Abstract

Taskinen, Ilkka
Cluster priors in the Bayesian modelling of fMRI data
Jyväskylä: University of Jyväskylä, 2001, 105 p.
ISBN 951-39-1059-8
ISSN 1457-8905
Diss.

Functional magnetic resonance imaging (fMRI) is a scanning technique for revealing haemodynamic changes connected with brain processing on the neuronal level. In neuropsychology, fMRI has been used in designed experiments together with controlled stimulation. fMRI data are temporal series of digital images corrupted by spatio-temporally correlated physiological processes and scanner noise. The statistical challenge in analysing fMRI data is to localize stimulus-related brain activation and estimate its characteristics. In this thesis, the focus is on spatial aspects of activations. A Bayesian approach is proposed and an *a priori* model which describes the clustering of activations is suggested. The prior is used to control the spatial extent, coherence and locations of clusters. Marked Gibbs point processes have been used to construct the prior. The prior is designed so that expert knowledge on the neuronal processing of interest can be incorporated into statistical analysis. To model the conditional distribution of observations, given the activations, Gaussian conditional autoregressive processes have been applied. Using these processes, heteroskedasticity and spatial autocorrelation in noise is accounted for. Inference is based on Markov chain Monte Carlo (MCMC) simulations of the posterior distribution. A modified version of an existing general simulation method for Gibbs point processes is devised to sample the posterior. Real fMRI data are analysed and the influence of different amounts of prior information on the uncertainty in activations is illustrated. An example of analysing synthetic data is provided to compare the new method with conventional nonparametric techniques. The conclusion is that, by adopting a structural approach, relevant features of activations can be accounted for leading to a potentially more efficient inference.

Key words: functional magnetic resonance imaging, marked point process, human brain, Bayesian modelling, Markov chain Monte Carlo, image analysis

Acknowledgements

My supervisors, Professor Antti Penttinen, Center for Mathematical and Computational Modeling (CMCM), and Professor Heikki Lyytinen, Department of Psychology, introduced me to the field of functional brain imaging. I want to thank them for their support during my work. I am also grateful to Ph.D. Niels Hartvig with whom I have had inspiring and detailed discussions on many topics of statistics, mathematics and functional imaging. My warm thanks are also due to Jennifer Hiemenz for examining numerous anatomical brain images and preparing anatomical data for further analysis. I appreciate the assistance of Jarmo Toivanen, the hospital physicist of Central Hospital of Central Finland, whose help was valuable in designing fMRI experiments and in transferring the data for my use. I also appreciate the financial support from the Academy of Finland at an early stage of this work ("Large structural statistical models and their estimation algorithms in the analysis of life history and spatial data", no. 3507). Mrs. Tuula Blåfield deserves special thanks for her linguistic comments, which have improved the final text. Thanks are also due to examiners Dr. Marie-Colette van Lieshout and Dr. Risto Ilmoniemi for their comments and suggestions for the final version. Finally, I want to thank all the members of the COMAS organization of our university and the Department of Mathematics and Statistics for providing me a pleasant atmosphere for working.

Jyväskylä, November 2001

Ilkka Taskinen

Contents

List of main notation

Spaces, sets and constants

$\mathcal{S}$	set of brain voxels	T	number of scans
$\mathcal{U}$	voxellated Euclidean space	$\boldsymbol{\lambda}$	contrast vector
$\mathcal{M}$	product space $\mathcal{S} \times \mathbb{R}$	$\tilde{\boldsymbol{\lambda}}$	extended contrast vector
$\mathcal{D}$	space of daughter configurations	β_d	prior mean of daughters
$\Omega(\cdot)$	exponential space of a set	ρ	spatial interaction radius

Configurations and parameters

η	cluster height	w	centre of a cluster
κ	unscaled cluster function	$\boldsymbol{d}$	daughter configuration
ζ	daughter height	$\boldsymbol{w}$	configuration of centres
v	relative location of a daughter	$\boldsymbol{x}$	configuration of clusters

Series, fields, functions and matrices

$Y = (Y_t)$	time series of images	Ψ	interaction function
$Y = (Y(s))$	voxel time series	$\boldsymbol{\delta} = (\delta_t(s))$	noise field
$D = (D_t)$	design series	$\boldsymbol{Z} = (Z(s))$	compressed data
$X = (X_t)$	haemodynamic response	W	spatial covariances
$\boldsymbol{X}$	matrix of explanatory series	V	temporal correlations
$\boldsymbol{I} = (I(s))$	indicator field in $\mathcal{S}$	$\boldsymbol{\epsilon} = (\epsilon(s))$	spatial noise field
$\phi(s)$	mean signal level at voxel s	$\boldsymbol{\varrho} = (\varrho(s))$	precision parameters
$\beta(s)$	stimulation effect at voxel s	$\boldsymbol{\tau} = (\tau_l)$	interaction parameter
$\boldsymbol{\alpha} = (\alpha(s))$	activation profile	$L(\boldsymbol{x})$	likelihood function
$\boldsymbol{I_\alpha}$	indicator field of $\boldsymbol{\alpha}$	$\tilde{L}(\boldsymbol{x})$	modified likelihood
$B_p(s)$	parent bell function	$\boldsymbol{\Phi}$	smoothing potential
$B_d(s)$	daughter bell function		

Densities and measures

ν	counting measure in $\mathcal{U}$	h	prior of a random centre
m_1	Lebesgue measure	h^*, ψ^*	corrected densities
h_ζ	prior of daughter height	$f_{\boldsymbol{w}}$	prior of cluster centres
h_v	prior of daughter centres	μ	Poisson process of clusters
λ_d	unit rate daughter process	φ	intensity measure of μ
f_d	density of daughters w.r.t. λ_d	$\lambda_{w,\eta}$	unit measure in $\mathcal{M}$
h_η	prior of cluster height	$\pi(\boldsymbol{x})$	cluster prior density w.r.t. μ
ψ	prior number of centres	$p(\boldsymbol{x})$	cluster posterior density w.r.t. μ

Miscellaneous

$n(\cdot)$	size of a configuration	$\lvert\cdot\rvert$	cardinality of a set
$\lVert\cdot\rVert$	Euclidean norm	$nz(\cdot)$	number of nonzero elements

1 Introduction

During the last two decades, functional neuroimaging (FNI) has become an important topic in biomedical statistics. FNI is a class of powerful techniques for scanning the human brain, and it serves as a means for revealing activation processes. Compared to the conventional methodology, FNI enables huge possibilities for neuroscientists and neuropsychologists working in the field of brain research. From the very beginning it has been recognized that data acquired using FNI modalities have special features, and therefore, new statistical approaches are required to provide answers to questions relevant for neuroscientists applying FNI. Our objective in this Ph.D. thesis is to introduce new statistical ideas for applying Bayesian estimation methods in the context of FNI.

Statistical problems in connection with the modelling of FNI data originate from neuroscientific questions. Thus, in order to understand the motivation behind many statistical issues in FNI, it is necessary to have some familiarity with basic concepts and notions of neuroscience. The starting point for most investigations of FNI data lies in the general theory of the sensorimotor and cognitive function of the human brain. The brain manipulates available information in a highly parallel manner. Specifically, the brain can be considered to consist of many distinct modules which can function more or less independently of each other. A generally accepted view among neuroscientists is that these modules are localizable to some extent. From this perspective, it is of interest to study how the brain is divided into several modules, how localizable they are and what brain areas are responsible for processing the given information (Frackowiak *et al.*, 1997).

A comparison with conventional methodologies, such as experimental psychology, lesion studies and electroencephalography (EEG), illustrates the potential in

FNI. In the following, these techniques are discussed together with their shortcomings.

In experimental psychology, behavioural measures are used to improve our understanding of the brain function. As an example, these measures can be used to study dependencies between cognitive processes. By definition, two given tasks are considered to be independent if they can be done at once as well as they would be done separately. Experiments may help researchers to understand phenomena on the cognitive level, but in order to find an interpretation on the neuronal level, additional assumptions are needed. There is some tendency to presume that independent processes are also isolated on the physiological level. If this presumption is accepted, results from experimental psychology can lead to conclusions of connections between neuronal nets in the human brain. Nevertheless, the brain is essentially treated like a black box, and it remains unclear how the brain functions as an organ.

Lesion studies belong to the category of psychological research where the objective is to investigate associations between behavioural measures and brain anatomy. Typically, test individuals in these studies have experienced a surgical operation in the past or have been injured accidentally. Each localized lesion and corresponding observed change in his or her behaviour forms (at least) a hypothetical link between an anatomical structure and some function. However, an obvious difficulty in examining damaged brains is that the studies cannot be designed. It can well be the case that a patient having a lesion in an area of interest has also other lesions around the cortex, the outer grey matter of the brain. In this situation, it is unclear what lesions impair the performance of test tasks and how they possibly interact. A matter of an additional concern exists. There is no guarantee that the functional organization is not affected by lesions themselves. If the organization is altered, the conclusions are not relevant for understanding the function of the normal brain.

Lesion studies usually aim at clarifying the nature of the organization in the normal brain. Sometimes the scientific interest is focused on finding explanations for abnormal behaviour and, further, on understanding different types of variation in the human population. In a situation like this, an immediate question is whether the variability in the behaviour is linked with the variability in some anatomical feature of the brain or not. To get insight into this problem, it is necessary to be able to make measurements of the physical sizes of relevant brain structures such as sulci, gyri, etc. In the past, *post mortem* studies were the only source of detailed anatomical knowledge, but during the last decades progress has been made. Currently, it is possible to examine brain structures noninvasively from live individuals, using anatomical imaging techniques. To exemplify, Leonard *et al.* (1993) studied group differences between dyslexics and controls by acquiring anatomical planar magnetic resonance (MR) images from prespecified slice positions. The authors made measurements from an area covering *planum temporale* and *planum parietale* and were able to find indications about association between the relative sizes of certain structures and dyslexic behaviour. This study and other ones of similar type show that anatomical scanning can be an enlightening method if the objective is in the search of anatomical correlates explaining some unusual behaviour. However, it does not seem to be an appropriate method to investigate how normal brains work.

EEG offers considerable advantages over the previous methods in that the electrical brain activity can be monitored. What is more, the temporal resolution of the recorded signals is high, allowing us to observe changes in the activity within the range of milliseconds. EEG signals are often recorded together with ordinary responses. A drawback is that the spatial localization of activations is problematic, and, typically, the precise origin of the signals remains undetected.

We conclude that all the methods mentioned have some major deficiencies in providing information upon the actual brain processing. The development of FNI has meant a breakthrough for neuroscientific methodology. With special scanning devices it is now possible to monitor spatio-temporal phenomena in the live human brain. Some of the techniques are totally noninvasive and allow the brain to be scanned repeatedly, which makes it possible to apply a variety of experimental designs. This is an important property for constructing detailed brain maps of functional organization.

There are several imaging modalities available for FNI experiments. An especially important modality, and the one that will receive our main attention in this thesis, is functional magnetic resonance imaging (fMRI). fMRI is based on nuclear magnetic resonance (NMR) and is an example of a vast collection of magnetic resonance (MR) techniques (Stark & Bradley, 1988). fMRI is a noninvasive and nonionising acquisition technique. Consequently, it allows test individuals to be scanned an unlimited number of times without any known risk for health. Because of these appealing features, fMRI is acknowledged to be one of the most promising measurement methods for neuroscience.

The justification for using fMRI to detect neuronal processing depends on complex couplings between neuronal activity, brain haemodynamics and magnetic characteristics of brain tissue. According to current neurobiological knowledge, neuronal activation is accompanied with changes in local haemodynamics such as a local increase in cerebral blood flow and blood oxygenation. The results of dynamics in microvasculature have been obtained by applying imaging spectroscopy to examine the exposed cortex of different animal species (Malonek & Grinvald, 1996). The observations are expected to be generalizable to humans, however. The most important haemodynamic changes pertain to oxyhemoglobin and deoxyhemoglobin concentrations, which are expected to affect the local magnetic characteristics of the tissue (Binder & Rao, 1994). Moreover, the radio frequency signals recorded by the scanner are sensitive to changes in homogeneity of the local magnetic fields of the brain, and thus a series of resonance signal intensities can be used to delineate brain activity (Figure 1.1). Due to the complexity of the scanning procedure, the

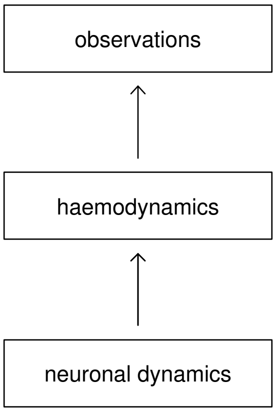

Figure 1.1 The fundamental coupling principle underlying the use of fMRI.

signal levels do not have an absolute meaning, but this is not a severe limitation in practice.

An example of other imaging modalities is positron emission tomography (PET). PET is an older technique than fMRI, and some of its features are inferior to those of magnetic resonance methods. First, the acquisition time for one scan can be about half a minute, and, second, the spatial resolution is usually lower than that of fMRI. Furthermore, the injection of radioactive contrast agents is required in PET studies limiting the total number of scans from a single individual. However, PET maps can provide an absolute measurement (unlike fMRI maps) which allows for investigating brain metabolism, for example. Both PET and fMRI are indirect measurement techniques by nature.

The efficient use of the fMRI technique in brain research is based on experimental designs. In experiments, a controlled stimulus is used to stimulate a test person. The recorded haemodynamic responses contain information on what parts of the brain are activated as a result of stimulation. fMRI data are a series of planar images $Y_1, \ldots, Y_T$ recorded at regular time intervals from chosen slice positions. Since each image consists of (rectangular) voxels, the whole data Y can be thought of as a collection of spatial time series $Y(s) = (Y_t(s))$. The MR signals are typically contaminated by scanner noise and uncontrolled physiological factors.

The objectives of the analysis of complex experimental fMRI data are usually localization of neuronal activation and estimation of activation magnitudes. In this thesis we shall concentrate on Bayesian estimation of spatial properties of stimulation effects. Our goal is to construct a prior distribution for brain activations which could be used to incorporate neuroscientific expert knowledge into statistical analysis of fMRI data. Our approach is based on a neuroscientific assumption that brain functions are localizable. We aim to model the set of activated voxels as a union of clusters. In our terminology, a cluster is a local area in the cortex. We assume that relevant expert knowledge can be expressed in terms of spatial extents and locations of clusters. The prior model has a hierarchical structure: we first model the locations of the strongest response in the clusters and after that haemodynamic effects, *given* the centre points. Our method provides applied researchers with an opportunity to utilize their experience when carrying out statistical analyses. It also challenges neuroscientists to specify numerically their prior conceptions of spatial characteristics of the brain function.

Several other priors for modelling fMRI data have been suggested earlier. The main ideas of some of these approaches are included in our proposal. Descombes *et al.* (1998b) emphasized the spatial smoothness of activation characteristics in the prior and applied nonparametric Bayesian smoothing methods (Besag *et al.*, 1995) for this purpose. Spatial smoothness is an ingredient also in our prior, but we do not need to utilize smoothing techniques. Hartvig & Jensen (2000) suggested that spatial contextuality and also the fraction of activated voxels should be controlled in the prior. Similar ideas play an important role also in our approach. In our model, the fraction of the activated areas is modelled by choosing a prior for the number and extent of the clusters. The proposal in Hartvig (1999) is closest to our work. In that work a configuration of activation centres is introduced and activation magnitudes in the vicinity of the centres are described using a parsimonious parametric model. Following Hartvig (1999), we utilize point processes in prior modelling. Our work differs from the approach in Hartvig (1999) in that we do not assume that activation magnitudes in the clusters can be modelled using simple parametric methods.

Our second step, after choosing a prior for activations, is to model the conditional distribution of signal measurements given the activations. This is not a straightforward task since the noise is usually spatio-temporally autocorrelated. In applications, the actual size of the correlations depends on the acquisition method. The presence of these correlations is often evident but, nevertheless, they have not always been accounted for in Bayesian analyses. Our opinion is that the careful modelling of the noise processes is an inseparable part of a responsible analysis of fMRI data. In this thesis, we show how Gaussian conditional autoregressive random fields can be fitted to the observed data, and we illustrate the consequences of omitting the dependence structure in the noise.

Following the Bayesian paradigm, we draw inference on brain activations from the *posterior distribution* of stimulation effects, which is the conditional distribution of activations given the data. We show in this thesis that, using a cluster prior, it is possible to make inference both on activation magnitudes and cluster centres. Also, since we have explicitly defined the concept of a cluster, we can carry out inference on other cluster characteristics than only centres. An appropriate measure of an activated cluster that could be considered is the integrated activation strength which describes the overall magnitude of the haemodynamic response.

There exist also many non-Bayesian statistical methods for fMRI problems. Statistical parametric mapping (SPM) techniques belong to this category, and they are probably the most widely used by neuroscientists today (Friston *et al.*, 1995a). The SPM consists of several statistical tests on the global null hypothesis that there are no activated voxels in the brain. The tests provide a means for finding an "estimate" for the set of activated voxels if the null hypothesis is rejected. In practice, the use of the tests is straightforward, but the distribution theory of the test statistics is rather involved and relies heavily on asymptotic properties of stationary Gaussian random fields. We shall consider the philosophy of SPM in some detail in this thesis in order to be able to relate SPM and Bayesian analyses.

The material of the thesis is organized in the following manner. Two sets of data, sound-stimulus data and word-stimulus data, are presented in Section 2, and typical characteristics of fMRI data are illustrated. The latter data will be used only to exemplify outlying observations, but the sound-stimulus data will be analysed in detail later in Section 7. After having introduced functional data, we review in Section 3 previous statistical methods for localizing activations and estimating haemodynamic responses to stimulus. We first review Bayesian techniques and then discuss the significance tests which are the core of the SPM methodology.

In Section 4, we introduce the concept of an activation profile and suggest how the clustering of activations in profiles can be modelled. We propose a marked Gibbs point process as a prior for a profile. In Section 5, we compress the original spatio-temporal data to *purely spatial* data and suggest how to model the conditional distribution of the compressed data given the activation profile. The choice of the prior and likelihood leads to a posterior distribution for a profile. As is typical in spatial applications, the posterior is high-dimensional and, therefore, standard numerical integration algorithms and conventional Monte Carlo methods (Ripley, 1987) are not practical for summarizing the posterior. To carry out inference, we construct in Section 6 a computationally intensive Markov chain Monte Carlo (MCMC) sampling algorithm for obtaining samples from the posterior.

The proposed Bayesian approach is applied in Section 7 to analyse the sound-stimulus data. In particular, we consider the sensitivity of the results to the choice of the prior. The priors used reflect different levels of prior knowledge on an auditory processing of interest. In Section 8, we compare structural and Markov random field priors by analysing synthetic data. The comparison is based on an index which measures the discrepancy between a true profile and a posterior distribution. In Section 9, we briefly consider alternative means for quantifying prior beliefs and for modelling the noise in the data. The purpose of the section is to clarify the role of several parameters of our prior and likelihood models by contrasting our models to some modified ones. Finally, Section 10 contains some observations related to our approach and fMRI statistics in general. The material in the Appendix consists of a few technical details which have been omitted from the main text.

2 Acquisition and characteristics of fMRI data

Functional MRI studies are conducted for finding answers to neuroscientific questions pertaining to the functional anatomy of the brain. These investigations consist of planning the experiment (the stimulation paradigm) and acquiring both functional and anatomical scans. We describe the data collection procedure and discuss the main properties of fMRI data. The discussion will form a basis for the statistical modelling in later sections. We illustrate some of the steps using example data from neuropsychological experiments. The present section will cover more details than are actually used in this thesis. Primarily, they are included for the sake of completeness and also to show the rich variety of aspects of fMRI data.

2.1 Acquisition of fMRI data

An experimental design includes a specification how test persons are stimulated during functional scanning. The use of a design usually aims at a high-level control over the brain processing during scanning. The planning of an experimental design is mainly based on psychological expertise. Perhaps the simplest method to stimulate an individual is to apply a periodic on/off-stimulus paradigm where two test conditions alternate during the scanning. Chapter 8 of Frackowiak *et al.* (1997) contains a comprehensive exposition on several types of study designs.

The planning of an experiment is not entirely a matter of neuropsychological thinking since there is always some measurement error (to a varying degree) in acquisitions. From the statistical viewpoint, a central issue is the construction of such designs that, when used, give the least variable estimates of changes in local brain haemodynamics. For example, in the situation of a simple on/off paradigm the task is to adjust the frequency of the stimulus. According to classical results in signal processing, the amount of information on stimulus effects is maximized if the spectral density of the noise is low at the forcing frequency of the stimulus paradigm and some of its harmonics; see Chapter 2 of Brillinger & Krishnaiah (1983).

The functional scans are acquired from selected slice positions relevant to the objective of the study. Modern scanners can be used for the whole brain imaging but unimportant slices can also be ignored. In the imaging procedure the chosen slices are scanned repeatedly in time. The dynamics of scanning depends on two parameters: the acquisition time for one image and the waiting time before starting a new acquisition. These time constants together characterize the temporal resolution.

The acquired functional MR images consist of pixels on a regular grid. The typical size ranges from 64×64 to 256×256. The field of view (FOV) together with the number of pixels determine the nominal spatial resolution. In functional imaging, the typical side length of a pixel varies between 1.0 and 3.0 mm. Since each pixel has a third dimension due to slice thickness, the term voxel (in place of a pixel) is also frequently used. Often, brain slices are chosen to be about 5.0 to 7.0 mm thick. Since the primary purpose of functional scanning is to investigate the temporal behaviour of the resonance signal in a brain location, it is more natural to consider the observed data as a collection of spatial time series rather than as a stack of images.

It is a common practice to acquire also whole brain *anatomical* MR scans when functional experiments are conducted. This provides an anatomical reference for any statistical conclusions concerning the functional images. The brain pixels can be segmented to white matter, grey matter and cerebro-spinal fluid (CSF) (Lange, 1996). For studying the brain function, it is the grey matter of the human cortex that the scientific interest is focused on since information is processed and manipulated there. Accordingly, the analysis of the data could be restricted to signals from the grey matter. However, the current practice has ignored the use of segmentations in this way. The main reason for considering three-dimensional volumes instead of two-dimensional cortical sheets is that nonnegligible inaccuracies occur in extracting the cortical surface from anatomical images and in registering it to functional images. We shall follow this usual convention.

As example data, we introduce here two series of functional scans from neuropsychological experiments: sound-stimulus data and word-stimulus data. The two data sets represent part of the dyslexia project of the Department of Psychology of Jyväskylä University in June 1997. The project was conducted to study differences in auditory processing between controls and dyslexic individuals, who have specific difficulties in reading and writing. Both data sets consist of 40 scans taken from a single sagittal brain slice from the left hemisphere of the same dyslexic person, the first test person scanned. The particular subject was an adult male. In each scanning session, the stimulus paradigm had the same alternating periodic pattern. In sound-stimulus sessions, the test persons alternatingly listened to the background noise of the scanner or a speech-like sound produced by manipulating human voice digitally (together with scanner noise). The manipulated sounds were generated by reducing the fast spectral components characteristic to speech, and the final stimulus consisted of pure sine wave tones. In word-stimulus sessions, the digital sounds were replaced by short spoken words. Both stimuli were presented to subjects binaurally using a computer playback system, a magnetically shielded transducer system and air conduction through paired plastic tubes (Vuorinen, 2000). Both stimuli began about 5 seconds before scanning to overcome the delay of haemodynamic response. The design is sketched in Figure 2.1.

In this particular dyslexia study, limitations on hardware made it necessary to examine functional data from one brain slice only. To choose the most relevant position for the functional slice, the functional scanning was preceded by anatomical MR imaging. The suitability of a slice was evaluated on the basis of anatomical structures identified on the slice. For studying dyslexia, examples of brain regions of interest are the planum region and the auditory cortex. The slice containing important anatomical structures was selected for the functional study. Figure 2.2 illustrates the anatomy from the selected functional slice position. When referring to the segmentation in Figure 2.2, we shall make use of the abbreviations that follow each region name. Since image voxels outside the brain are usually less interesting, we shall make use of a brain mask in all the figures of the thesis to extract the actual brain voxels. The technical details, such as the adjustments of the imaging device, are listed in Table 2.1.

2.2 Typical features of fMRI data

Stimulation effects are of primary interest in the acquired fMRI data. Before estimating these effects from the data, it is usually appropriate to preprocess the data to remove certain uninteresting features, which will be listed below.

Functional experiments are usually carefully arranged in order to optimize the quality of the data. To minimize disturbances in the MR signals recorded, auxiliary devices can be used to help the test persons to lie motionless while scanning (Kwong, 1995). Despite these efforts, observations tend to contain harmful features which must be removed during the data preprocessing steps. A typical disturbance is a gradual rigid head motion. To understand how it affects the signal, we note that the overall mean of voxel time series is dependent on the location of the voxel. If the head of a person rotates slowly during the scanning, any voxels for which the baselines of the surrounding voxels have large differences may contain trends. Head movements can be corrected by applying translations and rotations in the three-dimensional space so that the sums of squared differencies between corre-

Table 2.1 Values of the imaging parameters in the experiment.

	scanner	Siemens Impact Expert
	field strength	1.0 Tesla
	dim. of slices	5 mm $\times$ 220 mm $\times$ 220 mm
MRI	pulse sequence	spin echo
	time to repeat (TR)	730.0 ms
	time to echo (TE)	15.0 ms
	spatial resolution	512 $\times$ 512
	pixel size	0.43 mm $\times$ 0.43 mm
fMRI	pulse sequence	Turbo-Flash
	TR	90.0 ms
	TE	56.0 ms
	spatial resolution	256 $\times$ 256
	pixel size	0.86 mm $\times$ 0.86 mm
	acquisition time (TA)	14.0 s

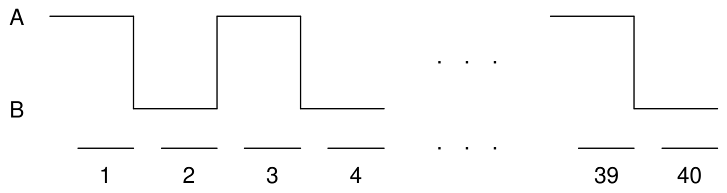

Figure 2.1 The design of the sound-stimulus experiment. Test condition A corresponds to listening to manipulated sounds. During condition B a test person listened to the background noise of the scanner. The scanning intervals are numbered as $1, 2, \ldots, 40$. After onset of either stimuli there was a wait of a few seconds before functional scanning; see the gaps between the scanning intervals.

sponding voxel intensities of a base image (a fixed functional scan, for example) and other functional images are minimized. The correction is based on all information contained in the rectangular MR images; even voxels outside the brain can be utilized. This strategy has a potential to give more precise displacement estimates than those calculated using information from brain tissue only. Otherwise, the signal intensities outside brain tissue are considered unimportant and are ignored from the actual statistical analysis since they do not provide any information on changes in brain haemodynamics. Inappropriate aligning techniques may influence the spectral properties of the signal. In fact, while aligning the images effectively removes low-frequency components, incorrect image alignment can increase the high-frequency components of the noise, as reported by Hajnal *et al.* (1994).

The search of outlying observations applies as well in fMRI data analysis as in most empirical studies. An informal but useful diagnostic method for detecting large deviations in brain signals is to visualize data using spatio-temporal displays. In Figure 2.3 signal series from the word-stimulus data are shown. A logically sound strategy to detect outliers is first to determine the time points that have been corrupted, and then examine numerically in what brain areas the voxel time series contain untypical values at the selected time points. In this particular data, the 18th value in the middle of the time series is clearly an outlying observation. The voxel series that were severely corrupted are highlighted on the right in Figure 2.3.

Outliers may arise in several different ways. In echo planar imaging (EPI) studies, the MR signal stabilizes only after a few seconds from the start, and, consequently, the first images are outliers which are usually discarded completely. In the single slice imaging, out-of-plane motion spoils the MR signals because aligning the functional images in the imaging plane does not lead to a reasonable correction. For the multislice imaging, out-of-plane motion is not as serious a problem. The presence of outliers can be regarded as a missing data problem. However, it seems that the tendency in fMRI research is to guarantee that experimental data are of high quality rather than to correct data for low-quality observations. In the dyslexia study, a conventional fixation device was used to provide moderate head restraint.

In group studies, data must be collected from several individuals. Pooling information over more than one subject necessitates the replacing of the natural 3D coordinate system of the brain by normalized coordinates since individual anatomy exhibits large variation. The change in coordinates transforms the individual brain anatomy to standardized anatomy. Several normalization techniques exist but perhaps the most popular is the piecewise-linear method published in Talairach & Tournoux (1988).

As shown above, certain data manipulation operations are recommended at a preprocessing stage. However, manipulation may increase the risk of invalidating the results of subsequent statistical analyses and may lead to biased inference. In fMRI problems there are risks since it is quite easy to create (applying software

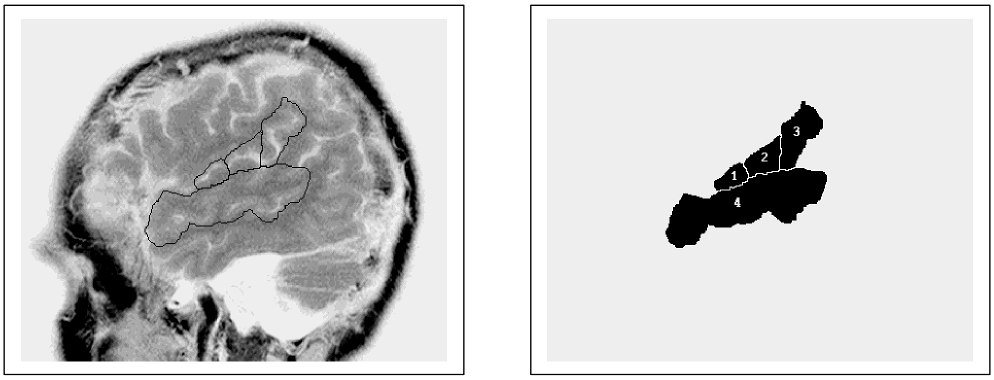

Figure 2.2 An anatomical image from the functional slice position (left) from which four regions relevant to the study were marked (right). The names of the regions: 1 transverse temporal gyrus (or primary auditory cortex) (TTG), 2 temporal bank of planum (PT), 3 parietal bank of planum (PP) and 4 superior temporal gyrus (STG).

