appendix D), an unbiased estimate estimate of $m(\mathbf{y}|\boldsymbol{\gamma})$ is:

$$\hat{m}_{BS}(\mathbf{y}|\boldsymbol{\gamma}) = \frac{G^{-1}\sum_{g=1}^{G} f(\mathbf{y}|\boldsymbol{\theta}_g)\pi(\boldsymbol{\theta}_g|\boldsymbol{\gamma})h(\boldsymbol{\theta}_g)}{J^{-1}\sum_{j=1}^{J} q(\boldsymbol{\theta}_j)h(\boldsymbol{\theta}_j)}.$$

This method requires the choice of an additional function $h(\boldsymbol{\theta})$, and the algorithm reduces to importance sampling when $h(\boldsymbol{\theta}) \equiv 1$. [Meng and Wong, 1996] show that certain choices of h can reduce the variance of the classical importance sampling estimate, and indicate iterative strategies to choose h. However, the choice of a proposal q remains an issue in our case, and it is not clear how this method performs for 'bad choices' (what happens for instance if q is chosen equal to the prior?). Thus, it is not clear how importance Monte-Carlo sampling strategies could be applied in a simple way to the present problem.

Harmonic mean estimator

The harmonic mean estimator (HME) in [Raftery et al., 2007] constitutes a very simple strategy to estimate the marginal likelihood from the output of any posterior sampling scheme, based on the *harmonic mean identity*:

$$\frac{1}{m(\mathbf{y}|\boldsymbol{\gamma})} = \int \frac{1}{f(\mathbf{y}|\boldsymbol{\theta})}\pi(\boldsymbol{\theta}|\mathbf{y},\boldsymbol{\gamma})\boldsymbol{d}\theta.$$

Thus, given a sample $(\boldsymbol{\theta}_j)_{1\leq j\leq J}$ from the posterior density, an unbiased estimate of the marginal likelihood is given by:

$$\hat{m}_{HM}(\mathbf{y}|\boldsymbol{\gamma}) = \left(J^{-1} \sum_{j=1}^{J} \frac{1}{f(\mathbf{y}|\boldsymbol{\theta}_j)} \right)^{-1}.$$

As with importance sampling, this method could easily be applied in our case, replacing $\boldsymbol{\theta}$ by $(\boldsymbol{\theta}, \mathbf{w})$, and sampling their joint posterior density using the sampling scheme described in Appendix D.

In spite of its appealing simplicity, the HME is known for its lack of numerical stability, as measured by the variance: $\mathrm{Var}[f(\mathbf{y}|\boldsymbol{\theta})^{-1}|\mathbf{y}]$, making its applicability hazardous. Indeed, this variance is determined by the second moment:

$$\begin{aligned} \mathrm{E}[f(\mathbf{y}|\boldsymbol{\theta})^{-2}|\mathbf{y}] &= \int \frac{1}{f(\mathbf{y}|\boldsymbol{\theta})^2} \pi(\boldsymbol{\theta}|\mathbf{y}, \boldsymbol{\gamma}) d\theta \\ &= \frac{1}{m(\mathbf{y}|\boldsymbol{\gamma})} \int \frac{\pi(\boldsymbol{\theta}|\boldsymbol{\gamma})}{f(\mathbf{y}|\boldsymbol{\theta})} d\theta. \end{aligned} \tag{5.14}$$

For the above integral to be finite, the prior distribution $\pi(\boldsymbol{\theta}|\boldsymbol{\gamma})$ must have lighter tails than the density $f(\mathbf{y}|\boldsymbol{\theta})$ (viewing the latter as a function of $\boldsymbol{\theta}$). Intuitively, this means that the HME is numerically stable when the prior is more sharply peaked, *i.e.*, provides more information on the parameter of interest, than the data! [Raftery et al., 2007] show for instance that this is not the case in the simple Gaussian model, with the usual conjugate inverse Gamma-Gaussian prior distribution on the Gaussian mean and variance (as defined in Section 5.4).

[Raftery et al., 2007] indicate possible ways of stabilizing $\hat{m}_{HM}(\mathbf{y}|\boldsymbol{\gamma})$, which consist in replacing $f(\mathbf{y}|\boldsymbol{\theta}_j)$ by a marginalized version $f(\mathbf{y}|h(\boldsymbol{\theta}_j))$, for an arbitrary function h. It is shown that in some cases the modified estimator has finite variance. However, the choice of h is strongly model-dependent, and we have found in our case no such convenient function ensuring finite variance for the HME.

Variational Bayes

The variational Bayes (VB) framework [Beal, 2003] can be used to compute a lower bound on the marginal likelihood. This bound originates in an identity similar to that

upon which the importance sampling approach is based,

$$m(\mathbf{y}|\boldsymbol{\gamma}) = \int q(\mathbf{z},\boldsymbol{\theta}) \frac{f(\mathbf{y},\mathbf{z},\boldsymbol{\theta}|\boldsymbol{\gamma})}{q(\mathbf{z},\boldsymbol{\theta})} d\mathbf{z} d\boldsymbol{\theta},$$

valid for any proposal distribution $q(\mathbf{z},\boldsymbol{\theta})$, with $\mathbf{z} = (\mathbf{x},\boldsymbol{\mu},\mathbf{w})$. Taking the logarithm on both sides, and applying Jensen's inequality yields (thanks to the concavity of the logarithm)

$$\log m(\mathbf{y}|\boldsymbol{\gamma}) \geq \int q(\mathbf{z},\boldsymbol{\theta}) \log \frac{f(\mathbf{y},\mathbf{z},\boldsymbol{\theta}|\boldsymbol{\gamma})}{q(\mathbf{z},\boldsymbol{\theta})} d\mathbf{z} d\boldsymbol{\theta}.$$

Thus we obtain a lower bound on the marginal likelihood, noted $\mathcal{F}_{\boldsymbol{\gamma}}(q)$ in the following. This inequality becomes an equality for $q(\mathbf{z},\boldsymbol{\theta}) = \pi(\mathbf{z},\boldsymbol{\theta}|\mathbf{y},\boldsymbol{\gamma})$, but of course this proposal cannot be used directly, because its normalizing constant is unknown, since it is precisely the marginal likelihood we wish to compute.

The goal of VB approaches is to maximize $\mathcal{F}_{\boldsymbol{\gamma}}(q)$ with respect to q, restricting the latter to a class $\mathcal{C}$ of functions such that the maximization may be performed efficiently. Typically, $\mathcal{C}$ is defined as a set of functions which are factored across several blocks of latent variables, for instance $\mathcal{C} = \{q|\, q(\mathbf{z},\boldsymbol{\theta}) = q_{\mathbf{z}}(\mathbf{z})\, q_{\boldsymbol{\theta}}(\boldsymbol{\theta})\}$. In this case, $\mathcal{F}_{\boldsymbol{\gamma}}(q)$ can be maximized using the VBEM algorithm, which alternates maximizations over $q_{\mathbf{z}}(\mathbf{z})$ and $q_{\boldsymbol{\theta}}(\boldsymbol{\theta})$, until convergence to a local maximum. It can be shown that iteration t of the algorithm is given by

$$q_{\mathbf{z}}^{(t+1)}(\mathbf{z}) \propto \exp\left[\int q_{\boldsymbol{\theta}}^{(t)}(\boldsymbol{\theta})\, \log f(\mathbf{y},\mathbf{z},\boldsymbol{\theta}|\boldsymbol{\gamma})\, d\boldsymbol{\theta}\right] \tag{5.15}$$

$$q_{\boldsymbol{\theta}}^{(t+1)}(\boldsymbol{\theta}) \propto \exp\left[\int q_{\mathbf{z}}^{(t+1)}(\mathbf{z})\, \log f(\mathbf{y},\mathbf{z},\boldsymbol{\theta}|\boldsymbol{\gamma})\, d\mathbf{z}\right]. \tag{5.16}$$

Several problems arise at this point. Even if the integral in (5.15) can be evaluated explicitly, their is little chance that the resulting expression of $q_{\mathbf{z}}^{(t+1)}(\mathbf{z})$ corresponds to a known distribution in the model with spatial uncertainty. Further factorizing $q_{\mathbf{z}}^{(t+1)}(\mathbf{z})$ does not resolve this issue, because it is due to the complicated relation between spatial displacements and the other variables, as explained in Appendix D. Because of that, the integral in (5.16) has no closed form, making the determination of $q_{\boldsymbol{\theta}}^{(t+1)}(\boldsymbol{\theta})$ problematic.

Thus, using a variational Bayes bound to approximate the marginal likelihood in the model with spatial uncertainty would most probably either require further approximations, or hybrid strategies combining MCMC and VB approaches [Forbes and Fort, 2007].

Though we do not dismiss the use of variational Bayes, it does not seem clear for the moment what advantage it would have over the other approaches, or how it could be implemented efficiently in our case.

Reversible jump MCMC

Another alternative would be to use the reversible jump MCMC (RJ-MCMC) algorithm [Green, 1995], which consists in sampling $\boldsymbol{\gamma}$ along with the other parameters. This approach allows to 'jump' from one model specified by $\boldsymbol{\gamma}$ to another, effectively changing the size of the parameter vector from one iteration to another. Thus only the models most relevant given the data are visited, rather than all the models, and a single run of the RJ-MCMC algorithm is necessary for the whole model selection procedure. The marginal are not computed directly in this approach; instead, the posterior probabilities of each model $\mathfrak{M}_{\boldsymbol{\gamma}}$ is estimated directly by the proportion of total iterations spent in $\mathfrak{M}_{\boldsymbol{\gamma}}$ by the sampling algorithm.

Though attractive, this technology has several drawbacks. [Chib and Jeliazkov, 2001] indicate that the algorithm can be quite complicated to tune in order to promote mixing across spaces of varying dimension. Also, each new model introduced in the sampling scheme must include a subset of the existing models, which artificially increases the parameter space. Finally, as noted in [Marin and Robert, 2007], the parameters of interest within each model (and the posterior probabilities $\pi(\boldsymbol{\gamma}|\mathbf{y})$) may be poorly estimated if the algorithm keeps jumping from one model to another. Hence, it is often necessary to use hybrid strategies, which alternate reversible jump and classical MCMC iterations. Finally, for each model to be well estimated, it seems that the RJ-MCMC would unavoidably require at least as many iterations as the sum of iterations needed to fit each model using more conventional MCMC techniques. Thus, it seems to us that the RJ-MCMC algorithm is less appropriate for model selection than for model averaging, and in particular to obtain estimates of the parameter $\boldsymbol{\gamma}$ which take into account the uncertainty on the choice of a model. An example of such an application in the context of fMRI group data analysis is given in [Xu et al., 2009], which is presented in Section 2.6.4. However, this is not our primary goal here.

5.5.1 Chib's approach

We have finally opted for Chib's method [Chib, 1995, Chib and Jeliazkov, 2001], which allows to compute the marginal likelihood from the output of virtually any posterior sampling scheme. It is efficient, simple and applicable in most cases. This approach is based on the basic marginal identity (BMI):

$$m(\mathbf{y}|\boldsymbol{\gamma}) \quad = \quad \frac{f(\mathbf{y}|\boldsymbol{\theta}^*)\pi(\boldsymbol{\theta}^*|\boldsymbol{\gamma})}{\pi(\boldsymbol{\theta}^*|\mathbf{y},\boldsymbol{\gamma})}, \tag{5.17}$$

which simply expresses the fact that the marginal likelihood is the normalizing constant of the posterior distribution. It is valid for any particular value $\boldsymbol{\theta}^*$ of the parameter, but it is advised to choose a high density point, such as the posterior mean or posterior maximum (MAP), where the different densities are more likely to be well estimated than in tails. In our model the MAP cannot be computed analytically, but can be numerically evaluated using the MCMC-SAEM algorithm detailed in Appendix E, while the posterior mean can be obtained from the Metropolis within Gibbs (MH-Gibbs) algorithm described in Appendix D.

Computationally, the main advantage of (5.17) is that it expresses the marginal likelihood $m(\mathbf{y}|\boldsymbol{\gamma})$ in terms of the posterior density $\pi(\boldsymbol{\theta}^*|\mathbf{y},\boldsymbol{\gamma})$, rather than the prior density, providing a way to use posterior sampling strategies, such as the Gibbs sampler, to estimate $m(\mathbf{y}|\boldsymbol{\gamma})$. Indeed, the posterior density can be written as

$$\pi(\boldsymbol{\theta}^*|\mathbf{y},\boldsymbol{\gamma}) \quad = \quad \int \pi(\boldsymbol{\theta}^*|\mathbf{z},\mathbf{y},\boldsymbol{\gamma})\pi(\mathbf{z}|\mathbf{y},\boldsymbol{\gamma})d\mathbf{z}, \tag{5.18}$$

where $\mathbf{z} = (\mathbf{x}, \mathbf{w}, \boldsymbol{\mu})$. Thus, given a sample $(\boldsymbol{\theta}_1, \mathbf{z}_1, \ldots, \boldsymbol{\theta}_J, \mathbf{z}_J)$ of the joint posterior density $\pi(\boldsymbol{\theta}, \mathbf{z}|\mathbf{y}, \boldsymbol{\gamma})$, obtained *e.g.* by a Gibbs sampler, the posterior density can be computed by *Rao-Blackwellisation* [Gelfand and Smith, 1990]:

$$\hat{\pi}(\boldsymbol{\theta}^*|\mathbf{y},\boldsymbol{\gamma}) \quad = \quad \frac{1}{J}\sum_{j=1}^{J} \pi(\boldsymbol{\theta}^*|\mathbf{z}_j,\mathbf{y},\boldsymbol{\gamma}), \tag{5.19}$$

since the $\mathbf{z}_j$'s are asymptotically drawn from the marginal $\pi(\mathbf{z}|\mathbf{y},\boldsymbol{\gamma})$. The obtained estimate is simulation consistent, *i.e.* it converges to the true value when $J \to \infty$.

This approach works nicely in the special case of no spatial uncertainty ($\sigma_S^2 = 0, \mathbf{w} = \mathbf{0}$), because then the likelihood function $f(\mathbf{y}|\boldsymbol{\theta}^*)$ is available in closed form (simply apply

the formulas in Appendix F with $\mathbf{w} = \mathbf{0}$). Furthermore, since in this case the regions are independent, as noted previously in Section 5.2, only $2N$ models need to be estimated instead of 2^N.

5.5.2 Likelihood under spatial uncertainty

Unfortunately, this appealingly simple method cannot be directly applied to the model with spatial uncertainty, because the conditional density $f(\mathbf{y}|\mathbf{w}, \boldsymbol{\theta}^*)$ only is available in closed form, but not the likelihood function:

$$f(\mathbf{y}|\boldsymbol{\theta}^*) = \int f(\mathbf{y}|\mathbf{w}, \boldsymbol{\theta}^*)\pi(\mathbf{w}|\boldsymbol{\theta}^*)d\mathbf{w}. \tag{5.20}$$

The difficulty in calculating (5.20) is essentially the same as in calculating the marginal likelihood: it is an integral on a high dimension space, with respect to the prior distribution $\pi(\mathbf{w}|\boldsymbol{\theta}^*)$, whereas significantly contributing values of the integrand are concentrated around the modes of $\pi(\mathbf{w}|\boldsymbol{\theta}^*, \mathbf{y})$. Therefore, we may apply Chib's method again, by writing the following alternative identity:

$$f(\mathbf{y}|\boldsymbol{\theta}^*) = \frac{f(\mathbf{y}|\mathbf{w}^*, \boldsymbol{\theta}^*)\pi(\mathbf{w}^*|\boldsymbol{\theta}^*)}{\pi(\mathbf{w}^*|\boldsymbol{\theta}^*, \mathbf{y})}, \tag{5.21}$$

valid for any value $\mathbf{w}^*$ of the elementary displacements. A high density value of the posterior ordinate $\pi(\mathbf{w}|\boldsymbol{\theta}^*, \mathbf{y})$ is however advisable, for accurate numerical evaluation of (5.21). Thus a reasonable choice is the conditional maximum a posterior $\mathbf{w}^*_{MAP} = \arg\max_{\mathbf{w}} \pi(\mathbf{w}|\boldsymbol{\theta}^*, \mathbf{y})$, which can be evaluated using the simulated annealing (SA) algorithm, as described in Appendix G. A simpler alternative, requiring less computations, is the posterior mean $\hat{\mathbf{w}} = \mathbb{E}[\mathbf{w}|\mathbf{y}]$, directly available from the output of the MH-Gibbs sampler in Appendix D as the average of sampled $\mathbf{w}$ values. However, it may be distant from the principal mode of $\pi(\mathbf{w}^*|\boldsymbol{\theta}^*, \mathbf{y})$, and result in a less stable estimator of (5.21).

Having chosen $\mathbf{w}^*$, both $f(\mathbf{y}|\mathbf{w}^*, \boldsymbol{\theta}^*)$ and $\pi(\mathbf{w}^*|\boldsymbol{\theta}^*)$ are available in closed form, but not the posterior ordinate $\pi(\mathbf{w}^*|\boldsymbol{\theta}^*, \mathbf{y})$. This difficulty cannot be solved using the approach in [Chib, 1995], which assumes that the hidden variables can be decomposed into blocks with conditional densities available in closed form. In the present case, $\mathbf{w}^*$ can be decomposed into blocks consisting of single elementary displacements $\mathbf{w}_{ib}$, but the conditional density of each block has no analytical expression, and must be sampled using a Metropolis-Hastings step, as described in Appendix G. A generalization of the method in [Chib, 1995] to this setting is developed in [Chib and Jeliazkov, 2001]. It consists in factorizing the posterior ordinate across blocks, according to:

$$\pi(\mathbf{w}^*|\boldsymbol{\theta}^*, \mathbf{y}) \quad = \quad \prod_{i=1}^{n} \prod_{b=1}^{B} \pi(\mathbf{w}^*_{ib}|\mathbf{w}^*_{-ib}, \boldsymbol{\theta}^*, \mathbf{y}),$$

where we write $\mathbf{w}_{-ib}$ to denote the blocks preceding ib in the lexical order, that is, the collection of blocks $\mathbf{w}_{i'b'}$ where $i' \leq i$, $b' \leq b$, and $(i', b') \neq (i, b)$.

Each reduced posterior ordinate $\pi(\mathbf{w}^*_{ib}|\mathbf{w}^*_{-ib}, \boldsymbol{\theta}^*, \mathbf{y})$ can then be evaluated from the output of a reduced run of the multiple block MH algorithm, conditional on $\mathbf{w}_{-ib}$, and an additional run where $\mathbf{w}_{ib}$ is added to the conditioning set. Justification and further details on the calculation of the reduced ordinates and the likelihood function are given in Appendix H.

5.6 Comparing different parcellations

The choice of a particular parcellation is an important issue, as our whole decision framework rests upon it. Yet, the definition of functionally homogeneous regions in the human brain remains an open issue. In practice, mis-specified parcellations may cause activated areas to cross several parcels, resulting in reduced sensitivity, and difficulty in interpreting the detected pattern.

Though resolving this issue is beyond the scope of this work, we remark that the same Bayesian model selection formalism used to select the functional network $\boldsymbol{\gamma}$ can be used to compare different candidate parcellations. This is done by considering the parcellation as a random variable to be estimated, rather than a fixed quantity. Thus, two given parcellations $\mathscr{P} = \{\mathcal{V}_1, \ldots, \mathcal{V}_N\}$ and $\mathscr{P}' = \{\mathcal{V}'_1, \ldots, \mathcal{V}'_N\}$ may be compared through their

posterior odds:

$$\frac{\pi(\mathscr{P}|\mathbf{y})}{\pi(\mathscr{P}'|\mathbf{y})} = \frac{m(\mathbf{y}|\mathscr{P})}{m(\mathbf{y}|\mathscr{P}')} \times \frac{\pi(\mathscr{P})}{\pi(\mathscr{P}')}, \tag{5.22}$$

where $\pi(\mathscr{P})$ is the prior probability assigned to parcellation $\mathscr{P}$, and $m(\mathbf{y}|\mathscr{P})$ the evidence for $\mathscr{P}$, given by:

$$m(\mathbf{y}|\mathscr{P}) = \sum_{\boldsymbol{\gamma}\in\{0,1\}^N} m(\mathbf{y}|\boldsymbol{\gamma},\mathscr{P})\pi(\boldsymbol{\gamma}|\mathscr{P}). \tag{5.23}$$

Here $\boldsymbol{\gamma}$ is the vector indicating possible functional networks, based on parcellation $\mathscr{P}$, $m(\mathbf{y}|\boldsymbol{\gamma},\mathscr{P})$ is the marginal likelihood defined by (5.8) and $\pi(\boldsymbol{\gamma}|\mathscr{P})$ the prior probability of network $\boldsymbol{\gamma}$, defined by (5.13).

Thus, selecting the most adequate parcellation among a set of candidates, to infer the functional pattern associated with a certain functional task, is a straightforward application of the variable selection framework introduced in the previous section. Furthermore, applied to anatomically defined regions, it provides a tool to investigate the link between brain anatomy and function.

5.7 2D toy example

We now illustrate our model selection approach on a simulated dataset. In Section 4.5.2, our goal was to evidence a stretching effect of the activations due to spatial variability, and its compensation through appropriate modeling. Going one step further in our analysis, we now show that this stretching effect may result in a bias toward false positives when testing the presence of activations within pre-defined regions. Furthermore, we show that this bias can be corrected when spatial uncertainty is accounted for.

We defined a synthetic activation pattern, within a 2D search volume of 24×24 voxels, consisting of a central activated disc, with uniform intensity value 5 (the background was set to 0) and a diameter of 7 voxels.

To simulate the data, this activation was deformed according to a displacement field $\mathbf{u}$, simulated under the model described in Section 4.3, with one control point in each voxel. The standard displacement was taken equal to $\sigma_S = 1.0$ voxels and the field smoothness

parameter was set to $\omega = 4.0$ voxels. Independent heteroscedastic Gaussian noise was then added to each voxel $\mathbf{v}$, with variance equal to $1 + \mathbf{s}^2(\mathbf{v})$, where $\mathbf{s}^2(\mathbf{v})/\varepsilon \sim \chi^2(1)$, ε being the noise level, set to 1.0 in this example. A total of $n = 30$ pairs $(\mathbf{y}_i, \mathbf{s}_i^2)$ of effect and variance maps were sampled in this fashion, as illustrated in Figure 5.2, and constitute a sample from the hierarchical model in Section 4.2 .

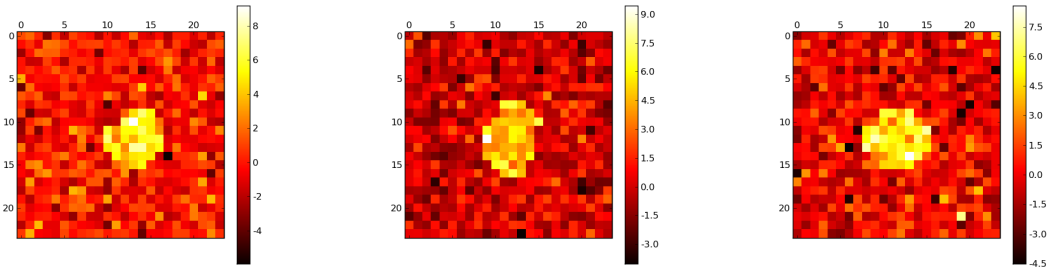


FIGURE 5.2: Three different simulated effect effect maps $\mathbf{y}_i$

Data analysis

The search volume was divided in two regions, corresponding to the the background ($\mathcal{V}_1$) and the active disc ($\mathcal{V}_2$). In this elementary situation, only 4 activation models are in competition, each of them corresponding to a value of the indicator variable $\boldsymbol{\gamma} \in \{0, 1\}^2$. We applied the Bayesian model selection approach described in 5.3 to recover the parcels with a nonzero mean population effect, both with and without modeling spatial uncertainty. In the model with spatial uncertainty, deformation fields were specified using a single control point in the center of the search volume, and the same regularity parameter $\omega = 4$ used to simulate the data.

