PARAMETRIC MODELLING OF FUNCTIONAL MAGNETIC RESONANCE IMAGING DATA

Niels Væver Hartvig

Ph.D. Thesis
Department of Theoretical Statistics
University of Aarhus

July 2000

Preface

In one of the first papers I read on functional magnetic resonance imaging (fMRI), it was stated that non-parametric tests seemed both necessary and sufficient to cope with significance testing in this area. This was slightly discouraging, given that I had just started a Ph.D. study in statistical modelling of fMRI data. Luckily time have proved that the statement was wrong: During the last four years there has been an incredible activity in this research area, and detailed models are now recognized as necessary for understanding fMRI data and for using these to gain new insight in the brain. The development of gradually more realistic and explicit models, as well as the refinement of analytical and computational methods, is still a research topic of much interest.

As is almost always the case, the practical problems have spawned theoretical statistical research as well. In the opinion of Keith Worsley, one of the most influential statisticians in the field, the impact of brain imaging data on statistics in the next century may be as large as the effect that agricultural field trials had in this. The most visible theoretical research has been within the topic of excursion sets of random fields, a theory which is used to assign significance to observed activation clusters in a brain image. Without doubt, other fields of spatial statistics will be influenced by these data also. There is an abundance of complex spatial and spatio-temporal problems in brain imaging, and there is a rich potential for applying current and developing new statistical methods to address these.

This paper, together with the enclosed manuscripts, constitute my Ph.D. thesis submitted to the Faculty of Science, University of Aarhus. The enclosed papers are written independently, and the purpose of this first paper, is to provide an introduction to the methods presented, and compare the results with those of other researchers in the field.

I wish to express my sincere gratitude to all, who have helped and encouraged me during the writing of this thesis, in particular to my supervisor Jens Ledet Jensen, who is an inexhaustible source of good ideas. Hans Stødkilde-Jørgensen, Jeppe Burchhardt and their colleagues at Aarhus University Hospital introduced me to the subject and have shared data and insight, for which I am very grateful. I wish to thank Lars Kai Hansen, Technical University of Denmark, and Antti Penttinen, University of Jyväskylä, and their colleagues and Ph.D. students for their hospitality during my stays. Finally I am indebted to my wife, Helle, for her endless patience, tolerance and support.

Århus, July 31, 2000. Niels Væver Hartvig

Contents

Accompanying papers

1 Introduction

The purpose of this overview paper is to 1) give a brief presentation of the statistical problems in functional magnetic resonance imaging (fMRI), 2) present the main results and methods of the accompanying papers and 3) give a critical exposition of results obtained by other researchers in the area. We start with an introduction to fMRI, and a motivation for the succeeding chapters.

1.1 Functional magnetic resonance imaging

1.1.1 Physical background

Magnetic resonance imaging (MRI) is a scanning technique, which was introduced to clinical medicine about twenty years ago. It is a unique and extremely flexible technique, which is regarded as one of the biggest advances in medical imaging since Röntgen's discovery of X-rays in 1895. Contrary to X-rays, MR is capable of producing anatomical images of soft tissue in the body, by exploiting the magnetic properties of hydrogen nuclei. The images have extremely good resolution, is acquired and displayed within milliseconds and the recording of the image is completely harmless to the subject, who is exposed only to a strong magnetic field.

The MR scanner creates anatomical images from the hydrogen density of the tissue. A further advantage of the scanner, however, is the possibility of measuring a range of other tissue-specific parameters; functional MRI is one example of this. fMRI uses the different magnetic properties of oxy- and deoxyhaemoglobin to visualize localized changes in blood flow, blood volume and blood oxygenation in the brain. These are in turn indicators for local changes in neural activity. By exposing a subject to controlled stimuli, which are carefully designed to affect only certain brain functions, it is possible to estimate the anatomical location of neurons involved in the corresponding functions. Brain function may then be mapped to brain anatomy by combining fMRI scans with anatomical scans obtained by conventional MRI.

The possibility of using MRI for neurofunctional studies of the brain was discovered less than ten years ago—one of the first experiments was published by Kwong *et al.* (1992). Since then the interest in fMRI has been enormous, and for good reasons: fMRI has a better spatial resolution than the older positron emission tomography (PET) technique, the temporal resolution is orders of magnitude better and fMRI is completely non-invasive. Formulated in popular terms, a PET scanner records photographs of brain function, while fMRI provides the neuroscientists with entire movies of the spatio-temporal activation processes.

Leaving the journalistic jargon aside, we will describe how an fMRI experiment is actually carried out. The scanner records an image of a slice of the brain of thickness about 5 mm. The image consists of 64×64 or 128×128 pixels (or voxels) of dimension 1.5–3 mm. The two terms “pixel” and “voxel” are abbreviations of respectively “picture element” and “volume element”, and will be used interchangeably. In most studies a collection of equi-distant slices is combined to form a pseudo-3D volume. “Pseudo”, because the slices are obtained sequentially in time, and hence the time difference between the top and bottom may be as large as several seconds, and because the slices are often not consecutive, but have a distance of several millimetres.

A volume of slices may be obtained within one or two seconds, and a sequence of such volumes are recorded while the subject is exposed to certain stimuli. This may for instance be a motor stimulus, related to the muscular movement of a hand, or a sensory stimulation presented as a flashing light or a sound. The stimulus is presented repeatedly with epochs of rest in between. Since brain activity may be assessed only relatively to the rest condition, the latter is designed as carefully as the stimulation condition, in order to illuminate a specific brain function. In a motor or sensory experiment, the subject may simply be asked to relax during rest, while in a cognitive experiment, the rest condition may be another stimulus, which is perceived differently by the mind. An example is the presentation of well-known words during stimulation and nonsense words during rest.

More comprehensive introductions to MR scanning and fMRI experiments may be found in Lange (1996) and Cohen and Bookheimer (1994), the paper by Lange and Zeger (1997) contains a good introduction also. The reader who is interested in the physics of MR scanning may look in Canet (1996) or Stark and Bradley (1992).

1.1.2 Data and pre-processing

The data obtained in an fMRI experiment is a time series of scans, together with covariates of the experiment. The latter includes the stimulation function, which indicates the periods of stimulation and baseline, the scan-parameters and sometimes also external measurements of pulse and respiratory rates.

Usually the head of the subject is fixed with a vacuum cushion or with foam paddings during scanning, but small heads movements are unavoidable. Before the statistical analysis, a motion-correction routine is hence almost always applied. Usually the scans are aligned individually to a reference scan under rotations and translations, and possibly also corrected for the delayed magnetization effects that previous movements may have on the present scan through the so-called spin history (Friston *et al.*, 1996b).

When data from a single subject is to be analysed, the detected activation may be mapped directly on an anatomical scan from the subject. However, when results must be generalized to a population or when data from different subjects are to be combined, it is often necessary to relate different brains to a standard atlas, this is known as stereotaxic normalization. Even if brains have almost the same topography, the anatomical variation is quite large and the normalization is hence a very difficult task. The standard approach is to use the coordinate system of Talairach and Tournoux (1988), which is defined by the position of characteristic anatomical landmarks. A brain is 'transformed' to the reference coordinates by aligning the landmarks of the brain and the atlas under piece-wise linear scalings of the axes. This normalization is far from perfect—major sulci (grooves between ridges on the cortical surface) may vary in position by 1-2 cm in a normal population (Talairach and Tournoux, 1988)—and the status as a standard atlas is mainly due to the lack of practical alternatives. More satisfying methods, which are also much more computer intensive, have been developed during the last years, based on for instance deformable and probabilistic atlases (Thompson *et al.*, 2000).

A good alternative to stereotaxic normalization, which is applicable in some studies, is to delineate regions of interest on an anatomical scan, and then compare relevant summary statistics of the activation in these regions between subjects. A problem with this approach is that it is difficult to obtain standard errors of the estimates to be compared, owing to the lack of a good model for the activation. This problem is addressed in a Bayesian framework in paper I.

Finally a common pre-processing step is to smooth the data spatially to increase "signal-to-noise" ratio. We prefer to view this as a spatial estimation procedure, and will hence discuss it in the context of spatial models in Chapter 3.

1.1.3 Statistical analysis

Ideally the experiment should be designed with a specific neuroscientific hypothesis in mind, and the purpose of the statistical analysis is to test the corresponding hypothesis. In general, however, it may be extremely difficult to test hypotheses which are not phrased in terms of the location of the activation, and hence the primary goal of the analysis is to estimate the position of the activation, that is regions of the brain, where the intensity correlates with the stimulation function.

Though the data may be viewed as a high-dimensional time series, it is common to separate the spatial and temporal analyses and perform the temporal analysis voxel-by-voxel. We show in Paper I that in some circumstances this corresponds to a sufficient reduction of the data, and no information is lost by separating the two steps. In general however, the separate analyses are made for simplification and for computational convenience.

Temporal analysis. The intensities of a single voxel is regarded as a one-dimensional time series, and an estimate of the "correlation" with the stimulation function is obtained. The central statistical problem at this stage is to formulate an appropriate time series model. To define the term "correlation" one must:

- Formulate a model for the haemodynamic response to neural activation. Owing to the lack of a commonly accepted biological model to explain the coupling of haemodynamic response to neural activation, this must be based on empirical studies.

- Formulate a model for the noise. The distribution of the noise may be very complex and may differ from one voxel to another. Model control is hard to perform, since thousands of time series are analysed in an automatic way.

- Choose a statistic to quantify the magnitude of activation in the time series. In the light of the remarks above, robustness to deviations from the model is a relevant issue.

The delicate balance between simplicity and sensitivity on one hand and flexibility and robustness on the other may be very difficult to obtain, and has received much attention. We review approaches to solving these issues in Chapter 2.

Spatial analysis. The voxel estimates are next viewed as a volume (or a map) which is analysed by a spatial model—commonly the aim is to classify voxels as active or non-active. Often the estimates are test-statistics which have a common distribution under the null-hypothesis of no activation anywhere. The temptation to view the spatial analysis as a hypothesis testing problem is obvious: The map is comprised of thousands of identical voxel-wise tests which may be rejected or accepted by thresholding the map. One problem here is that of multiple comparisons: If the voxel-wise test is made at level α, a fraction of α voxels will be classified as active by chance, when there is no activation anywhere, which is clearly unacceptable. The simplest solution is to correct the significance level to $\alpha/|V|$, where $|V|$ is the number of voxels. When all voxels are independent and $|V|$ is large, this so-called Bonferroni correction will yield a global significance level of α. When the voxels are correlated, however, the Bonferroni correction may be much too conservative. A better solution is to model the map as a correlated random field, and set the threshold such that the maximum of the field exceeds it with probability α, under the null-hypothesis of no activation. This is far from easy, since a theoretical expression for the distribution of the maximum of a random field is rarely available. Much research has been devoted to obtaining approximate expressions for the tails of this distribution, particularly for Gaussian random fields, but also for t-fields, χ^2-fields and F-fields. We discuss the random field theory in greater detail in Section 3.1.

The hypothesis testing approach has other problems than that of multiple comparisons. The fundamental problem is the lack of a model for the activation, i.e. there is no model for the distribution of the statistics under the alternate hypothesis, and no assumptions are made about the distribution of shape and size of activated regions. Without an explicit spatial model, concepts such as uncertainty of the estimated pattern or the testing of high-level hypotheses about the activation profile are very difficult to study. This problem is studied in several of the enclosed papers, by investigation of different spatial models. In Section 3.2 this work, and related approaches, are reviewed and compared with the hypothesis testing setup.

1.1.4 Why "parametric modelling"?

We use the term "parametric modelling" in the title to emphasize the focus of this thesis. There has been an immense activity in fMRI data analysis, and researchers with different scientific backgrounds have attacked the problems with the philosophy and tools of their own field. Our viewpoint will naturally be from statistics, and we will focus on probability based models for the data, through which significance statements can be made. Clearly the term "parametric" is not fully adequate for this class of methods, since non- or semi-parametric models may fulfill these criteria also, and we will discuss some of these approaches as well. Yet parametric models are characterized by the fact that they incorporate explicit knowledge of the phenomena under study. Often this allows for direct interpretation of parameters in terms of the physical processes generating the observations, and the possibility of formulating hypotheses of interest through simple restrictions on parameters. It is this type of structured models, we will focus on here.

Parametric models form the core of the tools for fMRI data analysis, since researchers most often wish to attach uncertainty estimates to their findings. There has, however, been much research in related areas, such as on non-parametric multivariate methods, which extracts relevant spatio-temporal features from data, and allows one to display data in enlightening formats. These may rarely be used to test hypotheses, but are very useful for hypotheses generation and for diagnosing unexpected features in the residuals of a model. We will not discuss any of these methods, but refer the reader to recent overview papers and references therein, such as Petersson *et al.* (1999a) and Lange *et al.* (1999).

1.2 Summary of enclosed papers

A brief summary of the enclosed papers are given below. Four of them (I-III,VI) are concerned with spatial and temporal models for fMRI data, while papers IV and V are theoretical papers. Paper IV describes a simulation algorithm used in Paper III, and Paper V studies asymptotical normality of the maximum likelihood estimator in state space models. We use state space models for modelling respectively noise components in Paper VI and the haemodynamic response function in Paper I. We will discuss the contents in greater detail and related approaches in the following chapters.

I **A stochastic geometry model for fMRI data.** This paper describes a high-level spatio-temporal model for the activation pattern. This is modelled spatially as a collection of Gaussian functions with unknown width, height and position, and temporally by a state space model. Inference in the model is based on MCMC and the paper describes an algorithm for simulating from the posterior distribution. *Submitted to Scandinavian Journal of Statistics.*

II **Spatial mixture modelling of fMRI data.** This paper proposes several *a priori* spatial models for the activation pattern in a small square or cube of voxels. The distribution of a test statistic is modelled under both the activation state and the non-activated state, and when using the spatial models proposed, the posterior distribution of a voxel being activated may be calculated in closed form. The models may also be used as marginal priors in image restoration problems. *Written jointly with Jens Ledet Jensen. Accepted for publication in Human Brain Mapping.*

III **Spatial deconvolution of the BOLD signal by a hierarchical model.** This paper describes a convolution model for the spatial haemodynamic effects in fMRI data. The neural activation pattern is modelled as an independent field of Gamma variates, which is next smoothed with a kernel representing the haemodynamic blurring. By MCMC methods we may estimate the neural activation pattern; effectively this corresponds to a spatial deconvolution of the data. *Unpublished manuscript.*

IV **Simulation of the Gamma-Normal distribution.** This paper studies a rejection sampling algorithm for simulating a distribution with density $f(x) \propto x^{\nu-1}\exp(-\alpha x - \beta x^2)$, $x > 0$. The algorithm is proved to be asymptotically optimal for certain limits of the parameters. The simulation algorithm is used to make inference in the convolution model of the previous paper. *Unpublished manuscript.*

V **Asymptotic normality of the Maximum Likelihood Estimator in state space models** In this paper we prove asymptotic normality of the maximum likelihood estimator in a certain class of stationary state space models, namely models where the latent process belongs to a compact space. The technique is based on a martingale central limit theorem. *Written jointly with Jens Ledet Jensen. Published in Annals of Statistics (1999),* ***27****(2), 514–535*

VI **Non-linear state space models with applications in functional magnetic resonance imaging.** This manuscript describes an approximative Kalman filter for non-linear state space models. The filter is based on sequential normality approximations, where moments are calculated by numerical integration. The method is used to estimate physiological fluctuation patterns and trends in fMRI time series of large veins. The estimated noise components are next used as confounds in a general linear model for the data. *Unpublished manuscript. Presented at 4th International Conference on Functional Mapping of the Human Brain, Montreal 1998, NeuroImage* ***7****(4), S592.*

2 Temporal modelling of fMRI data

Compared to other types of brain imaging data, fMRI data have special temporal aspects, which need explicit consideration and modelling. These may be divided into the following three categories: i) Dynamic properties of the haemodynamical effects which causes the measured signal, ii) the form of structured noise in the data, arising from physiological sources such as respiratory or cardiac cycles or from drifts in the signal level, and iii) the properties of random noise.

Clearly one of these issues cannot be studied without defining the other two. Nevertheless it is almost a tradition in the literature to focus on each of the different aspects separately, and we will try to make the same distinction below. In the first section we will discuss approaches to modelling the haemodynamic response to stimulation, next we turn to modelling of structured noise components, and finally to models for the random noise. In the last section we discuss a Bayesian approach where the three parts are fully integrated.

We will focus only on the temporal modelling in this chapter. With a few exceptions, all the models reviewed below treat the data as a collection of independent time series, one for each voxel, and do not assume any spatial structure. As mentioned in the previous chapter, the spatial modelling is traditionally introduced either implicitly in the preprocessing of the data or as a second step in the analysis. We will defer the discussion of spatial modelling to the next chapter.

Unless otherwise stated we will use the notation $\{Y_{it}\}_{i\in V,t=1,\ldots,n}$ for the spatio-temporal fMRI data, where V is the set of voxel indices, and $t = 1,\ldots,n$ is the scan number.

2.1 Models for the haemodynamic response

The functional MRI signal arises from the haemodynamic effects occurring in the vascular system concurrently with neural activation. The biological processes behind this are not fully understood, but the general structure of the signal has been described and reproduced in many studies. The haemodynamic response lags the neuronal activation with several seconds; it increases slowly to a peak value at about 6-8 seconds after a neuronal impulse, and then returns to baseline again. The shape of the response looks roughly like a Gamma density. Often a late undershoot is reported as well, in the sense that the signal drops below baseline for a period after the peak value before it returns to the baseline value.

The reason why this response can be measured is the different magnetic properties of oxygenated and deoxygenated blood. A relative increase in the amount of oxyhaemoglobin will cause a slower transversal dephasing of the nuclei spins, which is detected as a signal increase in T_2^* weighted MR sequences. Even if neural activation increases the local oxidative metabolism, causing a relative decrease in the amount of oxyhaemoglobin, this will be overcompensated by the increased blood flow and volume, which occurs after a few seconds. The result is thus a delayed increase in the level of oxyhaemoglobin, which is the contrast used in blood-oxygen-level-dependent (BOLD) fMRI.

Though there has been some attempts to model the physiological effects underlying the BOLD signal directly (Buxton *et al.*, 1998; Glover, 1999), most models for the haemodynamic response function (HRF) are empirical in nature, and assumptions have been tested directly on the observed data.

2.1.1 Periodic models

The first fMRI experiments were based on blocked periodic stimulation designs where the two epochs, stimulation and baseline, were repeated over time. A periodic model for the haemodynamic response is thus a natural choice. The simplest possible model for the HRF is a binary model which is 1 during stimulation and 0 during rest. Even if this seems simplistic, given the complex nature of the haemodynamic effects, it is the model underlying the simple two sample t-test, which is not uncommon in practical analysis. The model is sometimes refined by shifting the binary function 4-8 seconds in time, to accommodate the delay of the response.

Bandettini *et al.* (1993) and Lee *et al.* (1995) proposed to model the HRF as a sine-wave with the same period as the stimulation function. This resembles the observed response more than a binary function, and may furthermore be interpreted as a spectral analysis of the time series, where the power at the stimulation frequency is the parameter of interest. Bandettini *et al.* (1993) proposed to use the cross-correlation of the sinusoidal reference function and the time series as a test-statistic, which is equivalent to a t-statistic in a regression model with independent normal errors. The cross-correlation is still a very popular method, owing to its simplicity and good power. Bullmore *et al.* (1996) and Ardekani *et al.* (1999) took this model a step further, by modelling the response function as a linear combination of cosines and sines with the stimulation frequency and the first and second harmonics of this. This corresponds to a Fourier basis for the response function. These two papers also addressed more general noise models than Bandettini *et al.*, we will return to this issue later.

2.1.2 Linear models

Parallel to the study of periodic models, there has been much focus on convolution models for the haemodynamic response. These are less empirical since they attempt to explain the response from a biological viewpoint, and thus aim at a better understanding of the underlying processes. This in turn allows for generalizations to non-periodic experiments and other more detailed studies of the brain.

In a convolution model the haemodynamic response is given as

$$\varphi(t) = h \star \pi(t) = \sum_s h(t-s)\pi_s.$$

Here π_t is the stimulation function, which is 1 during stimulation and 0 during rest, and $h(t)$ is a model for the impulse response, also known as the transfer function. The latter may be interpreted as the haemodynamic response to a short stimulation of one time point. The assumptions underlying this model are that the response is time invariant (or stationary), in the sense that the impulse response $h(t)$ is the same for all stimulation time points, and that the impulse responses combine additively over prolonged periods of stimulation.

Motivated by observed responses, several choices for the model $h(t)$ have been proposed. Among these are the discrete Poisson density (Friston *et al.*, 1994), the Gaussian function (Friston *et al.*, 1995), the Gamma density (Lange and Zeger, 1997) and the shifted Gamma density (Boynton *et al.*, 1996). A parametric form for the impulse response allows for separate estimation of parameters in different voxels, by which differences in delay and shape of the response may be quantified. A more empirical and less arbitrary choice is to use the actually observed impulse response in other fMRI studies (Cohen, 1997). A step in the opposite direction is to use completely non-parametric models for the impulse response, as proposed by Nielsen *et al.* (1997) in a likelihood framework and by Højen-Sørensen *et al.* (2000) in a Bayesian framework.

For blocked paradigms, where the stimulus is presented during several scans, the convolved response $\varphi(t)$ will not be very sensitive to the exact shape of the transfer function $h(t)$. In recent years, however, there has been an increasing interest in so-called event-related paradigms, where the stimulus is only presented for a short period of time. Experiments may be designed with greater flexibility in this case, since many stimuli cannot be repeated over a block of scans. This may be the case in cognitive experiments, where for instance the immediate perception of a visual or auditory stimuli is of interest. More careful modelling is needed to extract the response from this type of data, which has led to refined models for the impulse response function. Glover (1999) and Worsley (2000a) model the latter as a difference between two Gamma densities, which represent respectively the initial increase and the later undershoot,

$$h(t) = \left(\frac{t}{p_1}\right)^{\alpha_1} \exp\left\{-\frac{t-p_1}{\beta_1}\right\} - c\left(\frac{t}{p_2}\right)^{\alpha_2} \exp\left\{-\frac{t-p_2}{\beta_2}\right\}.$$

Here $p_i = \alpha_i\beta_i$ is maximum point of the Gamma density, and may thus be interpreted as a delay parameter. Glover (1999) estimated parameters for an auditory response to $\alpha_1 = 6$, $\alpha_2 = 12$, $\beta_1 = \beta_2 = 0.9$ s and $c = 0.35$. A similar form was considered by Friston *et al.* (1998a), who also used a further refinement by combining $h(t)$ with its temporal derivate $\partial h/\partial t$. A linear combination of the two may, in a Taylor-like fashion, account for small voxel-wise differences in the delay of the response. A non-parametric approach, which is applicable when the stimulus events are presented periodically, was proposed by Josephs *et al.* (1997), who modelled the impulse response as a linear combination of 32 Fourier basis functions.

The approximate linearity of the haemodynamic response has been demonstrated empirically in visual stimulation studies by Boynton *et al.* (1996) and Dale and Buckner (1997). Boynton *et al.* varied both stimulus contrast and length, and illustrated that the response combines approximately linearly over time, and that the temporal profile of the haemodynamic response function does not change with contrast. The magnitude of the response was, however, a non-linear function of the stimulation contrast. This was hypothesized to result from the non-linear response of the neuronal system, which has been observed also with single-unit recordings. Their conclusion is thus that the haemodynamic system is approximately linear as a function of neuronal activation, but that the neuronal activation may depend non-linearly on the stimulation contrast.

2.1.3 Non-linear and non-stationary models

The evidence of linearity of the haemodynamic response should be contrasted to the work of Glover (1999), who found that the responses in the motor and auditory cortices were not linear; the magnitude of the response to very short stimuli (<1 s) was smaller than expected, and the response to long stimuli (about 16 s) showed less undershoot than predicted by a linear model. Vazquez and Noll (1998) found non-linear responses in the visual cortex, also for short stimuli (<4 s). Friston *et al.* (1998b) detected and quantified the non-linear properties of the response in the auditory cortex as a function of word presentation rate. They used a second order model for the HRF of the form

$$\varphi(t) = \sum_s h^1(t-s)\pi_s + \sum_{s,s'} h^2(t-s, t-s')\pi_s\pi_{s'},$$

where h^1 and h^2 are kernels modelling respectively first and second order effects. The authors assumed that the kernels were given by a linear combination of a small number of known basis functions,

$$h^1(t) = \sum_{i=1}^p g_i^1 b_i(t), \quad h^2(s,t) = \sum_{i,j=1}^p g_{ij}^2 b_i(s) b_j(t),$$

the response is then just a linear function of the unknown parameters $\{g_i^1\}$ and $\{g_{ij}^2\}$. The authors chose $p = 3$ and let $b_i(t)$ be a Gamma density with shape parameter 2^i. This is clearly a very important part of the model formulation, and some of their results may be sensitive to this choice. The model may either be interpreted as a so-called Volterra series expansion of a general non-linear system, or, perhaps more intuitively, as a non-linear function f of a linear system

$$\varphi(t) = f\left(\sum_s h^1(t-s)\pi_s\right).$$

By Taylor expanding f to the second order we get approximately the same representation as above. The authors hypothesize that the neuronal activity is linear as a function of word presentation rate, and that the non-linear effects are introduced by the haemodynamic effects. In this case f would represent non-linear effects of the haemodynamic system, for instance a non-linear relationship between flow and oxygen extraction fraction.

The assumption of stationarity of the haemodynamic response is clearly a restrictive one. In many experiments one can imagine that the response will change with general alertness or due to learning effects, and a constant model for the impulse response function $h(t)$ may not be applicable. A parametric method, which addresses non-stationarity, is that of sliding time-windows (Gaschler-Markefski *et al.*, 1997). Basically overlapping temporal blocks of the time series are analysed separately, and the results are combined to study the temporal development of the activation. It is difficult to combine results in different blocks in a rigorous way, and hence the authors use the method mainly descriptively in this sense.

Gössl *et al.* (2000) study a non-stationary response function in a state space model. A simple convolution model is used as a fixed regressor in the analysis, but the magnitude of the function is considered as a latent, time-varying function, which is estimated individually in each voxel. Since they also have an interesting noise model, we will postpone the detailed description of the model to the next section. Their approach is very appealing as it permits a study of the spatio-temporal activation pattern in a solid statistical framework. One problem, however, is that the main parameter of interest, the magnitude of activation, is not well-defined when the model response function is zero during epochs of rest.

A state space approach was also considered in Paper I, where the entire response function was modelled as a latent process. The response was assumed to be a random walk with drift given by a simple convolution model. This framework is rather flexible, and instead restrictions were imposed on the spatial pattern. Firstly we made the assumption that the response function was the same in all voxels, and secondly the magnitude of activation was described spatially by a collection of Gaussian functions with a minimal extent and height. We will describe the model in more detail in the context of spatial models in the next chapter. By MCMC techniques the posterior distribution of the response function given the data was obtained, and this indeed showed significant non-stationarities in a visual stimulation experiment.

2.2 Models for the systematic noise

Trends, drifts or fluctuations are often observed in the data and have been reported in many studies. It is in general unclear what causes these but several explanations are possible. Trends may be caused by instability in the scanner, by motion artifacts or possibly by slow variations in physiological parameters such as blood pressure. Fluctuations may be aliased physiological oscillations caused by the cardiac or respiratory cycle. The repetition time, or inter-scan time, is often around 1 or 2 seconds, which means that both cardiac and respiratory effects may be aliased.

2.2.1 Filtering of physiological fluctuations

A simple approach to reducing fluctuations caused by the aliased cardiac and respiratory rhythms was proposed by Biswal *et al.* (1996). They monitored the processes externally, and designed Gaussian band-reject filters to remove the corresponding frequencies of the time-series. Filtering introduces correlation in the series, and there is hence a trade-off between removing nuisance components and the cost of reducing degrees of freedom. Filtering may thus be worse than doing nothing. This was recognized by Buonocore and Maddock (1997) who designed Wiener filters to remove physiological fluctuations. Effectively a Wiener filter returns the residuals of the time series after a a linear least-squares estimate of the structured noise component has been removed. The method thus requires an estimate of structured as well as random noise components, or rather the spectral density of these. A group of voxels, which were dominated by either form of noise, were used to estimate these components directly from the data, hence no external measurements were required. Like the band-reject filters, Wiener filtering introduces correlation in the time series, and Buonocore and Maddock concluded that the filter was only useful if the stimulation frequency was contaminated by physiological noise. If this was not the case, the improvement in the detection of activation was not large enough to compensate for the reduction in the degrees of freedom.

The filtering methods rely on approximate periodicity of the physiological noise; at least the frequency band of this noise must be relatively small and constant over time. Due to aliasing effects this may not be a realistic assumption. A more robust and in fact very simple approach is proposed by Hu *et al.* (1995) who monitor the heart and breath rate externally. The time point of each scan is transformed to its relative position within the unit cardiac cycle, and the scans are re-shuffled in accordance to this. At each point in k-space[1], the observations are replaced by the residuals after fitting a periodic function to the unit cycle observations, and the scans are then shuffled back in their original order. Hence variations consistent with the cardiac rhythm are removed individually in each point in k-space. This acts as a preprocessing step of the raw scans, rather than a model for the noise, but seems more effective than applying filters of high-order. Only four parameters are fitted in each time-series in k-space, assuring a minimal reduction in the degrees of freedom. The main disadvantage is the requirement to monitor the physiological processes externally; besides the practical problems with this, the equipment may disturb the static magnetic field of the scanner and thus introduce artifacts in the MR signal. This was addressed by Le and Hu (1996) who obtained information on the physiological processes directly from the k-space data, however a fast imaging rate was required for this, which is not possible in all experiments.

2.2.2 Simple trend models

Even if trends may be considered as noise sources, it is often more convenient to model them in the mean value of the time series than to formulate models for

[1] The k-space is the frequency space in which the scans are acquired. The observed images are obtained by a 2-dimensional Fourier transform of the k-space image.

stochastic processes, which may exhibit the relevant features. The most simple approach is to model the trend by a linear term in the mean value space, a choice which is very common in the literature. Even if this may seem simplistic, it serves as a good approximation for many time series, and it is a more parsimonious choice than some of the more elaborated noise models. Obvious extensions are of course to include higher order polynomial terms or exponential functions of time.

Holmes *et al.* (1997) first proposed to include a basis of low frequency cosines in the mean value space, where the maximal frequency is chosen well below the stimulation frequency, say less than a half of the latter. When both sines and cosines are included in the mean value space this effectively corresponds to a high-pass filter for the data, but formulated in a statistical model. Holmes *et al.* proposed to use only cosine terms, however, presumably in order to trade off flexibility for a smaller dimension of the model.

Often a so-called global signal is included in the mean value also. This is simply the time series obtained by averaging all voxel time series. The motivation for this originally comes from PET data, where the global signal may be interpreted as global blood flow; in fMRI data, where the magnitude of the intensities have no direct physical interpretation, the global signal can only be interpreted as a global trend structure in the data. The biggest problem with the inclusion of a global signal is that it may be confounded with activation related signals, this will especially be the case if large areas are activated. This may lead to underestimation of the response in activated areas, and artificial detection of negative activation in non-active voxels. On the other hand, Zarahn *et al.* (1997) showed that inclusion of a global signal reduced the spatial correlation of voxels and hence also the variance of estimated activation patterns.

2.2.3 Semi-parametric trend models

Ardekani *et al.* (1999) took a semi-parametric approach to estimating global trends. Suppose we let Y_i denote the time series in voxel i, Y_i is then a vector with length equal to the number of scans, n. They considered a model of the form

$$Y_i = A\theta_i + B\phi_i + \varepsilon_i, \quad \varepsilon_i \sim N_n(0, \sigma^2 I_n),$$

where A is a known $n \times p$ matrix representing a basis for the haemodynamic response model (the authors used a trigonometric basis for a periodic response function), and B is an unknown $n \times q$ matrix with trend terms. All voxels were assumed to be independent and have equal variance. The column space of B, denoted the nuisance space, was estimated by maximum likelihood, subject to the constraint that the column spaces of B and A were orthogonal. The order q of the nuisance space was selected by a minimum AIC procedure, in their examples the authors selected $q = 2$. For this value of q, one would expect that the two columns of B correspond to an almost constant term and a global trend term. In this sense the model is closely related to models where a global signal term is included in the mean value space instead of the estimated trend term. However, unlike the global signal term, the trend terms here are orthogonal to the response function, avoiding confounding between the two.

In fact the authors showed that the MLE of B is given by the q first principal components of the residual data, after the estimated response-function has been removed in each voxel. (See Mardia *et al.* (1979) for a description of principal component analysis.) This gives a more intuitive understanding of the method; each estimated trend term is just a weighted average of the residual time series, where the weights constitute an eigenvector of the empirical spatial covariance matrix of the voxels. This also points to a weakness of the method, because the weight assigned to a given voxel time series will scale with the variance of the series. Contrary to the authors assumption, variance homogeneity across voxels is often not realistic, and voxels with high variance will hence tend to dominate in the estimate of B. This may, or may not, be an advantage, depending on how trends and variance are related in the data.

In Paper VI we described a method for estimating trends directly from the data also. Rather than restricting trends to be orthogonal to the response function, they were estimated only from voxels which did not correspond to neuronal tissue. In a concrete example, six voxels corresponding to sinus sagittalis, a large vein in the mid-sagittal line of the head, were selected for estimating the structured noise components, including the cardiac fluctuation effect. The hypothesis behind this was that trends corresponding to movement artifacts, signal drifts and slowly varying physiological processes would be detectable in a large vein, and could hence be estimated from the noise voxels, and used when analysing voxels corresponding to neuronal tissue.

Let $X_t = (X_{1t}, \ldots, X_{kt})$ denote the k-dimensional time series of the k noise voxels located in a large vein. These are assumed to consist of a fluctuation term with a common frequency, which is determined by the cardiac rhythm, and an individual trend term in each voxel. The model for X_t is thus

$$X_t = \mu_t + a\cos(v_t) + b\sin(v_t) + \varepsilon_t, \quad \varepsilon_t \sim N_k(0, \Sigma). \tag{2.1}$$

Here $\mu_t \in \mathbb{R}^k$ represent the trend terms, v_t is the phase at time t of the fluctuation term and a and b are k-dimensional vectors determining the magnitude and relative phase of the fluctuation in each series. By restricting $b_1 = 0$, v_t is the phase of X_{1t}. The trends and phase are assumed to be unobserved, and modelled as latent processes in a state space framework,

$$\begin{aligned} \mu_t &= \mu_{t-1} + \omega_t^\mu, & \omega_t^\mu &\sim N_k(0, \sigma_\mu^2 I_k), \\ \delta_t &= \delta + \rho(\delta_{t-1} - \delta) + \omega_t^\delta, & \omega_t^\delta &\sim N(0, \sigma_\delta^2), \\ v_t &= v_{t-1} + \delta_t. \end{aligned} \tag{2.2}$$

One may interpret the term δ_t as the aliased pulse rate at time t, which is allowed to vary around a long-term average δ.

Since v_t enters non-linearly in the observation equation the ordinary Kalman filter cannot be used. One may linearize the observation equation by a Taylor expansion, which would give the generalized Kalman filter (Fahrmeir and Tutz, 1994), but since the trigonometric functions are highly non-linear, this was found not to work well. Instead we proposed another approximative Kalman filter, based on sequential Gaussian approximations of the filtering distributions, where the moments of the latter were calculated by numerical integration. This may be used to obtain estimates of the latent processes, and to calculate approximations to the likelihood function and the residual variance; the latter was used to estimate parameters by numerical minimization. We will return to the approximate Kalman filter in Chapter 4.

In the applications, the model was found to yield an acceptable fit to the six noise series, and the estimated pulse rate corresponded well with external measurements. The estimated trend terms $\hat{\mu}_{jt}$ $j = 1, \ldots, k$, as well as sines and cosines of the estimated pulse phase $\hat{v}_t$ were included as regressors in a multiple regression model for the time series in any voxel. Denote the latter Y_{it}, where i indexes the voxel, the model was then,

$$Y_{it} = \mu_i + \sum_{j=1}^{k} \alpha_{ij} \hat{\mu}_{jt} + \sum_{j=1}^{3} (\gamma_{ij} \cos(j\hat{v}_t) + \lambda_{ij} \sin(j\hat{v}_t)) + a_i \varphi_t + \varepsilon_{it},$$

where $\{\varepsilon_{it}\}_t$ followed an AR(1) model. Here φ_t is the HRF, which was omitted in the study, since we only considered baseline data. We verified the fit of the model by studying several time series, which were contaminated by high levels of structured noise. The residuals from these series showed no significant deviations from a sequence of independent normal variables. We furthermore compared the model to models with a linear trend term and a cosine basis for the trends, but also with AR(1) errors. As measured by a minimum AIC criteria, the model above gave the best fit in respectively 92% and 88% of the voxels. The two other models did, however, not account for cardiac fluctuation, which is at least partly the reason for the improved fit.

The advantage of this method is that structural noise components are estimated directly from the data, without artificial regularity assumptions on these, such as restricting trends to be “sufficiently different” from the haemodynamic response. In the case where for instance motion artifacts are correlated with stimulus, the assumption of orthogonality between trends and the response may not hold. The cardiac pulsations are not restricted to be approximately periodic as in the filtering approaches, but is allowed to vary with the pulse rate, and no external measurements of the pulse are required. One of the weaknesses is the somewhat arbitrary choice of the noise pixels: It is not obvious how many one should choose, and how they should be chosen. This may be compared with the problem of choosing the form and number of basis functions in a general trend basis. In the application we have selected a group of three voxels in the front of the head and three in the back, in order to capture the different effects of movement in different parts of the image. This seems to work well, but it may be possible to use another number of voxels. The voxels were furthermore manually selected to ensure that they were located at the vein, it is not an easy task to design an automatic selection procedure which does this.

Gössl *et al.* (2000) consider an approach based on state space models as well. The model is of the form

$$Y_{it} = a_{it} + z_{it}b_{it} + \varepsilon_{it}, \quad \varepsilon_{it} \sim N(0, \sigma_i^2). \tag{2.3}$$

The noise components are temporally and spatially independent. Here $z_{it} = z_t(d_i, \theta_i)$ is a model for the haemodynamic response, obtained by convolving the stimulation function with a Poisson density with mean θ_i and lagged d_i time points. The terms a_{it} and b_{it} are latent processes which represent respectively a trend and the magnitude of the haemodynamic response. The model for these are

$$a_{it} = 2a_{it-1} - a_{it-2} + \zeta_{it}, \qquad \zeta_{it} \sim N(0, \sigma_{\zeta_i}^2), \tag{2.4}$$

$$b_{it} = 2b_{it-1} - b_{it-2} + \eta_{it}, \qquad \eta_{it} \sim N(0, \sigma_{\eta_i}^2). \tag{2.5}$$

The authors first obtain simple estimates for the parameters (d_i, θ_i) by minimizing the squared distance between Y_{it} and $z_t(d_i, \theta_i)$, $t = 1, \ldots, n$, and next use the EM algorithm to estimate the variances. The Kalman filter and smoother are used in the expectation step of the EM-algorithm, as well as to produce estimates $\hat{b}_{it}$ and variance estimates $\hat{\sigma}_{b_{it}}^2$ of the latter in the fitted model.

The general framework is very appealing. The model for the trend is flexible and intuitive, avoiding arbitrary choices on the form or number of basis functions, and the common assumption of temporal stationarity of the haemodynamic response is relaxed, allowing the researcher to study spatio-temporal activation patterns in the data. Ironically, the flexibility is also the main drawback of the method: The authors note themselves that model complexity directly influences inferential power, and suggest that if the temporal non-stationarity is only expected to be slight, then one should use a fixed response function instead. In fact their examples with real data suggest that the model does not improve the fit, compared to a simple multiple regression model of the form

$$Y_{it} = \mu_i + \alpha_{i1} t + \alpha_{i2} t^2 + z_{it} b_i + \varepsilon_{it}, \quad \varepsilon_{it} \sim N(0, \sigma_i^2),$$

where z_{it} is the same response function as above. In the case where $\alpha_{i2} = 0$ this is a parametric sub-model of the state space model obtained by restricting $\sigma^2_{\eta_{it}}$ and $\sigma^2_{\zeta_{it}}$ to zero. Thus in this case the state space model will always have a larger maximized likelihood value. In the general case where $\alpha_{i2} \in \mathbb{R}$, the regression model is strictly speaking not contained in the state space model, but the quadratic trend is a special case of the general trend term in the state space model. A rigorous comparison of the difference in maximum likelihood of non-parametric and parametric models is not possible, and hence it is difficult to determine whether the fit of the state space model is significantly better. The authors choose to quantify the fit by the R^2 value given by

$$R^2 = 1 - \frac{\sum_t (y_{it} - \hat{y}_{it})^2}{\sum_t (y_{it} - \bar{y}_i)^2},$$

where $\hat{y}_{it}$ is the estimated value of $E(y_{it})$. When fitting the models to a range of time series, the authors find that the R^2 value of the state space model is significantly larger than that of the regression model, when comparing the values by a paired t-test. On this basis, they claim that the state space model fits better than the regression model; this seems to be a very weak argument. Suppose, for the sake of argument, that we considered two nested parametric models. For a parametric model the R^2 measure may be written as $R^2 = 1 - Q^{2/n}$ where Q is the likelihood ratio for the hypothesis $H_0 : E(Y_{it}) = \mu_i$. Thus R^2 is an increasing function of the maximum likelihood value, and due to the nesting the biggest model will always have the largest R^2 value. The important issue is of course, whether the R^2 difference is larger than what can be explained by the added amount of parameters, which in the parametric case could be quantified by using the asymptotical χ^2-distribution of the difference in log-likelihoods.

For the state space model the situation is more complicated, since the maximized likelihood function has a more complicated expression, and the asymptotic theory does not apply. The principle is however the same: It is not surprising that a more complicated model yields a better fit in terms of the R^2 value, and this alone does not justify the model. Their examples furthermore illustrate, that the difference in R^2 values is not greater for real data, than for synthetic data generated with the stationary response of the regression model. It is hence not evident that the state space model really improves the fit compared to much simpler models, contrary to the claim made by the authors. When considering the computational burden of estimating parameters in the state space models as well, a simpler model may very well be favourable.

Gössl *et al.* consider the process of studentized values $\hat{b}_{it}/\hat{\sigma}_{b_{it}}$, and discuss the possibilities of studying the spatio-temporal development of activation through these, for instance in experiments where the activation pattern may change with attentional shifts. This is unarguably an interesting feature, but we think that one should be careful not to misinterpret the maps. The estimate $\hat{b}_{it}$ should be interpreted as the magnitude at time t of the model haemodynamic response in voxel i. It will be difficult to understand how b_{it} relates to values in neighbouring voxels, if the HRF of the voxels differ; one cannot make any direct interpretation of differences without considering the response functions also. A clear example of this is the problem of interpreting b_{it} when $z_{it} = 0$. We would prefer to investigate the processes $z_{it}\hat{b}_{it}$ which is more readily interpretable.

Purdon and Weisskoff (1998) proposed a related approach, by modelling the noise as a superposition of an AR(1) and a white noise process. One may think of this as a state space model, where the physiological noise is modelled as an AR(1) process, and the random scanner noise is superimposed as white noise. The authors did not consider a state space framework, but gave a direct and efficient way of performing the linear transformation $\Sigma^{-1/2}\varepsilon$, where $\Sigma = \mathrm{var}(\varepsilon)$, through a whitening filter. Once estimates of the variance parameters are obtained, likelihood inference in the model may hence be performed directly. They considered data with different inter-scan times (TR) and found that the AR(1) correlation decreased roughly as $\exp(-\mathrm{TR}/15\text{ sec})$. This was, however, based on the assumption that all time series in a certain region of interest had the same distribution, and the authors did not report any model control. Yet their model is partly verified by an extensive empirical study by Zarahn *et al.* (1997), who investigated a range of noise datasets. They assumed that all time series in a data set were identically distributed, and calculated an averaged periodogram for all time series. They demonstrated that a power spectrum corresponding to that of the AR(1) plus white noise process fitted well to the data. Interestingly, they also demonstrated that time series obtained from a silicone phantom contained correlated noise. Hence temporal correlation cannot solely be attributed to physiological processes, but also arises intrinsically in the scanner.

2.3 Models for the random noise

The random noise is basically what is left when haemodynamic response and structured noise components have been specified. Of course this may both be "genuine" random physiological noise and scanner noise, but also residual variance due to an imperfect model. Probably the most popular model is that of white Gaussian noise. Even if this is not always written explicitly, it is the implicit model underlying simple t-test methods or correlation methods, which are widely used. However, as we saw in the last section, a white noise model is not always a simplistic choice. The validity of the model depends on how one defines and models structured noise components.

Worsley and Friston (1995) studied a refinement of the white noise model, which is in fact built on the latter, but the inference in the model is designed to be robust to deviations from the white noise assumption. This was originally proposed by Friston *et al.* (1994) and Friston *et al.* (1995), but was brought into a solid statistical framework by Worsley and Friston (1995). Let Y_i denote the time series in voxel i of length n. The idea is to consider a linear model,

$$Y_i = A\theta_i + \varepsilon_i, \quad \varepsilon_i \sim N_n(0, \Sigma_i),$$

where A is a $n \times d$ design matrix with a model for the haemodynamic response, a constant mean value term and trend terms. Here Σ_i is an unspecified covariance matrix. Rather than estimating Σ_i and performing a usual maximum likelihood analysis, the authors take the opposite approach and smooth the data by convolving it with a kernel K,

$$KY_i = KA\theta_i + K\varepsilon_i, \quad K\varepsilon_i \sim N_n(0, K\Sigma_i K').$$

An estimator for θ_i is now given by

$$\hat{\theta}_i = (A'K'KA)^{-1}A'K'KY_i.$$

This is not as efficient as the maximum likelihood estimator, but it is an unbiased estimator which does not depend on the unknown variance Σ_i. The fundamental assumption is now that $K\Sigma_i K' \approx \sigma_i^2 KK'$. This is only exact if a white noise model is assumed, i.e. $\Sigma_i = \sigma_i^2 I_n$, but the idea is that by smoothing the data, the inference is less sensitive to deviations from the white noise model. Using this assumption, the variance of $\hat{\theta}_i$ may be directly obtained, and an unbiased estimator s_i^2 of σ_i^2 may be obtained from the residuals. Using Satterthwaite's method, an approximative distribution $s_i^2 \sim \sigma_i^2\chi^2(f)/f$ is obtained, where the so-called effective degrees of freedom f depend on the smoothing kernel K. Assuming furthermore that s_i^2 and $\hat{\theta}_i$ are independent, an approximate t-statistic is derived for the hypothesis that a specific constrast of the coordinates of θ_i are 0. We refer to Worsley and Friston (1995) for details.

The advantage of the method, is its robustness towards the non-specified covariance structure. The assumption which underlies many other approaches, of an identical covariance model in all voxels, will in general not hold, since the noise tend to depend on the properties of the underlying tissue. The framework above will in general be more robust to a wide range of actual noise distributions, which is clearly a big advantage when analysing thousands of time series in an automatic way. The framework can of course be extended by inserting an estimate of Σ_i, instead of assuming $\Sigma_i = \sigma_i^2 I_n$ (Zarahn *et al.*, 1997). This will result in an analysis that accounts partly for the specified covariance, but is robust to the variance of the estimator.

An important part of the method is the choice of the smoothing kernel K. Friston *et al.* (1995) chose this to resemble the impulse haemodynamic reponse by reference to the so-called Matched Filter Theorem. The latter is a theorem from the signal processing literature, which states that a signal embedded in white noise is optimally detected by convolving the data with a kernel shaped like the signal. It is however not obvious how one should interpret this in statistical terms. Worsley (2000b) views the method from a spectral point of view (as did also Friston *et al.* (1994) originally). In the spectral domain it may be seen as weighted least squares with weights proportional to the Fourier coefficients of the smoothing kernel. By choosing the kernel to equal the impulse response, frequencies of the time series which are dampened by convolution with the response, and hence contain little information on this, are thus given small weights.

Bullmore *et al.* (1996) consider an $\mathrm{AR}(1)$ model for the noise and a periodic haemodynamic response function modelled by the first three Fourier components of the stimulation frequency,

$$Y_{it} = \mu_i + \beta_i t + \sum_{k=1}^{3} (\gamma_{ik} \sin(k\omega t) + \delta_{ik} \cos(k\omega t)) + \varepsilon_{it},$$

where the noise process $\{\varepsilon_{it}\}_t$ is AR(1). Here $\omega = 2\pi/T$ is the frequency of the stimulation with an on/off period of length T. The authors verified the model by diagnostic plots of the auto-correlation and the normality of the residuals, however mainly using a single time series obtained by averaging 156 voxel time series. In order to detect activation they consider the fundamental power quotient,

$$\mathrm{FPQ}_i = \frac{\hat{\gamma}_{i1}^2 + \hat{\delta}_{i1}^2}{2s_i^2},$$

where s_i^2 is an estimate of the (approximate) common variance of $\hat{\gamma}_{i1}$ and $\hat{\delta}_{i1}$. This is closely related to using the periodogram at frequency ω as test-statistic, as suggested by Lee *et al.* (1995) and Bandettini *et al.* (1993), but the FPQ statistic here is based on an explicit parametric model for the signal and the noise. Under the null hypothesis $\gamma_i = \delta_i = 0$, the statistic has an approximate $\chi^2(2)/2$ distribution, which may be derived using the asymptotic normality of estimators. When studying baseline data, however, the authors found that the empirical distribution of the statistic had a different location and variance than the theoretical distribution, hence the approximation was not valid. This may either indicate that the model does not hold, contrary to the model control performed by the authors, or that there is not enough observations to use the asymptotic theory.

Bullmore *et al.* refrained from studying this issue further, but used a randomization test instead. They obtained the null-distribution of the statistic by randomly permuting the order of the observations in each observed time series, and calculating the FPQ statistic for each randomized time series. By reference to the permutation distribution, the significance of an observed FPQ value was calculated. Care should be taken, when interpreting this distribution: The null-hypothesis, no matter what test statistic is used, is that all permutations of the data, have the same distribution. In the present model this corresponds to the simultaneous hypothesis that: $\beta_i = 0$, $\gamma_{ik} = \delta_{ik} = 0$ for $k = 1, 2, 3$ and $\rho_i = 0$, where ρ_i is the AR(1) correlation. The alternative hypothesis is that just one of these conditions does not hold. Hence any specific inference about the activation must be based on non-statistical considerations; the permutation test can be used only to reject the null-hypothesis. Bullmore *et al.*, however, calculate critical values for the FPQ statistic from the permutations, and use these to make inference about the activation; this does not seem to be a valid interpretation of the permutation distribution.

Locascio *et al.* (1997) study a more comprehensive noise model by fitting ARMA models individually to each voxel time series. For a given voxel, the model for the time series Y_{it} is given by,

$$Y_{it} = \mu_i + \sum_{j=1}^{k} a_{ij} C_{jt} + \beta_{i1} t + \beta_{i2} t^2 + \varepsilon_{it},$$

where the noise process $\{\varepsilon_{it}\}_t$ follows an ARMA(p_i, q_i) model. The terms C_{jt} are indicators (or contrasts) of the states of possibly k different stimulation types, in the simlest case with only one task C_{1t} is just the stimulation function. The authors estimate the order individually in each voxel, by starting with an ARMA(3,3) model and then successively remove non-significant correlation terms. The parameters are estimated by conditional least squares, conditional on the assumed values of Y_{it} prior to the initial time point, and t-statistics and P-values for hypotheses of the form $a_{ij} = 0$ is calculated. Clearly this model is more flexible than an AR(1) model for all voxels, and a better fit to the time series is inevitable. Furthermore the authors verify the fit of the model in an automatic way, by requiring that the residuals pass a Box-Ljung test for white noise in each time series. An automatic and routinely applicable model control like this is a rare finding in the fMRI literature. The authors however note, that about 5% of the voxels does not pass this test and is discarded, it does not seem obvious then how one should model these. The authors use a simplistic binary model for the HRF, but their method is of course not restricted to this choice.

When making global inference in the image of P-values, Locascio *et al.* propose to solve the multiple comparison problem by a permutation test in the following sense: A random permutation of the time-points is applied simultaneously to the white-noise residuals of all voxels. Using the same contrast function C_{jt} as above, a t-test is calculated for the hypothesis that $a_{ij} = 0$ in the permuted residuals, a hypothesis which is true by construction. This produces an image of P-values, and by performing multiple random permutations a distribution of these images is obtained. The fundamental assumption is now: Under the null-hypothesis of no activation in any voxel, the multivariate distribution of the P-values is the same for the original data and for all permutations of the white-noise residuals. The authors make this assumption and use the permutation distribution to make inference in the original P-value image. The multiple comparison problem is handled by considering the distribution of the minimal P-value (see the paper for details). It does, however, not seem obvious that the assumption above is fulfilled. Firstly, the spatial correlation is different in the residual time series compared to the original ones, because different models are applied in different voxels. The authors stress that their method, unlike the Bonferroni correction, considers spatial correlation, but it is the correlation of the residuals, not the original data, which is accounted for. Secondly the original P-values and the permutation values are calculated under different models and estimation procedures. Even if the P-values have the same theoretical uniform distribution under the null-hypothesis, one could fear that their sensitivity to deviations from the true model may be different in different models. This effect may be especially problematic when the multivariate distribution of a whole image of P-values is considered. In fact we *know* that the white-noise model is wrong, since the residuals will allways be slightly correlated, hence there is strictly speaking no theoretical foundation for assuming that the P-values have identical distributions. Hence rather than being assumption-free, as most permutation methods, their method relies on assumptions which are known not to hold in general. The sensitivity of the results to departures from the assumptions is of course a different and more difficult story.