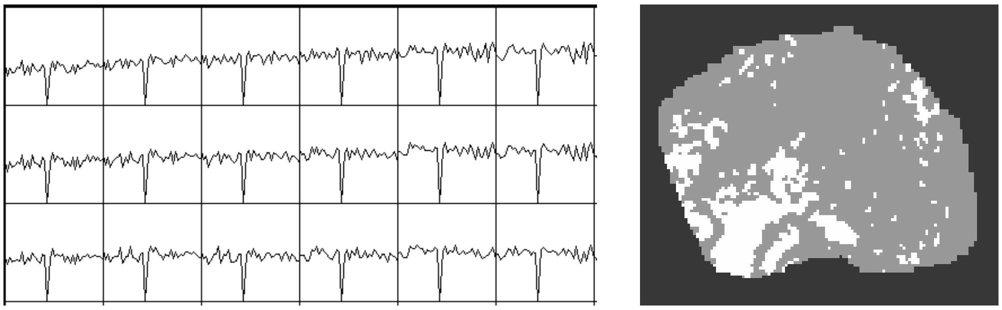

Figure 2.3 A 3×6 window where sudden head motion was visually detected from the voxel time series of the word-stimulus data at 18th time point (left). The contaminated voxels are painted white (right). A voxel was considered contaminated if the difference between the signal intensity (at 18th time point) and the mean of a series exceeded three times the standard deviation of the series.

packages) new modified data, visualize the manipulated intensities, and choose for detailed analysis the data that accords with one's prior expectations. As Lange (1996) has remarked: "*Voxels can be grouped, stretched, smoothed, replaced or otherwise altered in infinite number of ways to cause desired features to appear or disappear at will.*" One of the most important things how statistics could contribute to FNI should be removing such arbitrariness from the data analysis of functional images.

We shall now discuss the characteristics of brain activations in detail. The effects of controlled stimulation have both spatial and temporal properties. Neuroscientific theories, lesion studies and studies based on neuroimaging support the hypothesis that to some extent neural processing is segregated to brain regions especially reserved to some special function. It follows that the spatial locations of stimulus-related neuronal processing depend on the experimental design. As to temporal properties, the dynamical range of neurons in brain tissue is in milliseconds. The time homogeneity of the neural response is an assumption which fMRI analyses are typically heavily based on. If inhomogeneity arises, a possible reason may be an attentional effect, adaptation to stimulation, or learning. Usually, it is a design issue to guarantee the stability of the response in experiments. Sometimes, learning effects may be of interest themselves (Frackowiak *et al.*, 1997).

The haemodynamic responses have slightly different spatial and temporal properties from the neural ones. Spatially, the haemodynamic response is diffused a few millimeters around a neurally activated tissue (Malonek & Grinvald, 1996). Temporally, the haemodynamic response is slower than the neural response. In fMRI literature it is common to describe this slowness using concepts called *delay* and *dispersion*. The size of the delay determines how long it takes before the MR signal level starts to change as a result of neural activity. Dispersion, in turn, quantifies the speed at which the signal rises to its maximum and falls to its minimum level. The delay is usually between 4 and 10 seconds (Bandettini *et al.*, 1993; Friston *et al.*, 1994b). A challenging issue in fMRI research is that the size of delay and dispersion may vary spatially around the cortex (Lange & Zeger, 1997).

A natural means to estimate the temporal pattern of a haemodynamic response to stimulus is by fitting a convolution model. In on/off-experiments, the design can be described as a sequence $(D_t)_{t=1}^T$ of 1's and -1's. Then, a temporal pattern $(X_t(s))_{t=1}^T$ can be modelled as

$$X_t(s) = a_0(s)D_t + a_1(s)D_{t-1} + a_2(s)D_{t-2} + \dots , \tag{2.1}$$

where $a_0(s), a_1(s), \dots$ are some parametric weights. An advantage in convolution models is that there is an obvious relation between haemodynamic delays and dispersions and the model weights. We do not consider the modelling of temporal features of haemodynamic responses in detail in this thesis. Discussion on this topic can be found elsewhere in fMRI literature; see Rajapakse *et al.* (1998), for example. The temporal features are usually less interesting for brain mapping purposes, but taking them into account leads to potentially sharper localization. Figure 2.4 illustrates the temporal patterns that the local haemodynamics can take as a result of instabilities in neuronal processing.

Unknown delays and dispersions are not a problem in our sound-stimulus data. Our data were temporally sparsely acquired as the stimulation paradigm of the data shows (Figure 2.1). Therefore, unknown haemodynamic delays can be ignored in a natural way. In this thesis, our continuing assumption is that the temporal characteristics of haemodynamic responses are known. To simplify the presentation, we also presume that these temporal patterns are spatially invariant, i.e., $X_t(s) \equiv X_t$.

The interpretation of the single voxel time series is not always as straightforward as we have explained above. Sometimes external processes, such as rigid head motion, are associated with the stimulation paradigm. In that kind of situation, the head motion induces extra variability to the intensity time series which resembles activation but which must not be interpreted as such. Logically, this phenomenon is called *artefactual* activation. Hajnal *et al.* (1994) devised a simple method for detecting signs of stimulus-correlated motion. First, an image (which may be any one of the scans) is selected from the stack of functional images. Then, the selected image is matched with all other images of the functional series by calculating optimal translations and rotations. In this way, artificial data can be constructed, and the results obtained from the analysis of both the original and the artificial data can be compared. If the findings are highly consistent with each other, the reliability of the results from the original data is obviously questionable. Stimulus-correlated motion tends to increase the number of activated clusters although it can also have potential to hide true activations. Handling the correlated motion is basically a design issue since the interpretation of the results is always conditional on the assumption that the experiment is controlled by the design.

We shall next discuss some aspects of noise which contaminate MR measurements from the tissue. As already explained, rigid head motion distorts the MR signal. There are also other sources of noise, such as respiration and heart beating, which originate from normal body physiology. Both respiration and cardiac effects affect the CO_2 level and the O_2 consumption, which the resonance signal is sensitive to.

The strength of the noise variance varies among voxels in a slice as Figure 2.5 indicates. The proximity of large blood vessels tends to increase the variability of the noise (Kwong, 1995). There is some evidence that part of inhomogeneity in variances can be explained so that noise deviation actually scales with the level of the baseline signal. If this is the case in the data, it can be advantageous to analyse the data on a logarithmic scale. From Figure 2.6 it can be observed that the overall mean of the series and variance are related, but not strongly in our sound-stimulus data.

A sensible strategy for treating respiratory and cardiac effects is to measure the evolvement of these processes during the scanning and use them as covariates

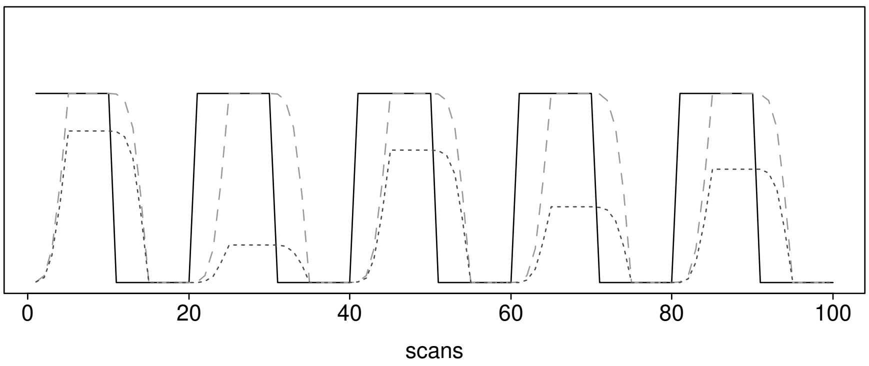

Figure 2.4 Simulated temporally homogeneous and inhomogeneous haemodynamic responses: stimulation paradigm (—), a homogeneous response pattern (– –) and an inhomogenous one (···). The homogeneous pattern was generated by applying the convolution formula (2.1) to an artificial stimulation series (or design series) of length 100. The inhomogeneous temporal pattern is a randomly scaled version of the homogeneous one.

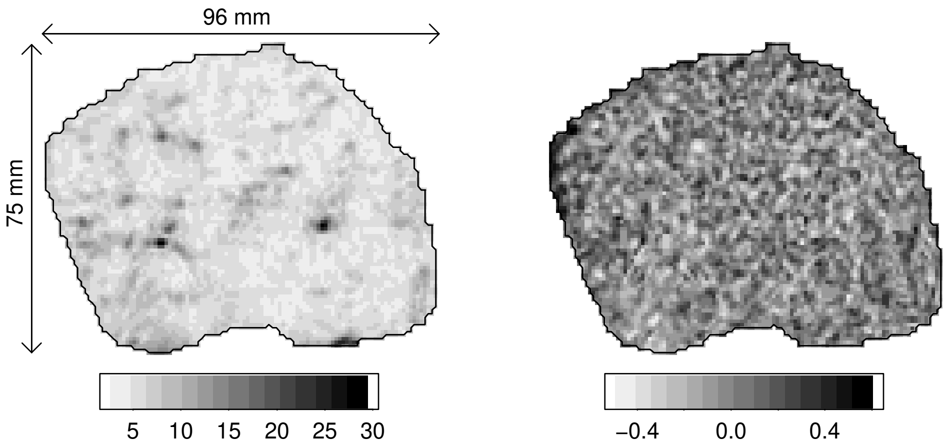

Figure 2.5 Voxelwise standard deviations of residual series in the sound-stimulus data (left) and voxelwise temporal autocorrelations at lag one in the same data (right). The residual series were calculated by fitting a linear model to data using the design series of the experiment as an explanatory series. On the left, the numerical values are measured in MR units of the scanner.

in statistical modelling. However, the inclusion of physiological covariates requires the establishing of a mathematical model linking the measured processes to the observed MR signals. This link is avoided if physiological effects on the MR signal are measured directly inside the brain. This can be done by locating structures of the cerebral blood circulation system and predicting physiological covariates from anatomically selected time series. A natural solution to the prediction problem is to assume that respiratory and cardiac effects have a fixed frequency or that at least this frequency varies smoothly in time (Petersen, 1998). A drawback in this approach is that problems with uncertainty of the locations of vessels are encountered. In general, spatial periodograms can be useful in indicating quasiperiodic physiological fluctuations in a brain slice (Lange & Zeger, 1997).

The background noise level of the magnet also affects the quality of fMRI data. To estimate the size of variation in the MR signal due to the scanner itself, we selected a collection of voxels outside the body of the test subject and extracted voxel time series from this set for further analysis. In this way, we estimated that the median standard deviation of the background noise was around 1 % of the overall MR signal level in brain tissue in the sound-stimulus data.

Typically, the noise is correlated between intensities of voxels close both in space and time. This spatio-temporal dependence structure is the joint effect of several endogenous physiological processes. In Figure 2.5, estimated voxelwise temporal autocorrelations of lag one in the sound-stimulus data are shown. The temporal correlations are weak, which is probably a consequence of the temporally somewhat sparse acquisition method. We observe that some voxels near anatomical boundaries exhibit a larger temporal dependence than voxels elsewhere. Time can be regarded as an additional dimension to spatial dimensions when modelling fMRI data. However, it has a slightly different role since it is far less natural to presume the stationarity of the spatial dependence structure than that of the temporal correlation structure. A reason for this is that spatial correlations may be local, that is, they may depend on the local brain anatomy. Nevertheless, most of the current practice has adopted the stationarity of spatial autocorrelations. Obviously, if space stationary dependence models are to be applied, they must be fitted to the *mean* correlation structure over the whole brain slice. By definition, the mean spatial correlation at some spatial lag l is the average of spatial correlations between all pairs of voxels which are separated by l. The mean spatial correlations in the four main directions (with respect to the coordinate system of the image) of noise series of the sound-stimulus data are shown in Figure 2.7. It can be seen from this figure that the strength of the correlations are nearly of the same size in all four directions. In other words, no anisotropic phenomena were detected when the mean structure was analysed.

The material of this section has covered several key issues in fMRI research: the role of experimental designs, data acquisition, preprocessing steps, activation processes, and the features of noise. Consequently, there are many aspects in fMRI problems that might deserve to be subject to statistical research. We shall limit the scope of the thesis to *spatial aspects* of fMRI data. In the following section we shall review part of the statistical research on fMRI problems and emphasize the aims and means of spatial modelling.

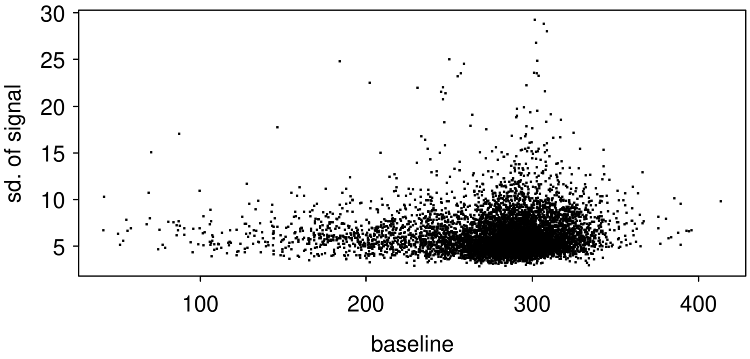

Figure 2.6 The standard deviation of the residual series of the sound-stimulus data plotted against the overall mean of observed voxel time series. Each point in the figure corresponds to one brain voxel. The residual series were calculated as explained in the caption of Figure 2.5.

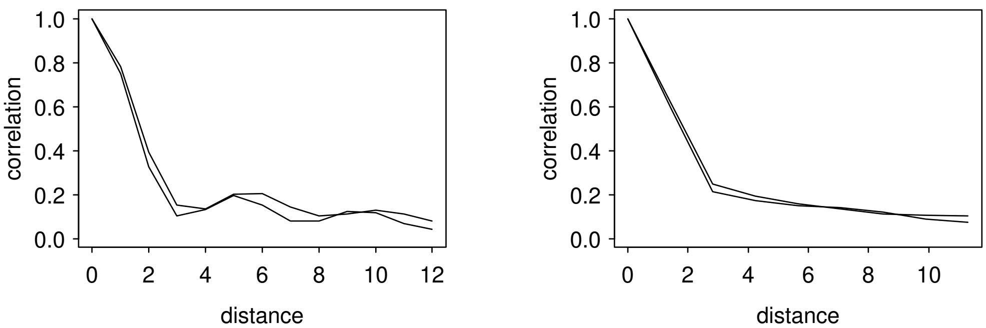

Figure 2.7 Empirical mean spatial autocorrelations in the sound-stimulus data in the coordinate directions of the image (left) and in the two main diagonal directions (right). On the horizontal axis, the distance is measured in voxels. Distance of 10 voxels corresponds to 8.6 mm.

3 A review on statistical methods for fMRI data

The purpose of the present section is to overview some previously suggested statistical analysis techniques for fMRI data. We shall discuss the assumptions, goals and computational methods of the papers emphasizing the treatment of spatial aspects in functional MRI data. Nearly all the methods proposed concern the problem of estimating the haemodynamic processes or localizing neuronal activations. The statistical methods differ from each other considerably and they will cover both Bayesian issues and frequentist significance tests. In addition to spatial modelling ideas, other topics that will be discussed in detail are the parallels and dissimilarities of the two fundamentally different approaches in the context of FNI.

3.1 Bayesian estimation of brain activations

Bayesian estimation of the presence of activation is considered among others in Everitt & Bullmore (1999) and Hartvig & Jensen (2000). In the presence/absence analysis an indicator field $\boldsymbol{I} = (I(s))$ is defined in the space of brain voxels $\mathcal{S}$ in order to use it to classify voxels to activated ($I(s) = 1$) and non-activated ($I(s) = 0$). Both papers suggest that each voxel time series $Y(s)$ of a fMRI data $Y = (Y(s))$ is reduced to a scalar-valued test statistic $Q(s)$ which is sensitive to stimulus-related changes. The key idea is to build a model for $f(Q(s)|I(s))$, the conditional density of the observation $Q(s)$ given the state of activation $I(s)$, and to describe prior beliefs in the indicator field I in terms of a prior distribution. This leads to a joint distribution model for $(\boldsymbol{Q}, \boldsymbol{I})$, and the properties of $\boldsymbol{I}$ after observing $\boldsymbol{Q}$ can be inferred from the posterior distribution $\boldsymbol{I} \,|\, \boldsymbol{Q}$.

In both papers, the components of $\boldsymbol{Q}$ were assumed to be conditionally independent given $\boldsymbol{I}$. Everitt & Bullmore (1999) made further assumptions and considered the components $I(s)$ independent and identically distributed a priori. In that case, the components of $\boldsymbol{I}$ are independent in the posterior and the marginal posterior probabilities $\Pr(I(s)=1|Q(s))$ can be calculated directly from the Bayes rule

$$\frac{f(Q(s)|I(s)=1)\Pr(I(s)=1)}{f(Q(s)|I(s)=1)\Pr(I(s)=1)+f(Q(s)|I(s)=0)\Pr(I(s)=0)}. \tag{3.1}$$

Hartvig & Jensen (2000) considered the approach by Everitt & Bullmore (1999) inefficient since the spatial clustering of activations was ignored from the prior model of $\boldsymbol{I}$. In principle, clustering can be accounted for by applying a spatially correlated prior model for $\boldsymbol{I}$. A problem with correlated priors is that there are no computationally efficient formulas available for calculating the marginal posterior probabilities of $\boldsymbol{I}$. The only solution is to apply computationally intensive Markov chain Monte Carlo (MCMC) techniques for estimating the posterior probabilities. The authors of Hartvig & Jensen (2000) recognized that it can be sufficient to carry out *local* inference around each voxel s if the objective is to make inference on the state of $I(s)$. In local inference, a neighbourhood $\mathcal{N}_s$ of a voxel s is chosen and a model is formulated only for subfield $I(\mathcal{N}_s)$ and a subset $Q(\mathcal{N}_s)$ of all observations $\boldsymbol{Q}$. The authors suggested several local *uniform* correlation prior models for the subfield $I(\mathcal{N}_s)$, proposed inference to be drawn from $I(\mathcal{N}_s)|Q(\mathcal{N}_s)$ and derived expressions in a closed form for the marginal posteriors $I(s)|Q(\mathcal{N}_s)$. Ideally, it would be more efficient to make use of all the data and condition on $\boldsymbol{Q}$ rather than only on $Q(\mathcal{N}_s)$. However, most of the information provided by the data is contained in $Q(\mathcal{N}_s)$, and, therefore, it can be sensible to ignore data outside the neighbourhood $\mathcal{N}_s$ of s and utilize cleverly constructed analytical approximations. To contrast the local approach with the more conventional global approach, it must be emphasized that MCMC simulations of the posterior of $\boldsymbol{I}$ can be used to calculate estimates for any probabilities and it is not necessary to restrict attention to marginals $I(s)$. In applications, the crucial issue is whether relevant questions can be answered in terms of marginal probabilities or not.

A common feature in the two papers is that the stochastic properties of noise are not fully accounted for. Assuming the conditional independence of the components of $\boldsymbol{Q}$, given $\boldsymbol{I}$, is equivalent to assuming that noise processes are spatially uncorrelated. Clearly, this is a simplification and is made for mathematical convenience. Since the noise in fMRI data is spatially correlated, it should be reflected in the statistical models.

Descombes *et al.* (1998b) focused on the calculation of Bayesian point estimates for several spatial and temporal characteristics of activation processes. They assumed that the properties of fMRI data can be described using a few spatial parameter fields, one of which is an indicator field $\boldsymbol{I}$ for activation. The aim was to calculate maximum a posteriori (MAP) estimates for these spatial parameters. The main idea in the construction of the prior for the parameters was to smooth parameter fields avoiding smoothing between activated and non-activated voxels and encouraging sharp boundaries for activated clusters. To this end, they employed nonparametric Bayesian smoothing techniques. The activation indicator field was assigned a contextual prior, and the way how the other parameter fields were smoothed depended on the segmentation defined by the indicator field. In other words, the suggested prior has a hierarchical structure. The point estimate of the activation indicator field provides a classification of the voxels. The choice of the likelihood function of the parameter maps was based on mathematical convenience and did not have a probabilistic justification. Salli *et al.* (2001) also employed the segmentation idea and showed that the sensitivity to weak activations increased as the result of "spatial contextual clustering".

Kornak *et al.* (1999) considered the Bayesian estimation of indicator field $\boldsymbol{I}$ and magnitudes of activations. The data Y were modelled as

$$Y_t(s) = \phi(s) + \beta(s)X_t(s) + \delta_t(s), \tag{3.2}$$

where $\phi(s)$ is an overall mean in voxel s, $X(s) = (X_t(s))_{t=1}^T$ is the temporal pattern of stimulus-related activation in voxel s, $\beta(s)$ represents the magnitude of response in s and $\boldsymbol{\delta} = (\delta_t(s))$ is zero mean noise term. Thus, the temporal patterns of activations were allowed to depend on location. The patterns $X(s)$ were estimated parametrically simultaneously with terms $\beta(s)$ from the data. To incorporate spatial prior knowledge, β was factorized as $\beta(s) = \beta^*(s)I(s)$, where β^* is a smooth field and $\boldsymbol{I}$ is a positively correlated indicator field. The Gaussian conditional autoregressive (CAR) field was used as the prior of β^* and a thresholded intrinsic Gaussian Markov random field as the prior of $\boldsymbol{I}$. Then, the estimated magnitudes $Z(s) = \hat{\beta}(s)$ were modelled as

$$Z(s) = \beta^*(s)I(s) + \epsilon(s), \tag{3.3}$$

where $\boldsymbol{\epsilon}$ is an *uncorrelated* zero mean Gaussian process with spatially homogeneous variance. The motivation for using a multiplicative model is of similar type as in the approach by Descombes *et al.* (1998b).

Taskinen (1998) used the model

$$Y_t(s) = \phi(s) + \beta(s)X_t + \delta_t(s), \tag{3.4}$$

where the temporal pattern of activation $(X_t)_{t=1}^T$ was assumed to be known and the same in all voxels. The approach differed from that of Kornak *et al.* (1999) in the way how the stochastic properties of the noise $\boldsymbol{\delta}$ were modelled. First, all fields $\delta_t = (\delta_t(s))$ were assumed to be independent and identically distributed, having a Gaussian CAR distribution. Second, the noise $\boldsymbol{\delta}$ was modelled as a *heteroskedastic* process using one precision parameter for each voxel. A pairwise difference smoothing prior (Besag *et al.*, 1995) was used to account for the smoothness in the coefficients $\beta(s)$. In a similar manner, the uncertainty due to many additional precision terms was accounted for by introducing another smoothing prior.

Hartvig (1999) suggested a spatio-temporal prior model for the activation process which can be used to make inference from both spatial and temporal properties of activations. The model for data Y is assumed to have the form

$$Y_t(s) = \phi(s) + \beta(s) X_t^* + \delta_t(s), \tag{3.5}$$

where $\beta(s)$ is the magnitude of activation in s, $X^* = (X_t^*)$ is the temporal pattern of activation (the same for all s), and $\boldsymbol{\delta} = (\delta_t(s))$ is a correlated error process. The temporal pattern X^* was assumed to be a sum $X_t^* = X_t + G_t$ where $X = (X_t)$ is a haemodynamically convolved design series and $G = (G_t)$ is a random effect describing the (possible) temporal instability of the activation process. It was suspected that the level of attention of a test person, for example, can produce departures from the time homogeneity of the MR response. The attention effect G was assumed to follow a Gaussian (zero mean) random walk with increments having a known variance.

The author suggested two models for magnitudes $\beta(s)$: one model where the magnitudes are restricted to be positive, and the other without any restrictions. We shall first consider the restricted model. The modelling of the spatial features of the approach was motivated by neuronal arguments. It was assumed that typically only some parts of the brain are involved in the neuronal processing of interest during the image acquisition. A set of points $a_1, \ldots, a_k$ (in the continuous version of the space of brain voxels) called centre points were introduced to represent the active brain areas in such a way that haemodynamic effects occur only in the proximity of the configuration $\{a_1, \ldots, a_k\}$. To construct the magnitude field corresponding to the centre configuration, a local spatial profile (Gaussian bell) was assigned to each centre. A Gaussian bell $B(s; b)$ is an exponential of a negative definitive square form of the argument vector s. Bell characteristics can be represented as a vector $b = (c, \theta, d)$ the components of which are eccentricity c, angle of rotation θ with respect to the centre of a bell, and area d of contourellipse at half height. In mathematical terms, the model for $\boldsymbol{\beta}$ is a superposition

$$\beta(s) = h_1 B(s - a_1; b_1) + \cdots + h_k B(s - a_k; b_k), \tag{3.6}$$

where $h_1, \ldots, h_k$ are some positive heights. The author claimed that "to some extent [the bells in the sum above] can be thought of as individual [activation] centres in the brain".

By the formulation (3.6), the map of magnitudes $\boldsymbol{\beta}$ is parameterized by the heights $h_1, \ldots, h_k$, centre points $a_1, \ldots, a_k$ and bell characteristics $b_1, \ldots, b_k$. These parameters can be represented as a marked point configuration $\boldsymbol{z} = \{z_1, \ldots, z_k\}$ where each point $z_i = (a_i; (b_i, h_i))$ is a combination of a location a_i and a mark (b_i, h_i). Depending on the stochastic properties of $\boldsymbol{z}$, different prior models for $\boldsymbol{\beta}$ result. The author suggested a pairwise interaction Gibbs process for $\boldsymbol{z}$ having density

$$\pi_{pos}(\boldsymbol{z}) \propto \prod_{i=1}^{k}[p_a(a_i)p_c(c_i)p_d(d_i)p_h(h_i)] \prod_{i\neq j} \Psi(z_i, z_j) \tag{3.7}$$

with respect to a unit rate Poisson process in the product space of the (continuous) search volume and the mark space. The densities p_a, p_c, p_d and p_h can be used to control the number of centres, their location and the bell characteristics. To complete the construction of the prior it is enough to specify these densities and the interaction function Ψ. A repulsive interaction function Ψ was used and the degree of repulsion was modelled to be dependent on the similarity of the Gaussian bells. The similarity of two bells, z_i and z_j, can be quantified by the Kullback J-divergence measure

$$d_J(z_i, z_j) = \int (f_i - f_j) \log \frac{f_i}{f_j}, \tag{3.8}$$

where f_i and f_j are the Gaussian densities corresponding to the bells. The suggested pairwise interaction function was then

$$\Psi(z_i, z_j) = 1 - \exp(-(d_J(z_i, z_j)/\rho)^p), \tag{3.9}$$

where ρ represents the radius of spatial interaction and p determines the type of decay of the interaction. It follows that the inhibition depends both on the bell characteristics b_i and b_j and on the locations a_i and a_j. On the other hand, the heights h_i and h_j do not play any role.

The prior for a general signed field of parameters $\boldsymbol{\beta}$ was defined as the distribution of the difference $\boldsymbol{\beta}^{(1)} - \boldsymbol{\beta}^{(2)}$ of two positive fields, $\boldsymbol{\beta}^{(1)}$ and $\boldsymbol{\beta}^{(2)}$. In order that $\boldsymbol{\beta}^{(1)}$ would correspond to the positive of part $\boldsymbol{\beta}^{+} = \max(\boldsymbol{\beta}, 0)$ of $\boldsymbol{\beta} = \boldsymbol{\beta}^{(1)} - \boldsymbol{\beta}^{(2)}$, it was required that the supports of $\boldsymbol{\beta}^{(1)}$ and $\boldsymbol{\beta}^{(2)}$ should not overlap much. This was accomplished by considering the joint distribution of two marked point processes, $\boldsymbol{z}^{(1)}$ and $\boldsymbol{z}^{(2)}$, which inhibit each other. Their joint density $\tilde{\pi}$ was given by

$$\tilde{\pi}(\boldsymbol{z}^{(1)}, \boldsymbol{z}^{(2)}) \propto \pi_{pos}(\boldsymbol{z}^{(1)})\pi_{pos}(\boldsymbol{z}^{(2)}) \exp\left(-\gamma \sum_{s} \beta^{(1)}(s)\beta^{(2)}(s)\right), \tag{3.10}$$

where γ is a positive penalty parameter controlling the strength of inhibition.

The noise in Y was assumed to be homoskedastic, and a separable model was used to model the spatio-temporal correlations of $\boldsymbol{\delta}$. This leads to a likelihood function for $\boldsymbol{\beta}$ and temporal (random) effect G. The resulting posterior is a Gibbs point process, and an MCMC simulation algorithm was devised by the author to obtain samples from the posterior. Essentially, the MCMC technique used was an application of a general Metropolis-Hastings algorithm for finite point processes introduced by Geyer & Møller (1994).

To adjust the densities p_h and p_d it was suggested that the heights and areas should have an uninformative prior but very small (in width or height) bells should be penalized. A similar penalization argument concerned heavily eccentric bells, which guides the choice of p_c. As to the value of γ, small values should be avoided. Hartvig (1999) showed that if the noise is modelled as an uncorrelated process, there is a value $\gamma = \gamma_0$ such that the two processes, $\boldsymbol{z}^{(1)}$ and $\boldsymbol{z}^{(2)}$, are independent in the posterior. In principle, the posterior independence of these processes permits inference to be drawn separately from the positive and negative parts of the $\boldsymbol{\beta}$ field. However, it was not shown whether the proposed value of γ is large enough in practice. Also, this suggestion depends on the number of scans, which is somewhat strange. The rationale behind the use of Ψ was not stated clearly. However, it is fairly easy to see how the interaction function influences the role of the prior. Plainly, if inhibition is weak, the bells may overlap each other and the points in the configuration $\boldsymbol{z} = \{z_1, \dots, z_k\}$ can be harder to interpret. At the other extreme, if inhibition is strong, $\boldsymbol{\beta}$ will be a superposition of distinct Gaussian bells. This is probably too restrictive a prior assumption since each bell is given by a simple parametric function.

The last two Bayesian papers that we shall review here are Descombes *et al.* (1998a) and Hartvig (2000). They differ slightly from the Bayesian proposals above in that the focus was neither on localizing activations nor on inferring magnitudes of activations. Instead, Descombes *et al.* (1998a) considered Bayesian methods for noise reduction to restore the complete spatio-temporal activation signal from a series of scans. Hartvig (2000), in turn, aimed at drawing inference on neuronal fields by deconvoluting profiles.

We discuss first the paper by Descombes *et al.* (1998a). In this approach, the activation signal was assumed to be smooth in space and time. The idea was to construct a robust smoothing prior for the activation process, calculate the maximum a posteriori (MAP) estimate of the activations and treat this estimate as a new, less contaminated data. To calculate the MAP estimate, a version of simulated annealing was applied for this purpose. The MAP estimate can then be treated as data which will be the subject for further analysis. No parametric assumptions on activation signals were made. In particular, the design plays no role in the procedure. The authors argued that the preceding noise reduction step before inferential steps tends to lead to improved inference. The suggested method is not genuinely Bayesian since it leads to an *improper* posterior distribution. However, from a purely operational point of view, this is not a complication since the objective is in calculating the posterior mode, not posterior probabilities. Petersson *et al.* (1999) argue that a Bayesian restoration of a spatio-temporal activation signal is not recommended since the effect of bias on the restored signal can be more difficult to understand than the bias created by a linear smoothing filter, which is commonly used in image analysis. The point is that it is not straightforward to model a stochastic relation between restored intensities and unobserved activations.