More precisely, the marginal likelihood in each model was estimated as explained in Section 5.5. Thus, the posterior density of the model parameters was maximized over 500 iterations of the MCMC-SAEM algorithm (see Appendix E), following 500 'burn-in' iterations. Next, the posterior density of the resulting MAP estimate was computed using 1000 Gibbs iterations, following a burn-in period of 100 iterations. This was sufficient to obtain the marginal likelihood in the model without spatial uncertainty, applying (5.17).

In the model with spatial uncertainty, the density of the deformation fields, conditional on the MAP estimate of model parameters, was maximized using 500 iterations of the SA

algorithm (see Appendix G). Finally, this density was computed as in section 5.5.2. As explained in Appendix H, each reduced run was sampled using a number of iterations inversely proportional to the number of sampled blocks, and used 3000 iterations for the single-block run. Computing the marginal likelihood of each model under spatial uncertainty took approximately $9mn$ on a PC with a clock rate of $1.33GHz$, against $20s$ seconds without spatial uncertainty. To quantify the variance of the Monte-Carlo estimates of the marginal likelihood values, we repeated all calculations 10 times on the same dataset.

Results

Presence of the stretching effect can be checked in Figure 5.3, where posterior estimates of the signal from one trial are shown. These were computed by averaging over the four models the posterior mean conditional on the MAP parameter estimates $\hat{\boldsymbol{\theta}}_{\boldsymbol{\gamma}}$, according to: $\hat{\boldsymbol{\mu}} = \sum_{\boldsymbol{\gamma}} \mathbb{E}[\boldsymbol{\mu}|\mathbf{y}, \hat{\boldsymbol{\theta}}_{\boldsymbol{\gamma}}]\pi(\boldsymbol{\gamma}|\mathbf{y})$ (using the posterior mean $\mathbb{E}[\boldsymbol{\mu}|\mathbf{y}, \boldsymbol{\gamma}]$ would have made more sense, but was not directly available from the output of our algorithm).

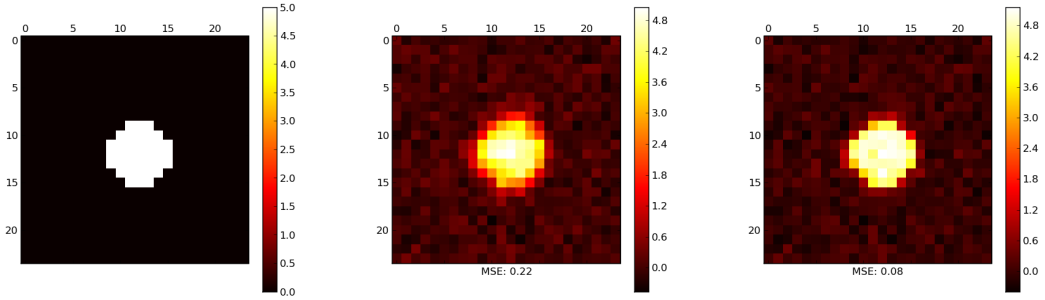


FIGURE 5.3: Posterior estimates of $\boldsymbol{\mu}$ (left), with (right) and without (center) modeling spatial uncertainty.

The mean and standard deviate over 10 trials of the marginal likelihood values of each model, are given in Table 5.1. As expected, when spatial uncertainty is unaccounted for, the maximum marginal likelihood is attained for $\boldsymbol{\gamma} = (1, 1)$, that is, activations are detected in the background, due to the stretching of the estimated activate disc. These values are numerically stable, as can be checked from their standard deviates.

The stretching effect is compensated in the model without spatial uncertainty, where the maximum marginal likelihood is achieved on the average for the correct model $\boldsymbol{\gamma} = (0, 1)$. Moreover, the marginal likelihood values are much higher, indicating the better fit

$\boldsymbol{\gamma}$	$m(\mathbf{y}\|\boldsymbol{\gamma}, \sigma_S = 0)$	$m(\mathbf{y}\|\boldsymbol{\gamma})$
$(0,0)$	-87882.68 ± 0.12	-33655.6 ± 21.6
$(1,0)$	-87868.39 ± 0.17	-33663.9 ± 12.7
$(0,1)$	-87748.58 ± 0.10	$\mathbf{-33644.0} \pm 13.2$
$(1,1)$	$\mathbf{-87733.29} \pm 0.18$	-33651.3 ± 16.1

TABLE 5.1: Log marginal likelihood values computed on 2D simulated data, over 10 trials. Results are given in the form: mean ± std. deviate.

obtained by modeling spatial uncertainty. However, the differences between the marginal likelihoods of the different models are now swamped in the variability of the Monte-Carlo estimates, so that the procedure does not systematically select the correct model on each trial.

This variability comes from the difficulties encountered in sampling the conditional density of the elementary displacements, as described in D. As a result, the Markov Chain is very sluggish in its exploration of the space of all possible deformation fields, and tends to get trapped in local maxima, an issue already noted in Sections 4.5 and 4.6. Because on each trial, the Markov chain gets trapped around a different mode, the resulting estimates of the marginal likelihood are highly variable. Increasing the number of iterations did not solve this issue, though in theory it would allow after sufficient time the chain to escape from each local mode and explore exhaustively the sampling space.

In conclusion, this numerical experiment demonstrates the potential benefits of modeling spatial uncertainty when testing regional hypotheses on fMRI data, using the model selection approach developed in Section 5.3. However, this approach turns out to be numerically unstable when modeling spatial uncertainty, even on the overly simplified toy dataset used here, with 2D data simulated under a high signal to noise ratio, and using smooth deformation fields with known smoothness.

5.8 Approximate inference using posterior modes.

To address the above mentioned numerical instability issue, we consider approximating the marginal likelihood $m(\mathbf{y}|\boldsymbol{\gamma})$ by the following likelihood, conditional on a particular

value $\hat{\mathbf{w}}$, which is the same for all networks $\boldsymbol{\gamma}$:

$$m(\mathbf{y}|\hat{\mathbf{w}}, \boldsymbol{\gamma}) = \int f(\mathbf{y}|\hat{\mathbf{w}}, \boldsymbol{\theta})\pi(\boldsymbol{\theta}|\boldsymbol{\gamma}). \tag{5.24}$$

By doing this, we avoid the difficult integration with respect to the displacements. This may be interpreted in terms of the Laplace approximation (see [Robert, 2007] for instance), which consists in approximating the unnormalized posterior density $m(\mathbf{y}|\mathbf{w}, \boldsymbol{\gamma})\pi(\mathbf{w}|\boldsymbol{\gamma})$ of $\mathbf{w}$ by an unnormalized Gaussian, obtained from a second order Taylor expansion of $\log m(\mathbf{y}|\mathbf{w}, \boldsymbol{\gamma})\pi(\mathbf{w}|\boldsymbol{\gamma})$ around its mode $\hat{\mathbf{w}}$.

Using this approximation would be overly expensive, because the posterior mode $\hat{\mathbf{w}}$ would need to be computed for each 2^N possible values of $\boldsymbol{\gamma}$, an unfeasible task for real datasets, since each model would need at least several minutes to be processed. Moreover, the validity of a Taylor expansion of $\log m(\mathbf{y}|\mathbf{w}, \boldsymbol{\gamma})\pi(\mathbf{w}|\boldsymbol{\gamma})$ is questionable in our case, since it is only piecewise continuous as a function of $\mathbf{w}$.

Hence, we simplify the Laplace approximation, replacing $m(\mathbf{y}|\mathbf{w}, \boldsymbol{\gamma})\pi(\mathbf{w}|\boldsymbol{\gamma})$ by a Dirac mass in the posterior mode $\hat{\mathbf{w}}$ instead of a Gaussian, and further assuming $\hat{\mathbf{w}}$ to be the same for all networks $\boldsymbol{\gamma}$.

We define $\hat{\mathbf{w}}$ as the following conditional posterior mode in the model without parcellation, defined in Chapter 4:

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \pi(\boldsymbol{\theta}|\mathbf{y}); \tag{5.25}$$

$$\hat{\mathbf{w}} = \arg\max_{\mathbf{w}} \pi(\mathbf{w}|\mathbf{y}, \hat{\boldsymbol{\theta}}). \tag{5.26}$$

We use the mode conditional on the MAP parameters $\hat{\boldsymbol{\theta}}$, rather than the unconditional mode, for simplicity, since it can be estimated using the SAEM and SA algorithm previously introduced (see Appendices E and D).

The conditional likelihood (5.24) can be evaluated exactly as the marginal likelihood under no spatial uncertainty. Indeed, because voxels are independent conditional on $\hat{\mathbf{w}}$, this expression can be factorized across regions:

$$m(\mathbf{y}|\hat{\mathbf{w}}, \boldsymbol{\gamma}) = \prod_{j=1}^{N} m(\mathbf{y}^j|\hat{\mathbf{w}}, \gamma_j).$$

where $\mathbf{y}^j = \{y_{i,k}; \mathbf{v}_k + \mathbf{u}_{i,k} \in \mathcal{V}_j\}$ is the subset of observations corresponding to voxels displaced into region j. Note that this set is random, since it is a function of $\mathbf{u}_{i,k}$, itself a function of the elementary displacements $\mathbf{w}$, as defined by (4.3). The conditional likelihood of region j is equal to:

$$m(\mathbf{y}^j|\hat{\mathbf{w}}, \gamma_j) = \int f(\mathbf{y}^j|\hat{\mathbf{w}}, \boldsymbol{\theta}_j)\pi(\boldsymbol{\theta}_j|\gamma_j)d\boldsymbol{\theta}_j,$$

where $\boldsymbol{\theta}_j = (\eta_j, \nu_j^2, \sigma_j^2)$ is the parameter vector for region j.

These conditional likelihoods may be evaluated separately, or equivalently, (5.24) can be computed for $\boldsymbol{\gamma} = \mathbf{0}_N$ and $\boldsymbol{\gamma} = \mathbf{1}_N$. This can be done by Chib's method (see Section 5.5.1), writing:

$$m(\mathbf{y}|\hat{\mathbf{w}}, \boldsymbol{\gamma}) = \frac{f(\mathbf{y}|\hat{\mathbf{w}}, \hat{\boldsymbol{\theta}}_{\boldsymbol{\gamma}})\pi(\hat{\boldsymbol{\theta}}_{\boldsymbol{\gamma}}|\boldsymbol{\gamma})}{\pi(\hat{\boldsymbol{\theta}}_{\boldsymbol{\gamma}}|\mathbf{y}, \hat{\mathbf{w}}, \boldsymbol{\gamma})}, \tag{5.27}$$

where we choose $\hat{\boldsymbol{\theta}}_{\boldsymbol{\gamma}}$ as the conditional MAP estimate: $\hat{\boldsymbol{\theta}}_{\boldsymbol{\gamma}} = \arg\max_{\boldsymbol{\theta}} \pi(\boldsymbol{\theta}|\mathbf{y}, \hat{\mathbf{w}}, \boldsymbol{\gamma})$. The posterior density $\pi(\hat{\boldsymbol{\theta}}_{\boldsymbol{\gamma}}|\mathbf{y}, \hat{\mathbf{w}}, \boldsymbol{\gamma})$ is obtained by the Rao-Blackwell method, as explained in Section 5.5.1, while the exact expression of the density $f(\mathbf{y}^j|\hat{\mathbf{w}}, \boldsymbol{\theta})$ is given in Appendix F.

Finally, the posterior density $\pi(\boldsymbol{\gamma}|\mathbf{y})$ of the functional network is approximated by $\pi(\boldsymbol{\gamma}|\mathbf{y}, \hat{\mathbf{w}})$, which can be factorized across regions into:

$$\pi(\boldsymbol{\gamma}|\hat{\mathbf{w}}, \mathbf{y}) = \prod_{j=1}^{N} \pi(\gamma_j|\mathbf{y}^j, \hat{\mathbf{w}}). \tag{5.28}$$

This posterior density is entirely determined by the posterior probabilities that each region is involved,

$$P_j = \pi(\boldsymbol{\gamma}_j = 1|\mathbf{y}^j, \hat{\mathbf{w}}) = 1 - \pi(\boldsymbol{\gamma}_j = 0|\mathbf{y}^j, \hat{\mathbf{w}}),$$

obtained from the conditional likelihoods as

$$P_j = \left(1 + \frac{m(\mathbf{y}^j|\hat{\mathbf{w}}, \gamma_j = 0)}{m(\mathbf{y}^j|\hat{\mathbf{w}}, \gamma_j = 1)} \times \frac{\pi(\gamma_j = 0)}{\pi(\gamma_j = 1)}\right)^{-1}. \tag{5.29}$$

Intuitively, conditioning on $\hat{\mathbf{w}}$ may be seen as a pre-processing step wherein the functional images are registered to a template $\boldsymbol{\mu}$, which is estimated at the same time. Thus, it provides a way to compensate for spatial normalization errors based on the functional

data. However, because the posterior variability of the elementary displacements around their mode is neglected, we anticipate good results only when this variability is reduced, that is when the data provides enough information for the elementary displacements to be well estimated.

5.8.1 2D toy example

We applied the above posterior mode approximation to the same dataset used in Section 5.7 . Our goal was to validate the approximation, and evaluate its numerical stability.

Data analysis

We used our posterior mode approximation to compute the posterior probabilities P_j of positive mean activations within each region $j = 1, 2$. More precisely, the MAP estimate $\hat{\boldsymbol{\theta}}$ of model parameters was first obtained without parcelling the search volume, using 500 iterations of the SAEM algorithm, following 500 burn-in iterations. Then, the conditional MAP estimate $\hat{\mathbf{w}}$ of the displacement fields was computed using 500 iterations of the SA algorithm. The conditional likelihoods $m(\mathbf{y}|\hat{\mathbf{w}}, \boldsymbol{\gamma})$ were then computed both under the null model, defined by $\gamma_j \equiv 0$, and under the full model, defined by $\gamma_j \equiv 1$. In each case, the conditional MAP estimate $\hat{\boldsymbol{\theta}}_{\boldsymbol{\gamma}}$ of model parameters was obtained from 1 000 iterations of the SAEM algorithm (including 500 burn-in iterations), and the conditional posterior $\pi(\hat{\boldsymbol{\theta}}_{\boldsymbol{\gamma}}|\mathbf{y}, \hat{\mathbf{w}}, \boldsymbol{\gamma})$ was computed from the output of 1 000 iterations of a MH-Gibbs algorithm, following a burn-in of 100. We then computed the log Bayes factors:

$$B_j = \log \frac{m(\mathbf{y}^j|\hat{\mathbf{w}}, \gamma_j = 1)}{m(\mathbf{y}^j|\hat{\mathbf{w}}, \gamma_j = 0)},$$

from which the P_j were directly deduced as

$$P_j = \left(1 + e^{-B_j}\right)^{-1},$$

following (5.29), adopting a uniform prior $\pi(\gamma_j = 0) = \pi(\gamma_j = 1)$ for the state of each region. We compared these results with those obtained under no spatial uncertainty,

setting $\mathbf{w} = \mathbf{0}$ and $\sigma_S^2 = 0$, noting that, in this case, the procedure is exact and equivalent to the one tested in Section 5.7. All computations were re-run 10 times to assess numerical stability.

Results

Presence of the stretching effect can be checked in Figure 5.4, where posterior estimates of the signal from one trial are shown. As previously, these were computed by averaging over the four models the posterior mean conditional on the MAP parameter estimates and the MAP elementary displacements, according to: $\hat{\boldsymbol{\mu}} = \sum_{\boldsymbol{\gamma}} \mathbb{E}[\boldsymbol{\mu}|\mathbf{y}, \hat{\boldsymbol{\theta}}_{\boldsymbol{\gamma}}, \hat{\mathbf{w}}]\pi(\boldsymbol{\gamma}|\mathbf{y}, \hat{\mathbf{w}})$. Also, it can be seen that the posterior mode $\hat{\mathbf{w}}$ provides in the present case a good estimate of the unknown displacements. Indeed, the posterior estimate of $\boldsymbol{\mu}$ conditional on $\hat{\mathbf{w}}$ has a mean square error (MSE) of 0.11, almost as low as that of the unconditional estimate tested in Section 5.7, which was equal to 0.08.

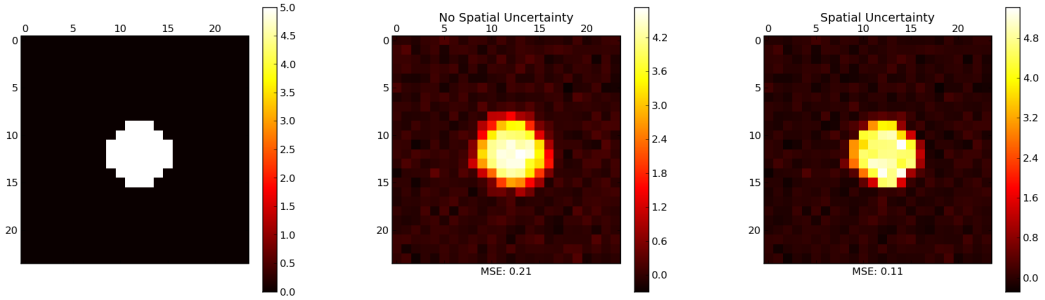


FIGURE 5.4: Posterior estimates of $\boldsymbol{\mu}$ (left), with (right) and without (center) modeling spatial uncertainty, using the posterior mode approximation

	Spatial uncertainty	No spatial uncertainty
Region j	B_j	B_j
1 (background)	-7.16 ± 0.27	29.23 ± 0.48
2 (disc)	3.21 ± 0.75	51.89 ± 0.01

TABLE 5.2: Results of the approximate model selection procedure on a single 2D simulated dataset, over 10 trials.
Results are given in the form: mean $\pm$ std. deviate.

The mean and standard deviates of the Bayes factor for each region is given in Table 5.2. Results for the model without spatial uncertainty are identical to those found in the previous experiment, since the procedures are in this case equivalent. They are also quantitatively similar in the model with spatial uncertainty: the background is correctly classified as inactive, with a negative log Bayes factor B_1, and the disc as active, with $B_2 > 0$. Furthermore, the variance of the log Bayes factors is much reduced with respect to that of the log marginal likelihoods (see Table 5.1), showing that our approximation provides better numerical stability.

In conclusion, this experiment shows that our posterior mode approximation works well in a favorable setting, were it leads to the same conclusions as the exact procedure, with a much reduced numerical variability. These good results can be attributed to the fact that the data is simulated with a high signal to noise ratio and warped using smooth deformation fields with known regularity, so that the posterior density of $\mathbf{w}$ is likely to be sharply peaked around its mode.

5.8.2 3D toy example

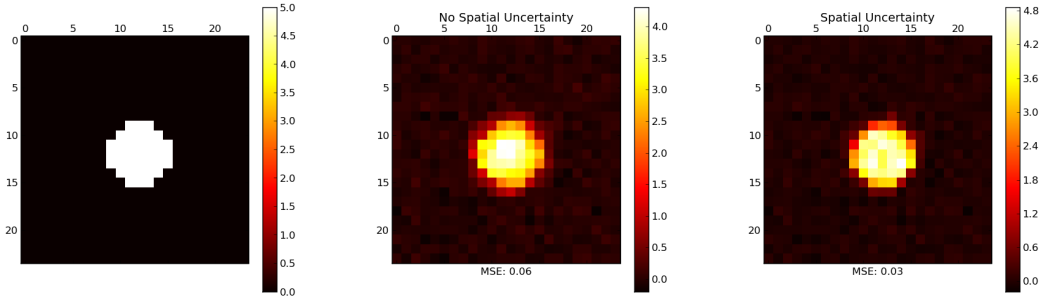


FIGURE 5.5: Posterior estimates of $\boldsymbol{\mu}$ (left), with (right) and without (center) modeling spatial uncertainty, using the posterior mode approximation

The previous results are encouraging, but does our posterior mode approximation work as well on 3D data? To answer this, we used a dataset simulated exactly as in Section 5.7, except that the 2D 24×24 search volume was replaced by a 3D $24 \times 24 \times 24$ search volume, and the central disc was replaced by a sphere with same diameter (7 voxels).

We used our posterior mode approximation to compute the Bayes factors for both regions (background and sphere), using the algorithm detailed in Section 5.8.1. We compared these results with those obtained under no spatial uncertainty, setting $\mathbf{w} = \mathbf{0}$ and $\sigma_S^2 = 0$,

	Spatial uncertainty	No spatial uncertainty
Region j	B_j	B_j
1 (background)	39.55 ± 1.2	80.47 ± 0.05
2 (sphere)	264.55 ± 6.82	237.06 ± 0.01

TABLE 5.3: Results of the approximate model selection procedure on a 3D simulated dataset, over 10 trials.

and repeated all calculations over 10 trials to assess numerical stability. Each trial took approximately one hour on a PC laptop with a clock rate of $1.33GHz$. Results for the 3D dataset are given in Table 5.3. As in Section 5.8.1, we report the log Bayes factor values B_j for each region j.

In terms of estimation, the posterior mode approximation again gives satisfying results, as seen in Figure 5.5, with a mean-square error (0.03) lower than that of the estimate in the model with no spatial uncertainty (0.06). In terms of model selection, results are less satisfying, since the background is selected as active, with a positive Bayes factor, both with and without spatial uncertainty.

One possible explanation is that estimating 3D displacement fields requires more information than is provided by the data (which seemed sufficient in the 2D case), so that our posterior mode approximation, though it reduces the spreading effect due to the mis-localization of individual activations, does not reduce it enough in order for the background to be found inactive. Consequently, the Bayes factor conditional on the most probable displacements is lower than when displacements are fixed to zero, but still positive.

5.8.3 Additional penalty on model fit

The above illustrations suggest that approximating the marginal likelihood by fixing the displacements to their most probable value works well on 2D data, but on 3D data fails to entirely compensate the warping of individual images and the ensuing swelling of the estimated activations. Consequently, the systematic bias toward false positives, found

when registration errors are not modeled, is reduced but still present when using the posterior mode approximation.