Lange and Zeger (1997) considered an even more general model, by assuming only that the noise process $\{\varepsilon_t\}_t$ is stationary and Gaussian. (For clarity we omit the voxel index i for a moment.) Then it has a spectral representation (Cox and

Miller, 1965),

$$\varepsilon_t = \int_{-\pi}^{\pi} e^{i\omega t}\, S(d\omega),$$

where $\{S(\omega)\}$ is a complex Gaussian process with orthogonal increments. The variance of S is given by the spectral density $g(\omega)$,

$$\mathrm{var}(S(\omega_2) - S(\omega_1)) = \int_{\omega_1}^{\omega_2} g(\omega)\, d\omega, \quad \omega_1 < \omega_2.$$

Suppose we observe $\varepsilon_1, \ldots, \varepsilon_n$ where n is odd. By approximating the above integral with a Riemann-Stieltjes sum with n terms, we may invert the relation and obtain increments of the S process by a discrete Fourier transform (DFT) of ε_t,

$$R_k = d_\varepsilon(\omega_k) = \frac{1}{n}\sum_{t=0}^{n-1} e^{-i\omega_k t}\varepsilon_t$$

where $\omega_k = 2\pi k/n$, $k \in \{-[n/2], \ldots, [n/2]\}$ and $R_k = S(\omega_k + \pi/n) - S(\omega_k - \pi/n)$. Here we adopt the notation of Lange and Zeger (1997) where d_f is the DFT of a function f. When $\{\varepsilon_t\}$ is real there is complex symmetry such that $R_k = \bar{R}_{-k}$, whence there is a bijection between $\varepsilon_1, \ldots, \varepsilon_n$ and $R_0, \ldots, R_{[n/2]}$. The latter will be an (approximately) independent sequence of Gaussian variables, with different variances determined by the spectral density.

After removal of a linear trend, the authors consider a convolution model of the form

$$Y_{it} = \beta_i[\lambda(\cdot;\theta_i) \star \pi](t) + \varepsilon_{it},$$

where $\{\varepsilon_{it}\}_t$ is a stationary Gaussian process. Here $\lambda(t;\theta_i)$ is the Gamma density with parameters $\theta_i \in \mathbb{R}_+^2$, and π_t is the stimulation function. When performing a DFT we approximately have

$$d_{Y_i}(\omega_k) = \beta_i d_\lambda(\omega_k;\theta_i) d_\pi(\omega_k) + d_{\varepsilon_i}(\omega_k), \quad k = 0, \ldots, [n/2].$$

The authors restrict themselves to periodic paradigms with, say, m on/off periods. Then $d_\pi(\omega_k) = 0$ unless k is a multiple of m, and hence the observations of the form $d_{Y_i}(\omega_{km})$ for $k \in \mathbb{N}$ are sufficient for (β_i, θ_i). In the example of the paper, the on and off periods have equal length, whence $d_\pi(\omega_k)$ is only non-zero when k is an odd multiple of m.

Since the Fourier transforms d_λ and d_π are known, the problem is reduced to a non-linear regression model, where the noise terms are complex Gaussian variables with unequal variances. In order to estimate (β_i, θ_i), the authors assume that the noise variances have a homogeneous spatial and temporal structure. In their application, the variance of $d_{\varepsilon_i}(\omega_k)$ is assumed to be homogeneous over a region of 11×11 voxels, and over the 5 closest frequencies, $\omega_{k'}$, $|k - k'| \leq 2$.

Once the parameters (β_i, θ_i) are estimated, the spatial correlation is modelled by fitting isotropic exponential or Gaussian correlation functions to the normalized residuals at each frequency. This in fact corresponds to a non-separable spatio-temporal covariance function, where noise components of different temporal frequencies have different spatial correlation. The variance estimates are used to calculate the covariance of β-estimates in different voxels, which in turn is used to calculate a χ^2-test statistic for specific hypotheses of the form $\beta_R = 0$, where R is a small region of voxels.

The approach is very elegant in the context of periodic stimulation paradigms: The transform to the Fourier domain reduces the convolution term to a simple product and the correlated noise process to independent observations. A non-parametric model for the noise really makes sense, at least as a benchmark model, given the complex and non-homogeneous nature of the time series. However, as mentioned above, the authors assume a specific homogeneity structure for the variance, which is *ad hoc* and difficult to interpret in terms of a specific model for the noise terms. Thus the general, non-parametric framework assumed when formulating the model is somewhat restricted in the estimation procedure, and the assumption of spatial homogeneity of the variance may in fact be violated in practice.

Another problem with the approach was raised by the discussants of the paper: Almost all information on the activation will be contained in the Fourier coefficient corresponding to the fundamental frequency, $d_{Y_i}(\omega_m)$. In essence the authors fit three mean-value parameters using this single complex variable, and concerns regarding identifiability are natural. The authors report troubles with convergence of their estimation algorithm, which may also be caused by this over-parametrization.

The model may be applied to non-periodic stimulation paradigms, but the extension is not straightforward, since the frequencies ω_k that does not contain any signal are used presently to estimate the noise spectra. In the case of general stimulation function it might be necessary to assume a specific covariance model, in which case there is no particular reason to work in the Fourier domain.

Marchini and Ripley (2000) consider the same approach as Lange and Zeger, but restrict themselves to simple sinusoidal response functions. They hence avoid the problems with overfitting described above. They propose an alternative way of estimating the spectral density, namely by fitting a smoothing spline to the the log-periodogram. Furthermore they only smooth the periodogram spatially over a 3×3 grid. They illustrate that this works well on real data and baseline data, but do not make any formal comparisons with the related approach of Lange and Zeger.

2.4 A Bayesian approach

In this last section we will review a Bayesian approach where the models for the three different components of the time series are integrated to an extent, that does not allow us to categorize them as above. Bayesian approaches are rare in fMRI analysis, presumably because of the typically larger computational burden of these methods. Frank *et al.* (1998) studied different Bayesian models, which were, however, mainly adaptations of well known models from the likelihood framework to the Bayesian paradigm.

Genovese (2000) took a fundamentally new approach by considering a model given by

$$Y_{it} = \mu_i(1 + a_t(\gamma_i, \theta_i)) + d_{it} + \varepsilon_{it}, \quad \varepsilon_{it} \sim N(0, \sigma_i^2),$$

where a_t is the activation profile, d_{it} is a trend term, and μ_i is the baseline mean. The parameter $\gamma_i = (\gamma_{i1}, \ldots, \gamma_{iC})$ is the magnitude of the activation in each of C different stimulation conditions and θ_i is an 8-dimensional parameter describing the shape of the haemodynamic response $b(t; \theta_i)$ to a single stimulation epoch. The latter is constructed from cubic splines, by a decomposition into a delay, a smoothly increasing and decreasing part and a post-stimulus undershoot. Responses from closely spaced epochs are combined additively or possibly sub-additively; in the additive case the model for the activation profile is $a_t(\gamma, \theta) = \sum_k \gamma_{c_k} b(t - t_k; \theta)$, where t_k is the starting time and c_k is the type of the k'th stimulation period. The drift term d_{it} is modelled by cubic splines.

Genovese takes a Bayesian approach and formulates prior distributions for all parameters. The γ_{ic}'s are a priori either 0 or positive, in the latter case γ_{ic} is exponential distributed with mean 2%. The drift term, interpreted as a continuous function $d_i(t)$, has a Sobolev prior

$$p(d_i) \propto \exp\left\{-\frac{1}{2\lambda}\left(\rho_{nc}\int_0^1 d_i(t)^2\,dt + \int_0^1 |d_i''|^2(t)\,dt\right)\right\}$$

where ρ_{nc} determines the relative penalization of magnitude and curvature. The author chooses $\rho_{nc} = 0.01$. The number and position of knots for the cubic spline basis are fixed when posterior maximization is performed, but is allowed to vary when sampling from the posterior by MCMC. The coordinates of θ_i have independent Gamma priors, which are selected by experience from earlier experiments. All voxels are assumed to be independent both in the prior and in the likelihood.

The inference in the model is either based on posterior maximization or on sampling by MCMC. The biggest problem in this context seems to be the fact that the model is composed of submodels of different dimension, obtained by letting different subsets of the γ_{ic}'s be 0, or by varying the number of knots used for the spline. This is either handled by reversible jump (Green, 1995) or by combining different submodels by weighting them with their respective posterior probabilities (i.e. their Bayes factors), obtained by Laplace approximations.

To illustrate the advantage of a Bayesian framework, Genovese addresses more complicated hypothesis than usually asked. An example is the monotonicity hypothesis $\gamma_{i,Tr} < \gamma_{i,T_1} \leq \gamma_{i,T_2} \leq \gamma_{i,T_3}$, where T_1, T_2 and T_3 are tasks of increasing difficulty, and Tr is a control task. The posterior probability of this event is calculated in each voxel and displayed as a map. We may note, that inference of this type may also be performed in a parametric framework by mapping a test statistic for the hypothesis in all voxels, as was also pointed out by Worsley (2000a). It is not clear how to correct for the multiple comparison problem in the Bayesian framework: The larger the search region, the greater is the chance that an estimate of the posterior probability will be large somewhere, and hence the threshold should somehow be dependent on the search region. In a Bayesian setting also, this problem is addressed in Paper I and II by using global information from all voxels to calculate posterior probabilities. There is no explicit or implicit spatial structure in Genovese's model to be used in this context; all voxels are independent.

The distribution of the size of activated clusters is also studied by simulation and compared under different conditions. This is a very complicated issue in most parametric models, but is very easy in the Bayesian framework, although time-consuming simulations are needed. One would, however, expect that the distribution of a spatial parameter like size of a cluster, is critically dependent on the (lack of) spatial structure in the model, and the results should be interpreted with this in mind.

The incorporation of prior information is very relevant in fMRI data analysis, and Genovese illustrates why, by choosing genuine priors which are based on previous studies as well as general experience with the data. fMRI data are different from many other types of data in the sense that they may be acquired in enormous quantities extremely quickly and the important features of the data are common to almost all experiments. Hence there *is* prior information available on both temporal and spatial aspects of the data, which should of course be used whenever possible. Genovese includes prior information on the temporal structure, but spatial priors, which can be included in the models discussed in Section 3.2, may turn out to be even more fruitful.

As a final point, we may note that there are some similarities between the model studied in Paper III and Genovese's model. Both uses Gamma-priors for the activation magnitude, Genovese however restricts himself to an exponential prior with mean 2%. The latter has a rather small standard deviation, and is probably chosen for computational convenience since it is conjugate with the normal distribution of the error terms, in the sense that the posterior of γ_{ic} will be a truncated normal. It seems relevant to study general Gamma priors, that express larger uncertainty about the activation level. The posterior distribution of γ_{ic} would then be of the Gamma-Normal form, which may be simulated by procedures described in Paper IV.

Notice that unlike many other models, Genovese scales the activation as a fraction of the baseline level. It is common to interpret activation levels like this, both because there is some evidence that variability tend to scale with intensity level and because the absolute intensity values of the MR scanner does not have any physical interpretation. An alternative is to assume an additive model but for the log-transformed data, which is the approach in for instance Paper I.

3 Spatial modelling of fMRI data

In this chapter we will focus on the spatial analysis which ensues after an image (or volume) of voxel-wise activation estimates have been obtained by a temporal model, as described in the previous chapter. In the brain mapping literature, this is known as a *statistical parametric map* (SPM). Unless otherwise stated, we will use the notation $X = \{X_i\}$ for the SPM, to distinguish it from the original spatio-temporal data $\{Y_{it}\}$. We will, however, follow the notation of the enclosed papers, which is unfortunately not entirely consistent with this rule.

Often X_i is scaled to have a standard normal or a t-distribution under the null-hypothesis of no activation. As discussed in the introduction it is therefore natural to view the analysis of X from a hypothesis testing point of view. The first section describes the random field theory, which is a very successful approach in this context. In the second section we turn to explicit spatial models for the distribution of the spatial activation pattern, where the main focus is on the models presented in Papers I, II and III.

3.1 Gaussian random field theory

This section contains a brief introduction to the random field theory and its application in brain imaging. Rather than to give a complete account of the theory, the intention is to discuss advantages and disadvantages, in order to provide a comparison with the explicit spatial models to be presented later. There are several overview papers which give a more complete introduction to the theory, see for instance Adler (1998) for a review, Worsley (1996) and Petersson *et al.* (1999b) for non-technical overviews, Worsley *et al.* (1996) and Cao and Worsley (1999) for summaries of the different formulas involved and Friston *et al.* (1996a) for a discussion of the interpretation of the inference.

3.1.1 Theoretical background

Perhaps the best way to understand the philosophy of the random field approach from a statistical modelling point of view, is to consider the setup of Siegmund and Worsley (1995) where a random field $\{Z(s), s \in C \subseteq \mathbb{R}^D\}$ is considered,

$$Z(s) = \xi \sigma_0^{-D/2} f(\sigma_0^{-1}(s - s_0)) + W(s). \tag{3.1}$$

Here W is a random field of white Gaussian noise, and $f(\cdot)$ is a known signal with unknown width $\sigma_0 \in [\sigma_1, \sigma_2]$, position $s_0 \in C$ and magnitude $\xi \geq 0$. For simplicity we will adopt the common assumption that $f(\cdot)$ is an isotropic Gaussian function,

$$f(s) = \pi^{-D/4} \exp(-\|s\|^2/2),$$

which is normalized such that $\int f(s)^2 \, ds = 1$. Let $X(s, \sigma)$ be $Z(s)$ convolved with the signal f,

$$X(s, \sigma) = \sigma^{-D/2} \int f(\sigma^{-1}(h - s)) \, dZ(h). \tag{3.2}$$

This convolution will partly be performed prior to the analysis, by the point spread function of the scanner, so the observed data may be considered as $\{X(s, \sigma_1)\}_s$, for $\sigma_1 > 0$. We may obtain $\{X(s, \sigma)\}_s$ for $\sigma > \sigma_1$ by additional smoothing. We wish to test the hypothesis that no signal is present in $\{X(s, \sigma_1)\}_s$, i.e. $H_0 : \xi = 0$. The authors show that the maximum likelihood estimate of (s_0, σ_0) is given by

$$(\hat{s}_0, \hat{\sigma}_0) = \underset{s \in C, \sigma \in [\sigma_1, \sigma_2]}{\operatorname{Arg\,max}} X(s, \sigma)$$

and that $\hat{\xi} = X(\hat{s}_0, \hat{\sigma}_0)$. Furthermore $-2 \log Q = X(\hat{s}_0, \hat{\sigma}_0)^2$, where Q is the likelihood ratio for the hypothesis $\xi = 0$. This means that the maximum value of X over the physical space C and the scale space $[\sigma_1, \sigma_2]$ should be used to test for the presence of the signal.

Most often σ_0 is assumed to be fixed and known, by prior knowledge of typical signal widths. For simplicity, we will make this assumption in the following. In this case the likelihood ratio test for the presence of the signal is given by $X(\hat{s}_0, \sigma_0)$, where $\hat{s}_0 = \operatorname{Arg\,max}_{s \in C} X(s, \sigma_0)$. The optimal test for $\xi = 0$ is thus obtained by smoothing the image with the signal itself, and then obtaining the maximum value over C. This illustrates what is known as the Matched Filter Theorem from the signal processing literature, which says that to optimally detect a signal embedded in an image of white noise, one should convolve the image with the signal itself.

Denote $X(s,\sigma_0)$ by $X(s)$. In order to test whether a signal $f(\cdot)$ is present or not, one must then obtain the distribution of $\max_s X(s)$ under the null-hypothesis that $X(s)$ has mean zero. Exact results on the distribution of the maximum of a random field of this type is not available, but an approximation was given by Adler (1981), which has later been refined by Worsley (1995a,b) and Siegmund and Worsley (1995), and extended to non-Gaussian random fields (Worsley, 1994, 1998). The approximation is based on the exact expression for the mean value of the Euler charactistic of an excursion set, based on the following intuition: For a threshold z, define the excursion set by $A_z = \{s \in C \mid X(s) \geq z\}$, and let $\chi(A_z)$ be the Euler characteristic of A_z. Basically the Euler characteristic counts the number of connected regions in the set A_z minus the number of holes and plus the number of hollows in these. For high thresholds, the holes and hollows tend to disappear, and there will be just one connected region if the maximum of the field is higher than z and none if the maximum is below z. Hence the Euler characteristic will approximate the indicator for the maximum of the field being above z,

$$\chi(A_z) \approx 1(\max_s X(s) \geq z),$$

which leads to the approximation $P(\max_s X(s) \geq z) \approx E(\chi(A_z))$. Under regularity conditions on X, a general expression for this mean value is given by

$$E(\chi(A_z)) = \sum_{i=0}^{D} \mu_i(C)\rho_i(z),$$

where $\mu_i(C)$ is proportional to the i-dimensional Minkowski functional of C, which is related to the geometry of C, and $\rho_i(z)$ is the so-called Euler characteristic intensity, which is given by the distribution of X. For a Gaussian random field in three dimensions, the dominating third term in the sum is given by

$$\mu_3(C)\rho_3(z) = |C|\lambda^{3/2}(2\pi)^{-2}(z^2-1)e^{-z^2/2},$$

where λ is a measure of the smoothness of the field, namely the variance of the partial derivatives. This expression was used by Worsley *et al.* (1992) as an approximation to $P(\max_s X(s) \geq z)$.

3.1.2 Estimation of the signal

Given the approximate expression for the tail-distribution of the likelihood ratio statistic, we may test the hypothesis $\xi = 0$. This is rejected if the maximum of the field exceeds a given threshold z_α, determined by the level α of the test. In this case, the signal has been *detected*, and the inferential problem is changed to characterizing or *estimating* the shape of the signal. It is in this step, that the largest weakness of the random field theory is found.

If the model (3.1) was really true, the maximum likelihood estimate of the signal would be given by the estimates $(\hat{\xi}, \hat{s}_0, \hat{\sigma}_0)$ and the known function $f(\cdot)$, and we would be done. In practice, however, the field may have more than one signal and they are not necessarily shaped like Gaussian functions. One does thus not really believe the model (3.1), but use it merely as a step in a heuristic argument. Instead the signal is estimated without reference to a specific model, for instance by the entire excursion set A_{z_α}. The rationale behind this estimate is that any voxel in the excursion set has a significant degree of activation, as measured by the threshold z_α, and is hence regarded as activated. Basically the image is thresholded to detect "significant tops".

The problem is of course, that the estimated signal does not even remotely reflect the original model (3.1). Firstly the theoretical signal rises from zero and takes all values from zero to $\xi\sigma_0^{-D/2}f(0)$, while the estimated signal will only have values larger than z_α. The traditional discussion about the significance level and the philosophy behind the Neyman-Pearson approach to statistical inference is extremely relevant in this situation: The experimenter may slide the significance level α up and down by pulling a bar on the computer screen, by which the excursion sets decreases and increases smoothly, and eventually the activation peaks may be followed all the way to zero. A problematic question is why the magic number $\alpha = 5\%$, say, is chosen, when clearly the activation pattern has more facets than what is revealed by the significant excursion set at this level.

Another problem is that the smoothing applied to the image σ_0 is optimal in the sense of maximizing the test statistic for activation, but not in terms of producing an excursion set, which reflects the true signal. On the contrary if the true signal has width σ_0, the detected signal in the field $X(s, \sigma_0)$ will have width $\sqrt{2}\sigma_0$, and it will hence be wider than the true signal. In practice this means that the estimated activation patterns will be too smooth, and activation may be present in unrealistic areas. In other words, the optimal smoothing is related to the detection the signal, not to the estimation.

The issues raised above are some of the "obvious pitfalls of answering what is really an estimation question by a hypothesis test" (Worsley, 1997). Procedures which are optimal in hypothesis testing need not be optimal in an estimation procedure, and inferential choices which are accepted when testing a hypothesis, may be subjective and *ad hoc* in an estimation framework.

3.1.3 Alternative test-statistics

As mentioned above, the signal may have characteristics which the excursion set A_{z_α} does not convey. To reveal these, alternative test-statistics have been proposed, and their approximate distribution obtained for a random field of type (3.2) under the null-hypothesis of no signal. One statistic is the maximal size of a connected region, or cluster, in the excursion set A_z (Friston *et al.*, 1994; Xiong *et al.*, 1995). The threshold z is chosen smaller than the significant threshold z_α above, and an entire cluster is declared signficant if its size is larger than what is expected by chance. This statistic is not derived as a likelihood ratio statistic in a specific model for the signal, but it is more powerful than the maximum height, for distributed non-focal signals (Friston *et al.*, 1994; Xiong *et al.*, 1995). An alternative is to combine the two tests, by declaring a peak significant using both the size and the height (Poline *et al.*, 1997). Friston *et al.* (1996a) takes the inference a step further and performs a set level inference, where a complete set of clusters is declared significant based on their number and size. Individual clusters may hence be insignificant, but the entire set of clusters is significant.

The estimates obtained in this way suffer from the same limitations as the ones discussed above: Characteristic features of the estimates are determined by the significance level of a hypothesis test, rather than from a model for the activation pattern. For the cluster size test, the activation level must always be higher than the threshold used to define the clusters, which is commonly chosen in an ad hoc way, and the estimated clusters will always be larger than the critical size.

3.1.4 Assumptions underlying the random field theory

The distributional results rely on the assumption that the observed discrete random field $\{X_i\}$ is a sufficiently good approximation to the continuous, L^2-differentiable field $X(s, \sigma)$ in (3.2). The critical assumptions are that i) the field must be stationary and Gaussian, ii) the discretization must be sufficiently fine compared to the width of the auto-correlation. Furthermore the distributional results are asymptotical, which are good approximations only for high limits of the threshold z, and the expressions depend on the smoothness of the field, which must be estimated.

The Gaussian distribution is crucial, since the extreme tail is used to determine the threshold. There is hence a considerable difference between a Gaussian distribution and a t-distribution with 20 degrees of freedom, for instance. The assumption that the field is smooth compared to the discretization is also quite restrictive, and is problematic in fMRI data, where the voxels are not very correlated. The typical solution to this problem is to smooth the data spatially with a

Gaussian kernel[1]. This corresponds to transforming the raw data $X(s, \sigma_1)$ to the smoothed field $X(s, \sigma_0)$. As mentioned earlier this transformation, and thus the choice of the smoothness parameter σ_0, is based on knowledge of typical signal widths, but often the concern that data must be smooth enough to comply with the theoretical framework influences the choice of smoothing also. Regarding the latter aspect, it is of course preferable to adapt the theory to the data, rather than vice versa.

We focus on the Gaussian random field framework here, but should note that a range of alternative methods have been proposed for assessing the distribution of the test-statistics, when the assumptions are not fulfilled. These are based on permutation tests (Holmes *et al.*, 1996; Bullmore *et al.*, 1999) or Monte Carlo simulations (Roland *et al.*, 1993; Poline and Mazoyer, 1993, 1994; Forman *et al.*, 1995; Ledberg *et al.*, 1998).

3.1.5 Smoothing as non-parametric estimation

As mentioned above, the spatial smoothing of the data is partly made in order to obtain the optimal test statistic for a signal and partly to ensure that the discrete data is a good approximation to a continuous random field. A third way of viewing smoothing is as a non-parametric estimation procedure, since the smoothed image may be interpreted as a kernel estimate of the true activation surface (assuming for the moment that the kernel is scaled to integrate to 1). An overview of kernel estimation for longitudinal data is given by Müller (1988).

Non-parametric regression methods are valuable data-exploratory tools, and they are superior to parametric methods if there is a lack of knowledge on the function of interest to propose a realistic parametric model. Even if this is partly the case for fMRI data, due to the complexity of the activation signal, we do have some knowledge of the physiological and neuronal processes, which may be included qualitatively in a parametric model. Also there is often substantial prior knowledge on the location of the activation, based on high-resolution anatomical scans, which should also be used in a model.

As is well-known the bandwidth of the kernel governs the trade-off between bias and variance of the estimate; the wider the kernel, the more biased and less variable will the activation surface be. The bias is proportional to the second derivative of the true surface, and will hence be high near peaks, which are the most interesting features of the fMRI signal. Choosing the correct bandwidth is therefore a difficult and important issue, which has received much attention in the kernel estimation literature. In fMRI, however, the degree of smoothing is often

[1] An alternative is to reduce voxel size by interpolation to a finer grid. The dimensionality of the data does, however, increase drastically with this approach.

selected by a subjective criteria, and it seems that none of the traditional methods, such as cross-validation, have been applied.

3.1.6 Why is the random field approach so popular?

As discussed above, there are problems with the random field theory when used for estimation of the activation pattern. Yet the random field approach is extremely popular in the brain imaging community, and it has developed into the golden standard by which to report ones result. The natural question, given the problems mentioned above, is why this is so.

The field of human brain mapping is still young and it is evolving rapidly. The first fMRI experiments were reported less than 10 years ago and the PET experiments predate these with only a few years. There has been a great deal of pioneerers enthusiasm in the field, due to the potential of non-invasively obtaining movies of the working human brain, and the analysis techniques have evolved very rapidly. The enormous amount of information in brain imaging data may be pre-processed, analysed and displayed in a multitude of ways, which may make it impossible to obtain a transparent view of the information presented. In order to have a solid scientific foothold in this evolving research environment, it is very important to force researchers to quantify the significance of their results in an objective way, and the random field theory provides a framework for doing exactly this.

There is a strong tradition in the medical world to report significance of obtained results in terms of p-values. These are familiar quantities, and researchers know they are on safe ground, when a p-value can be attached to their findings. This is what the random field theory provides. Even if a p-value provides only limited information, the concept of variance of an estimator of the activation volume is much more subtle, also for statisticians. It is a bit tricky to interpret and visualize the variance of a vector of dimension 65536, say.

The random field approach was originally used for PET data and it gained its popularity here. PET data are quite different from fMRI data, for example because the spatial resolution is much lower in PET, since the images are smooth at acquisition. Furthermore it is often necessary to combine data from different subjects in PET studies to obtain sufficient statistical power, which requires additional smoothing, in order to transform data to a common brain atlas. Hence both the signal and the noise tend to be much smoother in PET, than in MR data. This both means that the assumption of smoothness of the field is often naturally satisfied for PET data, and that detailed estimation of the spatial activation pattern is not relevant. On the other hand MR data has an excellent spatial resolution at acquisition, which is often sacrificed in order to use the random field theory.

Finally an important practical issue is the computational burden of an analysis technique and the degree to which it can be automated. Many, though not all, of the spatial models to be presented later, requires time consuming algorithms and a lot of user interaction, which makes them unsuitable for routine analyses. The simple formulas of the random field theory are hard to compete with in this respect.

3.2 Parametric spatial models

We will now turn to a review of different spatial models, that have been proposed for the activation pattern in fMRI data. They will be divided into three main categories, based on the three different types of spatial modelling, that the enclosed papers address.

3.2.1 Local models

By the term "local model" we here mean a model, which is specified through conditional or marginal distributions for small regions of the scan. The inferential aim in this approach is more modest than that underlying the high-level models described in the next section: One does not attempt to model the simultaneous distribution of the entire activation surface, but merely wishes to include the assumptions of coherency and smoothness in the estimate of the activation pattern.

In Paper II we propose a marginal mixture model for the spatial pattern of activation. The distribution of the pattern is specified on a small region of voxels, say a 3 $\times$ 3 $\times$ 3 block around the voxel of interest, and the neighbours are then used when calculating the posterior distribution of the centre voxel being activated. Unlike related local smoothing or filtering techniques, this approach is based on a parametric model for the data, which provides a better theoretical understanding of the method, and allows us to quantify results in probabilistic terms.

We will assume that the activation is described by an unobserved binary image $\{A_i\}_{i \in V}$, where $A_i = 1$ if voxel i is activated and $A_i = 0$ if not. The marginal distribution of the X_i's is a mixture of two components,

$$f(x) = pf(x \mid A = 1) + (1 - p)f(x \mid A = 0), \tag{3.3}$$

where A is the indicator variable for activation in the particular voxel and p is the global probability that a voxel is active. In a typical analysis of an SPM, only the null-distribution $f(x \mid A = 0)$ is explicitly given; this is the distribution of the test-statistic under the null-hypothesis of no activation in the particular voxel. Here we also require that the alternative distribution $f(x \mid A = 1)$ can be specified. The latter may be a non-central version of the null distribution or it may have an entirely different form, this will depend on the test-statistic used and on the distribution of activated voxels.

Let $C_i = \{i^0, i^1, \ldots, i^k\}$ denote the neighbourhood around voxel i, such that $i^0 = i$ is the centre, and $i^1, \ldots, i^k$ are the k neighbours. For typographical reasons we will use the notation X_i^j instead of X_{i^j}. Let $X_{C_i} = (X_i, X_i^1, \ldots, X_i^k)$ and let A_{C_i} be defined similarly. Suppose we have specified a prior distribution $P(A_{C_i} = a)$ for any configuration $a \in \{0,1\}^{k+1}$. The corresponding posterior is then given by

$$P(A_{C_i} = a \,|\, X_{C_i}) \propto P(X_{C_i} \,|\, A_{C_i} = a) P(A_{C_i} = a).$$

This may be marginalized to the posterior distribution of A_i by summing over the neighbour values. If we furthermore assume that the X_i's are conditionally independent given A_{C_i}, we arrive at the expression

$$\begin{aligned} &P(A_i = a^0 \,|\, X_{C_i}) \\ &\qquad \propto f(X_i \,|\, a^0) \sum_{a^1 \in \{0,1\}} \cdots \sum_{a^k \in \{0,1\}} \left(\prod_{j=1}^{k} f(X_i^j \,|\, a^j) \right) P(A_{C_i} = a). \qquad (3.4) \end{aligned}$$

It is very appealing to use this posterior distribution for classification of a voxel: It combines the observed values of the statistic, not only at voxel i, but also at the neighbouring voxels, and by the prior distribution for A_{C_i} we may include our knowledge of the activation pattern.

The important part of the model is the specification of the prior distribution for A_{C_i}. We propose three different models in the paper, ranging from an almost simplistic one to a more realistic one. They all reflect the idea, that activated areas tend to constitute a cluster of voxels, rather than a single isolated voxel. For a given configuration $a \in \{0,1\}^{k+1}$ we will let $s = \sum_{j=0}^{k} a^j$. The simplest model is then given by

$$P(A_{C_i} = a) = \begin{cases} q_0 & \text{if } s = 0, \\ q_1 & \text{if } s > 0. \end{cases}$$

Since all configurations with at least one activated voxel have the same probability, this is a kind of uninformative prior, which neither favours configurations with very large activation regions, nor isolated activated voxels. The model may be parametrized by the probability p of a voxel being active, given by $p = q_1 2^k$. An extension of the model is

$$P(A_{C_i} = a) = \begin{cases} q_0 & \text{if } s = 0, \\ \alpha\gamma^{s-1} & \text{if } s > 0. \end{cases} \qquad (3.5)$$

Here γ is a correlation parameter and the restriction $\gamma = 1$ corresponds to the previous model. Finally we also consider a model where both the number of active and non-active voxels enter in the prior, we refer to the paper for details.

At a first sight, the practical problem with this approach, namely the calculation of the sum (3.4), seems to be a severe limitation. If the neighbourhood is a $3 \times 3 \times 3$ cube, the sum has $2^{26} \approx 67$ million terms which is far too many for direct numerical calculation. Luckily, for many relevant models of $P(A_{C_i} = a)$, including the ones above, the sum may be given in closed form. The key point is the simple identity

$$\sum_{a^1 \in \{0,1\}} \cdots \sum_{a^k \in \{0,1\}} \prod_{j=1}^{k} f(X_i^j \mid a^j) = \prod_{j=1}^{k} \left(f(X_i^j \mid 0) + f(X_i^j \mid 1) \right),$$

which reduces the large sum to a simple product of k terms. Using various forms of this, we derive closed form expressions for the posterior probability $P(A_i = 1 \mid X_{C_i})$, for instance for the simple model we get

$$P(A_i = 1 \mid X_{C_i}) = \left\{ 1 + \frac{1}{v_i^0} \left[1 + \left(\frac{q_0}{q_1} - 1 \right) \left(\prod_{j=1}^{k} (1 + v_i^j) \right)^{-1} \right] \right\}^{-1}, \quad (3.6)$$

where

$$v_i^j = \frac{f(X_i^j \mid 1)}{f(X_i^j \mid 0)} \quad j = 0, 1, \ldots, k. \quad (3.7)$$

Notice that v_i^j is the likelihood ratio for the voxel i^j being active vs. not active. The formula (3.6) thus effectively combine the likelihood ratios from voxel i together with those of its neighbours to calculate the posterior probability of activation.

While the posterior distribution is calculated using only local information, global information is used for estimation. Denote the model parameters by (ϕ, ψ) where ϕ parametrizes the conditional distribution of X_{C_i} given A_{C_i}, and ψ parametrizes the marginal distribution of A_{C_i}. One possibility for estimation is to maximize the contrast function

$$\gamma(\phi, \psi) = \sum_{i \in V} \log f(X_{C_i}; \phi, \psi).$$

The density of X_{C_i} may be calculated analytically by the same technique as used when calculating the posterior distribution above. An alternative to the full contrast function, is to use only the marginal density (3.3) of X_i to estimate (ϕ, p),

$$\gamma_m(\phi, p) = \sum_{i \in V} \log f(X_i; \phi, p).$$

The simplest model for A_{C_i} has only one parameter, p, which may be estimated in this way. The correlation parameter γ of the extended model (3.5) may be estimated by the method of moments, using the empirical spatial covariance of the field. Both contrast functions are based on densities, hence the estimation equations will be unbiased, and the estimators will be consistent and asymptotically normal under mild regularity conditions on the spatial correlation of the process.

Everitt and Bullmore (1999) considered the same basic mixture model as above, but did not use any spatial information. Their model is in fact equivalent to assuming that all voxels are spatially independent, and their approach is thus a special case of our setup, obtained by the restriction $\alpha = \gamma/(1+\gamma)^k$ in (3.5). Not surprisingly spatial information is very important for obtaining a good estimate of a spatial pattern, and we found that our model reduced classification error with more than 40% in a synthetic fMRI dataset. When applying the models to true data, there was a striking difference between the estimated activation pattern in our model and in the non-spatial model, see Figure 4 of Paper II. In general the activated areas were larger, and single activated voxels were suppressed.

Kernel smoothing estimates. The marginal mixture model may also be compared with the usual filtering approach, where the data are spatially smoothed with a Gaussian kernel before calculating the summary image. We compared the estimates obtained with our method to those obtained by smoothing the data with a kernel of full-width-at-half-maximum (FWHM) 2 and 3 voxels respectively. The latter are commonly used kernel widths. On synthetic fMRI data, we found that our method was more powerful than FWHM 3 smoothing, at a given level of significance, while the FWHM 2 and our method had similar power. The obtained estimates were, however, qualitatively different, with our estimates being less smooth, this was especially observed with true data. Clearly we can only speculate what the "correct" activation pattern looks like for real data, but it is well known that smoothing produces a biased estimate, which may partly explain the difference.

Salli *et al.* (1999) proposed a so-called contextual clustering method, which is effectively a spatial mixture model, with an Ising prior. They consider the following setup,

$$X_i \mid A_i = 0 \sim N(0,1), \quad X_i \mid A_i = 1 \sim N(\mu,1),$$

where $\mu > 0$. The X_i's are independent given the indicator field $\{A_i\}$, and the latter has an Ising prior distribution,

$$p(A_i \mid A_{-i}) \propto \exp\left(\beta \sum_{j\sim i} 1(A_j = A_i)\right).$$

A neighbourhood consisting of the 26 closest voxels in a cube of size $3 \times 3 \times 3$ is used. The posterior probability of a voxel being activated is then,

$$P(A_i = 1 \mid A_{-i}, X)^{-1} = 1 + \exp\left\{\mu\left(-\frac{2\beta}{\mu}(U_i - N/2) + \mu/2 - X_i\right)\right\},$$

where $U_i = \#\{j \sim i \mid A_j = 1\}$. For fixed values of β and μ, the ICM algorithm is used to estimate the activation field. This corresponds to iteratively setting $A_i = 1$ if

$$X_i + \frac{2\beta}{\mu}(U_i - N/2) > \mu/2,$$

and $A_i = 0$ if not, sweeping over all voxels until convergence.

A problem in this approach is that the prior is symmetric wrt. active and non-active voxels. This is rarely realistic, in most cases there will be much fewer active voxels than non-active. To compensate for this, the parameters β and μ are adjusted, guided by simulation studies, to control the number of voxels which are wrongly classified as active. From a modelling point of view, this is somewhat artificial; the β parameter should reflect the smoothness of the underlying activation pattern, and μ should reflect the magnitude of activated voxels, neither of these have direct interpretation in terms of the probability of a voxel being active. The authors are aware of this, and argue that they use the model only as a hypothesis testing device, rather than as a model for the true pattern. It would, however, be easier to interpret results if the model was considered in terms of the distribution of the data. This could be obtained by adding a first-order term to the prior, governing the overall balance between active and non-active voxels,

$$p(A_i \mid A_{-i}) \propto \exp\left(p1(A_i = 1) + \beta \sum_{j \sim i} 1(A_j = A_i)\right).$$

The parameters may be estimated directly from the data, to reflect the properties of the true unknown activation. Unfortunately in a conditionally specified model, the parameters cannot be directly interpreted in terms of, for instance, the number of active voxels, as is possible for the marginal models in Paper II.

Descombes *et al.* (1998a) consider a Markov Random Field (MRF) model as well, however for the spatio-temporal data. The authors model the data $Y = \{Y_{it}\}$ after having subtracted the output of a temporal moving average filter in each voxel. Let $X = \{X_{it}\}$ denote the true underlying process. Because mean and trends have been removed from the data, X is considered to represent only the spatio-temporal activation process. In the case of isotropic voxels, the prior for X has the form

$$p(X) \propto \exp\left\{-\sum_{t=1}^{n}\sum_{i \in V}\left(\sum_{j \sim i}\beta\Phi(x_{it} - x_{jt}) + 2\beta\Phi(x_{it} - x_{it+1})\right)\right\},$$

where $\beta > 0$ and the interaction function is

$$\Phi(u) = -\frac{1}{1 + u^2/\delta^2}.$$

The spatial neighbours are given by the four closest voxels (the authors only model 2D scans). The so-called Φ-interaction function favours homogeneous regions in space and time, but does not smooth edges too much, since the function has an asymptote at 0 as $|u| \to \infty$. The edges which should be preserved in this situation, are the transitions between baseline and activation. A robust noise model is assumed, by letting the likelihood function be of the same form as the prior,

$$P(Y|X) \propto \exp\left\{ -\sum_{t=1}^{n} \sum_{i \in V} \Phi(x_{it} - y_{it}) \right\}.$$

By a simulated annealing algorithm the authors obtain a MAP estimate of X. They illustrate, using synthetic and real data, that the edges of activated area are much better preserved in this estimate than when smoothing the data with a Gauss kernel.

In principle inference can be made by simulating observations from the posterior distribution with an MCMC algorithm. The authors refrain from this, presumably because it is a very difficult task, due to the high dimensionality of X. Instead they propose to use the model only for restoring the activation pattern, and then proceed with an ordinary voxel-by-voxel analysis, treating the MAP estimate as the observed data. They acknowledge, however, that this is problematic, because the distribution of the MAP estimate is very different from that of the original data, and hence assumptions underlying usual analysis procedures do not hold.

As an alternative, a kind of meta-analysis is proposed in Descombes *et al.* (1998b). From the restored data empirical estimates of the HRF are obtained in each voxel, and distinctive features such as the maximum and delay are extracted. An MRF model is next constructed where these summary statistics are treated as observations, and a binary activation map is considered as a hidden process in a Bayesian setting. An MRF prior is formulated for this classification map, and interactions between the latter and the parameter maps are constructed. The model is, however, a somewhat *ad hoc* construction, which is not based on physical properties of the data, and the authors do not give any significance statements in the setup either.

Kornak *et al.* (1999) proposed another model based on MRF's, formulated for an SPM $\{X_i\}$. The model for the latter is, $X_i = A_i M_i + \varepsilon_i$, where $\{M_i\}$ is a conditional autoregressive Gaussian random field modelling the strength of activation, and $\{A_i\}$ is a binary field indicating whether a voxel is active or not. The correlation structure in the A field is modelled by letting $A_i = 1(W_i > 0)$ where $\{W_i\}$ is an intrinsic Gaussian random field. The framework is only sketched in a conference proceedings, and there seems to be no publicly available papers on the method yet.

Related approaches in statistical image analysis. The marginal mixture model of Paper II may be considered from a statistical image analysis viewpoint also. Pioneered by Geman and Geman (1984) and Besag (1986), Markov random field priors have been widely used in this area. Their popularity is due to the fact that knowledge of the local spatial structure in the image may be included in the model in an intuitive and mathematically elegant way, in terms of the pairwise interaction functions in a Gibbs model. It is well known, however, that the joint distribution of an MRF may possess unexpected effects, such as long range correlation. In statistical image analysis, these global properties are a nuisance, and local estimation methods, such as ICM, are designed to minimize the influence that the global structure of the model has on the estimate. The formulation of a model through marginal distributions is hence an alternative, where we ensure that only the specified local properties are used in the reconstruction.

We compared estimates of an image obtained by our method, and respectively the MAP and ICM estimates using an Ising prior. We considered classical datasets from the literature with both Gaussian and binary noise, and found that our method, using the model (3.5) on a 5×5 grid in the plane, performed comparably well with both other methods. An important difference, however, is that we give closed form expressions for the estimates, where the MRF approaches require iterative procedures.

Meloche and Zamar (1994) considered a framework very similar to ours. They also restored binary images using marginal models, and furthermore considered a very elegant non-parametric approach, where the distribution of the underlying image was completely unspecified. Using the method of moments they obtained unbiased estimators for the parameters. Their adaptive setup allowed them to restore quite different images within a single model, but instead they had to restrict themselves to small neighbourhoods of only four neighbours. The direct calculation of the sum (3.4) was thus not a problem for them, and they did not obtain closed form expressions for it.

An early reference is Hjort and Mohn (1984), who also studied a very similar approach. Their framework is reviewed in Hjort and Omre (1994). They considered a transition model, which, in our notation, has the form $P(A_{C_i} = a) = p(a^0) \prod_{j=1}^{k} p(a^j \mid a^0)$. This form allows analytical computation of the sum as in our setup. They also studied the extension to correlated noise, where the simple form of the posterior probability is maintained.

3.2.2 High-level models

It is often the case, that more high-level questions are asked than what can be answered by a simple spatial activation estimate. While a local model is attractive, because it is simple and computationally fast, global models hold the potential for more detailed inference about the spatio-temporal activation pattern.

Pioneered by the seminal paper of Grenander and Miller (1994), there has been an increasing interest in the field of statistical image analysis in formulating prior models for spatial patterns, which are based on high-level structures rather than low-level smoothness properties. Where image analysis mainly meant “restoration” in the eighties, the focus of the nineties have been to interpret the structure of an image in closed form. Examples in this line of thought are the works by Baddeley and van Lieshout (1993) and Rue and Hurn (1999).

In this section, we will discuss models for fMRI data, which are based on this high-level line of thought. The aim of the analysis with this type of models are often more ambitious than what can be achieved by local smoothness models. Unlike a local model, a high-level model may be parametrized by the number of activation foci and the extent of these, and inference on these parameters is therefore possible. Hence more specific hypothesis may be addressed, than merely “where is the activation?”. An example could be the hypothesis that the area of an activated region increases under one task compared to another. The price paid for this advantage is of course that stronger assumptions are made in the model. A critical point is thus to study how well data support these, or how robust the conclusions are to violation of the assumptions. In contrast to statistical image analysis, this is often a very difficult task in fMRI, since we never actually know what the true activation pattern is.

In Paper I we formulate a global model in a Bayesian framework for the spatio-temporal data $\{Y_{it}\}$. The fundamental assumption is that space and time are separable, in the sense that the temporal activation profile is the same in any voxel, only the magnitude changes from voxel to voxel. For simplicity, we only consider a two-dimensional model for a single slice.

Let $X = \{X_1, \ldots, X_n\}$ be a marked point process, which parametrizes the activation pattern. A point $X_j = (\mu_j, a_j, d_j, r_j, \theta_j)$ may to some extent be considered as a centre of activation, where the location is given by μ_j, and the four marks $(a_j, d_j, r_j, \theta_j)$ describe respectively the magnitude, area, eccentricity and angle of the centre. The magnitude of activation $\{A_i\}_{i\in V}$ is assumed to have a specific geometry, namely a sum of Gaussian functions,

$$A_i(X) = h(i; X_1) + \cdots + h(i; X_n)$$

where

$$h(i; X_j) = a_j \exp\left\{-\frac{\pi \log 2}{d_j}\left(\frac{\tilde{i}_1^2}{r_j/(1-r_j)} + \frac{\tilde{i}_2^2}{(1-r_j)/r_j}\right)\right\}$$

and $\tilde{i} = (\tilde{i}_1, \tilde{i}_2) = R(-\theta_j)(i - \mu_j)$ and $R(\theta)$ is a rotation with angle θ. This representation is motivated by the common assumptions of smoothness and spatial extent of the activation, and the idea is that a general smooth activation surface with few localized peaks, may be well approximated by a collection of Gaussian functions.

We have specific prior knowledge on the different parameters in the spatial model. The magnitude of the activation is typically about 2%-5%, we expect the activated areas to cover at least a few voxels, and we may even have a strong prior idea of where the activation will occur, based on previous experiments on the same subject or on general knowledge of the brain function under study. We will assume that the prior distribution of X has density with respect to the unit rate Poisson process of the form

$$p(x) \propto \prod_{i=1}^{n} \beta(\mu_i) \left(\prod_{i<j} \phi(x_i, x_j)\right) \prod_{j=1}^{n} \{p(a_j)p(d_j)p(r_j)\}.$$

Here $\beta(\cdot)$ is an intensity function, which may be constant if we have no knowledge of where the activation is likely to occur, $\phi(\cdot, \cdot)$ is a pairwise interaction function, which discourages centres to fall on top of each other, and $p(\cdot)$ is a generic notation for a prior distribution for the three mark parameters a_j, d_j and r_j. We have chosen truncated inverted Gamma distributions for a_j and d_j, which penalizes small values severely, but is fairly uninformative otherwise.

Suppose that a model φ_t for the HRF is assumed. We remove a linear trend in each time series prior to analysis, and thus assume that non-activated voxels have mean 0. Given X, the detrended data Y_{it} are modelled as,

$$Y_{it} = (A_i(X) + \eta_i)\varphi_t + \varepsilon_{it} \tag{3.8}$$

where $\varepsilon = \{\varepsilon_{it}\} \sim N_{|V| \times n}(0, \sigma^2\Gamma \otimes \Lambda)$, and $\eta = \{\eta_i\} \sim N_{|V|}(0, \tau^2 I_{|V|})$. The η-terms constitute a spatial process, which compensate for small differences in the true activation surface and the idealized description $A(X)$. Technically the purpose of the random surface is to regularize the spatial estimate, but intuitively we may think of the variance τ^2 as a measure of how well we expect the actual true surface to be represented by the simple structure of the model.

Let $\tilde{Y}_i$ denote the estimated regression coefficient wrt. φ in voxel i, $\tilde{Y}_i = Y_i'\Lambda^{-1}\varphi/\varphi'\Lambda^{-1}\varphi$, where $Y_i = (Y_{i1}, \ldots, Y_{in})'$ and $\varphi = (\varphi_1, \ldots, \varphi_n)'$. The regression image $\tilde{Y} = \{\tilde{Y}_i\}_{i \in V}$ then corresponds to the usual SPM in the present model. In the paper, we show that $\tilde{Y}$ is sufficient for X, and inference on the latter may hence be carried out in the spatial domain only. This illustrates that the classical approach of reducing the time series of scans to a single image, which is next analysed by a spatial model, is in fact a sufficient reduction of the data in this case. Also it provides an intuitive understanding of the spatio-temporal model; in the case where φ is known, it is in fact just a quite simple spatial model.

In the more general case, where φ is not known, we may assume a parametric model for this. In the paper we have illustrated the flexibility of the method by assuming a general state space model for φ, which allows for temporal variation in the HRF. The model is

$$\varphi_t = \lambda_t + \nu_t, \tag{3.9}$$

where λ_t is a fixed convolution model for the HRF, and $\{\nu_t\}$ is a random walk.

The inference in the model is centred on the posterior distribution of (X, φ). We may compute estimates of the posterior mean of different functions of interest, and assess their uncertainty by the posterior variance. Also the support that data gives to a specific hypothesis of interest may be quantified by posterior probabilities. In order calculate these integrals with respect to the posterior distribution, we have designed an MCMC algorithm for simulating the point process X given (Y, φ), based on the Geyer and Møller (1994) algorithm for general point processes. The posterior distribution of φ given (Y, X, η) is a simple normal distribution, which may be simulated directly by the Kalman smoother recursion.

Kiebel *et al.* (2000) proposed to decompose a spatial activation pattern in terms of Gaussian functions as well, however in a quite different framework. Their approach was based on a technique for extracting the cortical surface from anatomical MR scans, acquired simultaneously with the functional scans. The grey matter surface is described by a list of vertices and triangular faces, which is projected onto a plane to form a flattened map. The projection is designed to minimize distortions due to the curvature of the cortical surface. A set of spatial basis functions b_F^j, $j = 1, \ldots, N_p$ is defined on the flattened surface, where b_F^j is an isotropic Gaussian function with standard deviation w. The centres of the functions form a regular hexagonal grid in the plane, with distance d between points. The authors chose $w = 1$ mm and $d = 2$ mm, which gives in the order of 1500 basis functions in each hemisphere of the brain. A spatial activation pattern of the form $\sum_{j=1}^{N_p} \beta_j b_F^j$ may now be projected back onto the folded cortical surface, and then transformed to voxel space by integrating the activation contribution in each voxel. These operations affect the basis functions only, and there is hence a linear relationship between the volume of scans $Y_{\star t}$ and the magnitudes $\beta_{\star t} = (\beta_{1t}, \ldots, \beta_{N_p t})'$ at time t,

$$Y_{\star t} = A\beta_{\star t} + \varepsilon_{\star t}. \tag{3.10}$$

Here $\varepsilon_{\star t}$ represents scanner noise. Clearly this is a very appealing framework, in which to work: It is much more natural and realistic to formulate simple spatial models in the cortical surface space, than in the arbitrary space determined by the voxels. It is quite clear that in order to extract more and better physical knowledge of the brain from neuroimages, one should let the anatomy of the brain dictate the analysis, not the scanner equipment.

Kiebel *et al.* continue from the representation (3.10) by obtaining a ridge regression estimate of $\beta_{\star t}$ and next use a combination of singular value decomposition and canonical variate analysis to estimate the activation pattern in the β-space. The procedure is not backed by an explicit statistical model, and hence the authors do not make any significance statements. There is, however, a wide range of alternative and relevant methods for making model based inference in the β-space: The complicated 3D data are reduced to a regular two-dimensional hexagonal lattice, and any relevant lattice process from spatial statistics may be applied for this part of the analysis. The β-parameters could be considered as a spatio-temporal stochastic process in a Bayesian setting, and the assumptions underlying the stochastic geometry model above, of a certain spatial extent of the activations, would naturally be translated into a positive correlation between neighbouring sites in the lattice process. Prior information, regarding location of the activation, is easily incorporated in this framework as well, and the uncertainty of the β-process may be readily quantified by the posterior variance. A further advantage is that there would be a clearer distinction between scanner noise, represented by ε, and the physiological noise, which would be embedded in the model for the β-process. In short, we are of the opinion that spatio-temporal models formulated in the space of the cortical surface have many advantages compared to voxel based models, and there is no doubt that this is a promising and important research topic for the coming years.

The marked point process model could be formulated on the flattened map also. In practice this would not be a good solution, because the transformation of the activation pattern from the flattened map to voxel space would have to be recalculated for any update of the point process in the MCMC algorithm, which would be quite expensive in computer time. Instead it really makes sense from a practical point of view, to discretize the space into the amplitudes of the basis functions, and work in the lattice space instead.

Taskinen (2000) studies an extension of the marked point process model. In order to make the model more realistic, he models each activation site by a centre point, or a "mother" point, which has an associated Gaussian function with an unknown width and height. The mother point has a sequence of daughter points as well, these are modelled by a Poisson process, with an intensity that decreases with the distance from the mother point. Each daughter point contributes to the activation surface with a small Gaussian kernel with a fixed width. Clearly one can represent irregular activation areas more satisfactorily in this setup, and it is furthermore possible to interpret the mother points individually, unlike in the approach in Paper I. The setup is however quite complicated: The activation is described by a marked point process, where one of the marks is itself a point process. The task of designing an MCMC algorithm which moves efficiently in this vast state space is by no means straightforward, and hence one may be concerned about the simulation of the posterior distribution.

3.2.3 A deconvolution model

With Kiebel *et al.* (2000) as an exception, the models described above all focus on estimating the smooth haemodynamic effects of neural activation. It seems natural, however, to study approaches to solve the inverse problem, and estimate the underlying neural activation instead, which is the main parameter of interest. By formulating a model in terms of local smooth basis functions, Kiebel *et al.* effectively perform a kind of deconvolution of the data, and estimate activation on a neuronal level. In this approach, the shape and size of the basis functions are chosen in an *ad hoc* manner. Descombes *et al.* (1998b) propose to extent their Markov random field model with a convolution term, to model the haemodynamic effects directly, but they as well argue that the choice of the convolution kernel is a difficult one.

The deconvolution approach is also studied in Paper III, where a model is formulated for the an SPM $X = \{X\}_{i \in V}$. Let $\Gamma = \{\Gamma_i\}_{i \in V}$ denote the activation pattern on a neuronal level. This is modelled as a random field, defined on the discrete voxel space V. The variates Γ_i are a priori assumed to be independent and follow a mixture of a positive and a negative Gamma distribution and a distribution concentrated at 0. The haemodynamic response Λ is modelled by convolving this Γ-field by a kernel,

$$\Lambda_i = \sum_{j \in V} k_{ij} \Gamma_j,$$

where k_{ij} is a kernel on $V \times V$, which is normalized such that $\sum_{j \in V} k_{ij} = 1$. Finally the observed field is modelled by adding Gaussian noise to the haemodynamic response,

$$X = \Lambda + \varepsilon, \quad \varepsilon \sim N_{|V|}(0, \Sigma),$$

where Σ is a $|V| \times |V|$ covariance matrix.