Hartvig (2000) made an attempt to model coupling between spatial haemodynamic effects and the underlying neuronal field in order to make inference on the latter field. A basic assumption was that a linear model such as (3.4) captures the activation information from the data Y. As in Taskinen (1998), the temporal pattern of activations, $X = (X_t)$, is assumed to be known. Instead of modelling the original data Y in the analysis, the least squares estimate of $\boldsymbol{\beta}$ (denoted here by Z) was considered. The compressed data Z can be expressed as a sum $Z(s) = \beta(s) + \epsilon(s)$, where $\boldsymbol{\epsilon}$ is a spatial zero mean Gaussian process. The map $\boldsymbol{\beta}$ was modelled as a spatial convolution of a neuronal field $\Gamma = (\Gamma(s))$ by the formula

$$\beta(s) = \sum_{v} k(s, v)\Gamma(v), \tag{3.11}$$

where $k(\cdot, \cdot)$ is a convolution kernel. The motivation for considering spatial deconvolutions is that the haemodynamic process does not necessarily provide enough localized information on neuronal processing since the blood oxygen level dependent (BOLD) resonance signal is typically diffused a few millimetres in the microvasculature (Malonek & Grinvald, 1996). The variates $\Gamma(v)$ of the neuronal field Γ were modelled as independent and identically distributed in the prior. This prior, common to all variables $\Gamma(s)$, was defined as a mixture of three distributions in the following way. First, the conditionals $\Gamma(s)|\Gamma(s) > 0$ and $\Gamma(s)|\Gamma(s) < 0$ were specified. Second, mixing probabilities p^+ and p^- were assigned to them in such a way that the sum of p^+ and p^- is less than one. Then, the prior is the mixture of the two conditionals and the Dirac distribution where the weight of the Dirac probability measure is $1 - p^+ - p^-$. It follows that *a priori* a given $\Gamma(s)$ can be zero with positive probability.

In principle, available prior information on the neuronal field Γ can be taken into account when specifying the component distributions. To apply the convolution approach, an appropriate choice for the width of the smoothing kernel k is necessary, and recognizing this, Hartvig (2000) suggested a method for estimating the width as well as other hyperparameters from the data. Finally, a highly developed simulation technique was proposed for drawing inference from the posterior.

We comment here on two features of the approach above. Firstly, an uncorrelated model was fitted to the spatial noise process $\boldsymbol{\epsilon} = (\epsilon(s))$ in the paper. It follows that part of the errors will be included in $\boldsymbol{\beta}$ in the posterior. Consequently, it is expected that the sizes of spatial correlations estimated from the raw residuals of $\hat{\boldsymbol{\delta}}$ and from the model residuals $Z - \mathrm{E}(\boldsymbol{\beta}|Z)$ will not match. In fact, this is exactly what was observed in the diagnostic considerations of the analysis. This indicates that more sophisticated noise models should be used which account for the presence of spatial correlations in fMRI data. The second comment is related to the meaningfulness of the approach itself. A kind of drawback in this approach is that the Γ field does not have a much simpler structure than $\boldsymbol{\beta}$. Usually, the statistical models aim to find simple explanations. Another issue of concern is that the neuronal field was assigned a noninformative prior. However, since the fMRI data cannot provide direct information on any neuronal field, it follows that the posterior variability of Γ is not necessarily moderate. Naturally, neuronal fields are of considerable interest, but more informative priors for Γ should be used if possible.

As explained, our continuing theme is to discuss spatial phenomena in fMRI data. Nevertheless, it is important to note here that Bayesian methods have been utilized in fMRI studies also in many other ways, such as temporal modelling of voxel time series (Kershaw *et al.*, 1999; Kershaw *et al.*, 2001). Bayesian thinking has also proved to be useful in studies of other imaging modalities, in improving the quality of reconstructed PET images, for example (Alenius & Ruotsalainen, 1997; Sastry & Carson, 1997; Kao *et al.*, 1998). These topics are beyond our scope, and therefore we shall not pursue discussion on these applications here.

3.2 Random field tests on activations

Perhaps the most popular analysis technique for detecting activated brain voxels is the statistical parametric mapping (SPM) method, which provides several significance tests for localizing haemodynamic responses. At its simplest, the approach consists of fitting a linear model (3.4) to data Y and calculating a *statistical parametric map* (SPM), the least squares estimates of all $\beta(s)$ divided by their standard errors:

$$G(s) = \frac{\hat{\beta}(s)}{\text{s.e.}(\hat{\beta}(s))} \tag{3.12}$$

The essence of the method is to examine the map G, test the significance of some of its geometrical features and base the localization on the test results.

Early approaches towards analysing SPMs are contained in Friston *et al.* (1991) and Worsley *et al.* (1992), where voxelwise values of an SPM were referred to the quantiles of the distribution of the maximum statistic of a stationary Gaussian random field. If the observed maximum of the SPM is statistically significant, the global null hypothesis is rejected and the voxels, in which the SPM score exceeds the chosen quantile, are considered to be activated. Since each voxel can be assigned a P-value from the reference distribution, the approach is called *voxel-level* inference. The main contributions of these articles were the following. First, it was observed that the inference becomes unnecessarily conservative if spatial correlations between the SPM scores are ignored. Second, the authors were able to derive a useful approximation for the tail distribution of the maximum functional.

To detect brain activations, it is natural to consider also other properties of SPMs than only intensities in individual voxels. Friston *et al.* (1994a) set a height threshold t_h for an SPM G to create a binary map and analysed the spatial extents $S_1, \ldots, S_N$ of the resulting contiguous suprathreshold patterns (or clusters)

$C_1, \ldots, C_N$. In mathematical terms, this is stated as

$$\{s \mid G(s) > t_h\} = C_1 \cup \cdots \cup C_N. \tag{3.13}$$

Poline & Mazoyer (1994a) and Poline *et al.* (1997) extended this approach by combining evidence against the null hypothesis from both the spatial extent S_i and the peak height P_i of a cluster C_i. The way how the two characteristics of a cluster can be combined in testing is not unique, and recognizing this, the authors suggested two different combination ideas. In order to find out which clusters are statistically significant, the following two basic properties of suprathreshold patterns were applied. First, the clusters C_i are approximately independent of each other, and, second, the number of clusters, N, is approximately Poisson distributed (Adler, 1981). Using these principles, the distribution of the maximum of the clusterwise statistics was derived. This kind of *cluster-level* inference requires the specification of the height threshold t_h, and hence the testing of clusters is not unique in the same way as that of voxels.

In addition to the voxel- and cluster-level methods above, Friston *et al.* (1995a) introduced inference on the *set-level*. In this case, two thresholds are needed to specify the test, a height threshold t_h for SPM and a spatial extent threshold t_s for cluster sizes. The height threshold t_h determines the clusters $C_1, \ldots, C_N$ as in the cluster test, but now clusters having smaller extent than the extent threshold t_s are discarded, and the final test statistic is then the number of clusters (M, say) that survive in the second thresholding. The significance of an observed value of M can be calculated, noting that the total count N follows a Poisson distribution, and that the distribution of the extent S_i of a suprathreshold cluster C_i is approximately known (Friston *et al.*, 1994a). If the observed count is significant, the union of the survived clusters is an estimate of the set of activated voxels. Otherwise, the estimate is the empty set.

We conclude that hypothesis tests have been proposed on three levels of spatial hierarchy for SPMs: voxel-, cluster- and set-level.

The approximations for the distributions of the test statistics under the null hypothesis are valid if it can be assumed that the SPM of interest is a reasonable approximation of a differentiable strictly stationary Gaussian random field. Clearly, the probabilistic properties of an SPM are completely characterized by its spatial autocorrelation function and its variance, if the assumption is valid. If analytical approximations of the reference distributions are not accurate enough, Monte Carlo tests can be used at the expense of longer computation time. Part of the relevant results on Gaussian fields have also been generalized to other statistic fields, such as $t-$, $F-$ and χ^2-fields (Worsley, 1994; Cao, 1999).

To achieve the desired level of regularity for an SPM, it is typically necessary to smooth the original data linearly. The width of the smoothing kernel is a compromise: the new data should be smooth enough so that the theory of smooth Gaussian random fields is applicable, and at the same time, fine small-scale details of the original data should be retained. Since smooth maps $\boldsymbol{\beta}$ remain almost unchanged in a smoothing operation, one can argue that smoothing is safe when the true activations are quite smooth. It is important to notice that a smoothing operation provides a *nonparametric* estimate of activation magnitudes in a natural way. The level of smoothing is frequently reported as the full width half maximum (FWHM) of the smoothing kernel. The consequences of linear filtering on the results are twofold. On one hand, the sensitivity of a test usually increases against activation signals having the same width as the smoothing kernel. On the other hand, the significant activations may be displaced by a few voxels and the sensitivity to detect signals having different width may decrease. This means that the choice of the kernel width will be reflected in the results. One solution is to smooth the data using several kernels and simultaneously examine the data in many spatial scales, as suggested by Poline & Mazoyer (1994b). If this strategy is adopted, it is obvious that the appropriate range of kernel widths decreases as the number of scans increases.

Lange & Zeger (1997) worked with original unsmoothed data and advocated for focused (or regional) tests on brain activation. The inferential idea in focused inference can be described in the following way. Let $\hat{\boldsymbol{\beta}}$ be an estimate for $\boldsymbol{\beta}$, Σ the covariance matrix of $\hat{\boldsymbol{\beta}}$ and $\mathcal{R}$ a set of p brain voxels. To test a hypothesis $\boldsymbol{\beta}_{\mathcal{R}} = 0$, we can employ the test statistic

$$C_{\mathcal{R}} = \hat{\boldsymbol{\beta}}_{\mathcal{R}}^T \hat{\Sigma}_{\mathcal{R}}^{-1} \hat{\boldsymbol{\beta}}_{\mathcal{R}}, \tag{3.14}$$

where $\Sigma_{\mathcal{R}}$ is the covariance matrix of $\hat{\beta}_{\mathcal{R}}$. Assuming that the noise process is multivariate Gaussian, the statistic $C_{\mathcal{R}}$ follows approximately the chi square distribution with p degrees of freedom under the null hypothesis. The null distribution is exact if the covariance structure is estimated accurately. The distributional properties of $C_{\mathcal{R}}$ can be derived analytically and there is no need to apply asymptotics of Gaussian random fields. It is also important to observe that the topological or geometrical properties of the set $\mathcal{R}$ do not play any special role in the form of the statistic $C_{\mathcal{R}}$ or in its null distribution since these properties are implicitly contained in the covariance matrix $\Sigma_{\mathcal{R}}$. In contrast, the probability structure of a statistical parametric map over a set $\mathcal{R}$ is always closely related to the geometry of $\mathcal{R}$ in a way which can be quantified only approximately.

The statistic $C_{\mathcal{R}}$ resembles the S-statistic by Worsley *et al.* (1995). By definition, S is the sum of squares over the whole SPM. It follows that S equals to the C-type statistic if zeros are put to the off-diagonal elements of the covariance matrix Σ. The null behaviour of S is more complicated than that of $C_{\mathcal{R}}$. However, an approximative solution is available. In principle, the test (3.14) is straightforward to use. However, in case $\mathcal{R}$ is large, inverting $\hat{\Sigma}_{\mathcal{R}}$ may be computationally problematic. Thus, a regional test is practical only when $\mathcal{R}$ has a moderate size.

To evaluate the performance of the localization tests, two concepts, sensitivity and specificity, are commonly used in fMRI literature. Sensitivity is synonymous to the power of a test, and accordingly, a method is called sensitive if it is able to detect the occurrence of weak activation signals. Here, the word detection means that the presence of activation *somewhere* is observed ignoring the accuracy of localization. Specificity, in turn, is not an equally well-defined concept. In informal terms, a method has strong specificity if most of the voxels of the set of estimated activations are truly activated. Friston *et al.* (1995a) note that both cluster- and set-level inference have weaker specificity properties than voxel-level inference. The lack of specificity means in practice that detected brain activity cannot be attributed to individual voxels but only to a collection of voxels. This is a fundamental feature of spatial testing and cannot be avoided. The same authors also recognized the influence of the spatial characteristics of an activation map $\boldsymbol{\beta}$ on the detecting power. For example, voxel-level inference is powerful against activation maps containing high peaks, cluster-level inference in turn detects peaks with at least moderate spatial extent, and set-level inference is powerful if the map consists of many peaks. On the basis of simulation studies, Friston *et al.* (1995a) reported that *generally* the three test procedures are hierarchically related to each other in such a way that set-level inference is more powerful than cluster-level inference and that cluster-level infererence is more powerful than voxel-level inference. The word *generally* means here that complex stimulation paradigms are commonplace and that activation maps $\boldsymbol{\beta}$ are likely to have a spatially dispersed shape in such experiments. We conclude that sensitivity and specificity are usually conflicting requirements.

In the construction of the test (3.14), geometrical features of activations were not emphasized. In fact, $C_{\mathcal{R}}$ merely measures the overall magnitude of $\hat{\boldsymbol{\beta}}_{\mathcal{R}}$. It seems likely that a $C_{\mathcal{R}}$-based test is typically less powerful than SPM tests since it is not based on likely spatial properties of activations. On the other hand, the statistic $C_{\mathcal{R}}$ has potential sensitivity to spatially completely unstructured alternatives (for example, unsmooth situations) compared with the tests for SPMs. The prevalence of such activations in real data sets is questionable, however. Lange & Zeger (1997) did not comment on the potential power of their test. The degree of specificity of the regional test depends crucially on the size of the region of interest. If $\mathcal{R}$ is a large region and the observed value of the test statistic is significant, it is not clear what voxels are responsible for the result of the test, and thus the specificity of the test is necessarily low. A benefit from restricting the scope of inference to a subset is that then the test will not be overconservative. This also increases the specificity of the test procedure. To finish the discussion, it is worth mentioning that published arguments against focused testing exist: the use of pre-specified subregions has not received wide acceptance among applied scientists, and it has been criticized as being scientifically unreasonable (Discussion in Lange & Zeger (1997)).

3.3 Contrasting frequentist and Bayesian fMRI analyses

We have overviewed two philosophically dissimilar lines of statistical reasoning, frequentist and Bayesian, for carrying out inference in this section. Since there are two disparate ways of thinking, we here clarify the situation by contrasting these fundamental mainlines in the context of fMRI.

The core of the SPM methods is the nonparametric testing of global null hypotheses. The emphasis is on ingenious hypothesis testing, that is, on a clever choice of a test statistic. The SPM tests are not based on any explicit statistical model of an activation map, and consequently, the inferential procedure does not provide any model-based estimates for activations. This kind of concentration on hypothesis testing is sometimes called the "hypothesis testing fallacy" in fMRI literature. The SPM theory cannot provide any confidence for the detected activation pattern. The point is that the random field theory can give an approximation for the probability of detecting (false) activation when no activation is present. However, there is not any useful answer how the detected set of activations is related to the true set of activations if there are truly activated voxels in the search volume.

When applying Bayesian methods, the inference is drawn from the posterior distribution of an activation map. The posterior is constructed by suggesting a prior model and a likelihood function for the activation process. The prior describes our prior beliefs in activations. The likelihood, in turn, reflects the stochastic properties of the noise in imaging. Thus, both the noise and the activation phenomenon are subject to modelling. This differentiates Bayesian and SPM analyses since in SPM only the noise is addressed for stochastic modelling.

A frequentist approach attempting to model activations may run into difficulties since there are no obvious suitable *parsimonious* parametric models for describing spatial properties of haemodynamic effects. The importance of parsimonious models in estimation is a well-known statistical principle. Generally, models with an economic structure lead to less variable estimates for the effects of interest (Altham, 1984). Historically, the lack of activation models was the original motivation for considering random field techniques. However, adopting the Bayesian paradigm, one can allow a large collection of parameters, control the effective dimension of the parameter space by choosing appropriate priors, and thereby increase the efficiency of the inference on activations. In SPM, the only means to account for any prior knowledge is the choice of the significance test. We conclude that some connections do exist between the aforementioned SPM and Bayesian proposals but that they cannot be reconciled perfectly. In general, classical significance tests and posterior probabilities cannot be related in a satisfactory manner. A discussion on this topic can be found in Bernardo *et al.* (1992).

Finally, we remark that an important difference between SPM and Bayesian analyses is the way how spatial smoothing is carried out. The linear spatial smoothing of fMRI data, which precedes SPM analyses, is not necessary in those analyses which do not utilize probabilistic results of the theory of differentiable Gaussian fields. The elegance of the Bayesian approach is that spatial smoothness can be imposed on activation maps in the prior and the data itself is left unchanged.

3.4 Comment upon Bayesian methods for fMRI data

We have reviewed several Bayesian approaches for analysing fMRI data in this section. The level of prior modelling has varied considerably among the proposed Bayesian models. Both Everitt & Bullmore (1999) and Hartvig & Jensen (2000) modelled the prior probability that a voxel is activated in the search volume. The latter authors also accounted for spatial contextuality of brain activations. Descombes *et al.* (1998b), Kornak *et al.* (1999), and Taskinen (1998) emphasized a single feature of activations, the spatial smoothness of activation intensities, which was accomplished by applying nonparametric Bayesian smoothing techniques. The approach by Hartvig (1999), in turn, led to a high-level model since the basic elements of describing the activation are (low-dimensional) spatial objects instead of single voxels. Our view is that high-level approaches may be reasonable in the context of fMRI. To defend our claim, we refer to a discussion in Friston *et al.* (1995a) where the authors consider the influence of the experimental design on activation profiles. They remark:

> *"It should be noted ... that many experiments discount functional integration and attempt to elicit activity in one area that is functionally specialised for a single sensorimotor or cognitive process. In this instance, the signal may well comprise one (or a small number of) foci."*

Thus, in certain situations useful prior information is available, and this knowledge may be structural, for example the probable number of activation foci. Since high-level approaches are ideally suited for making full use of structural prior knowledge, we conclude that high-level prior models deserve more statistical attention. We expect that neuropsychological expertise can be utilized more efficiently if structural models are applied. Following Hartvig (1999), we shall consider prior models based on point processes in the next section.

4 Bayesian modelling of spatial activation processes

In the following three sections we shall develop a new Bayesian approach for modelling spatial aspects of haemodynamic responses to controlled stimulation, propose a statistical model for noise in voxel time series and construct a computationally intensive algorithm for drawing inference from the posterior distribution of brain activations. In this section, we consider the problem of constructing a prior model which could be used to describe a certain set of prior assumptions on activations. We shall first state these assumptions and then suggest a marked point process for prior modelling.

4.1 Activation profiles

In earlier sections we concentrated on the Bayesian estimation of a single stimulation effect. Sometimes it is desirable to contrast *several* effects since relevant neuroscientific hypotheses may involve comparisons among responses to different stimulation types. In order to combine information from several effects, it is advisable to acquire all functional data during the same scanning session. Otherwise, uncontrolled contaminating factors may influence the comparability of the functional images. In practice, several (m, say) designs can be concatenated to a single design which consists of m consecutive temporal blocks:

$$\text{design } 1 \quad \text{design } 2 \quad \cdots \quad \text{design } m. \tag{4.1}$$

Let us assume that each block is an on/off-design, where the stimulus paradigm consists of two alternating states, the only difference between the blocks being that the stimulation type is different. In this case, the common temporal characteristics of the designs can be represented as a series $D = (D_t)_{t=1}^p$ of 1's and -1's where p is the number of scans in a block. We presume that the corresponding temporal pattern of the haemodynamic response $X = (X_t)_{t=1}^p$ (common to all individual designs) is known, that is, suitable estimates for the convolution weights of the equation (2.1) are available. Assuming the design is balanced, the overall mean of the temporal pattern $(X_t)_{t=1}^p$ is zero. Then, the expected value of the observed series $Y(s)$ can be modelled using two parameters for each voxel:

$$\mathrm{E}\,Y_{t+p(k-1)}(s) = \phi^{(k)}(s) + \beta^{(k)}(s)X_t \tag{4.2}$$

where $k = 1, \ldots, m$ are the blocks, $t = 1, \ldots, p$ are points of time in each block, $\phi^{(k)}(s)$ is the mean level of the resonance signal in voxel s in block k, and $\beta^{(k)}(s)$ is the stimulation effect. To study contrasted effects, sums of the form

$$\alpha(s) = \boldsymbol{\lambda}^T \beta(s) = \lambda_1\beta^{(1)}(s) + \cdots + \lambda_m\beta^{(m)}(s) \tag{4.3}$$

can be estimated in each voxel, where the choice of the *contrast vector*

$$\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_m)^T \tag{4.4}$$

reflects the scientific hypothesis to be studied. Adopting the same nomenclature as in Friston *et al.* (1995b), we shall call sums (4.3) *profiles*. To exemplify, profiles such as

$$\boldsymbol{\beta}^{(1)},\ \boldsymbol{\beta}^{(1)} - \boldsymbol{\beta}^{(2)} \text{ and } \boldsymbol{\beta}^{(1)} - (1/2)(\boldsymbol{\beta}^{(2)} + \boldsymbol{\beta}^{(3)}) \tag{4.5}$$

are contrasts that might be of interest in an fMRI study (assuming $m \geq 3$).

If contrasted effects are to be estimated, one can choose whether to assign a prior for all the effects $\boldsymbol{\beta}^{(k)} = (\beta^{(k)}(s))$, $k = 1, \ldots, m$, or whether to construct directly a prior for a profile $\boldsymbol{\alpha} = (\alpha(s))$. We suggest that if there are a few profiles or just one which are of primary neuroscientific interest, it may be advantageous to adopt the latter procedure. The reason for this is that then the parameter space is potentially smaller, which may decrease the computational burden in our MCMC calculations (Section 6). In this thesis, we shall consider the Bayesian estimation of a single contrast and thus our aim is to construct a prior model for a profile.

A profile is a measure of a functional change in signal intensity. In this thesis, we assume that functional changes typically occur only in some brain regions and that they are *clustered*. More precisely, we presume that an indicator field

$$I_{\boldsymbol{\alpha}}(s) = 1_{\{\alpha(s)\neq 0\}}(s) \tag{4.6}$$

consists of clusters of ones. We shall also assume that these clusters are *coherent*. By this we mean that intensities $\alpha(s)$ and $\alpha(u)$ have the same sign whenever s and u are contained in the same cluster. In other words, we believe that in general it is plausible to expect neighbouring voxels to respond to stimuli coherently.

Our prior model for profiles is based on what can be known about clusters *a priori*. Our view is that neuroscientific experience from previous fMRI studies can provide useful information on the locations and number of clusters. Naturally, the probable locations and counts depend on the design and also on the contrast which defines the profile. For example, the number of activation foci may be small if the objective of the experiment is to highlight the functional segregation of the human brain, as Friston *et al.* (1995a) claim. Moreover, experience from lesion studies may provide valuable information on the organization of the brain function, see Shaywitz *et al.* (1995) and a commentary to it by Rugg (1995). Also, there is information available on the spatial extent of a cluster, that is, how large a cluster is likely to be. These arguments motivate us to consider a model whose distributional properties can be expressed in terms of locations, counts and extents. To this end, we suggest the following decomposition

$$\alpha(s) = \eta_1 \kappa_1 (s - w_1) + \cdots + \eta_n \kappa_n (s - w_n), \tag{4.7}$$

for a profile where $\boldsymbol{w} = \{w_1, \ldots, w_n\}$ is a configuration of voxels (i.e. points in a discrete space), $\kappa_1, \ldots, \kappa_n$ are spatial functions, and $\eta_1, \ldots, \eta_n$ are real-valued scaling factors. We shall utilize this decomposition in the following way. First, each term in the sum (4.7) will correspond to one cluster of the profile $\boldsymbol{\alpha}$. Consequently, the number of points in $\boldsymbol{w}$ equals the number of clusters in $\boldsymbol{\alpha}$. Second, a point w_i is to represent the location of the strongest response in a cluster. We shall call these points *cluster centres*. The role of cluster centres is interesting since they can be used to localize a spatial activation pattern using a single voxel. A motivating reason for emphasizing the importance of strongest responses (or local peaks in a profile) is that in statistical parametric mapping (SPM) analyses it is a common practice to report the positions of these voxels in standardized brain coordinates. Third, the functions κ_i will be used to model the coherence and the spatial extent of clusters. We accomplish this by using functions which are nonnegative and are null outside some neighbourhood of the origin. For consistency, we also require that functions κ_i attain the maximum at the origin. Then, the strongest responses are located at the points w_i. Finally, the terms η_i are used to scale the magnitudes of κ_i. Clearly, these scaling terms also determine the signs of clusters.

We shall first consider a model for κ. We describe haemodynamic effects around a cluster centre by the model

$$\kappa(s) = B_p(s) + \sum_{j=1}^{k} \zeta_j B_d(s - v_j), \tag{4.8}$$

which is a sum of a *parent bell* B_p and *daughter bells* B_d scaled by positive ζ_j and shifted by v_j. This construction is a step towards nonparametric Bayesian modelling compared with the approach by Hartvig (1999). The purpose of this formulation is to introduce a structure inside clusters which could be used to control some properties of clusters in the prior. We shall treat the bells B_p and B_d as fixed spatial objects, whereas the number k of the daughter bells and parameters ζ_j and v_j will be assigned priors. We assume that B_p and B_d are non-negative spatial functions which are centred at the origin and have the same maximum. In particular, the two bells could be isotropic and normalized Gaussian bells with possibly different spatial width. When using Gaussian bells, we shall cut the tails of the bells in such a way that bell values less than 5% of the bell maximum are rounded to zero. Clearly, this kind of thresholding determines the spatial extent or *radius* of the bell.

We explain now how the choice of k, ζ_j and v_j influences the shape and extent of κ. First, the spatial extent of a cluster is related to parameters v_j. If all v_j are small, the resulting cluster has small extent. Second, the shape of a cluster depends on the scaling factors ζ_j: if all the scaling terms ζ_j have small positive values, the shape of κ resembles that of the parent bell B_p. In particular, κ has a peak near the origin. We note that the spatial extent is smallest when all ζ_j vanish. Third, the number k of daughter bells describes the complexity of a cluster. Figure 4.1 illustrates our construction and exemplifies effects that the heights ζ_j and the spatial width of a Gaussian daughter bell B_d can have on the intensities $\kappa(s)$.

4.2 Prior modelling of profiles

To assign probabilistic properties to the sum (4.7), we treat the configuration $\boldsymbol{w}$, functions $\kappa_1, \dots, \kappa_n$ and scaling terms $\eta_1, \dots, \eta_n$ stochastically independent of each other *a priori*. The independence property is justified in our context since we do not expect any general relationships between the locations, spatial extents and magnitudes of clusters. We shall complete the construction of the prior by suggesting priors for $\boldsymbol{w}$, κ and η. We emphasize here that our prior has a hierarchical structure: the prior of $\boldsymbol{w}$ is a model for the locations of clusters and the conditional prior $\boldsymbol{\alpha} \mid \boldsymbol{w}$ is a model for stimulus-related effects, *given* the locations.

We shall assume that the number of daughters k is Poisson distributed with some mean β_d. Then, β_d represents our uncertainty on the complexity of a cluster. We denote by h_ζ the continuous prior density from which the scaling parameters $\zeta_1, \dots, \zeta_k$ are drawn *independently*. We suggest that h_ζ should be supported in the unit interval since our purpose is to model clusters with one main peak. The most uninformative choice is to let h_ζ follow the uniform distribution on the unit interval. With this extreme choice, the maximum of κ is not necessarily attained at the origin and the interpretation of the cluster centres w_i will be unclear. However, we can increase the prior probability that the parent bell forms the highest peak of a cluster by preferring small values for scaling factors ζ_j.

We shall denote by h_v the prior density from which the centres $v_1, \dots, v_k$ of the daughter bells are drawn independently of each other. Here, the reference measure of the density h_v is the counting measure ν in $\mathcal{U}$, the voxellated version of $\mathbb{R}^2$ (or $\mathbb{R}^3$ if the data consist of more than one brain slice; the set of brain voxels $\mathcal{S}$ is

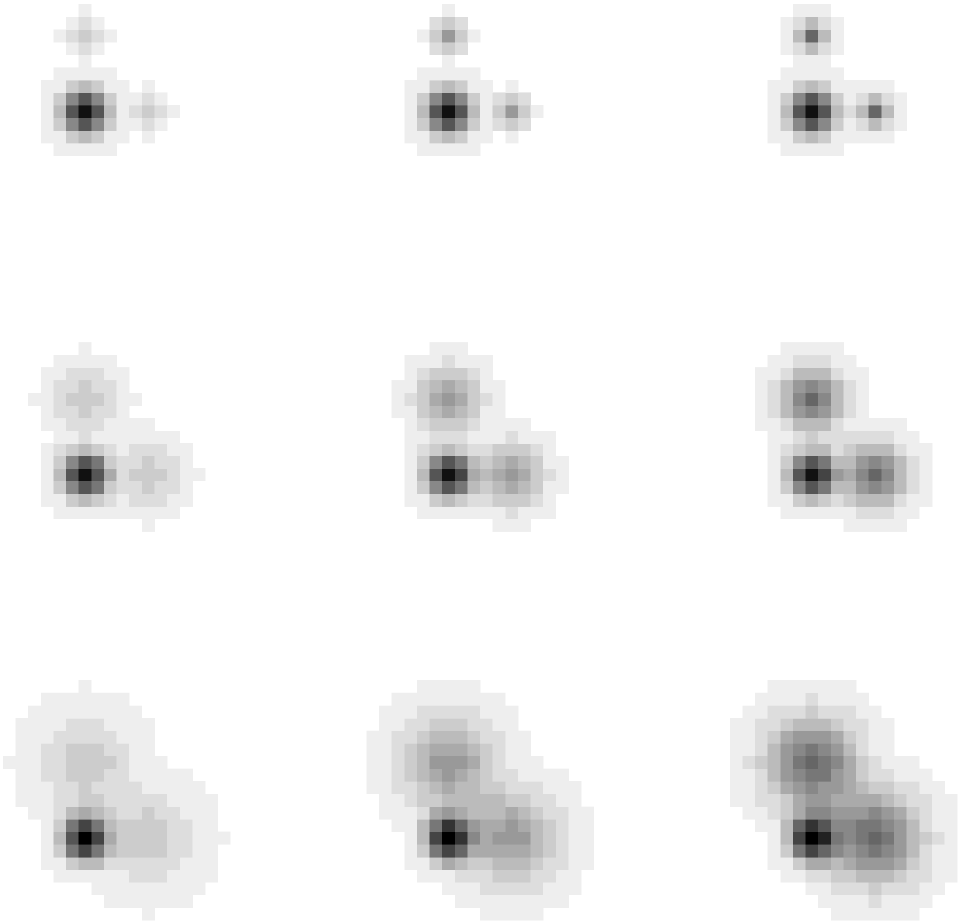

Figure 4.1 The intensities of nine different κ which all consist of one parent bell and two daughter bells ($k = 2$). The common heights of the daughter bells are 0.2 (left column), 0.4 (middle column) and 0.6 (right column). The radii of the daughter bells are 2 (top row), 4 (middle row) and 6 (bottom row) voxels. The daughter bells have the same relative position in all nine spatial patterns. The radius of the parent bell is three voxels. All the patterns are magnified in order to facilitate the visual inspection. In reality, the voxels have smaller size.

regarded as a subset of $\mathcal{U}$). We have now suggested a model for κ using local point configurations

$$\boldsymbol{d} = \{(v_1, \zeta_1), \ldots, (v_k, \zeta_k)\}, \tag{4.9}$$

which we shall call *daughter configurations*. It follows from the construction that $\boldsymbol{d}$ is an inhomogeneous Poisson point process with intensity measure $\beta_d h_v h_\zeta \, \mathrm{d}\,(\nu \times m_1)$. Clearly, the daughter point process has density

$$f_d(\boldsymbol{d}_i) = e^{1-\beta_d}\, \beta_d^{n(\boldsymbol{d}_i)} \tag{4.10}$$

with respect to a *unit* rate Poisson process λ_d in $\mathcal{U} \times \mathbb{R}^+$ whose intensity function is $h_v(v)h_\zeta(\zeta)$ with respect to $\nu \times m_1$. An advantage of the construction based on the sum (4.8) is that if the parent and daughter bells are smooth, there is no need for controlling the smoothness of the resulting profile in any way. In other words, the implicit prior requirement about the smoothness of profiles is automatically satisfied.