This bias can be seen as a form of data overfit, the number of parameters needed to model the data being overestimated, due to likelihood values inflated by displaced activations. Note that this bias would be automatically compensated when using the true log marginal likelihood $\log m(\mathbf{y}|\boldsymbol{\gamma})$ rather than the conditional approximation $\log m(\mathbf{y}|\hat{\mathbf{w}}, \boldsymbol{\gamma})$, because it would contain an additional term penalizing the overfit due to modeling spatial displacements:

$$\log m(\mathbf{y}|\boldsymbol{\gamma}) = \log m(\mathbf{y}|\hat{\mathbf{w}}, \boldsymbol{\gamma}) + \log \frac{\pi(\hat{\mathbf{w}})}{\pi(\hat{\mathbf{w}}|\mathbf{y}, \boldsymbol{\gamma})}.$$

However, the previous example has illustrated the fact that computing this penalizing term requires heavy MCMC calculations (specifically, the posterior ordinate $\pi(\hat{\mathbf{w}}|\mathbf{y}, \boldsymbol{\gamma})$), and in our case resulted in a far too important numerical variability to be of ay use. As a surrogate to this currently unatainable exact quantity, we consider compensating the bias by modifying B_j, adding a penalty to model fit, measured by the log likelihood ratio

$$LR_j \quad = \quad \log \frac{f(\mathbf{y}^j|\hat{\mathbf{w}}, \hat{\boldsymbol{\theta}}_{1j})}{f(\mathbf{y}^j|\hat{\mathbf{w}}, \hat{\boldsymbol{\theta}}_{0j})}, \tag{5.30}$$

where $\hat{\boldsymbol{\theta}}_{kj} = \arg\max_{\boldsymbol{\theta}_j} \pi(\boldsymbol{\theta}_j|\gamma_j = k) f(\mathbf{y}^j|\hat{\mathbf{w}}, \boldsymbol{\theta}_j)$ for $k = 0, 1$. This quantity is available as an output of the algorithm which computes the log Bayes factor B_j, since, following (5.27) it can be written as:

$$\begin{aligned} B_j &= \log \frac{m(\mathbf{y}^j|\hat{\mathbf{w}}, \gamma_j = 1)}{m(\mathbf{y}^j|\hat{\mathbf{w}}, \gamma_j = 0)} \\ &= \log \frac{f(\mathbf{y}^j|\hat{\mathbf{w}}, \hat{\boldsymbol{\theta}}_{1j})}{f(\mathbf{y}^j|\hat{\mathbf{w}}, \hat{\boldsymbol{\theta}}_{0j})} + \log \frac{\pi(\hat{\boldsymbol{\theta}}_{1j}|\gamma_j = 1)}{\pi(\hat{\boldsymbol{\theta}}_{0j}|\gamma_j = 0)} + \log \frac{\pi(\hat{\boldsymbol{\theta}}_{0j}|\mathbf{y}^j, \hat{\mathbf{w}}, \gamma_j = 0)}{\pi(\hat{\boldsymbol{\theta}}_{1j}|\mathbf{y}^j, \hat{\mathbf{w}}, \gamma_j = 1)} \\ &= LR_j + D_j. \end{aligned} \tag{5.31}$$

We expect LR_j to be always positive. Indeed, the MAP estimates $\hat{\boldsymbol{\theta}}_{kj}$, used to define it in (5.30) are likely to be very close to the maximum likelihood (ML) estimates, defined by $\hat{\boldsymbol{\theta}}_{kj}^{ML} = \arg\max_{\boldsymbol{\theta}_j} f(\mathbf{y}^j|\mathbf{w}, \boldsymbol{\theta}_j)$, given that we have chosen a prior $\pi(\boldsymbol{\theta}|\gamma_j)$ that is as

minimally informative as possible (see Section 5.4). Hence, it is reasonable to expect LR_j to be close to the log maximum likelihood ratio, a quantity that is always positive for nested models. Corroborating this, we have always observed in practice positive values for LR_j. Hence we have not felt the need to use the ML estimates instead of the MAP estimates of the parameter values, though this would be a possible alternative.

Using the positivity of LR_j, we define a lower bound on the Bayes factor B_j by:

$$\tilde{B}_j = c \times LR_j + D_j, \tag{5.32}$$

for a certain scale factor $c \in (0, 1)$, which determines the added penalty. Based on this quantity, we define the following lower bound on the posterior probability P_j of a nonzero mean activation within region j :

$$\tilde{P}_j \quad = \quad \left(1 - \tilde{B}_j \times \frac{\pi(\gamma_j = 0)}{\pi(\gamma_j = 1)}\right)^{-1}. \tag{5.33}$$

$\tilde{P}_j$ underestimates the probability of each region being involved in the task at hand. Thus, the bias it introduces is always conservative.

Calibration of the additional penalty on simulated data

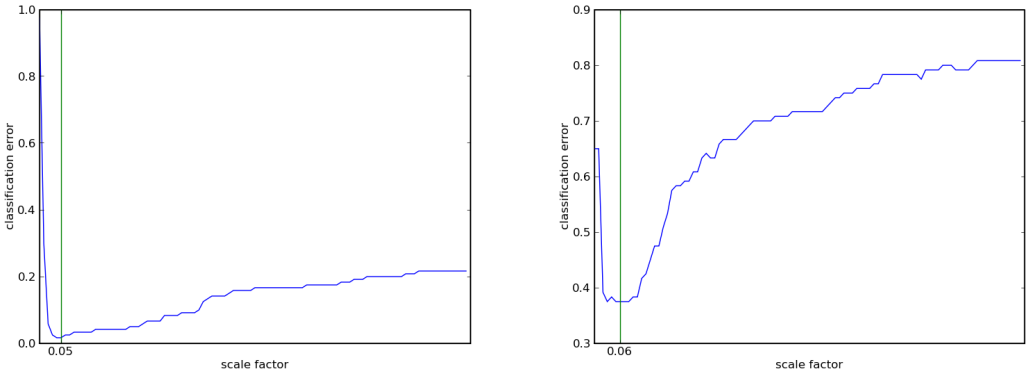


FIGURE 5.6: Optimization of the additional penalty on simulated datasets, for the posterior mode approximation using the model with spatial uncertainty (left), and using the model with no spatial uncertainty (right).

Ideally, we would like to choose the factor c small enough so that it will compensate for data overfit, but not too small, for active regions not to be missed. We perform this calibration using intensive calculations, by estimating the optimal value from a collection of datasets simulated under different parameter values.

As in Section 5.8.2, we defined a synthetic activation pattern, within a search volume of $24 \times 24 \times 24$ voxels, consisting of a central activated sphere, with uniform intensity value 5 (the background was set to 0) and a diameter of 7 voxels.

The data was then simulated using a field smoothness parameter $\omega = 3, 4$ or 5 and a noise level $\epsilon = 1, 2, 3$ of 4. For each of the 3×4 possible combinations of these parameters, we simulated 10 different datasets comprising $n = 30$ images, to which we applied the procedure in 5.8 to compute Bayes factor values $B_j = LR_j + D_j$ for regions $j = 1, 2$, using the same algorithm as in Sections 5.8.1 and 5.8.2.

Next, for all $c = 0, 0.01, 0.02, \ldots, 1$, we computed the modified criterion $\tilde{B}_j(g, c) = c \times LR_j(g) + D_j(g)$ for all 120 datasets, indexed by $g = 1, \ldots, 120$ and measured the proportion of corresponding mis-classified regions, given by:

$$R(c) = \sum_{g=1}^{120} \left\{ \mathbf{1}_{\tilde{B}_1(c,g)>0} + \mathbf{1}_{\tilde{B}_2(c,g)<0} \right\}.$$

Finally, we chose the value of c which minimized $R(c)$.

As illustrated in Figure 5.6, left, a minimum of 2 misclassified regions (out of 420) was obtained for $c = 0.05$. The surprisingly low number of classification errors obtained after optimizing c may suggest that it is the introduction of this factor, rather than the posterior mode approximation, that is effective in correcting data overfit.

To verify this assertion, we also computed for each dataset the Bayes factors B_j using the model without spatial uncertainty, computed the modified criterions $\tilde{B}_j$ for different values of c and chose the value optimizing the classification rate. As illustrated in Figure 5.6, the minimum number of mis-classified regions, obtained in this case for $c = 0.06$, was much higher, and equal to 45. This indicates that the good classification score achieved using the model with spatial uncertainty is due to the combination of the posterior mode approximation, as well as the additional penalty.

5.8.4 Phantom activations

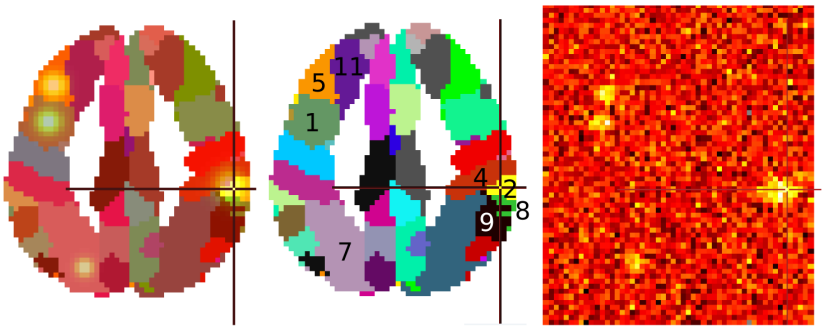


FIGURE 5.7: Data simulated using phantom activations. Slice $z = 37$mm in Talairach space. From left to right: Synthetic activation pattern (with CSA atlas in the background); CSA atlas (numbers correspond to region index in Table 5.4); simulated data example.

We conclude this chapter by applying our final procedure to a dataset which presents some similarities to real-life situations, and in particular, which was not simulated under the full hierarchical model (illustrated in Figure 5.1) used to analyze the data.

Data simulation and analysis

To start with, we defined a 'life-size' search volume of $45 \times 62 \times 52$ voxels, corresponding to the actual dimensions of true fMRI images. Then, we designed an artificial activation pattern, based on the cortical sulci atlas (CSA), developed in [Perrot et al., 2008], and which we used to analyze real fMRI data (see Chapter 6). Each activation was defined in terms of a peak location, from which the signal decreased radially according to a Gaussian kernel. Each signal peak was taken equal to 5. We placed two activations in neighboring regions, one at the intersection of several regions, and a smaller one inside the largest atlas region (see Figure 5.7, region 7).

$n = 40$ images were generated by warping this map according to the deformation model defined by (4.3), with one control point in each voxel, choosing $\omega = 4$ voxels and $\sigma_S = 2.0$ voxels. Homoscedastic noise was then added to each image according to (5.5), with $\sigma_j^2 \equiv 1$, and the $s_{i,k}^2$'s generated as independent chisquare variables.

We then applied our Bayesian model selection algorithm based on the posterior mode approximation described in Section 5.8 to this dataset, to compute for all regions $j = 1, \ldots, N$, the Bayes factor testing the presence of a nonzero regional mean η_j, still using

the deformation model defined in (4.3), except this time control points were restricted to a grid with regular spacing equal to γ along each axis. We used the algorithm detailed in Section 5.8.1 (except that $j = 1, \ldots, N$ in our case, with $N = 124$).

Finally, lower bounds $\tilde{P}_j$ on the posterior probabilities of nonzero regional means where computed, as explained in Section 5.8.3, using an additional penalty controlled by a factor c, which we tuned to the optimal value derived in 5.8.3. This penalty was used for the Bayes factor computed both in the model with and without spatial uncertainty.

Results

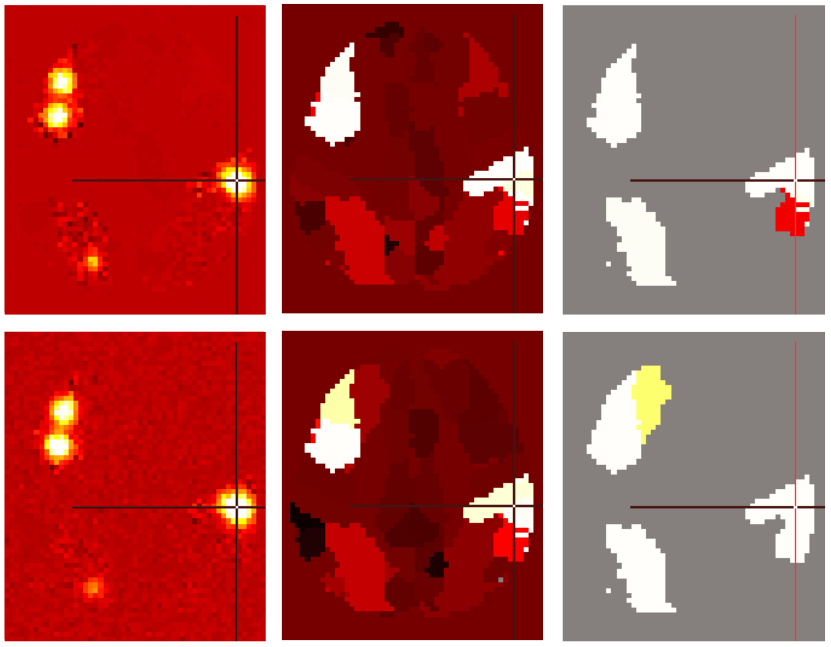


FIGURE 5.8: Statistical maps obtained by the model selection approach on 3D simulated data. Axial slice $z = 37$mm in Talairach space. First row contains results for the model with spatial uncertainty, bottom row for the model without. From left tor right: posterior estimate of the mean effect map $\boldsymbol{\mu}$, the regional mean effect $\boldsymbol{\eta}$, and lower bounds $\tilde{P}_j$ on the probabilities of a nonzero mean activation, restricted to detected regions ($\tilde{P}_j > 0.$).

Results from this simulation study are illustrated in Figure 5.7. We estimated the mean effect map $\boldsymbol{\mu}$ by its conditional posterior mean, averaged over models:

$$\hat{\boldsymbol{\mu}} = \sum_{\boldsymbol{\gamma}} \mathbb{E}(\boldsymbol{\mu}|\mathbf{y}, \hat{\mathbf{w}}, \hat{\boldsymbol{\theta}}_{\boldsymbol{\gamma}}) \tilde{\pi}(\boldsymbol{\gamma}|\mathbf{y}, \hat{\mathbf{w}}).$$

	$\tilde{P}_j$	$\hat{\eta}_j$	$\tilde{P}_j$	$\hat{\eta}_j$	$\bar{\boldsymbol{\mu}}_j$
Region	Spatial uncertainty		No spatial uncertainty		
1:	1.00	0.53	1.00	0.41	0.56
2:	1.00	0.26	1.00	0.19	0.27
3:	1.00	0.24	1.00	0.19	0.27
4:	1.00	0.25	1.00	0.20	0.26
5:	1.00	0.07	1.00	0.04	0.06
6:	1.00	0.26	1.00	0.21	0.26
7:	1.00	0.04	1.00	0.03	0.05
8:	1.00	0.33	0.99	0.23	0.34
9:	0.69	0.07	1.00	0.04	0.06
10:	*0.46*	*0.04*	1.00	0.06	0.05
11:	*0.00*	*0.00*	0.93	0.02	0.01

TABLE 5.4: Results of the model selection approach on 3D simulated data. Reported regions have an approximate posterior probability of being involved in the task greater than 0.5. $\hat{\eta}_j$ is the posterior estimate of the regional mean effect. $\bar{\boldsymbol{\mu}}_j$ is the average over region j of the group mean effect map $\boldsymbol{\mu}$.

where $\tilde{\pi}(\boldsymbol{\gamma}|\mathbf{y}, \hat{\mathbf{w}})$ is noted using a tilde to remind that it is computed using the lower bound $\tilde{P}_j$ instead of the posterior probability. It can be seen from Figure 5.7, left, that the map obtained using the spatial uncertainty is less noisy, especially in the background, indicating the better fit obtained conditional on the most probable displacements *a posteriori*.

Figure 5.8 shows that both algorithms detected similar regions, but that the model with no spatial uncertainty was less conservative. In particular, the latter detected activations in region 11, simply because it was located next to region 5 (using the labels from Figure 5.7) which contained one of the simulated activation peak. Conversely, this region was considered inactive (with $\tilde{P}j = 0.00$) using the model with spatial uncertainty.

This tendency is confirmed in Table 5.4, where we can see that two more regions were

detected in the model ignoring spatial displacements. Both methods detected activations in region 9, which was contaminated by activations from neighboring regions 8 and 2, though the approximate probability is lower in the model with spatial uncertainty.

Finally, we computed the following estimates of the regional means:

$$\hat{\eta}_j = \sum_{\boldsymbol{\gamma}} \hat{\eta}_{\boldsymbol{\gamma}} \tilde{\pi}(\boldsymbol{\gamma}|\mathbf{y}, \hat{\mathbf{w}}),$$

where $\eta_{\boldsymbol{\gamma}}$ is the MAP estimate of η, conditional on $\mathbf{w}$ and $\boldsymbol{\gamma}$. Still from table 5.4, it can be seen that, as an estimate of the average mean effect within region j, $\bar{\boldsymbol{\mu}}_j = \sum_{k \in \mathcal{V}_j} \mu_k$, $\hat{\eta}_j$ is more accurate using spatial uncertainty than not. In fact, the average relative error $\epsilon = \frac{1}{N} \sum_j |\hat{\eta}_j - \bar{\boldsymbol{\mu}}_j| / \bar{\boldsymbol{\mu}}_j$ was equal to 1% for the spatial uncertainty model against 8% for the other. This confirms the fact that our posterior mode approximation did provide a better fit to the data than obtained without accounting for displacements.

5.9 Discussion

Throughout this chapter, we have investigated the possibility of testing the presence of a nonzero mean effect within each region of a pre-defined parcellation of the search volume, while accounting for individual images being spatially deformed, according to unknown displacement fields. We proposed a Bayesian model selection approach to this problem, entailing the computation of the marginal likelihood for each possible observation model, specified by a partition of all regions into 'involved' (containing a nonzero mean effect) and 'inactive'.

We have shown how Monte-Carlo estimates of these marginal likelihoods can be obtained using MCMC techniques. However, this theoretically sound solution fails in practice, because the Monte-Carlo estimates of the marginal likelihoods are unstable numerically. We then considered an approximate procedure, which consists in assimilating the unknown spatial displacements to their most probable values *a posteriori*. This approximation substantially reduces the numerical instability, however we found that it re-introduces a certain amount of bias toward false positives, which is systematically present when using the model which ignores spatial displacements. We compensated this residual bias by introducing an additional penalty, calibrated on a large number of

simulated datasets. This last approximation gave satisfying results when validated on a final synthetic 3D dataset.

There is clearly some space for improvement here, in particular concerning the Monte-Carlo estimation of the exact marginal likelihood. We have seen that its numerical instability stems from the random-walk Metropolis-Hastings step used to sample the elementary displacements, which has a high rejection rate, and results in the Markov Chain getting stuck in local modes of the posterior density. Thus, a promising direction for future work would be to test alternative proposal densities, in view of ameliorating our posterior sampling scheme.

Another alternative we have not yet explored would be to integrate $\mathbf{x}$ and $\boldsymbol{\mu}$ analytically and sample directly from the joint posterior distribution $\pi(\boldsymbol{\theta}, \mathbf{w}|\mathbf{y}, \boldsymbol{\gamma})$, using two alterning Metropolis-Hastings steps. This would potentially speed-up the convergence of the Markov chain, and moreover allow to compute the unconditional MAP estimate of $\mathbf{w}$, using simulated annealing. This is required in Section 5.8, and we have so far used a conditional MAP estimate as a surrogate.

Such ameliorations would allow to investigate the behavior of the exact approach in more complex situations than provided by the simplistic 2D datasets studied in this chapter. This would also provide an additional way of validating our approximate approach, by comparing the results obtained by both methods.

However, we point that the exact approach would nevertheless be associated with a high computational complexity, since the number of possible models increases exponentially with the number of regions, due to the fact that regions are dependent when modeling spatial uncertainty. In contrast, the number of distinct models in the model without spatial uncertainty, or under the posterior mode approximation, is linear in the number of regions. Thus, the use of approximate techniques, such as the one we developed here, seems unavoidable in view of practical applications.

This raises additional questions concerning the generability of the additional penalty c tuning, which we have done here using a colection of simulated datasets. These datasets where generated in order to reflect some features of the real datasets we intended to analyze (number of observations, inter-subject variance). The resulting penalty was found to work well when applied to real data. However, it may very well be that the

optimal value for c varies with respect to those parameters we chose to fix during our simulation study, and may need to be adapted to different situations, such as a much smaller or a much bigger sample size. These questions need to be investigated thoroughly in order for the method proposed here to be widely applicable.

Chapter 6

Application to real fMRI data

Abstract

In this chapter, we apply the Bayesian model selection approach for fMRI group data analysis developed in Chapter 5 to a real fMRI dataset.

To validate our approach, we chose a paradigm based on extensively studied cognitive tasks (number and language processing), involving known brain regions. We show that our procedure successfully recovers in each case the complete functional network.

We also compare two different versions of our approach, both with and without modeling spatial uncertainty, along with the SPM-like approach, described in Chapter 2. In this way, we illustrate the shortcomings of standard voxel-based approaches which rely on the thresholding of a statistical map, and how these are overcome by the procedure we propose.

6.1 Data analysis

The data used here is extracted from the Localizer database [Pinel et al., 2007]. We used the same cohort of 38 subjects as in Chapter 3 and refer to Section 3.4 for a detailed description.

6.1.1 Individual data processing

Individual data analyses were conducted following the standard pipeline described in Section 2.2, using SPM5 (http://www.fil.ion.ucl.ac.uk/spm/). Data were submitted successively to motion correction, slice timing and normalization to the MNI template, and spatial smoothing using a $5 \times 5 \times 5$ mm^3 FWHM Gaussian filter. For each subject, BOLD contrast images were obtained from a fixed-effect analysis on all sessions.

6.1.2 Methods compared

For each studied contrast, we used the method developed in Chapter 5 to select the functional network most probably involved in the cognitive task under investigation, based on a fixed brain parcellation. We used to this end the cortical sulci atlas (CSA) developed in [Perrot et al., 2008], derived from the anatomical images of 63 subjects and comprising 125 regions which correspond to subdivisions of cortical sulci (see Figure 6.1).

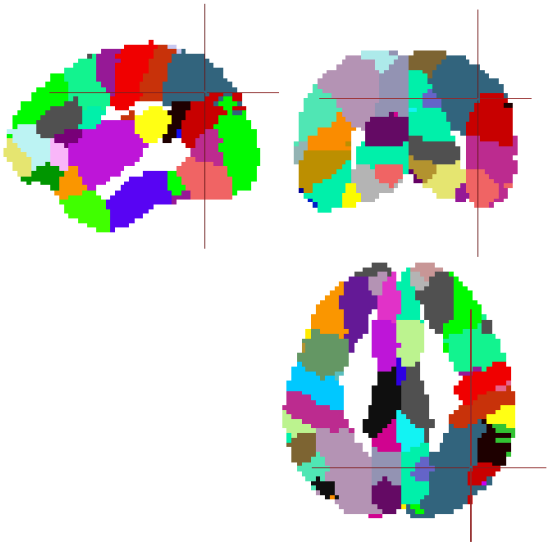


FIGURE 6.1: The cortical sulci atlas

More specifically, we used the posterior mode approximation described in Section 5.8 to compute the Bayes factor B_j to test the presence of a nonzero mean activation, for each region j. The algorithm was tuned as in the simulation study in Section 5.8.2. As explained in Section 5.8.3, these Bayes factors where modified through the use of an additional penalty, according to (5.32). The factor calibrating this penalty was taken equal to $c = 0.05$, which was found optimal from the simulation study in Section 5.8.3.

We compared the results obtained in the model with and without spatial uncertainty, the latter corresponding to the special case where $\sigma_S^2 = 0$, and the elementary displacements are frozen to zero.

We also included the results of the SPM-like approach, described in Section 2.3. In this case, the group analyses were restricted to the intersection of all subjects' whole-brain masks, comprising 43 367 voxels (no mask was used to restrict the analysis using our Bayesian model selection approach). First, a t-score map was computed from the 37 individual estimated effect maps. Then, a permutation test was used to compute three different cluster-forming thresholds, tuned to control the per-comparison error rate (PCER) respectively at 10^{-2}, 10^{-3} and 10^{-4} uncorrected. For each threshold, a second permutation test was used to determine the critical cluster size to guarantee a FWER control of 5% (see Section 2.4.2). Each detected cluster was labeled according to the region containing its maximum t-score; this is one of the procedures suggested in [Tzourio-Mazoyer et al., 2002].

6.1.3 Summarizing the inference

We chose to summarize the results of each procedure using the following statistics. The first one is the posterior estimate of the mean population effect, averaged across all possible networks, obtained as

$$\hat{\boldsymbol{\mu}} = \sum_{\boldsymbol{\gamma}} \mathbb{E}[\boldsymbol{\mu}|\mathbf{y}, \hat{\boldsymbol{\theta}}_{\boldsymbol{\gamma}}, \hat{\mathbf{w}}] \tilde{\pi}(\boldsymbol{\gamma}|\mathbf{y}, \hat{\mathbf{w}}),$$

using the notations introduced in Section 5.8. Here we note $\tilde{\pi}(\boldsymbol{\gamma}|\mathbf{y}, \hat{\mathbf{w}})$ using a tilde since we use the conservative approximation in Section 5.8.3.