The width and shape of the kernel is estimated by the method of moments. For a stationary model, we have that

$$\mathrm{cov}(X_i, X_l) = \sigma_{l-i} + \mathrm{var}(\Gamma_i) \sum_{j \in V} k_{ij} k_{lj},$$

hence the covariance of the X field will consist of a part determined by the random noise, and a part determined by the haemodynamic diffusion kernel. The noise covariance may be estimated very well from the original time series of scans, used to produce the summary image X, and hence we may use the empirical covariance of X to estimate the kernel k. For visual stimulation data, we found that an exponential kernel

$$k_l \propto \exp\left(-\|l\|/w\right), \quad l \in \mathbb{Z}^2,$$

with $w = 1.1$ mm fitted well to the empirical covariance at small lags. The model thus predicts that the vascular effects spread over a circle of diameter about 5 mm, which corresponds well to the figures of 3-6 mm reported in the literature (Malonek and Grinvald, 1996).

We study the posterior distribution of the Γ-field by MCMC simulations. An algorithm is proposed, which employs an auxiliary variable to decorrelate the Γ-field in the posterior. In order to perform Gibbs updates, we furthermore use a rejection sampling algorithm for simulating observations from the “Gamma-Normal” distribution, of the form $f(x) \propto x^{\alpha-1} \exp\{-\beta x - \gamma x^2\}$, $x > 0$. The algorithm uses different combinations of envelopes in each of four different regions of the parameter space. We describe the algorithm in detail in Paper IV and show that it is asymptotically optimal with certain limits of the parameters.

We estimate Γ by the posterior mean image, and illustrate how a more detailed and “sharper” image of the activation is obtained. The difference is particularly large, when compared to the non-parametric kernel density estimates, which takes the direct opposite approach and smooth the data.

3.3 Estimation or hypothesis testing?

The philosophy of parametric spatial models in Section 3.2 is very different from the hypothesis testing approach discussed in Section 3.1. The outcome of the analysis is an estimate of the activation surface, preferably with standard errors. There is no multiple comparisons problem as such, because one does not test multiple hypotheses. In contrast to the random field approach, we may summarize the advantages of parametric models as follows:

1. The data are not smoothed spatially, hence the detailed resolution of the MR scan is maintained. Instead the smoothing is performed implicitly by the model.

2. Prior knowledge, such as the anticipated location of the activation, may be included directly in the model.

3. The distributions of active as well as non-active voxels are modelled; in general this will make the analysis robust to assumptions of the noise model. In contrast, the random field approach defines activation by the extreme tail of the noise distribution, and it will hence be very sensitive to deviations from the model.

4. Parametric models allow the study of detailed hypotheses. One example is the question of whether the size of an activated region increases under a demanding task compared to a simpler task. Questions of this type may be closer related to the neuroscientific hypotheses of interest than the question of where the activation is.

5. By imposing structure on the spatial activation pattern, assumptions on the temporal response may be relaxed. As described in Chapter 2, there is much debate about the temporal properties of the HRF, and it therefore seems relevant to study this with as few assumptions as possible.

We have already touched upon some of the disadvantages of parametric models, which the random field approach does not have:

1. Most parametric methods are much more computer intensive than the random fields approach, and the analysis may be harder to automate.

2. The spatial activation pattern may be very complex, and most parametric models are too simplistic to fully describe the true scene. Model diagnostics and critical evaluation of assumptions are hence very important.

3. The medical world has a strong tradition of assigning p-values to observations made from data. Most likely many researchers prefer this measure of significance to the more difficult concept of uncertainty of an estimate of the activation pattern.

4. The thresholding in the random field setup is very intuitive. The result of a parametric model may be less transparent and more difficult to visualize in a simple way.

Above we view the two methods as competitors, and contrast their advantages and disadvantages, but this need not be the case in practice. A pragmatic solution would be to combine the best of the two worlds, and use a thresholding approach to detect significant activation and next analyse data by a parametric model in order to obtain confidence intervals for the size of the active regions, for instance. This would eliminate some doubts that the parametric estimate of the activation is really "significant".

The choice of method may also depend on the aim of the analysis. If the main interest is to detect an activation site, it may suffice to consider the location of a peak in the SPM, as this may be interpreted of the maximum likelihood estimate of the centre of the signal. In many functional studies however, the entire activation estimate is interpreted and studied in detail. This is especially the case when fMRI is used for pre-surgical planning. In this case the variance of the estimate is much more important than the protection against false positives—most patients would supposedly be more worried about false negatives.

Another example is the combination of results from fMRI studies with data from other brain imaging modalities such as magneto- or electroencephalography (MEG/EEG) (Liu *et al.*, 1998). The latter techniques record brain activation with a temporal resolution of milliseconds, but with a very poor spatial resolution. When the detailed spatial resolution of fMRI and the excellent temporal resolution of EEG/MEG can be successfully combined, a very powerful imaging technique is available. Also in this case an estimate of the activation from the fMRI data, with associated standard errors, is needed.

4 State space models

State space models is a very flexible class of time series models, which are becoming increasingly popular these years. This is both due to the wide range of applications and to the intuitive framework, which allows one to interpret variables and the structure of the model in a direct way. As discussed in the previous chapters, state space models have also found applications in fMRI data analysis, as flexible trend models and as models for the haemodynamic response. The aim of this chapter is not to give an introduction to the field as such, but to introduce the work on state space models presented in Paper V and VI. A short introduction to state space models and the Kalman filter, with the Canadian Lynx data (Elton and Nicholson, 1942) as a motivating example, may be found in Paper VI. Chapter 8 of Fahrmeir and Tutz (1994) is a concise exposition of the general theory, with special focus on generalized linear models, and also Künsch (1999) gives a good and short introduction. The book by West and Harrison (1989) provides a more detailed discussion, particularly from a Bayesian point of view, with many good examples.

4.1 Definition and examples

A state space model is a model for a sequence of dependent observation $\{Y_t\}$ where the dependence is modelled through an unobserved Markov process $\{X_t\}$. The latter is often referred to as the *latent process* or the *regime*. Given X the Y_t's are independent, and the conditional distribution of Y_t depends on X_t only. A general state space model may thus be formulated as,

$$\begin{aligned} Y_t \mid X_t &\sim g_\theta(y_t \mid x_t), \\ X_t \mid X_{t-1} &\sim \alpha_\theta(x_{t-1}, x_t), \end{aligned}$$

for $t = 1, \dots, n$ and $X_0 \sim \pi_\theta(x_0)$. Here $\alpha_\theta(\cdot, \cdot)$ is the transition density of the latent Markov process, $g_\theta(\cdot \mid \cdot)$ is the conditional density of Y_t given X_t, and $\theta \in \Theta \in \mathbb{R}^d$ is a parameter of the model.

The variables may have any dimension and the state spaces may be continuous or discrete. In the case where the latent process X takes values in a discrete finite space, the term *hidden Markov model* is often used instead of state space model.

An appealing feature of the model is the fact, that X may often be directly interpreted as a process which is driving the system, and which we are interested in estimating. This is the case in many applications in biology and finance. Alternatively the model may be a mathematical convenient way of formulating a specific correlation structure, for instance all ARMA processes may be formulated as state space models.

4.1.1 The linear Gaussian model

Suppose that $Y_t \in \mathbb{R}^k$ and $X_t \in \mathbb{R}^d$. The linear Gaussian state space model assumes the following form of the transition densities,

$$\begin{aligned} Y_t &= F_t X_t + \nu_t, \qquad & \nu_t &\sim N_k(0, V_t), \\ X_t &= G_t X_{t-1} + \omega_t, \qquad & \omega_t &\sim N_d(0, W_t), \end{aligned} \tag{4.1}$$

for $t = 1, 2, \dots, n$ and $X_0 \sim N_d(m_0, C_0)$. Here F_t and G_t are known $k \times d$ and $d \times d$ matrices respectively, and the error sequences $\{\nu_t\}$ and $\{\omega_t\}$ are serially and mutually independent. Despite the general formulation, this model is very easy to analyse, since all relevant conditional and joint distributions are Gaussian, and the corresponding moments may be calculated efficiently by the recursive Kalman filter[1] and smoother. As an example, the filter and smoother gives formulas for the moments of the conditional distributions $(X_t \mid Y_1, \ldots, Y_n)$, $(X_t \mid Y_1, \ldots, Y_t)$, $(X_t \mid X_{t+1}, Y_1, \ldots, Y_n)$ and $(Y_t \mid Y_1, \ldots, Y_{t-1})$ for $t = 1, \ldots, n$. In particular the likelihood function $p_\theta(Y_1, \ldots, Y_n)$ may easily be obtained.

4.1.2 Applications in fMRI analysis

We have described applications of the linear Gaussian state space model for the analysis for fMRI data in the earlier chapters. We will reconsider some of them here, as well as give a new example.

Gössl *et al.* (2000) considered the state space model (2.3) on page 17 for the trend and the haemodynamic response. For a particular voxel i, it may be rewritten in the above form by setting $X_t = (a_{it}, a_{it-1}, b_{it}, b_{it-1})$, $F_t = (1, 0, z_{it}, 0)$, $V_t = \sigma_i^2$,

$$G_t = \begin{pmatrix} 2 & -1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 2 & -1 \\ 0 & 0 & 1 & 0 \end{pmatrix} \quad W_t = \begin{pmatrix} \sigma_{\zeta_i}^2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & \sigma_{\eta_i}^2 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}.$$

The unknown parameters are V_t and W_t, which are estimated by the EM-algorithm. Here the Kalman filter and smoother are used to calculate the conditional mean and variance of X_t given $Y_1, \ldots, Y_n$, which are needed in the expectation step of the algorithm, and for estimating the haemodynamic response $z_{it}b_{it}$ by the posterior mean $E(z_{it}b_{it} \mid Y_{i1}, \ldots, Y_{in})$, with the estimated variance parameters inserted.

In Paper I the state space model (3.9) is considered for the haemodynamic response function. Let $Y_{\star t} = (Y_{it})_{i \in V}$ denote the vector of all voxel intensities at time t, and let the spatial activation magnitude $(A_i + \eta_i)_{i \in V}$ in the model (3.8) be denoted by A, we will condition on a known value of the latter for the moment. Suppose that the noise is temporally uncorrelated, $\Lambda = I_n$. The model is then of the form

$$\begin{aligned} Y_{\star t} &= A\varphi_t + \varepsilon_{\star t}, & \varepsilon_{\star t} &\sim N_{|V|}(0, \sigma^2\Gamma), \\ \varphi_t &= \varphi_{t-1} + \omega_t, & \omega_t &\sim N(\lambda_t - \lambda_{t-1}, \tau^2). \end{aligned}$$

[1] The term “filter” may be a bit confusing for a statistician: The Kalman filter has nothing to do with a filter in the probabilistic sense of a sequence of increasing σ-fields. A filter in this context is a device for estimating an underlying component X_t from a sequence of noisy observations $Y_1, \ldots, Y_t$.

where the noise sequences $\{\varepsilon_{\star t}\}_t$ and $\{\omega_t\}$ are mutually and serially uncorrelated. Besides the drift $\lambda_t - \lambda_{t-1}$ in the state equation, this is the same setup as in (4.1). In the paper, inference is made by an MCMC algorithm, where recursively the spatial pattern A and the temporal response φ is simulated. The formulas of the Kalman smoother allow direct simulation of the posterior distribution of φ given (Y, A): We simply start from the back by simulating φ_n, and then work our way backwards by simulating φ_t from the conditional distribution $(\varphi_t \mid \varphi_{t+1})$, which is given in closed form.

It is possible to extend the state space framework to some non-diagonal Λ-covariance matrices. In the paper, a first order autoregressive noise model is considered, this may be accommodated by augmenting the state process with the noise term ε_t, and modifying the observation equation accordingly.

Büchel and Friston (1998) present an application of a state space model in fMRI in a different context. The authors study effective connectivity, which is defined as the influence that one neural system exerts over another, and they are interested in how the effective connectivity may be modulated by attention. They conducted a visual fMRI study, where two different stimuli were presented, in both the subject was watching a screen with white dots emerging radially from the centre point. The subject was instructed to fixate on the centre. In the “attention” task the subject was instructed to “detect changes” in speed and in the “no-attention” to just look. The speed of the dots was constant, but psychophysical tests prior to scanning induced the anticipation of speed changes. The hypothesis of the experiment was that attention has a modulating effect on the connectivity between the motion sensitive area V5 in the visual cortex and the region known as the posterior parietal cortex (PP). Let Y_t and F_t denote the fMRI time series from respectively PP and V5. Their model was then,

$$\begin{aligned} Y_t &= F_t X_t + \varepsilon_t, & \varepsilon_t &\sim N(0, \sigma^2), \\ X_t &= X_{t-1} + \omega_t, & \omega_t &\sim N(0, \tau^2), \end{aligned}$$

where X_t is interpreted as an index of effective connectivity. By the Kalman filter and smoother, estimates of the smoothed values $\hat{X}_t = E(X_t \mid Y_1, \ldots, Y_n)$ were obtained. It was demonstrated that these were significantly higher during attention periods than during no-attention, which was interpreted as the influence attention has on the connection between V5 and PP. The estimated time course $\hat{X}_t$ was further shown to correlate well with the time series of a third region which was interpreted as the source of the modulation.

We discuss the work here as an interesting example of how the estimated latent process may be used to address a specific neuroscientific hypothesis. A different matter is that one should be very careful when interpreting the results from this statistical framework, which can only provide a measure of correlation. The interpretations in terms of *causal* dependencies, that the authors make, must be based on non-statistical arguments. The authors refer to previously verified hypotheses and general knowledge of connectivity of the brain for this. It is beyond most statisticians training to judge these claims, but it *is* the statisticians job to warn researchers about misinterpretation of detected dependencies, which is hereby done.

4.1.3 Non-linear models

Suppose instead of (4.1) the model is

$$\begin{aligned} Y_t &= h(X_t) + \nu_t, & \nu_t &\sim N_k(0, V_t), \\ X_t &= G_t X_{t-1} + \omega_t, & \omega_t &\sim N_d(0, W_t), \end{aligned} \tag{4.2}$$

where $h(\cdot) : \mathbb{R}^d \to \mathbb{R}^k$ is a non-linear function. In this situation, where the conditional mean of Y_t is not a linear function of X_t, the simple framework of the Kalman filter breaks down. The distributions are no longer Gaussian, and there are no simple formulas for the moments.

The typical approach to a problem of this kind, is to twist the model into the standard framework by simplifying assumptions. This is also the case for the earliest solutions, which are based on a linearization of h by a Taylor expansion. The approximate model is then linear and all distributions are normal. Several techniques of this flavour have been proposed, and carry names such as generalized Kalman filtering, extended Kalman filtering, non-linear filtering and so forth (West and Harrison, 1989).

An alternative is to keep the non-linear function h but to approximate the distributions of the latent process by normal distributions, where the moments are calculated numerically. This is the approach studied by Schnatter (1992), Frühwirth-Schnatter (1994) and in Paper VI. Let $Y_1^t = (Y_1, \dots, Y_t)$, and suppose that $X_{t-1} \mid Y_1^{t-1} \sim N(m_{t-1}, C_{t-1})$. Then $X_t \mid Y_1^{t-1} \sim N(a_t, R_t)$ where $a_t = G_t m_{t-1}$ and $R_t = G_t C_{t-1} G_t' + W_t$. By Bayes theorem, the posterior distribution $p(X_t \mid Y_1^t)$ is given by

$$p(X_t \mid Y_1^t) \propto p(Y_t \mid X_t) p(X_t \mid Y_1^{t-1}).$$

This will in general not be normal, but is approximated by a normal distribution. The moments of $X_t \mid Y_1^t$ are obtained by numerical integration with respect to the unnormalized density.

In the papers by Frühwirth-Schnatter a generalized linear model framework is assumed, where

$$g(E(Y_t \mid X_t)) = H \cdot X_t, \quad H \in \mathbb{R}^d,$$

for a monotone function g, and where Y_t may have any exponential family distribution. In Paper VI we restrict ourselves to normal distributions, but consider a different non-linear structure. We split the latent process into two components, $X_t = (X_{t,1}, X_{t,2})$, and assume that the model is linear if we condition on $X_{t,1}$. This allows us to use the ordinary Kalman filter conditionally on the value of $X_{t,1}$ and then only perform the numerical integration with respect to this variable. Since numerical integration is very inefficient for higher (more than three) dimensions, the proposed dimensionality reduction is a crucial step in many applications. The model for dynamic pulse rate in (2.1) and (2.2) has this form: Conditionally on the phase of the pulse v_t the model is linear. The numerical integration needs only be performed in one dimension in this case, even if the dimension of the state space is 8.

Numerical integration approaches with essentially no assumptions on the state space model have been studied by Kitagawa (1987) and West and Harrison (1989). They use a finite grid of points to approximate the density of the latent process $X_{t-1} \mid Y_1^{t-1}$ at each time point t. The updating from one posterior to the next may then be calculated directly. The number of grid points represents a compromise between speed and accuracy of the approximation. West and Harrison (1989) argue that the choice of the finite set of points is crucial, and designs an adaptive way of updating the set such that the grid evolves in time. For both methods, the computations may be very time consuming, especially if the dimension of the latent process is large.

The setup in Paper VI is in fact a special case of the framework named partial non-Gaussian state space by Shephard (1994), who designed a Gibbs sampler for making inference. The special structure of the model is advantageous also in this situation, since the ordinary Kalman smoother may be used for simulation of one part of the latent process conditionally on the other. There has been many other approaches to simulation based inference in state space models; see Durbin and Koopman (1997), Shephard and Pitt (1997) and the references therein.

4.2 Asymptotic normality of the MLE

In Paper V we prove that for a general state space model, with a probability that tends to 1 as $n \to \infty$, there exists a sequence $\{\hat{\theta}_n\}$ of (local) maximum points of the likelihood function, which is consistent and asymptotically normal,

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \to N(0, \mathcal{I}_0^{-1}), \quad P_{\theta_0}\text{-weakly as } n \to \infty.$$

In particular the maximum likelihood estimator is asymptotically normal, if it exists and is consistent. The information matrix $\mathcal{I}_0$ may be obtained as the limit of the observed information,

$$\frac{1}{n}\frac{\partial^2}{\partial\theta\partial\theta'}l_n(\hat{\theta}_n) \xrightarrow{P_0} -\mathcal{I}_0, \quad n \to \infty,$$

where $l_n(\theta)$ is the log-likelihood function corresponding to n observations. The results are valid for general stationary state space models, however to prove mixing results, we need bounds σ and M on the transition densities,

$$0 < \sigma \leq \alpha_\theta(x, z) \leq M < \infty \quad \text{for all } x, z \text{ and } \theta$$

(Assumption A1), as well as an upper bound on

$$\sup_\theta \sup_{x,x'} \frac{g_\theta(y \mid x)}{g_\theta(y \mid x')}$$

(Assumption A3). Typically these are only fulfilled if the state space of the latent process is compact.

The proof is based on the fact that the score function may be written as

$$\frac{\partial}{\partial\theta}l_n(\theta) = \sum_{t=1}^{n} \frac{\partial}{\partial\theta} \log p_\theta(Y_t \mid Y_1^{t-1}).$$

As $t \to \infty$ the terms in the sum will tend to a stationary martingale increment sequence, and the score function will hence approach a martingale in the limit as $n \to \infty$. The asymptotic normality is then obtained from a martingale central limit theorem.

Our proof is an extension of the work by Bickel *et al.* (1998), who consider the case of hidden Markov models, where the latent process takes values in a finite space. They in turn build on the proof by Baum and Petrie (1966), where both the latent and observed state spaces are finite. Besides these works, Leroux (1992) studied hidden Markov models, and proved consistency of the MLE. Recently Douc and Matias (1999) have studied a different technique for proving asymptotic normality, where the chain does not have to be stationary, as long as the model is time-invariant. Their assumptions are, however, very similar to the ones in Paper V, and they do hence not provide the final proof, for non-compact latent state spaces.

5 Concluding remarks

Let me conclude with a few general comments on fMRI data and statistics—as only a young researcher would be foolish enough to do. In many ways fMRI represents a new type of data, which is becoming increasingly common in applied statistics, and which may eventually force us to take a fresh look at some of the pillars of theoretical statistics.

Traditionally data is viewed as an almost sacred quantity of information in statistics. We are taught that one must never throw away data but only consider sufficient reductions, and we strive to find estimators which use the information with full efficiency. Besides the mathematical interest in deriving estimators with optimal properties, this is also due to the fact that data is often very expensive and difficult to obtain, and should be treated with this in mind. In fMRI studies a data set of several millions of observations is acquired within one minute. If something goes wrong, another set is acquired, virtually free of running costs. Clearly neuroscientific studies are designed carefully and needs much planning and consideration, but a data set is not a treasure which is passed from teacher to student in this world. Seen in this light one may very well question the practical relevance of the optimality conditions that are usually undisputed in theoretical statistics.

Another feature of fMRI data is the problem of determining what the data actually is. The raw observations of the MR scanner are samples of currents in a coil, which form an image in a Fourier space. The scanner automatically performs spatial filtering, Fourier transformation and possibly also spatial interpolation. Next different pre-processing routines are applied to correct for image artifacts and for movement effects. By the time the statistician actually gets the data, it has already been processed and analysed at several different stages. We should not regret this, because the pre-processing steps improve the quality of the data and makes the statisticians job much easier. But it is a bit worrying to know that it is *never* possible to replicate the conditions under which your data was generated. How should we interpret a frequentistic or asymptotic argument, when we know that there is no such thing as an exact replication of the same experiment?

Most likely many types of data of the next century will possess features like these, and they will maybe change the field of statistics as much as the data of this century has. Seen in this light, this thesis is only an infinitesimal step!

References

Adler, R.J. (1981) *The Geometry of Random Fields*. John Wiley & Sons.

Adler, R.J. (1998) On excursion sets, tube formulae, and maxima of random fields. *Annals of Applied Probability*. In press. See http://iew3.technion.ac.il:8080/~radler/publications.html.

Ardekani, B.A., Kershaw, J., Kashikura, K. and Kanno, I. (1999) Activation detection in functional MRI using subspace modeling and maximum likelihood estimation. *IEEE Trans. Med. Imaging*, **18**, 101–114.

Baddeley, A.J. and van Lieshout, M.N.M. (1993) Stochastic geometry models in high-level vision. In K.V. Mardia and G.K. Kanji (eds.), *Statistics and Images*, vol. 1, chap. 11, pp. 231–256, Appl. Statist.

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

Baum, L.E. and Petrie, T. (1966) Statistical inference for probabilistic functions of finite state Markov chains. *Ann. Math. Statist.*, **37**, 1554–1563.

Besag, J. (1986) On the statistical analysis of dirty pictures. *J. R. Statist. Soc. Ser. B*, **48**, 259–302.

Bickel, P.J., Ritov, Y. and Rydén, T. (1998) Asymptotic normality of the maximum-likelihood estimator for general hidden Markov models. *Ann. Statist.*, **26**, 1614–1635.

Biswal, B., DeYoe, E.A. and Hyde, J.S. (1996) Reduction of physiological fluctuations in fMRI using digital filters. *Magn. Reson. Med.*, **35**, 107–113.

Boynton, G.M., Engel, S.A., Glover, G.H. and Heeger, D.J. (1996) Linear systems analysis of functional magnetic resonance imaging in human V1. *J. Neurosci.*, **16**, 4207–4221.

Büchel, C. and Friston, K.J. (1998) Dynamic changes in effective connectivity characterized by variable parameter regression and Kalman filtering. *Human Brain Mapping*, **6**, 403–408.

Bullmore, E., Brammer, M., Williams, S.C., Rabe-Hesketh, S., Janot, N., David, A., Mellers, J., Howard, R. and Sham, P. (1996) Statistical methods of estimation and inference for functional MR image analysis. *Magn. Reson. Med.*, **35**, 261–277.

Bullmore, E.T., Suckling, J., Overmeyer, S., Rabe-Hesketh, S., Taylor, E. and Brammer, M.J. (1999) Global, voxel, and cluster tests, by theory and permutation, for a difference between two groups of structural MR images of the brain. *IEEE Trans. Med. Imag.*, **18**, 32–42.

Buonocore, M.H. and Maddock, R.J. (1997) Noise suppression digital filter for functional magnetic resonance imaging based on image reference data. *Magn. Reson. Med.*, **38**, 456–469.

Buxton, R.B., Wong, E.C. and Frank, L.R. (1998) Dynamics of blood flow and oxygenation changes during brain activation: the balloon model. *Magn. Reson. Med.*, **39**, 855–864.

Canet, D. (1996) *Nuclear Magnetic Resonance: Concepts and Methods*. Chichester: John Wiley.

Cao, J. and Worsley, K. (1999) Applications of random fields in human brain mapping. In *Proceedings of the Atelier sur la Statistique du Mappage du Cervau / Workshop on Statistics of Brain Mapping*, Centre de Recherche Mathématique, Université de Montréal. Juin 13-14, 1998.

Cohen, M.S. (1997) Parametric analysis of fMRI data using linear systems methods. *NeuroImage*, **6**, 93–103.

Cohen, M.S. and Bookheimer, S.Y. (1994) Localization of brain function using magnetic resonance imaging. *Trends in neurosciences*, **17**, 268–277.

Cox, D.R. and Miller, H.D. (1965) *The Theory of Stochastic Processes*. London: Chapman and Hall Ltd.

Dale, A.M. and Buckner, R.L. (1997) Selective averaging of rapidly presented individual trials using fMRI. *Hum Brain Mapping*, **5**, 329–340.

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

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

Douc, R. and Matias, C. (1999) Asymptotics of the maximum likelihood estimator for general hidden Markov models. Preprint, Université Paris Sud, Orsay.

Durbin, J. and Koopman, S.J. (1997) Monte Carlo maximum likelihood estimation for non-Gaussian state space models. *Biometrika*, **84**, 669–684.

Elton, C. and Nicholson, M. (1942) The ten-year cycle in numbers of the lynx in Canada. *J. Anim. Ecol.*, **11**, 215–244.

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

Fahrmeir, L. and Tutz, G. (1994) *Multivariate statistical modelling based on generalized linear models*. New York: Springer-Verlag. With contributions by Wolfgang Hennevogl.

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

Frank, L.R., Buxton, R.B. and Wong, E.C. (1998) Probabilistic analysis of functional magnetic resonance imaging data. *Magn. Reson. Med.*, **39**, 132–148.

Friston, K.J., Jezzard, P. and Turner, R. (1994) The analysis of functional MRI time-series. *Human Brain Mapping*, **1**, 153–171.

Friston, K.J., Holmes, A.P., Poline, J.B., Grasby, P.J., Williams, S.C.R., Frackowiak, R.S.J. and Turner, R. (1995) Analysis of fMRI time-series revisited. *NeuroImage*, **2**, 45–53.

Friston, K.J., Holmes, A., Poline, J.B., Price, C.J. and Frith, C.D. (1996a) Detecting activations in PET and fMRI: Levels of inference and power. *NeuroImage*, **4**, 223–235.

Friston, K.J., Williams, S.R., Howard, R., Frackowiak, R.S.J. and Turner, R. (1996b) Movement-related effects in fMRI time-series. *Magn. Reson. Med.*, **35**, 346–355.

Friston, K.J., Fletcher, P., Josephs, O., Holmes, A., Rugg, M.D. and Turner, R. (1998a) Event-related fMRI: Characterizing differential responses. *NeuroImage*, **7**, 30–40.

Friston, K.J., Josephs, O., Rees, G. and Turner, R. (1998b) Nonlinear event-related responses in fMRI. *Magn. Reson. Med.*, **39**, 41–52.

Frühwirth-Schnatter, S. (1994) Applied state space modelling of non-Gaussian time series using integration-based Kalman filtering. *Stat. Comp.*, **4**, 259–269.

Gaschler-Markefski, B., Baumgart, F., Tempelmann, C., Schindler, F., Stiller, D., Heinze, H.J. and Scheich, H. (1997) Statistical methods in functional magnetic resonance imaging with respect to nonstationary time-series auditory cortex activity. *Magn. Reson. Med.*, **38**, 811–820.

Geman, S. and Geman, D. (1984) Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Trans. Pattern Anal. Machine Intell.*, **6**, 721–741.

Genovese, C.R. (2000) A Bayesian time-course model for functional magnetic resonance imaging data (with discussion). *J. Amer. Statist. Assoc.* To appear. See http://www.stat.cmu.edu/˜genovese/papers/fmri/.

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

Glover, G.H. (1999) Deconvolution of impulse response in event-related BOLD fMRI. *NeuroImage*, **9**, 416–429.

Gössl, C., Auer, D.P. and Fahrmeir, L. (2000) Dynamic models in fMRI. *Magn. Reson. Med.*, **43**, 72–81.

Green, P.J. (1995) Reversible jump markov chain monte carlo computation and bayesian model determination. *Biometrika*, **82**, 711–732.

Grenander, U. and Miller, M.I. (1994) Representation of knowledge in complex systems. (with discussion). *J. R. Statist. Soc. Ser. B*, **56**, 549–603.

Hjort, N.L. and Mohn, E. (1984) A comparison of some contextual methods in remote sensing. In *Proceedings 18th international symposium on remote sensing of the environment*, pp. 1693–1702, CNES, Paris.

Hjort, N.L. and Omre, H. (1994) Topics in spatial statistics. *Scand. J. Statist.*, **21**, 289–357.

Højen-Sørensen, P.A.d.F.R., Hansen, L.K. and Rasmussen, C.E. (2000) Bayesian modelling of fMRI time series. In S.A. Solla, T.K. Leen and K.R. Müller (eds.), *Advances in Neural Information Processing Systems*, vol. 12. MIT Press.

Holmes, A.P., Blair, R.C., Watson, J.D.G. and Ford, I. (1996) Non-parametric analysis of statistic images from functional mapping experiments. *J. Cereb. Blood Flow Metab.*, **16**, 7–22.

Holmes, A.P., Josephs, O., Büchel, C. and Friston, K.J. (1997) Statistical modelling of low-frequency confounds in fMRI. *NeuroImage*, **5**, S480.

Hu, X., Lee, T.H., Parrish, T. and Erhard, P. (1995) Retrospective estimation and correction of physiological fluctuation in functional MRI. *Magn. Reson. Med.*, **34**, 201–212.

Josephs, O., Turner, R. and Friston, K. (1997) Event-related fMRI. *Human Brain Mapping*, **5**, 243–248.

Kiebel, S.J., Goebel, R. and Friston, K.J. (2000) Anatomically informed basis functions. *NeuroImage*, **11**, 656–667.

Kitagawa, G. (1987) Non-Gaussian state space modelling of non-stationary time series (with discussion). *J. Amer. Statist. Assoc.*, **82**, 1032–1063.

Kornak, J., Haggard, M.P. and 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 and I.L. Dryden (eds.), *Spatial Temporal Modelling and its Applications*. Leeds University Press.

Künsch, H.R. (1999) State space and hidden Markov models. In *Complex Stochastic Systems*, Semstat, Séminaire Européen de Statistique, EURANDOM, Eindhoven. See http://www-m4.mathematik.tu-muenchen.de/m4/lect-conf/semstat.html.

Kwong, K.K., Belliveau, J.W., Chesler, D.A., Goldberg, I.E. *et al.* (1992) Dynamic magnetic resonance imaging of human brain activity during primary sensory stimulation. *Proc. Natl. Acad. Sci. USA*, **89**, 5675–5679.

Lange, N. (1996) Tutorial in biostatistics. Statistical approaches to human brain mapping by functional magnetic resonance imaging. *Statistics in Medicine*, **15**, 389–428.

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

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

Le, T.H. and Hu, X. (1996) Retrospective estimation and correction of physiological artifacts in fMRI by direct extraction of physiological activity from MR data. *Magn. Reson. Med.*, **35**, 290–298.

Ledberg, A., Åkerman, S. and Roland, P.E. (1998) Estimation of the probabilities of 3D clusters in functional brain images. *NeuroImage*, **8**, 113–128.

Lee, A.T., Glover, G.H. and Meyer, C.H. (1995) Discrimination of large venous vessels in time-course spiral blood-oxygen-level-dependent magnetic-resonance functional neuroimaging. *Magn. Reson. Med.*, **33**, 745–754.

Leroux, B.G. (1992) Maximum-likelihood estimation for hidden Markov models. *Stoch. Proc. Appl.*, **40**, 127–143.

Liu, A.K., Belliveau, J.W. and Dale, A.M. (1998) Spatiotemporal imaging of human brain activity using functional MRI constrained magnetoencephalography data: Monte Carlo simulations. *Proc. Natl. Acad. Sci. USA*, **95**, 8945–8950.

Locascio, J.J., Jennings, P.J., Moore, C.I. and Corkin, S. (1997) Time series analysis in the time domain and resampling methods for studies of functional magnetic resonance brain imaging. *Human Brain Mapping*, **5**, 168–193.

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

Marchini, J.L. and Ripley, B.D. (2000) A new statistical approach to detecting significant activation in functional MRI. *NeuroImage*. To appear. See http://www.stats.ox.ac.uk/˜marchini/index.html.

Mardia, K.V., Kent, J.T. and Bibby, J.M. (1979) *Multivariate analysis*. London: Academic Press [Harcourt Brace Jovanovich Publishers]. Probability and Mathematical Statistics: A Series of Monographs and Textbooks.

Meloche, J. and Zamar, R.H. (1994) Binary-image restoration. *Canadian J. Statist.*, **22**, 335–355.

Müller, H.G. (1988) *Nonparametric Regression Analysis of Longitudinal Data*. Lecture Notes in Statistics. Springer-Verlag.

Nielsen, F.Å., Hansen, L.K., Toft, P., Goutte, C., Lange, N., Strother, S.C., Mørch, N., Svarer, C., Savoy, R., Rosen, B., Rostrup, E. and Born, P. (1997) Comparison of two convolution models for fMRI time series. *NeuroImage*, **5**, S473.

Petersson, K.M., Nichols, T.E., Poline, J.B. and Holmes, A.P. (1999a) Statistical limitations in functional neuroimaging. I. Non-inferential methods and statistical models. *Philos. Trans. R. Soc. Lond. B Biol. Sci.*, **354**, 1239–1260.

Petersson, K.M., Nichols, T.E., Poline, J.B. and Holmes, A.P. (1999b) 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. and Mazoyer, B.M. (1993) Analysis of individual positron emission tomography activation maps by detection of high signal-to-noise-ratio pixel clusters. *J. Cereb. Blood Flow Metab.*, **13**, 425–437.

Poline, J.B. and Mazoyer, B.M. (1994) Analysis of individual brain activation maps using hierarchical description and multiscale detection. *IEEE trans. med. imaging*, **13**, 702–710.

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

Purdon, P.L. and Weisskoff, R.M. (1998) Effect of temporal autocorrelation due to physiological noise and stimulus paradigm on voxel-level false-positive rates in fMRI. *Human Brain Mapping*, **6**, 239–249.

Roland, P.E., Levin, B., Kawashima, R. and Åkerman, S. (1993) Three-dimensional analysis of clustered voxels in 15O-butanol activation images. *Human Brain Mapping*, **1**, 3–19.

Rue, H. and Hurn, M.A. (1999) Bayesian object identification. *Biometrika*, **86**, 649–660.

Salli, E., Visa, A., Aronen, H.J., Korvenoja, A. and Katila, T. (1999) Statistical segmentation of fMRI activations using contextual clustering. In *Proc. of the Second International Conference on Medical Image Computing and Computer Assisted Intervention MICCAI'99*, vol. 1679 of *Lecture Notes in Computer Science*, pp. 481–488. Springer-Verlag.

Schnatter, S. (1992) Integration-based kalman-filtering for a dynamic generalized linear trend model. *Comp. Stat. Data analysis*, **13**, 447–459.

Shephard, N. (1994) Partial non-gaussian state space. *Biometrika*, **81**, 115–131.

Shephard, N. and Pitt, M.K. (1997) Likelihood analysis of non-Gaussian measurement time series. *Biometrika*, **84**, 653–667.

Siegmund, D.O. and Worsley, K.J. (1995) Testing for a signal with unknown location and scale in a stationary Gaussian random field. *Ann. Stat.*, **23**, 608–639.

Stark, D.D. and Bradley, W.G. (eds.) (1992) *Magnetic resonance imaging*. St. Louis, Missouri: Mosby Year Books, second edn.

Talairach, J. and Tournoux, P. (1988) *Co-planer Stereotaxic Atlas of the Human Brain*. New York: Thieme Medical Publishers.

Taskinen, I. (2000) Cluster priors in the Bayesian modelling of fMRI data. Ph. D. dissertation, Jyväskylä Studies in Computer Science, Economics and Statistics, Department of Statistics, University of Jyväskylä.

Thompson, P.M., Woods, R.P., Mega, M.S. and Toga, A.W. (2000) Mathematical/computational challenges in creating deformable and probabilistic atlases of the human brain. *Human Brain Mapping*, **9**, 81–92.

Vazquez, A.L. and Noll, D.C. (1998) Nonlinear aspects of the BOLD response in functional MRI. *NeuroImage*, **7**, 108–118.

West, M. and Harrison, J. (1989) *Bayesian Forecasting and Dynamic Models*. Springer-Verlag, New York.

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

Worsley, K.J. (1995a) Boundary corrections for the expected Euler characteristic of excursion sets of random fields, with an application to astrophysics. *Adv. Appl. Prob.*, **27**, 943–959.

Worsley, K.J. (1995b) Estimating the number of peaks in a random field using the Hadwiger characteristic of excursion sets, with applications to medical images. *The Annals of Statistics*, **23**, 640–669.

Worsley, K.J. (1996) The geomety of random images. *Chance*, **9**, 27–40.

Worsley, K.J. (1997) Comment on “Non-linear Fourier time series analysis for human brain mapping by functional magnetic resonance imaging”, by Lange and Zeger. *Appl. Statist.*, **46**, 25–26.

Worsley, K.J. (1998) Testing for signals with unknown location and scale in a χ^2 random field, with an application to fMRI. *Advanced in Applied Probability*. Accepted subject to revision. See http://www.math.mcgill.ca/˜keith/.

Worsley, K.J. (2000a) Comment on “A Bayesian time-course model for functional magnetic resonance imaging” by Chris Genovese. *J. Amer. Statist. Assoc.* To appear. See http://www.math.mcgill.ca/˜keith/.

Worsley, K.J. (2000b) Statistical analysis of activation images. In P.M. Matthews, P. Jezzard and S.M. Smith (eds.), *Functional Magnetic Resonance Imaging of the Brain: Methods for Neuroscience*, chap. 14, Oxford University Press. Submitted. See http://www.math.mcgill.ca/˜keith/.

Worsley, K.J. and Friston, K.J. (1995) Analysis of fMRI time-series revisited — again. *NeuroImage*, **2**, 173–181.

Worsley, K.J., Evans, A.C., Marret, S. and Neelin, P. (1992) A three-dimensional statistical analysis for cbf activation studies in human brain. *Journal of Cerebral Blood Flow and Metabolism*, **12**, 900–918.

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

Xiong, J., Gao, J.H., Lancaster, J.L. and Fox, P.T. (1995) Clustered pixels analysis for function MRI activation studies of the human brain. *Human Brain Mapping*, **3**, 287–301.

Zarahn, E., Aguirre, G.K. and D'Esposito, M. (1997) Empirical analyses of BOLD fMRI statistics. I. Spatially unsmoothed data collected under null-hypothesis conditions. *NeuroImage*, **5**, 179–197.

A stochastic geometry model for fMRI data

Niels Væver Hartvig[*]

University of Aarhus

Abstract

Functional magnetic resonance imaging (fMRI) is a principal method for mapping the human brain. fMRI data consist of a sequence of MR scans of the brain acquired during stimulation of specific cortical areas, and the purpose of analysing the data is to detect activated areas, i.e. areas where the intensity changes according to the stimulation paradigm. A common analysis procedure is to estimate the activity pattern non-parametricly by smoothing the data spatially. The focus is then on assessing significance of peaks or clusters in the smoothed activation surface by means of multiple hypothesis testing, rather than assessing the uncertainty of the estimated pattern itself. In this paper we formulate a more structured model for the spatial activation pattern. We achieve this by considering a stochastic geometry model where the activation surface is given by a sum of Gaussian functions, which to some extent can be thought of as individual centres of activation in the brain. The model is formulated in a Bayesian framework, where the prior distribution of the centres is given by a marked point process density. An advantage of this approach is that inference can be carried out by simulation techniques, and hence it is easy, though time consuming, to evaluate the uncertainty of the estimate or to test hypotheses of interest regarding the activation. Furthermore in this framework, we are able to model the temporal pattern of the activation with fewer assumptions than usually imposed. This reveals significant non-stationarities in the analysed data, which violate the common assumption of stationarity of the haemodynamic response.

Keywords: Functional magnetic resonance imaging; Stochastic geometry model; Marked point proces; Markov chain Monte Carlo; State space model; Correlogram

1 Introduction

Functional magnetic resonance imaging (fMRI) is a medical imaging technique where fast MR scanners are used to measure changes in blood oxygenation in the brain. The latter is known as the Blood Oxygen Level Dependent (BOLD) signal. These oxygenation changes

[*] Department of Mathematical Sciences, University of Aarhus, Ny Munkegade, DK-8000 Aarhus C, email: vaever@imf.au.dk

correlate with neural activity in the surrounding tissue, and hence fMRI is an indirect method for measuring activation in the brain. The technique is quite new, one of the first experiments was reported by Kwong *et al.* (1992) and since then the number of publications in the field has grown extremely fast. fMRI is a very attractive modality for imaging the brain, since it is non-invasive and has a good temporal and spatial resolution.

In a typical fMRI experiment a subjects brain is scanned while specific centres are stimulated, for instance the visual cortex can be activated by flashing a light in the eyes. The acquired data consist of a sequence of scans and the aim of the statistical analysis is to identify regions in the images, where the intensity changes according to the stimulus rhythm. A biostatistical introduction to the subject is given in Lange (1996), Lange and Zeger (1997) also contains a good introduction.

The analysis of the data is impeded by the uncertainty of the haemodynamic response to the stimulus. It is well known, that the response is delayed about 6 seconds and dispersed in time compared to the stimulus paradigm, but otherwise there is no general accepted biological model for the response, which can guide us when modelling the signal.

Another problem is the incorporation of spatial structure in the analysis. Of course the spatial activation pattern depends on the type of stimulation, and it is difficult to impose structure on this in a general setting. Instead, a common approach is to marginalize the analysis to a one dimensional time-series problem for each voxel in the scan, see for instance Worsley and Friston (1995), Lange and Zeger (1997) or Bullmore *et al.* (1996). The spatial structure of the data is included in a second step, when the image of activation estimates is convolved with a smoothing kernel to obtain a non-parametric estimate of the activation. In this approach there is no specific model for the spatial pattern of activation. Furthermore the focus is on assessing significance of peaks and clusters in the image by testing thousands of voxel-wise hypotheses simultanously.

In this paper we will focus on the issue of estimating the activation pattern, rather than testing multiple hypotheses. The model for the spatial pattern is based on two fundamental assumptions in the fMRI literature: 1) The activated areas have a spatial extent of several millimetres and 2) the activation pattern is “smooth”. Both assumptions are based on the haemodynamic origin of the signal: Even if neural activation is localized to a single voxel, say, the haemodynamic effects will occur in the surrounding venes, and will cover a larger area. We will incorporate these two assumptions in a stochastic geometry model based on marked point processes, see for instance Baddeley and van Lieshout (1993). This is done by modelling the spatial activation surface by a collection of Gaussian functions, which to some extent can be thought of as individual centres in the brain. The model is formulated in a Bayesian setting where the centres *a priori* are distributed as a marked point process; here the points are the locations of the centres and the marks describe the shape and height of the centres. The inference in the model is based on simulation techniques, by which we can estimate the posterior mean of functions of interest, such as the mean activation pattern.

The advantages compared to the common analysis procedure outlined above are many. 1) We don't have to smooth the data spatially, but can retain the detailed resolution of the MR scans. 2) We can assess the uncertainty of the estimated spatial pattern in a Bayesian framework. 3) We can quantify our belief in more specific hypotheses about the activation by estimating posterior probabilities in the model. 4) Finally we can model the haemodynamic response function, i.e. the temporal pattern of the activation, in a semi-parametric way, which allows for non-stationarities and non-linearities. With the latter approach explicit knowledge of the stimulation paradigm is not required, and we can hence estimate activation which is not time-locked to the stimulation rhythm.

The paper is organized as follows: In Section 2 we formulate the Bayesian model for the spatial activation pattern, and combine this with a simple initial model for the temporal response to obtain a spatio-temporal model. The temporal pattern is assumed to be known and described by a convolution model. This is somewhat restrictive, but it allows us to focus on the spatial pattern for a start, and discuss how we can simulate the latter from the posterior distribution. This is done by an MCMC algorithm, which is described in Section 3. In Section 4 we apply the model to a simulated data set, which is used for estimating prior parameters, and to visual stimulation data. In the next two sections we extend the model in different ways: In Section 5 we describe a state space model for the haemodynamic response function, and demonstrate its ability to model non-stationarities which are indeed present in the data. In Section 6 we extend the covariance structure to account for the spatio-temporal correlation, which is present in the noise. Finally we have a discussion in Section 7 and an appendix where theoretical properties of the MCMC algorithm are studied.

2 The model

2.1 Preprocessing of the data

Suppose the data consist of m scans, acquired with a stimulation paradigm $\pi_1, ..., \pi_m$, where $\pi_t = 1$ indicates stimulation and $\pi_t = 0$ no stimulation at time t. Typically the paradigm is arranged in blocks of, say, 10 scans with stimulation and 10 without. Let V be the set of voxels covering brain tissue, $V \subseteq S$, where S represents a 2 dimensional slice or a 3 dimensional volume of the brain. The dataset is hence given by a set of intensity measurements $Y = \{Y_{it}, i \in V, t = 1, \dots, m\}$.

The units of the intensities recorded by the MR scanner are arbitrary, and it is common in the literature to report variation of the signal in percent of baseline intensity. In order to consider variation of the intensity in different voxels on the same scale, we have log-transformed the data. Suppose for instance, that the measurement in a given voxel at time t is given by $Y_t = \mu(1 + \varepsilon_t)$, where ε_t is a deviation from the baseline intensity of the voxel. For small deviations we then have

$$\log Y_t = \log \mu + \log(1 + \varepsilon_t) \simeq \log \mu + \varepsilon_t,$$

and hence the magnitude of (structural and random) variations of the log data, can be compared between different timeseries. Furthermore, the unit of the deviations can be thought of as percent of baseline intensity.

We will preprocess the data, such that the images are aligned to correct for subject movement and has been corrected for trends. In our applications we have used a simple procedure, where each image is aligned to a reference image by minimizing the squared difference between the two images over all translations and rotations. As for the trend correction, we will consider the residuals after subtracting the mean and correcting for a linear trend in each individual time series. The presence of trends and low-frequency fluctuations in fMRI time series is often reported in the literature, though the processes which generate these are not well understood. Modelling these features as linear terms is necessarily an approximation, and more general models such as proposed by Holmes *et al.* (1997) and Petersen *et al.* (1998) may be applied. However, as will be described in Section 5, our aim is to model general temporal response patterns, and hence we are cautious not to remove any fluctuations caused by the haemodynamic response. A linear model is a good compromise in this context.

A basic assumption of the model is that the spatial and temporal patterns of the activation can be modelled separately. Considering an image or a volume of the activation magnitudes $A = \{A_i, i \in V\}$ and a timeseries $\varphi = \{\varphi_t, t = 1, \dots, m\}$ of the common temporal variation caused by the BOLD effect, we assume that the mean intensity measured in voxel i at time t is given by $EY_{it} = A_i \varphi_t$. We will now describe in detail how the spatial and temporal pattern are modelled.

2.2 A model for the spatial activation pattern

Consider first the case where data only represent a 2 dimensional slice of the brain, that is $V \subseteq S \subseteq \mathbb{R}^2$. The spatial activation pattern will be modelled as a collection of n "activation centres" $X = \{X_1, X_2, \dots, X_n\}$, each parametrized as $X_j = (\mu_j, a_j, d_j, r_j, \theta_j)$. The global pattern $A(X) = \{A_i(X) \mid i \in V\}$ is given by the superposition of n bells,

$$A_i(X) = h(i; X_1) + \cdots + h(i; X_n)$$

where

$$h(i; X_j) = a_j \exp\left\{ -\frac{\pi \log 2}{d_j} \left(\frac{\tilde{i}_1^2}{r_j/(1-r_j)} + \frac{\tilde{i}_2^2}{(1-r_j)/r_j} \right) \right\} \tag{1}$$

and $\tilde{i} = (\tilde{i}_1, \tilde{i}_2) = R(-\theta_j)(i - \mu_j)$. Here $R(\theta)$ is a rotation with angle θ. Hence $h(\cdot; X_j)$ is a Gaussian bell of height a_j centred at $\mu_j \in S$. The parameter $d_j \in R_+$ is the area of the contourellipse at half height, $r_j \in (0, 1)$ is a measure of the eccentricity of the ellipse, more precisely the ratio of the first principal axis and the sum of the two axes, and $\theta_j \in [-\pi/4, \pi/4]$ is the orientation of the ellipse. Notice that the angle is constrained to an interval of length $\pi/2$ to ensure identifiability of the parameters (r, θ).

In order for this specification to be meaningful, we need to restrict heights to be positive and incorporate some regularity in the point pattern. We will achieve this by formulating a prior model for X in the context of marked point processes, see for instance Møller (1999). Each centre $X_j = (\mu_j, a_j, d_j, r_j, \theta_j)$ is a point in $\mathcal{X} = S \times M$, where

$$M = [0, C_a] \times [0, C_d] \times (0, 1) \times [-\pi/4, \pi/4].$$

Here C_a and C_d are natural bounds for the height and area respectively. Though time series with negative activation amplitude is observed, we will initially assume that all bells have positive height. We will discuss later, how negative activation can be accounted for in the model.

Let $\mathcal{X}$ be equipped with the Borel σ-field $\mathcal{S} \times \mathcal{M}$ and the Lebesque measure $\lambda_2 \times \lambda_4$, and let Ω denote the exponential space over $\mathcal{X}$, that is the set of finite sets $\{x_1, \dots, x_n\}$ where $x_i \in \mathcal{X}$ for all i. The activation profile $X = \{X_1, \dots, X_n\}$ can then be interpreted as a point process in Ω or equivalently as a marked point process with point space S and mark space M. A priori we will assume that X has density wrt. the unit rate homogenous Poisson process on Ω of the form

$$f(x) \propto \beta^n \left(\prod_{i=1}^{n} \prod_{j=i+1}^{n} \phi(x_i, x_j) \right) \prod_{j=1}^{n} \{p(a_j)p(d_j)p(r_j)\}, \quad x = \{x_1, \dots, x_n\} \tag{2}$$

where $n = n(x)$ is the number of points in x and β is an intensity parameter. The pairwise interaction function ϕ introduces a regularity in X, discouraging configurations with centres placed "on top" of each other. A popular choice when modelling repulsive point patterns is the Strauss model with interaction radius $\rho > 0$ with respect to a metric $\delta(\cdot, \cdot)$ on $\mathcal{X}$. In this case ϕ is given by

$$\phi(\xi, \eta) = \gamma^{1(\delta(\xi,\eta)<\rho)}, \quad \xi, \eta \in \mathcal{X},$$

with $\gamma \in [0, 1]$ and with the convention that $0^0 = 1$. In our setup, however, we wish to impose a hard-core restriction, which prohibits pairs of centres with distances close to zero. The hard-core model with $\gamma = 0$ is not very suitable in this context, since the posterior distribution will be very sensitive to the choice of ρ. An appropriate alternative is the so-called very-soft-core model of Ogata and Tanemura (1984) with

$$\phi(\xi, \eta) = 1 - \exp\left\{-(\delta(\xi, \eta)/\rho)^p\right\}, \quad \xi, \eta \in \mathcal{X}, \quad p \geq 2. \tag{3}$$

The hard-core model is obtained by setting $p = \infty$, while finite values of p yield a continous interaction function which increases smoothly from 0 to 1 with the distance between two points. A plot of the interaction functions for different values of p can be seen in Figure 1.

The metric $\delta(\cdot, \cdot)$ should be defined such that two centres x_1 and x_2 are close, if they are close in space and have similar size and shape. One way of assessing this is by the *J-divergence* (Kullback, 1959) of the corresponding Gaussian functions: By rewriting the expression in (1), we find that the activation intensity $h(\cdot, x_j)$ induced by $x_j = (\mu_j, a_j, d_j, r_j, \theta_j)$ is given by $h(i; x_j) = a_j d_j f_j(i)/\log 2$, where $f_j(\cdot) = f(\cdot; \mu_j, \Sigma_j)$ is a multivariate normal density with mean μ_j and covariance matrix

$$\Sigma = \frac{d_j}{2\pi \log 2} R(\theta) \begin{pmatrix} \frac{r}{1-r} & 0 \\ 0 & \frac{1-r}{r} \end{pmatrix} R(-\theta).$$

The J-divergence between the two densities is now given by

$$\delta(x_1, x_2) = J(f_1, f_2) = \int (f_1(x) - f_2(x)) \log \frac{f_1(x)}{f_2(x)} \, \lambda_2(dx) \tag{4}$$

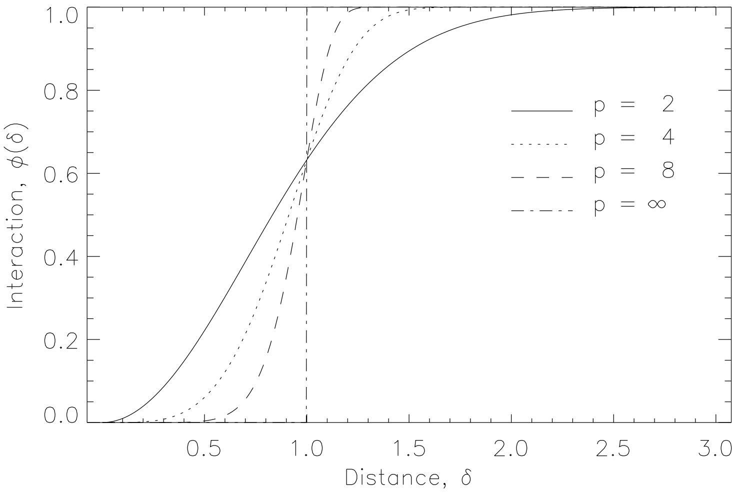


Figure 1: The soft-core interaction function $\phi(\xi, \eta)$ in (3) as a function of the distance $\delta(\xi, \eta)$, $\xi, \eta \in \mathcal{X}$. The parameter ρ equals 1.

and by inserting the means and variances we get

$$\delta(x_1, x_2) = -2 + \frac{1}{2}\left\{(\mu_1 - \mu_2)'(\Sigma_1^{-1} + \Sigma_2^{-1})(\mu_1 - \mu_2) + \mathrm{trace}(\Sigma_2^{-1}\Sigma_1 + \Sigma_1^{-1}\Sigma_2)\right\}. \quad (5)$$

Figure 2 is a plot illustrating distances between pairs of points with this metric. The expression can of course be rewritten in terms of the (d, r, θ)-parametrization of the Gaussian density.