The role of the parameters η_i is to scale the magnitude of clusters. In fact, in the model (4.7) we treat separately the strength of activation (using η) and the shape of the local activation surface (using κ). Hartvig (1999) applied a prior which penalizes Gaussian bells having magnitude close to zero. Our choice is to apply an informative prior on the counts $n(\boldsymbol{w})$ rather than on the heights. An uninformative prior density h_η for η (with respect to Lebesgue measure) could be the normal distribution with mean zero and variance suitably large.

We shall construct a prior model for the configuration $\boldsymbol{w}$ of cluster centres so that the prior can be used to control the locations and the number of the centres. In practice, the prior expectations on the locations of centres concern some anatomically defined regions in the brain. Let us suppose $\mathcal{R}$ is a brain region (or a union of regions) which is expected to exhibit stimulus-related activation. This belief can be quantified by introducing a density $h(s)$ in $\mathcal{S}$ (with respect to ν) which will represent the prior density of the location of a random point in $\boldsymbol{w}$. The simplest way to describe prior beliefs is to assume that h is constant in $\mathcal{R}$ and $\mathcal{S} \setminus \mathcal{R}$. Adopting this strategy, it is sufficient to specify the prior mean of

$$n(\boldsymbol{w} \cap \mathcal{R})/n(\boldsymbol{w}), \tag{4.11}$$

the portion of centres falling into $\mathcal{R}$. If there is no obvious reason to favour some brain regions, one can always use the ignorance choice $h(s) \propto 1$. The prior number of the centres can be modelled by introducing a discrete distribution ψ for the counts $n(\boldsymbol{w})$. In this context, a relevant aspect of the total count which could be modelled is a probable upper bound of $n(\boldsymbol{w})$. In other words, we specify how distributed the indicator field $I_{\boldsymbol{\alpha}}$ can be *a priori*. Using the density h and the distribution ψ, we can construct a simple tentative point process model for the centres by sampling the number of them from ψ and then drawing equally many centres independently from h. The resulting process has density

$$\boldsymbol{w} \mapsto e\, n(\boldsymbol{w})!\, \psi(n(\boldsymbol{w})) \tag{4.12}$$

with respect to an inhomogeneous unit rate Poisson process with intensity h.

In addition to locations and counts of centres, there is one more aspect of prior knowledge which can be used to decrease prior uncertainty on the configuration $\boldsymbol{w}$. In the context of medical imaging, it may be natural to expect that the inter-distance of two centres cannot be arbitrarily small. In particular, one may assume that the distances between centres typically exceed some physical distance ρ. Neuropsychological expertise can provide approximations for the probable intercentre distances. From the modelling point of view, this means that the centres inhibit each other *a priori*. To take inhibition into account, we modify the density (4.12) and weight it by multiplying the density by a product $\prod_{i<j} \Psi(w_i, w_j)$ of interaction terms, where Ψ is a spatial interaction function. We shall use the following interaction function:

$$\Psi(w_i, w_j) = \begin{cases} (\|w_i - w_j\|/\rho)^p & \|w_i - w_j\| \le \rho \\ 1 & \|w_i - w_j\| > \rho. \end{cases} \tag{4.13}$$

Here, ρ and p are positive parameters which determine the size of inhibition as a function of distance. The roles of these parameters are quite clear: as p increases to infinity, all pairwise distances less than ρ become increasingly heavily penalized, whereas distances more than ρ will not be penalized. The special case $p = \infty$ leads to the hard-core process. Generally, too large values of ρ limit the collection of possible profiles, whereas some prior information will be ignored if very small values of ρ are used. Our opinion is that in applications it can be more natural to use an inhibition technique based on Euclidean distances than one based on the Kullback J-divergence measure, see (3.9). For the background on point processes with pairwise interaction and finite point processes in general, we refer to van Lieshout (2000), Stoyan *et al.* (1995) and Daley & Vere-Jones (1988).

A slight difficulty is encountered when accounting for information on interdistances. The modified version of the density of the centre process does not have the desired stochastic properties mentioned above. The distribution of a randomly chosen centre will not any longer follow $h(s)$, and a similar remark applies to ψ. This problem can be solved by replacing h and ψ in (4.12) by corrected densities h^* and ψ^*. Analytical solutions for this correction operation do not seem to be available but we can get insight into the size of corrections by simulating the centre process. The simulations can be carried out using an MCMC algorithm by Geyer & Møller (1994). The details can be found in Appendix A.2. This leads to the prior density

$$f_{\boldsymbol{w}}(\boldsymbol{w}) \propto n(\boldsymbol{w})!\, \psi^*(n(\boldsymbol{w})) \prod_{i<j} \Psi(w_i, w_j) \tag{4.14}$$

with respect to the Poisson process with intensity h^*.

In our prior model formulation, a profile is parametrized by a *marked point configuration* $\boldsymbol{x} = \{x_1, x_2, \ldots, x_n\}$ where each point

$$x_i = \big(w_i; (\eta_i, \boldsymbol{d}_i)\big) \tag{4.15}$$

consists of two components: a location w_i and a mark $(\eta_i, \boldsymbol{d}_i)$. We show here that the process $\boldsymbol{x}$ can be regarded as a Gibbs point process, that is, it has a density with respect to a Poisson process. To this end, we introduce a new unit measure $\lambda_{w,\eta}$ in $\mathcal{M} = \mathcal{S} \times \mathbb{R}$ for which

$$\mathrm{d}\,\lambda_{w,\eta} = h^*(s) h_\eta(\eta)\ \mathrm{d}\,(\nu \times m_1). \tag{4.16}$$

Now, let μ be a Poisson process in the product space of $\mathcal{M}$ and the exponential space $\mathcal{D} = \Omega(\mathcal{U} \times \mathbb{R}^+)$ equipped with an intensity measure $\varphi = \lambda_{w,\eta} \times \lambda_d$. Then, the Gibbs process whose density with respect to μ is

$$\pi(\boldsymbol{x}) = f_{\boldsymbol{w}}(\boldsymbol{w}) \prod_{i=1}^{n(\boldsymbol{x})} f_d(\boldsymbol{d}_i) \tag{4.17}$$

$$\propto e^{n(\boldsymbol{x})(1-\beta_d)} n(\boldsymbol{x})!\ \psi^*(n(\boldsymbol{x}))\ \beta_d^{\sum_i n(\boldsymbol{d}_i)} \prod_{k<j} \Psi(w_k, w_j), \tag{4.18}$$

is actually equivalent to our cluster process. We note that a sample from μ is a random configuration $\boldsymbol{z} = \{z_1, z_2, \ldots, z_n\}$ where n, the number of points, follows the Poisson distribution with mean one, the locations w_i of the points $z_i = \big(w_i; (\eta_i, d_i)\big)$ are independent samples from h^* and the points are independently marked.

We emphasize that the proposed interaction between two clusters, x_i and x_j, depends only on the intercentre distance $\|w_i - w_j\|$ and not on the actual shape of the two clusters. Consequently, a change in the model of κ does not have any effect on the likely number of clusters and their probable locations if the centre process $\boldsymbol{w}$ is kept unchanged. In other words, we have separated intrinsic properties of clusters and spatial interaction between clusters from each other.

5 Modelling of the noise processes

In the previous section we introduced a prior distribution for profiles as a first step towards making posterior inference on haemodynamic effects induced by controlled stimulation. In the following, we establish a probability model for data, given the activation profile. Our objective is to consider noise models for fMRI data which would be practical in computations while being adequate descriptions of the distributional properties of errors in the observations.

5.1 Likelihood of an activation profile

We continue the discussion in subsection 4.1. In short, we assume that a concatenated design (4.1) has been applied and fMRI data Y have been acquired. Then, the data Y can be modelled as

$$Y(s) = \left[\boldsymbol{X}^{(\phi)}\ \boldsymbol{X}^{(\beta)}\right] \begin{bmatrix} \phi(s) \\ \beta(s) \end{bmatrix} + \delta(s), \tag{5.1}$$

where $\boldsymbol{X} = \left[\boldsymbol{X}^{(\phi)}\ \boldsymbol{X}^{(\beta)}\right]$ is the matrix of explanatory series corresponding to block effects and stimulation effects; the exact form of $\boldsymbol{X}$ can be found in Appendix A.1. We also assume that a contrast vector $\boldsymbol{\lambda} = (\lambda_1, \dots, \lambda_m)$ has been chosen and aim to find a likelihood function for the corresponding profile $\boldsymbol{\alpha} = \boldsymbol{\lambda}^T \boldsymbol{\beta}$. To begin with, we note that if the noise $\boldsymbol{\delta} = (\delta(s))$ is a Gaussian process and if, moreover, the space-time covariance structure of $\boldsymbol{\delta}$ is known, all information in data on $\boldsymbol{\alpha}$ is carried by its generalized least squares estimate (GLS) $\hat{\boldsymbol{\alpha}} = \boldsymbol{\lambda}^T \hat{\boldsymbol{\beta}}$. Therefore, once a model for $\boldsymbol{\delta}$ is fitted, we may compress the original spatio-temporal data Y to *purely spatial* data $\boldsymbol{Z} = \hat{\boldsymbol{\alpha}}$ without any loss of information. In the following, we shall consider a spatio-temporal Gaussian model for $\boldsymbol{\delta}$ and derive a formula for the conditional density of $\boldsymbol{Z} \mid \boldsymbol{\alpha}$ which provides us with a likelihood function for $\boldsymbol{\alpha}$.

When constructing correlated models in spatio-temporal problems, a straightforward possibility is to treat spatial and temporal covariances separately. Let us assume that the covariances of $\boldsymbol{\delta}$ can be represented as products

$$\operatorname{cov}(\delta_t(s), \delta_k(u)) = w_{su} v_{tk}, \tag{5.2}$$

where $W = (w_{su})$ is a spatial covariance matrix and $V = (v_{tk})$ is a temporal correlation matrix. Then, if we use notation $\tilde{\boldsymbol{\lambda}}^T = [0 \ \boldsymbol{\lambda}^T]$, the compression can be carried out in a voxelwise manner:

$$Z(s) = \tilde{\boldsymbol{\lambda}}^T (\boldsymbol{X}^T V^{-1} \boldsymbol{X})^{-1} \boldsymbol{X}^T V^{-1} Y(s), \tag{5.3}$$

as is shown in detail in Appendix A.1. Interestingly, the compression is independent of the choice of W. The compressed data $\boldsymbol{Z}$ is actually a sum $Z(s) = \alpha(s) + \epsilon(s)$, where $\boldsymbol{\epsilon} = (\epsilon(s))$ is a zero mean process. The relation between $\boldsymbol{\epsilon}$ and $\boldsymbol{\delta}$ is expressed in the identity

$$\epsilon(s) = \tilde{\boldsymbol{\lambda}}^T (\boldsymbol{X}^T V^{-1} \boldsymbol{X})^{-1} \boldsymbol{X}^T V^{-1} \delta(s), \tag{5.4}$$

and thus there is a simple relationship between the second-order moments of $\boldsymbol{\epsilon}$ and those of $\boldsymbol{\delta}$. In fact,

$$\operatorname{cor}(\epsilon(s), \epsilon(u)) = w_{su} / \sqrt{w_{ss} w_{uu}}, \tag{5.5}$$

$$\operatorname{var}(\epsilon(s)) = w_{ss} \tilde{\boldsymbol{\lambda}}^T (\boldsymbol{X}^T V^{-1} \boldsymbol{X})^{-1} \tilde{\boldsymbol{\lambda}}. \tag{5.6}$$

It follows from the separability assumption that the spatial correlation structure of $\boldsymbol{\epsilon}$ and each δ_t are equal. Also, the variances of $\boldsymbol{\epsilon}$ and each δ_t are proportional. To model spatial correlation, we would like to use a spatial process whose density can easily be expressed in a closed form. The reason for this is that we must be able to handle the posterior density of $\boldsymbol{\alpha}$, and to this end, the likelihood of $\boldsymbol{\alpha}$ must not involve complicated expressions, such as inverses of large correlation matrices. We choose to model $\boldsymbol{\epsilon}$ using a Gaussian conditional autoregressive process (CAR) (Besag, 1974; Ripley, 1981). In mathematical terms, the joint density function of $\boldsymbol{\epsilon}$ will be proportional to

$$\exp\Big(-\frac{1}{2}\big(\sum_s \varrho(s)^2 \epsilon(s)^2 + \sum_l \tau_l \sum_{s-u=l} \varrho(s)\varrho(u)\epsilon(s)\epsilon(u)\big)\Big), \tag{5.7}$$

where $\boldsymbol{\tau} = (\tau_l)$ is a spatial interaction parameter vector and $\boldsymbol{\varrho} = (\varrho(s))$ is a positive scaling (or precision) parameter. Here, $\boldsymbol{\tau}$ determines the spatial correlation structure of $\boldsymbol{\epsilon}$. The spatial parameter $\boldsymbol{\varrho}$ adjusts the variance of the $\boldsymbol{\epsilon}$ field, and using it we may model possible heteroskedasticity of the $\boldsymbol{\epsilon}$ process.

Formula (5.7) defines a valid density if the quadratic form in the parenthesis is negative definite. This imposes a restriction on the values that $\boldsymbol{\tau}$ can take. A sufficient condition is that

$$\sum_l |\tau_l| < 1 \quad \text{and} \quad \tau_l = \tau_{-l} \tag{5.8}$$

for all lags l (Ripley, 1981). In principle, the components of $\boldsymbol{\tau}$ may take both positive and negative values. However, all the variates $\epsilon(s)$ are positively associated if $\boldsymbol{\tau}$ is negative. Consequently, the fitted values $\hat{\tau}_l$ will typically be negative in our context since usually positive spatial autocorrelations are to be modelled. Clearly, the choice $\boldsymbol{\tau} = 0$ corresponds to spatial independence.

The application of the autoregressive model above means that we expect the spatial dependence structure to be nearly homogeneous over the search volume, i.e. independent of location. This follows from the formula for the density of the process since the strength of interaction between each pair $\epsilon(s)$ and $\epsilon(s+l)$ is governed by τ_l which does not depend on s. The model (5.7) is useful if it provides a suitable fit when using only a few nonzero interaction terms τ_l. Naturally, these terms correspond to small spatial lags l.

As to the temporal correlation matrix V, we do not impose any model assumptions on it, except we presume that $\boldsymbol{\delta}$ is stationary in time, that is, $v_{tk} \equiv v_{t-k}$. Since the number of voxels in the search volume is often high compared with the number of scans T, the use of parsimonious temporal models is not always necessary when estimating V. However, autoregressive (AR) processes frequently used in time series analysis provide a natural model family for V if a parametric approach seems appropriate.

Assuming $\boldsymbol{\varrho}$ and $\boldsymbol{\tau}$ are known, the likelihood function for $\boldsymbol{\alpha}$ has the form

$$\begin{aligned} L(\boldsymbol{\alpha} \mid \boldsymbol{Z}) \propto \exp & \left(-\frac{1}{2} \sum_s \varrho(s)^2 (z(s) - \alpha(s))^2 \right) \\ & \times \exp \left(-\frac{1}{2} \sum_l \tau_l \sum_{s-u=l} \varrho(s)\varrho(u)(z(s) - \alpha(s))(z(u) - \alpha(u)) \right). \end{aligned} \tag{5.9}$$

Since a profile $\boldsymbol{\alpha}$ is parametrized by a configuration $\boldsymbol{x}$, this leads to a likelihood for configurations, and we shall also use the notation $L(\boldsymbol{x} \mid \boldsymbol{Z})$ along with $L(\boldsymbol{\alpha} \mid \boldsymbol{Z})$, which should not cause any misunderstandings in the following sections.

As discussed in Section 2, measurements from brain haemodynamics are contaminated in several ways. It is not our purpose to decompose the noise and model some of the noise components separately. This would lead to a detailed description of the noise mechanism, but is not of primary interest. Instead, we confine ourselves to modelling the net effect of the different sources of the noise.

5.2 Estimation of spatial interaction and the precision parameters

To use the likelihood function (5.9) in posterior analysis, we estimate $\boldsymbol{\tau}$ and $\boldsymbol{\varrho}$ and substitute the estimates into the posterior. The estimation errors will have a negligible influence if the temporal degrees of freedom $T - 2m$ is large enough. We start the estimation procedure by fitting the temporal correlation matrix V and the spatial interaction parameter $\boldsymbol{\tau}$. Using the ordinary least squares residuals

$$\hat{\delta}(s) = (I - \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T)y(s), \tag{5.10}$$

a nonparametric estimate of V can be calculated by pooling all voxelwise information on temporal dependence. This leads to the estimate

$$\hat{v}_d = \frac{1}{|S|}\sum_s \frac{\sum_{t-k=d}\hat{\delta}_t(s)\hat{\delta}_k(s)}{\sum_t \hat{\delta}_t(s)^2} \tag{5.11}$$

for the autocorrelation at a time lag d. The natural order for estimating the model parameters $\boldsymbol{\tau}$ and $\boldsymbol{\varrho}$ is to estimate $\boldsymbol{\tau}$ first since it determines the spatial correlation structure. Once $\boldsymbol{\tau}$ is estimated, the scaling terms $\boldsymbol{\varrho}$ can be used to adjust the voxelwise variances of the CAR model to match with the voxelwise variability of the noise in the data. We suggest that the interaction parameter could be estimated by finding a value for $\boldsymbol{\tau}$ for which the mean spatial correlation (MSC) of the model (5.7)

$$\mathrm{MSC}(l) = \frac{1}{N_l}\sum_{s-u=l}\mathrm{cor}(\epsilon(s), \epsilon(u)) = \frac{1}{N_l}\sum_{s-u=l}\mathrm{cor}(\delta_t(s), \delta_t(u)) \tag{5.12}$$

matches suitably well with the empirical mean correlations

$$\widehat{\mathrm{MSC}}(l) = \frac{1}{N_l}\sum_{s-u=l}\frac{\sum_t \hat{\delta}_t(s)\hat{\delta}_t(u)}{\sqrt{\sum_t \hat{\delta}_t(s)^2\sum_t \hat{\delta}_t(u)^2}}. \tag{5.13}$$

Here, N_l is the number of pairs (s, u) such that $s - u = l$. Simulating the model (5.7) using several values for the strength of interaction, one can find the range of plausible values of $\boldsymbol{\tau}$. At this point, the scaling parameters of the model (5.7) can be chosen arbitrarily ($\boldsymbol{\varrho} \equiv 1$, say) since they do not have any effect on the spatial correlations. It is usually sufficient to consider only the first few spatial lags l since spatial autocorrelations tend to decay to zero for large l.

The variances of $\boldsymbol{\epsilon}$ and the precision parameters are related by

$$\varrho(s)^2 = \omega(s)^2/\,\mathrm{var}(\epsilon(s)), \tag{5.14}$$

where $\omega(s)^2$ is the variance of the unnormalized density

$$\exp\Big(-\frac{1}{2}(\sum_s \epsilon(s)^2 + \sum_l \tau_l \sum_{s-u=l} \epsilon(s)\epsilon(u))\Big). \tag{5.15}$$

In order to fit the precision parameters $\varrho(s)$, it is necessary to estimate the variance of $\epsilon(s)$ from the data Y and also the variance $\omega(s)^2$ of the model (5.15), where $\boldsymbol{\tau}$ is replaced by its estimate. Since the voxelwise variances of $Y_t(s)$ can be estimated by

$$\widehat{\mathrm{var}}(Y_t(s)|\mu, \beta) = \sum_{t=1}^{T} \hat{\delta}_t(s)^2/(T-2m), \tag{5.16}$$

we have

$$\widehat{\mathrm{var}}(\epsilon(s)) = \frac{1}{T-2m} \sum_{t=1}^{T} \hat{\delta}_t(s)^2 \; \tilde{\boldsymbol{\lambda}}^T (\boldsymbol{X}^T \hat{V}^{-1} \boldsymbol{X})^{-1} \tilde{\boldsymbol{\lambda}}. \tag{5.17}$$

The unconditional variances of CAR fields do not have useful analytical expressions. Therefore, $\omega(s)^2$ has to be calculated for each s using Monte Carlo methods. Since the autoregressive model (5.15) possesses the following simple formulae for conditional mean and variance,

$$\mathrm{E}(\epsilon(s)|\epsilon(-s)) = -(1/\varrho(s)) \sum_u \tau_{s-u} \varrho(u)\epsilon(u) \tag{5.18}$$

$$\mathrm{var}(\epsilon(s)|\epsilon(-s)) = 1/\varrho(s)^2, \tag{5.19}$$

it is straightforward to apply Gibbs sampling (Gilks *et al.*, 1996). The Monte Carlo estimate of $\omega(s)^2$ is then the empirical variance of the sample.

In fully Bayesian analyses, it is a rule to assign priors to all unknown nuisance parameters and thereby account for the (prior) uncertainty in them. In this section, we have chosen a different route since we have suggested the estimation of the noise parameters directly from the data. In other words, we have adopted an empirical Bayes approach. We shall come back to this topic later in subsection 9.2, where we discuss possibilities for fully Bayesian strategies.

6 Markov Chain Monte Carlo sampling

We suggested in Sections 4 and 5 that inference on a profile $\boldsymbol{\alpha}$ could be based on the posterior distribution of the profile, which is the conditional distribution of $\boldsymbol{\alpha}$ given the compressed data $\boldsymbol{Z}$. The density p of the posterior (with respect to μ) is proportional to the product of the likelihood function and the prior density:

$$p(\boldsymbol{x} \mid \boldsymbol{Z}) \propto L(\boldsymbol{x} \mid \boldsymbol{Z})\, \pi(\boldsymbol{x}). \tag{6.1}$$

We recall that μ is the marked Poisson process constructed in Section 4. For convenience, we suppress the dependence on $\boldsymbol{Z}$ and use the notation $p(\boldsymbol{x})$ and $L(\boldsymbol{x})$. In order to draw inferential conclusions, it is necessary to be able to calculate posterior probabilities of events of interest or, more generally, compute posterior means of functionals of marked point configurations $\boldsymbol{x}$. Then, the posterior inference is reduced to the calculation of integrals of interest with respect to the posterior measure. What makes this a nontrivial task is that in our case the posterior distribution has no analytical form simple enough for carrying out integration analytically or by applying standard numerical methods. The utilization of classical Monte Carlo methods is also problematic since there is no apparent direct technique available to obtain samples from the posterior. A solution for drawing samples is to apply Markov Chain Monte Carlo (MCMC) methods (Gilks *et al.*, 1996). Using MCMC, realizations $\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}, \ldots$ are simulated by constructing a stochastic transition mechanism in the parameter space. The transitions generate a Markov chain that converges to the target distribution.

In our point process approach, the dimension of the parameter space varies with the size n of the configuration $\boldsymbol{x} = \{x_1, \ldots, x_n\}$. Therefore, we cannot simply apply sampling techniques designed for fixed dimensional probability distributions. Green (1995) devised a general reversible jump MCMC algorithm for simulating distributions in which the dimension is random. Geyer & Møller (1994) focused on simulating finite Gibbs point processes in the *exponential* space (see Carter & Prenter (1972)). Both methods are suitable for handling variable dimensional problems. Here, we shall apply the latter technique for sampling our posterior $p(\boldsymbol{x})$.

6.1 MCMC sampling of the cluster posterior

To generate a Markov chain with our posterior p as the equilibrium distribution, we propose the following five basic transition types: updating of all the heights ζ of daughter bells of a random cluster, updating of the scaling factor η of a random cluster, displacement of a random cluster, removal/insertion of a random daughter from a random cluster and, finally, removal/insertion of a random cluster. Then, the actual transition measure can be constructed as a mixture of these five move types. In this thesis, we have been using the same selection probability 0.20 for all basic moves although other mixtures could also be used.

In Metropolis-Hastings sampling, an updating rule consists of a proposal step and the corresponding acceptance probability. The acceptance probabilities are always calculated as the minimum of a *Metropolis-Hastings ratio* (MHR) and one to ensure the convergence of the chain to the target distribution. In the following, we shall introduce several proposal distributions and derive formulas for the related Metropolis-Hastings ratios. The convergence of the sampler is discussed in Appendix A.4.

To update heights of daughter bells of a random cluster, we use proposals ζ_j^* for which $\log \zeta_j^*$ has a uniform distribution with mean $\log \zeta_j$. Then, the Metropolis-Hastings ratio is

$$MHR_\zeta = \frac{L(\boldsymbol{x}^*)}{L(\boldsymbol{x})} \prod_{j=1}^{k} \frac{h_\zeta(\zeta_j^*)}{h_\zeta(\zeta_j)} \prod_{j=1}^{k} \frac{\zeta_j^*}{\zeta_j}. \tag{6.2}$$

To propose a new scaling factor η^* for a random cluster, we sample a normal distribution with the present value η as the mean. In this case,

$$MHR_\eta = \frac{L(\boldsymbol{x}^*)}{L(\boldsymbol{x})} \frac{h_\eta(\eta^*)}{h_\eta(\eta)}. \tag{6.3}$$

Our third transition type is to move a randomly chosen cluster while keeping its other characteristics unchanged. We shall use a proposal distribution which favours small moves from the present location w_i to a new location w_i^*. We apply a distribution which has the density

$$q_{\mathcal{S}}(w_i^*|w_i) \propto 1_{\mathcal{S}\cap\mathcal{D}_i}(w_i^*) \exp(-r\|w_i^* - w_i\|^2) \tag{6.4}$$

with respect to the counting measure ν. Here, $\mathcal{D}_i = \{w_i^* \mid \|w_i^* - w_i\| \leq R\}$ is a disc centred at w_i, R is a fixed positive radius of $\mathcal{D}_i$ and r is a tuning parameter of this proposal technique. Large positive values of r encourage small moves from w_i to w_i^*. The density $q_{\mathcal{S}}$ is easily seen to be a conditional distribution of

$$q(w_i^*|w_i) \propto 1_{\mathcal{D}_i}(w_i^*) \exp(-r\|w_i - w_i^*\|^2) \tag{6.5}$$

conditioned on the event $\{w_i^* \in \mathcal{S}\}$. It follows that the identity

$$q_{\mathcal{S}}(w_i|w_i^*)/q_{\mathcal{S}}(w_i^*|w_i) = g(w_i)/g(w_i^*) \tag{6.6}$$

holds, where $g(w_i)$ is the probability that an unconditional realization belongs to $\mathcal{S}$ when the current location is w_i. If x_i is the cluster to be moved, the acceptance ratio is

$$MHR_w = \frac{L(\boldsymbol{x}^*)}{L(\boldsymbol{x})} \frac{h^*(w_i^*)}{h^*(w_i)} \frac{g(w_i)}{g(w_i^*)} \prod_{j \neq i} \frac{\Psi(w_i^*, w_j)}{\Psi(w_i, w_j)}. \tag{6.7}$$

In this thesis, we have confined ourselves to use only the choice $r = 0$.

The transition for changing the number of clusters is the most complicated transition type in our sampler. We employ here the updating mechanism by Geyer & Møller (1994), in which one new cluster is inserted to the current configuration or a cluster is removed from the configuration. We use the following special case of their general sampling algorithm. To update the configuration, we first make a random decision whether to propose an insertion of a new cluster or to delete an existing one, both choices being equally likely. If a decision of inserting a cluster is made, we propose a cluster ξ from a birth density $b(\xi|\,\boldsymbol{x})$ (with respect to intensity measure φ) which depends on the present configuration $\boldsymbol{x}$. If we choose to delete a cluster, we select randomly one cluster from $\boldsymbol{x}$ and propose that cluster to be removed. The acceptance ratio for inserting a cluster is

$$MHR_{c,i} = \frac{p(\boldsymbol{x}^*)}{p(\boldsymbol{x})} \frac{1/(n+1)}{b(\xi|\,\boldsymbol{x})}, \tag{6.8}$$

and for deleting a cluster

$$MHR_{c,r} = \frac{p(\boldsymbol{x}^*)}{p(\boldsymbol{x})} \frac{b(\xi|\,\boldsymbol{x}^*)}{1/n}, \tag{6.9}$$

where $n = n(\boldsymbol{x})$ and $\boldsymbol{x}^*$ is the proposed configuration. If there are no clusters ($\boldsymbol{x} = \emptyset$) and a removal is proposed, we do nothing.

The construction of a suitable birth density $b(\xi|\,\boldsymbol{x})$ requires some thought since a cluster ξ is a multi-dimensional object. It is difficult to suggest a proposal mechanism which possesses a convenient analytical expression and, at the same time, generates proposals that will be accepted reasonably often. We shall apply here a *sequential proposing* technique following Hartvig (1999). The sampling ideas in Hartvig (1999) are not directly applicable here since we have a different parameter space. We have modified the sequential technique and present here a related sampling method. Let $\boldsymbol{x} = \{x_1, \dots, x_n\}$ be the current configuration, where

$x_i = (w_i; (\eta_i, \boldsymbol{d}_i))$, and let $\xi = (w; (\eta, \boldsymbol{d}))$ be a new cluster which is to be proposed. Then, sequentiality means that we first propose $\boldsymbol{d}$, then w and finally η.

We apply a simple proposal mechanism for daughter configurations

$$\boldsymbol{d} = \{(\zeta_1, v_1), \dots, (\zeta_k, v_k)\} \tag{6.10}$$

and propose them from a Poisson process whose intensity measure is

$$\beta_d h_\zeta(\zeta) h_v(v)\; d(m_1 \times \nu). \tag{6.11}$$

We recall that β_d is the expected prior number of daughters in a cluster. The density of this proposal distribution with respect to λ_d is simply

$$q(\boldsymbol{d}) = e^{1-\beta_d} \beta_d^{n(\boldsymbol{d})} \tag{6.12}$$

and it does not depend on the present configuration $\boldsymbol{x}$.