Secondly, we computed the posterior estimate of regional means, also averaged across all possible networks:

$$\hat{\boldsymbol{\eta}} = \sum_{\boldsymbol{\gamma}} \hat{\boldsymbol{\eta}}_{\boldsymbol{\gamma}} \tilde{\pi}(\boldsymbol{\gamma}|\mathbf{y}, \hat{\mathbf{w}}),$$

where $\hat{\boldsymbol{\eta}}_{\boldsymbol{\gamma}} = \arg\max_{\boldsymbol{\eta}} \pi(\boldsymbol{\eta}|\mathbf{y}, \hat{\mathbf{w}}, \boldsymbol{\gamma})$.

Finally, the functional network selected by our algorithm is summed up by the map of approximate posterior probabilities $\tilde{P}_j$ that the regions j are involved (as defined by (5.33)). For clarity, we have only represented regions j detected as involved ($\tilde{P}_j > 0.5$),

and with a positive regional mean estimate $\hat{\eta}_j$, *i.e.*, regions detected as active for the task under study.

6.2 Number processing task

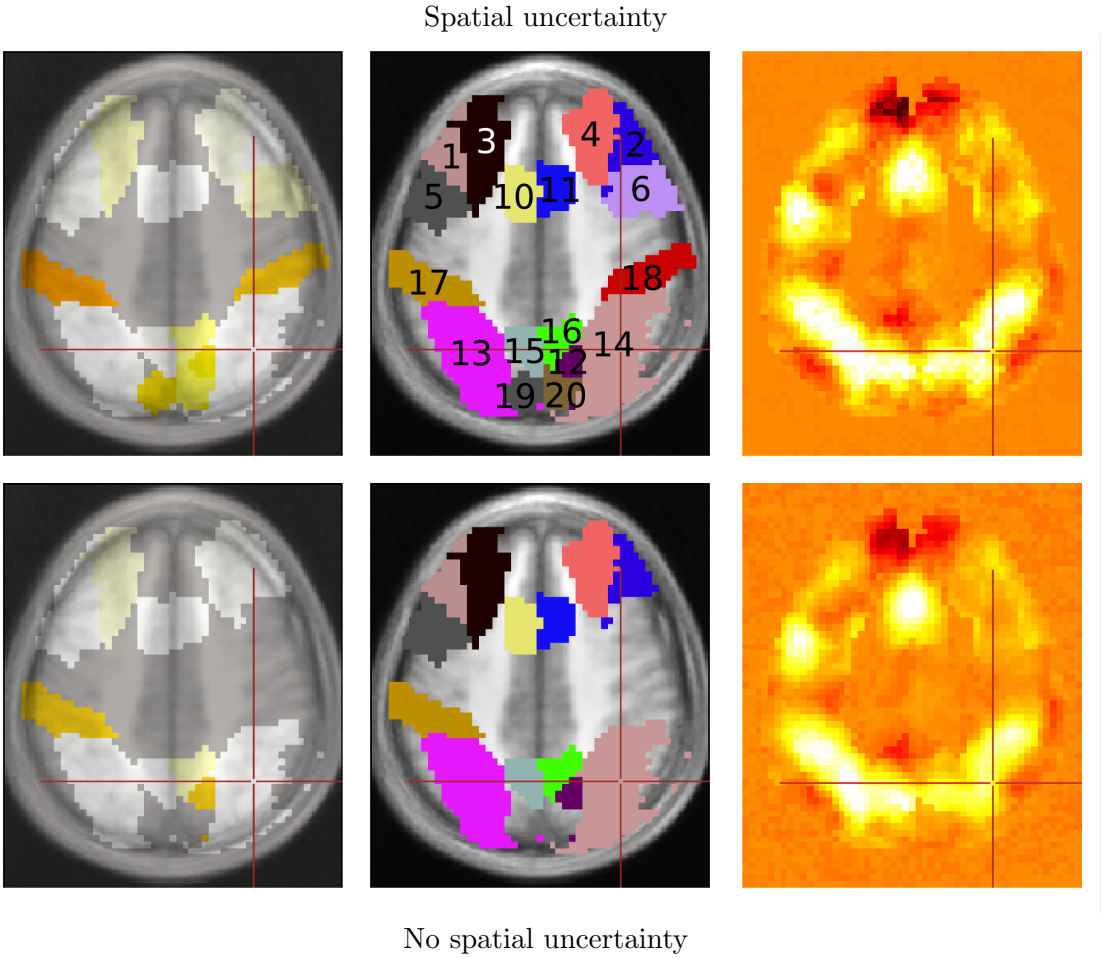


FIGURE 6.2: Number processing task, results of the model selection approach in axial slice $z = 37$mm in Talairach space, (top) using the model with spatial uncertainty, (bottom) with no spatial uncertainty. From left to right: approximate posterior probability map $\tilde{P}_j$, restricted to regions detected as activated ($\tilde{P}_j > 0.5$, $\hat{\eta}_j > 0$); labels of detected regions (numbers correspond to region index in Tables 6.1 and 6.2); posterior estimate of mean effect map $\hat{\boldsymbol{\mu}}$ (see Section 6.1.3). The first two maps are overlayed on the mean anatomical image of all subjects.

As can be seen from Figure 6.2, right, the estimated mean effect map $\hat{\boldsymbol{\mu}}$ is more contrasted using the model with spatial uncertainty than without, and slightly less noisy in the background. However, the regularizing effect observed previously (see Section 4.6), in a context where $\boldsymbol{\mu}$ was estimated by *marginalizing* out the elementary displacements $\mathbf{w}$, is absent here, since $\boldsymbol{\mu}$ is estimated *conditional* on a particular value $\hat{\mathbf{w}}$.

Sulcus/Fissure	$\tilde{P}_j$	$\hat{\eta}_j$	$\tilde{P}_j$	$\hat{\eta}_j$
	Spatial uncertainty		No spatial uncertainty	
Frontal lobe				
1: Left middle frontal	1.00	2.85	1.00	2.89
2: Right middle frontal*	1.00	3.47	1.00	3.20
3: Left superior frontal	0.89	2.21	0.92	2.11
4: Right superior frontal	0.96	1.87	0.98	1.85
5: Left inferior frontal*	1.00	4.48	1.00	4.04
6: Left middle precentral	0.99	4.53	1.00	4.50
7: Right middle precentral	0.90	1.99	*0.38*	*1.85*
8: Left inferior precentral	1.00	6.08	1.00	5.54
9: Right inferior precentral	0.92	2.83	0.58	2.54
10: Left anterior cingular	1.00	3.62	1.00	3.33
11: Right anterior cingular	1.00	4.26	1.00	4.02

TABLE 6.1: Number processing task, regions detected in the frontal lobe using the Bayesian model selection approach. Reported regions have a posterior probability of being involved in the task greater than 0.5, and constitute the most probable functional network given the data. $\hat{\eta}_j$ is the posterior estimate of the regional mean effect. Asterisks (*) mark regions that are found significant at 5% by the SPM-like approach (corrected cluster-level inference, cluster-forming threshold set to FPR $= 10^{-3}$), and correspond to the middle activation map in Figure 6.3.

The posterior probability and region label maps, illustrated in Figure 6.2 suggest that the network detected using both models is very similar, though more regions are detected using the model with spatial uncertainty, indicating an increased sensitivity.

The complete list of regions detected as activated ($\tilde{P}_j > 0.5$ and $\hat{\eta} > 0$) is given in Tables 6.1 and 6.2. Our method successfully detected the bilateral intra-parietal and fronto-cingular networks known to be active during number processing [Chochon et al., 1999, Dehaene et al., 2003]. Interestingly, the bilateral precuneus sulci where also detected. Although not considered as part of the core numerical system, the precuneus has been

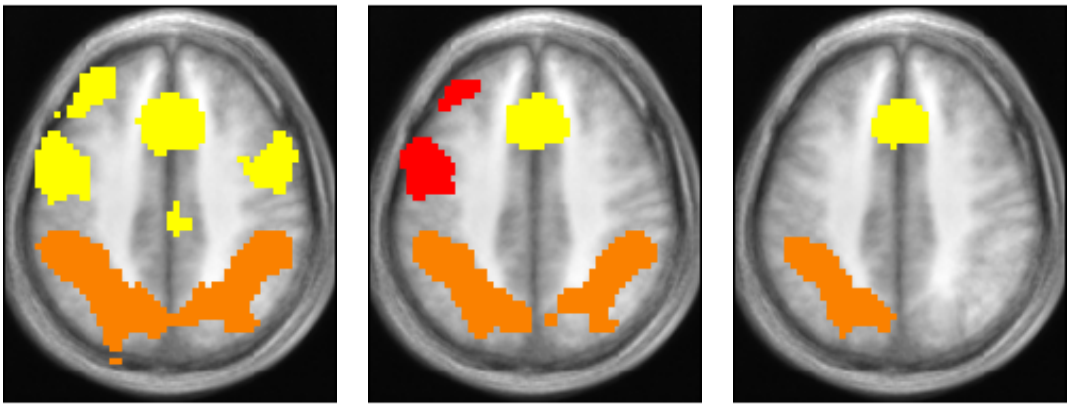
Sulcus/Fissure	$\tilde{P}_j$	$\hat{\eta}_j$	$\tilde{P}_j$	$\hat{\eta}_j$
	Spatial uncertainty		No spatial uncertainty	
Parietal Lobe				
12: Right transverse parietal	0.72	6.67	0.64	5.28
13: Left intra-parietal*	1.00	5.31	1.00	4.89
14: Right intra-parietal	1.00	3.50	1.00	2.95
15: Left precuneus	0.99	6.09	1.00	5.81
16: Right precuneus	0.82	4.10	0.88	3.74
17: Left postcentral intraparietal	0.53	2.35	0.64	2.14
18: Right postcentral intraparietal	0.61	2.07	*0.38*	*1.90*
19: Left parieto-occipital	0.68	1.75	*0.35*	*1.77*
20: Right parieto-occipital	0.78	1.53	*0.00*	*1.25*
Other				
21: Right callosal	0.97	2.10	0.59	1.99

TABLE 6.2: Number processing task, regions detected in the parietal lobe using the Bayesian model selection approach. Reported regions have a posterior probability of being involved in the task greater than 0.5, and constitute the most probable functional network given the data. $\hat{\eta}_j$ is the posterior estimate of the regional mean effect. Asterisks (*) mark regions that are found significant at 5% by the SPM-like approach (corrected cluster-level inference, cluster-forming threshold set to FPR $= 10^{-3}$), and correspond to the middle activation map in Figure 6.3.

linked to memory access and a wide range of high-level tasks [Cavanna and Trimble, 2006].

The networks detected with and without spatial uncertainty are very similar. However, 4 more regions were detected with spatial uncertainty; in particular, no activations were detected in the parieto-occipital region when neglecting spatial uncertainty. This is consistent with the fact that the estimated regional means $\hat{\boldsymbol{\mu}}$ are systematically higher under spatial uncertainty, and the higher contrast observed on the estimated mean effect map.

In contrast, only three activated clusters were detected by the SPM-like approach at the



PCER = 10^{-2} PCER = 10^{-3} PCER = 10^{-4}

FIGURE 6.3: Clusters detected at different cluster-forming thresholds, for the number processing task, in axial slices $z = 37$mm in Talairach, overlayed on the subjects' mean anatomical image in the background). The threshold is tuned to control the per-comparison error rate (PCER) respectively at 10^{-2}, 10^{-3} and 10^{-4} uncorrected. Each cluster surviving the FWER controlling-threshold at 5% is represented with a specific color.

chosen cluster-forming threshold. Each cluster contained over a thousand voxels, and extended over several atlas regions, hence merging several functionally distinct areas. Also, no activations were detected in the right frontal area. Using different thresholds could not solve these problems, as illustrated in Figure 6.3.

6.3 Language processing task

As in the previous case, a gain in the contrast of the estimated mean effect map $\hat{\boldsymbol{\mu}}$ associated with the model under spatial uncertainty can be observed in Figure 6.4, right. Apart from this, the regions detected in the displayed slice using both models were very similar, with a single additional region detected under spatial uncertainty.

The network detected by our method, shown in Table 6.3 was consistent with the known organization of language processing in the cerebral cortex, as described in [Pinel et al., 2007] for instance. This involves the superior temporal sulcus, with a dominance of the left hemisphere (which can be linked to the higher estimate of the regional mean observed here), and the left frontal areas (such as the precentral sulcus detected here) and the

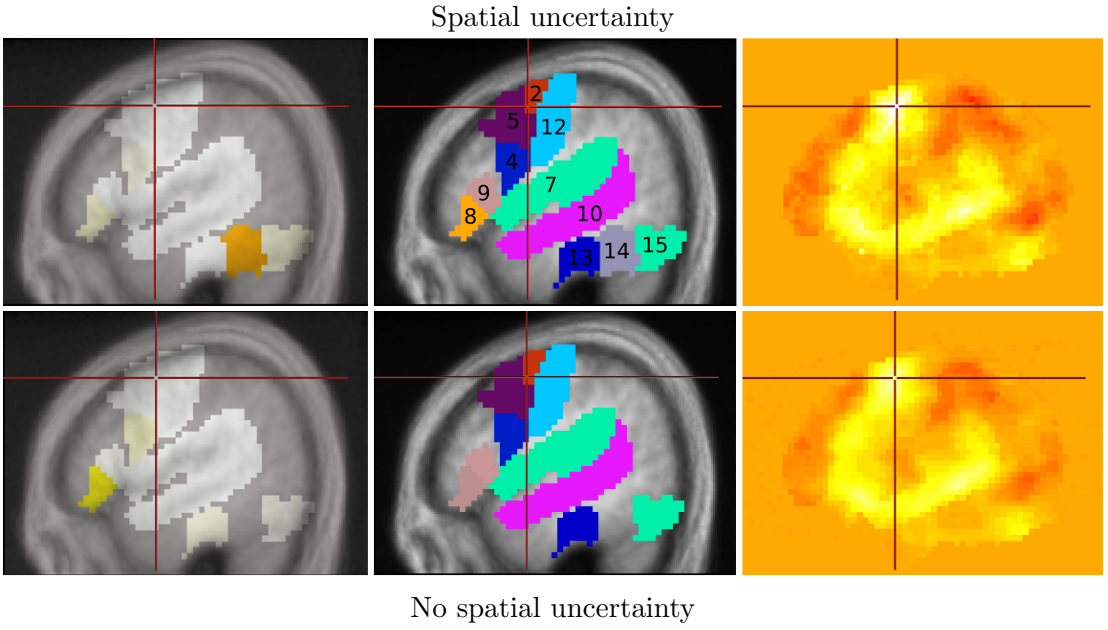


FIGURE 6.4: Language processing task, results of the model selection approach, in sagittal slice $x = -46$mm in Talairach space, (top) using the model with spatial uncertainty, (bottom) with no spatial uncertainty. From left to right: approximate posterior probabilities map $\tilde{P}_j$, restricted to regions detected as activated ($\tilde{P}_j > 0.5$, $\hat{\eta}_j > 0$); labels of detected regions (numbers correspond to region index in Table 6.3); posterior estimate of mean effect map $\hat{\boldsymbol{\mu}}$ (see Section 6.1.3). The first two maps are shown above the mean anatomical image of all subjects.

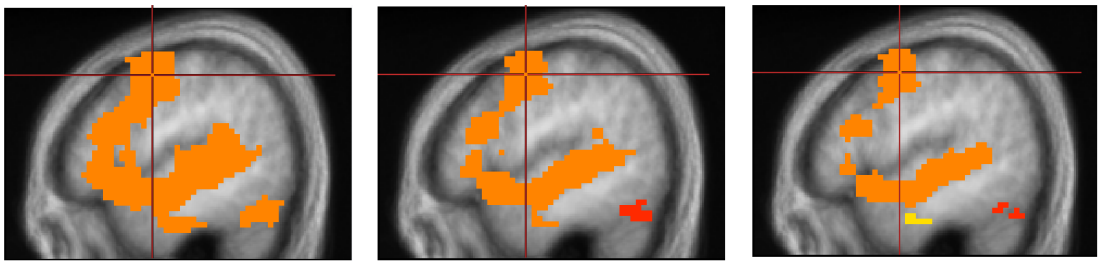


FIGURE 6.5: Clusters detected at different cluster-forming thresholds, for the language processing task, in the sagittal slice $x = -46$mm in Talairach,overlayed on the subjects' mean anatomical image in the background). The threshold is tuned to control the per-comparison error rate (PCER) respectively at 10^{-2}, 10^{-3} and 10^{-4} uncorrected. Each cluster surviving the FWER controlling-threshold at 5% is represented with a specific color.

supplementary motor area, part of the paracentral area detected here. Finally, activations were detected in the collateral fissure, which borders the lingual and the fusiform gyrus, both known to be involved in the visual recognition of words.

Again, there were some minor differences in the networks detected using both models,

Sulcus/Fissure	$\tilde{P}_j$	$\hat{\eta}_j$	$\tilde{P}_j$	$\hat{\eta}_j$
	Spatial uncertainty		No spatial uncertainty	
1: Left middle precentral	*0.05*	*2.19*	0.52	1.43
2: Left superior precentral	0.99	6.62	0.99	5.54
3: Right superior precentral	0.78	2.78	*0.00*	*2.23*
4: Left inferior precentral	0.97	3.64	0.91	3.33
5: Left middle precentral	0.99	5.11	0.99	4.74
6: Left paracentral	0.93	4.26	0.86	3.38
7: Left posterior sylvian	1.00	2.41	1.00	2.19
8: Left anterior sylvian	0.92	5.34	0.72	4.11
9: Left superior sylvian	1.00	5.64	0.98	5.07
10: Left superior temporal*	1.00	7.44	1.00	6.99
11: Right superior temporal*	1.00	4.5	1.00	3.94
12: Left central	1.00	2.0	1.00	1.74
13: Left anterior collateral	1.00	2.47	0.96	2.39
14: Left middle collateral	0.56	2.77	*0.17*	*2.51*
15: Left posterior collateral *	0.94	3.48	0.98	3.28

TABLE 6.3: Language processing task, regions detected using the Bayesian model selection approach. Reported regions have a posterior probability of being involved in the task greater than 0.5, and constitute the most probable functional network given the data. $\hat{\eta}_j$ is the posterior estimate of the regional mean effect. Asterisks (*) mark regions that are found significant at 5% by the SPM-like approach (corrected cluster-level inference, cluster-forming threshold set to FPR $= 10^{-3}$), and correspond to the middle activation map in Figure 6.5.

and the estimated regional means were generally higher in the model with spatial uncertainty. More importantly, the background, which was included in the list of regions, was detected as negatively active ($\tilde{P}_j = 1$, $\hat{\boldsymbol{\eta}} = -0.01$) in the model without spatial uncertainty, suggesting that some activations were projected outside the brain volume due to registration errors. This error was corrected when modeling spatially uncertainties.

In contrast, the SPM-like approach detected three clusters at the selected threshold (see Figure 6.5, middle). One of them (not represented) extended approximately over the right superior temporal sulcus, as defined in the CSA. Another one (in red), extended over several occipito-temporal parcels. Finally, a large cluster (in orange) containing over 1 000 voxels encompassed many different atlas regions, suggesting that it contained several functionally distinct areas. As in the case of number processing, this segmentation issue was not solved by varying the cluster-forming threshold.

6.4 Conclusion

In this chapter, we have validated the Bayesian model selection approach for fMRI group data analysis developed in Chapter 5 using a real dataset, by correctly recovering the brain functional networks associated to basic number and language processing, according to the description found in previous works. Furthermore, we have illustrated certain shortcomings of the classical, SPM-like approach, which fails to properly segment distinct functional regions, or misses them altogether, depending on the choice of the cluster-forming threshold.

Similar results were obtained using the models with and without spatial uncertainty. A probable explanation is the use of an additional penalty on data overfit in both cases, which prevented the false detection of inactive regions. However, the posterior mode approximation in the model with spatial uncertainty did improve the results, in that it allowed to detect additional regions, and enhanced the contrast of the estimated mean effect maps. This, together with the higher estimated regional mean effects, suggest that the "functional registration step", which consists in displacing the individual images according to the most probable displacements *a posteriori*, provided a better alignment of the activated regions of the different subjects.

Chapter 7

Conclusion

7.1 Main Results

Throughout this thesis, we have developed a new approach for the statistical analysis of multi-subject fMRI data, whose aim is to detect the cerebral regions involved in a certain cognitive task. Our objective was to address jointly certain limitations of the standard SPM-like approach, which are: the dependence on an arbitrary threshold to define clusters of potential activity; the lack of control over false negative risks; and the assumption that individual images are in perfect alignment.

In a first contribution, we revisited the adaptive thresholding technique developed in [Lavielle and Ludeña, 2007] by removing its dependence on a window parameter, making it more stable at low signal to noise ratios. This method provides an answer to the problem of choosing a detection threshold, while implicitly balancing false positive and false negative risks, by minimizing a model selection criterion rather than a multiple comparison type I error rate. It gave satisfying results when applied to individual and group activation maps, compared to Gamma-Gaussian mixture modeling [Beckmann et al., 2003b]. This technique may be best adapted to within-subject analysis however since it aims to detect individual voxels rather than regions, which are ultimately the objects of interest in group analysis.

The second contribution of this thesis consists in relaxing the assumption of perfect match between the estimated effect maps of the different subjects. To this end, we generalized the classical mass-univariate (voxelwise) model for fMRI group data analysis

by incorporating a set of hidden variables, representing the unknown registration errors, modeled as random deformation fields. In contrast, all previous approaches dealing with spatial uncertainty in fMRI data were feature-based, meaning that they aimed to match high-level features extracted from the individual activation maps. We proposed to estimate the map of group mean effects in a Bayesian setting by its posterior mean, showed the consistency of the joint posterior density of all model parameters, and designed a Metropolis-within Gibbs (MH-Gibbs) algorithm [Tierney, 1994] to draw samples from it, and compute Monte-Carlo estimates of the mean effect map.

Our simulation studies evidenced a stretching effect of the estimated activation pattern when the registration errors where unaccounted for, which caused neighboring activations to be merged. We also showed that this stretching effect could be substantially reduced when registration errors were modeled using our approach. When applied to real fMRI data, our method yielded estimates of the group effect map under spatial uncertainty that were both smoother and more contrasted than under no spatial uncertainty, an effect that could not be reproduced by linear isotropic smoothing. These encouraging results were obtained in spite of the slow mixing of our MH-Gibbs algorithm. In particular, the amplitude of the spatial displacements was under-estimated, suggesting some space for improvement.

Finally, we proposed a new paradigm for ROI-based fMRI group data analysis addressing jointly the three above-mentioned limitations of the SPM-like approach. Based on a Bayesian model selection framework, regions involved in the task under study are selected according to the posterior probabilities of a nonzero mean activation, given a pre-defined parcellation of the search volume into functionally homogeneous regions. Thus our approach is threshold-free, while allowing to incorporate prior information, provided that the parcellation is sensible. By controlling a Bayesian risk, our approach balances false positive and false negative risks, with weights than can be tuned depending on the application domain. Importantly, it is based on the same spatial uncertainty model as in Chapter 4, and thus accounts for the mis-alignment of individual images, due to inevitable registration errors. As a consequence, for each subject, the membership of a voxel to a given region is probabilistic rather than deterministic. This effectively allows to de-weight the contribution of activations to a region's mean level of activity, when they have been accidently projected into it by mis-registration.