The priors for a and d should be as uniform as possible, yet penalizing values close to zero. The inverted Gamma distribution is a suitable choice in this context, with its light tail near zero and its quite heavy tail for large values. Hence we will assume that $a^{-1} \sim \Gamma(2, \beta_a)$ and $d^{-1} \sim \Gamma(2, \beta_d)$ with the restrictions that $a \in (0, C_a]$ and $d \in (0, C_d]$. The density of d is

$$p(d) = \exp(\beta_d/C_d)(\beta_d/C_d + 1)^{-1}\beta_d^2 d^{-3}\exp(-\beta_d/d), \quad d \in (0, C_d].$$

In our application we set $\beta_a = 0.05$ and $\beta_d = 200$ mm^2, for comparison the voxels cover an area of 3.61 mm^2 in each slice in our data. As for the axis ratio r we wish to discourage very eccentric ellipses. This can be obtained by a Beta-prior, $r \sim$ Beta$(5, 5)$. Finally the angle θ is uniformly distributed on $[-\pi/4, \pi/4]$.

With this choice of prior, we assume that the intensity β of the centres is constant over V. An obvious refinement is to include covariate information on the underlying tissue and allow the intensity to depend on the location. Also the experimenter often has good prior knowledge of where the activation is likely to occur, which could be used, when specifying β.

This specification can straightforwardly be generalized to a 3-dimensional setting where $S \subseteq \mathbb{R}^3$. In this case a centre is given by $x = (\mu, a, d, r_1, r_2, \theta_1, \theta_2)$, and the contribution to

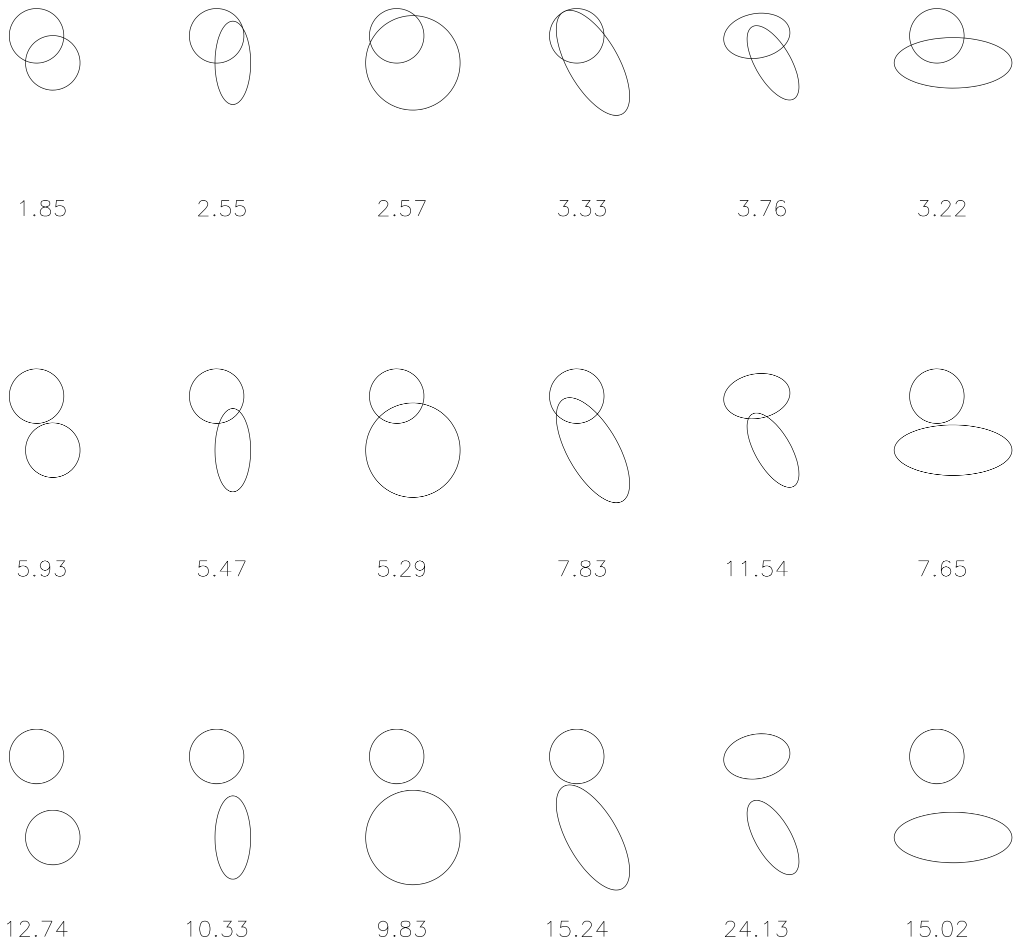


Figure 2: Examples of distances measured by the metric (5) on the product space of points and marks. Illustrated are pairs of countour ellipses at half height of the respective bells, together with their distance.

the activation volume is

$$h(i;x) = a \exp\left\{ -\log 2 \left(\frac{4\pi}{3d}\right)^{2/3} \left(\frac{\tilde{i}_1^2}{(r_1^2/r_2r_3)^{2/3}} + \frac{\tilde{i}_2^2}{(r_2^2/r_1r_3)^{2/3}} + \frac{\tilde{i}_3^2}{(r_3^2/r_1r_2)^{2/3}} \right) \right\}.$$

Here $r_3 = 1 - r_1 - r_2$, $r_i > 0$ for $i = 1, 2, 3$ and

$$\tilde{i} = (\tilde{i}_1, \tilde{i}_2, \tilde{i}_3) = \begin{pmatrix} \cos\theta_1 \cos\theta_2 & -\sin\theta_1 & -\cos\theta_1 \sin\theta_2 \\ \sin\theta_1 \cos\theta_2 & \cos\theta_1 & -\sin\theta_1 \sin\theta_2 \\ \sin\theta_2 & 0 & \cos\theta_2 \end{pmatrix} (i - \mu).$$

With this parametrization d is the volume of the contour ellipsoid at height $a/2$, and r_i is the ratio of the ith main axis and the sum of the three main axis. The angles θ_1 and θ_2 are the rotations in the xy-plane and xz-plane respectively, which are restricted to the interval $[-\pi/4, \pi/4]$. The natural extension of the priors is to assume that $(r_1, r_2) \sim D_2(5, 5)$ where D_2 is the two-dimensional Dirichlet distribution.

2.3 A model for the temporal pattern

In order to obtain a reasonably simple spatio-temporal structure in the model, we will assume that the temporal pattern $\varphi = \{\varphi_t, t = 1, \ldots, m\}$ is approximately the same for all voxels. Hence we assume that any voxel-wise differences in the delay is negligible, and we assume that the shape of the haemodynamic response is the same everywhere. This should be contrasted to, for instance, the approach in Lange and Zeger (1997), where differences from one voxel to another is explicitly accounted for. However, we will discuss later how the model can be extended in order to relax this assumption.

Initially we will follow the approach in Friston *et al.* (1995) and consider φ to be known and given by a convolution between the paradigm π and a Gaussian density of mean 6 seconds and variance 9 seconds^2, modelling the delay and dispersion of the signal. Hence

$$\varphi_t = \sum_i \pi_{t-i} \frac{T}{\sqrt{2\pi 3}} \exp(-\frac{(iT-6)^2}{18}), \tag{6}$$

where T is the repetition time, i.e. the time between two consecutive images. This is a rather simple model, and there is no particular reason for choosing a Gaussian density as the model for the impulse response function, neither is it obvious that the response is stationary. In Section 5 we will describe a more flexible semi-parametric model for the temporal pattern which does not require these assumptions. However the simulation procedure to be presented in the following section simplifies a great deal if we assume a known and fixed response, and we will thus start with this model.

2.4 Combining the spatial and temporal models

Given the centres X and the haemodynamic response function φ, the model for the intensity Y is,

$$Y_{it} = (A_i(x) + \eta_i)\varphi_t + \varepsilon_{it}, \quad \eta_i \sim N(0, \tau^2),\ \varepsilon_{it} \sim N(0, \sigma^2), \quad i \in V, t = 1, \ldots, m \tag{7}$$

where $\{\varepsilon_{it}\}$ and $\{\eta_i\}$ are independent white noise sequences. We assume a simple noise model, with the ε_{it}'s being independent, but more general covariance structures can be incorporated in a theoretically simple way, see Section 6. Also more complicated noise sources may be removed before the analysis, for instance by procedures in Le and Hu (1996) or Petersen *et al.* (1998).

The likelihood function in the model (7) is given by

$$p(Y|x) = (2\pi\sigma^2)^{-\frac{(m-1)|V|}{2}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i\in V}\sum_{t=1}^{m}\left(Y_{it} - \tilde{Y}_i\varphi_t\right)^2\right\} \times$$
$$(2\pi(\sigma^2 + \tau^2 \mathrm{ss}_\varphi))^{-\frac{V}{2}} \exp\left\{-\frac{1}{2(\sigma^2/\mathrm{ss}_\varphi + \tau^2)}\sum_{i\in V}\left(\tilde{Y}_i - A_i(x)\right)^2\right\}. \quad (8)$$

Here

$$\tilde{Y}_i = \sum_{t=1}^{m} Y_{it}\varphi_t/\mathrm{ss}_\varphi, \quad \mathrm{ss}_\varphi = \sum_{t=1}^{m}\varphi_t^2, \quad (9)$$

is the coefficient of the projection of $\{Y_{it}, t = 1, \ldots, m\}$ on the vectorspace $L = \mathrm{span}\{\varphi\}$. Notice that the likelihood function factorizes into two terms, involving only the projection of Y onto L and onto the orthogonal complement to L, respectively, with X only entering in the latter. Hence we find that $\{\tilde{Y}_i, i \in V\}$ is sufficient for X. The former is a regression image with the voxel-wise estimated activation amplitudes, this is also known as a *Statistical Parametric Map* (SPM) in the fMRI literature (Friston *et al.*, 1994). The estimation of the spatial pattern $A_i(x)$ can hence be viewed as a model based way of smoothing the SPM. This provides a link to more traditional methods, where the SPM is smoothed with a Gaussian filter, and afterwards regarded as a differentiable Gaussian random field for inference purposes. The model in this simplest setting hence provides an alternative estimate for the activation based on the raw SPM and a way of assessing the uncertainty of the estimate. In the more general setting described in Section 5, we can estimate φ semi-parametricly rather than assuming it is known, which is in general not possible in the traditional SPM approach.

The purpose of the random effect term η_i is to regularize the estimate of X. To see why this is necessary, consider the log posterior distribution of X, which up to an additive constant is given by

$$\log p(x|Y) = -\frac{1}{2(\sigma^2/\mathrm{ss}_\varphi + \tau^2)}\sum_{i\in V}\left(\tilde{Y}_i - A_i(x)\right)^2 + \log p(x).$$

Suppose for a moment that $\tau = 0$, corresponding to omitting the random effect η_i above. By inserting sufficiently many small bells, we can obtain a configuration where $A_i(x) = \tilde{Y}_i$ when the latter is positive, and $A_i(x) = 0$ elsewhere. This configuration will minimize the sum of squares above. Even if the prior density of such a pathological point configuration is very small, it will be the maximum aposteriori estimate in the limit as m, and hence ss_φ, tends to infinity, since the sum of squares will dominate in the limit. By assuming a fixed positive value for τ^2 this undesirable property of the posterior distribution is removed. Intuitively τ^2 is a measure of how well we expect the actual activation surface to be described by a reasonable collection of Gaussian functions, while the purpose of the prior for X is to quantify what we mean by a reasonable collection.

When applying the model, we will insert *ad hoc* estimates of σ^2 and τ^2. An unbiased and consistent estimator for σ^2 is given by

$$\hat{\sigma}^2 = \frac{1}{(m-1)|V|} \sum_{i \in V} \sum_{t=1}^{m} \left(Y_{it} - \tilde{Y}_i \varphi_t \right)^2 \sim \sigma^2 \chi^2(f)/f, \;\; f = (m-1)|V|. \tag{10}$$

As for τ^2, we will estimate $\sigma^2/\mathrm{ss}_\varphi + \tau^2$ by considering the regression coefficients $\tilde{Y}_i$. These are distributed as

$$\tilde{Y}_i \sim N(A_i(x), \sigma^2/\mathrm{ss}_\varphi + \tau^2), \quad i \in V,$$

with all $\tilde{Y}_i$'s independent. Letting ∂i denote the 9-voxel neighbourhood of i, we will let

$$\bar{Y}_i = \frac{1}{9} \sum_{j \in \partial i} \tilde{Y}_j \sim N(\bar{A}_i(x), \frac{1}{9}(\sigma^2/\mathrm{ss}_\varphi + \tau^2))$$

for $i \in V^\circ$, where $V^\circ = \{ i \in V \mid \partial i \subseteq V \}$. By assuming that the activation surface $A_i(x)$ can be approximated by a plane locally around i, we have that $A_i(x) = \bar{A}_i(x)$ and hence that

$$\frac{9}{8|V^\circ|} \sum_{i \in V^\circ} \left(\tilde{Y}_i - \bar{Y}_i \right)^2 \tag{11}$$

is an unbiased and consistent estimator for $\sigma^2/\mathrm{ss}_\varphi + \tau^2$. When the approximation is not exact, we will get a slight positive bias in the estimate for τ^2.

2.5 Modelling negative activation

So far we have only considered areas with increased intensity during stimulation, but in fact in some areas of the brain a parallel decrease in the intensity is observed. This is typically attributed to large veins or other types of non-neural tissue, and as such is not of primary interest in the analysis. However, in order to obtain a realistic model for the data, we need to consider this effect.

A natural way to proceed is to model the activation as $A^+ - A^-$, where A^+ and A^- are positive surfaces describing positive and negative activation respectively, and each has a prior similarly to the surface A described earlier. When doing this, we have to incorporate restrictions in the prior that separates the two surfaces for identifiability reasons. If not, the two surfaces may overlap to an extent where they cannot be individually identified as positive and negative activated areas, but rather positive and negative terms in a general surface, which is not necessarily given by sums of Gaussian functions. As a result of this the positive and negative centres will become highly correlated, and the interpretation of the activation surface becomes very difficult.

Suppose we let X^+ and X^- be two point processes modelled as described in Section 2.2, determining the postive and negative surfaces respectively. One way of separating the surfaces would be to model the prior as

$$p(X^+, X^-) \propto f(X^+)f(X^-)\exp(-\alpha \sum_{i\in V} A_i(X^+)A_i(X^-)),$$

where $f(\cdot)$ is the density in (2) and $\alpha > 0$. This prior allows some overlap between $A(X^+)$ and $A(X^-)$, but the last term penalizes configuration where $A_i(X^+)$ and $A_i(X^-)$ are both large for some $i \in V$. The parameter α determines the weight of the separation term in the prior. Suppose we let $\alpha^{-1} = \sigma^2/\mathrm{ss}_\varphi + \tau^2$, the variance of $\tilde{Y}_i$, such that $\sqrt{\alpha}A_i$ is given as units of standard deviation of $\tilde{Y}_i$. The posterior obtained with this prior is then

$$\begin{aligned} p(X^+, X^-|Y) &\propto \exp\left\{-\frac{1}{2(\sigma^2/\mathrm{ss}_\varphi + \tau^2)}\sum_{i\in V}\left(\tilde{Y}_i - [A_i(X^+) - A_i(X^-)]\right)^2\right\} \\ &\quad \times f(X^+)f(X^-)\exp\left\{-\frac{1}{\sigma^2/\mathrm{ss}_\varphi + \tau^2}\sum_{i\in V} A_i(X^+)A_i(X^-)\right\} \\ &\propto \exp\left\{-\frac{1}{2(\sigma^2/\mathrm{ss}_\varphi + \tau^2)}\sum_{i\in V}\left(\tilde{Y}_i - A_i(X^+)\right)^2\right\} f(X^+) \\ &\quad \times \exp\left\{-\frac{1}{2(\sigma^2/\mathrm{ss}_\varphi + \tau^2)}\sum_{i\in V}\left(\tilde{Y}_i + A_i(X^-)\right)^2\right\} f(X^-). \end{aligned}$$

This shows that X^+ and X^- are independent given the data Y, and the marginal distribution of X^+ is the same as that obtained when ignoring X^- as described in the previous sections. Hence if we choose to separate the surfaces by this choice of prior, we can make inference about X^+ and X^- in their respective marginal distributions, and afterwards combine estimates using the independence of the two point processes. Naturally this prior is only one suitable way of restricting the two activation patterns out of many. As an alternative one might model repulsion between points in the two point processes, or one could consider a hard-core restriction prohibiting the surfaces from overlapping more than a certain amount. In these cases the two point processes would not be independent in the posterior distribution. However, the independence argument above makes it plausible to make separate inference about positive and negative activation, or to only consider positive activation, which is the main parameter of interest. As there are considerable advantages of considering only one type of activation at a time, namely a reduction of the dimensionality of the point processes and improved properties of the simulation algorithm, we will henceforth marginalize the inference in this way.

3 Simulating from the posterior distribution

In order to explore the posterior distribution of the activation centres given the data, we have designed a Metropolis-Hastings algorithm based on the Geyer and Møller (1994) algorithm for general finite point processes. Let x be the current point configuration. We will then propose to 1) insert a new point, 2) remove an existing point or 3) change an existing point, with probabilities p_1, p_2 and p_3 respectively, where $p_1 + p_2 + p_3 = 1$. By "change an existing point" we mean that one of the coordinates of the point is changed, either the position or one of the mark-coordinates.

Let $q_m(x' \mid x)$ denote the proposal density of a new configuration x' based on the current configuration x with move type $m = 1, 2, 3$. The probability of accepting the move is then respectively

$$\alpha_1(x, x') = \min\left\{\frac{p(x' \mid Y) q_2(x|x') p_2}{p(x|Y) q_1(x'|x) p_1}, 1\right\},$$
$$\alpha_2(x, x') = \min\left\{\frac{p(x' \mid Y) q_1(x|x') p_1}{p(x|Y) q_2(x'|x) p_2}, 1\right\},$$
$$\alpha_3(x, x') = \min\left\{\frac{p(x' \mid Y) q_3(x|x')}{p(x|Y) q_3(x'|x)}, 1\right\}.$$

If the move is rejected, the Markov chain stays in x. The proposal distributions are described in detail in the following.

3.1 Insertion of a point

With probability p_1 we propose to add a new point $\xi = (\mu, a, d, r, \theta)$ to the existing point configuration $x = \{x_1, \ldots, x_n\}$. In order to obtain a reasonable acceptance rate for this move, we wish to perform a Gibbs-like update and sample the parameters from a density proportional to the Papangelou conditional intensity $p(x \cup \xi|Y)/p(x|Y)$. However this is a distribution on the 6 dimensional space of points and marks and it is not possible to simulate directly from it. Instead, we will propose the parameters (μ, a, d, r, θ) sequentially, hence the proposal $q_1(x \cup \xi|x)$ is a combination of the terms,

$$q_1(x \cup \xi|x) = q(\mu|x) q(a|\mu, x) q(d|\mu, a, x) q(r|\mu, a, d, x) q(\theta|\mu, a, d, r, x), \qquad (12)$$

where we use the generic symbol $q(\cdot|\cdot)$ for a proposal density. We will choose the proposal of a single parameter, a say, such that it resembles the conditional intensity of a point $(\mu, a, d_0, r_0, \theta_0)$ given the current configuration x, where (d_0, r_0, θ_0) are fixed typical values for the remaining parameters and μ is the proposed position of the point. In our applications we have chosen $a_0 = 0.01$, $d_0 = 50$ mm^2 (corresponding to about 14 voxels in our data), $r_0 = 0.5$ and $\theta_0 = 0$. Generally we simulate from discretized approximations to the conditional intensities, the details are given below.

Using (8) we find that when ignoring the priors, the Papangelou intensity of a new point ξ given x is

$$\frac{p(Y|x \cup \xi)}{p(Y|x)} = \exp\left\{-\frac{1}{2(\sigma^2/\mathrm{ss}_\varphi + \tau^2)}\left(\sum_{i\in V} h(i;\xi)^2 - 2\sum_{i\in V} h(i;\xi)(\tilde{Y}_i - A_i(x))\right)\right\}. \tag{13}$$

By approximating the discrete sum by an integral, we find,

$$\begin{aligned}\sum_{i\in V} h(i;\xi)^2 &\simeq \iint a^2 \exp\left\{-\frac{2\pi\log 2}{d}\left(\frac{x_\theta^2}{r/(1-r)} + \frac{y_\theta^2}{(1-r)/r}\right)\right\} dxdy/(v_x v_y)\\ &= \iint a^2 \exp\left\{-\frac{2\pi\log 2}{d}(x^2+y^2)\right\} dxdy/(v_x v_y)\\ &= a^2 d/(2\log 2 v_x v_y) = a^2\tilde{d}/(2\log 2),\end{aligned} \tag{14}$$

where v_x and v_y are the length of the voxelsides in mm's and $\tilde{d} = d/(v_x v_y)$ is the area measured in voxels. Above (x_θ, y_θ) represents a translation and rotation of (x, y), and the second equality follows since this transformation together with the coordinate scaling has Jacobian one.

When proposing the position μ we will fix the remaining parameters at $(a_0, d_0, r_0, \theta_0)$ and approximate the Papangelou intensity in (13) with a voxel-wise constant density;

$$q(\mu|x) \propto \exp\left\{\frac{1}{(\sigma^2/\mathrm{ss}_\varphi + \tau^2)}\sum_{i\in V} h(i;\mu, a_0, d_0, r_0, \theta_0)(\tilde{Y}_i - A_i(x))\right\} \quad \text{for } \mu \in V.$$

We will need to calculate a sum over V for all possible values of $\mu \in V$ in order to simulate from this density. Hence an order of $|V|^2$ iterations are required, which can be quite large; in most applications $|V|$ is around 5000. The computational burden can however be reduced, either by only performing the sum over a part of V, where $h(\cdot;\mu, a_0, d_0, r_0, \theta_0)$ is greater than a certain threshold, in which case the number of iterations is $O(|V|)$. Alternatively the convolution can be calculated in the Fourier domain, which requires $O(|V|\log_2|V|)$ iterations when a Fast Fourier Transform algorithm is used, see Press *et al.* (1992).

Considering (13) as a function of the height a, the proposal density is then

$$\begin{aligned}&q(a|\mu, x)\\ &\propto \exp\left\{-\frac{1}{2(\sigma^2/\mathrm{ss}_\varphi + \tau^2)}\left(\frac{a^2 d_0}{2\log 2 v_x v_y} - 2a\sum_{i\in V} h(i;\mu, 1, d_0, r_0, \theta_0)(\tilde{Y}_i - A_i(x))\right)\right\},\end{aligned} \tag{15}$$

which is a Gaussian distribution,

$$a|\mu, x \sim N\left(\frac{\sum_{i\in V} h(i;\mu, 1, d_0, r_0, \theta_0)(\tilde{Y}_i - A_i(x))}{\tilde{d}_0/(2\log 2)}, \frac{\sigma^2/\mathrm{ss}_\varphi + \tau^2}{\tilde{d}_0/(2\log 2)}\right),$$

restricted to the compact interval $(0, C_a]$. As for the three remaining parameters (d, r, θ) we will approximate the conditional intensity with a piecewise log-linear intensity, and sample from the corresponding distribution. When proposing d we will select a grid $(\delta_0, \dots, \delta_m)$ such that $\delta_0 = 0$, $\delta_m = C_d$ and let

$$q(d|\mu, a, x) \propto \exp\left\{p_{i-1} + \frac{p_i - p_{i-1}}{\delta_i - \delta_{i-1}}(d - \delta_{i-1})\right\} \quad \text{for } d \in (\delta_{i-1}, \delta_i],$$

where

$$p_i = -\frac{1}{2(\sigma^2/\mathrm{ss}_\varphi + \tau^2)}\left(\delta_i \frac{a^2}{2\log 2 v_x v_y} - 2\sum_{i \in V} h(i; \mu, a, \delta_i, r_0, \theta_0)(\tilde{Y}_i - A_i(x))\right) \\ - 3\log\delta_i - \beta_d/\delta_i, \quad (16)$$

for $i = 1, \dots, m-1$, $p_0 = p_1$ and $p_m = p_{m-1}$. Above the last two terms stem from the prior for d.

The expressions for $q(r|\mu, a, d, x)$ and $q(\theta|\mu, a, d, r, x)$ are derived similarly.

3.2 Removal of a point

With probability p_2 we propose to remove a point. If the current configuration x is empty we do nothing, otherwise we select the candidate between the points in x with equal probability $1/n(x)$.

3.3 Moving a point

With probability p_3 we propose to change a parameter of a randomly selected point. We choose one of the parameters μ, a, d, r or θ with equal probability and a new value is proposed by considering the conditional distribution of the parameter given the other parameters.

Suppose for instance that a point $\xi = (\mu, a, d, r, \theta) \in x$ has been selected and we wish to propose a new position μ' for ξ. Corresponding to the insertion of a new point above, we will then propose the position by simulating from a distribution which has voxel-wise constant density

$$q(\mu'|x) \propto \exp\left\{\frac{1}{(\sigma^2/\mathrm{ss}_\varphi + \tau^2)} \sum_{i \in V} h(i; \mu', a, d, r, \theta)(\tilde{Y}_i - A_i(x \backslash \xi))\right\}, \quad \mu' \in V.$$

For the parameters r, d and θ we consider a neighbourhood of the current value, and approximate the conditional density as in (16) above. In our application, we have chosen a neighbourhood of 100 mm^2 for d, 0.3 for r and 0.35 for θ.

Finally, the height a is simulated from a normal distribution as when proposing a new point,

$$a|x \sim N\left(\frac{\sum_{i \in V} h(i; \mu, 1, d, r, \theta)(\tilde{Y}_i - A_i(x \backslash \xi))}{\tilde{d}/(2\log 2)}, \frac{\sigma^2/\mathrm{ss}_\varphi + \tau^2}{\tilde{d}/(2\log 2)}\right).$$

4 Simulation study

Estimation of an activation surface can be carried out by simulating from the posterior distribution of the surface given the data. However, we are left with the problem of determining sensible values for the parameters of the prior for X, sensible in the sense that the estimated activation surface corresponds well with the underlying true surface. To this end we have generated a training data set by simulating from the model (7), with a known underlying activation pattern A. An image of the latter can be seen in Figure 3. The image was generated to mimic a "true" activation image, with coherent regions of activation of both small and moderate sizes. In each region the activation level in individual voxels were simulated from a normal distribution with a common mean. Finally the image was smoothed with a Gaussian kernel to obtain a smooth activation image. Naturally an image obtained in this way cannot be reproduced exactly by a single realization of the activation pattern of model, in this sense the training data is not different from real fMRI data sets. However by using the posterior mean of the activation pattern as an estimate of the latter, we can reproduce more general patterns than those represented by the prior model.

We will fix the parameters of the priors for d, a and r at the values given earlier. Hence we are left with the intensity β, the scaling parameter ρ and the order of the soft-core prior p. Since we need to perform an entire run of the MCMC algorithm for each combination of parameter values, it is only possible to perform a crude estimation where a few different values of each parameter are tried. For each set of parameters we produced 400000 samples from the MCMC algorithm and stored every 100'th sample. After an initial burn-in of 500 subsamples, the chain was judged to be stationary from plots of diagnostics of the simulated point patterns (not shown). We will measure the goodness-of-fit of the model by the posterior mean of the L^2 distance between the activation surface and the true surface. We will estimate this quantity by

$$\widehat{GOF} = \frac{1}{N}\sum_{j=1}^{N}\left\{\sum_{i\in V}(A_i(x^{(j)}) - A_i)^2\right\}^{1/2},$$

where $x^{(1)},\ldots,x^{(N)}$ is a sequence of simulations from the MCMC algorithm. The variance of this estimate was estimated by the method of batch means with batch sizes of 25. The variance estimate gives an idea of the level of uncertainty, but as it depends on the chosen batch size, it should be interpreted with care.

The true values for the standard deviations were respectively $\sigma = 0.03$ and $\tau = 0.005$. The estimates obtained by (10) and (11) were $\hat{\sigma} = 0.02996$ and $\hat{\tau} = 0.005074$. Table 4 shows the goodness-of-fit of the model with different parameters values. The model which yields the best fit is the one with $\beta = 0.01$, $\rho = 5$ and $p = 10$. Though the largest changes in the GOF measure occur when varying β, the four rows of $\beta = 10^{-4}$ indicates that some degree of regularity ($\rho > 0$) improves the goodness-of-fit. In the last column of the table is an estimate of the mean integrated activation, that is the integral of the activation surface. This can be considered as a summary statistic of the total level of activation.

Evidently the choice of prior parameters affects the final result to some extent. Usually, we would prefer to estimate the parameters in an empirical Bayes fashion, however as the maximum likelihood estimates can in general correspond to a prior that favours meaningless point configurations, cf. the discussion in Section 2.4, this is not an advisable strategy. The fully Bayesian approach, with hyperpriors on the parameters, is an alternative. However, it is not obvious how one should simulate the posterior distribution of the parameters, since the unknown normalization constant of the point process density would enter in the Metropolis-Hastings ratio. Instead we will fix the prior parameters at the values which yield the best fit in the simulation study. Though the result will to some extent depend on this choice, we note from the table that statistics of interst, such as the mean integrated activation, varies only litte, and in no systematic way, with the parameters.

β	ρ	p	$\widehat{GOF}$ (s.e.$\times 10^4$)	Int. act. (s.e.)
10^{-4}	5	2	0.1342 (2.22)	4.89 (0.029)
10^{-4}	20	2	0.1345 (2.59)	4.43 (0.019)
10^{-4}	5	10	0.1357 (2.16)	4.63 (0.022)
10^{-4}	20	10	0.1328 (1.97)	4.57 (0.018)
10^{-4}	0	-	0.1355 (3.01)	4.68 (0.022)
10^{-2}	5	10	0.1305 (2.87)	4.86 (0.026)
10^{-6}	5	10	0.1428 (2.00)	4.83 (0.025)

Table 1: Estimates of the goodness-of-fit of the model with different parameter values. Standard errors due to the simulation are given in parentheses. In the row with $\rho = 0$ no interaction between the points was included in the model. In the last column is an estimate of the mean integrated activation.

In Figure 3 is the estimate of the posterior mean activation image under the best model above. For comparison, the smoothed SPM estimate of the activation is also displayed in the figure. This is obtained by smoothing the regression image $\{\tilde{Y}_i\}$ with a Gaussian kernel of FWHM 3 voxels. The latter denotes the full width of half maximum of the smoothing kernel, this is the typical measure for the width of a smoothing kernel in the medical imaging literature. Both estimates tend to oversmooth the true image, due to the smoothness assumptions underlying them both, but the bias is largest for the SPM. This is more clearly seen in the plot in Figure 4 which shows the number of voxels with an activation level higher than a given threshold for the true image, the posterior mean activation image and the smoothed SPM. One cannot reduce the oversmoothing of the SPM by reducing the width of the smoothing kernel, since the theory of Gaussian random fields, used for making inference in the SPM, requires the discrete image to be a reasonable approximation to a differentiable spatial process.

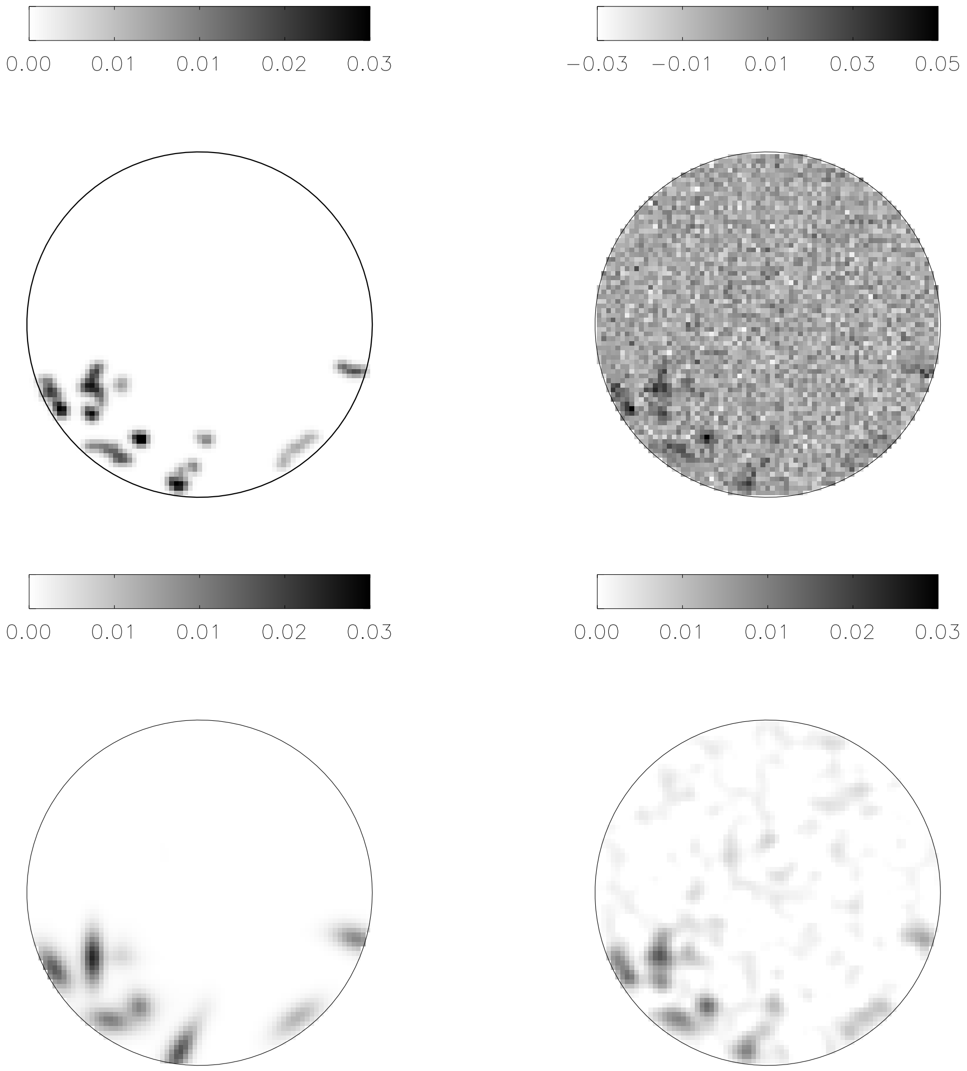


Figure 3: Top Left: The artificial activation image used for generating training data. Intensity values range from 0.0 to 0.04, but the image is clipped at 0.03 for display purposes. Top Right: The regression image $\{\tilde{Y}_i\}$ obtained from the training data. Bottom left: The estimate of the mean posterior activation with $\beta = 0.01$, $\rho = 5$, and $p = 10$. Bottom right: The regression image smoothed by a kernel of FWHM 3 voxels. Note that the color scale differs in the upper right image compared to the three others.

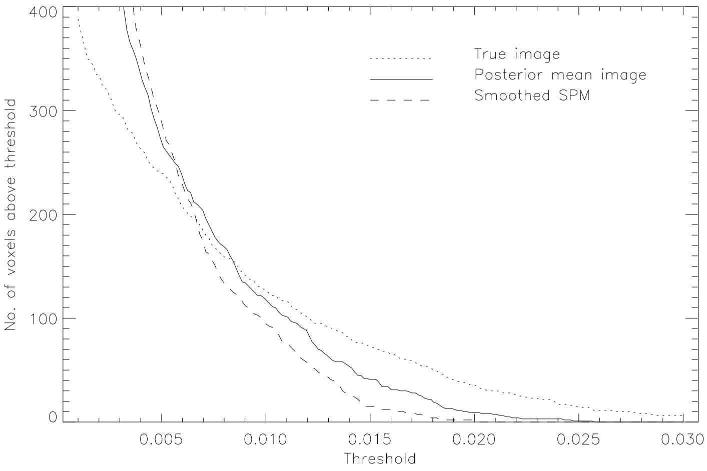


Figure 4: The number of voxels with activation level exceeding a given threshold. Shown is respectively the true image, the posterior mean activation image and the smoothed SPM.

4.1 An analysis of a visual stimulation dataset

We will apply the method to an fMRI dataset consisting of 90 scans, acquired while a light was periodicly flashed in the eye of the subject. The scans, which was obtained by a method denoted Echo-Planar Imaging (EPI), was recorded every 2 seconds during a 3 minute period. The stimulation was arranged in blocks of 20 seconds off, 20 second on, 20 seconds off etc., with 4 complete on-off cycles during the session. Each scans consists of 128 by 128 voxels each covering an area of 1.875 × 1.875 mm in a slice of thickness 5 mm.

In general the magnetization of the tissue will be highest in the initial scans, causing an increased intensity in the beginning of the time series. After a couple of scans an equilibrium is obtained, and the intensity stabilizes at a steady level, and we will hence discard the first 5 scans and only consider the remaining 85 in the analysis.

The variance estimates were $\hat{\sigma} = 0.0294$ and $\hat{\tau} = 0.00421$. We generated 1 million simulations from the MCMC algorithm and subsampled every 100'th sample. In Figure 5 is a plot of two diagnostics of the simulated point process, namely the number of points and the L^2 norm of the residual image, $\{\tilde{Y}_i - A_i(x)\}$. As can be seen from the plots there is an initial burn-in period of about 2000 subsamples, after which the chain stabilizes to a stationary level. However, as the auto-correlation plot for the number of points shows, the samples are somewhat correlated, and it would be worthwhile to improve the mixing properties of the algorithm to speed up convergence. The acceptance probabilities for the different movetypes are listed in Table 2.

In Figure 6 is a plot of the posterior mean activation image and the posterior standard deviation of the activation, calculated voxelwise. The back of the head is in the top of the images. It is evident from the images, that there are large areas of activation in the back of the brain, which corresponds to the location of the visual cortex, that processes visual

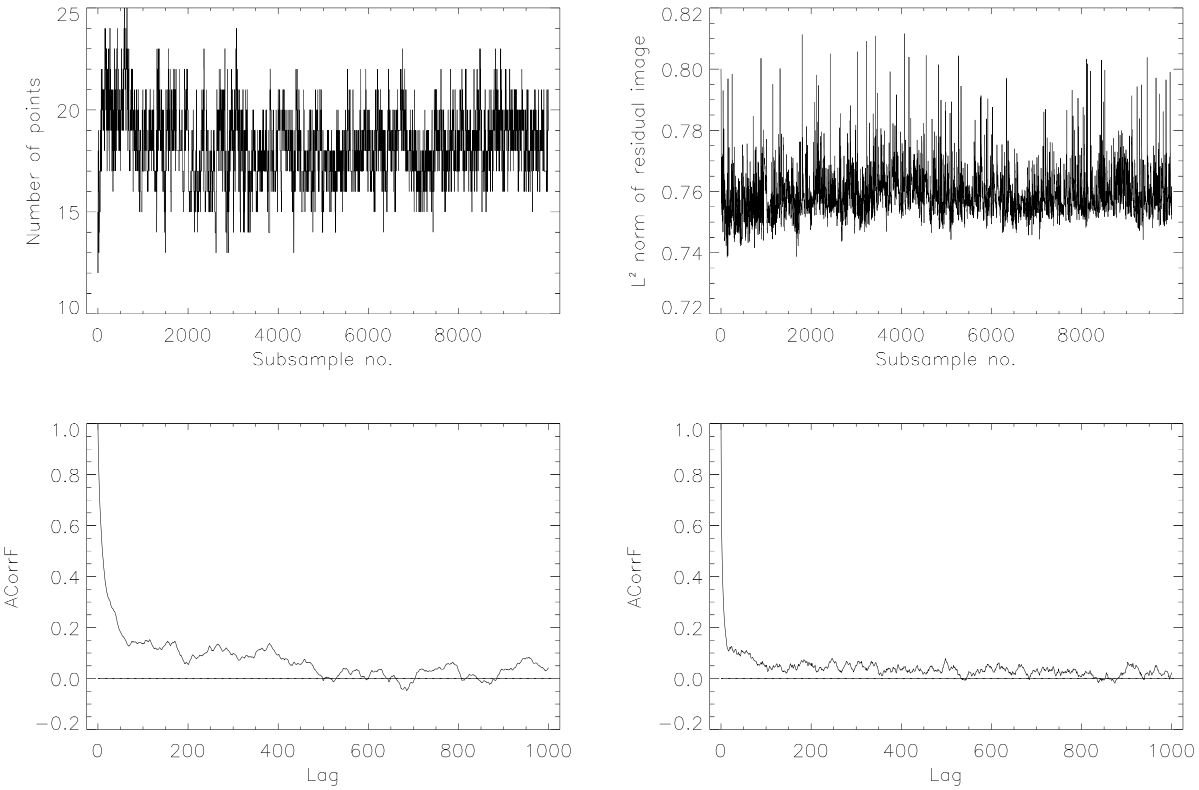


Figure 5: Diagnostics plots of the simulations obtained by subsampling every 100'th iteration of the MCMC algorithm. Shown is to the left the number of points in X and to the right the L^2 norm of the residual image, $\{Y_i - A_i(X)\}$. Below are auto-correlation plots of the two timeseries.

Move type	Acceptance (%) Independent noise	Acceptance (%) Correlated noise
Insert point	5.51	4.72
Delete point	5.53	4.75
Update position	15.78	11.35
Update height	43.92	25.98
Update area	35.63	23.76
Update angle	66.86	52.10
Update ratio	59.84	45.43

Table 2: Acceptance probabilities for the different move types in the MCMC algorithm. The correlated noise model will be described in Section 6.

impressions. For comparison is also the smoothed SPM, which appears to oversmooth the image, also in this example.

The posterior variance image gives some idea of the uncertainty of the activation estimate. The area which has large posterior variance turned out to posses some time series which were more noisy than the remaining ones, this lack of fit of the model is hence reflected in a larger uncertainty of the estimate in this area. Alternatively the uncertainty can be quantified by the posterior probabilities of individual voxels having activation level greater than a certain threshold, 0.009 say. The latter corresponds to the standard error of usual voxel-wise regression estimates for the activation level. The posterior probabilities are displayed as an image in the figure.

Often the interest is on a particular summary statistic, such as the activation area, measured in terms of number of activated voxels. For the current data set the estimate of the mean activation area is 491.0 voxels, and the standard deviation of the area is estimated to 26.0. More specific hypotheses about the activation pattern may be evaluated by estimating posterior probabilities of events of interest.

Finally we will estimate the shape of the response function, given the estimated activation surface. Recall that $Y_{it} = A_i\varphi_t + \varepsilon_{it}$, hence for known A the m.l.e. of φ_t is given by $\hat{\varphi}_t = \sum_i A_i Y_{it} / \sum_i A_i^2$ for $t = 1, \ldots, m$. By inserting the estimate of the posterior mean activation surface displayed in Figure 6, we get the estimate of φ plotted in Figure 7. Overlaid on the plot is the model response function given in (6). As can be seen from this plot, there are substantial differences between the model, and the observered response. The observed response does not appear to be stationary, for instance the last peak is higher than the 3 first, and the dip below baseline is more prominent after the first and third cycle than after the second. In the next section we will describe a method for modelling such non-stationarities in a semi-parametric setting.

5 A semi-parametric model for the haemodynamic response

Though the model for the haemodynamic response (6) is verified empirically to give a reasonable fit to the observed response, its limitations were demonstrated in the previous section. Several hypothesis have been proposed to explain the complex interplay between the local blood flow and oxygenation changes and the BOLD signal, yet there is still not concensus of the quantitative relationship between these. Clearly the choice of a convolution model with a Gaussian impulse response function is somewhat *ad hoc* in this context. Alternative models for the impulse response has been proposed, such as Gamma densities by Lange and Zeger (1997), Friston *et al.* (1998) and FIR filters by Nielsen *et al.* (1997). The question of whether a convolution model is appropriate is however not clear; in some circumstances the response is approximately linear (Dale and Buckner, 1997) while in others it is highly non-linear (Vazquez and Noll, 1998). Also a relevant question is whether the response is stationary over time, or if the response changes with general alertness and learning as suggested

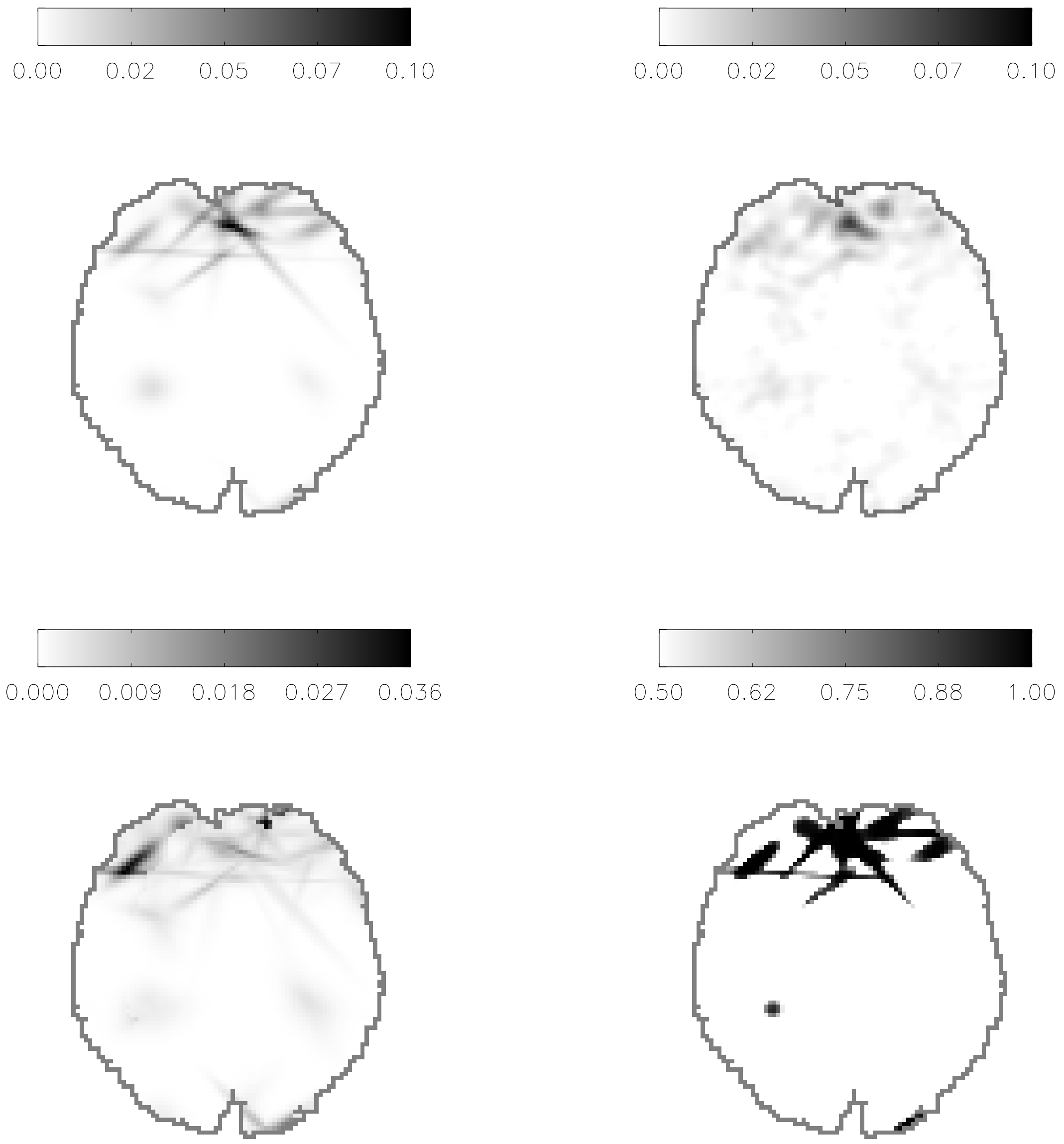


Figure 6: Top left: Monte Carlo estimate of posterior mean activation surface. Top right: Smoothed SPM estimate of activation surface. Bottom left: Estimate of voxelwise posterior standard deviation. Bottom right: Voxelwise posterior probability of activation level greater than 0.009. The images represent a slice of the brain, the upper part of the images correspond to the back. The grey line represents the surface of the cortex.

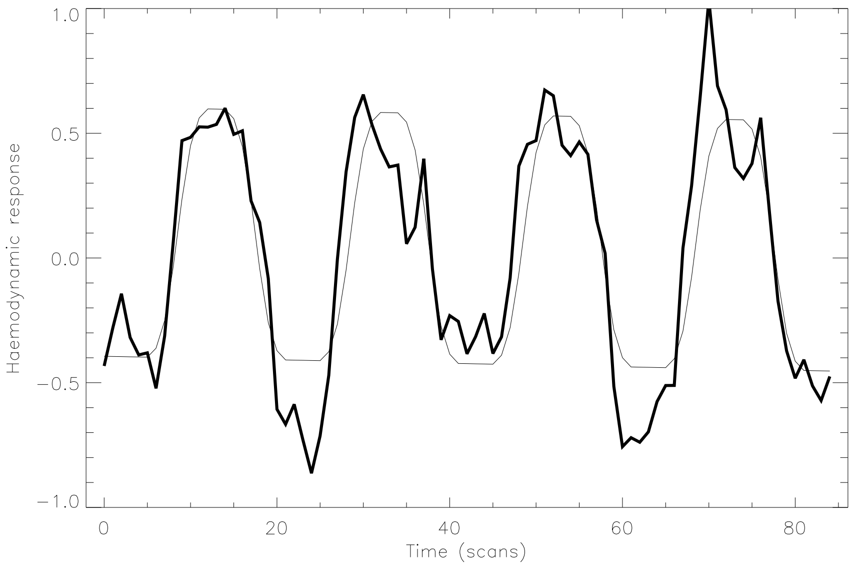


Figure 7: Thick line: Maximum likelihood estimate of the haemodynamic response function under the assumption that the spatial activation pattern is known and given by the estimate in Figure 6. Thin line: The model for the haemodynamic response in (6).

by Gaschler-Markefski *et al.* (1997).

Considering the complexity of temporal response, it seems very appealing to model the latter in a semi-parametric setting. In this framework we do not have to assume stationarity over time or additivity of the response. Instead we will assume a prior of the form

$$\varphi_t = \lambda_t + \nu_t, \quad \nu_t - \nu_{t-1} \sim N(0, \kappa^2), \quad t = 1, 2, \ldots, m. \tag{17}$$

where $\{\nu_t - \nu_{t-1}\}$ are independent and $\nu_0 = 0$. Here the mean λ_t is a simple model, such as that in (6) used in the previous section, reflecting the overall structure of the response. However φ is allowed to deviate a lot from the mean, via the random walk structure of the noise terms ν_t. The variance κ^2 governs the smoothness of $\varphi_t - \lambda_t$.

Combining this prior for φ with the prior for spatial activation pattern (2) we can make inference about (X, φ) through the simultaneous posterior distribution $P(X, \varphi | Y)$. For computational reasons we will in fact consider the posterior distribution of (X, Y, η) where $\eta = \{\eta_i, i \in V\}$ are the random intercepts in the model (7). This posterior is given by

$$p(X, \varphi, \eta | Y) \propto P(Y | X, \varphi, \eta) P(X) P(\varphi) P(\eta),$$

where the likelihood term is obtained by conditioning on η in (7). The variable η can be considered as an auxiliary variable in the simulation algorithm, since it is not of interest in itself, but simulation of the two other variables X and φ becomes much easier, when we condition on η.

We can generate a Markov chain which has the posterior as invariant distribution by a variable-at-a-time Metropolis-Hastings algorithm, where we iteratively update one parameter given the two others. When updating X, the proposals are as described earlier, though with the modification that we replace $A_i(x)$ with $A_i(x)+\eta_i$ and set $\tau^2=0$ in the formulas in Section 3, in order to account for the fact that we condition on η. A similar modification applies to the likelihood function in (8), when calculating the acceptance ratio.

When updating η, we will simulate directly from the conditional distribution given (Y,X,φ). This is hence a Gibbs update, which will always yield acceptance rates of 1. It can easily be verified that

$$\eta_i|Y,X,\varphi \sim N\left(\frac{\tau^2}{\tau^2+\sigma^2/\mathrm{ss}_\varphi}(\tilde{Y}_i - A_i(X)), \tau^2(1-\frac{\tau^2}{\tau^2+\sigma^2/\mathrm{ss}_\varphi})\right), \tag{18}$$

with all η_i's conditionally independent.

We will simulate directly from the conditional distribution of φ given (Y,X,η), as well, since this is also normal, where mean and variance can be calculated as follows. Let $Y_{\star t} = (Y_{it})_{i\in V}$ denote the image recorded at time t, regarded as a $|V|$-dimensional vector, and let $\varepsilon_{\star t}$ be defined correspondingly. In the following, all distributions are conditionally on $A=A(X)$ and η. Then the model (7) states that

$$Y_{\star t} = (A+\eta)\varphi_t + \varepsilon_{\star t}, \quad \varepsilon_{\star t} \sim N(0,\sigma^2 I_{|V|}), \quad t=1,\dots,m,$$

which combined with the prior (17) is a linear Gaussian state space model. Hence it is not difficult to see that if we condition on Y, φ has a Gaussian distribution. The literature on state space models is extensive, hence we will just give the formulas for the conditional mean and variance of φ as given by the Kalman smoother, and refer to West and Harrison (1989), for instance, for the proofs.