A natural method to propose a new location w is to utilize conditional posterior distributions. However, since η has not been proposed yet, we have to replace this scaling factor by some constant $\tilde{\eta}$ and draw a sample from the conditional posterior distribution $w|\,\boldsymbol{d}, \boldsymbol{x}, \tilde{\eta}$. The density of this distribution with respect to ν is proportional to

$$\begin{aligned} p(\boldsymbol{x} \cup (w; (\tilde{\eta}, \boldsymbol{d})))h^*(w) &\propto L(\boldsymbol{x} \cup (w; (\tilde{\eta}, \boldsymbol{d})))\pi(\boldsymbol{x} \cup (w; (\tilde{\eta}, \boldsymbol{d})))h^*(w) && (6.13)\\ &\propto L(\boldsymbol{x} \cup (w; (\tilde{\eta}, \boldsymbol{d})))h^*(w) \prod_{i=1}^{n} \Psi(w_i, w). && (6.14) \end{aligned}$$

In practice, empirical knowledge on the size of a probable percentual change in the MR signal level (due to neuronal activation) can suggest a range for reasonable values of $\tilde{\eta}$. To use the density above we should calculate the likelihood term $L(\boldsymbol{x} \cup (w; (\tilde{\eta}, \boldsymbol{d})))$ and consider sums like $\sum_{s-u=l} \varrho(s)\varrho(u)\epsilon(s)\kappa(u-w)$, which would be time-consuming. Here, the terms $\epsilon(s) = z(s) - \alpha(s)$ are the current residuals. To simplify updating calculations we introduce a modified likelihood $\tilde{L}$, for which

$$\tilde{L}(\boldsymbol{x}) \propto \exp\Big(-\frac{1}{2}\sum_s \tilde{\varrho}(s)^2 (z(s) - \alpha(s))^2\Big), \tag{6.15}$$

for all configurations $\boldsymbol{x}$ and where $\tilde{\varrho}(s)^2$ match with the inverse voxelwise variances $1/\operatorname{var}(Z(s)|\,\boldsymbol{x})$. This likelihood corresponds to a noise model which has the same voxelwise variances as the original model but now the noise is spatially independent. We shall propose w from the conditional distribution of the *modified* posterior $\tilde{p}$ where the original likelihood L is replaced by $\tilde{L}$. Thus, we use the density

$$q(w|\,\boldsymbol{x}, \boldsymbol{d}, \tilde{\eta}) \propto \tilde{L}(\boldsymbol{x} \cup (w; (\tilde{\eta}, \boldsymbol{d})))h^*(w) \prod_{i=1}^{n} \Psi(w_i, w) \tag{6.16}$$

to propose a new location. The calculation of the modified likelihood function $\tilde{L}(\boldsymbol{x} \cup (w; (\tilde{\eta}, \boldsymbol{d})))$ involves computation of the square sum $\sum_s \tilde{\varrho}(s)^2 (z(s) - \alpha^*(s))^2$, where $\alpha^*(s) = \alpha(s) + \tilde{\eta}\kappa(s-w)$ is a temporary profile. This, in turn, requires that the sums $\sum_s \tilde{\varrho}(s)^2 \epsilon(s)\kappa(s-w)$ and $\sum_s \tilde{\varrho}(s)^2 \kappa(s-w)^2$ are calculated for all locations w. The calculation of these convolutions is also a time-consuming operation but now it is easier to apply the fast Fourier transform (FFT) to speed up computations.

In the special case $\tilde{\eta} = 0$ we have a particularly simple form for proposing locations:

$$q(w|\, \boldsymbol{x}, \boldsymbol{d}, 0) \propto h^*(w) \prod_{i=1}^{n} \Psi(w_i, w). \tag{6.17}$$

In fact, this is the conditional prior density of a cluster location, given the other clusters $x_1, x_2, \ldots, x_n$ of the configuration. It is seen that in this case the proposal distribution ignores the form of the likelihood function and is computationally more efficient. On the other hand, probable cluster locations are not proposed as likely if $\tilde{\eta} = 0$. It is difficult to give general recommendations, but we anticipate that using the computationally more involved proposal density $q(w|\, \boldsymbol{x}, \boldsymbol{d}, \tilde{\eta})$, the chain reaches the equilibrium faster, and that the proposal density $q(w|\, \boldsymbol{x}, \boldsymbol{d}, 0)$ may be useful in exploring the posterior in the equilibrium.

After proposing the daughter configuration $\boldsymbol{d}$ and the new parent centre w, we finally propose the height η of the new cluster from the conditional modified posterior distribution given $\boldsymbol{x}$, $\boldsymbol{d}$ and w. The density (with respect to m_1) is in this case

$$q(\eta|\, \boldsymbol{x}, \boldsymbol{d}, w) \propto \tilde{p}(\boldsymbol{x} \cup (w; (\eta, \boldsymbol{d})))h_\eta(\eta) \propto \tilde{L}(\boldsymbol{x} \cup (w; (\eta, \boldsymbol{d})))h_\eta(\eta). \tag{6.18}$$

If h_η is a normal density, then $q(\eta|\, \boldsymbol{x}, \boldsymbol{d}, w)$ is also a normal density. It is probable that this proposal distribution is usually narrower than the one based on the original likelihood L since we ignore spatial correlations in the data.

For the calculation of the acceptance probabilities for inserting or removing a cluster, we need the density of the birth distribution $b(\xi|\, \boldsymbol{x})$ with respect to $\varphi = \lambda_{w,\eta} \times \lambda_d$. The birth density of a new cluster ξ with respect to φ will be

$$b(\xi|\, \boldsymbol{x}) = \frac{q(\eta|\, \boldsymbol{x}, \boldsymbol{d}, w)}{h_\eta(\eta)} \frac{q(w|\, \boldsymbol{x}, \boldsymbol{d}, \tilde{\eta})}{h^*(w)} q(\boldsymbol{d}). \tag{6.19}$$

The Metropolis-Hastings ratio for accepting an insertion of a proposed cluster ξ is

$$MHR_{c,i} = \frac{p(\boldsymbol{x}^*)}{p(\boldsymbol{x})} \frac{1/(n+1)}{b(\xi|\, \boldsymbol{x})} \tag{6.20}$$

$$= \frac{L(\boldsymbol{x}^*)}{L(\boldsymbol{x})} \frac{\psi^*(n+1)}{\psi^*(n)} \frac{e^{1-\beta_d} \beta_d^{n(\boldsymbol{d})} \prod_{i=1}^{n} \Psi(w, w_i)}{b(\xi|\, \boldsymbol{x})}, \tag{6.21}$$

where $\boldsymbol{x}^* = \boldsymbol{x} \cup \xi$, and for accepting a removal of a random cluster $\xi \in \boldsymbol{x}$ we have

$$MHR_{c,r} = \frac{p(\boldsymbol{x}^*)}{p(\boldsymbol{x})} \frac{b(\xi|\,\boldsymbol{x}^*)}{1/n} \tag{6.22}$$

$$= \frac{L(\boldsymbol{x}^*)}{L(\boldsymbol{x})} \frac{\psi^*(n-1)}{\psi^*(n)} \frac{b(\xi|\,\boldsymbol{x}^*)}{e^{1-\beta_d}\beta_d^{n(\boldsymbol{d})} \prod_{w_i \neq w} \Psi(w, w_i)}, \tag{6.23}$$

where $\boldsymbol{x}^* = \boldsymbol{x} \setminus \xi$. It is important to observe that although we have used a modified likelihood $\tilde{L}$ to propose clusters, we use the original likelihood L and posterior density p to correct the proposal distribution in order to maintain detailed balance. Another important detail is that $\tilde{\eta}$ can be considered random in the following sense. Since for each choice of $\tilde{\eta}$ the transition operation is reversible, we may first draw $\tilde{\eta}$ randomly and then update the configuration using the sampled value of $\tilde{\eta}$, ignoring the randomness of $\tilde{\eta}$.

The fifth transition type is to insert or remove a daughter from a randomly chosen cluster x_i. First, we decide whether to insert or remove a daughter. The two proposal types have an equal probability. Second, in case of an insertion we use our prior densities h_v and h_ζ for proposing a new location v and a new height ζ. The proposed configuration $\boldsymbol{x}^*$ will then be $\boldsymbol{x}$ with the new daughter $d = (\zeta, v)$ included in the chosen cluster x_i. The acceptance ratio is easy to calculate since the density of the birth distribution is identically one with respect to the intensity measure of λ_d. In fact,

$$MHR_{d,i} = \frac{L(\boldsymbol{x}^*)}{L(\boldsymbol{x})} \frac{\pi(\boldsymbol{x}^*)}{\pi(\boldsymbol{x})} \frac{1/(k+1)}{1} = \frac{L(\boldsymbol{x}^*)}{L(\boldsymbol{x})} \frac{\beta_d}{k+1}, \tag{6.24}$$

if the number of daughters in the selected cluster x_i is k. In case of a removal we pick randomly one daughter from the cluster and our proposal $\boldsymbol{x}^*$ will then be $\boldsymbol{x}$ with the chosen daughter deleted from x_i. The acceptance ratio is now

$$MHR_{d,r} = \frac{L(\boldsymbol{x}^*)}{L(\boldsymbol{x})} \frac{k}{\beta_d}. \tag{6.25}$$

If there are no daughters in the chosen cluster and a removal is proposed, we do nothing. Furthermore, if there are no clusters, that is $\boldsymbol{x} = \emptyset$, we also do nothing.

6.2 MCMC estimation

The MCMC estimation of functionals of the posterior distribution p is based on the assumption that the simulated Markov chain is in equilibrium. In practice, we cannot simulate the chain in equilibrium since $\boldsymbol{x}^{(0)}$ cannot be drawn from p. We can simulate only the conditional values of the chain given the initial state $\boldsymbol{x}^{(0)}$. As a consequence, the distributional properties of the realizations $\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}, \ldots$ can be far from p since the serial dependence in the chain is typically non-negligible. However, the chain will asymptotically converge to the target p. Therefore, the chain will be nearly stationary after some initial period, which is commonly called *burn-in time*. To avoid sampling bias, it is a common practice to approximate the burn-in time K and discard the realizations $\boldsymbol{x}^{(1)}, \dots, \boldsymbol{x}^{(K)}$. To get an estimate for the burn-in time, it is useful to monitor the behaviour of the chain. There are usually many aspects of the chain that could be monitored, but two natural candidates are the values of the log likelihood function and log prior. The formulae for these functionals are given by the equations

$$\begin{aligned} \log \pi(\boldsymbol{x}) = n(\boldsymbol{x})(1-\beta_d) + \log \psi^*(n(\boldsymbol{x})) + \sum_{i=1}^{n(\boldsymbol{x})} \Big[n(d_k) \log \beta_d + \log i \Big] \\ + \sum_{i<j} \log \Psi(w_i, w_j) + \text{const.}, \end{aligned} \tag{6.26}$$

and

$$\begin{aligned} \log L(\boldsymbol{x}) = & -\frac{1}{2} \sum_s \varrho(s)^2 (z(s) - \alpha(s))^2 \\ & - \frac{1}{2} \sum_l \tau_l \sum_{s-u=l} \varrho(s)\varrho(u)(z(s)-\alpha(s))(z(u)-\alpha(u)) + \text{const.}, \end{aligned} \tag{6.27}$$

which contain two unknown constants that can be ignored. Clearly, $\log p(\boldsymbol{x})$ is the sum of $\log L(\boldsymbol{x})$ and $\log \pi(\boldsymbol{x})$ up to an additive constant.

After the burn-in time K has been approximated, the mean μ_g of a random variable $g(\boldsymbol{x})$ can be estimated by

$$\hat{\mu}_g = (1/(N-K)) \sum_{n=K+1}^{N} g(\boldsymbol{x}^{(n)}). \tag{6.28}$$

Also, quantiles of $g(\boldsymbol{x})$ can be estimated by empirical quantiles of realizations $g(\boldsymbol{x}^{(K+1)}), \dots, g(\boldsymbol{x}^{(N)})$. The calculation of the sampling variance of the estimate $\hat{\mu}_g$ is based on the estimation of the covariances $\gamma_k = \text{cov}(g(\boldsymbol{x}^{(n)}), g(\boldsymbol{x}^{(n+k)}))$. Asymptotically,

$$\widehat{\text{var}(\hat{\mu}_g)} = \Big(\gamma_0 + 2 \sum_{k=1}^{\infty} \gamma_k \Big) / (N-K). \tag{6.29}$$

Therefore, a naive method to estimate $\text{var}(\hat{\mu}_g)$ would be to sum all estimated covariances $\hat{\gamma}_k$. However, this estimate lacks consistency (Priestley, 1981). An appropriate strategy is to allow the number of covariance terms (to be summed) to depend on the length of the realized chain. Geyer (1992) proposed that an estimate of the sum (6.29) should contain only terms up to $\hat{\gamma}_{2m+1}$ where the truncation value m is the largest integer for which the sequence $\hat{\gamma}_2 + \hat{\gamma}_3, \ldots, \hat{\gamma}_{2m} + \hat{\gamma}_{2m+1}$ is positive and monotonically decreasing. The resulting standard error can be used to construct Monte Carlo confidence interval for μ_g since the estimate (6.28) is consistent and asymptotically normally distributed (Kipnis & Varadhan, 1986).

We end this section by suggesting two elementary computational techniques to speed up sampling calculations. The first one concerns the updating of the profile $\boldsymbol{\alpha}$, which is important since the current value of $\boldsymbol{\alpha}$ is needed for computing acceptance probabilities. We note that an updating operation can change the value of $\boldsymbol{\alpha}$ only inside some cluster. Clusters, in turn, are *local* spatial objects. Therefore, each transition affects $\boldsymbol{\alpha}$ only locally and it is enough to update only part of $\boldsymbol{\alpha}$ instead of deriving it from $\boldsymbol{x}$ (using (4.7) and (4.8)) each time $\boldsymbol{x}$ happens to change. The second computational idea concerns the time-efficiency when monitoring the likelihood function $L(\boldsymbol{x})$. A useful observation is that the likelihood ratios $L(\boldsymbol{x}^*)/L(\boldsymbol{x})$ are always calculated when computing the acceptance probability for a proposed new state $\boldsymbol{x}^*$. Therefore, no extra calculations are required to keep track of the values of the likelihood. However, to avoid accumulation of rounding errors, it is advisable to compute both the profile $\boldsymbol{\alpha}$ and the log likelihood from scratch at regular intervals.

7 Analysis of the sound-stimulus data

In Sections 4, 5 and 6 we suggested a Bayesian technique for accounting for subjective prior knowledge and presented a method for sampling the posterior. The development of ideas proceeded on a theoretical level, and we discussed the modelling concepts in general. The purpose of this section is to show how our Bayesian machinery can be used in a real application and to answer specific inferential questions. We shall illustrate the proposed new Bayesian approach by applying it to the analysis of the sound-stimulus data. Some basic properties of these data were already described in Section 2. We recall that the auditory stimulus had a periodic pattern and consisted of two alternating test conditions. In this experiment, there is only one natural contrast $\boldsymbol{\alpha}$ to be examined: the difference in the mean resonance signal levels between the two states of the brain. Consequently, the intensities of the profile $\boldsymbol{\alpha}$ will now simply be the amplitudes of the auditory activation process in the scanned brain slice. We recall also that these data were acquired from a *single* functional slice position.

The auditory experiment was designed in such a way that only a few regions in the brain were expected to be involved in processing the test task. In fact, the experiment could be considered an example of brain studies mentioned by Friston and his co-workers in the motivating citation at the end of Section 3 of this thesis. This kind of subjective knowledge could be modelled in a natural way by assigning a suitable prior to the number of activated clusters. Another source of useful neuropsychological prior knowledge is that certain brain regions more likely participate in the neural processing of interest. We shall consider here the regions TTG, PT, PP and STG (Figure 2.2). These regions were identified from an anatomical scan by an expert (Dr. Jennifer Hiemenz, University of Georgia, personal communication). To illustrate the utilization of information on the locations of brain

Table 7.1 Areas of brain regions.

Region name	Area (in voxels)
TTG	145
PT	227
PP	422
STG	1377
Union $\mathcal{R}$	794
All brain voxels	7530

activations, we shall assume that the union of TTG, PT and PP has a significant role in processing the specific type of auditory stimulus. The symbol $\mathcal{R}$ shall be used for this union. The areas of these regions (measured in voxels) are listed in Table 7.1. We observe that the region $\mathcal{R}$ contains about one tenth of all the brain voxels on the slice.

The total number of functional scans in the sound-stimulus data is 40. We do not aim to analyse the complete data here. Instead, we shall consider the first four scans and make inference on the profile $\boldsymbol{\alpha}$ based on these functional images. The reason for this is that in this thesis we are mainly interested in considering situations where the posterior is sensitive to subjective prior knowledge (to some extent at least). If a series of 40 scans is analysed, the domination of the data will be more pronounced. We shall, however, use all the data when estimating the spatio-temporal characteristics of the noise in the series of scans. Consequently, the application in this section can be considered a special case of applying a concatenated design. The data are visualized in Figure 7.1.

7.1 The choice of the model

Priors

We shall apply three different priors describing different levels of prior knowledge on the parent configuration. In prior A, the number of parent centres is assumed to be distributed uniformly from 0 to 10. In prior B, the total count of these centres is assumed to have a narrower distribution, namely from 0 to 5. The two priors, A and B, will be uninformative with respect to the locations of the centres. The prior C will be adjusted so that the number of centres is restricted between 0 and 5 (as in the case of prior B) and the location of a random centre is assumed to belong to region $\mathcal{R}$ with probability 0.50. The probability 0.50 is quite high proportioned to the fact that region $\mathcal{R}$ contains only about one tenth of the

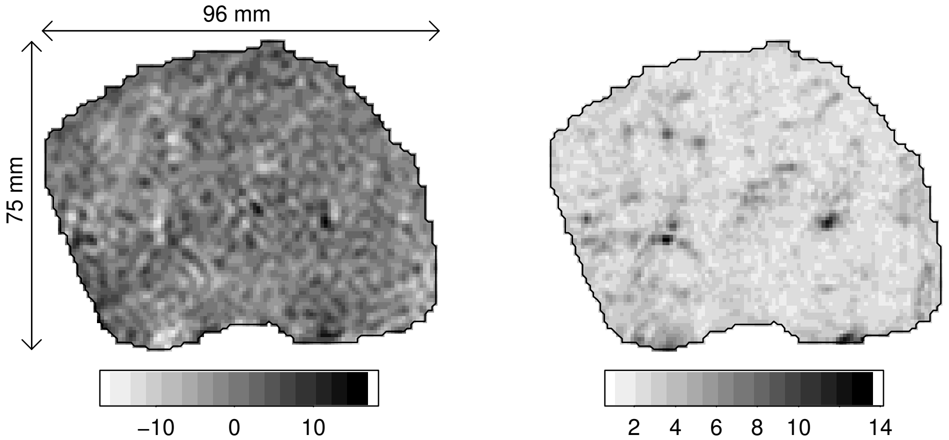

Figure 7.1 The compressed data $\boldsymbol{Z}$ (left) and the estimated voxelwise standard deviations of $\boldsymbol{Z}$ given the activations (right).

voxels. Therefore, when applying the third prior, the centres will favour region $\mathcal{R}$. The reason for employing several priors is that we want to illustrate how sensitive the posterior is to priors used in typical applications. The different adjustments of the prior are used in an illustrative purpose but they are close to choices which could be used by an applied scientist.

Our cluster model contains several hyperparameters. Our aim is to select such values for these parameters that the clusters will have acceptable physical dimensions (Malonek & Grinvald, 1996). To begin with, the parent bell and the daughter bell are chosen to be isotropic Gaussian bells having a common radius of spatial extent. A plausible choice for this radius is four voxels (3.4 mm). We assume *a priori* that the parent bells cannot overlap one another. To accomplish this, we shall use a hardcore spatial interaction function Ψ where the radius of inhibition is eight voxels. We note that daughter bells of different parents may overlap although the parent bells do not. The corrected forms for the prior count distribution ψ of parent points and the density h of the location of a random parent can be computed as shown in Appendix A.2.

Our choice for daughter locations v is uniform on voxels that have distance to the origin at most 4 voxels (3.4 mm). Thus, the maximum spatial extent for any cluster measured from the parent centre is 8 voxels (6.9 mm). As the prior for scaling factors ζ of daughter bells we shall take the uniform distribution at the unit interval $(0,1)$. This choice reflects a prior belief that clusters are not strongly peaked spatial objects. In fact, uniform distributions at intervals like $(0,0.5)$ or $(0,0.1)$ would favour more peaked clusters in the profile since then the role of the parent bell would be emphasized. As to the prior number of daughters in a

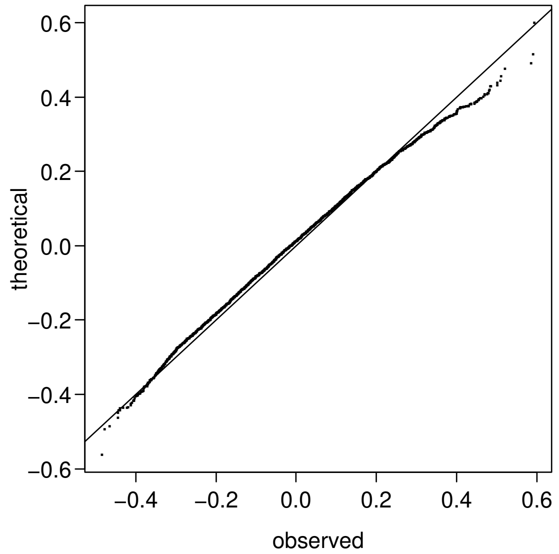

Figure 7.2 A Q-Q plot for observed and theoretical temporal correlations of lag one.

cluster, our purpose is to be economical in the overall number of parameters in the posterior. Our choice is to take the Poisson mean of the number of daughters to be 4.0. Finally, we assign a prior for a scaling factor η. Following the logical development of the material in Section 4, our aim is to be relatively uninformative with respect to the magnitudes of activations. Data analytic considerations suggest that the zero-mean normal distribution having standard deviation 6.0 serves for this purpose.

Likelihood

To fit the parameters of the likelihood function to the data, we first considered the strength of temporal dependence in the series of scans. The empirical temporal autocorrelations were small in most voxels and, moreover, a comparison with an artificial temporally *uncorrelated* data showed that the distributions of empirical and theoretical autocorrelations match quite well (Figure 7.2). The probable reason for this is that the functional data was temporally sparsely sampled (see Figure 2.1). We conclude that the error time series $\delta(s)$ can be treated as temporally uncorrelated noise in this particular data.

We fitted the spatial correlation structure of the Gaussian noise model (5.7) using only lags of length one in horizontal and vertical directions. The curves in

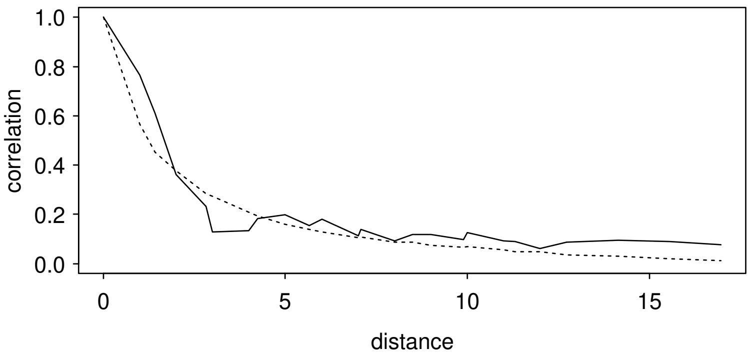

Figure 7.3 Observed (solid curve) and fitted (dotted curve) mean spatial correlations. The distance is measured in voxels and the distance of 15 voxels corresponds to about 13 mm.

Figure 7.3 represent the empirical mean spatial correlations for each lag and the corresponding correlations computed from simulated realizations of the Gaussian model. It is seen that the choice $\tau = -0.249$ is a plausible estimate for the strength of spatial dependence in the data. The precision parameters $\boldsymbol{\varrho} = (\varrho(s))$ were calculated by the formula (5.14) using simulations from the fitted autoregressive model (5.15) and noise variance estimates from the data. The fitted values of the precision parameters are visualized in Figure 7.4.

7.2 Sampling

We shall apply the MCMC sampling method suggested in Section 6 to calculate Monte Carlo estimates for various functionals of the posterior. To carry out posterior simulation, we have to replace the tuning parameters of the proposal distributions with known numerical values. In a sense, the actual choice of these values does not matter because the sampler has the same convergence properties irrespective of the choice of the parameters of the proposal distributions (Appendix A.4). However, the level of mixing of the resulting Markov chain may crucially depend on the tuning of the proposal mechanism.

To avoid mixing problems, we performed a few test runs and monitored the acceptance probabilities of the proposed moves of the chain. We aimed to propose large moves while maintaining reasonably high acceptance rates. As a result, we arrived at the following sampling scheme. A new height η^* of a cluster is proposed from the normal distribution having the current height η as the mean and standard

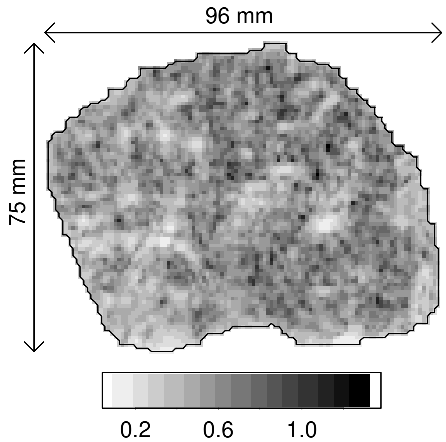

Figure 7.4 The map of the fitted precision parameters $\hat{\boldsymbol{\varrho}} = (\hat{\varrho}(s))$.

deviation 0.5. A displacement for the location w^* of the parent of a cluster is drawn uniformly over voxels less than three voxels away from the current location w. New heights ζ^* of daughter bells are proposed so that $\log \zeta^*$ is uniform on $(\log \zeta - 0.3, \log \zeta + 0.3)$. The location of a new daughter is proposed uniformly from voxels less than four voxels away from the origin. Also, the proposed height of the new daughter bell is uniformly distributed at the unit interval $(0, 1)$.

The proposal mechanism for generating new clusters is adjusted as follows. The number of daughters to a new cluster is proposed from the Poisson distribution with mean 4.0. The locations and heights of daughters are proposed uniformly as in the case of inserting/removing single daughters from an existing cluster. For proposing the location of a new cluster, a height $\tilde{\eta}$ must be chosen. We select these heights randomly from a set of 100 values ranging from -10.0 to 10.0. We recall that the randomness of $\tilde{\eta}$ can be ignored as explained in Section 6. Table 7.2 illustrates the mean acceptance probabilities of each of the five move types.

Test runs revealed that the sampler approached equilibrium in 50000 sweeps when starting from the empty configuration. This configuration is usually a poor initial state for the chain, but otherwise its use is very convenient since then no extra estimation procedure is needed to provide values for initialization. To speed up calculations we implemented the two computational ideas which were mentioned at the end of Section 6. It turned out that in order to avoid rounding errors it is sufficient to compute the profile and the log likelihood from scratch at every 100th and 300th update, respectively. The sample paths (after the burn-in) of log prior and log likelihood are visualized in Figure 7.5. Monitoring results indicated that realizations are strongly autocorrelated. Therefore, we subsampled only every 20th update for posterior analysis.

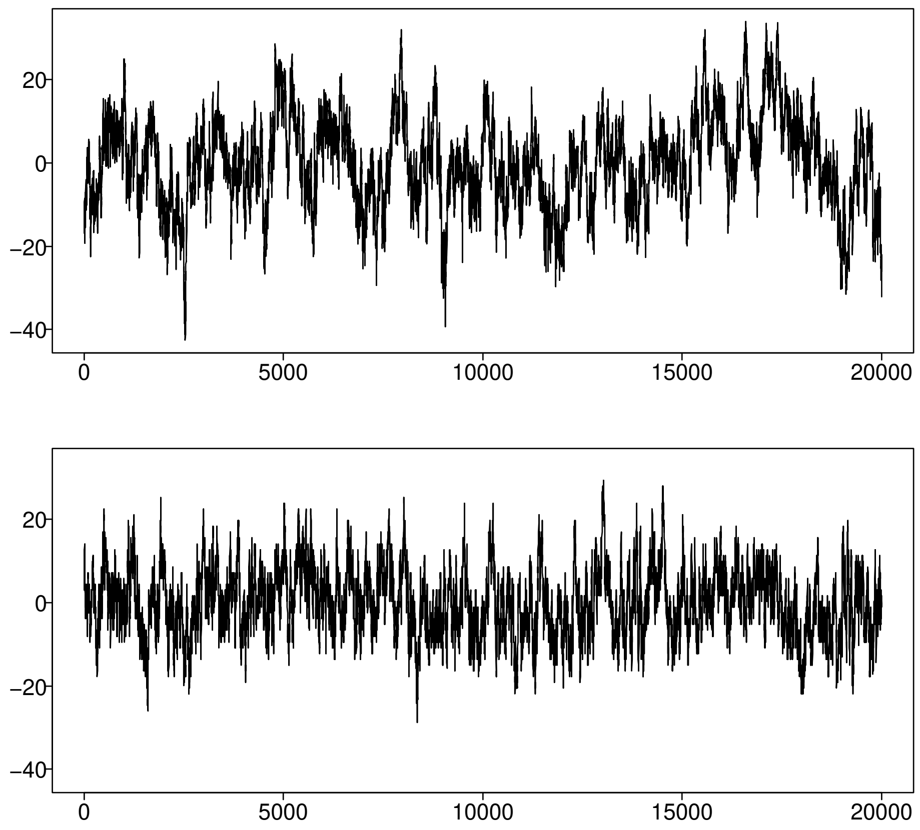

Figure 7.5 The behaviour of the log likelihood (top) and log prior (below) during MCMC simulation. The series are mean-corrected. The subsampling rate was 20 and, therefore, the length of the series is 20000 (instead of 400000).

Table 7.2 Acceptance probabilities.

Move type	Mean acceptance rate
Parent height	0.808
Parent location	0.092
Daughter height	0.466
Daughter add/del	0.482
Cluster add/del	0.004

7.3 Results

We simulated the three posteriors for priors A, B and C 400000 sweeps after the burn-in. To compare the posteriors we have juxtaposed their summary maps in Figure 7.6. The Monte Carlo standard errors were calculated for the posterior mean maps. The sizes of these errors ranged between 0.1 and 0.3 in those voxels where the estimation error was expected to be largest. In all of these voxels, the standard errors were from 1.0 to 4.0 % of the estimated voxel means. It can be seen from the posterior deviation maps that the posterior variability is non-negligible in the same areas where the posterior mean of the profile is not vanishing. Also, a visual inspection reveals that the overall variability slightly decreases when the prior becomes stronger.

The interpretation of the posterior mean maps requires some care. For example, the number of "hills" and "valleys" in the profile can well be somewhat larger than the typical number of clusters in a (posterior) configuration $\boldsymbol{x}$. The clusters of $\boldsymbol{x}$ may appear and disappear in different locations during the simulation, which explains this peculiar phenomenon. We want to stress here that although the posterior mean map is useful for describing the strength of brain activation voxelwise, it does not necessarily resemble a typical posterior profile $\boldsymbol{\alpha}$. In this sense, the *maximum a posteriori* (MAP) estimate of the posterior could be an interesting alternative for the mean when carrying out Bayesian point estimation. We shall not study MAP estimates here, however.

We have now illustrated what the effects of incorporating different levels of prior knowledge can be on the inference. From now on, we shall confine ourselves to examining the posterior A.

The two posterior moments, the mean and standard deviation, provide us with some information on the posterior uncertainty in activation magnitudes. It is an appealing feature of the Bayesian paradigm that it allows for investigating uncertainty directly in terms of probabilities. The left of Figure 7.7 shows voxelwise posterior probabilities that the profile intensities $\alpha(s)$ are positive.