This approach requires to evaluate the marginal likelihood of each model, defined as a partition of regions into involved and inactive. These marginal likelihoods are not available in closed form, but can be evaluated numerically. We chose to use Chib's method [Chib, 1995, Chib and Jeliazkov, 2001]. This required the specification of a MCMC-SAEM algorithm [Kuhn and Lavielle, 2004], adapted from the MH-Gibbs sampler above, to derive the MAP estimates of the model parameters, and a simulated annealing scheme to obtain the posterior mode of the displacement field density, conditional on these parameters.

Results on both simulated and real fMRI data show that the previously evidenced stretching effect may cause inactive regions to be contaminated by neighboring activations and be detected as active by mistake. This bias toward false positives is reduced when modeling spatial uncertainties. However, the marginal likelihood estimate in the model with spatial uncertainty turned out to be numerically instable, presumably because of the slow mixing observed when sampling the displacement fields. We proposed an approximate procedure, which consisted in fixing the displacement fields to their conditional MAP value, found by the above-mentioned simulated annealing algorithm. This approximation proved effective in stabilizing numerically the output of the algorithm, but did not entirely remove the bias toward false positives observed when neglecting spatial displacements. We compensated this residual bias by including an additional penalty on model fit, resulting in a lower bound on the probability of each region being involved in the task at hand. This final procedure was validated on both simulated and real fMRI datasets.

7.2 Perspectives

Many promising directions can be envisioned for future work, some of which have been mentioned earlier. To start with, the Metropolis-Hasting algorithm used to sample the displacement fields has a high rejection rate and mixes very slowly, as discussed in Appendix D, and Sections 4.7, 5.7, and 5.9. This results in the numerical instability of our model selection procedure, and needs to be dealt with, for instance by adopting alternative proposal densities to the isotropic and stationary Gaussian random distribution we use here. Indeed, the posterior density of the displacements is likely to be constrained along certain preferential directions, especially in highly contrasted regions of the mean

effect map $\boldsymbol{\mu}$, such as the interface between an active parcel and an inactive one. Also, we expect the posterior density to be more peaked in these regions than in less contrasted areas, such as the background. Generalizing the proposal's covariance structure, and allowing it to vary across control points, would therefore seem a reasonable direction toward which extending our work. Integrating out analytically $\mathbf{x}$ and $\boldsymbol{\mu}$ instead of simulating them could also promote mixing of the Markov chain and constitute another promising line of work.

Another issue requiring further investigation concerns the additional penalty fit we found necessary to include in our approximate procedure. This requires the tuning of a certain scale factor c, which we have done using simulated datasets. Though this was found to work well on a real-life application, it is necessary to study carefully the stability of the optimal value on a wider range of simulated datasets, in order to determine its possible dependence on certain critical variables, such as sample size or inter-subject variance.

In a broader perspective, the framework we propose here for fMRI group inference can be extended to address a wide variety of situations. For instance, we have focused on the characterization of the activation pattern of a single population of subjects. As previously discussed in Section 2.3.2, our model could be extended to compare different populations, such as a certain category of patients and healthy subjects, or to correlate fMRI activations with certain interest covariates. This means specifying the mean population effect, or, in terms of the regional response model (5.1), the mean regional effect, as a linear term in the regressor variables of interest. There is an increasing need for such models, particularly in studies combining neuroimaging and genetics, such as the IMAGEN project http://www.imagen-europe.com/, which constitute a new and promising trend in neurosciences.

Another exciting prospect, mentioned in Section 5.2, is to extend the model selection framework to infer the functional connectivity of the specified parcels. Following [Bowman et al., 2008], this would imply modeling the covariance structure of the regional means. Identification of functional networks in this setting would involve both selecting the regions involved in the task at hand, and the pairs of regions functionally correlated. The ensuing model selection problem would be much more challenging than the one addressed in the present work, the number of possible models involved being increased by the number of potential interactions (in contrast, [Bowman et al., 2008]

simply estimates the covariance matrix of the regional means, without performing any further inference). Furthermore, reducing this complexity would no longer be possible, even by adopting an approximation such as the one introduced here (see Section 5.8).

In another direction, our approach can be generalized to probabilistic instead of deterministic parcellations. This would especially make sense when using anatomical atlases, some of which are probabilistic to account for the inter-subject anatomical variability, such as the the CSA atlas [Perrot et al., 2008] (in Chapter 6, we used a deterministic version of this atlas, based on the most probable label for each voxel). Interestingly, this means that the labels defining the membership of each voxel would have a joint prior distribution, given by the probabilistic parcellation, and consequently also a posterior distribution. In other terms, a Bayesian estimate of the parcellation itself would be available, informed by the prior parcellation. Thus, our framework could be used to revisit group-level parcellation approaches, such as developed in [Flandin, 2004, Thirion et al., 2006c]. It is also an alternative answer to the choice of a parcellation, which we discuss in 5.6.

Finally, the methodology we have developed is well-adapted to the analysis of surface-based data (see Section 2.6.3), for the following reasons. First, because cortical structures are better matched across subjects using surface-based registration than classical volume-based affine registration [Fischl et al., 1999], we may expect better defined anatomical ROIs. The superiority of surface-based registration may not be as obvious when compared to nonlinear volume-based registration methods, such as those compared in [Klein et al., 2009]. However, to our knowledge, no comparison of surface versus volume based nonlinear registration has been conducted up to now.

On the other hand, we have seen that these approaches require projecting the fMRI data of each subject on its cortical surface, which cannot be done in a straightforward way, making the localization of activations on the cortical surface uncertain. Thus, the spatial uncertainty model we have developed to account for registration errors would still be necessary to account for these projection uncertainties, even though registration errors may be less important. Furthermore, the simulation study in Section 5.8.1 suggests a better behavior of our posterior mode approximation on 2D datasets than on 3D datasets, making this application to surface data a promising prospect.

Appendix A

Elements of multiple testing theory

A.1 Generalities on multiple testing

Consider the problem of testing simultaneously m distinct null hypotheses. In the context of neuroimaging, there can be one hypothesis per voxel, as defined in Section 2.3.3, or one hypothesis per cluster identified above a certain threshold. For all $k = 1, \ldots, m$, the test is based on a certain decision statistic T_k. $\mathcal{H}_k = 0$ indicates that the k-th null hypothesis is true, and $\mathcal{H}_k = 1$ that it is false.

The quantities of interest for multiple testing are then summarized in Table A.1, following [Benjamini and Hochberg, 1995]. We note $\mathcal{M}_0 = \{k : \mathcal{H}_k = 0\}$ and $\mathcal{M}_1 = \{k : \mathcal{H}_k = 1\}$ the sets of true null and false null hypotheses, respectively, and their number: $m_0 = |\mathcal{M}_0|, m_1 = |\mathcal{M}_1|$, which are unknown. $\mathcal{M} = \{1, \ldots, m\} = \mathcal{M}_0 \bigcup \mathcal{M}_1$ is the number of tested hypotheses. The only observed variables are the number of rejected null hypotheses R, and the number of accepted null hypotheses, $R - m$. The goal of any multiple testing procedure is to minimize both the number V of type I errors, or false positives, and the number T of type II errors, or false negatives.

	# accepted	# rejected	
# True null hypotheses	U	V	m_0
# False null hypotheses	T	S	m_1
t	W	R	m

TABLE A.1: Number of errors in a multiple testing problem.

Type I error rates

A common strategy for limiting the number of errors is to maximize the statistical power of the tests while controlling a certain type I error rate at a given level α. The error rates most often used, as defined in [Ge et al., 2003], are the following:

- *False positive rate* (FPR). It is the expectation of the proportion of type I errors among all the tests:

$$\text{FPR} = E(V)/m \tag{A.1}$$

- *Family-wise error rate*(FWER). It is the probability of at least one type I error:

$$\text{FWER} = P[V > 0] \tag{A.2}$$

- *False discovery rate*(FDR). It is the expectation of the proportion of type I errors among all rejected null hypotheses. When no null hypotheses have been rejected, this proportion is set to 0, yielding:

$$\text{FDR} = E\left[\frac{V}{R}1_{R>0}\right] \tag{A.3}$$

These quantities can be compared through the following inequality [Ge et al., 2003]:

$$\text{FPR} \leq \text{FDR} \leq \text{FWER}.$$

Thus, if the FWER is controlled at a given level α, *i.e.*, if FWER $\leq \alpha$, then the other rates are also automatically controlled at the same level. FWER is the most stringent criterion possible for multiple testing. As such, it is very much used in neuroimaging,

where a high level of confidence is required to report a brain region as involved in a certain task. Conversely, the FPR involves no correction at all for multiple comparisons, since it is controlled at a given level α as soon as the individual hypoteses $\mathcal{H}_k = 0$ are tested at level α.

Exact, weak and strong control

The error rates defined in the previous section are implicitely defined conditionally on the intersection of all null hypotheses: $\mathcal{H}_{\mathcal{M}_0} = \bigcap_{k\in\mathcal{M}_0}\{\mathcal{H}_k = 0\}$, as noted in [Ge et al., 2003]. Control on a given error rate conditional on the true intersection $\mathcal{H}_{\mathcal{M}_0}$ of null hypotheses is called *exact control*. For instance, a multiple comparison procedure has exact control on the FWER if it can control the quantity: $P[V > 0|\mathcal{H}_{\mathcal{M}_0}]$ at any given level α.

However the set $\mathcal{M}_0$ is unknown, so the error rates are generally computed, and controlled, under the global null hypothesis $\mathcal{H}_{\mathcal{M}} = \bigcap_{k\in\mathcal{M}}\{\mathcal{H}_k = 0\}$, referred to as a *weak control*. It is important to note that weak control alone does not in general imply exact control. Hence, a multiple testing procedure which has only weak control is in general not a valid procedure, since its results hold only under the assumption that all null hypotheses are true.

Finally, *strong control* means control for every possible choice of $\mathcal{M}_0$. For instance, strong control of the FWER means to be able to control the quantity $\max_{\mathcal{M}'\subseteq\mathcal{M}} P[V > 0|\mathcal{H}_{\mathcal{M}'}]$ at any given level α. Strong control implies both weak and exact control, and is therefore sufficient to define a valid mutiple comparison procedure.

Subset pivotality and p-values

Subset pivotality is a central property in multiple testing. It is used to ensure that a multiple comparison procedure having weak control over a certain error rate also has strong control. It is usually defined as follows [Westfall and Young, 1993]:

Definition A.1 (Subset Pivotality)**.** The joint distribution of the test statistics $(T_1, \ldots, T_m)$ is said to have the *subset pivotality* condition if for all subset $\mathcal{K} \subset \mathcal{M}$, the joint distribution of $(T_k)_{k\in\mathcal{K}}$ is the same under the global null hypothesis $\mathcal{H}_{\mathcal{M}} = \bigcap_{k\in\mathcal{M}}\{\mathcal{H}_k = 0\}$ as under the restriction $\mathcal{H}_{\mathcal{K}} = \bigcap_{k\in\mathcal{K}}\{\mathcal{H}_k = 0\}$.

An immediate consequence of this definition is that, under subset pivotality, the joint distribution of $(T_k)_{k\in\mathcal{K}}$ is the same under any intersection of null hypotheses including $\mathcal{H}_\mathcal{K}$. Informally, subset pivotality implies that the test of each given null hypothesis $\mathcal{H}_k = 0$ can be done independently of the status of all other hypotheses $\mathcal{H}_{k'}$, which is a quite natural requirement. For instance, if each null hypothesis $\mathcal{H}_k = 0$ concerns a data vector D_k, if the D_k's are independent, and T_k is a function of D_k only, then the subset pivotality property is trivially verified.

Otherwise, there is no general characterization of subset pivotality. We will see that most multiple testing procedures that have strong control rely on this property.

To illustrate the importance of this notion, consider the p-values, which are another useful tool in the multiple testing setting. They are traditonally defined as follows:

Definition A.2 (p-values)**.** Note $(t_1, \ldots, t_m)$ the observed values of the test statistics $(T_1, \ldots, T_m)$. For $k = 1, \ldots, m$, The *p-value* p_k associated with t_k is given by:

$$p_k = P[T_k > t_k | \mathcal{H}_k = 0]. \tag{A.4}$$

Thus, p_k is the lowest level α at which $\mathcal{H}_k = 0$ is rejected on the basis of t_k.

Note that, under subset pivotality, the quantity defined in (A.4) is well-defined, since the probability in the right term takes the same value under any intersection of null hypotheses including $\mathcal{H}_k = 0$, and in particular under the global null. On the other hand, if subset pivotality is not verified, then the probability of $T_k > t_k$ depends not only on $\mathcal{H}_k$, but on the state of potentially all other hypotheses $\mathcal{H}_{k'}$, hence the quantity in (A.4) has more than one possible value, and is consequently ill-defined. This is an important point, though scarcely mentioned in the multiple testing llitterature. It shows that multiple comparison procedures based on p-values often rely implicitely on the subset pivotality property.

Proposition A.3. Strong control of the maxT procedure

If the joint distribution of the test statistics $(T_k)_{1\leq k\leq d}$ verifies the subset pivotality property (Definition A.1), The maxT procedure, defined in Section 2.4.1, has strong control over the family wise error rate, i.e.*:*

$$FWER \quad \leq \quad P\left[\max_{k\in\mathcal{M}} T_k > u | \mathcal{H}_\mathcal{M}\right]. \tag{A.5}$$

Proof. This follows directly from the definitions:

$$\begin{aligned} FWER &= P[V > 0 | H_{\mathcal{M}_0}] \\ &= P[\exists k \in \mathcal{M}_0, T_k > u | \mathcal{H}_{\mathcal{M}_0}] \\ &= P\left[\max_{k \in \mathcal{M}_0} T_k > u | \mathcal{H}_{\mathcal{M}_0}\right] \\ &= P\left[\max_{k \in \mathcal{M}_0} T_k > u | \mathcal{H}_{\mathcal{M}}\right] \\ &\leq P\left[\max_{k \in \mathcal{M}} T_k > u | \mathcal{H}_{\mathcal{M}}\right], \end{aligned} \tag{A.6}$$

where the the fourth equality holds under subset pivotality □

A.2 Strong control of the maxT and maximum cluster size tests

Based on the definitions in Section 2.4.1, we now show that the test on the maximum suprathreshold cluster also has strong control:

Proposition A.4. Strong control of the maximum cluster size test *If the joint distribution of the test statistics* $(T_k)_{1 \leq k \leq d}$ *verifies the subset pivotality property, then the test on the maximum cluster size has strong control over the family-wise error rate:*

$$FWER \leq P[\max_{C_i \subseteq \mathcal{M}} \sharp C_i > N | \mathcal{H}_{\mathcal{M}}].$$

Proof. The cluster-level family-wise error rate (2.10) can be expressed as:

$$\begin{aligned} FWER &= P[\max_{C_i \subseteq \mathcal{M}_0} \sharp C_i > N | \mathcal{H}_{\mathcal{M}_0}] \\ &\leq P[\max_{C_i \subseteq \mathcal{M}} \sharp C_i \cap \mathcal{M}_0 > N | \mathcal{H}_{\mathcal{M}_0}]. \end{aligned} \tag{A.7}$$

This last event depends only $(T_k)_{k \in \mathcal{M}_0}$, whereas the first one depends on whether $C_i \subseteq \mathcal{M}_0$ for each suprathreshold clusters C_i, an event depending on statistics $T_{k'}$ in all

neighboring voxels k' of C_i, some of whom may be outside $\mathcal{M}_0$ (a voxel is said to be neighboring C_i if it is neighbor to a voxel contained in C_i).

Hence, we may apply subset pivotality, yielding:

$$\begin{aligned} FWER &\leq P[\max_{C_i \subseteq \mathcal{M}} \sharp C_i \cap \mathcal{M}_0 > N | \mathcal{H}_{\mathcal{M}}]. \\ &\leq P[\max_{C_i \subseteq \mathcal{M}} \sharp C_i | \mathcal{H}_{\mathcal{M}}] \square \end{aligned} \tag{A.8}$$

Appendix B

Proof of the consistency of the random threshold procedure

Following [Lavielle and Ludeña, 2007], we first recall some notations. Set $u_i = y_i$ for $i \in I_{k_n^\star}$ and $v_i = y_i$ for $i \notin I_{k_n^\star}$; notice that (v_i) is a sample from the distribution F_ϵ. Let $(u_{(i)})_{1 \leq i \leq k_n^\star}$ and $(v_{(i)})_{1 \leq i \leq n-k_n^\star}$ be the sequences $(|u_i|)$ and $(|v_i|)$ in decreasing order. Let Ω_n be the subset of Ω where $v_{(1)} < \alpha_n/2$ and $u_{(k_n^\star)} > \alpha_n/2$.

A first lemma in [Lavielle and Ludeña, 2007] shows that $P(\Omega_n) \to 1$, *i.e.*, the collections $(u_{(i)})$ and $(v_{(i)})$ are stochastically in order with high probability. The proof can then be restricted to Ω_n.

Now, let $\mathbb{E}_k(T_{k,j})$ and $Q_{k,j}$ be defined as in Equation (3.1). Using Proposition 3.1, we have:

$$\begin{aligned} \mathbb{E}_k(T_{k,j}) &= j(1 + \sum_{i=j+1}^{n-k} 1/i); \\ Q_{k,j} &= \frac{\mathbb{E}_k(T_{k,j})}{\mathbb{E}_k(T_{k,n-k})} T_{k,n-k} \\ &= B_{k,j,n} T_{k,n-k}. \end{aligned}$$

Also, let $a_i = \mathbb{E}_0(X_{(i)}) = \sum_{\ell=i}^{n} 1/\ell$. Equation (3.2) can be shown separately for $k > k_n^\star$ and $k < k_n^\star$. Since the two cases are treated similarly, we will restrict ourselves here to the case $k > k_n^\star$. On Ω_n :

$$\begin{aligned} T_{k,j} - Q_{k,j} &= T_{k,j} - B_{k,j,n} T_{k,n-k} \\ &= \left(T_{k,j} - \mathbb{E}_{k_n^\star}(T_{k,j})\right) - B_{k,j,n}\left(T_{k,n-k} - \mathbb{E}_{k_n^\star}(T_{k,n-k})\right) \\ &\quad + \mathbb{E}_{k_n^\star}(T_{k,j}) - B_{k,j,n}\mathbb{E}_{k_n^\star}(T_{k,n-k}) \\ &= R_{k,j} + S_{k,j}. \end{aligned}$$

Thus $T_{k,j} - Q_{k,j}$ is decomposed into a random part $R_{k,j}$ and a deterministic part $S_{k,j}$. Over Ω_n, $R_{k,j}$ is a function of $v_{(k)}, \ldots, v_{(n-k_n^\star)}$. Before going further, we now recall the following result:

Let $X_{(1)} \geq \ldots \geq X_{(n)}$ be an ordered sequence of independent $Exp(1)$ random variables. For $1 \leq j \leq n$, let $T_j = \sum_{i=1}^{j} X_{(i)}$. Introduce for $t \in [0,1]$ the random process $d_n(t) = T_{[nt]} - \mathbb{E}(T_{[nt]}|T_n)$. Then it is shown in [Lavielle and Ludeña, 2007] that $\frac{1}{\sqrt{n}} d_n(t)$, as a process indexed on $t \in [0,1]$, converges in distribution to a certain zero mean Gaussian process Δ.

To use this result, let $k = [tn]$ and $j = [sn]$, for $0 < t < 1 - c$ and $0 < s < 1 - t^\star - t$, for c in [**AF3**]. Then $\frac{1}{\sqrt{n-k}}(T_{k,j} - Q_{k,j})\mathbf{1}_{\Omega_n} = \frac{1}{\sqrt{n-k}}(T_{[tn],[sn]} - Q_{[tn],[sn]})\mathbf{1}_{\Omega_n}$, as a process indexed by $(t,s) \in (0,1)^2$, converges in distribution to the zero-mean Gaussian process:

$$\Gamma_{t,s} = \sqrt{\frac{1-t^\star}{1-t}}\left[\Delta\left(\frac{t+s-t^\star}{1-t^\star}\right) - \Delta\left(\frac{t-t^\star}{1-t^\star}\right)\right].$$

similarly, $\frac{1}{\sqrt{n-k}} B_{k,j,n}\mathbb{E}_{k_n^\star}(T_{k,n-k})\mathbf{1}_{\Omega_n}$ converges in distribution to another zero-mean Gaussian process, and so does their sum, $\frac{1}{\sqrt{n-k}} R_{k,j}\mathbf{1}_{\Omega_n}$.

On the other hand,

$$S_{k,j} = \sum_{i=1}^{k-k_n^\star} (a_{i+j} - a_i + B_{k,j,n}(a_{i+n-k} - a_i)), \tag{B.1}$$

so that there exists a constant $\gamma > 0$, which depends on c in [**AF3**], such that for all $n \geq 1,\ k_n^\star < k \leq n - K_n$, we have $\sup_{1\leq j\leq n-k} |S_{k,j}| \geq \gamma(k - k_n^\star)$. Finally we use the following inequality:

$$\mathbb{P}_{k_n^\star}(\hat{k}_n - k_n^\star > nu_n) \quad \leq \quad \mathbb{P}(\eta_{k_n^\star} > \inf_{k-k_n^\star>nu_n} \eta_k).$$

From Equation (B.1), $S_{k_n^\star,j} = 0$, hence it follows that:

$$\begin{aligned}\sqrt{n-k_n^\star}\,\eta_{k_n^\star} &= \sup_{1\leq j\leq n-k_n^\star} |R_{k_n^\star,j} + S_{k_n^\star,j}| \\ &= \sup_{1\leq j\leq n-k_n^\star} |R_{k_n^\star,j}| \\ &\leq \sup_{k\geq k_n^\star} \sup_{1\leq j\leq n-k} |R_{k,j}|.\end{aligned}$$

On the other hand,

$$\begin{aligned}\sqrt{n-k} \inf_{k-k_n^\star>nu_n} \eta_k &= \inf_{k-k_n^\star>nu_n} \sup_{1\leq j\leq n-k} |R_{k,j} + S_{k,j}| \\ &\geq \inf_{k-k_n^\star>nu_n} \sup_{1\leq j\leq n-k} |S_{k,j}| - \sup_{k\geq k_n^\star} \sup_{1\leq j\leq n-k} |R_{k,j}|,\end{aligned}$$

so that we have:

$$\begin{aligned}\mathbb{P}_{k_n^\star}(\hat{k}_n - k_n^\star > nu_n) &\leq \mathbb{P}(C \sup_{k\geq k_n^\star} \sup_{1\leq j\leq n-k} |R_{k,j}| \geq \inf_{k-k_n^\star>nu_n} \sup_{1\leq j\leq n-k} |S_{k,j}|) + \mathbb{P}(\Omega_n^c) \\ &\leq \mathbb{P}(C \sup_{k\geq k_n^\star} \sup_{1\leq j\leq n-k} |R_{k,j}| \geq \gamma n u_n) + \mathbb{P}(\Omega_n^c),\end{aligned}$$

where C is a constant which depends on c in [**AF3**]. This last probability vanishes as n goes to infinity, due to the weak convergence of $R_{k,j}\mathbf{1}_{\Omega_n}$ □

Appendix C

Identifiability in the model with spatial uncertainty

We show here the identifiability of the parameters in the hierarchical model introduced in Chapter 5, $\boldsymbol{\theta} = (\boldsymbol{\eta}, \boldsymbol{\nu}^2, \boldsymbol{\sigma}^2, \sigma_S^2)$. Identificability in the model in Chapter 4 requires a slightly different proof, since the parameter vector is defined by $\boldsymbol{\theta} = (\boldsymbol{\mu}, \sigma^2, \sigma_S^2)$. We have not included it here however, as it uses very similar arguments.