Let $D_t = \sigma\{Y_{\star 1},\dots,Y_{\star t}\}$ denote the information up to time t and suppose that $\varphi_{t-1}|D_{t-1} \sim N(\mu_{t-1}, C_{t-1})$. This is true for $t=1$ when we consider the initial distribution of φ_0 as a degenerate normal distribution concentrated at 0. The Kalman filter then gives that φ_t given D_t is also normal, $\varphi_t|D_t \sim N(\mu_t, C_t)$, where

$$\begin{aligned}
C_t^{-1} &= \|A+\eta\|^2/\sigma^2 + (C_{t-1}+\kappa^2)^{-1},\\
\mu_t &= \mu_{t-1} + \lambda_t - \lambda_{t-1} + \frac{C_t}{\sigma^2}(A+\eta)'(Y_{\star t} - (A+\eta)(\mu_{t-1}+\lambda_t-\lambda_{t-1})).
\end{aligned}$$

Here prime denotes the transpose matrix.

In order to simulate from the distribution of φ_t given $Y=D_m$, we will also consider the Kalman smoother. Suppose that $\varphi_{t+1}|\varphi_{t+2}, D_m \sim N(\bar{\mu}_{t+1}, \bar{C}_{t+1})$. By the recursion above, this is true for $t+1=m$ with $\bar{\mu}_m = \mu_m$ and $\bar{C}_m = C_m$. Then we have that $\varphi_t|\varphi_{t+1}, D_m \sim N(\bar{\mu}_t, \bar{C}_t)$, where

$$\bar{C}_t = C_t - \frac{C_t^2}{C_t+\kappa^2}, \quad \bar{\mu}_t = \mu_t + \frac{C_t}{C_t+\kappa^2}(\varphi_{t+1} - \mu_t - \lambda_{t+1} + \lambda_t). \tag{19}$$

In fact this is the conditional distribution of φ_t given $\varphi_{t+1}, \dots, \varphi_m, D_m$, by conditional independence of φ_t and $\varphi_{t+2}, \dots, \varphi_m$, and hence we can use this recursion to simulate $\varphi | D_m$: We simply simulate the φ_t's one at a time, starting from the back with φ_m.

We note here, that a collection of response functions could be modelled by allowing a multidimensional φ. We could assign different functions to different groups of centres, and in this way account for regional differences in the response. The formulas above would be slightly more complicated, but at least for moderate dimensions of φ the recursive simulation routine would still be very efficient.

In Figure 8 is a plot based 1000000 simulations of this Markov chain. We considered the visual stimulation data of the previous section, though preprocessed in a slightly different way, as we removed some low-frequency trends with very large magnitude from the data, in order to stabilize the algorithm. The plot illustrates the estimated posterior mean $E(\varphi|Y)$ with confidence limits for φ based on the posterior variance. For comparison is an overlay of the initial model (6). The plot shows the same deviations from stationarity as was indicated by Figure 7.

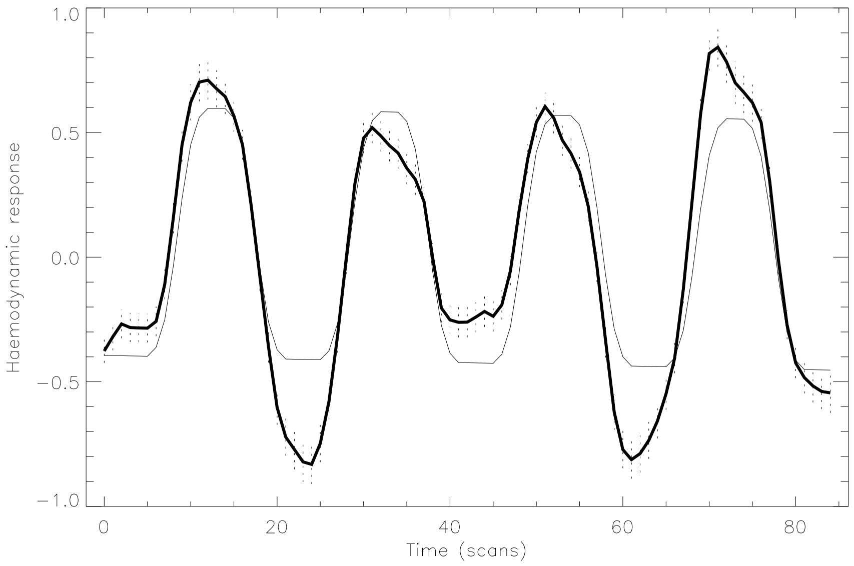


Figure 8: Monte Carlo estimate of the posterior mean of the haemodynamic response function based on 10000 subsamples of 1000000 simulations. Overlaid is pointwise 95%-confidence regions based on the estimated posterior variance. The thin line is the prior mean of the response given by the convolution model in (6).

One consequence of modelling φ in this way, is that the paradigm is only vaguely included in the model, in the sense that the response function is not time-locked to the paradigm, but is allowed to drift by the random walk structure. It might seem unwise to ignore a relevant covariate like this, however, in some experiments the *actual* paradigm is not directly controllable by the experimenter and hence precise information of this is not available. For instance in memory processing or other mental stimulation experiments, it is not possible to end the stimulation at an exact time point. Furthermore with this formulation, we may detect subtle activation patterns, which depends on the paradigm in more complex ways. An example of the latter is the XOR signal of Lange *et al.* (1999).

6 Accounting for correlated noise

The initial model in (7) assumed that the noise was uncorrelated both temporally and spatially. This is necessarily a somewhat optimistic assumption. The noise sources in fMRI data are both of physiological and physical origin. The pixel values are constructed by inverse Fourier transforms of a sequence of mesurements of currents in a coil over a short time period. Hence there is no physical separation of the pixels, which could justify independence. The temporal correlation is likely to arise from physiological sources, but also intrinsicly in the MR scanner.

In order to investigate the correlation of the noise, we will consider the residuals in the model (7), $r_{it} = Y_{it} - \tilde{Y}_i \varphi_t$, where we assume the response function φ is known. The empirical temporal and spatial correlograms are respectively,

$$\hat{\gamma}_i(l) = \frac{\sum_{t=1}^{m-l} r_{i,t} r_{i,t+l}}{\sum_{t=1}^{m} r_{i,t}^2} \quad l = 1, 2, \ldots, m-1, \ i \in V,$$
$$\hat{\lambda}_t(k) = \frac{\sum_{i \in V_k} r_{i,t} r_{i+k,t}}{\sum_{i \in V} r_{i,t}^2}, \quad k \in \mathbb{Z}^2, t = 1, \ldots, m,$$

where $V_k = \{i \in V | i + k \in V\}$. We estimate the correlograms voxel-by-voxel respectively scan-by-scan in order to assess whether the correlation is stationary over voxels and scans. In Figure 9 is a plot of $\hat{\gamma}_i(1)$ as a function of $i \in V$ with estimated 95%-confidence bounds based on a global AR(1) model $\gamma_i(1) = \gamma(1)$ and $\gamma_i(l) = 0$ for $l > 1$ for all i. Displayed is also a plot of $\lambda_t((1,0))$ as a function of t with estimated 95%-confidence bounds based on the spatial model described below. About 10% of the points fall outside the confidence bounds in each plot, which indicates that the temporal (spatial) correlation structure is not the same in all voxels (scans). This is not really surprising, as it merely reflects the inhomogeneity of the underlying tissue. The observation suggests that a non-separable covariance model, which allows for the different temporal structures, should be fitted to the data. One such model is that proposed by Lange and Zeger (1997), where the voxel time series are considered in the frequency domain, and different spatial covariance models are fitted to different frequencies. However, since we need to invert (or Cholesky decompose) the large spatio-temporal covariance matrix in order to calculate the likelihood function, we will have to restrict ourselves to reasonably simple covariance structures. For this reason we will only consider a separable model in the following sense. Let $\varepsilon = \{\varepsilon_{it}, i \in V, t = 1, \ldots, m\}$ be the noise terms in (7) regarded as a $|V| \times m$ matrix, then

$$\varepsilon \sim N_{|V| \times m}(0, \sigma^2 \Gamma \otimes \Lambda).$$

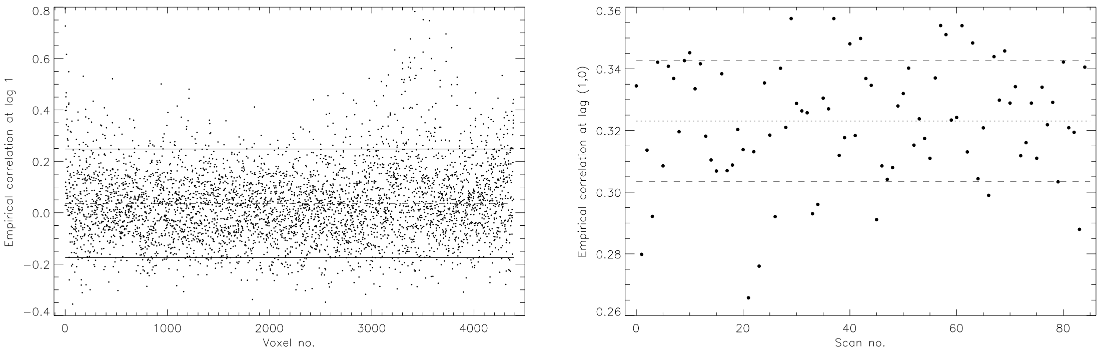


Figure 9: Left: The empirical temporal correlations at lag 1, $\hat{\gamma}_i(1)$, as a function of voxel number. The 95%-confidence bounds are based on a common AR(1) model for all voxels. Right: The empirical spatial correlation at lag (1,0), $\hat{\lambda}_t((1,0))$, as a function of scan number. The 95%-confidence bounds are based on a common spatial model for all scans (see text.)

where $\otimes$ denotes the Kronecker product and where Γ and Λ are $|V| \times |V|$ and $m \times m$ correlation matrices.

Global correlation estimates are obtained by the averages $\hat{\gamma}(l) = |V|^{-1} \sum_{i \in V} \hat{\gamma}_i(l)$ and $\hat{\lambda}(k) = m^{-1} \sum_{t=1}^{m} \hat{\lambda}_t(k)$. A plot of the empirical temporal correlations can be seen in Figure 10. The fitted AR(1) model with $\hat{\gamma}(1) = 0.0367$ gives a reasonable fit to the observed correlations.

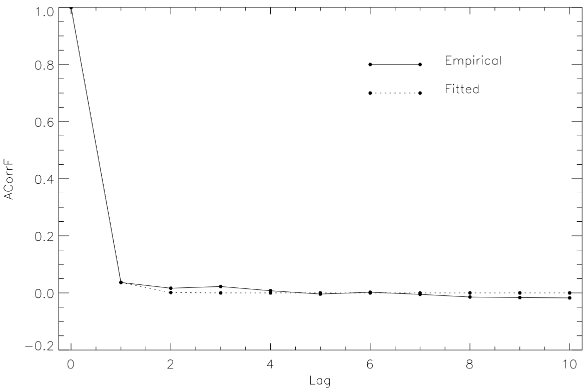


Figure 10: The empirical temporal correlogram for all voxels with the fitted AR(1) correlogram.

A plot of the empirical spatial correlation can be seen in Figure 11. The correlation is clearly non-isotropic and furthermore there is evidence of negative correlation at lag 2 voxels. The plot indicates that observations further than a distance of 2 voxels apart are almost uncorrelated, which suggests a moving average type model,

$$\varepsilon_i = \sum_{j \in D} g_j U_{i+j}, \quad i \in V,$$

where $\{U_j, j \in \mathbb{Z}^2\}$ is white noise and D is some neighbourhood of the origin. Here and in the following we will only consider the spatial covariance structure of a single scan and hence ignore the temporal index t in the notation. The parameters $g_j, j \in D$ can be estimated by fitting the model to the empirical covariances and an estimate of the spatial correlation matrix Γ may be calculated. The problem with this approach, however, is that one needs to invert Γ, or at least compute the Cholesky decomposition $\Gamma = LL'$ where L is lower triangular, in order to calculate the likelihood function. In our data set there are more than 4000 voxels constituting V which makes it very demanding to decompose the correlation matrix. As a practical alternative to the above model, we propose to parametrize the Cholesky square root L rather than Γ itself, and hence consider the model $\varepsilon \sim N_{|V|}(0, \sigma^2 LL')$ where L is parametrized as follows. Let $\tilde{L} = \{\tilde{l}_{ij}\}$ be a lower triangular matrix, such that

$$\tilde{l}_{ij} = \begin{cases} \tilde{l}_{i-j} & \text{if } i > j,\ i - j \in D, \\ 1 & \text{if } i = j, \\ 0 & \text{else,} \end{cases}$$

where D is a neighbourhood of the origin, not necessarily equal to the neighbourhood in the moving average model above. Then we let $L = \{l_{ij}\}$ be given by

$$l_{ij} = \frac{\tilde{l}_{ij}}{\left(\sum_{j \le i} \tilde{l}_{ij}^2\right)^{1/2}} \quad i, j \in V$$

The normalization above ensures that LL' is correlation matrix. Notice that when ignoring edge-effects, the model will be stationary since $L_{ij} = L_{i-j}$.

In this formulation Γ can be calculated by a matrix product, and expressions such as $z'\Gamma^{-1}z$ for $z \in \mathbb{R}^{|V|}$, which enters in the likelihood function, can be calculated by

$$z'\Gamma^{-1}z = \|L^{-1}z\|^2 = \|v\|^2, \tag{20}$$

where v is the solution to $Lv = z$ which can be obtained easily due to the lower-triangularity of L. Notice that we have to order the indices in $|V|$ when expressing the spatial correlation as the matrix Γ. The correlation between ε_i and ε_j is given by

$$\operatorname{corr}(\varepsilon_i, \varepsilon_j) = \sum_{k \le \min(i,j)} l_{i-k} l_{j-k}. \tag{21}$$

Hence the model has the peculiar property, that the covariance structure depends on the ordering of the indices. From a theoretical point of view this is difficult to accept, since the ordering is arbitrarily chosen. From a practical point of view, however, the ordering is chosen in any natural way, and the model is judged by how well it fits data. We will demonstrate in a moment that the model fits data well, and since the computational advantages by working with L rather than Γ or Γ^{-1} are considerable, we favour this method.

The parameters may be estimated by fitting the implied correlation (21) to the empirical covariance,

$$\hat{l} = \text{argmin} \sum_{\{j \in \mathbb{Z}^2 : D \cap (D-j) \neq \emptyset\}} \left(\sum_{k \leq \min(0,j)} l_{-k} l_{j-k} - \hat{\gamma}(j) \right)^2 .$$

We have chosen the natural lexicographic ordering of the voxel indices (x, y) and have parametrized the model by letting D include 3. order neighbours, which gives 6 free parameters. In Figure 11 is a plot of the empirical correlation along the 4 equiangular directions $(0, \pi/4, \pi/2, \pi)$, together with the fitted correlation. Clearly the model fits the data quite well.

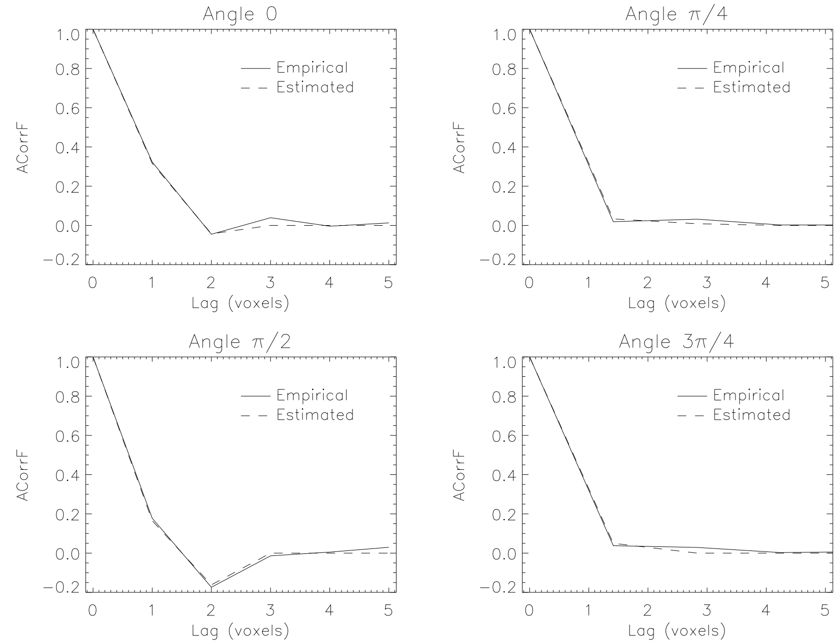


Figure 11: Empirical and fitted spatial correlations along the four equiangular directions $(0, \pi/4, \pi/2, 3\pi/4)$

Incorporating estimates of Γ and Λ in the model (7) is quite straightforward if we assume that these are reasonable precise estimates, and hence can be regarded as fixed. Consider the data Y as a $|V| \times m$ matrix, and let $Y^\circ = Y(M^{-1})'$, and $\varphi^\circ = M^{-1}\varphi$, where M is the lower triangular Cholesky square root of Λ. Then the conditional likelihood function, where we condition on the values of η, is given by

$$p(Y|x,\varphi,\eta) = (2\pi\sigma^2)^{-\frac{m|V|}{2}}|\Gamma|^{-\frac{m}{2}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{t=1}^{m}\|L^{-1}(Y^\circ_{\star t} - \tilde{Y}^\circ\varphi^\circ_t)\|^2\right\} \\ \times \exp\left\{-\frac{1}{2\sigma^2/\mathrm{ss}^\circ_\varphi}\|L^{-1}(\tilde{Y}^\circ - A - \eta)\|^2\right\}, \quad (22)$$

where

$$\tilde{Y}^\circ_i = \sum_{t=1}^{m} Y^\circ_{it}\varphi^\circ_t/\mathrm{ss}^\circ_\varphi, \quad \mathrm{ss}^\circ_\varphi = \sum_{t=1}^{m}\varphi^{\circ 2}_t,$$

is defined equivalently to $\tilde{Y}$ in (9). Recall from Section 3 and 5 that the proposal distributions for updating X and φ in the MCMC algorithm were based on the model with independent noise. A big advantage from an application point of view is that these need not be changed when we incorporate correlated noise terms, if we are willing to accept slightly worse mixing properties. By simply substituting the expressions for the likelihood ratio in the Metropolis-Hastings ratio, we ensure that the chain converges to the correct posterior distribution. Inference on X can hence be made by simulating (X, η) iteratively, where the distribution of $\eta|X, Y, \varphi$ is as in (18).

We considered the visual stimulation data again, and simulated 1000000 samples of X using the same MCMC algorithm as described in Section 3, but with the modified likelihood function. As expected, the acceptance rates decreased a bit, see Table 2. The effect of accounting for correlation in the noise shows both in the activity estimate itself and in an increased uncertainty of the latter. The mean activation image is illustrated in Figure 12. The largest difference, compared to the activation image obtained with the uncorrelated noise model in Figure 6, is the circular region in the back of the brain, which is much larger in the current image. As an example of how the variance of the estimate increases, we may consider the number of activated voxels. The mean and standard deviation of this are estimated to 543.1 respectively 31.4, the corresponding figures from the uncorrelated model in Section 4.1 were 491.0 and 26.0.

7 Discussion

We have proposed a spatio-temporal model for fMRI data which explicitly accounts for the fact that the signal changes are locally coherent in both space and time. This assumption is often implicitly included in the analysis of fMRI data, when spatial and temporal filtering are applied prior to the analysis, but rarely included explicitly in a model. The relation (8) shows that in the simplest setting the procedure is effectively fitting ellipsoids of different sizes and orientations to a regression image, and assessing the significance of these. The random field theory has counterparts to this procedure, namely the search for local maxima in both scale and space, Siegmund and Worsley (1995), and in the space of ellipses with different orientation and shape, Shafie *et al.* (1998). The method is, however, fundamentally

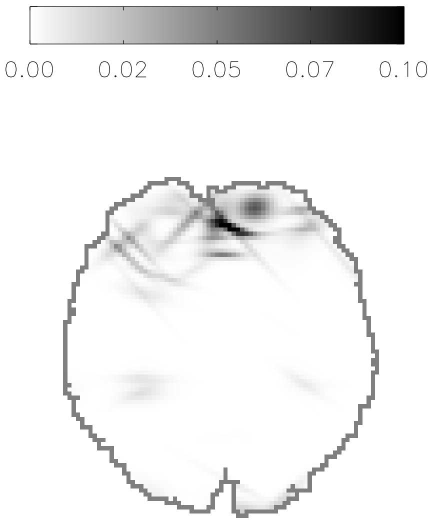


Figure 12: The estimate of the posterior mean activation image in the correlated noise model.

different from the random field approach. The latter provides a framework for testing the null-hypothesis of no activation in each individual voxel with correction for the large number of tests performed. As was pointed out by Keith Worsley in the discussion of Lange and Zeger (1997), what is really an estimation problem is hence answered by a large number of statistical tests, with corresponding conceptual and mathematical problems. With the proposed method the focus is shifted towards estimating the activation pattern by use of standard Bayesian methods, rather than testing for activation in individual voxels.

Assessing the uncertaincy of the estimated activated pattern is theoretically easy by considering posterior variances. This allows us to evaluate the significance of hypothesis of interest within single subjects. Alternatively we may record estimates and standard errors of relevant features of the activation in different experiments, and use this when comparing different groups of subjects. This might for instance be the mean activity level in a certain region of the brain, where the latter could be identified individually in each subject from high-resolution anatomical scans acquired simultaneously with the fMRI scans.

The approach of modelling the temporal response in a non-parametric setting with few assumptions seems appealing to us, given the uncertainty about the nature of the haemodynamic effects in different stimulation types. Also the fact that the modelled response depends only vaquely on the specified paradigm is an advantage when analysing data where the actual paradigm is difficult to determine. Naturally the method has it's limitations. Firstly the prior we have formulated, restricts the response to be sufficiently smooth, and one could imagine that this is not the case in the recent event-related pardigms Buckner (1998), where several stimulation types are rapidly interchanged. To analyse these data with our method, an alternative prior should be formulated, possibly by incorporating the paradigm and assuming some sort of stationarity.

Another limitation is the assumption, that the response is the same everywhere in the brain. Authors such as Lee *et al.* (1995) and Kornak *et al.* (1999) have found, by fitting simple parametric response functions to fMRI time series, that the delay can vary with a few seconds over the activated regions. Though the semi-parametric model is limited by the assumption of constant delays, it is advantageous in the sense that it can capture more general response patterns than those proposed by these authors. An obvious way of relaxing the assumption of constant delay and shape is by working with a collection of response functions, and assigning different functions to different centres, hence we would only assume that the response is locally similar. In this formulation we would in fact search for any spatial regions of similar temporal pattern, and not only paradigm related patterns.

Acknowledgements

The author is grateful for inspiring and constructive discussions with Jens Ledet Jensen and Rasmus Waagepetersen. Thanks also to Hans Stødkilde-Jørgensen from the MR-ResearchCentre at Skejby Sygehus for kindly providing the data.

A Ergodicity properties of the Markov chain

Theoretical results for the algorithm described in section 3 are studied in Geyer and Møller (1994), Møller (1999) and Geyer (1999). Geyer has a stability condition, namely that the point process has bounded conditional intensities, which implies Harris recurrence and geometric ergodicity of the algorithm, that he studies. This situation is, however, slightly different, since our proposal density for inserting a new point $q_2(\xi \cup x|x)$ is not constant, as in Geyer's algorithm, but has a rather complicated structure. However, when restricting the support of the prior in a natural way $q_2(\xi \cup x|x)$ is bounded below, which turns out to be sufficient to apply Geyer's method.

Proposition 1 *There exists a constant M such that*

$$p(x \cup \xi|Y) \leq Mp(x|Y) \quad \forall x \in \Omega, \xi \in \mathcal{X}.$$

Proof Recall that $\phi(x_i, x_j)$ given by (3) is less than 1, hence we have,

$$\frac{p(x \cup \xi|Y)}{p(x|Y)} = \frac{p(x \cup \xi)}{p(x)}\frac{p(Y|x \cup \xi)}{p(Y|x)} = \beta \prod_{\eta \in x} \phi(\xi, \eta) p(a, d, r)$$
$$\times \exp\left\{-\frac{1}{2(\sigma^2/\mathrm{ss}_\varphi + \tau^2)}\left(\sum_{i \in V} h(i;\xi)^2 - 2\sum_{i \in V} h(i;\xi)(\tilde{Y}_i - A_i(x))\right)\right\}$$
$$\leq c \exp\left\{\frac{1}{(\sigma^2/\mathrm{ss}_\varphi + \tau^2)}\sum_{i \in V} h(i;\xi)\tilde{Y}_i\right\}$$
$$\leq c \exp\left\{\frac{1}{(\sigma^2/\mathrm{ss}_\varphi + \tau^2)}\left(C_d C_a^2/(2\log 2\, v_x v_y)\sum_{i \in V}\tilde{Y}_i^2\right)^{1/2}\right\}.$$

Here the last inequality follows from (14). Let M denote the expression in the last line, this does not depend on x og ξ and the proof is complete. □

In the rest of this section we will restrict the support of the prior to the region D given by

$$D = \{x \in \Omega | \sum_{j=1}^{n(x)} h(i; x_j) < C_a \, \forall i \in V\}.$$

This assumption says that not only is C_a a natural upper bound for the height of individual activation bells, but also for the image obtained by combining all bells. As C_a was chosen arbitrarily large, this is not a restriction in practice.

Proposition 2 *There exists a $\delta > 0$ such that*

$$q_1(\xi \cup x|x) \geq \delta \quad \forall \xi \in \mathcal{X}, x \in \Omega \textit{ such that } x \cup \xi \in D.$$

Proof We will show that each of the factors in (12) is bounded below. Considering $q(\mu|x)$ we have

$$q(\mu|x) = Z^{-1} \exp\left\{\frac{1}{\sigma^2/\mathrm{ss}_\varphi + \tau^2}\sum_{i \in V} h(i; \mu, a_0, d_0, r_0, \theta_0)(\tilde{Y}_i - A_i(x))\right\},$$

where Z is the normalizing constant, that is the sum over $\mu \in V$ of the last term. By an evaluation such as that in the proof of Proposition 1 we have that this term is bounded above by a finite constant c_1, hence $Z^{-1} \geq (|V|c_1)^{-1}$. By the assumption that $x \cup \xi \in D$ we have that $A_i(x) < C_a$ for all $i \in V$ such that the last term is bounded below by a positive constant c_2. Hence we have $q(\mu|x) \geq c_2/(|V|c_1)$. A similar evaluation of $q(a|\mu, x)$, combined with the fact that the proposal is restricted to a bounded interval, shows that $q(a|\mu, x) \geq c_3$ for a positive constant c_3 and all x, μ and a. Likewise p_i defined in (16) is bounded below and above for all i, and hence so is $q(d|a, \mu, x)$. The proposals for r and θ are equivalent to the one for d. □

PROPOSITION 3 *The algorithm simulates a Markov chain that is Harris recurrent and geometrically ergodic.*

These properties are desirable, since they ensure that the chain will converge to the correct stationary distribution geometrically fast, such that a central limit theorem holds.

PROOF By Propositions 1 and 2 the probability of accepting an upstep can be dominated as follows,

$$\min\left\{1, \frac{p(x \cup \xi|Y)}{p(x|Y)} \frac{1}{(n+1)q_1(x \cup \xi|x)}\right\} \leq \frac{M}{\delta(n+1)}.$$

As the number of points n tends to infinity, the expression on the right hand side tends to zero. The probability of accepting a downstep is

$$\min\left\{1, \frac{p(x|Y)}{p(x \cup \xi|Y)}(n+1)q_1(x \cup \xi|x)\right\} \geq 1,$$

for n large enough. Hence if the number of points gets very large, the propability of accepting a further upsted is almost zero while we will allways accept a downstep. This guarantees a drift towards a smaller number of points which again implies geometrically ergodicity. We refer to the arguments in the proofs of Propositions 2 and 3 in Geyer (1999) for details. □

References

Baddeley, A.J. and van Lieshout, M.N.M. (1993) Stochastic geometry models in high-level vision. In K.V. Mardia and G.K. Kanji (eds.), *Statistics and Images*, vol. 1, chap. 11, pp. 231–256, Appl. Statist.

Barndorff-Nielsen, O., Kendall, W. and Lieshout, M. (eds.) (1999) *Stochastic Geometry. Likelihood and Computation.* Chapman & Hall/CRC.

Buckner, R.L. (1998) Event-related fMRI and the hemodynamic response. *Human Brain Mapping*, **6**, 373–377.

Bullmore, E., Brammer, M., Williams, S.C., Rabe-Hesketh, S., Janot, N., David, A., Mellers, J., Howard, R. and Sham, P. (1996) Statistical methods of estimation and inference for functional MR image analysis. *Magn. Reson. Med.*, **35**, 261–277.

Dale, A.M. and Buckner, R.L. (1997) Selective averaging of individual trials using fMRI. *NeuroImage*, **5**, S47.

Friston, K.J., Jezzard, P. and Turner, R. (1994) The analysis of functional MRI time-series. *Human Brain Mapping*, **1**, 153–171.

Friston, K.J., Holmes, A.P., Poline, J.B., Grasby, P.J., Williams, S.C.R., Frackowiak, R.S.J. and Turner, R. (1995) Analysis of fMRI time-series revisited. *NeuroImage*, **2**, 45–53.

Friston, K.J., Josephs, O., Rees, G. and Turner, R. (1998) Nonlinear event-related responses in fMRI. *Magn. Reson. Med.*, **39**, 41–52.

Gaschler-Markefski, B., Baumgart, F., Tempelmann, C., Schindler, F., Stiller, D., Heinze, H.J. and Scheich, H. (1997) Statistical methods in functional magnetic resonance imaging with respect to nonstationary time-series auditory cortex activity. *Magn. Reson. Med.*, **38**, 811–820.

Geyer, C. (1999) Likelihood inference for spatial point processes. In Barndorff-Nielsen *et al.* (1999), chap. 3.

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

Holmes, A.P., Josephs, O., Büchel, C. and Friston, K.J. (1997) Statistical modelling of low-frequency confounds in fMRI. *NeuroImage*, **5**, S480.

Kornak, J., Haggard, M.P. and 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 and I.L. Dryden (eds.), *Spatial Temporal Modelling and its Applications*. Leeds University Press.

Kullback, S. (1959) *Information Theory and Statistics*. John Wiley & Sons, Inc.

Kwong, K.K., Belliveau, J.W., Chesler, D.A., Goldberg, I.E. *et al.* (1992) Dynamic magnetic resonance imaging of human brain activity during primary sensory stimulation. *Proc. Natl. Acad. Sci. USA*, **89**, 5675–5679.

Lange, N. (1996) Tutorial in biostatistics. Statistical approaches to human brain mapping by functional magnetic resonance imaging. *Statistics in Medicine*, **15**, 389–428.

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

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

Le, T.H. and Hu, X. (1996) Retrospective estimation and correction of physiological artifacts in fMRI by direct extraction of physiological activity from MR data. *Magn. Reson. Med.*, **35**, 290–298.

Lee, A.T., Glover, G.H. and Meyer, C.H. (1995) Discrimination of large venous vessels in time-course spiral blood-oxygen-level-dependent magnetic-resonance functional neuroimaging. *Magn. Reson. Med.*, **33**, 745–754.

Møller, J. (1999) Markov chain Monte Carlo and spatial point processes. In Barndorff-Nielsen *et al.* (1999), chap. 4.

Nielsen, F.Å., Hansen, L.K., Toft, P., Goutte, C., Lange, N., Strother, S.C., Mørch, N., Svarer, C., Savoy, R., Rosen, B., Rostrup, E. and Born, P. (1997) Comparison of two convolution models for fMRI time series. *NeuroImage*, **5**, S473.

Ogata, Y. and Tanemura, M. (1984) Likelihood analysis of spatial point patterns. *J. R. Statist. Soc. B*, **46**, 496–518.

Petersen, N.V., Jensen, J.L., Burchhardt, J. and Stødkilde-Jørgensen, H. (1998) State space models for physiological noise in fMRI time series. *NeuroImage*, **7**, S592.

Press, W.H., Teukolsky, S.A., Vetterling, W.T. and Flannery, B.P. (1992) *Numerical Recipes in C.* Cambridge University Press, second edn.

Shafie, K., Worsley, K.J., Wolforth, M. and Evans, A.C. (1998) Rotation space: Detecting functional activation by searching over rotated and scaled filters. *NeuroImage*, **7**, S755.

Siegmund, D.O. and Worsley, K.J. (1995) Testing for a signal with unknown location and scale in a stationary Gaussian random field. *Ann. Stat.*, **23**, 608–639.

Vazquez, A.L. and Noll, D.C. (1998) Nonlinear aspects of the BOLD response in functional MRI. *NeuroImage*, **7**, 108–118.

West, M. and Harrison, J. (1989) *Bayesian Forecasting and Dynamic Models.* Springer-Verlag, New York.

Worsley, K.J. and Friston, K.J. (1995) Analysis of fMRI time-series revisited — again. *NeuroImage*, **2**, 173–181.

Spatial mixture modelling of fMRI data

Niels Væver Hartvig and Jens Ledet Jensen[*]

Department of Theoretical Statistics
Department of Mathematical Sciences and MaPhySto[*]
University of Aarhus

June 2, 2000
Revised version

Corresponding author:

Niels Væver Hartvig
Department of Mathematical Sciences
University of Aarhus
Ny Munkegade
DK-8000 Aarhus C
Denmark

Phone: +45 8942 3439
Fax: +45 8613 1769
e-mail: vaever@imf.au.dk

Running title: Spatial mixture modelling of fMRI data

Keywords: functional magnetic resonance imaging, spatial model, mixture model, image analysis

Copyright notice: This is a preprint of an article accepted for publication in Human Brain Mapping © 2000 Wiley-Liss, Inc.

[*]Centre for Mathematical Physics and Stochastics, funded by a grant from the Danish National Research Foundation.

Abstract

Recently Everitt and Bullmore (1999) proposed a mixture model for a test statistic for activation in fMRI data. The distribution of the statistic was divided into two components; one for non-activated voxels and one for activated voxels. In this framework one can calculate a posterior probability for a voxel being activated, which provides a more natural basis for thresholding the statistic image, than that based on p-values. In this article, we extend the method of Everitt and Bullmore to account for spatial coherency of activated regions. We achieve this by formulating a model for the activation in a small region of voxels, and use this spatial structure when calculating the posterior probability of a voxel being activated. We have investigated several choices of spatial models, but find that they all work equally well for brain imaging data. We applied the model to synthetic data from statistical image analysis, a synthetic fMRI data set and to visual stimulation data. Our conclusion is that the method improves the estimation of the activation pattern significantly, compared to the non-spatial model and to smoothing the data with a kernel of FWHM 3 voxels. The difference between FWHM 2 smoothing and our method were more modest.

1 Introduction

In the literature on analysis of functional magnetic resonance imaging (fMRI) data the focus is primarily on the temporal aspect. Perhaps the most common analysis scheme is to treat voxel time series separately, and estimate the activation level voxel by voxel. This framework ranges from simple t-tests and correlation methods to more detailed models for the haemodynamic response, and models which account for correlated noise. The latter encompasses generalized linear models and time series models. A few papers which fall in this category are Bandettini *et al.* (1993), Bullmore *et al.* (1996), Worsley and Friston (1995) and Lange and Zeger (1997), but we refer to an overview paper, like Lange *et al.* (1999), for the long list of references which should be cited in this context.

The spatial properties of the data are rarely modelled with the same care as is given the temporal ones: Common approaches are either to assume spatial independence, or to smooth data spatially with a Gaussian kernel. The latter approach has been studied primarily by Keith Worsley in a series of papers, see for instance Worsley *et al.* (1996). Smoothing the data spatially is in fact equivalent to using a non-parametric model for the spatial activation pattern, assuming only smoothness of the latter (Müller, 1988). It should hence be viewed as an estimation procedure which is optimal in this model, but there is no general statistical reason for smoothing. On the contrary smoothing may produce a biased estimate, by displacing activation peaks and underestimating the height of the latter (Descombes *et al.*, 1998; Hartvig, 1999).

Even if explicit spatial models are rare, the value of including spatial information in the analysis has been recognized for many years. Commonly this is achieved by assessing significance of activation by the size of supra-threshold clusters. This was first suggested by Poline and Mazoyer (1993) and have later been studied from a theoretical point of view (Friston *et al.*, 1994; Poline *et al.*, 1997), using Monte Carlo methods (Forman *et al.*, 1995) and permutation methods (Bullmore *et al.*, 1999). In our minds the important distinction here, is that of a spatial model and the inference made in this. Even if cluster size is used as a measure of significance, the estimated pattern is still a product of the underlying model used to produce the clusters. Also in this context, the non-parametric smoothing model seems to be the typical choice.

Recently Descombes *et al.* (1998) proposed a Markov random field model for the spatio-temporal activation pattern and used this for estimation of the latter. Their assumption is that the activation pattern is spatially coherent, yet may possess sharp boundaries between different regions, and the model introduces this explicitly in the estimation procedure. Assessment of uncertainty and significance is not straightforward in this framework, since it requires simulations of the posterior distribution of the spatio-temporal activation pattern. In principle this may be done with Markov chain Monte Carlo (MCMC) techniques (Gilks *et al.*, 1996), but since the state space of the spatio-temporal activation pattern is enormous it is a time-consuming and far from trivial task. Instead the authors suggested to use the procedure only as a preprocessing step, and did not use the model for making explicit inference on the activation.

The dimensionality of the activation pattern is much reduced in Hartvig (1999) where stronger assumptions are made. Specificly the activation is modelled as a collection of centres with Gaussian shape, but with unknown extent and height. This enables inclusion of prior information directly in the model, and simulation of the posterior distribution is possible by MCMC. However also in this context the need to perform lengthy simulations is a limitation of the method.

The problems of the two last approaches perhaps explain the lack of spatial models: 1) It is somewhat difficult to formulate the general idea of coherency of activated regions in a specific model, which is still general enough to model the range of patterns observed in brain data. 2) Most spatial models are analytically intractable, and statistical inference must rely on simulation methods, which are time-consuming and often requires a lot of user interaction. The latter makes them less suitable for routine use. In this paper we try to bridge the gap between formulating a spatial model which has some realistic properties, and the computational feasibility, which makes it applicable in a routine analysis. The idea is to formulate the model through the marginal distribution on a small grid of voxels, for instance a 3 by 3 region in the slice.

Though the model may be used as the spatial part of a spatio-temporal model, we will only consider the problem of estimating the activation pattern based on a single summary image (or volume) of voxel-wise activation estimates, also known as a statistical parametric map (SPM). Let $\{x_i\}$ denote the latter, where i indexes the voxels. Recently Everitt and Bullmore (1999) (henceforth denoted EB) suggested a marginal analysis of such an image. Let A_i be the indicator for voxel i being activated. The approach of EB is to calculate the conditional probability $P(A_i = 1|x_i)$ for each voxel, and use the latter to estimate the activated areas. In order to calculate this, they specify the distribution of activated and non-activated voxels, i.e. the conditional distributions $p(x_i|A_i = 1)$ and $p(x_i|A_i = 0)$, as well as the probability $P(A_i = 1)$. The method does not use any spatial properties of the data.

What we propose in this article is to keep the simplicity of the approach in EB, but to extend it in such a way that spatial interaction is partly taken into account. Instead of using $P(A_i = 1|x_i)$ we suggest to use $P(A_i = 1|x_{C_i})$, where C_i is voxel i together with the neighbouring voxels. The idea is that activated areas tend to constitute a group of at least a few voxels, hence voxel i has a higher chance of being activated if both voxel i and some of its neighbours have high values. Conversely the activation probability is small if x_i is high, but all the neighbours has small values. The main problem in this approach becomes the specification of the marginal probabilities of the activation A_{C_i} in the region C_i. We propose three different models for these probabilities, ranging from a very simple one to a more realistic one. Common to all is that the probability of a voxel being activated has a simple expression, which can be easily calculated.

2 Theory

In the two first subsections we present an overview of the method and the spatial models for the activation pattern. The third subsection is on estimation of parameters in the model, and is more technical than the two first. The reader who is most interested in the general concept and examples of application of the model may skip this third subsection on a first reading.

2.1 Overview of the mixture model

As mentioned in the introduction, we assume that a statistical parametric map $\{x_i\}$ is given, and we wish to derive a posterior probability that a voxel is activated using this map. In the following we will describe how a local model for the activation pattern around a voxel i, can be used to incorporate spatial information in the posterior probability. In order to simplify notation we will drop the voxel index i from the notation.

Suppose we consider k neighbours around voxel i. Typically these would be the 8 neighbours in a 3×3 square in the slice or the 26 neighbours in a $3 \times 3 \times 3$ cube in a volume of slices with voxel i in the centre. We will let C denote the set of $k+1$ voxels given by voxel i together with the k neighbours.

We will let A be an indicator for the event that voxel i is activated, in the sense that $A = 0$ means that there is no activation in voxel i and $A = 1$ means that the voxel is activated. Likewise, we will let $A^1, \dots, A^k$ be a vector of indicators for activation in the k neighbours. We index the A's by a superscript to avoid confusion with the usual voxel subscript. Finally $A_C = (A, A^1, \dots, A^k)$ is the vector of all activation indicators in C. We will consider this as a vector of unobserved stochastic variables, and formulate a model for it's distribution. Thus for each vector $a_C = \{0,1\}^{k+1}$ we specify the prior probability $P(A_C = a_C)$ that the activation configuration takes a particular value. Different choices of models, which reflect the idea that activated areas tend to constitute a cluster of voxels, are proposed in the next section.

Rather than observing A_C, we observe $x_C = (x, x^1, \ldots, x^k)$, the values of the test statistic for activation in the different voxels. Like before x is the value for voxel i, and $x^1, \ldots, x^k$ are the values for the neighbours. The usual hypothesis testing approach assumes a specific model for x given that the voxel is not activated, for instance that this is a normal variable with zero mean and unit variance. In our setup, we require that one can also specify the alternative distribution, i.e. the distribution of x given $A = 1$. In EB the statistics are fundamental power quotients (FPQ), which have respectively a central and a non-central χ^2-distribution under the two activation states. In our Example 3, the test statistics are the estimated activity level from a regression analysis, and it is natural to take $(x|A = 0) \sim N(0, \sigma^2)$. When the voxel is activated, $A = 1$, it is not so clear what the proper distribution is. We find that the range of different activation levels are described well by a Gamma distribution, $(x|A = 1) \sim \Gamma(\lambda, \beta)$. Denote the distribution of x_C given $A_C = a_C$ by the density $f(x_C \,|\, a_C)$.

When these two parts of the model are specified it is straightforward to calculate the posterior probability of an activation configuration a_C given the data x_C, since, by Bayes rule, this is given by

$$P(A_C = a_C \,|\, x_C) \propto f(x_C \,|\, a_C) P(A_C = a_C).$$

Thus the posterior probability that the activation pattern A_C equals a_C is simply proportional to the likelihood of observing x_C given $A_C = a_C$ times the prior probability of $A_C = a_C$. In particular, one may calculate the probability that voxel i is active or not, irrespectively of the neighbours, by summing over the neighbouring states,

$$P(A = a \,|\, x_C) \propto \sum_{a^1=0}^{1} \cdots \sum_{a^k=0}^{1} f(x_C \,|\, a_C) P(A_C = a_C), \tag{1}$$

where $a_C = (a, a^1, \ldots, a^k)$. The constant of proportionality can be determined from the fact that the probabilities $P(A = 0 \,|\, x_C)$ and $P(A = 1 \,|\, x_C)$ must sum to one.

The problem with using this approach in practice is the calculation of the sum in (1), which has 2^k terms. In the situation with a 3×3×3 neighbourhood cube, the sum thus has 2^{26} or about 67 million terms, and since we must calculate this for each voxel in the volume, we are facing an order of 10^{13} iterations. Even though the computations may be performed in parallel, this is of course hopelessly too many in practice. The main contribution of our method, is that we propose models for $P(A_C = a_C)$, which are able to model clustered activation, but where the sum may be calculated analytically. Thus we obtain a simple, closed form expression for the posterior probability that a voxel is activated, which may be calculated almost instantly. These are given for each of the three models in the following sections, see equations (4), (16) and (19).

We will assume in the following that the statistics x_C are independent given the true activation pattern A_C. Thus the density of x_C given A_C can be written as,

$$f(x_C \mid A_C = a_C) = f(x \mid a) \prod_{j=1}^{k} f(x^j \mid a^j),$$

where $f(x \mid a)$ is the density of x given $A = a$.

2.2 Models for the marginal probabilities

In this section we give three choices for the marginal probabilities $P(A_C = a_C)$, $a_C = (a, a^1, \ldots, a^k) \in \{0,1\}^{k+1}$. For an activation configuration a_C, we will let $s = a + a^1 + \cdots a^k$, that is the number of ones in a_C.

2.2.1 Model 1

Perhaps the most simple choice is to take

$$P(A_C = a_C) = \begin{cases} q_0 & \text{if } s = 0, \\ q_1 & \text{if } s > 0. \end{cases} \tag{2}$$

Since there are 2^{k+1} values of a_C we must have $q_0 = 1 - (2^{k+1} - 1)q_1$ in order that the probabilities sum to one. Thus this distribution has only one parameter and a natural way of interpreting this parameter is through the probability p of a voxel being activated. This gives $p = q_1 2^k$ or

$$q_1 = p2^{-k} \quad \text{and} \quad q_0 = 1 - (2 - 2^{-k})p. \tag{3}$$

Notice that in the model there is equal probability of observing a configuration with all ones and one with only ones in a corner of the region C, for instance. If we expected the activated areas to be large coherent regions, the former probability should be larger than the second, whereas if we expected the areas to be of moderate size but with long boundaries, the second probability should be larger than the first. The above model hence represents the situation that we neither believe that activated regions consist of single voxels nor that they are very large.

We will illustrate in this simple situation how the posterior probability in (1) may be calculated. We shall be using the equality

$$\sum_{a^1=0}^{1} \cdots \sum_{a^k=0}^{1} \left(\prod_{j=1}^{k} f(x^j|a^j) \right) = \prod_{j=1}^{k} \left\{ f(x^j|0) + f(x^j|1) \right\}.$$

Let η denote the above product. When $a = 0$ the expression in (1) is

$$\begin{aligned} P(A = 0 \,|\, x_C) &\propto f(x \,|\, 0) \sum_{a^1=0}^{1} \cdots \sum_{a^k=0}^{1} \left(\prod_{j=1}^{k} f(x^j|a^j) \right) P(A_C = a_C) \\ &= f(x \,|\, 0) \left(q_1\eta + (q_0 - q_1) \prod_{j=1}^{k} f(x^j \,|\, 0) \right), \end{aligned}$$

and when $a = 1$ we simply get

$$P(A = 1 \,|\, x_C) \propto f(x \,|\, 1) q_1 \eta.$$

Since the two probabilities must sum to one, we find,

$$\begin{aligned} P(A = 1|x_C) &= \frac{f(x|1)q_1\eta}{f(x|1)q_1\eta + f(x|0) \left(q_1\eta + (q_0 - q_1) \prod_{j=1}^{k} f(x^j|0) \right)} \\ &= \left\{ 1 + \frac{1}{v} \left[1 + \left(\frac{q_0}{q_1} - 1 \right) \left(\prod_{j=1}^{k} (1 + v^j) \right)^{-1} \right] \right\}^{-1}, \end{aligned} \qquad (4)$$

where

$$v = \frac{f(x \,|\, 1)}{f(x \,|\, 0)}, \quad v^j = \frac{f(x^j|1)}{f(x^j|0)} \quad j = 1, \ldots, k. \qquad (5)$$

Notice that v is the likelihood ratio for the voxel being active vs. not active. The formula (4) thus effectively combine the likelihood ratios from voxel i together with those of its neighbours to calculate the posterior probability of activation. The formula shows in a direct way the difference to the approach in EB. If all the neighbours are non-activated then (4) will typically be of the order

$$\left\{ 1 + \frac{f(x|0)}{f(x|1)} \frac{q_0}{q_1} \right\}^{-1}$$

whereas if at least one neighbour is activated the order is typically

$$\left\{1+\frac{f(x|0)}{f(x|1)}\right\}^{-1}.$$

For illustration let us consider the case where $p = 0.02$ and $k = 8$. Then $q_0/q_1 = 12289$, and if $f(x|0)/f(x|1) \approx \exp(-8)$ then the first term is 0.20 whereas the second expression is 0.9997.

2.2.2 Model 2

Another simple choice of $P(A_C = a_C)$ is

$$P(A_C = a_C) = \begin{cases} q_0 & \text{if } s = 0, \\ \alpha\gamma^{s-1} & \text{if } s > 0. \end{cases} \tag{6}$$

Here $\gamma = 1$ gives back the model 1 in (2), whereas the restriction $\alpha = \gamma/(1+\gamma)^{k+1}$ corresponds to the model where the voxels are independent and the probability of a voxel being activated is $\gamma/(1+\gamma)$. The latter is equivalent to the model in EB.

The model may be parametrized by the probability p of a voxel being active, which is given as $p = \alpha(1+\gamma)^k$, and by γ. The latter is a measure of correlation of neighbouring activation sites. The last parameter q_0 is given by the constraint that the probabilities must sum to one. The posterior probability of activation may be derived in the same way as in model 1, the expression is given in (16) in the appendix.

2.2.3 Model 3

Finally, we will consider a model of the form (6), but being more symmetric with respect to activated and non-activated voxels. We will consider the model

$$P(A_C = a_C) = \begin{cases} q_0 & \text{if } s = 0, \\ \alpha_1\gamma_1^{s-1} + \alpha_2\gamma_2^{s-k} & \text{if } 1 \le s \le k, \\ q_1 & \text{if } s = k+1. \end{cases} \tag{7}$$

The model may be parametrized by the probability p of a voxel being active, and 4 other parameters describing the correlation between voxels. The relation between parameters may be found in the appendix, as may the expression for the probability that a voxel is active (18).

2.3 Estimation of parameters

For estimation purposes, we will now study the whole volume of voxels, rather than just a single voxel. For this reason, we will let the notation depend explicitly on the voxel index. Rather than just using x_C, we will let x_{C_i} denote the vector of observations in the region C_i around voxel i. The elements of the vector are denoted by $x_{C_i} = (x_i^0, x_i^1, \ldots, x_i^k)$, thus x_i^0 refers to the statistic x_i in voxel i, and $x_i^1, \ldots, x_i^k$ to the statistic in the k neighbours of i. Similarly A_C is changed to $A_{C_i} = (A_i^0, A_i^1, \ldots, A_i^k)$ and the likelihood ratios (5) are denoted v_i^j, where

$$v_i^j = \frac{f(x_i^j \mid 1)}{f(x_i^j \mid 0)}, \quad j = 0, 1, \ldots, k, \ i \in V.$$

Within the model we can calculate the marginal density of x_{C_i}. We denote this by $f(x_{C_i}; \phi, \psi)$, where ϕ parametrizes the conditional distribution of x_{C_i} given A_{C_i}, and ψ parametrizes the marginal distribution of A_{C_i}. Thus

$$f(x_{C_i}; \phi, \psi) = \sum_{a_C \in \{0,1\}^{k+1}} f(x_{C_i} | A_{C_i} = a_C; \phi) P(A_{C_i} = a_C; \psi).$$

A possibility for estimating the parameters (ϕ, ψ) is to maximize the contrast function

$$\gamma(\phi, \psi) = \sum_{i \in V} \log f(x_{C_i}; \phi, \psi). \tag{8}$$

This is related to maximum likelihood estimation, in particular the estimators will be asymptotically normal distributed under conditions where the maximum likelihood estimators are. For model 2, and hence also for model 1 by setting $\gamma = 1$, we get

$$\begin{aligned} &f(x_{C_i}; \phi, \gamma, \alpha) \\ &\quad = \prod_{j=0}^{k} f(x_i^j | 0; \phi) \left\{ \frac{\alpha}{\gamma} \prod_{j=0}^{k} (1 + \gamma v_i^j(\phi)) + 1 - \frac{\alpha(1+\gamma)^{k+1}}{\gamma} \right\}. \end{aligned} \tag{9}$$

The formula for model 3 is given in (19) in the appendix.

Usually, though, we will take a more simple approach instead of using (8). We propose to use only the marginal distribution of x_i to estimate ϕ and the fraction of activated voxels p. The marginal density of x_i is a mixture density

$$f(x; \phi, p) = (1-p) f(x \mid 0; \phi) + p f(x \mid 1; \phi), \tag{10}$$

We thus maximize the contrast function

$$\gamma_m(\phi, p) = \sum_{i \in V} \log f(x_i; \phi, p) \tag{11}$$

to estimate ϕ and p. Under model 1 all parameters have been estimated this way.