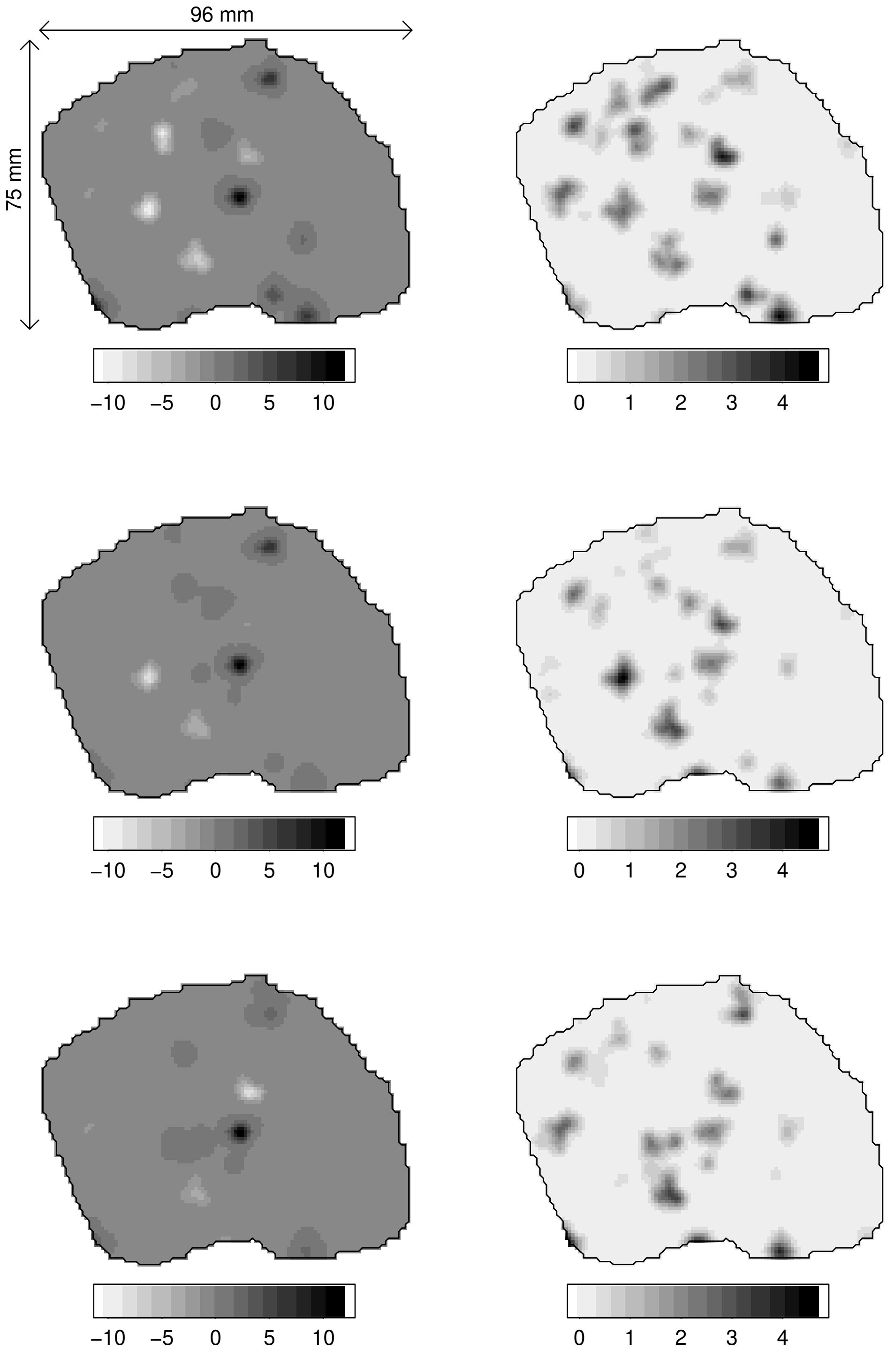

Figure 7.6 Voxelwise means (left column) and standard deviations (right column): posterior A (top row), B (middle row) and C (bottom row).

Here, a high posterior probability in a voxel indicates that the voxel was activated when the test person was stimulated using digital sounds. We note that the intensities $\alpha(s)$ are not normally distributed in the posterior, and thus the probabilities cannot be expressed using the posterior moments.

In addition to considering the activation magnitudes, it is also possible to examine the configuration of the centres of the parent bells. The right of Figure 7.7 shows the map of probabilities that a voxel belongs to the support of some parent bell. As is easily observed, the number of local patterns in this map is larger than ten. This is related to our previous comment upon clusters and the way how their locations are random in the posterior.

Up to this point, we have been constructing maps for summarizing our posterior. We shall now consider two specific inferential questions which can be answered using single probability statements. The problem which we study first concerns brain activation in given regions. Let us suppose a region of interest can be identified from anatomical scans. Then, it may be natural to consider the mean activation in that region in order to infer from the role of that particular part of the brain in neuronal processing. The mean activation magnitude over a region has a one-dimensional posterior distribution from which the inference can easily be drawn. An illustration of this is shown in Figure 7.8. The Monte Carlo estimate for the posterior probability that the mean activation over region $\mathcal{R}$ is positive is 0.94. In general, the size of a region may have some influence on the interpretability of the results. The regional mean is a less accurate predictor for the whole regional profile for large regions than for smaller ones. Thus, regions should be chosen as small as possible in applications.

Our second problem is related to the sizes of the clusters, that is, to the elements x_i in configurations $\boldsymbol{x} = \{x_1, \dots, x_n\}$. The size of a cluster is a potentially relevant notion in fMRI research. Basically, there are two separate means for measuring the size of a cluster: the spatial extent and the highest activation magnitude of a cluster. Interestingly, the parametrization of our cluster model can be used to construct a simple combined measure of extent and magnitude. We suggest that the size of a cluster x_i could be assessed using the integral of $\eta_i \kappa_i(s)$ over U. Clearly,

$$\int \eta_i \kappa_i(s)\ \mathrm{d}\nu(s) = \eta_i \int B_p(s)\ \mathrm{d}\nu(s) + \eta_i \sum_{j=1}^{k_i} \zeta_{j,i} \int B_d(s)\ \mathrm{d}\nu(s), \qquad (7.1)$$

where k_i is the number of daughter bells in x_i. We shall call (7.1) the *integrated activation magnitude*. The integrated activation magnitude is more informative than the extent and maximum magnitude of a cluster together since the integrated measure provides information also on activation magnitudes in the neighbouring voxels of the cluster centre. The calculation of the proposed measure is illustrated in Figure 7.9. Integrated magnitudes were calculated for each cluster of MCMC samples $\boldsymbol{x}$. The extreme values, the lowest and the highest, were used to estimate the posterior distributions of the lowest and the highest integrated activation magnitudes. Naturally, one could choose to utilize the integrated measures of *all* the clusters in a configuration (and not to ignore all but the two extreme ones) to produce other summaries of the posterior of $\boldsymbol{x}$ as well. However, a problem is what kind of descriptors of the set of integrals (7.1) would be appropriate in real applications.

A minor defect in our definition of an integrated measure of the cluster size is that it is not edge-corrected. In other words, an incorrect value results if a cluster is located near the boundary of the brain slice. We shall not consider edge-corrected versions of (7.1) in this thesis.

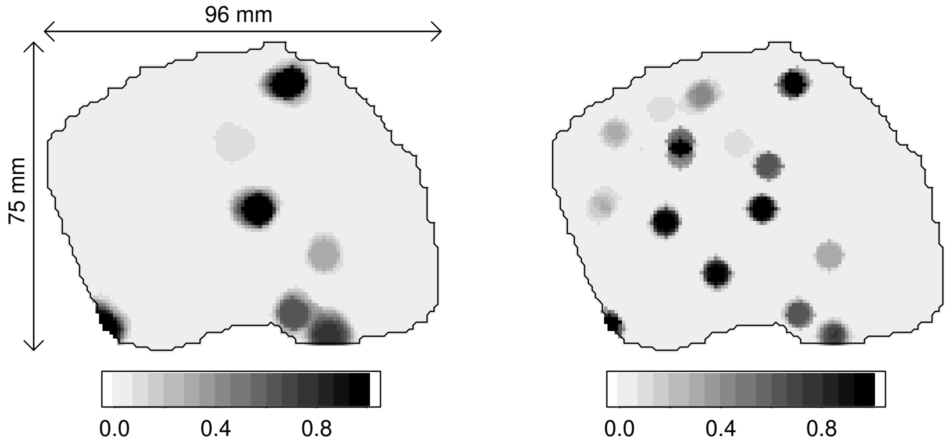

Figure 7.7 Posterior probabilities that an activation magnitude is positive (left). Posterior probabilities for belonging to a parent bell (right).

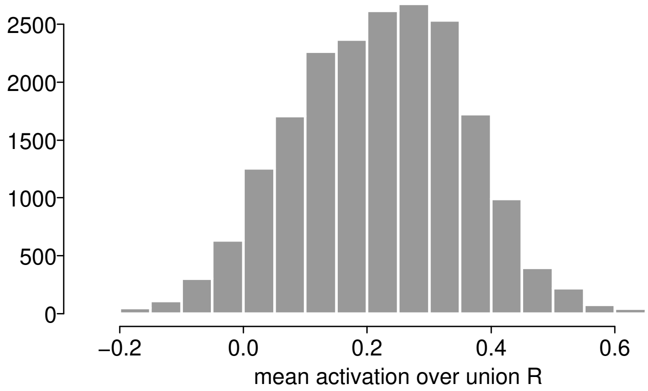

Figure 7.8 The histogram of the posterior samples of the regional mean over the union $\mathcal{R}$ of TTG, PT and PP. The number of MCMC samples is shown on the vertical axis. The total number of samples was 20000.

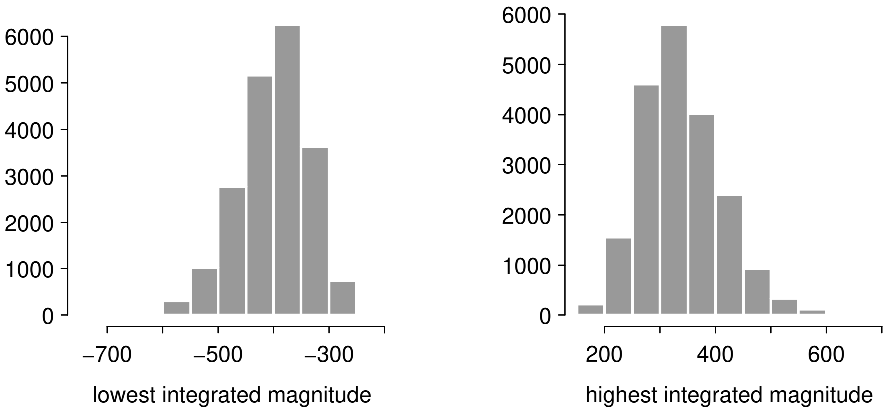

Figure 7.9 Histograms of the obtained realizations of the lowest and highest integrated magnitudes among clusters in the profile. The number of MCMC samples is shown on the vertical axis. The total number of samples was 20000.

8 Analysis of synthetic data

8.1 Nonparametric Bayesian smoothing

The cluster prior model was developed to be a means for incorporating relevant subjective prior knowledge on an activation process into the statistical analysis of fMRI data. In the previous section, we illustrated the use of the cluster approach in an application. We used the prior for controlling the probable number of clusters, their locations and extents. Consequently, the adjustment of the prior was quite involved. A natural question is: how do the posteriors change if the cluster prior is replaced by a more conventional Markov random field prior? Interestingly, Bayesian image analysis has to offer a whole class of prior models, the *pairwise difference smoothing* priors, which have a parsimonious hyperparameter structure (Besag, 1989). These priors are (improper) Markov random fields with density

$$\pi_M(\boldsymbol{\alpha}) \propto \exp\left(-\gamma_1 \sum_{s \sim u} \Phi\left(\frac{\alpha(s) - \alpha(u)}{\gamma_2}\right)\right), \tag{8.1}$$

where Φ is an increasing non-negative and even smoothing potential and γ_1 and γ_2 are positive constants. A basic concept of pairwise smoothing is the notion of a neighbourhood relation '$\sim$'. Typically, the neighbourhood of a pixel s is defined to consist of pixels spatially close to s. The motivation in the use of pairwise difference priors is that if spatially close pixels are expected to have similar attributes, a prior like (8.1) can be used to penalize large differences between neighbouring pixels. The control parameters γ_1 and γ_2 can be used to tune the properties of the difference prior. In informal terms, γ_1 determines the level of smoothing and γ_2 accounts for the scale of data.

Depending on the application, various choices of Φ have been suggested in the literature. Convex potentials have been used to support gradually changing surfaces in the prior, whereas concave potentials have been used to encourage abrupt changes in otherwise smooth images (Besag *et al.*, 1995; Geman & Reynolds, 1992). Among the convex Φ, a popular choice is the Gaussian potential $\Phi_G(x) = x^2$. A less heavy penalization for large differences in neighbouring intensities is provided by the Huber potential

$$\Phi_H(x) = \begin{cases} x^2 & |x| \leq 1 \\ 2|x| - 1 & |x| > 1, \end{cases} \tag{8.2}$$

which is also convex (Künsch, 1994). An example of a concave potential is

$$\Phi_{Ge}(x) = \frac{|x|}{1 + |x|}\,, \tag{8.3}$$

which was suggested by Geman & Reynolds (1992). An interesting feature of the Geman potential is that it can be used to implicitly model the effects of an unobserved edge process. In particular, the level of statistical modelling can be kept simple since it is not necessary to formulate probabilistic properties of sharp edges. The graphs of these examples of potentials are shown in Figure 8.1.

It is interesting to carry out comparisons between pairwise difference models and our cluster approach and assess how the posteriors differ from each other. In the context of estimating activation profiles, there is no single choice among the above potentials which would obviously be more suitable than the others. On one hand, a convex smoothing technique (like Gaussian) could be a natural candidate. However, a special property of the Gaussian prior is that it penalizes large differences heavily. In practice, this means that small clusters may be smoothed away from posterior realizations. In other words, there is a risk for oversmoothing. On the other hand, there is some justification for applying a concave smoother such as the Geman prior. Each cluster might be viewed as having a boundary curve, and these curves could be considered a kind of edges which would be preserved in concave smoothing. However, there hardly exist any *sharp* edges in activation profiles. Because of these controversies, we choose to apply the *absolute value* potential $\Phi_A(x) = |x|$ since it is an intermediate choice between both convex and concave potentials. As a sidenote, we remark that asymptotically the Geman prior tends to our choice as γ_2 increases.

To compare the different modelling strategies, we want to relate the resulting posteriors to the true state of activations in data and check how well they predict the underlying true activation profile. To this end, we shall consider synthetic data instead of real functional measurements. The synthetic data were generated by corrupting a known profile $\boldsymbol{\alpha}^{(0)}$ with additive scanning noise. The search volume is a rectangular region having width and height 50 and 30 voxels, respectively. The profile consists of three exactly similar clusters where each cluster is a superposition

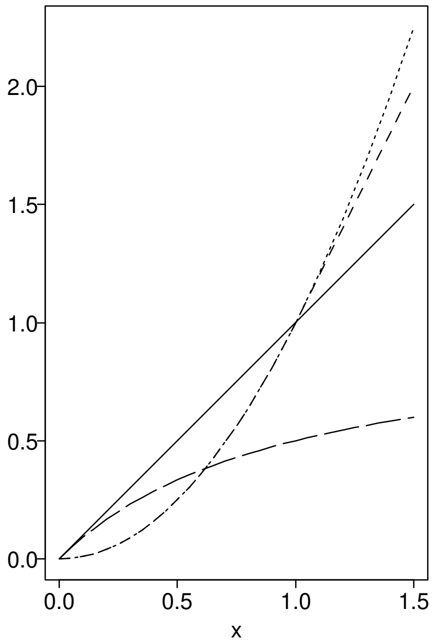

Figure 8.1 Graphs of smoothing potentials: absolute value (—), Gaussian (···), Huber (- -) and Geman (— —).

of one parent bell and one daughter bell. The radius of the parent and daughter bell is three voxels. The daughter bells are located four voxels to the right from their parent centres. Moreover, the height of each daughter bell is half from that of their parents. The profile and the data are visualized in Figure 8.2.

The noise was acquired in a real fMRI experiment where echo-planar imaging (EPI) technique was used. The spatial resolution of the functional images was 128 $\times$ 128. Using a mask with size 50 $\times$ 30 voxels, part of the voxel series was extracted from the complete data. The additive noise field was constructed by compressing these series in the spirit of Section 5. After the compression, the noise field was standardized dividing each intensity by its standard deviation in order to have unit variance in all voxels. The reason for normalizing the noise was to facilitate the examination of the posterior maps. We note that simulated noise fields drawn from known random fields could also be utilized in a comparison study like this. Now, our autoregressive model for noise will also be evaluated, and we gain information on how our results are related to the unobserved activation process in real problems.

We shall apply two cluster priors, A and B, in the present study. In prior A, the prior count of parents is uniform from 0 to 10. This range is limited to be between 0 and 5 in prior B. In both of these priors we shall be ignorant with respect to

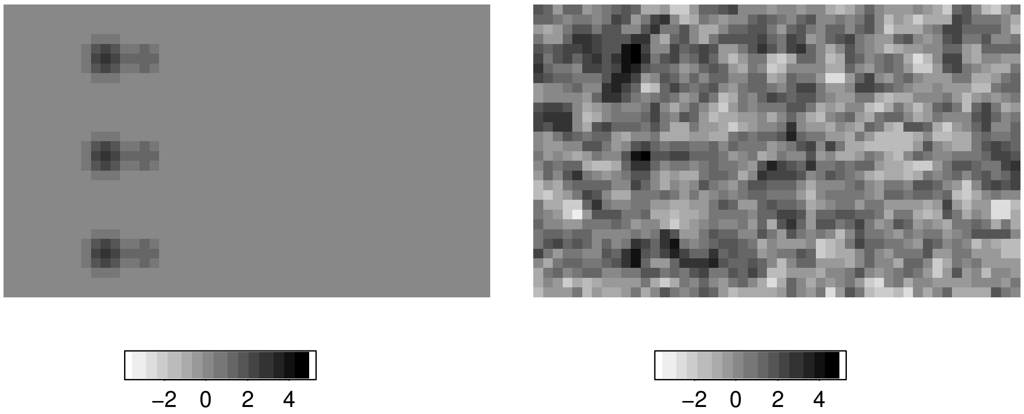

Figure 8.2 The underlying true profile of the synthetic data (left) and the data itself (right).

the locations of centres. The common radius of the parent and daughter bells is chosen three voxels instead of four (as was chosen in the previous section) since now the spatial resolution is lower. Using the same logic, the inhibition radius was changed to six voxels instead of eight. Otherwise, the priors A and B are similar with the priors of Section 7. For example, the parent bells cannot overlap each other *a priori*. We shall denote the corresponding posterior distributions also by A and B, which should not cause any confusion.

To use the pairwise difference prior we fix γ_2 to one (for obvious identifiability reasons). Thus,

$$\pi_M(\boldsymbol{\alpha}) \propto \exp\bigl(-\gamma_1 \sum_{s\sim u} |\alpha(s) - \alpha(u)|\bigr), \tag{8.4}$$

and, therefore, the only control parameter is γ_1, the level of smoothing. We do not aim to estimate this parameter from the data in any way. Instead, we try a few values of γ_1 and examine the behaviour of the resulting posteriors. A reasonable strategy is to avoid oversmoothing while aiming to decrease the posterior variability by increasing γ_1.

We fitted to the data essentially the same noise model which was applied in the analysis of the sound-stimulus data. The only difference is that the space of voxels $\mathcal{S}$ is the 50×30 rectangle instead of a slice from the human brain. The interaction and precision parameters were calculated in a similar fashion as in Section 7. The spatial interaction parameter τ was estimated to be about -0.24.

To compare the posteriors mentioned above, it is necessary to make use of MCMC sampling. We applied the same sampling scheme of the cluster posterior as in Section 7, except that a displacement of a location of a parent was proposed from a neighbourhood which was one voxel smaller than the one used earlier. The MCMC sampling of the Markov posterior is quite straightforward since the posterior is a Markov random field. The details of the sampling can be found in Appendix A.3. We simulated Markov posteriors using a few values for γ_1. In these simulations the length of the burn-in period was 15000 sweeps and the chains were run for 300000 sweeps towards equilibrium. The cases $\gamma_1 = 1.0$ and $\gamma_1 = 2.0$ will be illustrated here and are denoted by C and D, respectively.

The simulation results are shown in Figures 8.3, 8.4 and 8.5. These maps indicate that the use of nonparametric Bayesian smoothing and point process prior models result in strikingly different posteriors. First, the mean map of the posterior C (Figure 8.3) is not as patterned as the mean of A and B. In particular, one cannot obtain any useful details of activations by inspecting the mean of C or D. The mean of A and B immediately suggest locations where activations might occur. Second, the voxelwise standard deviations of C and D are relatively homogeneous (Figure 8.4). However, in A and B the posterior variance is highly inhomogeneous and tends to be larger in the same locations where the posterior mean is nonzero. This phenomenon was observed also in Section 7. Comparing B and C in Figure 8.4, it can be seen that the variability of the posterior B is smaller than the variability of the posterior C in a large fraction of voxels.

Finally, maps of expected differences,

$$s \mapsto \sqrt{\mathrm{E}(\alpha(s) - \alpha^{(0)}(s))^2}, \tag{8.5}$$

possess clearly dissimilar features (Figure 8.5). These maps are of interest in a comparison study like this since they reveal information on the concentration of our posterior distributions around the true profile. An overall measure of concentration is the root mean squared deviation I_c:

$$I_c = \sqrt{(1/|S|)\sum_s \mathrm{E}\big(\alpha(s) - \alpha^{(0)}(s)\big)^2}. \tag{8.6}$$

Measured in this manner, the posteriors C and D were less concentrated than A and B. In fact, these indices of concentration were 0.50, 0.42, 0.83, and 0.60 for A, B, C, and D, respectively. As expected, for the posterior B, I_c was smaller than for A. The difference is explained by the fact that the prior of B is more informative than that of A.

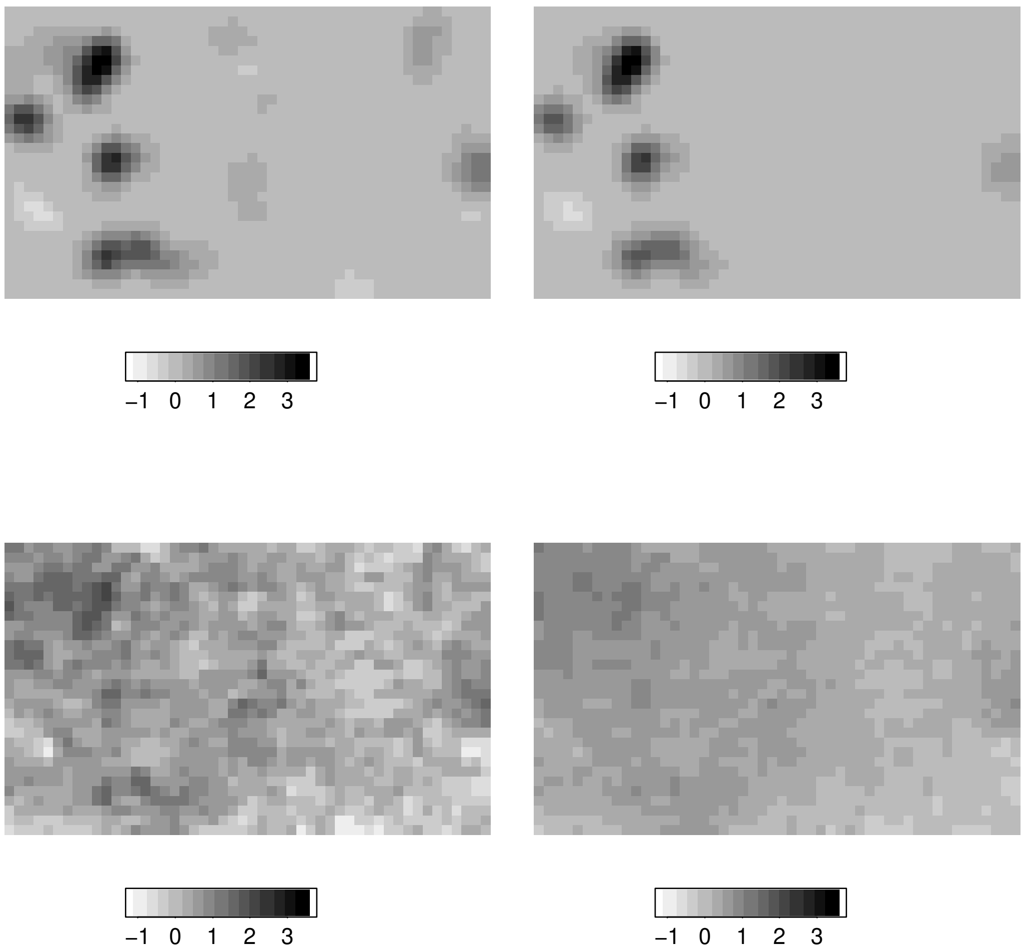

Figure 8.3 Mean maps for posteriors A (top left), B (top right), C (bottom left) and D (bottom right); see the text for the definitions.

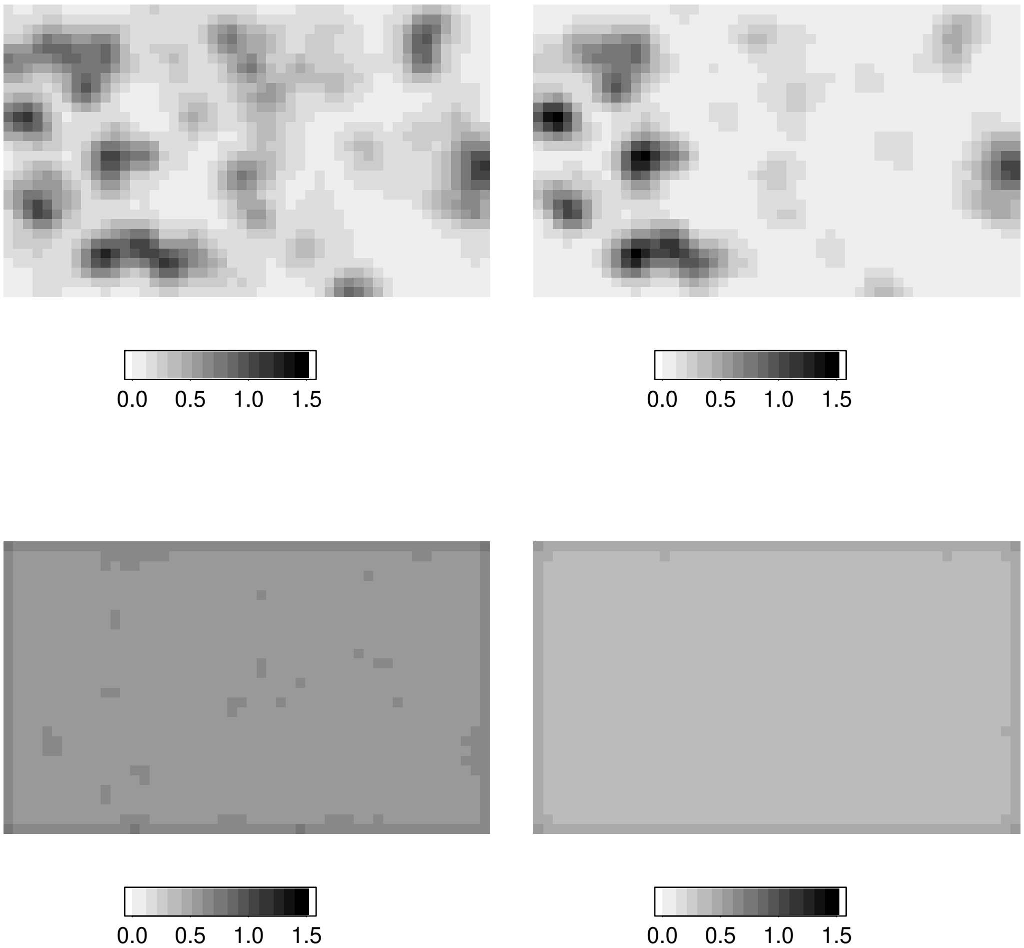

Figure 8.4 Maps of posterior standard deviations. The ordering of the maps is the same as in Figure 8.3.

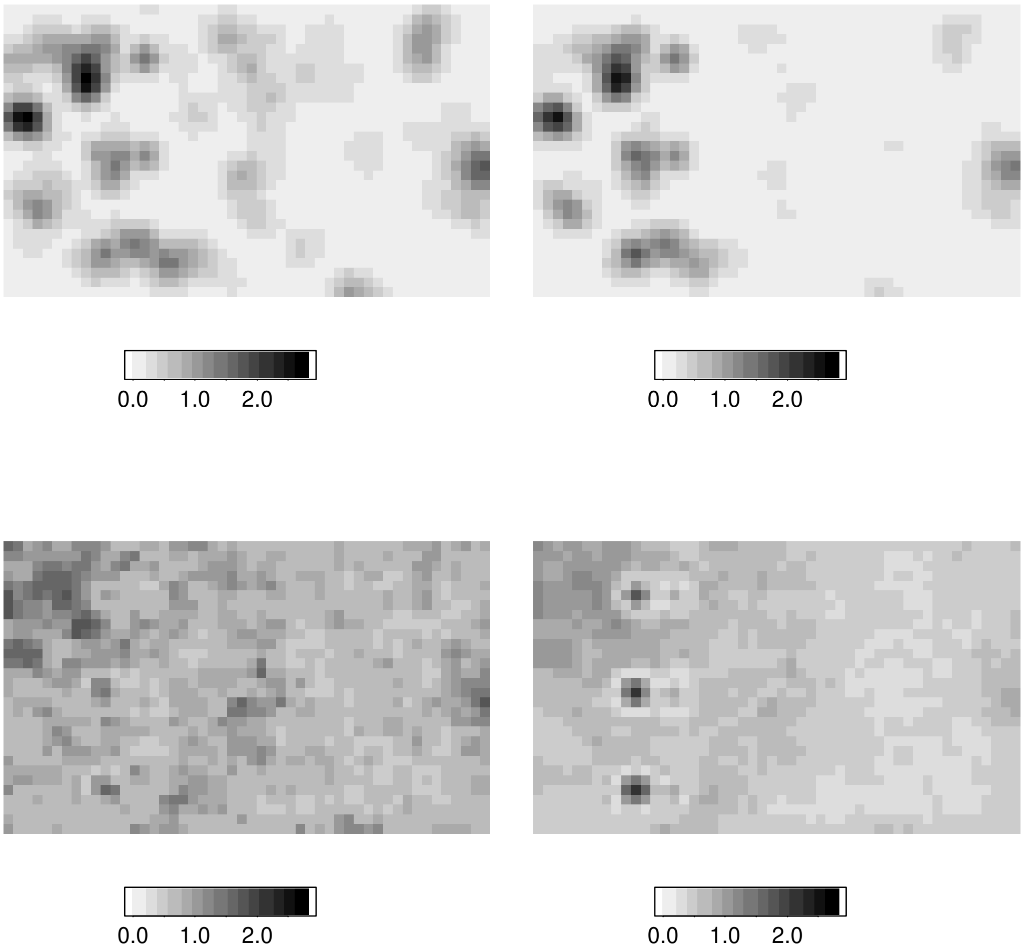

Figure 8.5 Maps of expected differences. The ordering of the maps is the same as in Figure 8.3.