Theorem C.1. (Identifiability in the Regionalized Model) *The full hierarchical model specified by (2.5), (4.2), (4.3), (4.4), (5.1) is identifiable.*

Proof. This result may be re-phrased by saying that, if $\boldsymbol{\theta} \neq \boldsymbol{\theta}'$, then $f(\mathbf{y}|\boldsymbol{\theta}) \neq f(\mathbf{y}|\boldsymbol{\theta}')$ in at least one point $\mathbf{y}$. As noted previously in Appendix F, the conditional density $f(\mathbf{y}|\mathbf{w}, \boldsymbol{\theta})$ is multivariate Gaussian. It depends on $\mathbf{w}$ only through the definition of the voxelwise blocks

$$I_k = \left\{ (i,l)_{\substack{1 \leq i \leq n \\ 1 \leq l \leq d}} \middle| \varphi_i(l) = k \right\},$$

containing for each k the indices of observations $y_{i,l}$ displaced to voxel k. The collection of all voxelwise blocks $\mathbf{I} = (I_1, \ldots, I_d)$ constitutes a partition of the cartesian product $\{1, \ldots, n\} \times \{1, \ldots, d\}$ into d subsets. Noting $\mathcal{I}$ the set of all possible such partitions, we see that the PDF can be re-written as the mixture:

$$f(\mathbf{y}|\boldsymbol{\theta}) = \sum_{\mathbf{I} \in \mathcal{I}} \pi(\mathbf{I}|\sigma_S^2) f(\mathbf{y}|\mathbf{I}, \boldsymbol{\eta}, \boldsymbol{\nu}^2, \boldsymbol{\sigma}^2),$$

where $\pi(\mathbf{I}|\sigma_S^2) = \int_{\mathbf{w}\in\mathbf{I}} \pi(\mathbf{w}|\sigma_S^2)d\mathbf{w}$ is the probability of block configuration $\mathbf{I}$, and $f(\mathbf{y}|\mathbf{I}, \boldsymbol{\eta}, \boldsymbol{\nu}^2, \boldsymbol{\sigma}^2)$ is multivariate Gaussian, with mean and covariance matrix given by functions of $(\boldsymbol{\eta}, \mathbf{I})$ and $(\boldsymbol{\sigma}^2, \mathbf{I})$, respectively. Note that these functions depend on $\mathbf{w}$ (through $\mathbf{I}$), and also that they may not be one-to-one. For instance, the observations may well be all displaced into a single region j_0, in which case the regional parameters $(\eta_j, \nu_j^2, \sigma_j^2)$ of all other regions have no effect on the data.

Because any finite family of distinct Gaussian PDFs is linearly independent, and the mixing weights $\pi(\mathbf{I}|\sigma_S^2)$ in the above display are always strictly positive, the only way we can have $f(\mathbf{y}|\boldsymbol{\theta}) = f(\mathbf{y}|\boldsymbol{\theta}')$ in all data points $\mathbf{y}$ is if there exists a permutation δ of $\mathcal{I}$, such that for all $\mathbf{I} \in \mathcal{I}$ and all $\mathbf{y} \in \mathbb{R}^{nd}$:

$$\begin{aligned} \pi(\mathbf{I}|\sigma_S^2) &= \pi(\delta(\mathbf{I})|\sigma_S'^2) \\ f(\mathbf{y}|\mathbf{I}, \boldsymbol{\eta}, \boldsymbol{\nu}^2, \boldsymbol{\sigma}^2) &= f(\mathbf{y}|\delta(\mathbf{I}), \boldsymbol{\eta}', \boldsymbol{\nu}'^2, \boldsymbol{\sigma}'^2). \end{aligned} \tag{C.1}$$

This the analog of the 'label-switching' phenomenon in classical finite mixture models. We now show that such label-switching is impossible, because the partitions $\mathbf{I}$ are not equally probable. Note $\mathbf{I}_0$ the most probable block configuration,

$$\mathbf{I}_0 = \arg\max_{\mathbf{I}\in\mathcal{I}} \pi(\mathbf{I}|\sigma_S^2).$$

Because $\mathbf{w}$ is a zero-mean Gaussian with spherical covariance (4.4), this most probable configuration is obtained for $\mathbf{w} = \mathbf{0}$, so that $\mathbf{I}_0 = (I_k^0)_{1\leq k\leq d}$, where for all k: $I_k^0 = \{(1,k), \ldots, (n,k)\}$. Likewise, $\delta(\mathbf{I}_0) = \mathbf{I}_0$, hence from C.1 it follows that for all $\mathbf{y}$:

$$f(\mathbf{y}|\mathbf{I}_0, \boldsymbol{\eta}, \boldsymbol{\nu}^2, \boldsymbol{\sigma}^2) = f(\mathbf{y}|\mathbf{I}_0, \boldsymbol{\eta}', \boldsymbol{\nu}'^2, \boldsymbol{\sigma}'^2),$$

which implies that $(\boldsymbol{\eta}, \boldsymbol{\nu}^2, \boldsymbol{\sigma}^2) = (\boldsymbol{\eta}', \boldsymbol{\nu}'^2, \boldsymbol{\sigma}'^2)$.

Finally, $\mathbf{I}_0$ is a star domain in that if $\mathbf{w} \in \mathbf{I}_0$, then the segment $\{\lambda\mathbf{w}, \lambda \in (0,1)\}$ lies inside $\mathbf{I}_0$. Thus $\pi(\mathbf{I}_0|\sigma_S^2)$ is a strictly decreasing function of σ_S^2, and since $\pi(\mathbf{I}_0|\sigma_S^2) = \pi(\mathbf{I}_0|\sigma_S'^2)$, we have $\sigma_S^2 = \sigma_S'^2$ □

Appendix D

Sampling the posterior density using a Metropolis within Gibbs algorithm

As described at the beginning of Section 5.5, the approach in [Chib, 1995] assumes that a MCMC algorithm has been devised to sample the posterior density of the model parameters . As expressed in (5.18), this density is obtained as a marginal of the joint posterior density:

$$\begin{aligned} \pi(\mathbf{z}, \boldsymbol{\theta}|\mathbf{y}, \boldsymbol{\gamma}) &= \pi(\mathbf{x}, \mathbf{w}, \boldsymbol{\mu}, \boldsymbol{\eta}, \boldsymbol{\nu}, \boldsymbol{\sigma}^2, \sigma_S^2|\mathbf{y}, \boldsymbol{\gamma}) \\ &\propto f(\mathbf{y}|\mathbf{x})\pi(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\sigma}^2, \mathbf{w})\pi(\boldsymbol{\mu}|\boldsymbol{\eta}, \boldsymbol{\nu}^2)\pi(\mathbf{w}|\sigma_S^2)\pi(\boldsymbol{\eta}|\boldsymbol{\nu}^2)\pi(\boldsymbol{\nu}^2, \boldsymbol{\sigma}^2, \sigma_S^2). \end{aligned} \tag{D.1}$$

We use the Gibbs sampler algorithm [Geman and Geman, 1984] to generate a sequence of samples from this joint density by sampling successively each of the six following blocks: $\mathbf{x}$, $\mathbf{w}$, $\boldsymbol{\mu}$, $(\boldsymbol{\eta}, \boldsymbol{\nu})$, $\boldsymbol{\sigma}^2$, and σ_S^2, conditionally on all others. The conditional density of each block is obtained from the complete density (D.1), by treating all other blocks as constants. These conditional densities are available in closed form, except for the elementary displacements $\mathbf{w}$, and are given as follows.

Hidden effects. For each subject i and each voxel k, such that the displaced voxel $\varphi_i(k)$ is in region j, it follows from (4.1) and (4.2) that:

$$x_{i,k}|\ldots \sim \mathcal{N}\left(\frac{\sigma_j^2\, y_{i,k} + s_{i,k}^2\, \mu_{\varphi_i(k)}}{\sigma_j^2 + s_{i,k}^2}, \frac{\sigma_j^2\, s_{i,k}^2}{\sigma_j^2 + s_{i,k}^2}\right). \tag{D.2}$$

Population mean effect. According to (4.2 and 5.1), for all voxel k in region j,

$$\mu_k|\ldots \sim \mathcal{N}\left(\frac{\nu_j^2\, m_k + s_k^2\, \eta_j}{\nu_j^2 + s_k^2}, \frac{\nu_j^2\, S_k^2}{\nu_j^2 + S_k^2}\right), \tag{D.3}$$

where $m_k = n_k^{-1} \sum_{\phi_i(l)=k} x_{i,k}$ and $S_k^2 = n_k^{-1} \sum_{\phi_i(l)=k} (x_{i,k} - m_k)^2$ are the empirical mean and variance of the effects centered on μ_k, and $n_k = \sharp\{(i,k); \phi_i(l) = k\}$ the number of these.

Regional parameters. For each region j, (5.1), (5.11) and (5.11) yield:

$$\eta_j|\nu_j^2, \gamma_j = 0, \ldots \sim \delta(0); \tag{D.4}$$

$$\eta_j|\nu_j^2, \gamma_j = 1, \ldots \sim \mathcal{N}\left(\frac{\sum_{\ell_k=j} \mu_k}{\lambda + d_j}, \frac{\nu_j^2}{\lambda + d_j}\right); \tag{D.5}$$

$$\nu_j^2|\ldots \sim \mathcal{IG}\left(\alpha + d_j,\, \beta + \frac{1}{2}\sum_{\ell_k=j} \mu_k^2 - \gamma_j \frac{\left(\sum_{\ell_k=j} \mu_k\right)^2}{2(\lambda + d_j)}\right), \tag{D.6}$$

where $d_j = \sharp\{k; \ell_k = j\}$ is the size of region j.

Population variance. For each region j, the posterior density of σ_j^2 is computed from (4.2) and (5.12), resulting in:

$$\sigma_j^2|\ldots \sim \mathcal{IG}\left(\alpha + \frac{1}{2}\sum_{\ell_k=j} n_k,\, \beta + \frac{1}{2}\sum_{\ell_k=j}\sum_{\varphi_i(l)=k} (x_{i,l} - \mu_k)^2\right). \tag{D.7}$$

Spatial variance. The posterior density of the spatial variance σ_S^2 is given by (4.3) and (5.12):

$$\sigma_S^2|\ldots \sim \mathcal{IG}\left(\alpha + \frac{3nB}{2},\, \beta + \frac{\sum_{i,b} \| \mathbf{w}_{i,b} \|^2}{2}\right). \tag{D.8}$$

Elementary displacements. For each subject i and each control point b, according to (4.2) and (4.3):

$$\mathbf{w}_{i,b}|\ldots \quad \propto \quad \pi(\mathbf{w}_{i,b}|\sigma_S^2)\pi(\mathbf{x}_i|\boldsymbol{\mu},\boldsymbol{\sigma}^2,\mathbf{w}_i), \tag{D.9}$$

where the prior conditional densities of $\mathbf{w}_{i,b}$ and $\mathbf{x}_i$ are defined in (4.4) and (4.2). This density cannot be computed analytically, let alone directly sampled. Indeed, $\pi(\mathbf{x}_i|\boldsymbol{\mu},\boldsymbol{\sigma}^2,\mathbf{w}_i)$, as a function of $\mathbf{w}_i$, is piecewise constant, due to the discrete interpolation of the mean population map (see Section 4.2). Instead, we sample each elementary displacement $\mathbf{w}_{i,b}$ using the Metropolis-Hastings algorithm [Hastings, 1970], wherein a candidate value $\mathbf{w}'_{i,b}$ is sampled from any proposal density $q(\mathbf{w}'_{i,b})$ (which may depend on the other variables), and accepted with probability:

$$\alpha \quad = \quad \min\left\{1, \frac{\pi(\mathbf{w}'_{i,b}|\sigma_S^2)\pi(\mathbf{x}_i|\boldsymbol{\mu},\boldsymbol{\sigma}^2,\mathbf{w}'_i)}{\pi(\mathbf{w}_{i,b}|\sigma_S^2)\pi(\mathbf{x}_i|\boldsymbol{\mu},\boldsymbol{\sigma}^2,\mathbf{w}_i)}\frac{q(\mathbf{w}_{i,b})}{q(\mathbf{w}'_{i,b})}\right\}. \tag{D.10}$$

In the above expression, $\mathbf{w}'_i$ stands for the vector $\mathbf{w}_i$, where $\mathbf{w}_{i,b}$ has been replaced by the candidate value $\mathbf{w}'_{i,b}$. This means that our Gibbs sampler is actually a *Metropolis within Gibbs* algorithm (see for instance [Tierney, 1994]). We used a random walk proposal $q(\mathbf{w}'_{i,b}) = \mathcal{N}(\mathbf{w}'_{i,b}; \mathbf{w}_{i,b}, \sigma_{RW}^2\mathbf{I}_3)$, with a proposal variance σ_{RW}^2 tuned during the burn-in period to obtain an acceptance rate close to 0.1. A higher acceptance rate of 0.25, as advocated in [Roberts et al., 1997], proved a bad idea in our case, as it resulted in such a low proposal variance σ_{RW}^2 that the Markov Chain remained almost unmoving.

Appendix E

Maximization of the posterior density using a MCMC-SAEM algorithm

We start by recalling basic facts concerning the MCMC-SAEM coupling developed in [Kuhn and Lavielle, 2004], as described hereafter. This extension of the SAEM algorithm allows to find the maximum likelihood, or the maximum a posteriori, in mixture models where the conditional distribution of the hidden variables can only be sampled by an MCMC algorithm, as is our case. The main prerequisite in [Kuhn and Lavielle, 2004] is that the complete likelihood belongs to the curved exponential family. In our case, this means that:

$$\begin{aligned} f(\mathbf{y},\mathbf{z}|\boldsymbol{\theta}) &= f(\mathbf{y}|\mathbf{z})\pi(\mathbf{z}|\boldsymbol{\theta}) \\ &= \exp - \left\{ \psi(\boldsymbol{\theta}) + \left\langle S(\mathbf{y},\mathbf{z}); \phi(\boldsymbol{\theta}) \right\rangle \right\}, \end{aligned} \tag{E.1}$$

where $\langle \cdot;\cdot \rangle$ denotes the scalar product. Since we consider MAP estimation here, the above likelihood must be mutiplied by the prior density. Due to its conditionally conjugate form, its expression is very close to that of the likelihood:

$$\pi(\boldsymbol{\theta}) = \exp - \left\{ \psi_\pi(\boldsymbol{\theta}) + \left\langle s_\pi; \phi(\boldsymbol{\theta}) \right\rangle \right\}. \tag{E.2}$$

Second, it is assumed that the hidden variables $\mathbf{z}$ can be simulated using a transition kernel $\pi_{\boldsymbol{\theta}}(\cdot|\cdot)$ whose stationary distribution is the conditional density $\pi(\mathbf{z}|\mathbf{y},\boldsymbol{\theta})$, for any value of $\boldsymbol{\theta}$. Then the k-th iteration of the MCMC-SAEM algorithm contains the following steps:

- *Simulation.* Draw a new value $\mathbf{z}_k$ from the current one $\mathbf{z}_{k-1}$, according to: $\mathbf{z}_k \sim \pi_{\boldsymbol{\theta}_{k-1}}(\mathbf{z}_k|\mathbf{z}_{k-1})$;
- *Stochastic Averaging.* Update the expected value s^{k-1} of the sufficient statistics by computing the weighted sum: $s^k = s^{k-1} + c_k\left(S(\mathbf{y},\mathbf{z}_k) - s^{k-1}\right)$;
- *Maximization.* Update the parameter value $\boldsymbol{\theta}_{k-1}$ by computing: $\boldsymbol{\theta}_k = \arg\min_{\boldsymbol{\theta}}\left\{\psi(\boldsymbol{\theta}) + \psi_\pi(\boldsymbol{\theta}) + \langle s^k + s_\pi; \phi(\boldsymbol{\theta})\rangle\right\}$.

The sequence of positive coefficients $(c_k)_{k\geq 0}$ must verify: $\sum_k c_k = \infty$ but $\sum_k c_k^2 < \infty$. Then, under certain regularity conditions concerning the functions $\phi(\cdot), S(\cdot,\cdot)$, and $\psi(\cdot)$, the sequence $(\boldsymbol{\theta}_k)_{k\geq 0}$ converges to a local maximum of the posterior density. The rationale underlying the SAEM algoritm is that the stochastic averaging step provides a Monte-Carlo estimate of the expected complete likelihood $\mathbb{E}[f(\mathbf{y},\mathbf{z},\boldsymbol{\theta})|\mathbf{y},\boldsymbol{\theta}_{k-1}]$ when it cannot be computed analytically, as in the E-step of a standard EM algorithm.

In our case, the likelihood and prior can be factorized across regions, except for the part depending on the spatial parameter σ_S^2, so the log-product of the right members in (E.1) and (E.2) can conveniently be written as

$$\begin{aligned} -\log f(\mathbf{y},\mathbf{z},\boldsymbol{\theta}) = C(\mathbf{y},\mathbf{z}) + \psi_S(\sigma_S^2) + S_S(\mathbf{w})\phi_S(\sigma_S^2) \\ + \textstyle\sum_{j=1}^N \left[\psi_j(\boldsymbol{\theta}_j) + \left\langle S_j(\mathbf{z}); \phi_j(\boldsymbol{\theta}_j)\right\rangle\right], \end{aligned} \tag{E.3}$$

where the term $C(\mathbf{y},\mathbf{z})$ is independent from $\boldsymbol{\theta}$ and therefore irrelevant for the maximization,

$$\psi_S(\sigma_S^2) = \left(\alpha + 1 + \tfrac{3nB}{2}\right)\log\sigma_S^2; \quad S_S(\mathbf{w}) = \beta + \frac{\sum_{i,b}\|\mathbf{w}_{i,b}\|^2}{2}; \quad \phi_S(\sigma_S^2) = \sigma_S^{-2}; \tag{E.4}$$

and for all $j = 1, \ldots, N$:

$$\psi_j(\boldsymbol{\theta}_j) = (\alpha+1)\log\sigma_j^2 + \left(\alpha + 1 + \frac{d_j+\gamma_j}{2}\right)\log\nu_j^2 + \gamma_j\frac{(d_j+\lambda_j)\eta_j^2}{2\nu_j^2}; \quad \text{(E.5)}$$

$$S_j(\mathbf{z}) = \frac{1}{2}\left(\sum_{\ell_k=j} n_k; \sum_{\ell_k=j}\sum_{\varphi_i(l)=k}(x_{i,l}-\mu_k)^2;\ 2\beta + \sum_{\ell_k=j}\mu_k^2;\ 2\gamma_j\sum_{\ell_k=j}\mu_k\right) \quad \text{(E.6)}$$

$$\phi_j(\boldsymbol{\theta}_j) = \left(\log\sigma_j^2;\ \sigma_j^{-2};\ \nu_j^2;\ -\gamma_j\frac{\eta_j}{\nu_j^2}\right). \quad \text{(E.7)}$$

Thus, the MCMC-SAEM algorithm starts with initial values $\mathbf{z}_0, \boldsymbol{\theta}_0, s_S^0, (s_j^0)_{1\leq j\leq N}$ and iterates the following steps for $k = 1, \ldots, K$:

- *Simulation.* Update the hidden variables $\mathbf{z}_k = (\mathbf{x}_k, \mathbf{w}_k, \boldsymbol{\mu}_k)$ by simulating them conditionnally on the current values $(\mathbf{z}_{k-1}, \boldsymbol{\theta}_{k-1})$, as described in Appendix D.

- *Stochastic Averaging.* Update the expected values $s_S^{k-1}, (s_j^{k-1})_{1\leq j\leq N}$ of the sufficient statistics by computing $s_S^k = s_S^{k-1} + c_k\left(S_S(\mathbf{w}_k) - s_S^{k-1}\right)$, and for all j : $s_j^k = s_j^{k-1} + c_k\left(S_j(\mathbf{z}_k) - s_j^{k-1}\right)$;

- *Maximization.* Update the parameter value $\boldsymbol{\theta}_{k-1}$ according to:

$$\sigma_{S,k}^2 = \frac{s_S^k}{\alpha + 1 + \frac{3nB}{2}} \quad \text{(E.8)}$$

$$\sigma_{j,k}^2 = \frac{s_{j,2}^k}{\alpha + 1 + s_{j,1}^k} \quad \text{(E.9)}$$

$$\nu_{j,k}^2 = \frac{s_{j,3}^k - \frac{1}{2}\left(s_{j,4}^k\right)^2/(d_j+\lambda_j)}{\alpha + 1 + (d_j+\gamma_j)/2} \quad \text{(E.10)}$$

$$\eta_{j,k} = \frac{s_{j,4}^k}{d_j+\lambda_j}. \quad \text{(E.11)}$$

To avoid getting trapped in a local maximum far from the global maximum, the stochastic averaging coefficients are set to $c_k = 1$ for iterations $k = 1, \ldots, K_0$, corresponding to a 'burn-in' period, during which the algorithm explores the parameter space. Subsequently, convergence to a local maximum is obtained by choosing $c_{K_0+k} = \frac{1}{k}$ for $k = 1, \ldots, K - K_0$.

Appendix F

Likelihood expression conditional on the displacements

Having estimated $\hat{\boldsymbol{\theta}}$ using the SAEM algorithm above, we now address the computation of the likelihood for region j, conditional on the elementary displacements $\mathbf{w}$ and on the indicator variable. This will be used to compute the marginal likelihood approximation given by (5.27). The choice of a particular value $\mathbf{w}^\star$ is addressed in the forthcoming Appendix G, as it uses the expression for the conditional likelihood derived here.

From (5.5), it follows that the observations can be separated in conditionally independent blocks, depending on the voxel k they are displaced to. Thus, noting $\tilde{\mathbf{y}}_k = (y_{il}, \phi_i(l) = k)$ the vector of observations displaced to voxel k, sorted *e.g.* in lexicographic order, we have:

$$f(\mathbf{y}_{\mathcal{V}_j}|\mathbf{w}, \hat{\boldsymbol{\theta}}_j) \quad = \quad \prod_{\ell_k = j} f(\tilde{\mathbf{y}}_k|\mathbf{w}, \hat{\boldsymbol{\theta}}_j),$$

where

$$\tilde{\mathbf{y}}_k|\mathbf{w}, \hat{\boldsymbol{\theta}}_j \quad \sim \quad \mathcal{N}(\eta_j \mathbf{1}_{n_k}, \boldsymbol{\Sigma}_k),$$

and the covariance matrix $\boldsymbol{\Sigma}_k$ is equal to:

$$\boldsymbol{\Sigma}_k \quad = \quad \nu_j^2 \mathbf{1}_{nk}' \mathbf{1}_{nk} + \sigma_j^2 \mathbf{I}_{n_k} + \mathrm{diag}(s_{il}^2)_{i,l;\,\phi_i(l)=k}.$$

Consequently, the log-conditional likelihood boils down to the sum for all voxels in region j of the explicit quantities:

$$\begin{aligned}&\log f(\tilde{\mathbf{y}}_k|\mathbf{w},\hat{\boldsymbol{\theta}}_j)\\&= -\frac{nk}{2}\log(2\pi) - \frac{1}{2}\log|\boldsymbol{\Sigma}_k| - \frac{1}{2}(\tilde{\mathbf{y}}_k - \eta_j\mathbf{1}_{n_k})'\boldsymbol{\Sigma}_k^{-1}(\tilde{\mathbf{y}}_k - \eta_j\mathbf{1}_{n_k}).\end{aligned} \tag{F.1}$$

The determinant and inverse of $\boldsymbol{\Sigma}_k$ are easily computed thanks to the matrix determinant lemma, and the Shermann-Morrison lemma (see [Golub and Van Loan, 1996] for instance):

$$|\mathbf{A} + \mathbf{u}\mathbf{u}'| = (1 + \mathbf{u}'\mathbf{A}^{-1}\mathbf{u})|\mathbf{A}|; \tag{F.2}$$

$$(\mathbf{A} + \mathbf{u}\mathbf{u}')^{-1} = \mathbf{A}^{-1} - \frac{\mathbf{A}^{-1}\mathbf{u}\mathbf{u}'\mathbf{A}^{-1}}{1 + \mathbf{u}'\mathbf{A}^{-1}\mathbf{u}}, \tag{F.3}$$

valid for any invertible $n \times n$ matrix $\mathbf{A}$ and any $n \times 1$ vector $\mathbf{u}$, such that $\mathbf{u}'\mathbf{A}^{-1}\mathbf{u} > -1$.