When $P(A_{C_i} = a_{C_i})$ is given by model 2 we still estimate $p = \alpha(1+\gamma)^k$ from (11). The remaining parameter γ may then be estimated from the empirical covariance of $\{x_i\}$: Suppose, for example, that $(x \mid A = 0) \sim N(0, \sigma^2)$, and $(x \mid A = 1) \sim N(1, \sigma^2)$. Then the covariance of x_i and x_j is given by

$$\begin{aligned}\mathrm{Cov}(x_i, x_j) &= P(A_i = A_j = 1) - P(A_i = 1)P(A_j = 1) \\ &= P(A_i = A_j = 1) - p^2. \end{aligned} \tag{12}$$

If j is a neighbour to i, say neighbour number 1, we may derive the first probability as

$$\begin{aligned}P(A_i = A_j = 1) &= \sum_{a^j \in \{0,1\}, j=2,\ldots,k} P(A_{C_i} = a_C) = \alpha\gamma \sum_{a^j \in \{0,1\}, j=2,\ldots,k} \gamma^{a^2+\cdots+a^k} \\ &= \alpha\gamma(1+\gamma)^{k-1} = p\frac{\gamma}{1+\gamma}. \end{aligned} \tag{13}$$

Notice that the two expression above does not depend on the position of the neighbour j. Suppose an estimate $\hat{C}$ of the covariance $\mathrm{Cov}(x_i, x_j)$ is given. This may be combined with the estimate $\hat{p}$ of p to form an estimate of γ by the equations above,

$$\hat{\gamma} = \frac{b}{1-b} \qquad \text{where } b = \hat{C}\hat{p}^{-1} + \hat{p}. \tag{14}$$

Since the covariance is the same for all neighbours, we may combine estimates of the covariance at different spatial lags within the neighbourhood, to form the estimate $\hat{C}$. In practice in our examples (where we consider respectively 3×3 and 5×5 neighbourhoods) we have used the eight nearest neighbours to estimate the covariance,

$$\hat{C} = \frac{1}{4}(\hat{C}_{(1,0)} + \hat{C}_{(1,1)} + \hat{C}_{(0,1)} + \hat{C}_{(-1,1)}).$$

Here $\hat{C}_l$ is the correlogram for the spatial lag l (Cressie, 1991),

$$\hat{C}_l = \frac{1}{N_l} \sum_{j \in V, j+l \in V} (x_j - \bar{x}.)(x_{j+l} - \bar{x}.),$$

where V denotes the set of brain voxels, N_l is the number of terms in the sum, and $\bar{x}.$ is the average of the x_i's.

Notice that the probability in (13) only depends on the model for A_C, and hence applies whenever model 2 is considered. This is not true for the covariance in (12), which depends on the distribution of x given A. In the setup above we have considered a statistic which is distributed as $(x \mid A = 0) \sim N(0, \sigma^2)$ and $(x \mid A = 1) \sim N(1, \sigma^2)$. When more generally $(x \mid A = 1) \sim N(\mu, \sigma^2)$ where $\mu > 0$, we obtain the setup above by scaling x by μ^{-1}. When the distribution of x is not normal, one needs to calculate the covariance in (12) for the distribution considered. A general formula, which applies whenever x_i and x_j are conditionally independent given A_i and A_j, is given by

$$\mathrm{Cov}(x_i, x_j) = \mathrm{Cov}(E(x_i \mid A_i), E(x_j \mid A_j)).$$

Usually it is straightforward to calculate the right hand side above. This is the approach used in Example 3, where x_i has a Gamma distribution when $A_i = 1$.

As for the model 3, this has 4 free parameters when p is given. Moment estimators may be derived for these as above, but we will refrain from this since the equations get more complicated. Instead we will estimate the remaining parameters from (8).

In our examples below we have used the simplex method to maximize the contrast functions (Press *et al.*, 1992). The standard errors of the maximum contrast estimators may be obtained by general asymptotic theory, see for instance Heyde (1997). Cressie (1991) provides formulas for the standard error of $\hat{C}_l$. Presently we have no formal way of including the uncertainty of the parameters in the analysis, it is, however, our experience, that the posterior probability maps were quite robust to the observed variations in the parameters. In fact, as we will show in the Example 2 and 3, they are quite robust to the choice of model.

3 Simulations and applications

We will illustrate the method by applying it to two synthetic data sets, where the truth is known, and a visual stimulation data set. For the synthetic data, we may quantify results by respectively classification error, statistical power or true positive rate (TPR) and level of significance or false positive rate (FPR). For a given threshold, the classification error is estimated as the number of misclassified voxels (either type I or type II errors), divided by the total number of voxels. The TPR is estimated as the as the number of active voxels classified as active, divided by the total number of active voxels. The FPR is estimated as the number of non-active voxels which are classified as active, divided by the total number of non-active voxels.

3.1 Example 1: Image restoration data

We will apply the models to a classical problem in statistical image analysis, namely the restoration of an unknown true image based on a degraded version of it. Techniques for achieving this are applied in many areas where images are recorded or transmitted with noise, including remote sensing images, satellite images and medical images. In functional brain imaging the problem is more complex than in the setting above: It is not as evident what the "true scene" is or which geometric characteristics it has, and the noise sources are far more complex than in image restoration problems. It still serves a purpose, however, to study how the models perform in this more simple problem, in order to understand the characteristics of the models, before moving on to more complex data.

We will consider two images. The first (denoted Image I) is the 64×64 binary image of an 'A' by Greig *et al.* (1989), see Figure 1. The image is corrupted with binary noise, where a pixel A_i with probability q is replaced by $1 - A_i$. The probability densities of the degraded pixel X_i given the true value A_i are then

$$\begin{aligned} f(x \mid A = 0) &= q^x (1-q)^{1-x}, \quad x \in \{0,1\}, \\ f(x \mid A = 1) &= (1-q)^x q^{1-x}, \quad x \in \{0,1\}. \end{aligned}$$

The error rate q was set to 25%. Five independently corrupted images were produced, in order to assess the variability of the estimates. The results are summarized in Table 1 and some of the image estimates are displayed in Figure 1.

The second image (Image II) is the binary image displayed in Fig. 4a of Besag (1986). The image was corrupted by adding white Gaussian noise with standard deviation 0.9105. In this setting the densities of a pixel X_i given A_i are

$$\begin{aligned} f(x \mid A = 0) &= \frac{1}{\sqrt{2\pi}\tau} e^{-\frac{1}{2\tau^2} x^2}, \quad x \in \mathbb{R}, \\ f(x \mid A = 1) &= \frac{1}{\sqrt{2\pi}\tau} e^{-\frac{1}{2\tau^2} (x-1)^2}, \quad x \in \mathbb{R}, \end{aligned}$$

where $\tau = 0.9105$. We produced five independent noisy images to assess the variability of estimates. The results are given in Table 1.

For each model, the parameters were estimated both by maximizing the contrast function (8) and, for model 1 and 2, by the simple estimators described in Section 2.3. Since the results were almost similar, we give only the figures for the maximum-constrast estimates. In practice we recommend that the simple estimators should be used when possible, since they are much easier to obtain, and give almost as good results.

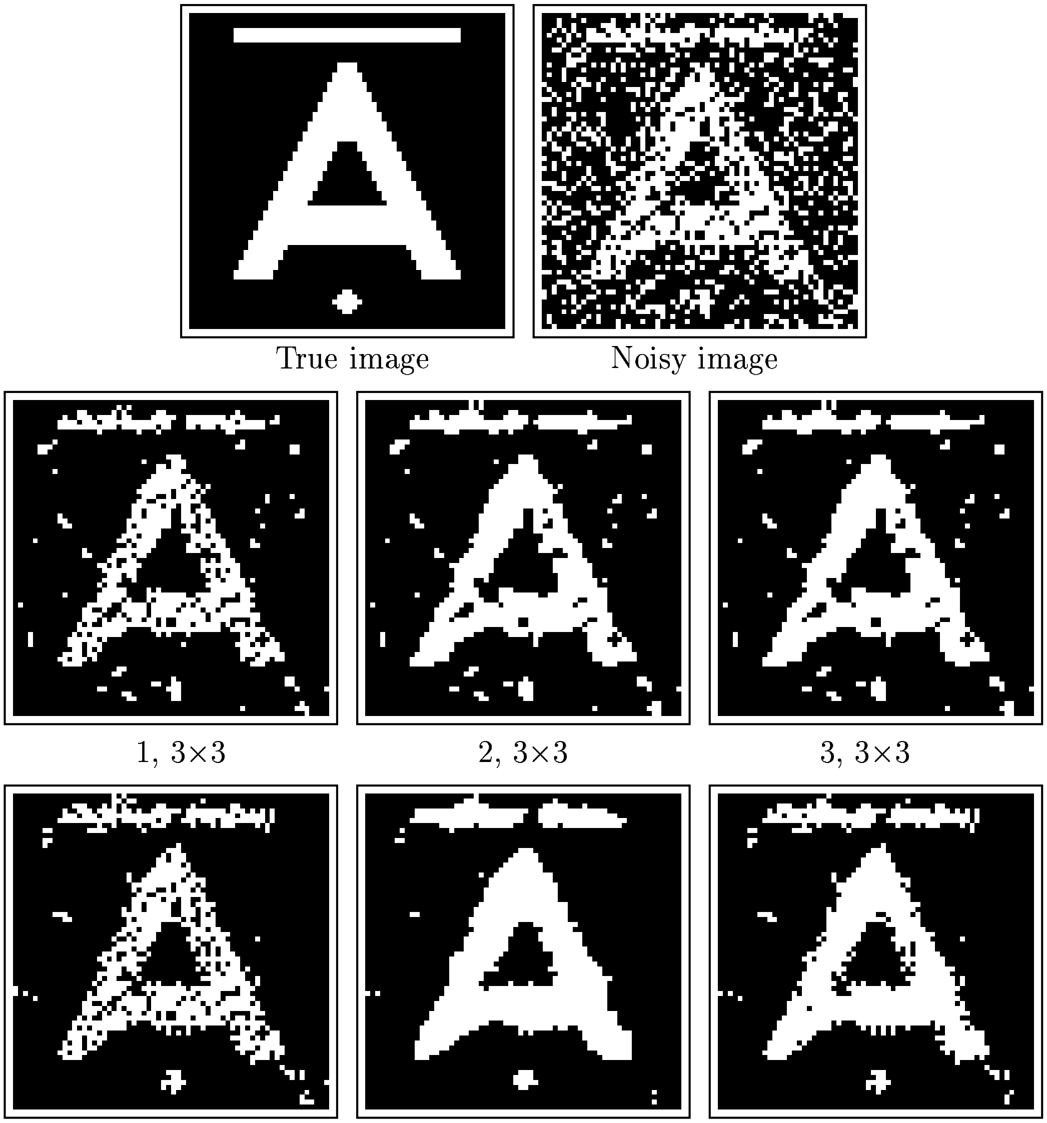
We calculated the posterior probability of $A = 1$ given X_C in each pixel, and the estimate of the true image was obtained by thresholding the probability image at 0.5. The estimates for one of the noisy versions of image I can be seen in Figure 1.

The estimated classification error and its standard error are listed in the second and third column of Table 1. The first column lists the models used in this example. The models 1, 2 and 3 of Section 2.2 were applied, respectively defined on a 3 by 3 pixel region and on a 5 by 5 region. For comparison, we have reproduced the classification errors of the maximum a posteriori (MAP) estimate and the iterated conditional modes (ICM) estimate, which can be found in Greig *et al.* (1989). These two estimates are based on the same global model for the true image, but only the local properties of the model are used with ICM.

Table 1: Estimated classification errors for the three models and the ICM and MAP estimates, based on 5 independent simulations of the degraded image. Image I refers to the true image in Figure 1, degraded with binary noise. Image II refers to the image in Fig. 4a in Besag (1986), degraded with Gaussian noise. All figures are in percent, standard errors of estimates are given in parentheses.

Model	Class. error	
	Image I	Image II
1, 3×3	10.0 (0.3)	14.6 (0.3)
1, 5×5	9.4 (0.2)	12.2 (0.2)
2, 3×3	7.6 (0.3)	9.0 (0.4)
2, 5×5	5.9 (0.8)	6.4 (0.2)
3, 3×3	7.6 (0.3)	9.0 (0.4)
3, 5×5	6.1 (0.3)	6.2 (0.3)
MAP	5.2 (0.2)	5.5 (0.2)
ICM	6.3 (0.4)	6.4 (0.1)

The table shows that model 1 performs worse than model 2 and 3, which is also clear from Figure 1. It is also clear that the 5 by 5 region models are superior in this setting, which is not surprising since the true images are quite regular with large patches of either black or white. We might suspect that the 3 by 3 models will be more appropriate in brain imaging, where the true scene is not as regular. Model 2 and 3 perform almost equally well, hence we prefer model 2, since this only has two parameters.



True image

Noisy image

1, 3×3

2, 3×3

3, 3×3

1, 5×5

2, 5×5

3, 5×5

Figure 1: Comparison of spatial mixture models. Top row: Image I and degraded version. Middle row: Estimates of the true image based on model 1, 2 and 3 applied to a 3×3 pixel region. Bottom row: Same as above, but with the models defined on a 5×5 region.

Model 2 performs well compared to the ICM and MAP methods also. There are several practical differences between these and our model: Firstly, it is more computationally intensive to obtain the ICM and MAP estimates, than our posterior probability images. The latter are calculated in closed form, while the ICM and MAP procedures require iterative algorithms. Secondly, the MAP and ICM procedures depend on a smoothing parameter which, especially for the MAP estimate, is crucial for the reconstructed image. In this case, the value of the smoothing parameter was based on the true image, which is of course not possible in practice. On the contrary the parameters of model 2 are estimated directly from the observed image. Seen in this light, our model seems to be an attractive alternative to the traditional methods. It is however not as flexible as the ICM approach, which can be generalized for instance to multicolour settings.

3.2 Example 2: Simulated fMRI data

In order to study the performance on data which are closer related to brain imaging problems than the ones in Example 1, we have applied the methods to a synthetic fMRI data set. We used the data set of Lange *et al.* (1999), which was generated from 72 baseline EPI scans that were temporally resampled to 384 scans[1]. We refer to the paper for a full description of the data, but will repeat the basic properties here. A region of 24 by 12 voxels is considered, and in each voxel the time series is linearly detrended. Denote the residual time series by Y_{it}, where i indexes voxels $i = 1, \ldots, V$ and t indexes scan $t = 1, \ldots, T$. Here $V = 288$ and $T = 384$. Artificial activation was added to obtain the actual data Z_{it}, say, by the model

$$Z_{it} = b_i x_t + Y_{it},$$

where the magnitude of activation b_i is given by

$$b_i = m s_{Y,i}.$$

Here $s^2_{Y,i}$ is an estimate of σ^2_i, the variance of Y_{it}, given by

$$s^2_{Y,i} = \frac{1}{T-1}\sum_{t=1}^{T}(Y_{it} - \bar{Y}_{i\cdot})^2, \quad \bar{Y}_{i\cdot} = \frac{1}{T}\sum_{t=1}^{T} Y_{it}.$$

The temporal activation pattern x_t is a simple binary function, where $x_t = 0$ when off and $x_t = 1$ when on, for $t = 1, \ldots, T$. The function is periodic

[1] The data may be obtained from the address http://pet.med.va.gov:8080/plurality.

with 8 runs, each of length 48 scans with 12 scans off, 24 on and 12 off. The ratio m of the activation magnitude to standard deviation was chosen to be positive and constant in the two connected regions of size 25 and 37 voxels depicted in Figure 2, and zero elsewhere. According to Lange *et al.* a value of $m = 0.15$ was chosen in the activated areas, however when estimating m directly from the data by a regression analysis (when the true activation pattern is known), we obtain $\hat{m} = 0.43$ with a standard error of 0.015. The value of m is not important for the present study, however.

In order to make the estimation problem a bit harder than in the paper, we divided the data into 4 subsets, each of length 96 scans. We estimated the spatial activation pattern from a single subset at a time, and used the empirical variation over the four subsets to evaluate the uncertainty of our results.

Consider a voxel time series at voxel i, Z_{it}, for $t = 1, \dots, T_0$, $T_0 = 96$. We tested for activation by a t-test. More specificly, the estimate of the activation level is given by

$$\hat{b}_i = \frac{1}{SSD_x} \sum_{t=1}^{T_0} Z_{it}(x_t - \bar{x}.), \quad SSD_x = \sum_{t=1}^{T_0} (x_t - \bar{x}.)^2,$$

and the variance of Z_{it} is estimated by

$$s_i^2 = \frac{1}{T_0 - 2} \sum_{t=1}^{T_0} (Z_{it} - \bar{Z}_{i\cdot} - \hat{b}_i x_t)^2 \sim \sigma_i^2 \chi^2(T_0 - 2)/(T_0 - 2).$$

Here $\chi^2(f)$ denotes the χ^2-distribution with f degrees of freedom. Then the statistic

$$X_i = \frac{\hat{b}_i}{\sqrt{s_i^2 / SSD_x}} \quad i = 1, \dots, V,$$

has a t-distribution with $T_0 - 2 = 94$ degrees of freedom, if the voxel is not activated. Since the degrees of freedom are quite large, it is reasonable to make the approximation that the variance estimates are exact, $s_{Y,i}^2 = s_i^2 = \sigma_i^2$, whence we get a normal distribution for X_i,

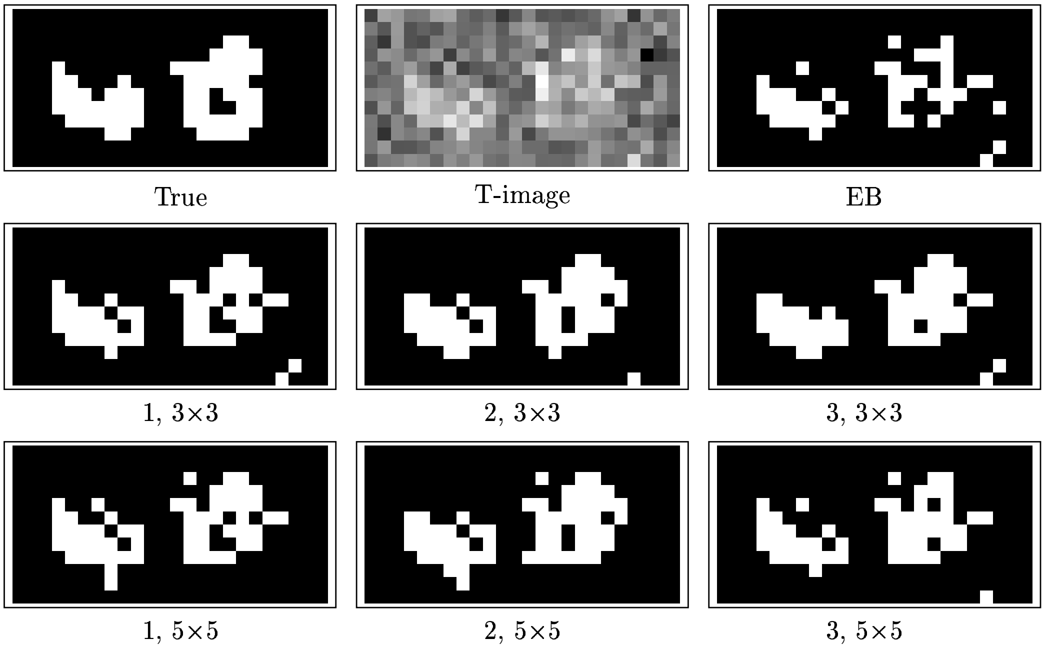
$$X_i \sim \begin{cases} N(\mu, 1), & \text{if } i \text{ is activated,} \\ N(0, 1), & \text{if } i \text{ is not activated,} \end{cases}$$

where $\mu = m\sqrt{SSD_x}$. The image of test statistics $\{X_i\}$ hence follows a mixture distribution, where the mean is positive when the voxel is activated and zero when not, and the setup is as in Section 2.1 with

$$p(x \mid A = 0) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}, \quad p(x \mid A = 1) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-\mu)^2}, \quad x \in \mathbb{R}.$$

We have assumed here that the temporal correlation is zero, which is necessarily an optimistic assumption. Temporal correlation will affect the variance of $\hat{b}_i$, but not the mean, and will lead to a higher variance of the statistic X_i, than stated above.

Figure 2 displays the image of t-statistics for the first of the four subdatasets. The posterior probability that a voxel is activated was calculated using the simple mixture model without spatial interaction, i.e. the setup of EB, and the models 1, 2 and 3. The image of posterior probabilities was thresholded at 0.5, which is a natural level when specifying a neutral balance between type I and II errors. The thresholded activation images are displayed in Figure 2. Clearly the spatial models (1, 2, 3) represent the true activation pattern much more closely than the simple mixture model. When using the latter, we effectively threshold the raw t-statistic image at a certain level, while at the spatial models we use information in neighbouring voxels, when classifying a voxel.



True T-image EB

1, 3×3 2, 3×3 3, 3×3

1, 5×5 2, 5×5 3, 5×5

Figure 2: Activation images for the first of four subsets of the synthetic data-set. Top left and middle: True binary activation image and observed t-statistics image. The remaining are thresholded posterior probability images for the different models. EB: Everitt and Bullmore's mixture model. 1, 2 and 3: Models 1, 2 and 3 defined on a 3×3 region or a 5×5 region. The images were thresholded at posterior probability 0.5.

In Table 2 the models are compared quantitatively by their ability to classify voxels correctly, and by the TPR at a given level of significance (FPR). The threshold was adjusted to yield an empirical FPR of 5% and 1% respectively in each image, and the TPR of this level was calculated. While the TPR estimates provide an idea of the strength of the classification test, they are mainly of theoretical interest, since the threshold used was calculated *given* the true activation pattern. On the contrary to this, the classification error measures reproducibility of the true pattern, when a practical and objective threshold is applied.

Table 2: Comparison of models for the synthetic fMRI data in Figure 2. From left to right are estimates of classification error for the thresholded images and TPR for images thresholded at a FPR of 5% and at 1% respectively. All figures are in percent. Standard errors of estimates, expressing the variability over the four sub-datasets, are given in parentheses.

Model	Class. error	TPR (level 5%)	TPR (level 1%)
EB	11.0 (0.7)	66.1 (2.3)	46.8 (5.0)
1, 3×3	6.3 (0.5)	88.3 (0.8)	65.7 (4.0)
1, 5×5	7.0 (0.3)	85.1 (2.0)	57.7 (6.3)
2, 3×3	6.3 (0.8)	90.7 (1.4)	72.5 (2.4)
2, 5×5	6.6 (0.8)	84.3 (2.9)	74.6 (3.5)
3, 3×3	6.3 (0.7)	87.5 (2.3)	66.5 (3.5)
3, 5×5	7.4 (0.3)	82.7 (3.0)	51.6 (7.6)

The table confirms the impression from Figure 2: The simple mixture model has the worst classification error and the lowest power. The three spatial models perform almost equally well, and a grid of 3 by 3 voxels gives the best result for this data. If the activated areas were larger than these, the 5 by 5 model might be more suitable, however this activation pattern seems reasonably representative for real data, and hence we recommend the 3 by 3 model to be used in practice. When considering the power, model 2 is slightly superior to the models 1 and 3, though this is not significant. Model 1 and 2 are furthermore preferable to model 3, since they have only 1 and 2 parameters respectively.

We may conclude that model 2 applied to a 3 by 3 neighbourhood is preferable in this situation: The statistical power is more than 90% at a significance level of 5%, and the mis-classification is reduced by more than 40% compared to the simple mixture model.

We will compare the performance of model 2 with a non-parametric model, where the activation is estimated by smoothing the data spatially with a a Gaussian kernel of full width at half maximum (FWHM) 2 and 3 voxels respectively, before calculating the t-statistic image. This is perhaps the most common way of including spatial information in the analysis of fMRI data, and usually the smooth t-image is thresholded using the random fields theory (Worsley *et al.*, 1996). Voxels may then be classified either on the basis of peak height or on cluster size. However, our aim here is *not* to compare results from thresholding based on random fields theory with that based on posterior probabilities. We think this is difficult, since the underlying principles and assumptions are fundamentally different. Rather we wish to compare the *estimates* of spatial activation pattern obtained by the two models. For this reason, we have thresholded the activation images in a comparable way, namely at the level which yields an actual FPR of 5% and 1% respectively, based on the true activation pattern. Figure 3 displays the estimated activation patterns.

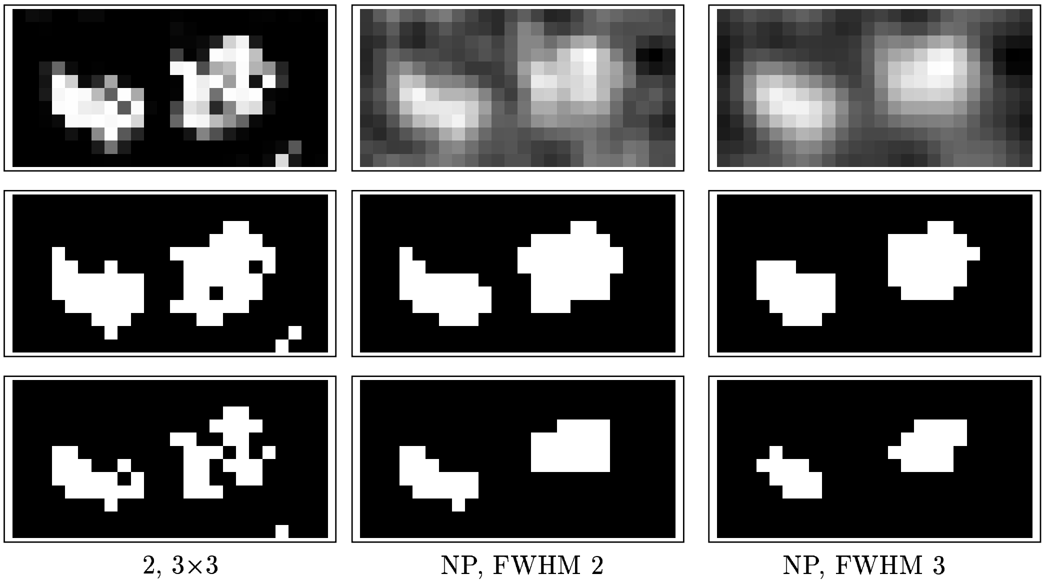


Figure 3: Activation images for the first of four subsets of the synthetic data-set. From left to right: Model 2 defined on a 3×3 region and the non-parametric model with FWHM 2 and 3 voxels respectively. Top row: Original activation images. Below: Images thresholded at empirical FPR 5% (middle) and 1% (bottom).

From the first row, we see that the distinction between noise and activation is dramatically different on the posterior probability scale compared to the t-image scale. EB made similar observations when comparing p-values and posterior probabilities. The two last rows show that the non-parametric model yields estimates which are smoother than the true regions, while the regions of model 2 are more irregular and have more holes. The estimated TPR for the non-parametric model are given in Table 3. By comparing this with Table 2, we see that model 2 reproduces the true activation best, as it has the highest TPR for each level of FPR. The difference is only significant for FWHM 3.

Table 3: Estimates of TPR for non-parametric activation images in Figure 3 thresholded at a FPR of 5% and at 1% respectively. All figures are in percent. Standard errors of estimates are given in parentheses.

Model	TPR (level 5%)	TPR (level 1%)
NP, FWHM 2	89.5 (2.0)	66.9 (2.8)
NP, FWHM 3	78.6 (3.4)	46.0 (6.7)

3.3 Example 3: Visual stimulation fMRI data

We finally considered a visual stimulation data set acquired with T_2^* weighted EPI on a 1.5 T scanner at the MR Research Centre, Aarhus University Hospital in Denmark. The data consist of 90 128×128 scans (5×1.875×1.875 mm voxels) for each slice, with a TR of 2 sec. 5 oblique slices were acquired in axial-coronal direction through the visual cortex. The stimulus was a 7Hz flashing light, which was presented in a blocked paradigm of 10 scans off, 10 scans on etc. starting an ending with an off-period. The first 5 scans were discarded, and we selected one of the slices for this analysis.

The scans were realigned by minimizing the squared distance of each scan to a reference scan under rotations and translations. Next we log-transformed the data and masked 4389 brain-voxels out. A linear model was fitted individually to each voxel time-series. The mean value space was spanned by a linear trend and a model for the haemodynamic response function given by a convolution of the paradigm with a Gaussian function with mean 6 sec. and variance 9 sec^2. The estimated activation amplitude was divided by its standard error to yield an image of t-statistics. The latter is displayed in the first panel in Figure 4.

We did not account for correlation in the time-series, whence we expect the variance of the statistics to be larger than the theoretical variance of the

t-distribution. We investigated the empirical distribution of the set $\{x_i\}$ of 4389 statistics, and found that a mixture of three components fitted well to this. Two of these were Gamma distributions, modelling respectively positive and negative BOLD effects, and one was a Normal distribution modelling the noise. The fitted density was

$$f(x) = p_0 f_N(x; 0, \sigma^2) + p_- f_\Gamma(-x; \lambda_-, \beta_-) + p_+ f_\Gamma(x; \lambda_+, \beta_+), \qquad (15)$$

where $f_N(\cdot; \mu, \sigma^2)$ denotes the density of a normal distribution with mean μ and variance $\sigma^2 > 0$, and $f_\Gamma(\cdot; \lambda, \beta)$ is the density of a Gamma distribution with mean λ/β and variance λ/β^2,

$$f_\Gamma(x; \lambda, \beta) = \frac{\beta^\lambda}{\Gamma(\lambda)} x^{\lambda-1} e^{-\beta x}, \quad x > 0,\ \lambda > 0, \beta > 0.$$

With the requirement that $p_0+p_++p_- = 1$, there are 7 free parameters, which were estimated by maximizing the likelihood function under the restriction that

$$E(X \mid X > 0) = \frac{\sum_{i=1}^V x_i 1(x_i > 0)}{\sum_{i=1}^V 1(x_i > 0)},$$

i.e. the mean of X given that it is positive, must equal the empirical mean of the positive x_i's. It is well known, that the likelihood function may be unbounded in mixture models, and the latter restriction was imposed to reduce the parameter space to finite likelihood-values. The estimates are given in Table 4.

Table 4: Parameter estimates for the distribution (15) of $\{x_i\}$.

$\hat{\sigma}$ =1.5160	$\hat{p}_+ = 0.0502$	$\hat{p}_- = 0.0081$
	$\hat{\lambda}_+ = 6.2349$	$\hat{\lambda}_- = 56.923$
	$\hat{\beta}_+ = 0.9433$	$\hat{\beta}_- =$10.2526

We are only interested in detecting positive activation in this example. Therefore we write $f(x)$ as

$$f(x) = (1 - p_+) f(x \mid A = 0) + p_+ f(x \mid A = 1),$$

where

$$f(x \mid A = 0) = \frac{p_0}{p_0 + p_-} f_N(x; 0, \sigma^2) + \frac{p_-}{p_0 + p_-} f_\Gamma(-x; \lambda_-, \beta_-)$$

is the null-distribution and

$$f(x \mid A = 1) = f_\Gamma(x; \lambda_+, \beta_+)$$

is the distribution of x, given that the voxel is positively activated.

The setup is hence as in Section 2.1, only here the null-distribution represents both no activation and negative BOLD effects. As an alternative to the Gamma distribution for positive activation, one could consider the sum of a Gamma and a Normal distribution, to account for the fact that the activation level is observed with noise. The density of the latter is, however, not available in closed form, and since the distributions are very similar at the present noise level, we have chosen a single Gamma.

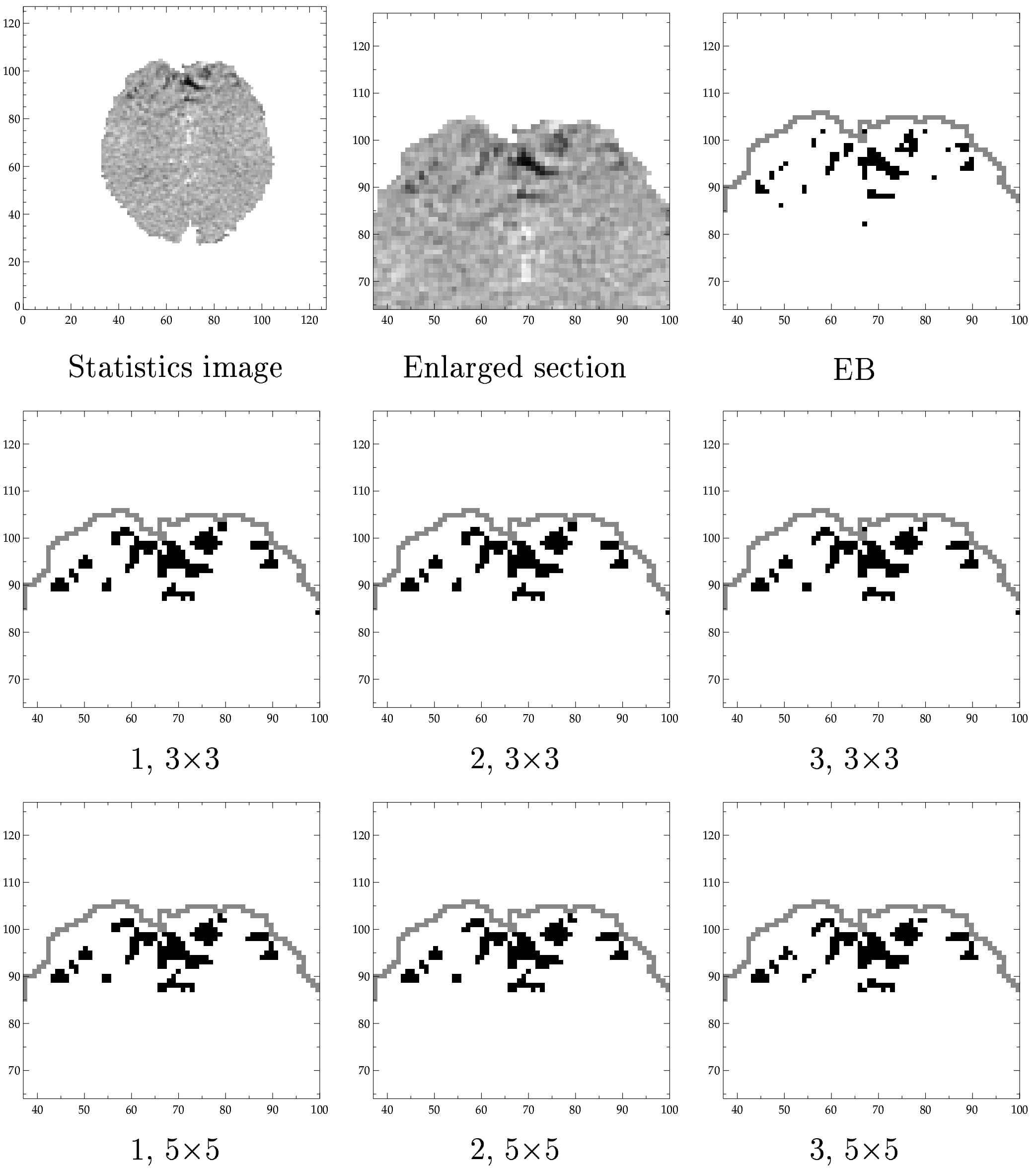
Figure 4 shows the image of statistics $\{x_i\}$ and enlarged sections of thresholded posterior probability maps for the non-spatial mixture model (EB), and for the different models in Section 2.2. The images were thresholded at 0.5. Like in the previous section, there is hardly any difference between the different spatial models, but there is a striking difference between the EB model and the others. In general the activated areas are larger with the spatial models and small (i.e. single-voxel) areas are suppressed. Clearly we can only speculate whether these estimates are closer to the truth or not. However, the simulated data of the previous section suggest that for activated areas of a certain size, the spatial model gives a significantly improved estimate. The idea that activation should have a certain spatial extent is the rationale behind spatial smoothing and other filtering techniques, and hence also this methodology.

In Figure 5 we have displayed the estimate, one gets by smoothing the original data before calculating the statistical image. We have no directly comparable way of thresholding this image, instead we have thresholded the image at three different levels. The mixture model estimates have some similarities with these activation patterns, but clearly the latter are much smoother. Again we can only speculate what is closest to the truth. It is, however, well known (Müller, 1988) that a kernel smoothing estimate will be biased, in the sense that the estimate will be smoother than the underlying signal. This is a likely explanation for the difference in smoothness.

4 Discussion

4.1 Conceptual summary

We have proposed a spatial mixture model for a statistical parametric map $\{x_i\}$. The idea is to model the distribution of x_i both when the voxel is



Statistics image

Enlarged section

EB

1, 3×3

2, 3×3

3, 3×3

1, 5×5

2, 5×5

3, 5×5

Figure 4: Comparison of estimated activation patterns for the different mixture models of the visual stimulation data. Top left and middle: Raw image of t-statistics and an enlarged section of this. The remaining panels are posterior probability images thresholded at 0.5. Top right: non-spatial mixture model. Middle and last row: Models 1, 2 and 3 defined on respectively a 3×3 voxel region (middle row) and 5×5 voxel region (last row.)

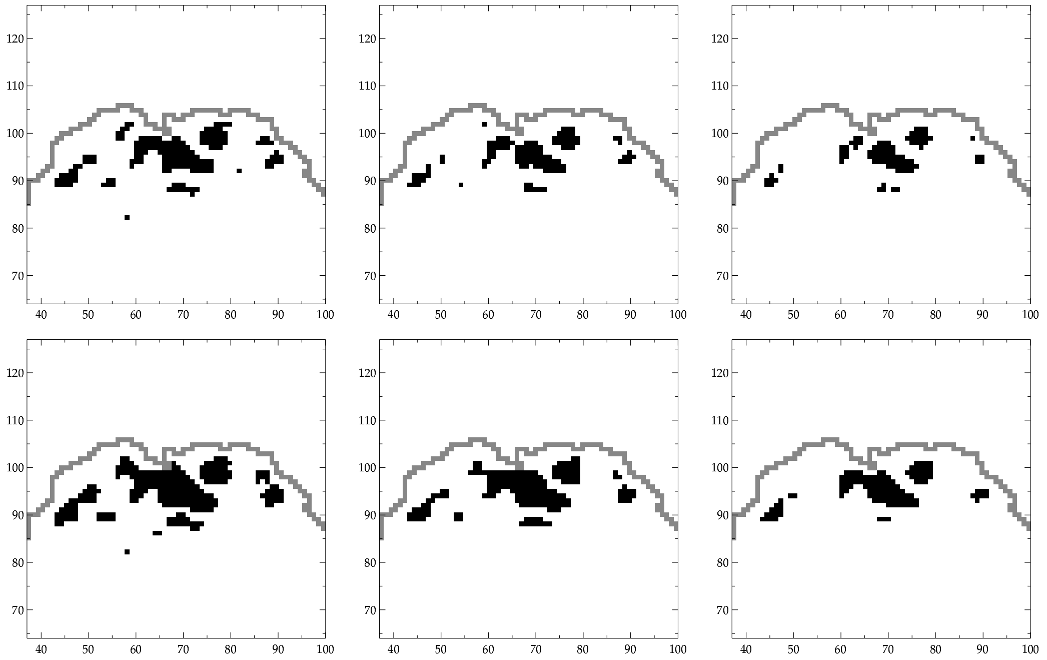


Figure 5: Images of t-statistics based on the visual stimulation data smoothed spatially with a Gaussian kernel of FHWM 2 voxels (top row) and 3 voxels (bottom row). The images are thresholded at 5.0 (left), 6.0 (middle) and 7.0 (right).

not active and when it is. Typically the non-activated distribution is known, this is the usual null-distribution of the SPM. The activation distribution might either be a simple non-central version of the null-distribution, as in Example 2, or a completely different distribution, which models the range of different activation strengths observed in the data, as in Example 3. The activation pattern is described by an unobserved volume of binary indicators, $\{A_i\}$. We suggest three different prior models for this pattern, which reflect the property that activation tends to occur in clusters rather than individual pixels. By formulating the models locally on a small region of pixels, it is possible to obtain a closed form expression for the posterior probability that a voxel is activated, given the values of the SPM in a region around the voxel.

To use the method in practice all one needs to specify are the two distributions of the test statistic. As in an ordinary analysis, the choice of test statistic influences the sensitivity, but there are no restrictions on the class of statistics that may be employed: The only requirement is that one can specify parametric distributions for the two activation states.

The proposed models account to some extent for the spatial structure of the underlying activation pattern. We found that the three different models worked almost equally well on synthetic and real fMRI data. In fact we tested 2 more advanced models also, but they gave similar results. (The models are described in a research report by the authors.) We recommend model 2 to be used in practice: It has only two parameters, with natural interpretations: One is p, the probability of a voxel being activated. An estimate of p is a global measure of the fraction of activated voxels, which is of interest in itself. The other is γ, which is a measure of the correlation of the true activation field. The parameters may easily be estimated directly from the data.

When only modelling a single slice with 1.9 mm voxels, we found that a 3×3 neighbourhood worked well. When a volume of slices is considered the neighbourhood could be extended with the two voxels directly below and above the centre or to a $3 \times 3 \times 3$ cube. This should of course depend on the interslice distance.

4.2 Comparison with existing methods

The methodology extends that of EB, who proposed a non-spatial mixture model. In fact the EB model is a special case of our analysis scheme, as it is contained in model 2. We found significant improvements in sensitivity on synthetic fMRI data compared to the non-spatial mixture model: The sensitivity increased from 66% to 91% at a FPR of 5%, and the mis-classification rate of the 0.5-thresholded images was reduced from 11% to 6%. The analysis of visual stimulation data indicated similar improvements.

When applied to synthetic fMRI data, our method was more sensitive than smoothing the data with a kernel of FWHM 3 voxels, but the sensitivity of the FWHM 2 smoothing was similar to ours. However the non-parametric smoothing model seems to produce estimates which are more smooth, than the ones obtained with our method. As mentioned in Example 3, this could be explained by the bias in the kernel smoothing estimate. One argument used for smoothing data is the Matched Filter Theorem (Rosenfeld and Kak, 1982). This states that in order to maximize signal-to-noise ratio at a specific point in an image, one should convolve the image with a kernel which has the same shape as the signal at that point. This is a statement about *detecting* a signal. When one wants to *estimate* the signal or some features of it, this is not necessarily an optimal strategy because of the bias introduced. On the contrary a parametric model, if correct, yields estimates which are less biased and more efficient. Clearly our model is not "correct", but we would like to emphasize the difference between parametric and non-parametric modelling. Furthermore the choice of the smoothing parameter, i.e. the FWHM of the kernel, is always a critical point in non-parametric estimation. It seems that for fMRI data, this parameter is often chosen in an *ad hoc* manner. With our method, the "smoothing parameter" (such as the parameter γ of model 2) is estimated directly from the data itself.

The assumptions underlying mixture modelling seem more natural and transparent to us, than those underlying the random fields theory. We expect a priori to find basically two different types of voxels, activated and non-activated, and a model for the data should reflect this. The inference in the model is fundamentally different from the usual hypothesis testing framework. In the latter, what is really an estimation problem, is answered by a hypothesis test (Worsley, 1997). The main problem is then the protection against false positives, with the large number of tests performed. In our approach we estimate the proportion of active voxels p, and use this to determine the posterior probability that a voxel is activated. As may be seen from (3) and (4) the probability that $A = 1$ tends to 0 as p tends to 0. This may be regarded as our way of handling multiple comparisons: If the size of the volume is increased, but the number of active voxels is fixed, p will decrease, and hence so will the posterior probability that a voxel is activated. For a fixed amount of activation, a larger search volume hence yields a more conservative analysis than a small.

Another advantage compared to the random fields framework is the robustness to misspecification of the model. To illustrate this, we replaced the normal distribution in Example 3 with a t-distribution with 20 degrees of freedom. The thresholded activation images were almost identical, with only a few voxels changing state. This is not surprising, since the two distributions are almost equivalent for our purposes. On the contrary, the random field theory relies on the extreme tail of the distribution, whence there is non-negligible difference between a $t(20)$-distribution and the normal distribution in this framework.

The method may be particularly relevant in applications where signal estimation is more important than signal detection. This is the case for instance when fMRI is used for pre-surgical planning, where the protection against false negatives is more important than false positives. Another example is when the results of an fMRI study are combined with data from other modalities, such as to regularize the inverse problem of MEG/EEG (Liu *et al.*, 1998).

During the review process of this paper, we realized that the idea of using local models for the true image in restoration problems is not new. Meloche and Zamar (1994) used an approach which is almost similar to ours, and they also derived moment estimators for parameters of the true image model. Meloche and Zamar considered a more general framework, where they estimated the probabilities $P(A_C = a_C)$ non-parametrically in a very elegant way. We restrict our attention to parametric models which are realistic from a brain imaging point of view, and this gives us the big advantage of being able to calculate the posterior probability in closed form. As mentioned earlier this point is crucial for the applicability of the method in practice. Furthermore Meloche and Zamar only consider models of the form $(x \mid A = 0) \sim N(0, \sigma^2)$ and $(x \mid A = 1) \sim N(1, \sigma^2)$, where our setup is completely general.

We have assumed throughout the paper that the observations are uncorrelated *given* the true activation pattern. Some spatial correlation can be detected in the noise in fMRI data, and hence this assumption will often be violated. The correlation of the signal is, however, much larger than that of the noise, and hence we have accounted for most of the correlation in the data by the model for the activation pattern. In some models, one may extend the methodology to correlated noise by estimating the spatial correlation first, and incorporating this in the expression for $f(x_C \mid a_C)$. Assuming stationarity of the correlation, this may be estimated from the residual time series, see for instance Hartvig (1999). Clearly the computations get more complicated then, as the closed form expression for the posterior probability is lost.

From a mathematical point of view, a natural question is whether there exist global models for the whole set of voxels, which have marginal distributions given by the models in this paper. This is in fact the case, since all three models have the property, that the structure of the model is maintained when reducing to marginal distributions. Considering model 2, for instance, this means that if we formulate the model on the whole set of voxels, the marginal distribution of a 3 by 3 region will be the same as that obtained by formulating the model on this region only. This also means that edge-effects may be handled in a rigorous way, by simply reducing the number of neighbours k, when calculating the probability of activation in boundary voxels.

5 Conclusion

We have formulated a simple mixture model for fMRI data which captures most of the spatial structure of the underlying activation pattern. The spatial model has two parameters, which are directly interpretable and may be estimated from the data. The expression for the posterior probability that a voxel is activated is given in closed form. Rather than the usual hypothesis testing, the focus of the method is estimation of the activation, which seems more natural in many applications.

In order to use this method, one needs only specify the null-distribution and the distribution of activated voxels. These can be any distributions. The resulting activation image is a posterior probability image, which may be thresholded in an intuitive way, without the need for correcting for multiple comparisons. Alternatively, one may display the un-thresholded probability map, which shows a clear distinction between estimated activation and baseline.

Acknowledgement

We would like to thank Hans Stødkilde-Jørgensen at the MR Research Centre, Aarhus University Hospital for sharing data and insight. We would also like to thank the anonymous reviewers for valuable suggestions.

A Appendix

We will derive the formulas for the posterior probability that a voxel is activated in model 2 and model 3 in the following.

A.1 Model 2

For given p and γ, q_0 is determined by

$$q_0 = 1 - \alpha \frac{(1+\gamma)^{k+1} - 1}{\gamma}.$$

Using the same technique as in model 1, we find

$$P(A=1|x_C) = \left\{ 1 + \frac{1}{v}\left[\gamma^{-1} + \frac{1 - \alpha(1+\gamma)^{k+1}/\gamma}{\alpha} \left(\prod_{j=1}^{k} (1+\gamma v^j) \right)^{-1} \right] \right\}^{-1}. \tag{16}$$

A.2 Model 3

In this model we have

$$1 = q_0 + q_1 + \frac{\alpha_1}{\gamma_1}\{(1+\gamma_1)^{k+1} - 1 - \gamma_1^{k+1}\} + \frac{\alpha_2}{\gamma_2^k}\{(1+\gamma_2)^{k+1} - 1 - \gamma_2^{k+1}\}, \tag{17}$$

$$p = q_1 + \alpha_1\{(1+\gamma_1)^k - \gamma_1^k\} + \frac{\alpha_2}{\gamma_2^{k-1}}\{(1+\gamma_2)^k - \gamma_2^k\},$$

where p is the probability of a voxel being activated. Instead of (16) we find

$$P(A=1|x_C) = \left\{ 1 + \frac{1}{v}\frac{N}{D} \right\}^{-1}, \tag{18}$$

where

$$N = \frac{\alpha_1}{\gamma_1}\prod_{j=1}^{k}(1+\gamma_1 v^j) + \frac{\alpha_2}{\gamma_2^k}\prod_{j=1}^{k}(1+\gamma_2 v^j) + q_0 - (\frac{\alpha_1}{\gamma_1} + \frac{\alpha_2}{\gamma_2^k}),$$

$$D = \alpha_1 \prod_{j=1}^{k}(1+\gamma_1 v^j) + \frac{\alpha_2}{\gamma_2^{k-1}}\prod_{j=1}^{k}(1+\gamma_2 v^j) + \{q_1 - (\alpha_1\gamma_1^k + \alpha_2\gamma_2)\}\prod_{j=1}^{k} v^j.$$

For model 3 the marginal density of x_{C_i}, used in the constrast function (8), is

$$\begin{aligned} f(x_{C_i};\phi,\psi) \\ = \prod_{j=0}^{k}\{f(x_i^j|0;\phi)\Bigg\{ &\frac{\alpha_1}{\gamma_1}\prod_{j=0}^{k}(1+\gamma_1 v_i^j(\phi)) + \frac{\alpha_2}{\gamma_2^k}\prod_{j=0}^{k}(1+\gamma_2 v_i^j(\phi)) \\ &+ q_0 - (\frac{\alpha_1}{\gamma_1} + \frac{\alpha_2}{\gamma_2^k}) + \{q_1 - (\alpha_1\gamma_1^k + \alpha_2\gamma_2)\}\prod_{j=0}^{k} v_i^j(\phi) \Bigg\}, \end{aligned} \tag{19}$$

with $\psi = (\alpha_1, \alpha_2, \gamma_1, \gamma_2, q_1)$ and q_0 given by the constraint in (17).

References

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

Besag, J. (1986) On the statistical analysis of dirty pictures. *J. R. Statist. Soc. Ser. B*, **48**, 259–302.

Bullmore, E., Brammer, M., Williams, S.C., Rabe-Hesketh, S., Janot, N., David, A., Mellers, J., Howard, R. and Sham, P. (1996) Statistical methods of estimation and inference for functional MR image analysis. *Magn. Reson. Med.*, **35**, 261–277.

Bullmore, E.T., Suckling, J., Overmeyer, S., Rabe-Hesketh, S., Taylor, E. and Brammer, M.J. (1999) Global, voxel, and cluster tests, by theory and permutation, for a difference between two groups of structural MR images of the brain. *IEEE Trans. Med. Imag.*, **18**, 32–42.

Cressie, N.A.C. (1991) *Statistics for spatial data.* Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics. New York: John Wiley & Sons, Inc.

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

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

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

Friston, K.J., Jezzard, P. and Turner, R. (1994) The analysis of functional MRI time-series. *Human Brain Mapping*, **1**, 153–171.

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

Greig, D.M., Porteous, B.T. and Seheult, A.H. (1989) Exact maximum a posteriori estimation for binary images. *J. R. Statist. Soc. Ser. B*, **51**, 271–279.

Hartvig, N.V. (1999) A stochastic geometry model for fMRI data. Research report 410, Department of Theoretical Statistics, University of Aarhus. *Submitted for publication.*

Heyde, C.C. (1997) *Quasi-likelihood and its application.* New York: Springer-Verlag. A general approach to optimal parameter estimation.

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

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

Liu, A.K., Belliveau, J.W. and Dale, A.M. (1998) Spatiotemporal imaging of human brain activity using functional MRI constrained magnetoencephalography data: Monte Carlo simulations. *Proc. Natl. Acad. Sci. USA*, **95**, 8945–8950.

Meloche, J. and Zamar, R.H. (1994) Binary-image restoration. *Canadian J. Statist.*, **22**, 335–355.

Müller, H.G. (1988) *Nonparametric Regression Analysis of Longitudinal Data.* Lecture Notes in Statistics. Springer-Verlag.

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

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

Press, W.H., Teukolsky, S.A., Vetterling, W.T. and Flannery, B.P. (1992) *Numerical Recipes in C.* Cambridge University Press, second edn.

Rosenfeld, A. and Kak, A.C. (1982) *Digital Picture Processing*, vol. 2. Academic Press, Orlando.

Worsley, K.J. (1997) Comment on "Non-linear Fourier time series analysis for human brain mapping by functional magnetic resonance imaging", by Lange and Zeger. *Appl. Statist.*, **46**, 25–26.

Worsley, K.J. and Friston, K.J. (1995) Analysis of fMRI time-series revisited — again. *NeuroImage*, **2**, 173–181.

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

Spatial deconvolution of the BOLD signal by a hierarchical model

Niels Væver Hartvig[*]

University of Aarhus

July 10, 2000

Abstract

We propose a hierarchical model for deconvolving the spatial haemodynamic effects in fMRI data. The activation on a neuronal level is modelled by a field of independent variables, which is smoothed by a kernel to represent the coupling between the neural activation and local changes in blood oxygenation. The smoothed image is finally overlaid by Gaussian noise to model an observed statistical parametric map (SPM). In this framework we may estimate the shape and width of the haemodynamic diffusion kernel directly from the data. The inference in the model is centered on simulation techniques, and we formulate a Markov chain Monte Carlo algorithm for simulating observations of the posterior distribution of the activation field. The model is fitted to visual stimulation data, and we illustrate how the estimated activation image is much more detailed than the usual estimate, obtained by smoothing the SPM.

Keywords: Functional magnetic resonance imaging; Bayesian deconvolution; Spatial model; Markov chain Monte Carlo; Haemodynamic response; Correlogram

1 Introduction

In functional magnetic resonance imaging (fMRI) the spatial and temporal properties of the haemodynamic effects as a function of neural activation have often been investigated for the purpose of including these in a model for the data. A common approach is to model the temporal response as a linear, stationary system (Lange and Zeger, 1997;

[*] Department of Theoretical Statistics, Department of Mathematical Sciences, University of Aarhus, Ny Munkegade, DK-8000 Aarhus C, Denmark. e-mail: vaever@imf.au.dk

Friston *et al.*, 1994; Cohen, 1997; Rajapakse *et al.*, 1998). Assuming that neural activation follows the stimulation function, the convolution of the latter with the impulse response function (or haemodynamic response function) then completely describes the temporal response. This seems to be reasonable in many experiments (Cohen, 1997; Dale and Buckner, 1997), but unreasonable in others (Glover, 1999; Vazquez and Noll, 1998; Friston *et al.*, 1998). In an analysis of this type, one effectively makes inference on the temporal activation pattern on a neural level, exploiting the simple model between neural and haemodynamic activation.