The simulation results support our belief that in the context of fMRI, the use of high-level spatial models for brain activations can be more appropriate than a straightforward use of conventional Markov random field priors. One powerful feature of the cluster approach is that the spatial organization of activated brain voxels can be incorporated into the statistical analysis. Nonparametric smoothing techniques do not allow for such a possibility. Another interesting dissimilarity between the two prior modelling strategies also exists: the cluster prior can be used to model *negative association* between profile intensities, whereas pairwise difference priors always impose *positive associations* between intensities. This can be deduced as follows. Let us assume that two voxels, s_1 and s_2, are given from the search volume $\mathcal{S}$. Then, for a Markov prior the conditional prior probability

$$\Pr\big(\,\alpha(s_1)\ >\ 0 \mid \alpha(s_2)\,\big) \tag{8.7}$$

is an increasing function of $\alpha(s_2)$. For a cluster prior, in turn, this conditional probability can be larger when $\alpha(s_2)$ is zero than when $\alpha(s_2)$ is positive. In her thesis, van Lieshout (1994) discusses the notions of negative and positive associations generally in the context of *object processes*.

The bell model of Hartvig (1999) is closely related to our approach since it is also based on point processes. It would be natural to contrast the bell model with our approach. However, we shall not investigate the influence of prior models any more in this thesis. Nevertheless, it seems that some differences can be observed between the two approaches without carrying out any further simulations. Our view is that the bell prior focuses on small-scale modelling of clustering of voxels, whereas our cluster prior aims at a large-scale control of clustering. For example, let us assume that it is expected that there are at most about five activated regions in the cortex and that there is an upper bound for the spatial extent of these regions. In this instance, the bell prior can be used to describe only the probable range of the number of Gaussian bells but not that there are approximately *five* regions. Using the cluster prior, the probable number of regions provides the prior for the number of cluster centres and the extent of the regions guides the adjustment of the daughter process. We remark that if the inhibition in the bell model is strong, all the properties of clustering can be controlled perfectly. However, in that case the clusters will be restricted to have a strict parametric form, which may be criticized.

8.2 The role of spatial correlations in the scanning noise

In this section, we have focused on considering different prior models for activations and evaluating the posterior moments. We end this section by carrying out one more synthetic analysis to study the consequences of using an incorrect likelihood function.

In Section 3, we reviewed several Bayesian strategies to estimate brain activations and remarked that in some proposals the scanning noise was treated as a random process without any spatial autocorrelation structure. We investigated the importance of utilizing spatially correlated noise models by replacing the original likelihood $L(\boldsymbol{\alpha})$ by the modified likelihood function $\tilde{L}(\boldsymbol{\alpha})$ from Section 6:

$$\tilde{L}(\boldsymbol{\alpha}) \propto \exp\Big(-(1/2)\sum_{s} \tilde{\varrho}(s)^2 \big(z(s) - \alpha(s)\big)^2\Big). \tag{8.8}$$

We recall that substituting the modified likelihood $\tilde{L}$ to the formula of a posterior density corresponds to the use of a spatially independent noise model. To evaluate the role of the noise model, we employed the prior A and analysed the synthetic data once again. It is immediately observed from Figure 8.6 that both the means and standard deviations are altered when the likelihood is changed: the number of hills in the mean maps increases, and there is also an apparent decrease in the overall posterior variance. These results indicate that the spatial coherence of fMRI noise should not be ignored in the statistical analysis of functional data. If the spatial autocorrelation in noise is not accounted for, the posterior variability will be underestimated and the inferential conclusions tend to be too optimistic.

The ultimate reason for the relevance of modelling autocorrelations in fMRI analyses is that the interaction strength is comparable with the spatial extent of brain activations. If the activations covered large areas in a brain slice, the modelling of noise would be an issue of less importance. As this is not the case in fMRI, studies on exploring the statistical properties of noise and on modelling them are justified.

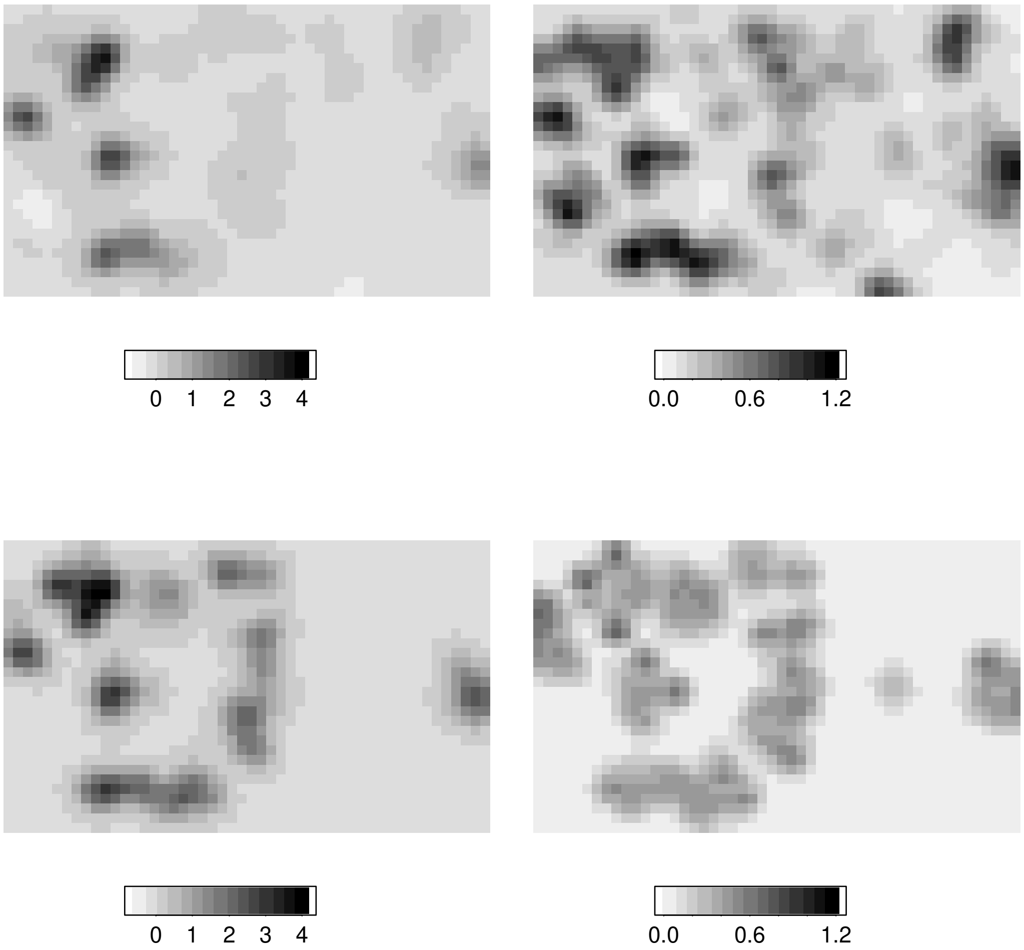

Figure 8.6 Upper row: posterior mean (left) and standard deviation (right) of A. Lower row: the moments of the modified posterior (based on the uncorrelated noise model) in the same order as in the upper row.

9 Discussion on modelling and computation

In Sections 7 and 8 we illustrated applications of our Bayesian model introduced in Sections 4 and 5. In this section, we comment upon the techniques used and discuss some alternative ideas that could be applied in prior modelling, in modelling noise processes and in MCMC simulations.

9.1 Prior distributions

The role of the parent points

In our formulation a profile is a superposition of parent and daughter bells in the following manner:

$$\alpha(s) = \eta_1\kappa_1(s - w_1) + \cdots + \eta_n\kappa_n(s - w_n) \tag{9.1}$$

$$= \sum_{i=1}^{n} \eta_i B_p(s - w_i) + \sum_{i=1}^{n}\sum_{j=1}^{n_i} \eta_i\zeta_{ij} B_d(s - (w_i + v_{ij})). \tag{9.2}$$

A natural question is what kind of a profile model could be constructed simply by defining a profile as a superposition of daughter bells and omitting the parents completely. Clearly, this could result in a less complex prior model for a profile. We show here that the parents have an important role in our model and that superpositions of only one type of bells do not have similar prior properties as the present model.

If the parent bells are omitted from the sum (9.1), we shall have an alternative model for profiles:

$$\tilde{\alpha}(s) = \sum_{i=1}^{n} \sum_{j=1}^{n_i} \eta_i \zeta_{ij} B_d(s - (w_i + v_{ij})). \tag{9.3}$$

A feature of this model is that the role of the parent centres is less pronounced. For example, $\tilde{\alpha}(w_i)$ may be null for some centres w_i. This does not fit well with our definition of the centre of a cluster. In our thinking, the purpose of the centre is to represent the location of the strongest haemodynamic response to stimulation. Another modification of the original profile model is to replace the products $\eta_i \zeta_{ij}$ that are used to scale the daughter bells by independent random variates. For example, one might model profiles by

$$\tilde{\alpha}(s) = \tilde{\eta}_1 B_d(s - \tilde{v}_1) + \cdots + \tilde{\eta}_m B_d(s - \tilde{v}_m), \tag{9.4}$$

where $\tilde{\eta}_i$ are (signed) independent variates and $\{\tilde{v}_1, \tilde{v}_2, \ldots, \tilde{v}_m\}$ follows some clustered point process. A defect in this approach is that the scaling factors $\tilde{\eta}_i$ are uncorrelated by construction, and so the signs $\tilde{\eta}_i$ and $\tilde{\eta}_j$ of two bells are not related in any way to the interdistance $\|\tilde{v}_i - \tilde{v}_j\|$ between the bells. Therefore, the resulting clustered spatial patterns are not as coherent *a priori* as in our formulation.

We conclude that the consequences of omitting some central parts of our construction can be considerable. An important feature of our model is that the daughter points are always nested in a parent configuration, that is, each daughter bell is linked directly to her parent bell. As a sidenote, we remark that *a priori* the complete configuration of all daughter centres $w_i + v_{ij}$ is actually a sample from a doubly stochastic Poisson process. In fact, the intensity function $\tilde{h}(s)$ of the daughter process, given a parent configuration $\boldsymbol{w}$, is

$$\tilde{h}(s) = h_v(s - w_1) + \cdots + h_v(s - w_n). \tag{9.5}$$

These kind of processes are also called *Cox processes* (Stoyan *et al.*, 1995).

Modelling of κ

A fundamental element in our prior for an activation profile is the concept of a cluster. We have been applying a specific model for a cluster. We expect, however, that other competitive models exist for this purpose. It is possible that many other prior models for κ which reflect the idea that κ is non-negative with its centre at the origin and with some prespecified maximal extent give similar posterior inference on profiles. We believe that, in general, local point configurations may be useful for parametrizing clusters. In our formulation, the daughter configurations $\boldsymbol{d}$ served for this purpose.

Modelling of interaction

We used an interaction function Ψ in the cluster prior (4.17) in order to force the centres w_i of x_i to inhibit mutually. We recall our choice for Ψ:

$$\Psi(x_i, x_j) = \begin{cases} (\|w_i - w_j\|/\rho)^p & \|w_i - w_j\| \le \rho, \\ 1 & \|w_i - w_j\| > \rho. \end{cases} \tag{9.6}$$

The inhibition between clusters was made to be dependent only on the intercentre distances. There are also some other interesting means of modelling the interaction. Baddeley & Lieshout (1993) list a variety of ways to encourage inhibiting configurations. Let

$$\mathcal{S}(x_i) = w_i + \{s \mid \kappa_i(s) \neq 0\} \tag{9.7}$$

denote the space occupied by a cluster x_i. An alternative choice is to make $\Psi(x_i, x_j)$ depend on whether the sets $\mathcal{S}(x_i)$ and $\mathcal{S}(x_j)$ overlap or not. Then, Ψ could be modelled by

$$\Psi(x_i, x_j) = \begin{cases} \vartheta & \mathcal{S}(x_i) \cap \mathcal{S}(x_j) \neq \emptyset, \\ 1 & \mathcal{S}(x_i) \cap \mathcal{S}(x_j) = \emptyset, \end{cases} \tag{9.8}$$

where ϑ is a constant between 0 and 1. It follows that in this formulation configurations are penalized according to the count of (unordered) pairs of overlapping clusters. A slightly different choice for the interaction function could be

$$\Psi(x_i, x_j) = \vartheta^{|\mathcal{S}(x_i) \cap \mathcal{S}(x_j)|}, \tag{9.9}$$

where $|\mathcal{S}(x_i) \cap \mathcal{S}(x_j)|$ is the area (or volume) of the intersection. With this interaction function, the strength of inhibition between two clusters depends on how much they overlap. A common feature of the interaction functions above is that they lead to priors with only pairwise interactions. If the product $\prod_{i<j} \Psi(x_i, x_j)$ in the prior (4.17) is replaced by

$$\varsigma^{|\mathcal{S}(x_1) \cup \mathcal{S}(x_2) \cup \cdots \cup \mathcal{S}(x_n)|}, \tag{9.10}$$

where ς is a constant, we get a point process with higher-order interactions. To favour configurations x with distinct clusters, the constant ς should be fixed to be larger than 1.

The cluster model is closely related to the object processes. The forms of interaction discussed in Baddeley & Lieshout (1993) do not assume any parametrization unlike our suggestion. Our approach is easier to implement since we have to calculate neither areas of pairwise intersections of sets nor areas of unions of planar sets.

Modelling of null intensities

Our last comment upon priors concerns the Markov random field models discussed in Section 8. In the analysis of the synthetic data, we used pairwise difference priors to encourage spatially smooth profiles. As explained in the previous sections, our belief is that the profiles are not only smooth but also strongly spatially organized. Therefore, we expect that an activation profile may contain a considerable amount of null intensities. However, a smoothing prior (8.1) cannot be used to model the presence of nulls in an image. We suggest here how to modify the density (8.1) so that it could be used to model *both* smoothness *and* the null intensities.

Let us consider an *atomic* smoothing prior which has density

$$\pi^*(\boldsymbol{\alpha}) \propto \sigma^{-nz(\boldsymbol{\alpha})} \exp\Big(-\gamma_1 \sum_{s\sim u} |\alpha(s) - \alpha(u)|/\sigma\Big) \ \exp\Big(-(1/2)\sum_s (\gamma_3 \alpha(s)/\sigma)^2\Big) \tag{9.11}$$

with respect to an *atomic* reference measure $\prod_s m^*(d\alpha(s))$ where $m^* = \delta_0 + m$ and δ_0 is the Dirac measure with unit mass at the origin. Here, $nz(\boldsymbol{\alpha})$ is the number of *nonzero* components in $\boldsymbol{\alpha}$, and σ is a scale parameter of this model (Appendix A.3). To understand the role of γ_3, it is convenient to consider the case $\gamma_1 = 0$. Then, the components $\alpha(s)$ of $\boldsymbol{\alpha}$ are independent and the density (9.11) is proportional to

$$\sigma^{-nz(\boldsymbol{\alpha})} \exp\Big(-(1/2)\sum_s (\gamma_3 \alpha(s)/\sigma)^2\Big). \tag{9.12}$$

Now, γ_3 controls the number of null intensities. In fact, if ϕ is a (chosen) probability level, the choice

$$\gamma_3 = \phi\sqrt{2\pi}/(1-\phi) \tag{9.13}$$

corresponds to a spatial prior model for which a component $\alpha(s)$ is null with probability ϕ. Also, the conditional variance of $(\alpha(s)|\alpha(s) \neq 0)$ is σ^2/γ_3^2 for all s. It is obvious that large positive values of γ_1 enforce spatial ordering to intensity values $\alpha(s)$. Unfortunately, when γ_1 is not zero, the interpretation of γ_3 is not as clear as in (9.12) because of the interactions among the components $\alpha(s)$. In order to adjust the fraction of null intensities (that is, non-activated voxels) a priori, it is necessary to simulate the density (9.11) to find a corrected value for γ_3.

We expect that high-level strategies are more suitable than pixel-based low level models in fMRI problems. Nevertheless, a simple model like (9.11) is easy to implement and results of high-level analyses can be contrasted with those of less involved analyses.

9.2 Likelihood

Our likelihood function is based on a Gaussian conditional autoregressive random field (CAR) model. We recall the density of the spatial noise field $\boldsymbol{\epsilon}$ is proportional to

$$\exp\Big(-\frac{1}{2}\Big(\sum_s \varrho(s)^2\epsilon(s)^2 + \sum_l \tau_l \sum_{s-u=l} \varrho(s)\varrho(u)\epsilon(s)\epsilon(u)\Big)\Big) \tag{9.14}$$

We shall now discuss the Bayesian estimation of the parameters of this model and possibilities to generalize the noise model.

In Section 5, we suggested how the (nuisance) parameters $\boldsymbol{\varrho}$ and $\boldsymbol{\tau}$ can be estimated from the data. We also remarked that the uncertainty in these estimates can usually be ignored. The reason for this is that in typical fMRI studies the number of scans T is so large that there are enough degrees of freedom for a reasonably accurate estimation. However, if the scanning series are short, it may be appropriate to account for the uncertainty in $\boldsymbol{\varrho}$ and $\boldsymbol{\tau}$. In principle, the uncertainty could be taken into account by assigning priors to $\boldsymbol{\varrho}$ and $\boldsymbol{\tau}$. For example, a pairwise difference smoothing prior could be used for $\boldsymbol{\varrho}$ to discourage abrupt spatial changes in $\varrho(s)$. While this kind of fully Bayesian approach would be philosophically satisfactory, computational difficulties will arise. The updating of $\boldsymbol{\tau}$ will be a problematic issue since the normalizing constant of the density (9.14) depends in a complicated way on $\boldsymbol{\tau}$ and is actually analytically intractable. Therefore, to update $\boldsymbol{\tau}$, computationally intensive techniques for estimating normalizing factors must be applied (Higdon, 1994). Interestingly, the scaling parameter $\boldsymbol{\varrho}$ can be updated without any difficulties (if $\boldsymbol{\tau}$ is kept fixed) since then the normalizing constant of (9.14) is proportional to the product $\prod_s \varrho(s)$ (Taskinen, 1998).

An underlying assumption in the use of the noise model (9.14) is that the spatial correlation structure in fMRI data is almost homogeneous. In other words, the size of spatial correlations is at the same level all around the brain slices. It follows that, irrespective of the degree of nonstationarity in data, the model will be fitted to match the average size of the spatial correlations. What is interesting is that the CAR model *could* be modified to accomodate nonstationary spatial correlations. Replacing the interaction parameter $\boldsymbol{\tau} = (\tau_l)$ by a new interaction term $\boldsymbol{\tau} = (\tau_{su})$ in (9.14), we could define a new spatial process, the density of which is proportional to

$$\exp\Big(-\frac{1}{2}\Big(\sum_s \varrho(s)^2\epsilon(s)^2 + \sum_{s,u} \tau_{su}\varrho(s)\varrho(u)\epsilon(s)\epsilon(u)\Big)\Big). \tag{9.15}$$

In order to apply this more general noise model, it would be valuable to have some auxiliary spatial covariate information which could be utilized when estimating the interaction structure. At present, there does not seem to exist any imaging modality which could give such information. Therefore, we have chosen to consider only homogeneous noise models in this thesis.

An important aspect of the CAR model is that it is a Markov random field meaning that the noise $\epsilon(s)$ in a voxel s depends on other noise terms only through those terms $\epsilon(u)$ where u is close to s. These voxels u constitute the *neighbourhood* of s and it is determined by the spatial lags l. This conditional independence is not important *per se*, but it simplifies the updating formulas in our MCMC computations. We also note that the Markov property of the noise model has an important role when spatially unstructured smoothing priors (8.1) for profiles are used. In that case, the posterior is also a Markov random field, which facilitates the sampling of the posterior.

9.3 Sampling issues

Our posterior $p(\boldsymbol{x})$ is a high-dimensional nonstandard distribution, and, apparently, there is no direct simulation method for it. Recognizing this, we applied a special case of the Geyer-Møller simulation algorithm. An alternative to the MCMC algorithm of Section 6 could be an application of the importance reweighting technique (Gilks *et al.*, 1996). In our situation, importance sampling could be used in the following way. Let us suppose that g is some relevant function of $\boldsymbol{x}$ and that the expected value of $g(\boldsymbol{x})$ must be calculated. To this end, a collection of samples $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_M$ could be drawn from some density $p^*(\boldsymbol{x})$ and the posterior expectation of $g(\boldsymbol{x})$ could be estimated from the importance reweighting formula

$$\widehat{\mathrm{E}_p(g(\boldsymbol{x}))} \;=\; (1/M) \sum_{m=1}^{M} g(\boldsymbol{x}^{(m)}) p(\boldsymbol{x}^{(m)}) / p^*(\boldsymbol{x}^{(m)}). \tag{9.16}$$

What is appealing, our prior density $\pi(\boldsymbol{x})$ (see (4.17)) could be used as the density p^* since π is easy to simulate. We recall that the simulation of π consists of two steps. First, a configuration of parent centres $\boldsymbol{w} = \{w_1, \ldots, w_n\}$ is simulated from an inhibitory point process (see Appendix A.2). Second, marks $(\eta_i, \boldsymbol{d}_i)$ are drawn from the mark distribution independently of each other and are assigned to each centre w_i. Thus, the simulation procedure is efficient. A disappointing feature of this algorithm is that the standard errors of the estimates (9.16) will be large unless the prior π and the posterior p are nearly identical. This happens exactly when the data contains only a small amount of information. We expect that this requirement severely limits the use of importance sampling.

As already mentioned, our MCMC algorithm is a special case of the general Geyer-Møller algorithm. Our sampler could be modified at least in two ways. First, when proposing a change in the number of clusters, we considered insertions and deletions equally likely. In other words, an insertion was always proposed with probability 0.50. However, the probability to propose a new cluster could depend on the current configuration $\boldsymbol{x}$, and, in particular, on the prior probability $\psi(n(\boldsymbol{x}))$. A minor correction to the Metropolis-Hastings acceptance probabilities is sufficient if the proposal mechanism is modified in this way. Second, when we had decided to propose a cluster to be deleted, each cluster had the same probability to be removed. In other words, the removal mechanism did not depend on the characteristics of the clusters x_i. Alternatively, we could, for example, propose to delete more often clusters x_i which have a small height η. Then, small clusters would be proposed to be deleted with higher probability than large ones. Exact recommendations for tuning the sampler are very difficult to give since the posterior depends on data, and the properties of the data may vary. There are also other methods for simulating point processes than the Metropolis-Hastings algorithm by Geyer & Møller (1994). Spatial birth-and-death processes provide an important class of Markov chains that could also be utilized (Preston, 1975). An interesting difference between a Metropolis-Hastings sampler and a birth-and-death sampler deserves to be mentioned here. When a Metropolis-Hastings algorithm is used, the number of points $n(\boldsymbol{x})$ may well remain unchanged. On the contrary, it is possible to construct a birth-and-death sampler for which the number of points always either increases or decreases when updating (Stoyan *et al.*, 1995).

10 Statistics and fMRI: concluding remarks

The main contribution in this thesis is a new Bayesian modelling technique for estimating activation profiles in the human brain. The proposed Bayesian approach is a structural model based on marked Gibbs point processes. There are three main reasons for using a structural strategy to model profiles. First, the experimental designs applied often provide useful prior information on the neuronal processing of interest. Second, expert knowledge on various characteristics of brain activations can be incorporated into statistical analysis conveniently. Third, posterior inference can be made both on activation profiles and related parameters like centres of neuronal activation. Our view is that the suggested prior model can be regarded as a candidate solution to the problem of searching for "models parametrizing the activation foci", an issue posed by Petersson *et al.* (1999).

Our focus was to consider only profiles with respect to which the likelihood (or data) is not highly informative. If a profile can be estimated from the data with high precision, the application of sophisticated Bayesian techniques becomes more or less inappropriate since the posterior will not be sensitive to the prior. Considering the utilization of designs such as (4.1), it is obvious that the standard errors of (classical) estimates of $\alpha(s)$ depend on p, the number of scans per an individual design, and that they decrease as p increases. Since the total number of scans $T = mp$ is not usually very large in fMRI experiments, it follows that the errors in the estimates cannot be very small, particularly when m is large. For a Bayesian, this means that as m increases (keeping T fixed) the domination of data decreases in the posterior and that the posterior becomes more sensitive to prior information on the activation processes. Consequently, when applying concatenated designs, Bayesian inferential conclusions may differ from those from classical analyses, which cannot utilize prior knowledge as elegantly.

The Bayesian paradigm provides a natural framework for modelling activation processes in neuroimaging problems. In this context, an especially elegant feature of Bayesian thinking is that it is not necessary to assume the number of parameters to be known, typical in frequentist analyses. The number of parameters or objects can be assigned a prior probability distribution which makes it possible to combine efficiency and flexibility, the ability to describe varying forms of activations. Sometimes, Bayesian techniques can be viewed as ways for controlling the effective degrees of freedom (Green & Silverman, 1994), and nonparametric Bayesian smoothing, for example, can be seen from this perspective.

When analysing image data using Bayesian methods, a structured approach can be advantageous both at the prior modelling stage and in inferential steps. First, the structure facilitates incorporating prior information into the subsequent inference. Second, the parameters of the model can be thought of extracting information from the image of interest. All the questions we are interested in can then be directly answered using the posterior of the parameters. In particular, the inference principles are unified. The subjective nature of the Bayesian paradigm is sometimes considered to be a matter of concern. However, as argued by Bayesian statisticians, all model-building is based on personal (or subjective) judgement to some extent, and thus in a strict sense "objective inference" does not really exist. Bayesian approaches are becoming more popular in natural sciences – this tendency being supported by increases in computer power.

The development of statistical methods for FNI should follow advances in neuroscientific methodology. For example, most functional investigations now are off-line studies, which means that data are processed after a scanning operation has been completed. As noted by Lange *et al.* (1999), a benefit in an off-line processing is that the same data can be inspected using several different strategies, which potentially reveals more information on the brain function than any single analysis. Interestingly, on-line studies may become more common in the next few years. The idea of real-time processing has many advantages, of which Cox *et al.* (1995) point out a few. For example, the quality of data can be directly assessed and reacquisitions can be carried out if found necessary. What seems promising from the neuroscientific point of view, the experimental design can be adjusted during the acquisition, and it is possible to arrange the experiment to be interactive in the sense that the stimulus will depend on the brain response history of a test person. In the pioneering work by Cox *et al.* (1995), a recursive computational method was suggested for revealing activation processes in almost real-time. It remains to be seen in what form Bayesian ideas can be utilized in on-line analyses. Our proposal, in its present form, is not adequate for such a use.

As neuroimaging continues its rapid development, new questions will certainly arise, and fast on-line processing comprises perhaps just one challenge among many. It is expected that FNI will be an inspiring field for applied scientists and that the interplay between FNI and statistical science will be fruitful for both also in the future.

A Appendix

A.1 Generalized least squares estimation

In Section 5, it was claimed that the generalized least squares (GLS) estimation of $\phi(s)$ and $\beta(s)$ can be carried out in a voxelwise manner under the separability assumption. In other words, the estimates of these parameters depend only on the observations in the voxel time series $Y(s)$. To show this, let $\varpi(s) = (\phi(s), \beta(s))$ be a single parameter containing all the voxel parameters. We shall treat the voxel time series $Y(s)$ and the noise time series $\delta(s)$ as column vectors:

$$Y(s) = (Y_1(s), \dots, Y_T(s))^T, \tag{A.1}$$

$$\delta(s) = (\delta_1(s), \dots, \delta_T(s))^T. \tag{A.2}$$

The data Y, the coefficients $\boldsymbol{\varpi}$ and the noise term $\boldsymbol{\delta}$ will be represented as vectors

$$Y = (Y(s_1)^T, \dots, Y(s_{|S|})^T)^T, \tag{A.3}$$

$$\boldsymbol{\varpi} = (\varpi(s_1)^T, \dots, \varpi(s_{|S|})^T)^T, \tag{A.4}$$

$$\boldsymbol{\delta} = (\delta(s_1)^T, \dots, \delta(s_{|S|})^T)^T. \tag{A.5}$$

Let $J = (1, \dots, 1)^T$ be column vector of size p. Then, the matrix $\boldsymbol{X}$ in the equation (5.1) is

$$\boldsymbol{X} = \left[\boldsymbol{X}^{(\phi)}\, \boldsymbol{X}^{(\beta)}\right] = \left[I_m \otimes J \;\, I_m \otimes X\right], \tag{A.6}$$

where I_m is the identity matrix of size m and $X = (X_t)_{t=1}^{p}$. A spatio-temporal version of our linear model (5.1) can be expressed as

$$Y = \tilde{\boldsymbol{X}}\, \boldsymbol{\varpi} + \boldsymbol{\delta}, \tag{A.7}$$

where $\tilde{\boldsymbol{X}}$ is the Kronecker product $I \otimes \boldsymbol{X}$ of the identity matrix I of size $|S|$ and $\boldsymbol{X}$. Assuming the space-time covariance matrix C of $\boldsymbol{\delta}$ is known, the generalized least squares estimate for $\boldsymbol{\varpi}$ is given by

$$\hat{\boldsymbol{\varpi}} = (\tilde{\boldsymbol{X}}^T C^{-1} \tilde{\boldsymbol{X}})^{-1} \tilde{\boldsymbol{X}}^T C^{-1} Y \tag{A.8}$$

If C is separable, that is, $C = W \otimes V$, the estimate of $\boldsymbol{\varpi}$ is

$$(\tilde{\boldsymbol{X}}^T C^{-1} \tilde{\boldsymbol{X}})^{-1} \tilde{\boldsymbol{X}}^T C^{-1} Y = \tag{A.9}$$
$$((I \otimes \boldsymbol{X}^T)(W^{-1} \otimes V^{-1})(I \otimes \boldsymbol{X}))^{-1} (I \otimes \boldsymbol{X}^T)(W^{-1} \otimes V^{-1}) Y = \tag{A.10}$$
$$(I \otimes ((\boldsymbol{X}^T V^{-1} \boldsymbol{X})^{-1} \boldsymbol{X}^T V^{-1})) Y \tag{A.11}$$

and therefore

$$\hat{\varpi}(s) = (\boldsymbol{X}^T V^{-1} \boldsymbol{X})^{-1} \boldsymbol{X}^T V^{-1} Y(s). \tag{A.12}$$

A.2 MCMC estimation of ψ^* and h^*

In Section 4 we introduced a parent centre process with density

$$f_{\boldsymbol{w}}(\boldsymbol{w}) \propto n(\boldsymbol{w})!\, \psi^*(n(\boldsymbol{w})) \prod_{i<j} \Psi(w_i, w_j) \tag{A.13}$$

with respect to an inhomogenous unit rate Poisson process with an intensity function h^*. We show here how MCMC methods can be used to find a discrete density ψ^* and an intensity function h^* such that the number of centres $n(\boldsymbol{w})$ and the location of a random centre w follow prespecified ψ and h.