We apply these results to $\mathbf{A} = \sigma_j^2\mathbf{I}_{n_k} + \text{diag}(s_{i,l}^2)_{i,l;\,\phi_i(l)=k}$ and $\mathbf{u} = \nu_j\mathbf{1}_{nk}$. Using (F.2), we obtain:

$$|\boldsymbol{\Sigma}_k| = \left(1 + \nu_j^2 \sum_{i,l;\,\phi_i(l)=k} \frac{1}{\sigma_j^2 + s_{il}^2}\right) \prod_{i,l;\,\phi_i(l)=k} (\sigma_j^2 + s_{il}^2). \tag{F.4}$$

Next, defining $\mathbf{X} = \tilde{\mathbf{y}}_k - \eta_j\mathbf{1}_{n_k}$, we apply (F.3), yielding:

$$\begin{aligned}\mathbf{X}'(\mathbf{A} + \mathbf{u}\mathbf{u}')^{-1}\mathbf{X} &= \mathbf{X}'\left(\mathbf{A}^{-1} - \frac{\mathbf{A}^{-1}\mathbf{u}\mathbf{u}'\mathbf{A}^{-1}}{1 + \mathbf{u}'\mathbf{A}^{-1}\mathbf{u}}\right)\mathbf{X}\\&= \mathbf{X}'\mathbf{A}^{-1}\mathbf{X} - \frac{(\mathbf{X}'\mathbf{A}^{-1}\mathbf{u})^2}{1 + \mathbf{u}'\mathbf{A}^{-1}\mathbf{u}},\end{aligned}$$

so that

$$(\tilde{\mathbf{y}}_k - \eta_j\mathbf{1}_{n_k})'\boldsymbol{\Sigma}_k^{-1}(\tilde{\mathbf{y}}_k - \eta_j\mathbf{1}_{n_k}) = \sum_{i,l;\,\phi_i(l)=k} \frac{(y_{il} - \eta_j)^2}{\sigma_j^2 + s_{il}^2} - \frac{\left(\sum_{i,l;\,\phi_i(l)=k} \frac{y_{il} - \eta_j}{\sigma_j^2 + s_{il}^2}\right)^2}{\nu_j^{-2} + \sum_{i,l;\,\phi_i(l)=k} \frac{1}{\sigma_j^2 + s_{il}^2}}. \tag{F.5}$$

Hence the exact value of $\log f(\tilde{\mathbf{y}}_k|\mathbf{w}, \hat{\boldsymbol{\theta}}_j)$ can be computed following (F.1), (F.4) and (F.5).

Appendix G

Most probable displacement field *a posteriori* by simulated annealing

As explained in Section 5.5, the marginal likelihood is approximated by conditioning on a certain value of the elementary displacements $\mathbf{w}$, defined as the most probable value given the data and the estimated parameter $\hat{\theta}$:

$$\hat{\mathbf{w}} \quad = \quad \arg\max_{\mathbf{w}} \pi(\mathbf{w}|\mathbf{y}, \hat{\boldsymbol{\theta}}). \tag{G.1}$$

The conditional posterior density is proportional to:

$$\pi(\mathbf{w}|\mathbf{y}, \hat{\boldsymbol{\theta}}) \quad \propto \quad f(\mathbf{y}|\mathbf{w}, \hat{\boldsymbol{\theta}})\pi(\mathbf{w}|\sigma_S^2), \tag{G.2}$$

given the *a priori* independence of $\mathbf{w}$ from all other variables conditional on σ_S^2. Though the right member of (G.2) can be computed explicitely, using the formulas derived in Section F, its maximization with respect to $\mathbf{w}$ is difficult, because of the high number of dimensions involved, and of the complexity of the objective function. In particular, it is neither differentiable nor convex, thus prohibiting the use of standard gradient methods. Instead, we use the Simulated Annealing (SA) algorithm [Kirkpatrick et al., 1983], as described below, which is well-adapted to this setting.

Given the objective function $\pi(\mathbf{w}|\mathbf{y}, \hat{\boldsymbol{\theta}})$ we wish to maximize, we define for all $\alpha > 0$ the modified density $\pi_\alpha(\mathbf{w}) \propto \pi^\alpha(\mathbf{w}|\mathbf{y}, \hat{\boldsymbol{\theta}})$. These densities share the same modes, but differ in the contrast of their landscape, which increases with α. At the limit $\alpha \to \infty$, π_α converges weakly to a combination of Dirac mass in its modes.

The SA algorithm works by simulating successive values $(\mathbf{w}^t)_{t\geq 1}$ from transition kernels $K_{\alpha_t}(\cdot, \cdot)$ with stationary distributions π_{α_t}, where α_t increases progressively, so that the resulting Markov chain at first explores many possible states, and then gets attracted with more and more strength toward a mode of π_{α_t}. This heuristic can be straightened by showing that for any *cooling schedule* $(\alpha_t)_{t\geq 1}$ such that $\alpha_t \to \infty$, then $\mathbf{w}^t$ converges almost surely to the global maximum $\pi(\mathbf{w}|\mathbf{y}, \hat{\boldsymbol{\theta}})$, as $t \to \infty$ [Granville et al., 1994].

As in Appendix D, each elementary displacement $\mathbf{w}_{i,b}$ is updated in turn under the target density π_α, using a M-H step, by simulating a candidate $\mathbf{w}'_{i,b}$ from the random-walk proposal

$$q_\alpha(\mathbf{w}'_{i,b}|\mathbf{w}_{i,b}) \quad = \quad \mathcal{N}\left(\mathbf{w}'_{i,b}; \mathbf{w}_{i,b}, \alpha^{-1}\sigma^2_{RW}\mathbf{I}_3\right). \tag{G.3}$$

The proposal is then accepted with probability

$$\begin{aligned} \tilde{\alpha} &= \min\left\{1, \frac{\pi_\alpha(\mathbf{w}')}{\pi_\alpha(\mathbf{w})}\right\} \\ &= \left\{1, \frac{f^\alpha(\mathbf{y}|\mathbf{w}', \hat{\boldsymbol{\theta}})}{f^\alpha(\mathbf{y}|\mathbf{w}, \hat{\boldsymbol{\theta}})} \frac{\mathcal{N}(\mathbf{w}'_{i,b}; \mathbf{0}, \sigma_S^2/\alpha\mathbf{I}_3)}{\mathcal{N}(\mathbf{w}_{i,b}; \mathbf{0}, \sigma_S^2/\alpha\mathbf{I}_3)}\right\}, \end{aligned}$$

where $\mathbf{w}'$ is obtained from the current value $\mathbf{w}$, replacing $\mathbf{w}_{i,b}$ by $\mathbf{w}'_{i,b}$, and $f^\alpha(\mathbf{y}|\mathbf{w}', \hat{\boldsymbol{\theta}})$ is computed from (F.1), (F.4) and (F.5) in Section F.

We used the cooling schedule: $\alpha_t = \tau^{-t}$, where $\tau \in (0, 1)$ is the cooling rate. In practice, we found that setting $\tau = 99\%$ and letting the algorithm run for $T = 100$ iterations gave satisfying results.

Appendix H

Likelihood under spatial uncertainty, by Chib's method

Following the notations introduced in Section 5.5.2, we start by showing how to compute the reduced posterior ordinates $\pi(\mathbf{w}^*_{ib}|\mathbf{w}^*_{-ib}, \boldsymbol{\theta}^*, \mathbf{y})$, for $i = 1, \ldots, n$ and $b = 1, \ldots, B$, following [Chib and Jeliazkov, 2001]. First, note that the full conditional density of $\mathbf{w}^*_{ib}$ is given by:

$$\pi(\mathbf{w}^*_{ib}|\mathbf{w}^*_{-ib}, \mathbf{w}^{+ib}, \boldsymbol{\theta}^*, \mathbf{y}) \propto f(\mathbf{y}|\mathbf{w}^*_{ib}, \mathbf{w}^*_{-ib}, \mathbf{w}^{+ib}, \boldsymbol{\theta}^*)\pi(\mathbf{w}^*_{ib}, \mathbf{w}^*_{-ib}, \mathbf{w}^{+ib}|\boldsymbol{\theta}^*),$$

where $\mathbf{w}^{+ib}$ denotes the collection of blocks outside $\mathbf{w}_{-ib}$ and distinct from $\mathbf{w}_{ib}$, that is, of blocks $\mathbf{w}_{i'b'}$ for $i' \geq i$, $b' \geq b$, and $(i', b') \neq (i, b)$. As described in Appendix G, This conditional density can be sampled by the MH algorithm, with proposal

$$q(\mathbf{w}_{ib}, \mathbf{w}'_{ib}|\mathbf{w}^*_{-ib}, \mathbf{w}^{+ib}, \boldsymbol{\theta}^*, \mathbf{y}),$$

and acceptance rate:

$$\alpha(\mathbf{w}_{ib}, \mathbf{w}'_{ib}|\mathbf{w}^*_{-ib}, \mathbf{w}^{+ib}, \boldsymbol{\theta}^*, \mathbf{y})$$

$$= \min\left\{1, \frac{\pi(\mathbf{w}'_{ib}|\mathbf{w}^*_{-ib}, \mathbf{w}^{+ib}, \boldsymbol{\theta}^*, \mathbf{y})}{\pi(\mathbf{w}_{ib}|\mathbf{w}^*_{-ib}, \mathbf{w}^{+ib}, \boldsymbol{\theta}^*, \mathbf{y})} \times \frac{q(\mathbf{w}'_{ib}, \mathbf{w}_{ib}|\mathbf{w}^*_{-ib}, \mathbf{w}^{+ib}, \boldsymbol{\theta}^*, \mathbf{y})}{q(\mathbf{w}_{ib}, \mathbf{w}'_{ib}|\mathbf{w}^*_{-ib}, \mathbf{w}^{+ib}, \boldsymbol{\theta}^*, \mathbf{y})}\right\}$$

(in fact, given that the proposal density q is symmetric, the ratio to the right in the above display simplifies to 1).

It can be verified by direct calculation that the transition kernel

$$\pi(\mathbf{w}^*_{ib}, \mathbf{w}_{ib}|\mathbf{w}^*_{-ib}, \mathbf{w}^{+ib}, \boldsymbol{\theta}^*, \mathbf{y})$$

$$= q(\mathbf{w}^*_{ib}, \mathbf{w}_{ib}|\mathbf{w}^*_{-ib}, \mathbf{w}^{+ib}, \boldsymbol{\theta}^*, \mathbf{y})\alpha(\mathbf{w}^*_{ib}, \mathbf{w}_{ib}|\mathbf{w}^*_{-ib}, \mathbf{w}^{+ib}, \boldsymbol{\theta}^*, \mathbf{y})$$

verifies the *local reversibility condition*:

$$\begin{aligned}&\pi(\mathbf{w}^*_{ib}, \mathbf{w}_{ib}|\mathbf{w}^*_{-ib}, \mathbf{w}^{+ib}, \boldsymbol{\theta}^*, \mathbf{y})\pi(\mathbf{w}^*_{ib}|\mathbf{w}^*_{-ib}, \mathbf{w}^{+ib}, \boldsymbol{\theta}^*, \mathbf{y})\\ &= \pi(\mathbf{w}_{ib}, \mathbf{w}^*_{ib}|\mathbf{w}^*_{-ib}, \mathbf{w}^{+ib}, \boldsymbol{\theta}^*, \mathbf{y})\pi(\mathbf{w}_{ib}|\mathbf{w}^*_{-ib}, \mathbf{w}^{+ib}, \boldsymbol{\theta}^*, \mathbf{y}).\end{aligned} \tag{H.1}$$

Multiply both sides of this equation by $\pi(\mathbf{w}^{+ib}|\mathbf{w}^*_{-ib}, \boldsymbol{\theta}^*, \mathbf{y})$, and integrate over both $\mathbf{w}^{+ib}$ and $\mathbf{w}_{ib}$, to obtain:

$$\int \pi(\mathbf{w}^*_{ib}, \mathbf{w}_{ib}|\mathbf{w}^*_{-ib}, \mathbf{w}^{+ib}, \boldsymbol{\theta}^*, \mathbf{y})\pi(\mathbf{w}^*_{ib}, \mathbf{w}^{+ib}|\mathbf{w}^*_{-ib}, \boldsymbol{\theta}^*, \mathbf{y})\mathbf{dw}^{+ib}\mathbf{dw}_{ib}$$

$$= \int \pi(\mathbf{w}_{ib}, \mathbf{w}^*_{ib}|\mathbf{w}^*_{-ib}, \mathbf{w}^{+ib}, \boldsymbol{\theta}^*, \mathbf{y})\pi(\mathbf{w}_{ib}, \mathbf{w}^{+ib}|\mathbf{w}^*_{-ib}, \boldsymbol{\theta}^*, \mathbf{y})\mathbf{dw}^{+ib}\mathbf{dw}_{ib}.$$

Next, express $\pi(\mathbf{w}^*_{ib}, \mathbf{w}^{+ib}|\mathbf{w}^*_{-ib}, \boldsymbol{\theta}^*, \mathbf{y})$ as:

$\pi(\mathbf{w}^*_{ib}|\mathbf{w}^*_{-ib}, \boldsymbol{\theta}^*, \mathbf{y})\pi(\mathbf{w}^{+ib}|\mathbf{w}^*_{-ib}, \mathbf{w}^*_{ib}, \boldsymbol{\theta}^*, \mathbf{y})$, and apply (H.1), yielding:

$$\pi(\mathbf{w}^*_{ib}|\mathbf{w}^*_{-ib}, \boldsymbol{\theta}^*, \mathbf{y}) = \frac{\mathbb{E}_1[q(\mathbf{w}_{ib}, \mathbf{w}^*_{ib}|\mathbf{w}^*_{-ib}, \mathbf{w}^{+ib}, \boldsymbol{\theta}^*, \mathbf{y})\alpha(\mathbf{w}_{ib}, \mathbf{w}^*_{ib}|\mathbf{w}^*_{-ib}, \mathbf{w}^{+ib}, \boldsymbol{\theta}^*, \mathbf{y})]}{\mathbb{E}_2[\alpha(\mathbf{w}^*_{ib}, \mathbf{w}_{ib}|\mathbf{w}^*_{-ib}, \mathbf{w}^{+ib}, \boldsymbol{\theta}^*, \mathbf{y})]}, \tag{H.2}$$

where $\mathbb{E}_1$ is the expectation with respect to the conditional posterior $\pi(\mathbf{w}_{ib}, \mathbf{w}^{+ib}|\mathbf{w}^*_{-ib}, \boldsymbol{\theta}^*, \mathbf{y})$, and $\mathbb{E}_2$ with respect to $q(\mathbf{w}^*_{ib}, \mathbf{w}_{ib}|\mathbf{w}^*_{-ib}, \mathbf{w}^{+ib}, \boldsymbol{\theta}^*, \mathbf{y})\pi(\mathbf{w}^{+ib}|\mathbf{w}^*_{-ib}, \mathbf{w}^*_{ib}, \boldsymbol{\theta}^*, \mathbf{y})$.

Both integrals can be estimated from the output of reduced multiple-block MH runs, as follows:

1. To estimate the denominator, sample the conditional posterior: $\pi(\mathbf{w}_{ib}, \mathbf{w}^{+ib}|\mathbf{w}^*_{-ib}, \boldsymbol{\theta}^*, \mathbf{y})$. Let $(\mathbf{w}^{(g)}_{ib}, \mathbf{w}^{+ib,(g)})_{1\leq g\leq G}$ stand for the generated sample.

2. Next, include $\mathbf{w}^*_{-ib}$ in the conditioning set, and generate a sample $(\mathbf{w}^{+ib,(j)})_{1\leq j\leq J}$ from the reduced conditional posterior $\pi(\mathbf{w}^{+ib}|\mathbf{w}^*_{-ib}, \mathbf{w}^*_{ib}, \boldsymbol{\theta}^*, \mathbf{y})$. For each j, draw $\mathbf{w}^{(j)}_{ib}$ from the proposal density $q(\mathbf{w}^*_{ib}, \mathbf{w}^{(j)}_{ib}|\mathbf{w}^*_{-ib}, \mathbf{w}^{+\boldsymbol{ib}(j)}, \boldsymbol{\theta}^*, \mathbf{y})$. Then, form the sample $(\mathbf{w}^{(j)}_{ib}, \mathbf{w}^{+ib,(j)})_{1\leq j\leq J}$.

3. Estimate the conditional posterior density in (H.2) by:

$$\hat{\pi}(\mathbf{w}^*_{ib}|\mathbf{w}^*_{-ib}, \boldsymbol{\theta}^*, \mathbf{y})$$

$$=$$

$$\frac{G^{-1}\sum_{g=1}^{G} q(\mathbf{w}^{(g)}_{ib}, \mathbf{w}^*_{ib}|\mathbf{w}^*_{-ib}, \mathbf{w}^{+ib,(g)}, \boldsymbol{\theta}^*, \mathbf{y})\alpha(\mathbf{w}^{(g)}_{ib}, \mathbf{w}^*_{ib}|\mathbf{w}^*_{-ib}, \mathbf{w}^{+ib,(g)}, \boldsymbol{\theta}^*, \mathbf{y})}{J^{-1}\sum_{j=1}^{J} \alpha(\mathbf{w}^*_{ib}, \mathbf{w}^{(j)}_{ib}|\mathbf{w}^*_{-ib}, \mathbf{w}^{+ib,(j)}, \boldsymbol{\theta}^*, \mathbf{y})}.$$

Repeat steps **1** − **3** for $i = 1, \ldots, n$ and $b = 1, \ldots, B$. Finally, estimate the likelihood in the log scale as:

$$\log \hat{f}(\mathbf{y}|\boldsymbol{\theta}^*, \boldsymbol{\gamma}) = \log f(\mathbf{y}|\mathbf{w}^*, \boldsymbol{\theta}^*, \boldsymbol{\gamma}) + \log \pi(\mathbf{w}^*|\boldsymbol{\theta}^*) - \sum_{i=1}^{n}\sum_{b=1}^{B} \log \hat{\pi}(\mathbf{w}^*_{ib}|\mathbf{w}^*_{-ib}, \boldsymbol{\theta}^*, \mathbf{y}).$$

It can be noted that the values $(\mathbf{w}^{+ib,(j)})_{1\leq j\leq J}$ generated in step **2** are also produced in step **1** of the next reduced run. This means that $n \times B$ reduced runs are necessary instead of $2n \times B$. In practical application, we found that the successive reduced runs of the multiple block MH algorithm required fine tuning; indeed, the variability of the sampled transition kernel values $\pi(\mathbf{w}^{(g)}_{ib}, \mathbf{w}^*_{ib}|\mathbf{w}^*_{-ib}, \mathbf{w}^{+ib,(g)}, \boldsymbol{\theta}^*, \mathbf{y})$ increased with the number of hidden variables included in the conditioning set $\boldsymbol{\theta}^*$. A possible explanation to this is that the modes of the reduced conditional distributions were more and more distant from the mode $\mathbf{w}^*$ of the full conditional, where the kernel was evaluated. We found

empirically that using a number of iterations for each run inversely proportional to the number of sampled blocks worked best.

Bibliography

[Allassonnière et al., 2007] Allassonnière, S., Amit, Y., and A., T. (2007). Towards a coherent statistical framework for dense deformable template estimation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(1):3–29.

[Allassonnière et al., 2008] Allassonnière, S., E., K., and A., T. (2008). Map estimation of statistical deformable templates via nonlinear mixed effects models: Deterministic and stochastic approaches. In *2nd Workshop on Mathematical Foundations of Computational Anatomy*, New York, USA. MICCAI.

[Anderson and Legendre, 1999] Anderson, M. J. and Legendre, P. (1999). An empirical comparison of permutation methods for tests of partial regression coefficients in a linear model. *Journal of statistical computing and simulation*, 62:271–303.

[Andrade et al., 2001] Andrade, A., Kherif, F., Mangin, J.-F., Worsley, K., Paradis, A.-L., Simon, O., Dehaene, S., and Poline, J.-B. (2001). Detection of fMRI activation using cortical surface mapping. *Human Brain Mapping*, 12:79–93.

[Ashburner and Friston, 1999] Ashburner, J. and Friston, K. (1999). Nonlinear spatial normalization using basis functions. *Human Brain Mapping*, 7(4):254–66.

[Beal, 2003] Beal, M. (2003). *Variational algorithms for approximate Bayesian inference*. PhD thesis, University College of London, London, United Kingdom.

[Beckmann et al., 2003a] Beckmann, C., Jenkinson, M., and Smith, S. (2003a). General multi-level linear modelling for group analysis in fMRI. *Neuroimage*, 20:1052–1063.

[Beckmann and Smith, 2004] Beckmann, C. and Smith, S. (2004). Probabilistic independent component analysis for functional magnetic resonance imaging. *IEEE Transactions on Medical Imaging*, 23(2):137–152.

[Beckmann et al., 2003b] Beckmann, C., Woolrich, M., and Smith, S. (2003b). Gaussian / gamma mixture modelling of ica/glm spatial maps. In *Ninth Int. Conf. on Functional Mapping of the Human Brain.*

[Benjamini and Hochberg, 1995] Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society*, 57(1):289–300.

[Bowman et al., 2008] Bowman, D. F., Caffo, B., Bassett, S. S., and Kilts, C. (2008). A bayesian hierarchical framework for spatial modeling of fmri data. *NeuroImage*, 39(1):146–156.

[Boynton et al., 1996] Boynton, G. M., Engel, S. A., Glover, G. H., and Heeger, D. J. (1996). Linear systems analysis of functional magnetic resonance imaging in human V1. *The Journal of Neuroscience*, 16:4207–4221.

[Brett et al., 2002] Brett, M., Johnsrude, I., and Owen, A. (2002). The problem of functional localization in the human brain. *Nature Reviews Neuroscience*, 3(3):243–249.

[Brown, 1992] Brown, L. G. (1992). A survey of image registration techniques. *ACM Comput. Surv.*, 24(4):325–376.

[Buxton and Frank, 1997] Buxton, R. and Frank, L. (1997). A model for the coupling between cerebral blood flow and oxygen metabolism during neural stimulation. *Journal of Cerebral Blood Flow Metabolism*, 17(1):64–72.

[Cade, 2005] Cade, B. S. (2005). Linear models: Permutation methods. In Everitt, B. S. and Howell, D. C., editors, *Encyclopedia of Statistics in Behavioral Science*, volume 2, pages 1049–1054. Wiley.

[Cavanna and Trimble, 2006] Cavanna, A. E. and Trimble, M. R. (2006). The precuneus: a review of its functional anatomy and behavioural correlates. *Brain*, 129(3):564–583.

[Chib, 1995] Chib, S. (1995). Marginal likelihood from the Gibbs output. *Journal of American Statistical Association*, 90:1313–1321.