While temporal convolution models are widely applied, spatial models of the same flavour seem much rarer. A common assumption is that the haemodynamic effects are spatially more smooth and dispersed than neural activation; this is based on the fact that the haemodynamic effects occur in the surrounding veins, and will consequently be less localized. This assumption motivates spatial filtering of the data (Siegmund and Worsley, 1995; Worsley and Friston, 1995; Lowe and Sorenson, 1997) or explicit spatial models (Kiebel *et al.*, 2000; Descombes *et al.*, 1998a; Hartvig, 1999; Hartvig and Jensen, 2000). However, following the temporal deconvolution approaches, it seems relevant to try to solve the inverse problem of estimating localized neural activation directly, rather than the resulting haemodynamic effects. Basically this would yield activation maps that are "sharper" than the smoothed maps, and would—to the extent that the assumptions of the model hold—be interpretable on a neural level. Assuming additivity and stationarity of the haemodynamic response, this corresponds to performing a spatial deconvolution of the data. While this cannot be performed directly, due to the high noise level in the data, we may perform it indirectly via a Bayesian model.

This has previously been suggested by Descombes *et al.* (1998b), who modelled the activation pattern as smooth, coherent regions, with possible non-smooth boundaries, by a Markov random field model. The authors discussed the possible extention to a convolution model, where the haemodynamic effects are explicitly modelled, but noted that the choice of the convolution kernel was problematic. We will address this issue by deriving the shape and width of the kernel directly from the data. The kernel fits our data well, and the width corresponds to measurements of the extent of the haemodynamic response reported in the literature.

Bayesian models are widely applied in medical imaging, examples are models for simple Gamma camera images (Besag *et al.*, 1995), PET images (Higdon, 1998) and ultrasound images (Husby *et al.*, 1999). They provide a general framework for modelling unobserved or partially observed variables together with the stochastic process that generates the observed data. Let Γ denote the unobserved neuron activation field. This is assigned an *a priori* distribution, $p(\Gamma)$ which reflects the expected properties of the field, before observing the data. The prior may be "uninformative", in the sense that little subjective information is built into the model, or it may represent substantial prior knowledge about the activation area under study, for instance obtained from previous experiments on the same subject. The likelihood function $p(X \mid \Gamma)$ is the distribution of the data X given the unobserved field Γ, i.e. a model for the haemodynamic effects and the noise. The inference on Γ is then based on the *a posteriori* distribution

$$p(\Gamma \mid X) \propto p(X \mid \Gamma)p(\Gamma).$$

This is the probability distribution of the Γ-field given the data, which combines prior knowledge with the information contained in the data. The result of a Bayesian analysis is thus a whole distribution, which allows us to estimate different functions of interest, such as the mean activation field or the proportion of activated voxels, as well as assessing the uncertainty of the estimates. On a higher analysis level, we may quantify how well data support a specific neuroscientific hypothesis of interest simply by calculating the posterior probability for the corresponding event.

In this paper, we will assume *a priori* that the neuron field consists of independent variates, which may either be zero, or have a non-zero Gamma distributed value. The Gamma model is motivated by the simple idea that in a stimulation experiment, the distribution of neural activity over the brain may range from values very close zero to very large values. At the next level the haemodynamic diffusion effects are modelled by smoothing this field with a kernel. The kernel may be non-stationary and non-isotropic, allowing for incorporation of anatomical covariates which describe the local tissue properties. A model for a single regression image or statistical parametric map (SPM) is finally obtained by adding noise to the smoothed field.

The most critical assumption is the spatial additivity of the haemodynamic effects. This assumption is made implicitly in other models as well (Kiebel *et al.*, 2000; Hartvig, 1999), but it remains to be studied how good approximation it is to the true effects. However, the Bayesian convolution model is robust to minor non-additive effects, since these may be absorbed in the noise process.

The posterior distribution $p(\Gamma \mid X)$ is not directly assessible, since it is only known up to a constant of proportionality. Instead we may explore it by the simulation technique known as Markov chain Monte Carlo (MCMC) (Gilks *et al.*, 1996; Tierney, 1994). The idea of this technique is to simulate a Markov chain of observations, which asymptotically has the correct distribution. Originally invented in physics (Metropolis *et al.*, 1953), this methodology is becoming a major tool in statistics these years due to its flexibility and applicability, enabling the researcher to analyse very complex models. We construct an MCMC algorithm based on an auxiliary variable, which decorrelates the observed field X and the underlying neuron field Γ, enabling efficient simulation of the latter.

The construction of the model is very similar to that used by Wolpert and Ickstadt (1998), who used smoothed Gamma random fields as intensity measures for a Poisson process. They used the conjugacy of the Gamma and Poisson distributions to design an elegant algorithm for simulating the underlying Gamma field given the data. This is not possible in our case, but our decorrelation technique described above, is quite similar to the one they proposed.

The paper is organized as follows: We will first describe the model and discuss how to parametrize it in the stationary and possibly non-stationary case. Next we will give a brief introduction to MCMC methods, and describe the algorithm used for making inference in the model. The performance and limitations of the method are illustrated on a visual stimulation data set in Section 3, and finally we discuss our results in Section 4.

2 The spatial convolution model

2.1 The basic model

The model may be formulated as a spatio-temporal model, but for simplicity we will only consider a spatial model for a summary image (or volume) $X = \{X_i\}_{i \in V}$ in this paper. Typically this would be a regression image, or a statistical parametric map, of voxel-wise estimated activation levels with respect to a given haemodynamic response function. Here V denotes the set of brain voxels. At the top level of the model we have a field $\Gamma = \{\Gamma_i\}_{i \in V}$ of independent variates, which intuitively represent activation on a neural level. The independence assumption implies that we make no a priori assumptions of interaction between neurons in different voxels. We will assume a Gamma distribution for the activation level given that this is positive,

$$\Gamma_i \,|\, \Gamma_i > 0 \sim \Gamma(\alpha_+, \beta_+), \quad i \in V.$$

The Gamma distribution allows the levels of activation to vary from values close to zero to very large values. We will also allow "negative" activation,

$$|\Gamma_i| \,|\, \Gamma_i < 0 \sim \Gamma(\alpha_-, \beta_-), \quad i \in V.$$

Negative activation levels have no obvious neuroscientific interpretations, but we will include them as a way of modelling negative BOLD effects. We will assume that the level is either zero, positive or negative with the probabilities

$$P(\Gamma_i = 0) = p_{i0}, \quad P(\Gamma_i > 0) = p_{i+}, \quad P(\Gamma_i < 0) = p_{i-},$$

where $p_{i0} + p_{i-} + p_{i+} = 1$. In the case where these probabilities does not depend on i, we may interpret p_+ as the proportion of voxels with neural activation.

At the next level we will let $\Lambda = \{\Lambda_i\}$ represent the BOLD effect caused by the neural activation, that is a smoothed version of the Γ-field,

$$\Lambda_i = \sum_{j \in V} k_{ij} \Gamma_j, \quad i \in V,$$

where $k_{ij} = k(i, j)$ for a kernel k on $V \times V$. The kernel will be normalized such that $\sum_j k_{ij} = 1$ for all $i \in V$. Here we make the assumption of additivity, i.e. that haemodynamic effects from different voxels combine additively. As mentioned earlier this assumption is difficult to verify empirically, but is made for convenience here. The kernel need not be stationary, in the sense that k_{ij} only depends on (i,j) through $i-j$. In the case where local tissue characteristics are available we may hence incorporate this covariate information in the diffusion kernel.

Finally the observed image $X = \{X_i\}$ is modelled as Λ overlaid with Gaussian noise,

$$X = \Lambda + \varepsilon, \quad \varepsilon \sim N_{|V|}(0, \Sigma).$$

Here $\Sigma = \{\sigma_{ij}\}$ is a $|V| \times |V|$ covariance matrix. In the simple setting where Σ is a diagonal matrix, the X_i's are conditionally independent given Γ.

The moments of X are given by

$$EX_i = \sum_{j \in V} k_{ij} E(\Gamma_j), \quad i \in V, \tag{1}$$

$$\operatorname{cov}(X_i, X_l) = \sigma_{il} + \sum_{j \in V} k_{ij} k_{lj} \operatorname{var}(\Gamma_j) \quad i, l \in V. \tag{2}$$

Here

$$E(\Gamma_j) = p_{j+}\frac{\alpha_+}{\beta_+} - p_{j-}\frac{\alpha_-}{\beta_-}, \tag{3}$$

and

$$\operatorname{var}(\Gamma_j) = p_{j+}\frac{\alpha_+(1+\alpha_+)}{\beta_+^2} + p_{j-}\frac{\alpha_-(1+\alpha_-)}{\beta_-^2} - \left(p_{j+}\frac{\alpha_+}{\beta_+} - p_{j-}\frac{\alpha_-}{\beta_-}\right)^2. \tag{4}$$

Notice that a part of the correlation between the variables can be attributed to the noise, and a part to the underlying haemodynamic effects. In particular when Γ and the kernel are stationary, the covariance function consists of the kernel convolved with itself plus the noise covariance.

In the following, we will assume that Σ is diagonal, $\Sigma = \operatorname{diag}(\sigma_1^2, \ldots, \sigma_{|V|}^2)$. This is no restriction when the covariance matrix is known, or a good estimate of this is available. Suppose namely that Σ is a general covariance matrix with Cholesky decomposition $\Sigma = LL'$. The class of distributions given by the model is closed under linear transformations, and we can thus decorrelate the data by working with $\tilde{X} = L^{-1}X$ instead of X. Letting $K = \{k_{ij}\}$ denote the matrix of original kernel values, the model is then $\tilde{X} = \tilde{\Lambda} + \tilde{\varepsilon}$, where $\tilde{\varepsilon}$ are *iid.* standard normal variates and $\tilde{\Lambda} = L^{-1}K\Gamma$. The kernel matrix for the transformed model is thus changed to $L^{-1}K$ but the underlying random field Γ is the same, and inference on the latter can hence be based on the transformed data $\tilde{X}$.

2.2 Parameters and covariates

In order to complete the setup, we must specify the parameters of the Gamma field, the kernel k and the covariance matrix Σ. In general the noise covariance may be well estimated from the original set of scans, used to produce the regression image, and we will hence assume that Σ is known. We will denote the parameters of the Gamma field by θ.

When specifying parameters, we may take either a likelihood (or empirical Bayes) approach, considering the parameters as fixed unknown quantities, or a Bayesian approach, where the parameters are considered as unobserved random variables with a prior distribution. When uninformative priors are used, the two methods will lead to similar estimates, hence the main argument for using one method or the other is a philosophical one, interpreting probabilities either from a frequentist or from a subjective point of view. A practical advantage of the Bayesian methodology, however, is a computational one. When estimating parameters in the likelihood setup in the present case, one has to use MCMC maximum likehood estimation (Geyer, 1994), which requires both a good initial guess of the parameters to use for calculating an approximation to the likelihood function, as well as a maximization of the latter. Both of these operations involve separate numerical procedures, as well as the MCMC algorithm. When using the Bayesian strategy a single MCMC algorithm suffices to all inferential purposes. Furthermore inclusion of parameter uncertainty in the estimate of the activation field is straightforward in a Bayes setup, but more cumbersome in the likelihood setting.

We will assume a priori that the mean and variance of the Γ-variates have a log-normal distribution, with a high variance to express our lack of prior information about these. Specificly, we have set,

$$\log(\alpha_+/\beta_+) \sim N(0,\kappa^2), \quad \log(\alpha_+/\beta_+^2) \sim N(0,\kappa^2),$$

where $\kappa = 1000$, and similarly for the negative Gamma field. A more informative alternative would be to let the priors represent typical observed activation levels, i.e. in the order of 5%-10% of the baseline intensity.

The kernel k may be estimated from the theoretical covariance (2) of the X field, assuming stationarity of the latter. The covariance $\mathrm{cov}(X_i, X_{i+l})$ may then be estimated by the empirical covariance, known as the correlogram (Cressie, 1991),

$$\hat{C}(l) = \frac{1}{|V|}\sum_{i\in V_l}(X_i - \bar{X})(X_{i+l} - \bar{X}), \quad l \in \mathbb{Z}^3, \tag{5}$$

where $V_l = \{i \in V | i + l \in V\} = V \cap (V - l)$. Since the correlogram converges to the true covariance as the number of pixel tends to infinity, by equation (2) we have

$$\hat{C}(l) \simeq \hat{\sigma}_l + \mathrm{var}(\Gamma)\sum_{j\in V} k_{ij}k_{i+l,j}, \quad l \in \mathbb{Z}^3. \tag{6}$$

Here $\mathrm{var}(\Gamma) = \mathrm{var}(\Gamma_i)$ and $\sigma_l = \mathrm{cov}(\varepsilon_i, \varepsilon_{i+l})$, $i \in V$, denote the variances of the stationary processes. In the absence of anatomical information to use for modelling the kernel, we will assume a stationary and isotropic form

$$k_l^\tau \propto \exp\left\{-\left(\|l\|/\tau\right)^a\right\}, \quad l \in \mathbb{Z}^3, \text{ for } a = 1, 2,$$

corresponding to respectively an exponential or Gaussian kernel. The kernel width τ may be estimated by least squares, minimizing

$$\sum_{l \in L} \left(C_{\theta,\tau}(l) - \hat{C}(l) \right)^2,$$

where L is a set of lags in $\mathbb{Z}^3$, and $C_{\theta,\tau}(l)$ is the theoretical covariance at the given kernel width.

Unless we have specific covariate or prior information, it is natural to assume a stationary Γ-field, where $p_{i+} = p_+, p_{i-} = p_-$ for all $i \in V$. For given values of the kernel and the parameters of the Γ-field, we will let p_+ and p_- be given by a moment estimator. We will use the empirical mean and variance of the X-field, set this equal to the theoretical values in (1) and (2), and solve for p_+ and p_-, using the expressions (3) and (4), to obtain the estimates. In our application we have more than 4000 observation in an image, and hence the variance of the empirical moments of X will be quite small.

More generally, however, it is often the case that external information is available with the data, such as tissue classifiers obtained from high-resolution anatomical scans, quantitative or qualitative prior knowledge of the brain functions under study or other covariates which can potentially explain variations in the neural activation level. Denote these covariates by $\eta_{i1} \in \mathbb{R}^{d_1}$ for each voxel $i \in V$. They are naturally incorporated in the parameters for the neuronal activation field, to reflect a higher a priori expectation of activation. We may employ the framework of generalized linear models for this (McCullagh and Nelder, 1983) and obtain the model

$$\operatorname{logit} p_{i+} = \pi_+ + B_1^T \eta_{i1}, \quad B_1 \in \mathbb{R}^{d_1} \quad i \in V.$$

Here π_+ determines the overall probability of observing neural activation, and the covariates modify this level locally.

Similarly we may have covariates $\eta_{i2} \in \mathbb{R}^{d_2}$ which explain local changes in the haemodynamic response, an example of the latter is the position of the cortical surface, recently used by Kiebel *et al.* (2000). These may be included in the model for the kernel $k(i, j) = k(i, j; B_2, \eta_{i2})$ in an appropriate way; in the case of cortical surface information the kernel may be rotated with the tangential plane to the cortical surface, thereby modelling a haemodynamic effect that diffuses over the surface. Due to the complexity of the final model, inference on the hyperparameters $B = (B_1, B_2)$ are most easily performed by adopting a fully Bayesian paradigm, assuming prior distributions for these as well.

We will restrict ourselves to the stationary model in the following.

3 Inference by simulation

Let ξ denote the vector of all the unknown variables, that is the Γ-variates and the parameters of the prior θ. In order to make inference on the underlying neuron field, we wish to study the conditional distribution of ξ given X. This is not analytically tractable, hence we will do this by simulating a sequence of observations from it. The distribution cannot be simulated directly by simple techniques, both because of the dependency between Γ and the hyperparameters θ and because the data introduces correlation in the underlying Γ-field. Instead we may apply MCMC, where a Markov chain of variates $\xi^{(k)} \in V$, $k = 1, 2, \ldots$ is generated with a transition density which fulfils the so-called detailed balance condition (see e.g. Gilks *et al.*, 1996). Under mild regularity conditions, we then have that in the limit as k tends to infinity, $\xi^{(k)}$ will have a stationary distribution which equals the conditional distribution of ξ given X. In particular the ergodic theorem ensures that for any function $f(\cdot)$,

$$\frac{1}{K}\sum_{k=1}^{K} f(\xi^{(k)}) \to E(f(\xi) \,|\, X) \quad \text{as } K \to \infty.$$

The flexibility in choosing the function $f(\cdot)$ illustrates the power of the MCMC approach: We may let $f(\xi) = \Gamma$ to estimate the posterior mean of the neural activation field $E(\Gamma \,|\, X)$, and we may combine this with an estimate of $E(\Gamma^2 \,|\, X)$ to form an estimate of the posterior variance of Γ. The latter may be used to make significance statements about the observed activation. An alternative would be to let $f(\xi) = (\mathbf{1}(\Gamma_i > 0), i \in V)$, i.e. an indicator field for whether Γ_i is greater than 0. This would produce a map of the posterior probability that the voxel is activated. Thresholding the probability map at a natural level, typically 0.5, would produce a map of significantly activated voxels. Finally we may also study functions related to a specific neuroscientific hypothesis of interest, for instance that the activation in one area of the brain has a greater extent than in an other. By letting f be the indicator for this event, we may obtain a posterior probability which quantifies how well data support the hypothesis.

Also the posterior mean of θ may be of interest. For instance we may use the simulations of p_+ to estimate the global amount of activation, together with a credibility interval for this.

3.1 Metropolis-Hastings algorithm

We refer to Gilks *et al.* (1996), Geyer (1999) or Green (2000) for a general introduction to MCMC methods, but will briefly describe the structure of the algorithm in the following. We will use the Metropolis-Hastings algorithm, which generates the Markov chain $\xi^{(k)}$, $k = 1, 2, \ldots$ in the following way. At time k we propose a new value ξ' from $\xi^{(k)}$ by simulating an observation from a proposal distribution $q(\xi^{(k)}, \cdot)$. This move is accepted with a probability given by the Metropolis-Hasting ratio,

$$R = \min\left(1, \frac{p(\xi' \mid X)q(\xi', \xi^{(k)})}{p(\xi^{(k)} \mid X)q(\xi^{(k)}, \xi')}\right).$$

If the move is accepted, we let $\xi^{(k+1)} = \xi'$, if not we set $\xi^{(k+1)} = \xi^{(k)}$. This acceptance step ensures that the Markov chain has the distribution $p(\xi \mid X)$ as stationary distribution, irrespectively of the choice of the proposal distribution $q(\cdot,\cdot)$. Most commonly the proposal is chosen such that only one part of the vector ξ is updated at a time, and one then iterates between updating different components of the vector, either systematically or in random order. This is the case for the so-called Gibbs sampler, where the ith coordinate ξ_i is updated by proposing a new value from the conditional distribution of ξ_i given (ξ_{-i}, X), here $\xi_{-i} = (\xi_j, j \neq i)$. With this proposal, the Metropolis-Hastings ratio is always 1, so the move is always accepted. Below we will combine both Gibbs and Metropolis-Hastings updates.

3.2 Auxiliary variables

In the current setting, we need to propose updates of Γ given (θ, X). In principle one may perform Gibbs updates here, sweeping through the coordinates of V and simulating Γ_i conditionally on (Γ_{-i}, X, θ). The problem with this approach, however, is that the variables in the Gamma field may be very correlated, especially if the kernel k is wide. When only one variable is updated at a time, this may lead to an algorithm with very correlated samples, which in turn implies that the convergence to the stationary distribution is too slow. This is a well known problem with single-site updating which have also been described by Husby *et al.* (1999) and Higdon (1998).

This is the reason for introducing a so-called auxiliary variable Z which will be used in the simulations. We will simulate a Markov chain of (Γ, θ, Z) variables, which has stationary distribution $p(\Gamma, \theta, Z \mid X) = p(\Gamma, \theta \mid Z)p(Z \mid X)$. Note that (Γ, θ) will still have the correct limit distribution. The point in introducing Z, is that Γ and X are conditionally independent given Z, and the Γ-variables are independent given Z. This makes $p(\Gamma, \theta \mid Z)$ much easier to simulate from, than $p(\Gamma, \theta \mid X)$.

The auxiliary variable is given by $Z = \{Z_{ij}, i \in V, j \in R_i\}$, where $R_i = \{j \in V \mid k_{ij} > 0\}$. Conditionally on Γ, we will let these be independent and distributed as

$$Z_{ij} \mid \Gamma \sim N(k_{ij}\Gamma_j, k_{ij}\sigma_i^2), \quad i \in V, j \in R_i.$$

Setting $X_i = \sum_{j \in R_i} Z_{ij}$ yields the same model for X as before, but here X and Γ are conditionally independent given Z. This enables us to run a Gibbs sampler of (Z, Γ), where we recursively simulate $\Gamma \mid Z$ and $Z \mid (\Gamma, X)$. The two following results provides the conditional distributions needed for this.

Result 3.1 *The Γ_i's are conditionally independent given Z, and Γ_i has conditional density*

$$p(\gamma_i \mid Z) \propto p(\gamma_i) \exp\left\{ \Big(\sum_{j:i\in R_j} Z_{ji}/\sigma_j^2 \Big) \gamma_i - \frac{1}{2} \Big(\sum_{j:i\in R_j} k_{ji}/\sigma_j^2 \Big) \gamma_i^2 \right\}, \quad \gamma_i > 0. \qquad (7)$$

where $p(\gamma_i)$ is the prior distribution of Γ_i given by

$$p(\gamma_i) \propto 1(\gamma_i = 0)p_0 + p_+ 1(\gamma_i > 0) \frac{(\beta_+)^{\alpha_+}}{\Gamma(\alpha_+)} \gamma_i^{\alpha_+ - 1} e^{-\beta_+ \gamma_i}$$
$$+ p_- 1(\gamma_i < 0) \frac{(\beta_-)^{\alpha_-}}{\Gamma(\alpha_-)} (-\gamma_i)^{\alpha_- - 1} e^{\beta_- \gamma_i}$$

Result 3.2 *Let $Z_i = (Z_{ij})_{j\in R_i}$. Conditionally on X and Γ, $\{Z_i\}_{i\in V}$ are independent and Z_i has a multivariate normal distribution with $E(Z_{ij} \mid X, \Gamma) = k_{ij}(\Gamma_j + X_i - \Lambda_i)$, $\mathrm{var}(Z_{ij} \mid X, \Gamma) = \sigma_i^2 k_{ij}(1 - k_{ij})$ and $\mathrm{cov}(Z_{ij}, Z_{ik} \mid X, \Gamma) = -\sigma_i^2 k_{ij} k_{ik}$, $j \neq k$. The distribution is degenerated, since $\sum_{j\in R_i} Z_{ij} = X_i$.*

Both results may verified by direct calculations.

Remark 3.1 Notice that we assume that the kernel is normalized such that $\sum_{j\in V} k_{ij} = 1$ for all i. Thus edge-effects are treated by scaling the kernel near the edges of V. When $V \subseteq \mathbb{Z}^3$ we will typically define the kernel k on $V \times V$ by a normalized kernel $\tilde{k}$ on $\mathbb{Z}^3 \times \mathbb{Z}^3$, letting $k_{ij} = \tilde{k}_{ij}/n_i$ for $i, j \in V$, where $n_i = \sum_{j\in V} \tilde{k}_{ij}$. When implementing the algorithm, the auxiliary variables Z_{ij} are then most naturally defined on the extended neighbourhood $j \in \tilde{R}_i = \{j \in \mathbb{Z}^3 \mid \tilde{k}_{ij} > 0\}$ since the dimension and conditional covariance of the vector Z_i is then the same for all $i \in V$. Thus we let $Z_{ij} \mid \Gamma \sim N(k_{ij}\Gamma_j, \tilde{k}_{ij}\sigma_i^2)$ for $j \in \tilde{R}_i \cap V$ and $Z_{ij} \mid \Gamma \sim N(0, \tilde{k}_{ij}\sigma_i^2)$ for $j \in \tilde{R}_i \setminus V$. The theorems above apply in this situation also, with slight modifications: In (7) σ_j^2 should be replaced by $n_j\sigma_j^2$. In the variance and covariance formulas in Result 3.2, k_{ij} should be replaced by $\tilde{k}_{ij}$, and the conditional mean should be replaced by $E(Z_{ij} \mid X, \Gamma) = k_{ij}\Gamma_j + \tilde{k}_{ij}(X_i - \Lambda_i)$ if $j \in V$ and $E(Z_{ij} \mid X, \Gamma) = \tilde{k}_{ij}(X_i - \Lambda_i)$ if $j \notin V$.

3.3 Proposal moves

For a given Z, we will in the following let b_i and c_i denote parameters of the distribution (7),

$$b_i = \sum_{j:i\in R_j} Z_{ji}/\sigma_j^2, \qquad c_i^{-2} = \Big(\sum_{j:i\in R_j} k_{ji}/\sigma_j^2 \Big).$$

We propose the following moves in the algorithm.

- Update of Z given (Γ, X). This is a Gibbs update where for each $i \in V$, we simulate $Z_i = (Z_{ij}, j \in \tilde{R}_i)$ directly from its conditional distribution as given in Result 3.2.

- Update of Γ given (Z, θ). The coordinates are independent by Result 3.1. Each coordinate will by updated by one of the following three moves:

  - If $\gamma_i > 0$: With probability $\exp(-(c_i b_i)^2/2)$ we propose a move to 0, denoted move 1. The reverse move from $\gamma_i = 0$ to a positive value is defined below. Together they yield the Metropolis-Hastings ratio
  $$R_1(\gamma_i, 0) = \frac{p_0 \Gamma(\alpha_+)}{p_+ {\beta_+}^{\alpha_+} \gamma_i^{\alpha_+ - 1} \exp(-\beta_+ \gamma_i) \sqrt{2\pi} c_i}. \qquad (8)$$
  With probability $1 - \exp(-(c_i b_i)^2/2)$ we propose a move of type 3 to a new positive value by simulating from the conditional distribution $p(\gamma_i \mid Z, \theta, \Gamma_i > 0)$. The latter has distribution
  $$p(\gamma_i \mid Z, \theta, \Gamma_i > 0) \propto \gamma_i^{\alpha_+ - 1} \exp\left((b_i - \beta_+)\gamma_i - \frac{1}{2c_i^2}\gamma_i^2\right), \quad \gamma_i > 0. \qquad (9)$$
  This last move is a Gibbs update, which has Metropolis-Hastings ratio 1.
  - If $\gamma_i < 0$ we propose similar moves as above: With probability $\exp(-(c_i b_i)^2/2)$ we propose a move to 0, denoted move 2. The reverse move from 0 to $\gamma_i < 0$ is described below. The Metropolis-Hastings ratio is
  $$R_2(\gamma_i, 0) = \frac{p_0 \Gamma(\alpha_-)}{p_- {\beta_-}^{\alpha_-} (-\gamma_i)^{\alpha_- - 1} \exp(\beta_- \gamma_i) \sqrt{2\pi} c_i}. \qquad (10)$$
  Else we propose a move of type 4 to a negative value by a Gibbs update,
  $$p(\gamma_i \mid Z, \theta, \Gamma_i < 0) \propto (-\gamma_i)^{\alpha_- - 1} \exp\left((b_i + \beta_-)\gamma_i - \frac{1}{2c_i^2}\gamma_i^2\right), \quad \gamma_i < 0.$$

  - If $\gamma_i = 0$: With probability $\Phi(c_i b_i)$ we propose a move of type 1 to $\gamma_i > 0$ and with probability $\Phi(-c_i b_i)$ a move of type 2 to $\gamma_i < 0$. Here $\Phi(\cdot)$ denotes the standard normal distribution function. When a move to a positive value is proposed, we simulate the new value γ_i' from a truncated normal,
  $$q(0, \gamma_i') = \frac{1}{\Phi(c_i b_i)\sqrt{2\pi} c_i} \exp\left(-\frac{1}{2c_i^2}(\gamma_i' - c_i^2 b_i)^2\right), \gamma_i' > 0.$$
  The Metropolis-Hastings ratio $R_1(0, \gamma_i')$ is the inverse of (8), with γ_i' inserted instead of γ_i. The type 2 move is similar, with
  $$q(0, \gamma_i') = \frac{1}{\Phi(-c_i b_i)\sqrt{2\pi} c_i} \exp\left(-\frac{1}{2c_i^2}(\gamma_i' - c_i^2 b_i)^2\right), \gamma_i' < 0,$$

and Metropolis-Hastings ratio $R_2(0, \gamma_i')$ equal to the inverse of (10) with γ_i' inserted instead of γ_i.

- Update of θ given Γ.
  - We propose randomly to update either the mean α_+/β_+ or the variance α_+/β_+^2 of the positive Γ-variates, or one of these two parameters for the negative Γ-variates. Let φ denote the parameter chosen. The new value φ' is proposed as $\varphi + \epsilon$, $\epsilon \sim N(0, \tau_\varphi)$, and we let θ' be the new value of θ obtained by recalculating p_+ and p_- for the new value of φ. The Metropolis-Hastings ratio for accepting the move is

    $$R_4 = \frac{p(\Gamma \mid \theta')p(\theta')}{p(\Gamma \mid \theta)p(\theta)},$$

    where $p(\theta)$ is the prior for θ. The variance τ_φ may depend on which of the four parameters that is chosen for updating.

The transitions from $\Gamma_i = 0$ to $\Gamma_i > 0$, for instance, are effectively dimension changing steps (Green 1995). Notice that we need to simulate from several different types of distributions in the proposals above. Algorithms for the normal, Gamma and truncated normal distributions may be found in Ripley (1987) or Devroye (1986). When updating Γ given Z, we have to draw random samples from the unnormalized density in (9). For $\alpha = 1$ this is a normal distribution truncated below at 0, but for other values it is not a standard distribution and the shape of the density depends very much on the parameters, in the same way that the Gamma distribution depends on the shape parameter. Since we must draw millions of samples from this density in a typical analysis, we need an efficient simulation algorithm. For this purpose we have designed a rejection sampling scheme (see e.g. Ripley, 1987), where a range of different envelopes are employed to ensure efficient sampling for all parameter values. The details of the algorithm are described in the separate paper Hartvig (2000), where also asymptotical optimality properties of the algorithm is studied.

3.4 Assessing convergence

As mentioned earlier the Markov chain is only asymptotically stationary and an important question when using MCMC is to assess how many simulations is needed before the chain has reached convergence. Though there exists several diagnostics for this (Gilks *et al.*, 1996), the only fail-proof method to detect if the chain has reached stationarity, is the so-called perfect simulation (Propp and Wilson, 1996). In practice the simplest diagnostic tool is to monitor time series of the simulations, and visually determine whether these have reached a stationary distribution or not. Also one may start the chain in different initial states and monitor the time series to see when the effect of the initial state has disappeared. In the same manner we may use the ergodic theorem and consider the sequential averages of a function $f(\xi)$,

$$\frac{1}{K}\sum_{k=1}^{K} f(\xi^{(k)}) \quad K = 1, \ldots, N. \tag{11}$$

By plotting the average as a function of K for different initial states of the Markov chain, we may detect when the effect of the initial state on the estimate is negligible.

4 Results

4.1 Visual stimulation data

We will apply the model to a visual stimulation data set. The data consist of 90 sets of MRI scans acquired with echo-planar imaging, each set consisting of 5 oblique slices in axial-coronal direction. The interscan time is 2 seconds. One scan consists of 128 by 128 voxels each covering a physical region of $1.875 \times 1.875 \times 5$ mm^3 and the interslice distance is 2.5 mm. The scans were acquired during a visual stimulation task, where the subject was watching a 7 Hz flashing light. The stimulus was presented in a periodical on-off paradigm with 10 scans off, 10 scans on etc., starting and ending with 10 scans of off-period. The magnetization level had stabilized after the first 5 scans, and these first scane were discarded from the analysis. We selected one slice covering the visual cortex for this analysis.

4.2 Calculation of activation image

We preprocess the data to obtain a single activation image to be analysed with the model. First we realign the images, by minimizing the L^2 distance between each scan and a reference scan under rigid transformations in the plane. Roughly 4500 voxels corresponding to brain tissue are masked out, and is considered as the set V. After log-transforming the data, we estimate the level of activation individually in each voxel, by fitting a linear normal model to the time series, where the mean value space is spanned by a constant term, a linear trend term and a model for the temporal haemodynamic response. The latter is a convolution of the paradigm and a Gaussian function with delay 6 sec. and variance 9 sec.^{2} to model the delay and dispersion of the response. The temporal correlation in the model was modelled by a first order autoregressive model. We scale the estimated regression coefficient of the haemodynamic response function to have unit variance and consider this as the level of activation in the voxel.

The above procedure produces the image $X = \{X_i\}_{i \in V}$ of voxel-wise estimated activation levels, to be analyzed by the model. An image of X may be seen in Figure 6.

4.3 Estimation of kernel width

We will ignore edge effects in the estimation, and assume a stationary kernel $k_{ij} = k_{i-j}$. We used the residual time series, i.e. the original time series minus the fitted mean values, to estimate the spatial covariance of the noise, $\{\hat{\sigma}_l, l \in \mathbb{Z}^2\}$, assuming stationarity. This was estimated by the correlogram (5) based on the residual scans. The correlogram was calculated scan by scan and averaged over the 85 scans, as we found almost identical covariance structures of the noise in all scans. Likewise the correlogram $\hat{C}(l)$ of the activation map X was calculated to estimate the covariance caused by the haemodynamic effects. In Figure 1 is a plot of $\hat{\sigma}_l$ and $\hat{C}(l) - \hat{\sigma}_l$ for different directions in $\mathbb{Z}^2$. The noise is almost uncorrelated for lags larger than 6 mm (3 voxels). The covariance caused by the haemodynamic response has much greater extent, especially along the direction with angle 0, which is the horizontal direction in the images in Figure 6. It is not evident what causes this anisotropy, but symmetries in the activation pattern in the two hemispheres of the brain is a likely explanation.

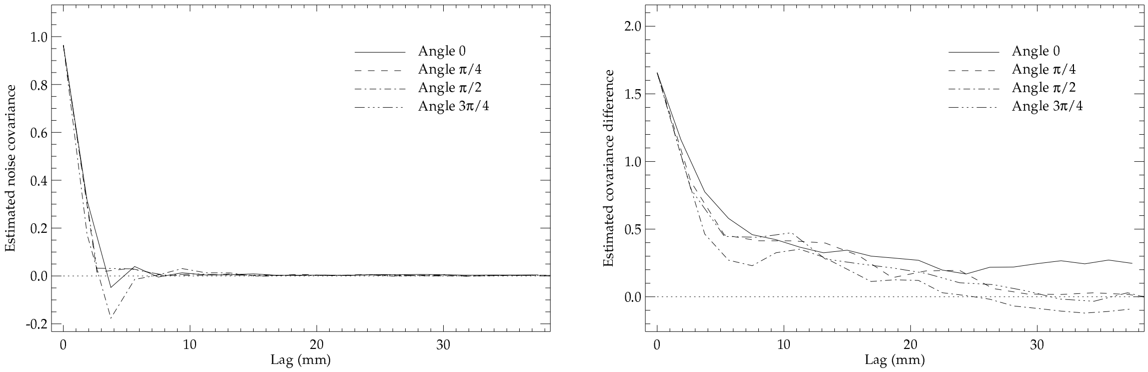


Figure 1: Left: Plot of the estimated noise covariance $\hat{\sigma}_l$ along 4 equiangular directions in the scan. Right: Same as left, but for the estimated covariance difference $\hat{C}(l) - \hat{\sigma}_l$.

Another distinctive feature of the covariance function is a “hump” at around lag 10 mm. This suggests, that the covariance can be decomposed into two parts, a quickly decaying part near 0, and a more flat part for greater lags. Most likely the flat part corresponds to global structure in the data, either large scale haemodynamic effects or neuronal activation patterns with a spatial structure, while the quickly decaying part may be attributed to local haemodynamic effects. Since we are only modelling local haemodynamic effects with the present model, we are most interested in the steep part of the covariance function.

Previously Matérn (1960) modelled similar covariograms by a mixture of two components, and though he was studying area distributions and volumes of trees, the idea is very natural in our framework also: Suppose we still have a stationary model, but rather than one underlying Gamma field, we have two, Γ^1 and Γ^2 independent of each other. The model is then

$$X = \Lambda^1 + \Lambda^2 + \varepsilon, \quad \text{where } \Lambda_i^p = \sum_{j \in V} k_{i-j}^{\tau_p} \Gamma_j^p, \;\; p = 1, 2. \tag{12}$$

Here the two kernel widths τ_1 and τ_2 represent two different types of haemodynamic responses. The covariance is then

$$C_{(\tau_1,\tau_2)}(l) = \operatorname{cov}(X_i, X_{i+l}) = \sigma_l + \sum_{p=1}^{2} \left\{ \operatorname{var}(\Gamma_i^p) \sum_{j \in \mathbb{Z}^2} k_j^{\tau_p} k_{j+l}^{\tau_p} \right\}, i \in V, l \in \mathbb{Z}^2. \tag{13}$$

Thus the covariance is a sum of two parts, each given by the convolution of a kernel with itself. Notice that by construction, this is a valid covariance function (i.e. positive definite) for all kernels k. Following the procedure described earlier, we fitted this model to the empirical covariance difference $\hat{C}(l) - \hat{\sigma}_l$, modelling the kernels as

$$k_l^\tau \propto \exp\left\{- \left(\|l\|/\tau\right)^a\right\}, \quad l \in \mathbb{Z}^3, \text{ for } a = 1, 2,$$

corresponding to respectively exponential or Gaussian kernels. The parameters (τ_1, τ_2) were estimated by least squares, minimizing

$$\sum_{l \in L} \left(C_{(\tau_1,\tau_2)}(l) - \hat{C}(l) \right)^2,$$

where L is a set of lags in $\mathbb{Z}^2$. The maximal lag used was 20 voxels.

In Figure 2 is a plot of the empirical covariances with the fit of the exponential and Gaussian kernels. The estimated parameters is listed in Table 1. The two models fit almost equally well as measured by the L^2 distance above, with the Gaussian kernels giving a slightly better fit. However it is clear from the plots, that the exponential kernels gives a better fit for small lags, and since it is important to get a good fit near the origin, we will use the exponential model.

Table 1: Estimated kernel widths.

Exponential		Gaussian	
$\hat{\tau}_1$	1.057 mm	$\hat{\tau}_1$	1.812 mm
$\hat{\tau}_2$	7.191 mm	$\hat{\tau}_2$	14.623 mm

We will only consider inference for the activation field corresponding to the narrow kernel, hence we will use the width $\tau = 1.057$ mm in the model. We will return to

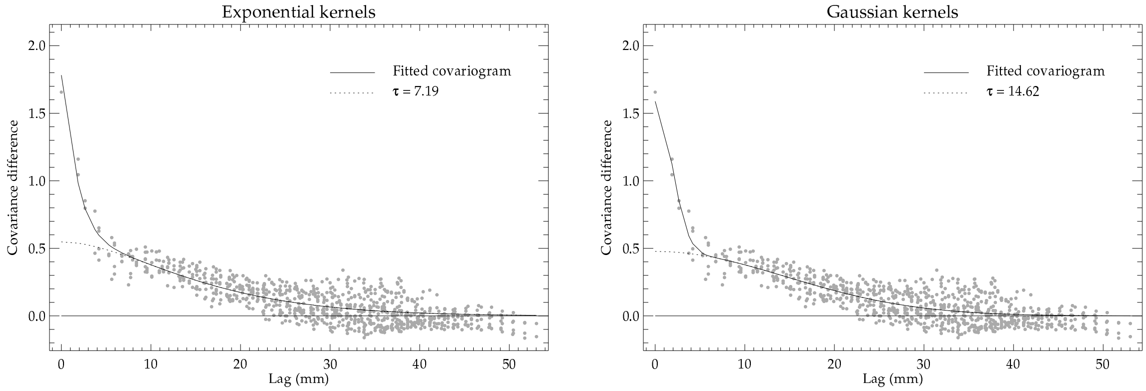


Figure 2: Fitted covariograms in the form of (13). Displayed is the values of the empirical covariogram for a range of lags in $\mathbb{Z}^2$, with the fitted covariogram, i.e. the sum of the two kernels convolved with themselves (full line) and the widest of the two kernels convolved with itself (dotted line). Left: Exponential kernels. Right: Gaussian kernels.

the point of making simultaneous inference on both kernels in the discussion. We may note that with this kernel, the modelled haemodynamic effects spread over an area of diameter about 6 mm, which corresponds well to figures reported in the literature.

4.4 Estimation of neuronal activation field

We considered the model with both positive and negative Gamma fields, with unknown shape and scale parameters. For computational simplicity, and given the short correlation length of the noise field, we chose a model with independent noise, where the noise variance was held fixed at the theoretical value $\sigma = 1.0$.

We iterated the MCMC algorithm 100000 times, at each step choosing randomly to update either the Gamma field, the matrix of auxiliary variables or one of the hyperparameters. We subsampled every 100'th observation, to obtain a time series of 1000 values. In Figure 3 is a plot of respectively the sum over all brain voxels of the absolute value of the Gamma variates and the number of positive variates. The plots indicate that the subsampled Markov chain stabilizes after about 100 iterations, and the remaining 900 samples are hence considered as simulations from the stationary distribution. The two variables are plotted together in Figure 4. This plot displays the same features as Figure 3, with an initial burn-in period and a stationary "cloud" of points. The high degree of posterior correlation between the variables is also visible. In order to assure that the initial state had no effect on the stationary level of the chain, we started the chain in two different states. Figure 5 shows the plot of the sequential average (11) of the mean of the positive Gamma variables, α_+/β_+. The averages converge to the same value despite the starting point, as predicted by the ergodic theorem. This supports the assumption that the chain has converged at the present number of iterations.

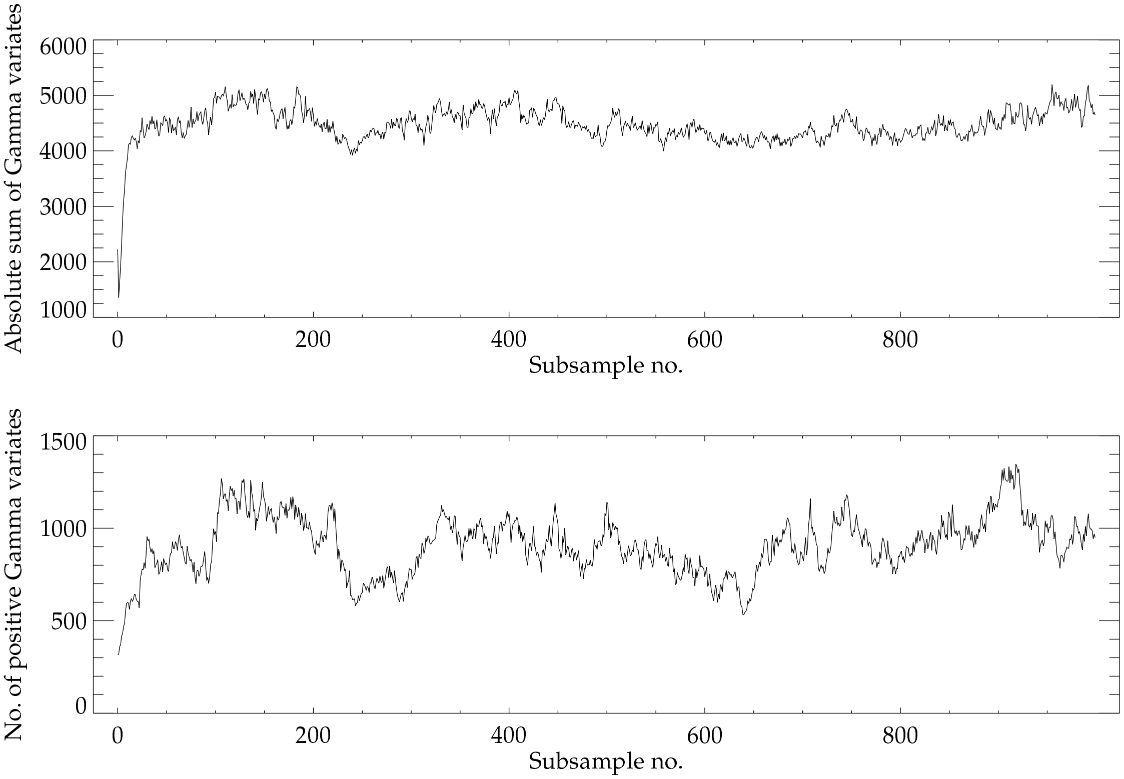


Figure 3: Time series of 1000 subsamples from the MCMC simulations.

Figure 6 displays the raw activation image X, together with posterior mean images of the Λ-field, that is the estimate of the haemodynamic activation pattern, and the neuronal activation field Γ. While the estimate of Λ is clearly much less noisy than the raw image X, the spatial resolution of the two are the same, since the data are not smoothed spatially. Visually this means, that the activation pattern is not smoothed out in the Λ image. The Γ-field may be considered as a denoised and decorrelated version of the original image. The figure illustrates how the spatial extent of the estimated activation foci are reduced and the activation pattern is sharpened in the Γ-image, compared to the Λ image.

For comparison Figure 7 displays activation images obtained by smoothing the data with a Gaussian kernel of FWHM 2 and 3 voxels, before the activation image is calculated. The latter corresponds to the usual SPM's. The estimated activation pattern is much smoother than both the posterior mean estimate of the haemodynamic and neuronal activation fields in Figure 6. In particular the images illustrate how larger areas of activation in the smoothed maps may correspond to a single or just a few active voxels

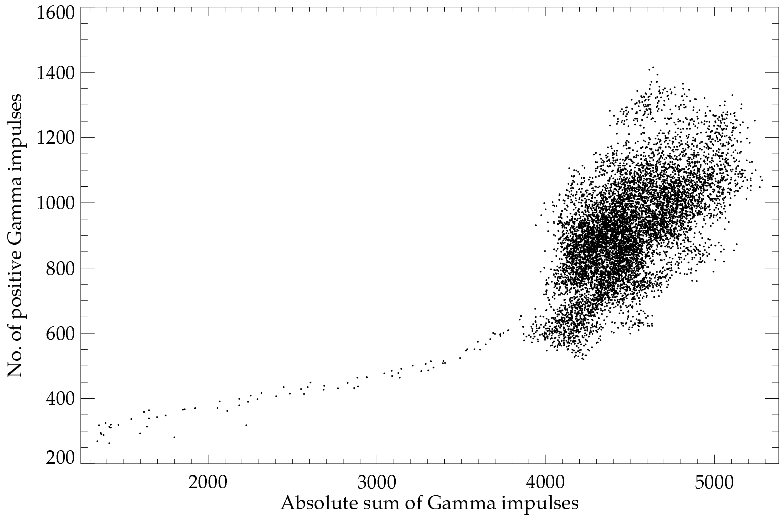


Figure 4: Plot of the number of positive Gamma variates against the absolute sum of the Gamma variates in the MCMC simulation.

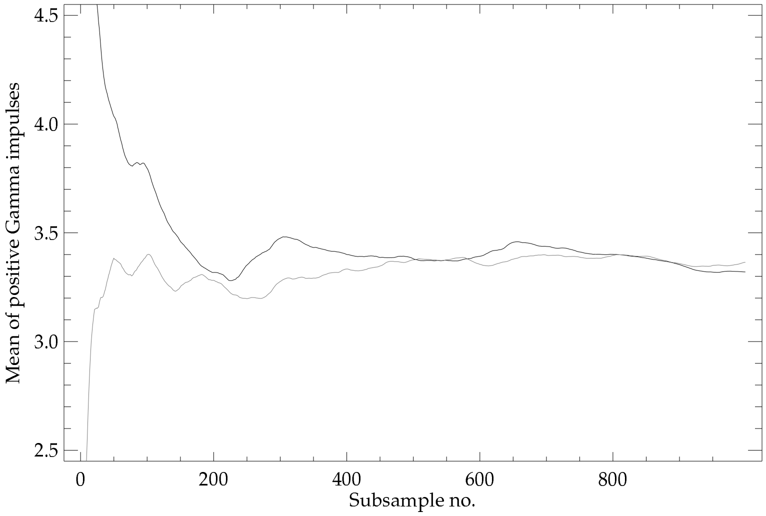


Figure 5: Plot of the sequential average (11) of the mean of the positive Gamma distribution. The two lines correspond to two runs of the MCMC algorithm with different initial values.

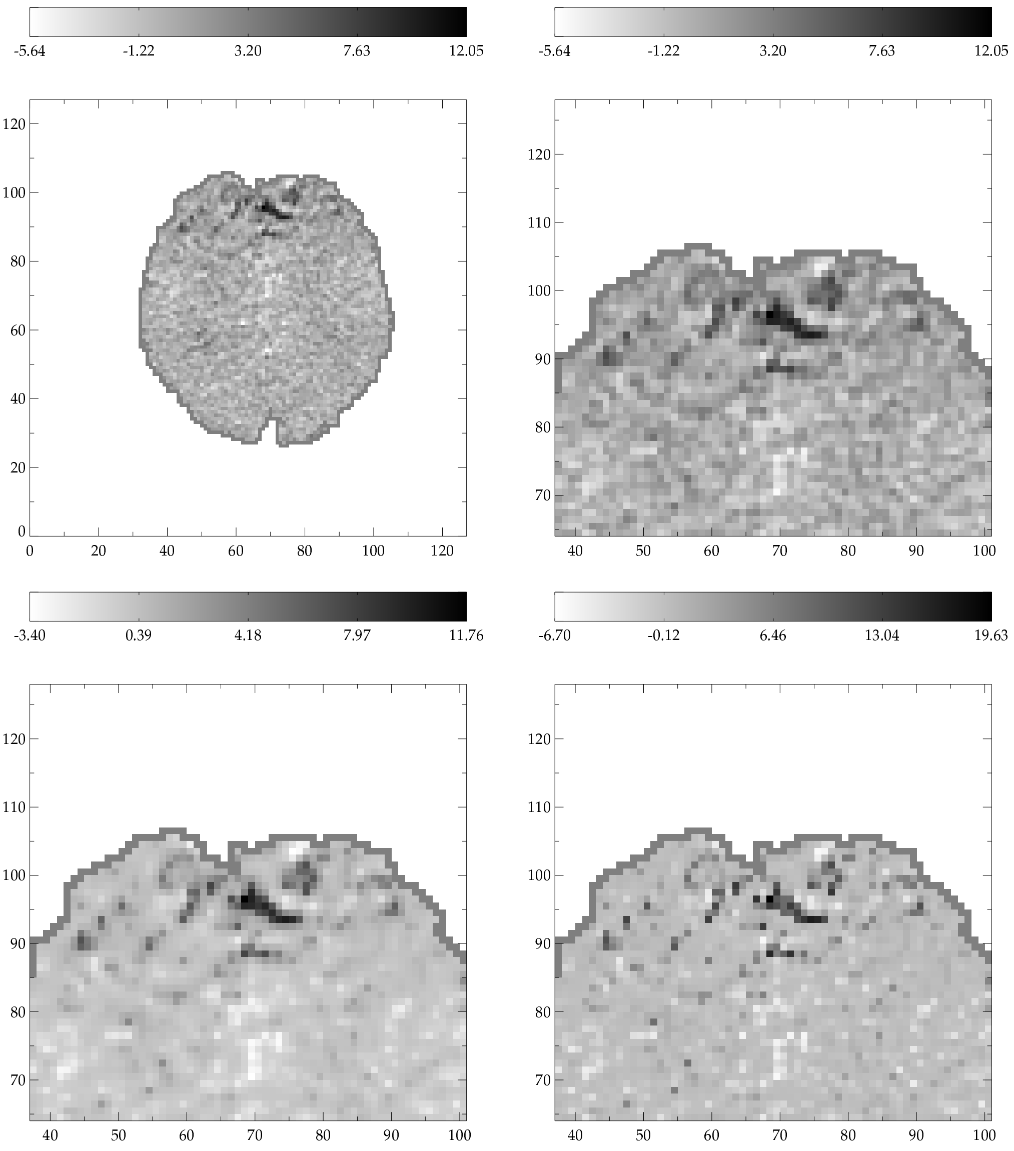


Figure 6: Top left: The image of voxel-wise estimated activation levels $\{X_i\}$. Top Right: An enlarged 64×64 subregion covering the visual cortex. Bottom: Estimated activation patterns. Left: Posterior mean of the smoothed activation field, Λ, based on 100000 simulations from the MCMC algorithm. Right: Posterior mean of the uncorrelated activation field, Γ. The left image may be interpreted as a haemodynamic activation image, while the Γ-image is a decorrelated version of this, which may be interpreted on a neuronal level.

in the Γ-activation map. Clearly, we can only speculate how the "true" pattern looks in this case. The argument for smoothing the data is based detecting the signal with optimal sensitivity, however when the aim is to estimate the signal, there is no general statistical argument for smoothing the data. On the contrary it is well known (Müller, 1988) that a non-parametric smoothing estimate produces a biased estimate, which is more smooth than the true field.

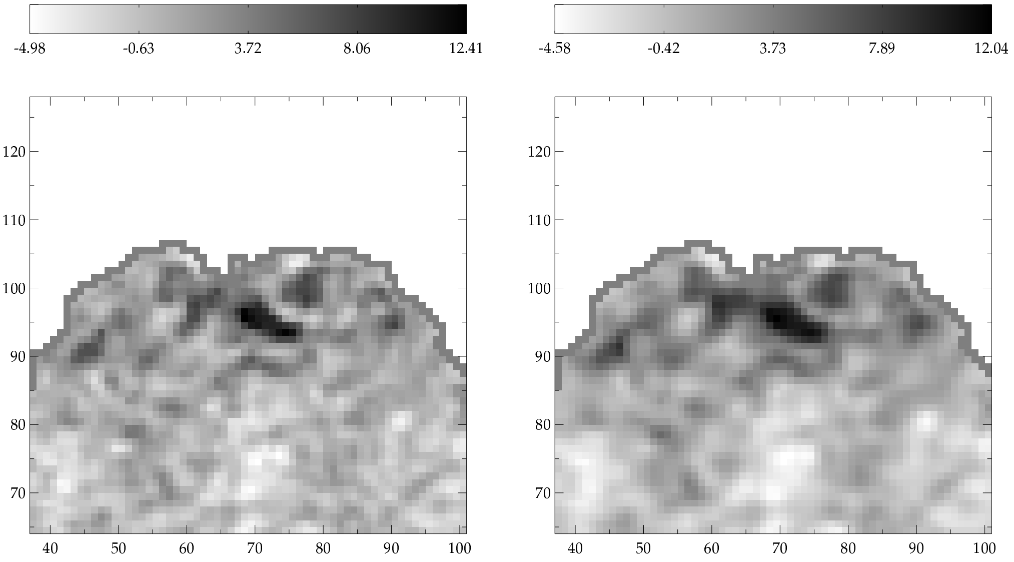


Figure 7: Smoothed activation estimates. Left: Activation map based on data smoothed with a Gaussian kernel with FWHM 2. Right: As left, but with FWHM 3.