We discuss first how to sample the density (A.13). The sampling scheme of Geyer & Møller (1994) can be applied here. We shall use two updating rules: displacement of a random centre and insertion/removal of a random centre. Let $\boldsymbol{w} = \{w_1, w_2, \ldots, w_n\}$ be the current configuration. To displace a random centre w_i, we propose a new location w^* for it from h^*. The acceptance ratio is then

$$MHR_d = \prod_{j \neq i} \frac{\Psi(w^*, w_j)}{\Psi(w_i, w_j)}. \tag{A.14}$$

If $\boldsymbol{w}$ is empty, we do nothing. To change the number of centres in the configuration $\boldsymbol{w}$, we first choose whether to insert or remove, both operations being equally likely. If we decide to insert a new centre, we sample a centre w^* from h^*. The acceptance ratio is

$$MHR_i = \frac{\psi^*(n+1)}{\psi^*(n)} \prod_{j=1}^{n} \Psi(w^*, w_j). \tag{A.15}$$

In case of removal, we select a centre w_i from the configuration $\boldsymbol{w}$ at random. The acceptance ratio is now

$$MHR_r = \frac{\psi^*(n-1)}{\psi^*(n)\prod_{j=1}^{n}\Psi(w^*, w_j)}. \tag{A.16}$$

If the configuration $\boldsymbol{w}$ is empty, we do nothing.

The corrected values for ψ^* and h^* can be estimated by trial and error. First, some initial values are chosen. For example, $\psi^{(0)} = \psi$ and $h^{(0)} = h$. Then, samples of the centre process are used to estimate $\tilde{\psi}^{(1)}$, the distribution of counts of the model, and $\tilde{h}^{(1)}$, the density of a random centre. These MCMC estimates can be used to find new iterated values, $\psi^{(1)}$ and $h^{(1)}$, and the procedure can be repeated m times until $\tilde{\psi}^{(m)} \approx \psi$ and $\tilde{h}^{(m)} \approx h$. It is useful to recognize that the ratio $\psi^{(k-1)}\psi/\tilde{\psi}^{(k)}$ can provide a suitable value for $\psi^{(k)}$ at iterate k. Although this kind of estimation method is quite elementary, it is usually practical since the simulation of the model (A.13) is typically very time-efficient.

A.3 Some notes on Markov random fields

Sampling of Markov posteriors

In Section 8 we simulated the posterior distribution of a profile $\boldsymbol{\alpha}$ which was a Markov random field. We used single-voxel updates in sampling. For the MCMC sampling of Markov posteriors we used the proposal density

$$q(\alpha(s)^*|\alpha(s)) \propto \exp\big(-(1/2)(\alpha(s)^* - \alpha(s))^2/\sigma_0^2\big), \tag{A.17}$$

where σ_0 is a tuning constant of the proposal density. The Metropolis-Hastings ratio is the product of the likelihood ratio $L(\alpha(s)^*)/L(\alpha(s))$, prior ratio $\pi(\alpha(s)^*)/\pi(\alpha(s))$ and the proposal ratio $q(\alpha(s)|\alpha(s)^*)/q(\alpha(s)^*|\alpha(s))$. Since the proposal mechanism q is symmetric, the last ratio is identically one. The prior ratio $\pi(\alpha(s)^*)/\pi(\alpha(s))$ is

$$\exp\Big(\gamma_1 \sum_{s\sim u}\big(|\alpha(s) - \alpha(u)| - |\alpha(s)^* - \alpha(u)|\big)\Big). \tag{A.18}$$

Atomic models

In Section 9 we suggested an atomic smoothing prior (see equation (9.11)). It was claimed that the parameter σ is a scaling parameter of the density π^*. To show this, we prove the following simple lemma.

Lemma Let $m_k^* = \prod_{i=1}^k(\delta + m)$ be a product measure and let $nz(\boldsymbol{\theta})$ be the number of nonzero components of $\boldsymbol{\theta}$. Then

$$\int \sigma^{-nz(\boldsymbol{\theta})} g(\boldsymbol{\theta}/\sigma)\; dm_k^*(\boldsymbol{\theta}) = \int g(\boldsymbol{\theta})\; dm_k^*(\boldsymbol{\theta}) \tag{A.19}$$

for all positive σ and positive functions g.

Proof We proceed using induction. The case $k = 1$ is trivial. Let $\boldsymbol{\theta} = (\tilde{\theta}, \theta)$ where $\tilde{\theta} = (\theta_1, \ldots, \theta_k)$. Assume that the formula holds for k. Noting that $nz(\boldsymbol{\theta}) = nz(\tilde{\theta}) + nz(\theta)$, we get

$$\int \sigma^{-nz(\boldsymbol{\theta})} g(\boldsymbol{\theta}/\sigma)\; dm_{k+1}^*(\boldsymbol{\theta}) = \int \sigma^{-nz(\tilde{\theta})} \left(\int \sigma^{-nz(\theta)} g(\tilde{\theta}/\sigma, \theta/\sigma)\; dm^*(\theta) \right) dm_k^*(\tilde{\theta}) \tag{A.20}$$

$$= \int \sigma^{-nz(\tilde{\theta})} \left(\int g(\tilde{\theta}/\sigma, \theta)\; dm^*(\theta) \right) dm_k^*(\tilde{\theta}) \tag{A.21}$$

$$= \int \left(\int \sigma^{-nz(\tilde{\theta})} g(\tilde{\theta}/\sigma, \theta)\; dm_k^*(\tilde{\theta}) \right) dm^*(\theta) \tag{A.22}$$

$$= \int \left(\int g(\tilde{\theta}, \theta)\; dm_k^*(\tilde{\theta}) \right) dm^*(\theta) \tag{A.23}$$

$$= \int g(\boldsymbol{\theta})\; dm_{k+1}^*(\boldsymbol{\theta}), \tag{A.24}$$

which proves the lemma.

Now, let f be the density (with respect to m_k^*) of a random variable $\boldsymbol{\theta}$. Then,

$$\Pr(\sigma\boldsymbol{\theta} \in \mathcal{E}) = \Pr(\boldsymbol{\theta} \in \sigma^{-1}\mathcal{E}) = \int_{\sigma^{-1}\mathcal{E}} f(\boldsymbol{\theta})\; dm_k^*(\boldsymbol{\theta}) \tag{A.25}$$

$$= \int_{\mathbb{R}^k} 1_{\sigma^{-1}\mathcal{E}}(\boldsymbol{\theta}) f(\boldsymbol{\theta})\; dm_k^*(\boldsymbol{\theta}) \tag{A.26}$$

$$= \int_{\mathbb{R}^k} 1_{\sigma^{-1}\mathcal{E}}(\boldsymbol{\theta}/\sigma) f(\boldsymbol{\theta}/\sigma) \sigma^{-nz(\boldsymbol{\theta})}\; dm_k^*(\boldsymbol{\theta}) \tag{A.27}$$

$$= \int_{\mathcal{E}} \sigma^{-nz(\boldsymbol{\theta})} f(\boldsymbol{\theta}/\sigma)\; dm_k^*(\boldsymbol{\theta}), \tag{A.28}$$

which confirms our claim.

For MCMC sampling of an atomic posterior, the same proposal density (A.17) can be used but now with respect to m^*. Then, our proposal $\alpha(s)^*$ comes from a certain mixture of a normal distribution and Dirac degenerate distribution. In fact, after some computations it follows that $\alpha(s)^*$ is zero with probability

$$\frac{1}{1 + \sigma_0\sqrt{2\pi}\exp(\frac{1}{2}\alpha(s)^2/\sigma_0^2)} \tag{A.29}$$

or otherwise a sample from $N(\alpha(s), \sigma_0^2)$. Also, this choice leads to the proposal ratio

$$\frac{q(\alpha(s)|\alpha(s)^*)}{q(\alpha(s)^*|\alpha(s))} = \frac{\sigma_0\sqrt{2\pi} + \exp(-\frac{1}{2}\alpha(s)^2/\sigma_0^2)}{\sigma_0\sqrt{2\pi} + \exp(-\frac{1}{2}(\alpha(s)^*)^2/\sigma_0^2)}. \tag{A.30}$$

It is interesting that now the proposal ratio is not identically one although we are using centred Gaussian densities. This is a consequence of using a reference measure that contains atoms. The prior ratio $\pi^*(\boldsymbol{\alpha}^*)/\pi^*(\boldsymbol{\alpha})$ of an atomic smoothing prior differs slightly from that of a conventional smoothing prior. If the absolute value potential is used, the prior ratio is

$$\begin{aligned} &\sigma^p \exp\Big(\gamma_1/\sigma \sum_{s\sim u}\big(|\alpha(s) - \alpha(u)| - |\alpha(s)^* - \alpha(u)|\big)\Big) \\ &\times \exp\Big((1/2)\gamma_3^2((\alpha(s)^*)^2 - \alpha(s)^2)/\sigma^2\Big), \end{aligned} \tag{A.31}$$

where the exponent p is 1, 0 or -1 depending on whether the number of nonzero components of $\boldsymbol{\alpha}^*$ is smaller, equal to or higher than that of $\boldsymbol{\alpha}$.

A.4 Convergence results

We shall here justify the convergence of our MCMC sampler constructed in Section 6. Our sampler is a mixture of five transition rules. We can update the height η of a random cluster, heights ζ_j of all daughters in a random cluster, location w of a random parent centre, insert/remove a daughter from a random cluster and finally insert/remove a random cluster. According to the results of Geyer & Møller (1994), the last updating rule is reversible and irreducible. Therefore, to prove the convergence, it is enough to show that the first four rules are reversible. We prove here that inserting/removing daughters and updating heights of daughters are reversible transitions. The proofs for the other rules are very similar. To prove the reversibility, let $x = (w; (\eta, \boldsymbol{d}))$ be a cluster from a configuration $\boldsymbol{x}$. We aim to show that we can update the conditional posterior distribution of $\boldsymbol{d}$, given $\boldsymbol{x} \setminus x$, w and η, in a reversible manner. Let us denote $p((\boldsymbol{x} \setminus x) \cup (w; (\eta, \boldsymbol{d})))$ by $f(\boldsymbol{d})$. Then, we have to show that the integral

$$\int_{\mathcal{D}} \int_{\mathcal{D}} \chi_{\mathcal{A}\times\mathcal{B}}(\boldsymbol{d}, \boldsymbol{d}^*) A(\boldsymbol{d}^* \,|\, \boldsymbol{d}) f(\boldsymbol{d}) \; dP(\boldsymbol{d}^* \,|\, \boldsymbol{d}) d\lambda_d(\boldsymbol{d}) \tag{A.32}$$

is symmetric in $\mathcal{A}$ and $\mathcal{B}$ where $P(\cdot|\,\boldsymbol{d})$ is one of the two proposal measures for daughter configurations $\boldsymbol{d}$, $A(\boldsymbol{d}^* \,|\, \boldsymbol{d})$ is the corresponding acceptance probability and $\mathcal{D} = \Omega(\mathcal{U} \times \mathbb{R}^+)$.

To begin with, we note that removing and deleting daughters from a fixed cluster x is clearly reversible with respect to $f(\boldsymbol{d})$ since this rule is based on the Geyer-Møller acceptance probabilities. We shall now consider the updating of heights of daughter bells in more detail. To this end, let $\mathcal{A}$ and $\mathcal{B}$ be two collections of configurations with the same number of points in the configurations, say k. In this case, the proposal measure P reduces to a k-dimensional proposal density $q(\zeta^*|\zeta)$ and the integral (A.32) has form

$$\int_{\mathcal{D}} \int_{R^k} \chi_{\mathcal{A}\times\mathcal{B}}(\boldsymbol{d}, \boldsymbol{d}^*) A(\boldsymbol{d}^* \mid \boldsymbol{d}) f(\boldsymbol{d}) q(\zeta^*|\zeta)\ d\zeta_1^* \ldots d\zeta_k^*\ d\lambda_d(\boldsymbol{d}). \tag{A.33}$$

The roles of $\mathcal{A}$ and $\mathcal{B}$ are clearly symmetric if they are both empty sets or contain only the empty configuration ($k = 0$). Let $k > 0$ and use notation $g(\boldsymbol{d}, \boldsymbol{d}^*) = A(\boldsymbol{d}^* \mid \boldsymbol{d}) f(\boldsymbol{d}) q(\zeta^*|\zeta)$. Then the integral

$$\int_{\mathcal{D}} \int_{R^k} \chi_{\mathcal{A}\times\mathcal{B}}(\boldsymbol{d}, \boldsymbol{d}^*) g(\boldsymbol{d}, \boldsymbol{d}^*)\ d\zeta_1^* \ldots d\zeta_k^*\ d\lambda_d(\boldsymbol{d}) \tag{A.34}$$

equals to

$$\frac{e^{-1}}{k!} \int_{(R^2)^k} \left[\int_{R^{2k}} \chi_{\mathcal{A}\times\mathcal{B}}(\boldsymbol{d}, \boldsymbol{d}^*) g(\boldsymbol{d}, \boldsymbol{d}^*) h_\zeta(\zeta_1) \cdots h_\zeta(\zeta_k)\ d\zeta_1^* \cdots d\zeta_k^*\ d\zeta_1 \cdots d\zeta_k \right] \\ h_v(v_1) \cdots h_v(v_k)\ dv_1 \cdots dv_k \tag{A.35}$$

The reversibility condition is satisfied if

$$g(\boldsymbol{d}, \boldsymbol{d}^*) h_\zeta(\zeta_1) \cdots h_\zeta(\zeta_k) \tag{A.36}$$

is symmetric in (ζ, ζ^*) for any given v. This holds if we use a Metropolis-Hastings type acceptance probability

$$A(\boldsymbol{d}^* \mid \boldsymbol{d}) = \min\left(1, \frac{f(\boldsymbol{d}^*) q(\zeta|\zeta^*) h_\zeta(\zeta_1^*) \cdots h_\zeta(\zeta_k^*)}{f(\boldsymbol{d}) q(\zeta^*|\zeta) h_\zeta(\zeta_1) \cdots h_\zeta(\zeta_k)}\right). \tag{A.37}$$

In conclusion, we have shown that the reversibility holds if a fixed cluster x is considered. It follows from this that the reversibility property holds also (now with respect to $p(\boldsymbol{x})$) when the cluster x is chosen *randomly* from $\boldsymbol{x}$.

Bibliography

Adler, R. J., (1981): *The geometry of random fields*. Wiley Series in Probability and Mathematical Statistics. Wiley.

Alenius, S. & Ruotsalainen, U., (1997): Bayesian image reconstruction for emission tomography based on median root prior. *Europ. J. Nuc. Med.*, **24**(3), 258–265.

Altham, P. M. E., (1984): Improving the precision of estimation by fitting a model. *J. R. Statist. Soc. Ser. B*, **46**(1), 118–119.

Baddeley, A. & Lieshout, M., (1993): *Stochastic geometry models in high-level vision*, volume 1 of *Statistics and Images*, chapter 11, pages 231–256. Abingdon: Carfax.

Bandettini, P. A., Jesmanowicz, A., Wong, E. C. & Hyde, J. S., (1993): Processing strategies for the time-course data sets in functional MRI of the human brain. *Magn. Reson. Med.*, **30**, 161–173.

Bernardo, J., Berger, J., Dawid, A. & Smith, A., editors, (1992): *Bayesian Statistics 4*. Oxford: Clarendon Press.

Besag, J., (1989): Towards Bayesian image analysis. *J. Appl. Statist.*, **16**, 395–407.

Besag, J., Green, P. J., Higdon, D. & Mengersen, K., (1995): Bayesian computation and stochastic systems. *Statist. Sci.*, **10**, 3–66.

Besag, J. E., (1974): Spatial interaction and the statistical analysis of lattice systems (with discussion). *J. R. Statist. Soc. Ser. B*, **36**, 192–236.

Binder, J. R. & Rao, S. M., (1994): Human brain mapping with functional magnetic resonance imaging. In A. Kertesz, editor, *Localization and neuroimaging in neuropsychology*, chapter 7, pages 185–211. San Diego: Academic Press.

Brillinger, D. R. & Krishnaiah, P. K., editors, (1983): *Handbook of Statistics 3: Time Series in the Frequency Domain*. Amsterdam: North-Holland.

Cao, J., (1999): The size of the connected components of excursion sets of χ^2, t and F fields. *Adv. Appl. Prob.*, **31**(3), 579–595.

Carter, D. S. & Prenter, P. M., (1972): Exponential spaces and counting processes. *Z. Wahr. verw. Geb.*, **21**, 1–19.

Cox, R. W., Jesmanowicz, A. & Hyde, J. S., (1995): Real-time functional magnetic resonance imaging. *Magn. Reson. Med.*, **33**, 230–236.

Daley, D. J. & Vere-Jones, D., (1988): *An introduction to the theory of point processes*. New York: Springer.

Descombes, X., Kruggel, F. & von Cramon, D., (1998a): fMRI signal restoration using a spatio-temporal Markov random field preserving transitions. *Neuroimage*, **8**, 340–349.

Descombes, X., Kruggel, F. & von Cramon, D., (1998b): Spatio-temporal fMRI analysis using Markov random fields. *IEEE Trans. Med. Imag.*, **17**(6), 1028–1039.

Everitt, B. S. & Bullmore, E. T., (1999): Mixture model mapping of brain activation in functional magnetic resonance images. *Hum. Brain Mapp.*, **7**, 1–14.

Frackowiak, R. S. J., Friston, K. J., Frith, C. D., Dolan, R. J. & Mazziotta, J. C., (1997): *Human brain function*. Toronto: Academic Press.

Friston, K., Holmes, A., Poline, J., Price, C. & Frith, C., (1995a): Detecting activations in PET and fMRI: levels of inference and power. *Neuroimage*, **4**, 223–235.

Friston, K., Worsley, K., Frackowiak, R., Mazziotta, J. & Evans, A., (1994a): Assessing the significance of focal activations using their spatial extent. *Hum. Brain Mapp.*, **1**, 214–220.

Friston, K. J., Frith, C. D., Liddle, P. F. & Frackowiak, R. S., (1991): Comparing functional (PET) images: the assessment of significant change. *J. Cerebr. Blood Flow Metab.*, **11**, 690–699.

Friston, K. J., Holmes, A. P., Worsley, K. J., Poline, J. P., Frith, C. D. & Frackowiak, R. S. J., (1995b): Statistical parametric maps in functional imaging: a general linear approach. *Hum. Brain Mapp.*, **2**, 189–210.

Friston, K. J., Jezzard, P. & Turner, R., (1994b): Analysis of functional MRI time-series. *Hum. Brain Mapp.*, **1**, 153–171.

Geman, D. & Reynolds, G., (1992): Constrained restoration and the recovery of discontinuities. *IEEE Trans. Patt. Anal. Mach. Intell.*, **14**(3), 367–383.

Geyer, C. J., (1992): Practical Markov chain Monte Carlo. *Statist. Sci.*, **7**(4), 473–511.

Geyer, C. J. & Møller, J., (1994): Simulation procedures and likelihood inference for spatial point processes. *Scand. J. Statist.*, **21**(4), 359–373.

Gilks, W. R., Richardson, S. & Spiegelhalter, D. J., editors, (1996): *Markov chain Monte Carlo in practice*. London: Chapman & Hall.

Green, P. J., (1995): Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, **82**(4), 711–732.

Green, P. J. & Silverman, B. W., (1994): *Nonparametric regression and generalized linear models: a roughness penalty approach*. Number 58 in Monographs on statistics and applied probability. London: Chapman & Hall.

Hajnal, J. V., Myers, R., Oatridge, A., Schwieso, J. E., Young, I. R. & Bydder, G. M., (1994): Artifacts due to stimulus correlated motion in functional imaging of the brain. *Magn. Reson. Med.*, **31**, 283–291.

Hartvig, N., (1999): A stochastic geometry model for fMRI data. *Scand. J. Statist.*. To appear.

Hartvig, N., (2000): Spatial deconvolution of the BOLD signal by a hierarchical model. Unpublished manuscript.

Hartvig, N. & Jensen, J., (2000): Spatial mixture modelling of fMRI data. *Hum. Brain Mapp.*, **11**(4), 233–248.

Higdon, D., (1994): Spatial applications of Markov chain Monte Carlo for Bayesian inference. Ph.D. thesis, Department of Statistics, University of Washington.

Kao, C. M., Pan, X., Chen, C. T. & Wong, W. H., (1998): Image restoration and reconstruction with a Bayesian approach. *Med. Phys.*, **25**(5), 600–613.

Kershaw, J., Ardekani, B. A. & Kanno, I., (1999): Application of Bayesian inference to fMRI data analysis. *IEEE Trans. Med. Imag.*, **18**(12), 1138–1153.

Kershaw, J., Kashikura, K., Zhang, X., Abe, S. & Kanno, I., (2001): Bayesian technique for investigating linearity in event-related BOLD fMRI. *Magn. Reson. Med.*, **45**(6), 1081–1094.

Kipnis, C. & Varadhan, S. R. S., (1986): Central limit theorem for additive functionals of reversible Markov processes and applications to simple exclusions. *Comm. Math. Phys.*, **104**, 1–19.

Künsch, H. R., (1994): Robust priors for smoothing and image restoration. *Ann. Inst. Statist. Math.*, **46**, 1–19.

Kornak, J., Haggard, M. P. & O'Hagan, A., (1999): Parameterisation of the BOLD haemodynamic response in fMRI incorporated within a Bayesian multiplicative Markov random field model for efficient spatial inference. In K. V. Mardia, R. G. Aykroyd & I. L. Dryden, editors, *Spatial temporal modelling and its applications*, pages 27–30. Leeds University Press.

Kwong, K. K., (1995): Functional magnetic resonance imaging with echo planar imaging. *Magn. Reson. Quart.*, **11**(1), 1–20.

Lange, N., (1996): Tutorial in biostatistics: statistical approaches to human brain mapping by functional magnetic resonance imaging. *Statist. Med.*, **15**, 389–428.

Lange, N., Strother, S. C., Anderson, J. R., Nielsen, F. A., Holmes, A. P., Kolenda, T., Savoy, R. & Hansen, L. K., (1999): Plurality and resemblance in fMRI data analysis. *Neuroimage*, **10**, 282–303.

Lange, N. & Zeger, S., (1997): Non-linear Fourier time series analysis for human brain mapping by functional magnetic resonance imaging. *Appl. Statist.*, **46**(1), 1–29.

Leonard, C. M., Voeller, K. K. S., Lombardino, L. J., Morris, M. K., Hynd, G. W. & Alexander, A. W., (1993): Anomalous cerebral structure in dyslexia revealed with magnetic resonance imaging. *Archives of Neurology*, **50**, 461–469.

Malonek, D. & Grinvald, A., (1996): Interactions between electrical activity and cortical microcirculation revealed by imaging spectroscopy: implications for functional brain mapping. *Science*, **272**, 551–554.

Petersen, N. V., (1998): Non-linear state space models with applications in functional magnetic resonance imaging. Unpublished manuscript.

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

Poline, J. B. & Mazoyer, B. M., (1994a): Analysis of individual brain activation maps using hierarchical description and multi-scale detection. *IEEE Trans. Med. Imag.*, **13**, 702–710.

Poline, J. B. & Mazoyer, B. M., (1994b): Enhanced detection in brain activation maps using a multifiltering approach. *J. Cerebr. Blood Flow Metab.*, **14**, 639–641.

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

Preston, C. J., (1975): Spatial birth-and-death processes. *Bull. Int. Statist. Inst.*, **46**(2), 371–391.

Priestley, M. B., (1981): *Spectral analysis and time series*, volume 1 of *Probability and Mathematical Statistics*. New York: Academic Press.

Rajapakse, J. C., Kruggel, F., Maisog, J. M. & von Cramon, D., (1998): Modeling hemodynamic response for analysis of functional MRI time series. *Hum. Brain Mapp.*, **6**(4), 283–300.

Ripley, B., (1981): *Spatial statistics*. New York: Wiley.

Ripley, B. D., (1987): *Stochastic simulation*. Series in probability and mathematical statistics. New York: Wiley.

Rugg, M., (1995): La différence vive. *Nature*, **373**, 561–562.

Salli, E., Korvenoja, A., Visa, A., Katila, T. & Aronen, H. J., (2001): Reproducibility of fMRI: effect of the use of contextual information. *Neuroimage*, **13**(3), 459–471.

Sastry, S. & Carson, R. E., (1997): Multimodality Bayesian algorithm for image reconstruction in positron emission tomography: a tissue composition model. *IEEE Trans. Med. Imag.*, **16**(6), 750–761.

Shaywitz, B. A., Shaywitz, S. E., Pugh, K. R., Constable, R. T., Skudlarski, P., Fulbright, R. K., Bronen, R. A., Fletcher, J. M., Shankweiler, D. P., Katz, L. & Gore, J. C., (1995): Sex differences in the functional organization of the brain for language. *Nature*, **373**, 607–609.

Stark, D. D. & Bradley, W. G., editors, (1988): *Magnetic resonance imaging*. C. V. Mosby Company.

Stoyan, D., Kendall, W. & Mecke, J., (1995): *Stochastic geometry and its applications*. Chicester: Wiley, 2nd edition.

Talairach, J. & Tournoux, P., (1988): *Co-planar stereotaxic atlas of the human brain*. New York: Thieme Medical Publishers.

Taskinen, I., (1998): Aivojen funktionaalisten magneettiresonanssikuvien bayesiläinen tilastoanalyysi. Unpublished Lic. Thesis.

van Lieshout, M. N. M., (1994): Stochastic geometry models in image analysis and spatial statistics. Ph.D. thesis, Amsterdam.

van Lieshout, M. N. M., (2000): *Markov point processes and their applications*. Imperial College Press.

Vuorinen, T., (2000): Processing of speech and non-speech sounds in language-related cortex compared between dyslexic and normal-reading adults: fMRI study. Master's thesis, Department of Psychology, University of Jyväskylä.

Worsley, K. J., (1994): Local maxima and the expected Euler characteristics of excursion sets of χ^2, F and t fields. *Adv. Appl. Prob.*, **26**, 13–42.

Worsley, K. J., Evans, A. C., Marrett, S. & Neelin, P., (1992): A three-dimensional statistical analysis for rCBF activation studies in human brain. *J. Cerebr. Blood Flow Metab.*, **12**, 900–918.

Worsley, K. J., Poline, J. B., Vandal, A. C. & Friston, K. J., (1995): Tests for distributed nonfocal brain activations. *Neuroimage*, **2**, 183–194.

Yhteenveto

Funktionaalinen magneettiresonanssikuvantaminen (fMRI) on aivotoimintojen tutkimusta varten kehitetty mittausmenetelmä, joka on viime vuosina vakiinnuttanut paikkansa neuropsykologisissa koeasetelmissa. fMRI:n avulla aivokudoksen hermosolujen aktiivisuutta voidaan havaita epäsuorasti seuraamalla aivokuoren verenkierrossa ilmeneviä muutoksia. Menetelmän avulla pyritään paikallistamaan kontrolloitujen ärsykkeiden synnyttämiä aivovasteita. Resonanssiaineisto koostuu sarjasta viipalemaisia pikselikuvia, joissa tilaresoluutio on tyypillisesti noin 128×128. Yleensä aineisto on luontevinta tulkita kokoelmaksi spatiaalisia aikasarjoja, koska tavoitteena on tarkkailla resonanssisignaalin kulkua ajassa kussakin pikselissä. Aineistojen tilastollista analysointia vaikeuttavat aivojen hemodynamiikan epätäsmällinen tuntemus, havainnointia häiritsevät satunnaiset fysiologiset tekijät ja itse mittauslaitteen synnyttämä kuvauskohina.

Väitöstutkimuksessa on rajauduttu tarkastelemaan aktivaatioprosessien spatiaalisten piirteiden tilastollista inferenssiä. Keskeisenä tuloksena esitetään bayesläiseen tilastotieteeseen perustuva malli aivojen aktivaatioprofiilien analysointia varten. Aktivaatioprofiilin priorimallilla pyritään kuvailemaan aktivaatioiden klusteroitumista. Lähestymistapa on epäparametrinen siinä mielessä, että priorilla ei pyritä säätelemään klusterin muotoa. Priorilla halutaan korostaa klustereiden oletettavaa spatiaalista laajuutta ja aktivaatioiden koherenttisuutta. Mallin erityisenä ominaisuutena voidaan pitää klustereiden keskusten muodostaman pistekuvion pitämistä yhtenä parametrina. Käytetty priorimalli konstruoidaan merkkisten Gibbsin pisteprosessien avulla.

Väitöstyössä korostetaan profiilin uskottavuusfunktion valinnan merkitystä, erityisesti kohinan spatiaalisen riippuvuusrakenteen mallintamista. Kohinan malleina on käytetty gaussisia ehdollisesti autoregressiivisiä prosesseja. Korrelaatiorakenteiden lisäksi uskottavuusfunktiossa otetaan huomioon aivojen fysiologisiin prosesseihin liittyvä kohinan heteroskedastisuus.

Tilastollinen päättely nojautuu aktivaatioprofiilin posteriorijakauman simulointiin. Simulointeja varten työssä on johdettu yleisestä Gibbsin pisteprosesseille esitetystä MCMC-menetelmästä tarkoitukseen soveltuva algoritmi. Simuloinnissa on erikoispiirteenä parametriavaruuden dimension muuttuminen päivitysten yhteydessä.

Bayesläistä mallia on sovellettu lukihäiriötutkimuksen yhden koehenkilön osa-aineiston analysointiin, jossa havainnollistetaan ennakkotiedon laadun ja määrän vaikutusta tilastolliseen epävarmuuteen aktivaatioprofiilista. Työssä tarkastellaan myös synteettistä aineistoa ehdotetun priorimallin vertaamiseksi muihin prioreihin. Vertailukohteena on käytetty epäparametrisia tasoitusprioreja, jotka poikkeavat hyperparametrirakenteen suhteen ratkaisevasti klusteripriorista. Simulointituloksista ilmenee, että tavanomaisin epäparametrisin priorein ei voida kuvailla kaikkia aktivaatioprofiileille olennaisia piirteitä, tärkeimpänä aktivaatioiden voimakasta spatiaalista organisoitumista. Väitöksessä todetaan, että (spatiaalisessa mielessä) rakenteellisten mallien avulla fMRI-analyysiin voidaan tuoda joustavasti yksityiskohtaista asiantuntijatietoa tutkittavasta aivotoiminnosta. Lisäksi työssä korostetaan koeasetelman vaikutusta posterioripäättelyn sensitiivisyyteen priorin suhteen ja todetaan, että bayes-päättely voi olla erityisen hyödyllistä, jos tutkittavassa profiilissa halutaan kontrastoida useiden eri ärsykkeiden vasteita.