[Chib and Jeliazkov, 2001] Chib, S. and Jeliazkov, I. (2001). Marginal likelihood from the Metropolis-Hastings output. *Journal of American Statistical Association*, 96(453):270–281.

[Chochon et al., 1999] Chochon, F., Cohen, L., Van De Moortele, P. F., and Dehaene, S. (1999). Differential contributions of the left and right inferior parietal lobules to number processing. *J. Cognitive Neuroscience*, 11(6):617–630.

[Coulon et al., 2001] Coulon, O., Mangin, J.-F., Poline, J.-B., Frouin, V., and Bloch, I. (2001). Group analysis of individual activation maps using 3d scale-space primal sketches and a markovian random field. In Moore, M., editor, *Spatial statistics. Methodological aspects and some applications*, volume 159 of *Lecture annotes in Statistics*, pages 201–211. Springer Verlag, New York, NY.

[Dehaene et al., 2003] Dehaene, S., Piazza, M., Pinel, P., and Cohen, L. (2003). Three parietal circuits for number processing. *Cognitive Neuropsychology*, 20:487–506.

[Dempster et al., 1977] Dempster, A., Laird, A., and Rubin, D. (1977). Maximum likelihood from incomplete data via the em algorithm (with discussion). *Journal of the Royal Statistical Society, Series B*, 39:1–38.

[Donnet, 2006] Donnet, S. (2006). *Inversion de Données IRMf. Estimation et Sélection de Modèles*. PhD thesis, Université de Paris–Sud, Orsay , France.

[Donnet et al., 2006] Donnet, S., Lavielle, M., and Poline, J.-B. (2006). Are fMRI event-related response constant in time? A model selection answer. *Neuroimage*, pages 1169–1176.

[Essen, 2005] Essen, D. C. V. (2005). A population-average, landmark- and surface-based (pals) atlas of human cerebral cortex. *Neuroimage*, 28(3):635–662.

[Fischl et al., 1999] Fischl, B., Sereno, M. I., Tootell, R. B., and Dale, A. M. (1999). High-resolution intersubject averaging and a coordinate system for the cortical surface. *Human Brain Mapping*, 8(4):272–284.

[Flandin, 2004] Flandin, G. (2004). *Utilisation d'informations géométriques pour l'analyse statistique des données d'IRM fonctionnelle*. PhD thesis, Université de Nice-Sophia Antipolis.

[Forbes and Fort, 2007] Forbes, F. and Fort, G. (2007). Combining Monte Carlo and mean-field like methods for inference in hidden Markov random fields. *IEEE Transactions on Image Processing*, 16(3):824–837.

[Friston et al., 1995] Friston, K., Ashburner, J., Frith, C., Poline, J.-B., Heather, J., and Frackowiak, R. (1995). Spatial registration and normalization of images. *Human Brain Mapping*, 3(3):165–189.

[Friston and Buechel, 2000] Friston, K. and Buechel, C. (2000). Attentional modulation of effective connectivity from V2 to V5/MT in humans. *Proceedings of the National Academy of Sciences of the United States of America*, 97(13):7591–7596.

[Friston et al., 2002a] Friston, K., Glaser, D. E., Henson, R. N. A., Kiebel, S., Phillips, C., and Ashburner, J. (2002a). Classical and Bayesian inference in neuroimaging: Applications. *Neuroimage*, 16(2):484–512.

[Friston et al., 2002b] Friston, K., Penny, W., Phillips, C., Kiebel, S., Hinton, G., and Ashburner, J. (2002b). Classical and bayesian inference in neuroimaging: Theory. *Neuroimage*, 16(2):465–483.

[Friston, 1997] Friston, K. J. (1997). *Human Brain Function*, chapter 2, pages 25–42. Academic Press.

[Friston et al., 2000] Friston, K. J., Mechelli, A., Turner, R., and Price, C. J. (2000). Nonlinear responses in fMRI: the balloon model, Volterra kernels, and other hemodynamics. *Neuroimage*, 12:466–477.

[Ge et al., 2003] Ge, Y., Dudoit, S., and Speed, T. (2003). Resampling-based multiple testing for microarray data analysis. Technical report, Department of Statistics, University of California, Berkeley 2. Division of Biostatistics, University of California, Berkeley 3.

[Gelfand and Smith, 1990] Gelfand, A. E. and Smith, A. F. M. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 85(410):398–409.

[Geman and Geman, 1984] Geman, S. and Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, PAMI-6(6):721–741.

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

[Golub and Van Loan, 1996] Golub, G. H. and Van Loan, C. F. (1996). *Matrix computations*. The Johns Hopkins University Press, Baltimore, Maryland, Third edition.

[Good, 2005] Good, P. (2005). *Permutation, Parametric, and Bootstrap Tests of Hypotheses*. Springer, 3rd edition edition.

[Granville et al., 1994] Granville, V., Krivánek, M., and Rasson, J.-P. (1994). Simulated annealing : a proof of convergence. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 16(6):652–656.

[Green, 1995] Green, P. (1995). Reversible jump markov chain monte carlo computation and bayesian model determination. *Biometrika*, 82:711–732.

[Grenander, 1993] Grenander, U. (1993). *General pattern theory - A mathematical study of regular structures*. Clarendon Press, Oxford.

[Gössl et al., 2001] Gössl, C., Auer, D. P., and Fahrmeir, L. (2001). Bayesian spatiotemporal modeling of the hemodynamic response function in BOLD fMRI. *Biometrics*, 57:554–562.

[Hajnal, 2001] Hajnal, J. V. (2001). *Medical Image Registration*. CRC.

[Harrison et al., 2008] Harrison, L. M., Penny, W., Daunizeau, J., and Friston, K. J. (2008). Diffusion-based spatial priors for functional magnetic resonance images. *Neuroimage*, 41(2):408–423.

[Hastings, 1970] Hastings, W. K. (1970). Monte carlo sampling methods using markov chains and their applications. *Biometrika*, 57(1):97–109.

[Hayasaka and Nichols, 2003] Hayasaka, S. and Nichols, T. (2003). Validating Cluster Size Inference: Random Field and Permutation Methods. *Neuroimage*, 20(4):2343–2356.

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

[Hellier et al., 2003] Hellier, P., Barillot, C., Corouge, I., Gibaud, B., Le Goualher., G., Collins, D. L., Evans, A., Malandain, G., Ayache, N., Christensen, G. E., and Johnson, H. J. (2003). Retrospective evaluation of intersubject brain registration. *IEEE Transactions on Medical Imaging*, 22(9):1120–1130.

[Hobert and Casella, 1996] Hobert, J. P. and Casella, G. (1996). The effect of improper priors on Gibbs sampling in hierarchical linear mixed models. *Journal of the American Statistical Association*, 91(436).

[Holmes et al., 1996] Holmes, A., Blair, R., Watson, J., and Ford, I. (1996). Nonparametric analysis of statistic images from functional mapping experiments. *Journal of Cerebral Blood Flow Metabolism*, 16:7–22.

[Jordan, 2003] Jordan, M. I. (2003). Graphical models. Technical report, Computer Science Division and Department of Statistics, University of California, Berkeley.

[Joshi et al., 2004] Joshi, S., Davis, B., Jomier, M., and Gerig, G. (2004). Unbiased diffeomorphic atlas construction for computational anatomy. *NeuroImage*, 23:S151–S160.

[Keller et al., 2009] Keller, M., Lavielle, M., Perrot, M., and Roche, A. (2009). Anatomically Informed Bayesian Model Selection for fMRI Group Data Analysis. In *12th International Conference on Medical Image Computing and Computer Assised Intervention*, London, U.K.

[Keller et al., 2007] Keller, M., Mériaux, S., Roche, A., Pinel, P., and Thirion, B. (2007). A mixed-effect statistic for two-sample group analysis in fmri. In *Human Brain Mapping*, Chicago, USA.

[Keller and Roche, 2008] Keller, M. and Roche, A. (2008). Increased sensitivity in fMRI group analysis using mixed-effect modeling. In *5th International Symposium on Biomedical Imaging*, pages 548–551, Paris, France.

[Keller et al., 2008] Keller, M., Roche, A., Tucholka, A., and B.Thirion (2008). Dealing with spatial normalization errors in fMRI group inference using hierarchical modeling. *Statistica Sinica*, 18(4):1357–1374.

[Kim et al., 2005] Kim, H. Y., Giacomantone, J., and Cho, Z. H. (2005). Robust anisotropic diffusion to produce enhanced statistical parametric map from noisy fMRI. *Computer Vision and Image Understanding*, 99(3):435–452.

[Kim and Smyth, 2006] Kim, S. and Smyth, P. (2006). Hierarchical Dirichlet processes with random effects. In *Advances in Neural Information Processing Systems*, Vancouver.

[Kirkpatrick et al., 1983] Kirkpatrick, S., Gellat, C. D., and Vecchi, M. P. (1983). Optimization by simulated annealing. *Science*, 220:671–680.

[Klein et al., 2009] Klein, A., Andersson, J., Ardekani, B. A., Ashburner, J., Avants, B., Chiang, M.-C., Christensen, G. E., Collins, L., Hellier, P., Song, J. H., Jenkinson, M., Lepage, C., Rueckert, D., Thompson, P., Vercauteren, T., Woods, R. P., Mann, J. J., and Parsey, R. V. (2009). Evaluation of 14 nonlinear deformation algorithms applied to human brain mri registration. *Neuroimage*.

[Kuhn and Lavielle, 2004] Kuhn, E. and Lavielle, M. (2004). Coupling a stochastic approximation of EM with a MCMC procedure. *ESAIM P&S*, 8:115–131.

[Lavielle and Ludeña, 2007] Lavielle, M. and Ludeña, C. (2007). Random threshold for linear model selection. *ESAIM: Probability and Statistics*.

[Lehmann, 1986] Lehmann, E. (1986). *Testing Statistical Hypotheses*. Springer texts in statistics. Springer.

[Leporé et al., 2008] Leporé, N., Brun, C., Chou, Y.-Y., Lee, A. D., Barysheva, M., de Zubicaray, G. I., Meredith, M., McMahon, K. L., Wright, M. J., Toga, A. W., and Thompson, P. M. (2008). Multi-atlas tensor-based morphometry and its application to a genetic study of 92 twins. In *Mathematical Foundations of Computational Anatomy workshop of MICCAI 2008 conference.*

[Maintz and Viergever, 1998] Maintz and Viergever, M. A. (1998). A survey of medical image registration. *Medical Image Analysis*, 2(1):1–36.

[Makni et al., 2008] Makni, S., Idier, J., Vincent, T., Thirion, B., Dehaene-Lambertz, G., and Ciuciu, P. (2008). A fully Bayesian approach to the parcel-based detection-estimation of brain activity in fMRI. *Neuroimage*, 41(3):941–969.

[Marin and Robert, 2007] Marin, J.-M. and Robert, C. P. (2007). *Bayesian Core: A Practical Approach to Computational Bayesian Statistics (Springer Texts in Statistics)*. Springer.

[Mcculloch and Searle, 2001] Mcculloch, C. E. and Searle, S. R. (2001). *Generalized, Linear, and Mixed Models (Wiley Series in Probability and Statistics)*. Wiley-Interscience, 1 edition.

[McGillem and Svedlow, 1976] McGillem, C. and Svedlow, M. (1976). Image registration error variance as a measure of overlay quality. *GeoEl*, 14(1):44–49.

[Meng and Wong, 1996] Meng, X. and Wong, W. (1996). Simulating ratios of normalizing constants via a simple identity: a theoretical exploration. *Statistica Sinica*, 6:831–860.

[Mériaux et al., 2006] Mériaux, S., Roche, A., Dehaene-Lambertz, G., Thirion, B., and Poline, J.-B. (2006). Combined permutation test and mixed-effect model for group average analysis in fMRI. *Human Brain Mapping*, 27(5):402–410.

[Natarajan and Kass, 2000] Natarajan, R. and Kass, R. E. (2000). Reference bayesian methods for generalized linear mixed models. *Journal of American Statistical Association*, 95(449):227–237.

[Natarajan and McCulloch, 1998] Natarajan, R. and McCulloch, C. E. (1998). Gibbs sampling with diffuse proper priors: A valid approach to data-driven inference? *Journal of Computational and Graphical Statistics*, 7(3):267–277.

[Nichols and Hayasaka, 2003] Nichols, T. and Hayasaka, S. (2003). Controlling the Familywise Error Rate in Functional Neuroimaging: A Comparative Review. *Statistical Methods in Medical Research*, 12(5):419–446.

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

[Ogawa et al., 1990] Ogawa, S., Lee, T., Kay, A., and Tank, D. (1990). Brain magnetic resonance imaging with contrast dependent on blood oxygenation. *Proceedings of the National Academy of Sciences of the United States of America*, 87(24):9868–9872.

[Operto et al., 2008a] Operto, G., Bulot, R., Anton, J.-L., and Coulon, O. (2008a). Projection of fmri data onto the cortical surface using anatomically-informed convolution kernels. *Neuroimage*, 39(1):127–135.

[Operto et al., 2008b] Operto, G., Clouchoux, C., Bulot, R., Anton, J.-L., and Coulon, O. (2008b). Surface-based structural group analysis of fmri data. In *MICCAI '08: Proceedings of the 11th international conference on Medical Image Computing and Computer-Assisted Intervention - Part I*, pages 959–966, Berlin, Heidelberg. Springer-Verlag.

[Pacifico et al., 2004] Pacifico, M. P., Genovese, C., Verdinelli, I., and Wasserman, L. (2004). False discovery control for random fields. *Journal of the American Statistical Association*, 99(468):1002–1014.

[Penny et al., 2007] Penny, W., Flandin, G., and Trujillo-Bareto, N. (2007). Bayesian Comparison of Spatially Regularised General Linear Models. *Human Brain Mapping*, 28(4):275–293.

[Penny and Friston, 2003] Penny, W. and Friston, K. (2003). Mixtures of general linear models for functional neuroimaging. *IEEE Transactions on Medical Imaging*, 22(4):504–514.

[Penny et al., 2003] Penny, W. D., Kiebel, S., and Friston, K. J. (2003). Variational Bayesian inference for fMRI time series. *Neuroimage*, 19(3):727–741.

[Perlbarg et al., 2008] Perlbarg, V., Marrelec, G., Doyon, J., Pélégrini-Issac, M., Lehéricy, S., and Benali, H. (2008). NEDICA: detection of group functional networks in fmri using spatial independent component analysis. In *5th International Symposium on Biomedical Imaging*, pages 1247–1250, Paris, France.

[Perrot et al., 2008] Perrot, M., Rivière, D., and Mangin, J.-F. (2008). Identifying cortical sulci from localizations, shape and local organization. In *5th International Symposium on Biomedical Imaging*, pages 420–423, Paris, France.

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

[Poldrack, 2007] Poldrack, R. (2007). Regions of interest analysis for fmri. *Social Cognitive and Affective Neuroscience (SCAN)*, 2:67–70.

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

[Price, 2000] Price, C. J. (2000). The anatomy of language: contributions from functional neuroimaging. *Journal of Anatomy*, 197(3):335–359.

[Raftery et al., 2007] Raftery, A. E., Newton, M. A., Satagopan, J. M., and Krivitsky, P. N. (2007). Estimating the integrated likelihood via posterior simulation using the harmonic mean identity. In Bernardo, J., Bayarri, M., Berger, O., David, A., Heckermann, D., Smith, A., and West, M., editors, *Bayesian statistics 8*, pages 1–45. Oxford University Press.

[Raichle, 1994] Raichle, M. (1994). La visualisation de la pensee. *Pour La Science*, 200:52–59.

[Robert, 2007] Robert, C. P. (2007). *The Bayesian Choice: From Decision-Theoretic Foundations to Computational Implementation (Springer Texts in Statistics)*. Springer Verlag, New York, 2nd ed. 2001. 2nd printing edition.

[Roberts et al., 1997] Roberts, G. O., Gelman, A., and Gilks, W. R. (1997). Weak convergence and optimal scaling of random walk metropolis algorithms. *Annals of Applied Probability*, 7:110–120.

[Roche et al., 2004] Roche, A., Lahaye, P.-J., and Poline, J.-B. (2004). Incremental activation detection in fMRI series using kalman filtering. In *Proceedings 2st International Symposium on Biomedical Imaging*, pages 376–379, Arlington, VA.

[Roche et al., 2007] Roche, A., Mériaux, S., Keller, M., and Thirion, B. (2007). Mixed-effects statistics for group analysis in fMRI: A nonparametric maximum likelihood approach. *Neuroimage*, 38:501–510.

[Sabuncu et al., 2008] Sabuncu, M. R., Balci, S. K., and Golland, P. (2008). Discovering modes of an image population through mixture modeling. *Medical image computing and computer-assisted intervention : MICCAI International Conference on Medical Image Computing and Computer-Assisted Intervention*, 11(Pt 2):381–389.

[Smith and Fahrmeir, 2007] Smith, D. and Fahrmeir, L. (2007). Spatial Bayesian variable selection with application to Functional Magnetic Resonance Imaging. *Journal of American Statistical Association*, 102(478):417–431.

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

[Spiegelhalter et al., 1996] Spiegelhalter, Thomas, Best, and Gilks (1996). BUGS 0.5: Bayesian Inference Using Gibbs Sampling - Manual. Technical report, MRC Biostatistics Unit, Cambridge.

[Strasser and Weber, 1999] Strasser, H. and Weber, C. (1999). The asymptotic theory of permutation statistics. *Mathematical Methods of Statistics*, 2:220–250.

[Subsol, 1995] Subsol, G. (1995). *Construction automatique d'atlas anatomiques morphométriques à partir d'images médicales tridimensionnelles*. PhD thesis, Ecole Centrale de Paris.

[Talairach and Tournoux, 1988] Talairach, J. and Tournoux, P. (1988). *Co–Planar Stereotaxic Atlas of the Human Brain. 3-Dimensional Proportional System : An Approach to Cerebral Imaging*. Thieme Medical Publishers, Inc., Georg Thieme Verlag, Stuttgart, New York.

[Taylor and Worsley, 2007] Taylor, J. E. and Worsley, K. J. (2007). Detecting sparse signals in random fields, with an application to brain mapping. *Journal of the American Statistical Association*, 102(479,):913–928.

[Taylor et al., 2007] Taylor, J. E., Worsley, K. J., and Gosselin, F. (2007). Maxima of discretely sampled random fields, with an application to 'bubbles'. *Biometrika*, 94(1):1–18.

[Thirion et al., 2006a] Thirion, B., Dodel, S., and Poline, J.-B. (2006a). Detection of signal synchronizations in resting-state fMRI datasets. *Neuroimage*, 29(1):321–327.

[Thirion et al., 2006b] Thirion, B., Duchesnay, E., Hubbard, E., Dubois, J., Poline, J.-B., and Dehaene, S. (2006b). Reading the brain visual system as an inverse problem. In *Proceedings 3th International Symposium on Biomedical Imaging*, Washington, USA.

[Thirion et al., 2006c] Thirion, B., Flandin, G., Pinel, P., Roche, A., Ciuciu, P., and Poline, J.-B. (2006c). Dealing with the shortcomings of spatial normalization: Multi-subject parcellation of fMRI datasets. *Human Brain Mapping*, 27(8):678–693.

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

[Thirion et al., 2007b] Thirion, B., Tucholka, A., Keller, M., Pinel, P., Roche, A., Mangin, J.-F., and Poline, J.-B. (2007b). High level group analysis of FMRI data based on Dirichlet process mixture models. In *International Conference on Information Processing in Medical Imaging*, volume 4584 of *LNCS*, pages 482–494.

[Tierney, 1994] Tierney, L. (1994). Markov Chains for Exploring Posterior Distributions. *The Annals of Statistics*, 22(4):1701–1728.

[Toga, 1999] Toga, A. W. (1999). *Brain Warping*. Academic Press.

[Tucholka et al., 2008a] Tucholka, A., Thirion, B., Perrot, M., Pinel, P., Mangin, J.-F., and Poline, J.-B. (2008a). Probabilistic anatomo-functional parcellation of the cortex: how many regions? In *11thProceedings MICCAI, LNCS Springer Verlag*, New-York, USA.

[Tucholka et al., 2008b] Tucholka, A., Thirion, B., Pinel, P., Poline, J.-B., and Mangin, J.-F. (2008b). Triangulating cortical functional networks with anatomical landmarks. In *5th International Symposium on Biomedical Imaging*, pages 612–615, Paris, France.

[Tzourio-Mazoyer et al., 2002] Tzourio-Mazoyer, N., Landeau, B., Papathanassiou, D., Crivello, F., Etard, O., Delcroix, N., Mazoyer, B., and Joliot, M. (2002). Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. *Neuroimage*, 15(1):273–89.

[Vaever Hartvig, 2000] Vaever Hartvig, N. (2000). *Parametric modelling of functional magnetic resonance imaging data*. PhD thesis, University of Aarhus, Aarhus, Danemark.

[Van den Elsen et al., 1993] Van den Elsen, P. A., Pol, E. D., and Viergever, M. A. (1993). Medical image matching: A review with classification. *IEEE Engineering on Medicine and Biology*, 12(1):26–38.

[van der Vaart, 2000] van der Vaart, A. W. (2000). *Asymptotic Statistics (Cambridge Series in Statistical and Probabilistic Mathematics)*. Cambridge University Press.

[Westfall and Young, 1993] Westfall, P. H. and Young, S. S. (1993). *Resampling-based multiple testing: Examples and methods for p-value adjustment*. John Wiley & Sons, New York.

[Woolrich, 2008] Woolrich, M. (2008). Robust group analysis using outlier inference. *NeuroImage*, 41(2):286–301.

[Woolrich and Behrens, 2006] Woolrich, M. and Behrens, T. (2006). Variational Bayes inference of spatial mixture models for segmentation. *IEEE Transactions on Medical Imaging*, 25(10):1380–1391.

[Woolrich et al., 2004a] Woolrich, M., Behrens, T., Beckmann, C., Jenkinson, M., and Smith, S. (2004a). Multi-level linear modelling for fMRI group analysis using Bayesian inference. *Neuroimage*, 21(4):1732–1747.

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

[Woolrich et al., 2004b] Woolrich, M., Jenkinson, M., Brady, J. M., and Smith, S. (2004b). Constrained linear basis set for HRF modelling using variational Bayes. *Neuroimage*, 21(4):1748–1761.

[Woolrich et al., 2001] Woolrich, M., Ripley, B., Brady, M., and Smith, S. (2001). Temporal autocorrelation in univariate linear modelling of fMRI data. *Neuroimage*, 14(6):1370–1386.

[Worsley, 1994] Worsley, K. (1994). Local maxima and the expected Euler characteristic of excursion sets of χ^2, f, and t fields. *Adv. Appl. Prob.*, 26:13–42.

[Worsley et al., 2002] Worsley, K., Liao, C., Aston, J., Petre, V., Duncan, G., Morales, F., and Evans, A. (2002). A general statistical analysis for fMRI data. *Neuroimage*, 15(1):1–15.

[Worsley, 2005] Worsley, K. J. (2005). An improved theoretical P value for SPMs based on discrete local maxima. *Neuroimage*, 28(4):1056–1062.

[Xu et al., 2009] Xu, L., Johnson, T., Nee, D., and Nichols, T. (2009). Modeling inter-subject variability in fmri activation location: A bayesian hierarchical spatial model. *Biometrics*.

[Zarahn et al., 1997] Zarahn, E., Aguirre, G. K., and D'Esposito, M. (1997). Empirical analysis of BOLD fMRI statistics. I. Spatially unsmoothed data collected under null-hypothesis conditions. *Neuroimage*, 5(3):179–197.

[Zitova and Flusser, 2003] Zitova, B. and Flusser, J. (2003). Image registration methods: a survey. *Image and Vision Computing*, 21(11):977–1000.