The inference on the activation field may be based on two measures: One is the posterior probability that a voxel has a positive value, $P(\Gamma_i > 0 \,|\, X)$, $i \in V$. A natural threshold of this image is at 0.5, representing a neutral balance between type I and type II errors. The other is the mean of the Γ-variate, which represents the level of activation. This two are combined in the left panel in Figure 8, which shows the posterior mean field, masked such that only voxels which have posterior probability of being positive greater than .5 is displayed.

On a higher level we may summarize the posterior distribution of the proportion of activated voxels p_+, the total activation mass $\sum_{i \in V} |\Gamma_i|$ or other parameters of interest. These distributions are a concise representation of the activation, which allows us to compare aspects of different datasets in a rigorous way, without normalizing the data to a standard brain atlas. In Figure 9 are plots of the simulations of different parameters in the model, together with histograms of the empirical distribution. The difference between the mean and range of positive and negative BOLD effects are clearly visible in the distributions of the means and variances of the two Γ-fields.

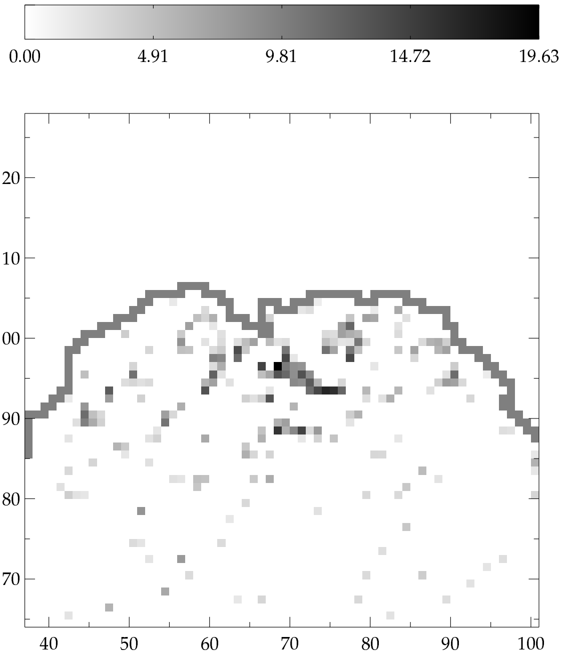


Figure 8: Posterior mean of the Γ-field. The image is masked, such that only voxels with posterior probability of being positive greater than 0.5 is displayed.

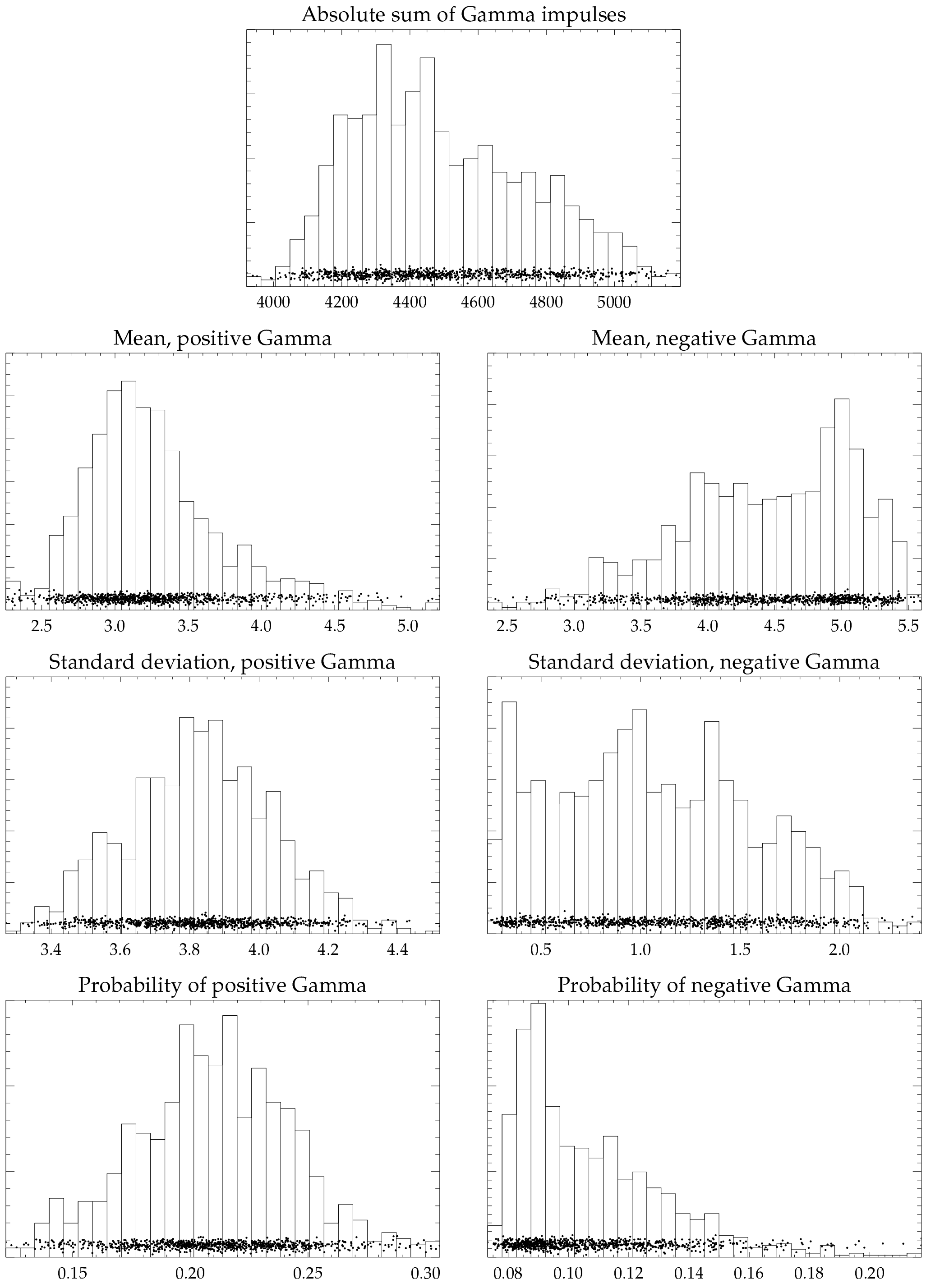


Figure 9: Simulations of the posterior distribution of different parameters in the model given the visual stimulation data.

For model validation, we may study how well the modelled activation surface captures the actual activation pattern. For this we considered the residual activation image $R = X - \hat{E}(\Lambda \mid X)$, i.e. the data X minus the estimated posterior mean of Λ. If the model is correct, this should be an image resembling the noise, in particular the spatial correlation should correspond to that of the noise. In Figure 10 are plots of the corellogram of this residual image, which may be compared with the correlogram of the noise in Figure 1. We see that the correlation functions are very similar, which means that the signal is well represented by $\hat{E}(\Lambda \mid X)$. A noticeable feature however, is that the covariance of the residuals are less than that of the noise, which indicates some degree of overfitting.

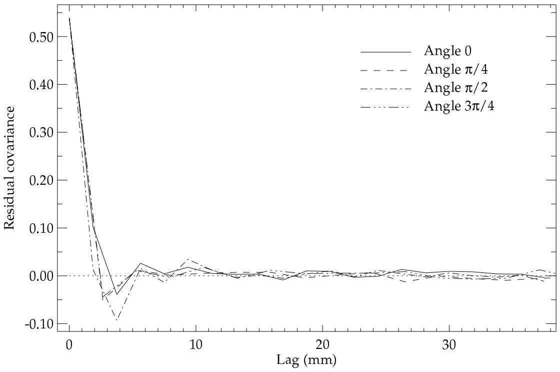


Figure 10: Correlogram of the residual image $R = X - \hat{E}(\Lambda \mid X)$.

5 Discussion

By formulating a Bayesian model for the haemodynamic response, we may perform a spatial deconvolution of the fMRI data. We assume that the haemodynamic response is additive, however as noted earlier the framework is not as sensitive to minor non-additive effects as direct deconvolution. The model for the underlying neural activation pattern is very flexible: We impose no dependency on neighbouring voxels, and assume a broad class of distributions, namely Gamma distributions, for the range of positive activation levels. The motivation for this was to be as uninformative as possible of the activation pattern under study. If substantial prior information is available, which describe possible dependencies in the activation on a neuronal level, they may be included in the model.

The deconvolved patterns in Figure 8 showed a much more localized activation pattern than the original images, and in particular than the pre-smoothed images. In particular we may note that activation detectable in a single or a few voxels in the deconvolved image, may correspond to a much larger area in the smoothed images. Clearly the idea underlying the two methods are very different: When smoothing the data, the main focus is to optimize the probability of detecting a signal of a certain shape and size. On the contrary our aim here is to try to solve the inverse problem of estimating the activation pattern on a neuronal level by a deconvolution.

Stationarity of the haemodynamic response is not required, in the sense that the diffusion kernel is allowed to depend on the underlying tissue. In the present application, however, we have only investigated the stationary case. We estimated the shape and width of the kernel directly from the data by the method of moments, and found that an exponential kernel of width about 1.1 mm fitted the data best. With this model the vascular effects spread over a circle of radius about 3 mm, which corresponds well to the figures reported by Malonek and Grinvald (1996). Using optical imaging, they observed a spread of the vascular response of about 3-5 millimeters.

We found that an exponential shaped kernel fitted the data better than Gaussian kernels. In passing we may note that a Gaussian kernel is often chosen for filtering the data, by reference to the Matched Filter Theorem. For the present data, however, an exponential filtering kernel seemed to match the signal better.

The covariance structure of the data was best represented by a sum of two exponential kernels convolved with themselves. We interpret the narrow kernel as corresponding to local vascular effects, and hence we only include this in the model. Yet the correlogram of the residuals in Figure 10 demonstrates, that the observed spatial covariance structure is well captured by the model. The observed long-range correlation in the data may either correspond to dependencies within the neuronal activation field or to large scale haemodynamic effects arising for instance from larger veins. One possibility of exploring this further is to consider a model with two independent Γ-fields with different kernels, as suggested in (12), and make simultaneous inference on the two fields. In this way it may be possible to separate large-scale and short-range haemodynamic effects, and effectively only make inference on the latter. We have made some initial investigations of this sort of model, and have found promising results. To some extent, the effects of the wide and narrow kernels seem to correspond to different areas, suggesting an interpretation in terms of different types of vascular responses.

There is some degree of overfitting in the model. This can be observed in the mean Γ-field in Figure 8, where some of the active voxels with small posterior mean values seem to corresponds to noise, and from the covariance of the residuals in Figure 10, which is smaller than that of the noise. Finally the estimated value of p_+ of about 21% seems too high to be interpretable as the relative number of active voxels. There is very little regularizing structure in the model, which may explain this overfitting: We assume no prior dependency between the Γ-variates, and the kernel k is relatively narrow, hence only limited spatial regularization is introduced in the Λ-field. Though it is possible to restrict the model by assuming some specific correlation structure on the Γ-field, we are reluctant to doing this, since it is difficult to formulate realistic models for this. The study of interaction and connectivity on the neuronal level is a separate and complex issue, and formulating simple structures in this context seems very problematic. A more promising alternative would be to use a robust noise model, which would be less sensitive to outliers.

In order to shed more light on this issue, the model could be validated on baseline datasets without stimulation, which is also the subject of current work. One problem with this, however, is that the model is explicitly defined for activation datasets, in the sense that the parameters of Γ-fields will be undefined for $p_+ = p_- = 0$. In practice this may lead to numerical problems and instability of the MCMC algorithm.

For the reasons described above, we would hesitate to interpret the detected activation directly in terms of the fitted parameters or to test detailed hypotheses by simulations. The method should be validated more firmly before this level of inference can be performed. The model may, however, be used for spatial deconvolution to obtain much more detailed activation estimates, than when smoothing the data. As the focus of the brain studies move from simple detection of activation to higher level interpretation of the latter, estimation of the activation pattern seems increasingly relevant.

Acknowledgments

The author is grateful for many inspiring discussion with Antti Penttinen, Department of Statistics, University of Jyäskylä, Finland. This work was initiated while the author was visiting Dr. Penttinen, supported by a grant from NorFA. The data used in the example was kindly provided by Hans Stødkilde-Jørgensen from the MR Research Centre, Aarhus University Hospital.

References

Besag, J. *et al.* (1995) Bayesian computation and stochastic systems. (with comments and a reply by the authors.). *Statist. Sci.*, **10**, 3–66.

Cohen, M.S. (1997) Parametric analysis of fMRI data using linear systems methods. *NeuroImage*, **6**, 93–103.

Cressie, N.A.C. (1991) *Statistics for spatial data.* Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics. New York: John Wiley & Sons, Inc.

Dale, A.M. and Buckner, R.L. (1997) Selective averaging of rapidly presented individual trials using fMRI. *Hum Brain Mapping*, **5**, 329–340.

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

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

Devroye, L. (1986) *Non-Uniform Random Variate Generation.* New York: Springer-Verlag.

Friston, K.J., Jezzard, P. and Turner, R. (1994) The analysis of functional MRI time-series. *Human Brain Mapping*, **1**, 153–171.

Friston, K.J., Josephs, O., Rees, G. and Turner, R. (1998) Nonlinear event-related responses in fMRI. *Magn. Reson. Med.*, **39**, 41–52.

Geyer, C. (1999) Likelihood inference for spatial point processes. In O. Barndorff-Nielsen, W. Kendall and M. Lieshout (eds.), *Stochastic Geometry. Likelihood and Computation*, chap. 3, Chapman & Hall/CRC.

Geyer, C.J. (1994) On the convergence of Monte Carlo maximum likelihood calculations. *J. R. Statist. Soc. Ser. B*, **56**, 261–274.

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

Glover, G.H. (1999) Deconvolution of impulse response in event-related BOLD fMRI. *NeuroImage*, **9**, 416–429.

Green, P.J. (2000) *A primer on Markov chain Monte Carlo.* To appear.

Hartvig, N.V. (1999) A stochastic geometry model for fMRI data. Research report 410, Department of Theoretical Statistics, University of Aarhus. *Submitted for publication.*

Hartvig, N.V. (2000) Simulation of the Gamma-Normal distribution. Unpublished manuscript.

Hartvig, N.V. and Jensen, J.L. (2000) Spatial mixture modelling of fMRI data. *Human Brain Mapping.* Accepted for publication.

Higdon, D.M. (1998) Auxiliary variable methods for Markov chain Monte Carlo with applications. *J. Amer. Statist. Assoc.*, **93**.

Husby, O., Lie, T., Langø, T., Hokland, J. and Rue, H. (1999) Bayesian 2D deconvolution: A model for diffuse ultrasound scattering. Preprint Statistics 20/1999, NTNU, Trondheim, Norway.

Kiebel, S.J., Goebel, R. and Friston, K.J. (2000) Anatomically informed basis functions. *NeuroImage*, **11**, 656–667.

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

Lowe, M.J. and Sorenson, J.A. (1997) Spatially filtering functional magnetic resonance imaging data. *Magn. Reson. Med.*, **37**, 723–729.

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

Matérn, B. (1960) *Spatial Variation. Stochastic models and their application to some problems in forest surveys and other sampling investigations.* Band 49, 5. Meddelanden frn Statens Skogsforskningsinstitut. Stockholm.

McCullagh, P. and Nelder, J.A. (1983) *Generalized linear models.* Monographs on Statistics and Applied Probability. London-New York: Chapman & Hall.

Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, M.N. and Teller, E. (1953) Equations of state calculations by fast computing machines. *Journal of Chemical Physics*, **21**, 1087–1091.

Müller, H.G. (1988) *Nonparametric Regression Analysis of Longitudinal Data.* Lecture Notes in Statistics. Springer-Verlag.

Propp, J.G. and Wilson, D.B. (1996) Exact sampling with coupled Markov chains and applications to statistical mechanics. *Random structures and algorithms*, **9**, 223–252.

Rajapakse, J.C. *et al.* (1998) Modeling hemodynamic response for analysis of functional MRI time-series. *Human Brain Mapping*, **6**, 283–300.

Ripley, B.D. (1987) *Stochastic simulation.* Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics. New York: John Wiley & Sons, Inc.

Siegmund, D.O. and Worsley, K.J. (1995) Testing for a signal with unknown location and scale in a stationary Gaussian random field. *Ann. Stat.*, **23**, 608–639.

Tierney, L. (1994) Markov chains for exploring posterior distributions. *Ann. Statist.*, **22**, 1701–1762.

Vazquez, A.L. and Noll, D.C. (1998) Nonlinear aspects of the BOLD response in functional MRI. *NeuroImage*, **7**, 108–118.

Wolpert, R.L. and Ickstadt, K. (1998) Poisson/gamma random field models for spatial statistics. *Biometrika*, **85**, 251–267.

Worsley, K.J. and Friston, K.J. (1995) Analysis of fMRI time-series revisited — again. *NeuroImage*, **2**, 173–181.

Simulation of the Gamma-Normal Distribution

Niels Væver Hartvig[*]
University of Aarhus

February 15, 2000

Abstract

We develop rejection sampling algorithms for simulating from the "Gamma-normal" density (1). This distribution arises naturally in problems where Gamma and normal distributions are combined, with our motivation stemming from human brain mapping. We perform a detailed study of the efficiency of the algorithms, as measured by the acceptance rate. While the algorithms are fairly simple, we show that in most cases they are asymptotically optimal.

AMS 2000 subject classification. Primary 65C10; secondary 68U20.
Keywords: Random numbers, simulation, Gamma distribution, normal distribution.

1 Introduction

In this paper we will consider the problem of simulating random variables with density

$$p(x;\alpha,\beta,\nu) = x^{\nu-1}\exp(-\alpha x-\beta x^2)/C(\alpha,\beta,\nu), \quad x>0. \tag{1}$$

Here $C(\alpha,\beta,\nu)$ is the normalizing constant,

$$C(\alpha,\beta,\nu) = \int_0^\infty x^{\nu-1}\exp(-\alpha x-\beta x^2)\,dx,$$

which is finite for $\alpha\in\mathbb{R},\beta>0,\nu>0$ and for $\alpha>0,\beta=0,\nu>0$.

Our motivation for studying this problem comes from functional magnetic resonance imaging, see Hartvig (2000). In that paper we consider a spatial model for neuronal activation in the brain, during periodic stimulation of a specific cortical centre. The activation levels are modelled as a Gamma random field, which is smoothed by a kernel to represent the blood oxygenation effects, that are detected by the MR scanner. By corrupting this image

[*] Department of Mathematical Sciences, University of Aarhus, Ny Munkegade, DK-8000 Aarhus C, email: vaever@imf.au.dk

with additive Gaussian noise, we obtain a model for the observed scans. The primary interest is to make inference on the underlying Gamma field from the recorded scans, and to this end we wish to draw random samples from the conditional distribution of the Gamma field given the scans. The latter turns out to be of the form (1), which is intuitively clear from the following simple example: Consider a model where $\theta \sim \Gamma(\nu, \lambda)$ and $X|\theta \sim N(\theta, \sigma^2)$. The conditional distribution of θ given X is of the form (1) with $\alpha = \lambda - X/\sigma^2$ and $\beta^{-1} = 2\sigma^2$.

The ability to simulate efficiently from this density is hence important in this model, as well as in other missing data problems where Gamma and normal distributions are combined. Another example is stochastic frontier models in econometry, where the unobserved efficiency of a firm is modelled by a Gamma distribution, see Ritter and Simar (1997) or Koop *et al.* (1995).

The distribution (1) was first introduced by Toranzos (1952) as a generalization of the Pearson family of frequency functions. Toranzos main motivation was to introduce an asymmetric bell-shaped density on the positive real axis, given by the sub-family with $\nu > 1$. He gave recurrent moment formulas and proposed to use these for parameter estimation. Recently Castillo and Puig (1997) considered maximum likelihood estimation, viewing the distribution as a general exponential family which encompasses the truncated normal, Gamma and Rayleigh distributions. The authors used this nesting to develop tests for departure from any of the three distributions.

From a Bayesian perspective the distribution is of interest, as it is a conjugate prior whenever the Gamma or normal distribution are. Hence the family is a flexible class of prior distributions for a positive parameter, such as the positive mean or the inverse variance of a normal variable, the inverse scale parameter of a Gamma variable or the mean of a Poisson variable.

Simulation from this family have mainly been considered in the case $\nu = 1$, where the distribution is a truncated normal. Devroye (1986) lists several rejection sampling algorithms for simulating from the tail of the normal distribution. More generally Koop *et al.* (1995) describe a rejection sampling algorithm for the case $\nu \in \{1, 2, 3\}$, within a Gibbs sampling framework. Here we give rejection sampling algorithms which encompasses all values of $\nu > 0$, and we study the expected number of iterations in the algorithm. The latter is equal to the rejection constant (see Devroye, 1986) which, in our case, is a continuous function of the parameters, and is hence bounded on compact sets. We derive the asymptotic limit of the rejection constant, as the parameters tend to the limit of the parameter space. In most cases the rejection constant tends to 1, i.e. the algorithm is asymptotically optimal. We also give tables of the rejection constant calculated numerically for different finite parameter values.

Notice first the special case where $\beta = 0$. Then (1) is the density of the $\Gamma(\nu, \alpha)$ distribution, which can be simulated efficiently by the algorithms found in Ripley (1987) or Devroye (1986). We can simulate easily also when $\alpha = 0$, as the distribution is then given by $\sqrt{X}$, where $X \sim \Gamma(\nu/2, \beta)$. Hence we will consider the case where $\beta > 0$, $\alpha \neq 0$ and $\nu > 0$. It suffices to consider the two densities

$$p_+(x;\nu,\sigma) = x^{\nu-1}\exp\Big(-\frac{1}{2\sigma^2}(x+1)^2\Big)/C_+(\nu,\sigma), \quad x>0, \tag{2}$$

$$p_-(x;\nu,\sigma) = x^{\nu-1}\exp\Big(-\frac{1}{2\sigma^2}(x-1)^2\Big)/C_-(\nu,\sigma), \quad x>0, \tag{3}$$

for $\nu > 0$ and $\sigma > 0$. For if we let $\sigma^2 = 2\beta/\alpha^2$ and generate $Y \sim p_-(\cdot;\nu,\sigma)$ if $\alpha < 0$ or $Y \sim p_+(\cdot;\nu,\sigma)$ if $\alpha > 0$, then $X = Y|\alpha|/(2\beta)$ has the desired distribution.

We will divide the simulation of p_+ and p_- into the cases $0 < \nu \leq 1$ and $\nu > 1$. Section 2 and 3 give algorithms for p_+ and Section 4 and 5 for p_-. Finally, we have deferred all proofs to an appendix.

2 Simulation of p_+ with $0 < \nu \leq 1$

We have at least two good envelopes for p_+ in this case, which can be used for rejection sampling. The first, g_1, is derived by

$$p_+(x) \leq x^{\nu-1}\exp\Big(-\frac{1}{2\sigma^2}(x^2+1)\Big)/C_+(\nu,\sigma) = c_1 g_1(x),$$

where

$$c_1 = \Gamma(\tfrac{\nu}{2})2^{\frac{\nu}{2}-1}\sigma^\nu e^{-\frac{1}{2\sigma^2}}/C_+(\nu,\sigma), \tag{4}$$

and

$$g_1(x) = \frac{2}{(2\sigma^2)^{\nu/2}\Gamma(\frac{\nu}{2})}x^{\nu-1}\exp\Big(-\frac{1}{2\sigma^2}x^2\Big), \quad x>0.$$

This is the density of $\sqrt{Z}$ where $Z \sim \Gamma(\frac{\nu}{2}, \frac{1}{2\sigma^2})$. The second is given by

$$g_2(x) = \frac{2e^{\frac{1}{2\sigma^2}}}{(2\sigma^2)^\nu\Gamma(\nu)}(x+1)(x+2)^{\nu-1}x^{\nu-1}\exp(-\frac{1}{2\sigma^2}(x+1)^2), \quad x>0, \tag{5}$$

which is the density of $\sqrt{Z+1}-1$ if $Z \sim \Gamma(\nu, 1/(2\sigma^2))$. We find that

$$p_+(x) = 2^{-1}e^{-\frac{1}{2\sigma^2}}(2\sigma^2)^\nu\Gamma(\nu)(x+1)^{-1}(x+2)^{1-\nu}g_2(x)/C_+(\nu,\sigma) \leq c_2 g_2(x),$$

where

$$c_2 = \sigma^{2\nu}\Gamma(\nu)e^{-\frac{1}{2\sigma^2}}/C_+(\nu,\sigma). \tag{6}$$

The rejection constants c_1 and c_2 should be as small as possible to obtain a good acceptance rate. Thus for fixed ν we should use g_1 when $c_1 \leq c_2$ or when

$$1 \geq \frac{c_1}{c_2} = \frac{\Gamma(\frac{\nu}{2})}{\Gamma(\nu)}2^{\nu/2-1}\sigma^{-\nu} = \frac{\sqrt{\pi}}{\Gamma(\frac{\nu+1}{2})}2^{-\nu/2}\sigma^{-\nu}. \tag{7}$$

Here we have used the formula (Abramowitz and Stegun, 1965, p. 256)

$$\frac{\Gamma(\frac{\nu}{2})}{\Gamma(\nu)} = \sqrt{2\pi}2^{1/2-\nu}\frac{1}{\Gamma(\frac{\nu+1}{2})}.$$

We hence find that g_1 should be used when

$$\sigma \geq \left(\frac{\sqrt{\pi}}{\Gamma(\frac{\nu+1}{2})}\right)^{1/\nu}\frac{1}{\sqrt{2}} \simeq 1.873 - 0.965\nu + 0.355\nu^2$$

The approximation, found by a least squares fit, is quite accurate for ν in the range from 0 to 1, with an absolute error less than 0.01.

Let here and in the following U denote a uniform variate on $(0, 1)$, the algorithm is thus:

Algorithm 1. p_+, $0 < \nu \leq 1$

```
1. Initialize: If $\sigma \geq 1.873 - 0.965\nu + 0.355\nu^2$ go to 2, else go to 4.
2. Generate $Z \sim \Gamma(\frac{\nu}{2}, 1)$ and let $X = \sigma\sqrt{2Z}$.
3. Generate $U$. If $U > \exp(-X/\sigma^2)$ go to 2. Otherwise return $X$.
4. Generate $Z \sim \Gamma(\nu, 1)$ and let $X = \sqrt{2\sigma^2 Z + 1} - 1$.
5. Generate $U$. If $U > 2^{\nu-1}(X + 1)^{-1}(X + 2)^{1-\nu}$ go to 4. Otherwise return $X$.
```

The algorithm has a good acceptance rate for all values of the parameters. Writing $c_i = c_i(\nu, \sigma)$ to express the dependence on (ν, σ) explicitly, we have the following optimality properties.

PROPOSITION 1 *The rejection constants satisfy:*

1) For all $\nu > 0$, $c_1(\nu, \sigma)$ is decreasing in σ and tends to 1 as $\sigma \to \infty$.

2) For all $\nu > 0$, $c_2(\nu, \sigma)$ is increasing in σ and tends to 1 as $\sigma \to 0$.

3) For all $\sigma > 0$, $c_1(\nu, \sigma) \wedge c_2(\nu, \sigma)$ is bounded by a function $k(\nu)$ given in (13). The latter is increasing with ν, $k(\nu) \to 1$ as $\nu \to 0$ and $k(1) = 2.36$.

We may note, that the actual rejection constants are somewhat smaller than the bound $k(\nu)$ obtained above. In Table 1 are listed the maximal rejection constants for different values of ν.

3 Simulation of p_+ with $\nu > 1$

For $\nu > 1$ we will proceed as in the case $\nu \leq 1$, dominating $p_+(x)$ by two envelopes, one for small values of σ and one for large. For small values of σ we will use g_2 in (5), which also

Table 1: Maximal rejection constants for algorithm 1 for different values of ν.

ν	$\max_\sigma c_1 \wedge c_2$
0.1	1.05
0.2	1.11
0.3	1.17
0.4	1.24
0.5	1.30
0.6	1.37
0.7	1.45
0.8	1.53
0.9	1.61
1.0	1.70

in this case is an asymptotically optimal envelope as σ tends to 0. For large values of σ we will use an envelope of the form

$$g_3(x) = \frac{2^{1-\frac{\alpha}{2}}}{\sigma^\alpha \Gamma(\frac{\alpha}{2})} x^{\alpha-1} e^{-\frac{1}{2\sigma^2}x^2}, \quad x > 0,$$

i.e. the density of $\sqrt{Z}$, where $Z \sim \Gamma(\frac{\alpha}{2}, \frac{1}{2\sigma^2})$. A scaled version of g_3 can dominate p_+ when $\alpha \leq \nu$. Choosing α such that the modes of p_+ and g_3 coincide, we find that

$$\alpha = \frac{1}{2\sigma^2} - \sqrt{\left(\frac{1}{2\sigma^2}\right)^2 + \frac{\nu-1}{\sigma^2}} + \nu,$$

which satisfies $\alpha \leq \nu$. By choosing the rejection constant c_3 as small as possible such that $p_+(x) \leq c_3 g_3(x)$, we get

$$c_3 = 2^{\frac{\alpha}{2}-1}\Gamma\left(\tfrac{\alpha}{2}\right)(\nu-\alpha)^{\nu-\alpha}\sigma^{2\nu-\alpha}e^{-(\nu-\alpha)-\frac{1}{2\sigma^2}}/C_+(\nu,\sigma).$$

As before, we should use g_3 rather than g_2 when

$$1 \geq \frac{c_3}{c_2} = \frac{\Gamma(\frac{\alpha}{2})}{\Gamma(\nu)} 2^{\frac{\alpha}{2}-1}\sigma^{-\alpha}(\nu-\alpha)^{\nu-\alpha}e^{-(\nu-\alpha)}. \tag{8}$$

When $\nu = 1$ this reduces to $\sigma \geq \sqrt{\pi/2}$, and one may show that for any $\sigma > 0$, $c_3/c_2 \to 0$ as $\nu \to \infty$. Hence for any σ, the envelope g_3 should be used for ν sufficiently large.

The algorithm is then as follows:

Algorithm 2. p_+, $1 < \nu < \infty$

1. Initialize: Let $t_1 = 1/(2\sigma^2)$, $t_2 = \sqrt{t_1^2 + 2t_1(\nu-1)} - t_1$, $\alpha = \nu - t_2$, $t_3 = \alpha/2$, $t_4 = t_2 \log(t_2)$ and $t_5 = \log \sigma^2$. Let $t_6 = \log\Gamma(t_3) - \log\Gamma(\nu) + 0.6931472(t_3 - 1) - t_3 t_5 + t_4 - t_2$. If $t_6 \leq 0$ then go to 2 else go to 4.

2. Generate $Z \sim \Gamma(t_3, 1)$ and let $X = \sigma\sqrt{2Z}$.
3. Generate U. If $\log U > -2t_1 X - t_4 + t_2(\log X + 1 - t_5)$ then go to 2. Else return X.
4. Generate $Z \sim \Gamma(\nu, 1)$ and let $X = \sqrt{2\sigma^2 Z + 1} - 1$.
5. Generate U. If $U > 2^{\nu-1}(X+1)^{-1}(X+2)^{1-\nu}$ go to 4. Otherwise return X.

We have the following asymptotical optimality properties.

PROPOSITION 2 *The rejection constant $c_3(\nu, \sigma)$ satisfies:*

1) For all $\nu \geq 1$, $c_3(\nu, \sigma)$ tends to 1 as $\sigma \to \infty$.

2) For all $\sigma > 0$, $c_3(\nu, \sigma)$ tends to 1 as $\nu \to \infty$.

Proposition 1 and 2 combined tell us, that the rejection constant is 1 in the limit as either $\nu \to \infty$, $\sigma \to 0$ or $\sigma \to \infty$. Hence the algorithm is asymptotically optimal in these limits. The propositions do not describe the asymptotic behaviour of $\max_\sigma c_2(\nu, \sigma) \wedge c_3(\nu, \sigma)$ as ν tends to infinity, which is difficult to obtain since the maximum is not available in closed form. There is, however, some numerical evidence, that the maximal rejection constant may be unbounded along this curve in the parameter space. In Table 2 below we have listed rejection constants for different values of ν.

Table 2: Rejection constants for algorithm 2

ν	$\max_\sigma c_2 \wedge c_3$		$c_2 \wedge c_3$	
		$\sigma = 0.1$	$\sigma = 1.0$	$\sigma = 10.0$
1.0	1.78	1.01	1.53	1.14
1.5	1.55	1.02	1.31	1.05
2.0	1.55	1.03	1.24	1.04
3.0	1.63	1.06	1.18	1.02
5.0	1.79	1.15	1.14	1.01
10.0	2.06	1.64	1.09	1.01
15.0	2.31	2.13	1.07	1.01
25.0	2.67	1.83	1.05	1.00

4 Simulation of p_- with $0 < \nu \leq 1$

This distribution is the most difficult of the four cases, as the shape of the density varies very much with the parameters, and there is no simple limit distribution as the parameters tend to the boundary of the parameter space. This makes it difficult to construct envelopes, which are asymptotically optimal.

When $\nu < 1$ the density tends to infinity at zero, and may or may not have a mode, depending on the value of σ. When $\sigma^2 < \frac{1}{4(1-\nu)}$ the density has a local minimum at $x = \frac{1}{2} - \sqrt{\frac{1}{4} + \sigma^2(\nu - 1)}$ and a mode at $x = \frac{1}{2} + \sqrt{\frac{1}{4} + \sigma^2(\nu - 1)}$. As $\sigma \to 0$ the mode will tend to 1, and the density is almost a truncated normal density except for the infinite pole near 0.

A good envelope should be of the form $x^{\nu-1}$ close to 0 while being almost normal near 1. Of course, it should be easy to simulate from it as well. This is the motivation for dominating p_- in the following way

$$p_-(x) \leq \left(qx^{\nu-1}\mathbf{1}(0 < x < 1) + e^{-\frac{1}{2\sigma^2}(x-1)^2}\right)/C_-(\nu,\sigma), \quad x \in \mathbb{R},$$

for some $q > 0$. The minimal q for which the inequality holds is not available in closed form, but a valid q is given by

$$q(\sigma) = \begin{cases} \frac{\sigma}{\sqrt{e}} & 0 < \sigma < 1, \\ e^{-\frac{1}{2\sigma^2}} & \sigma \geq 1. \end{cases}$$

We will hence let $g_4(x)$ be the density

$$g_4(x) = \frac{1}{\frac{q(\sigma)}{\nu} + \sqrt{2\pi}\sigma}\left(q(\sigma)x^{\nu-1}\mathbf{1}(0 < x < 1) + e^{-\frac{1}{2\sigma^2}(x-1)^2}\right), \quad x \in \mathbb{R},$$

which is a mixture between a normal density, and the density $\nu x^{\nu-1}$, $0 < x < 1$. The latter is the density of $U^{\frac{1}{\nu}}$ when U is uniform on $(0,1)$. The rejection constant is then

$$c_4(\nu,\sigma) = \left(\frac{q(\sigma)}{\nu} + \sqrt{2\pi}\sigma\right)/C_-(\nu,\sigma). \tag{9}$$

We may generate random variables from the normal density by for instance the Box-Muller method, see Devroye (1986). The algorithm is thus:

Algorithm 3. p_-, $0 < \nu \leq 1$

1. Initialize: If $\sigma < 1$ let $q = 0.6065307\,\sigma$ else let $q = \exp(-\frac{1}{2\sigma^2})$. Let $t_1 = q/\nu$ $t_2 = t_1 + 2.506628\,\sigma$.
2. Generate U_1. If $t_2U_1 < t_1$ then go to 3, else go to 5.
3. Generate U_2. Let $X = U_2^{\frac{1}{\nu}}$.
4. Generate U_3. If $U_3^{-1} \leq q\exp\left(\frac{1}{2\sigma^2}(X-1)^2\right) + X/U_2$ then go to 2, else return X.
5. Generate $X \sim N(1,\sigma^2)$. If $X \leq 0$ go to 2.
6. Generate U_4. Let $t_3 = X^{1-\nu}$. If $X < 1$ then $t_3 = t_3 + q\exp\left(\frac{1}{2\sigma^2}(X-1)^2\right)$. If $U_4^{-1} \leq t_3$ then go to 2, else return X.

As mentioned earlier, the algorithm is not optimal. In fact the rejection constant is unbounded as $\sigma \to \infty$ or as $\sigma \to 0$ and $\nu \to 0$, as the following proposition states. Since p_- doesn't tend to any simple distribution in these limits, it seems hard to determine an optimal envelope. However, the rejection constant is acceptable for typical parameter values, as is shown in Table 3. Furthermore, the algorithm may be speeded up by introducing squeezing steps, where the rejection conditions in step 4 and 6 are pretested by simpler expressions, before the exponentials are calculated.

PROPOSITION 3 *The rejection constant c_4 satisfies:*

1) For all $0 < \nu \leq 1$, $c_4(\nu, \sigma) \to (2\pi e)^{-\frac{1}{2}} \nu^{-1} + 1$ for $\sigma \to 0$.

2) For all $0 < \nu \leq 1$, $c_4(\nu, \sigma) = \sqrt{2\pi}\sigma^{1-\nu}2^{1-\frac{\nu}{2}}\Gamma(\frac{\nu}{2})^{-1}(1 + o(1))$ for $\sigma \to \infty$.

3) For all $\sigma > 0$, $c_4(\nu, \sigma) \to \Phi(\frac{1}{\sigma})^{-1} \left(\frac{q(\sigma)}{\sqrt{2\pi}\sigma} + 1\right)$ as $\nu \to 1$. The limiting expression tends to $(2\pi e)^{-1/2} + 1$ as $\sigma \to 0$ and to 2 as $\sigma \to \infty$.

4) For all $\sigma \geq 1$, $c_4(\nu, \sigma) \to 1$ as $\nu \to 0$, and for all $\sigma < 1$, $c_4(\nu, \sigma) \to \sigma e^{\frac{1}{2\sigma^2} - \frac{1}{2}}$ as $\nu \to 0$.

Table 3: Rejection constants for algorithm 3 for different values of ν and σ.

ν	$\sigma = 0.1$	$\sigma = 1.0$	$\sigma = 10.0$
0.1	3.39	2.64	4.71
0.2	2.19	1.89	4.42
0.3	1.80	1.69	4.32
0.4	1.60	1.61	4.14
0.5	1.48	1.58	3.84
0.6	1.40	1.56	3.48
0.7	1.34	1.55	3.08
0.8	1.30	1.53	2.67
0.9	1.27	1.51	2.28
1.0	1.24	1.48	1.93

5 Simulation of p_- with $\nu > 1$

In this situation we will use a single envelope for all values of ν and σ. The envelope g_5 is a normal density,

$$g_5(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}, \quad x \in \mathbb{R},$$

where the mean μ is chosen to match the mode of p_-, i.e.

$$\mu = \frac{1}{2} + \sqrt{\frac{1}{4} + \sigma^2(\nu - 1)}.$$

A scaled version of g_5 can dominate p_- when $\mu > 1$, which is fulfilled in this setting where $\nu > 1$ and $\sigma > 0$. By choosing the rejection constant c_5 as small possible subject to $p_-(x) \leq c_5 g_5(x)$ for all x, we get

$$c_5 = \sqrt{2\pi}\sigma \left(\frac{\sigma^2(\nu - 1)}{\mu - 1} \right)^{\nu - 1} e^{\frac{1}{2\sigma^2}(\mu^2 - 1) - (\nu - 1)} / C_-(\nu, \sigma). \tag{10}$$

The algorithm is thus:

Algorithm 4. p_-**,** $\nu > 1$

1. Initialize: Let $\mu = \frac{1}{2} + \sqrt{\frac{1}{4} + \sigma^2(\nu - 1)}$. Let $t_1 = 1 - 2\log\sigma - \log(\nu - 1) + \log(\mu - 1)$ and $t_2 = (\mu - 1)/\sigma^2$.
2. Generate $X \sim N(\mu, \sigma^2)$. If $X \leq 0$ then go to 2.
3. Generate U. If $\log U > (\nu - 1)(t_1 + \log X) - X t_2$ then go to 2. Else return X.

The algorithm is quite simple, compared to the two-envelope procedures developed in the earlier chapters. As might be expected the cost of this simplicity, is that the algorithm is not asymptotically optimal in all limits of the parameters. The algorithm does, however, yield asymptotical rejection constants which are less than 2, and it is hence reasonably fast, if not optimal, in the limit. The details are given in the following proposition.

Proposition 4 *The rejection constant c_5 satisfies:*

1) For all $\nu > 1$, $c_5(\nu, \sigma) \to 1$ as $\sigma \to 0$.

2) For all $\nu > 1$, $c_5(\nu, \sigma) \to \varphi(\nu)$ as $\sigma \to \infty$, where $\varphi(\nu)$ is a continuous function of ν with $\varphi(\nu) \to 2$ as $\nu \to 1$ and $\varphi(\nu) \to \sqrt{2}$ as $\nu \to \infty$.

3) For all $\sigma > 0$, $c_5(\nu, \sigma) \to \sqrt{2}$ as $\nu \to \infty$.

4) For all $\sigma > 0$, $c_5(\nu, \sigma) \to \Phi(\frac{1}{\sigma})^{-1}$ as $\nu \to 1$.

Acknowledgements

I wish to thank Jens Ledet Jensen for helpful comments. This work was initiated during an inspiring visit to Antti Penttinen at University of Jyväskylä. The visit was funded by a NorFA mobility scholarship.

Table 4: Rejection constants for algorithm 4

ν	$\sigma = 0.1$	$\sigma = 1.0$	$\sigma = 10.0$
1.0	1.00	1.20	1.85
1.5	1.00	1.21	1.54
2.0	1.00	1.23	1.48
3.0	1.01	1.26	1.45
5.0	1.02	1.29	1.43
10.0	1.04	1.32	1.42
15.0	1.05	1.34	1.41
25.0	1.08	1.35	1.41

A Proofs

Proof (Proposition 1) The first claim follows by rewriting the normalizing constant as

$$\begin{aligned} C_+(\nu,\sigma) &= \int_0^\infty x^{\nu-1} e^{-\frac{1}{2\sigma^2}(x+1)^2}\,dx \\ &= e^{-\frac{1}{2\sigma^2}} \sigma^\nu 2^{\nu/2} \int_0^\infty y^{\nu-1} e^{-y^2} e^{-\frac{\sqrt{2}}{\sigma}y}\,dy \\ &= e^{-\frac{1}{2\sigma^2}} \sigma^\nu 2^{\nu/2-1} \Gamma(\tfrac{\nu}{2}) E\big[e^{-\frac{\sqrt{2Z}}{\sigma}}\big], \quad Z \sim \Gamma(\tfrac{\nu}{2},1). \end{aligned} \tag{11}$$

Here we make the substitution $x = \sqrt{2}\sigma y$ in the second line. Inserting this in the expression for c_1 in (4) we find

$$c_1(\nu,\sigma) = E\big[e^{-\frac{\sqrt{2Z}}{\sigma}}\big]^{-1}.$$

By dominated convergence, the mean tends to 1 as $\sigma \to \infty$, which proves 1). If we instead substitute x with $\sigma^2 y$ above, we find that

$$C_+(\nu,\sigma) = e^{-\frac{1}{2\sigma^2}} \sigma^{2\nu} \Gamma(\nu) E\big[e^{-\frac{\sigma^2 V^2}{2}}\big], \quad V \sim \Gamma(\nu,1), \tag{12}$$

and hence

$$c_2(\nu,\sigma) = E\big[e^{-\frac{\sigma^2 V^2}{2}}\big]^{-1}.$$

Again by dominated convergence, the mean value tends to 1 as $\sigma \to 0$ and 2) is shown.

In order to prove 3) we will apply Jensen's inequality to the mean values above, and obtain

$$\begin{aligned} c_1(\nu,\sigma)^{-1} = E[e^{-\frac{\sqrt{2Z}}{\sigma}}] &\geq \exp\left(-\frac{\sqrt{2}}{\sigma}E[\sqrt{Z}]\right) = \exp\left(-\frac{\sqrt{2}}{\sigma}\frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})}\right), \\ c_2(\nu,\sigma)^{-1} = E[e^{-\frac{\sigma^2 V^2}{2}}] &\geq \exp\left(-\frac{\sigma^2}{2}E[V^2]\right) = \exp\left(-\frac{\sigma^2}{2}(\nu+\nu^2)\right). \end{aligned}$$

For fixed ν we hence have

$$c_1(\nu,\sigma) \wedge c_2(\nu,\sigma) \leq \exp\left(\frac{\sqrt{2}}{\sigma}\frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})} \wedge \frac{\sigma^2}{2}(\nu+\nu^2)\right) \leq \exp\left\{\left(\frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})}\right)^{\frac{2}{3}}(\nu(1+\nu))^{\frac{1}{3}}\right\}, \tag{13}$$

where the last inequality follows by maximizing over σ. By letting $k(\nu)$ be the last expression we see that 3) holds. □

Proof (Proposition 2) Using the expression for $C_+(\nu,\sigma)$ in (11), we find that

$$c_3(\nu,\sigma) = \left[\frac{\sigma(\nu-\alpha)}{\sqrt{2}e}\right]^{\nu-\alpha}\frac{\Gamma(\frac{\alpha}{2})}{\Gamma(\frac{\nu}{2})}\left(E\left[e^{-\frac{\sqrt{2Z_\nu}}{\sigma}}\right]\right)^{-1},$$

where $Z_\nu \sim \Gamma(\frac{\nu}{2},1)$. By the dominated convergence theorem we see that the mean value term tends to 1 as $\sigma \to \infty$. It can be directly verified that: 1) α is an increasing function of σ, $\alpha \to 1$ as $\sigma \to 0$ and $\alpha \to \nu$ as $\sigma \to \infty$, and 2) $\sigma(\nu-\alpha)$ is an increasing function of σ, which tends to 0 as $\sigma \to 0$ and to $\sqrt{\nu-1}$ as $\sigma \to \infty$. Hence we have proved 1).

To prove 2) we rewrite the expression for $c_3(\nu,\sigma)$ above, using Stirling's formula for the Gamma function (Abramowitz and Stegun, 1965, p. 257)

$$\Gamma(z) = e^{-z}z^{z-\frac{1}{2}}\sqrt{2\pi}(1+o(1)), \quad z \to \infty, \tag{14}$$

whence we get

$$c_3(\nu,\sigma) = e^{-\frac{\nu-\alpha}{2}}\left(\frac{\alpha}{\nu}\right)^{\frac{\alpha-1}{2}}\left(\frac{\sigma(\nu-\alpha)}{\sqrt{\nu}}\right)^{\nu-\alpha}\left(E\left[e^{-\frac{\sqrt{2Z_\nu}}{\sigma}}\right]\right)^{-1}(1+o(1)).$$

By the central limit theorem we have that $(Z_\nu - \frac{\nu}{2})/\sqrt{\nu/2} \xrightarrow{\mathcal{D}} N(0,1)$ as $\nu \to \infty$, and by the continuous mapping theorem we then have $\sqrt{Z_\nu} - \sqrt{\frac{\nu}{2}} \xrightarrow{\mathcal{D}} N(0,\frac{1}{4})$. Provided that the Laplace transform converges as well, we have that

$$E(e^{-\frac{\sqrt{2Z_\nu}}{\sigma}}) = e^{-\frac{\sqrt{\nu}}{\sigma}+\frac{1}{4\sigma^2}}(1+o(1)). \tag{15}$$

The Laplace transform at $s \in \mathbb{R}$ converges if $\sup_\nu E\left[e^{s(\sqrt{Z_\nu}-\sqrt{\frac{\nu}{2}})}\right] < \infty$ (Hoffmann-Jørgensen, 1994, 5.14). It suffices to show that the transform is bounded in the limit as $\nu \to \infty$. By concavity of $\sqrt{x}$ we have

$$\sqrt{x} - \sqrt{y} \leq \frac{1}{2\sqrt{y}}(x-y), \quad \text{for } 0 < y < x.$$

Using this inequality we find

$$\begin{aligned}
E\left[e^{s(\sqrt{Z_\nu}-\sqrt{\frac{\nu}{2}})}\right] &\leq P(Z_\nu \leq \tfrac{\nu}{2}) + E\left[\mathbf{1}(Z_\nu > \tfrac{\nu}{2})e^{s(\sqrt{Z_\nu}-\sqrt{\frac{\nu}{2}})}\right] \\
&\leq 1 + E\left[e^{\frac{s}{\sqrt{2\nu}}(Z_\nu - \frac{\nu}{2})}\right] \\
&= 1 + \left(1 - \frac{s}{\sqrt{2\nu}}\right)^{-\frac{\nu}{2}} e^{\frac{-s\sqrt{\nu}}{2\sqrt{2}}} \\
&= 1 + \exp\left\{\frac{\nu}{2}\left(\frac{s}{\sqrt{2\nu}} + \frac{1}{2}\frac{s^2}{2\nu} + o(\nu^{-1})\right) - \frac{s\sqrt{\nu}}{2\sqrt{2}}\right\} \\
&= 1 + e^{\frac{s^2}{8}+o(1)}.
\end{aligned}$$

Here the Laplace transform of Z_ν in the second line is finite for $\sqrt{\nu} > s/\sqrt{2}$. We see that the transform is asymptotically bounded, which was to be shown.

By inserting the expression for α and using the expansion $\log(1+x) = x - \frac{1}{2}x^2 + o(x^2)$ as $x \to 0$, with $x = \frac{\alpha}{\nu} - 1$, it is straightforward to show that

$$e^{-\frac{\nu-\alpha}{2}}\left(\frac{\alpha}{\nu}\right)^{\frac{\alpha-1}{2}} = e^{\frac{3}{4\sigma^2}-\frac{\sqrt{\nu}}{\sigma}}(1+o(1)).$$

Likewise we find that

$$\left(\frac{\sigma(\nu-\alpha)}{\sqrt{\nu}}\right)^{\nu-\alpha} = e^{-\frac{1}{2\sigma^2}}(1+o(1)).$$

Combining the these two results and (15) we find that $c_3(\nu,\sigma) = 1 + o(1)$, which proves 2). □

In order to prove Proposition 3 and 4, we will formulate the following lemma.

Lemma 1 *For all $\nu > 0$ and $\sigma > 0$ we have*

$$C_-(\nu,\sigma) = e^{-\frac{1}{2\sigma^2}}\sigma^\nu 2^{\frac{\nu}{2}-1}\Gamma(\tfrac{\nu}{2})E\left[e^{\frac{\sqrt{2Z}}{\sigma}}\right], \tag{16}$$

where $Z \sim \Gamma(\frac{\nu}{2},1)$. We furthermore have,

1) $C_-(\nu,\sigma) = \sqrt{2\pi}\sigma(1+o(1))$ as $\sigma \to 0$.

2) $C_-(\nu,\sigma) = e^{-\frac{1}{4\sigma^2}+\frac{\sqrt{\nu}}{\sigma}}\sigma^\nu 2^{\frac{\nu}{2}-1}\Gamma(\frac{\nu}{2})(1+o(1))$ as $\nu \to \infty$.

3) $C_-(\nu,\sigma) \to \Phi(\frac{1}{\sigma})\sqrt{2\pi}\sigma$ as $\nu \to 1$.

4) $C_-(\nu,\sigma) = e^{-\frac{1}{2\sigma^2}}\nu^{-1}(1+o(1))$ as $\nu \to 0$.

Proof By definition $C_-(\nu, \sigma)$ is given by

$$C_-(\nu, \sigma) = \int_0^\infty x^{\nu-1} \exp\left(-\frac{1}{2\sigma^2}(x-1)^2\right) dx. \tag{17}$$

The first expression follows directly by substituting x with $\sqrt{2}\sigma z$ in the integral.

In order to prove 1) we write $C_-(\nu, \sigma) = \sqrt{2\pi}\sigma E[h(X_\sigma)]$ where $X_\sigma \sim N(1, \sigma^2)$ and

$$h(x) = \begin{cases} x^{\nu-1} & x > 0, \\ 0 & x \leq 0. \end{cases}$$

By Chebychev's inequality $X_\sigma \to 1$ in probability as $\sigma \to 0$, and we must show that also $E[|h(X_\sigma) - 1|] \to 0$. By (Hoffmann-Jørgensen, 1994, 3.25.6) this is true if $\{|h(X_\sigma)|, 0 < \sigma < 1\}$ is uniformly integrable. Consider first $\nu > 1$. Then $\frac{[\nu]}{\nu-1} > 1$ and

$$\begin{aligned} E\left[|h(X_\sigma)|^{\frac{[\nu]}{\nu-1}}\right] &\leq E\left[|X_\sigma|^{[\nu]}\right] \leq \sum_{n=0}^{[\nu]} \binom{[\nu]}{n} E\left[|X_\sigma - 1|^n\right] \\ &= \sum_{n=0}^{[\nu]} \binom{[\nu]}{n} a_n \sigma^n < \sum_{n=0}^{[\nu]} \binom{[\nu]}{n} a_n \ \forall\, \sigma < 1, \end{aligned}$$

where $a_n = 2^{n/2}\pi^{-1/2}\Gamma(\frac{n+1}{2})$. By (Hoffmann-Jørgensen, 1994, 3.24.5) this implies uniform integrability. When $0 < \nu \leq 1$ and $a > 1$ we have

$$\begin{aligned} \sup_{\sigma<1} E\left[|h(X_\sigma)|\mathbf{1}(|h(X_\sigma)| > a)\right] &= \sup_{\sigma<1} \int_0^{a^{\frac{1}{\nu-1}}} x^{\nu-1} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(x-1)^2}\, dx \\ &\leq \sup_{\sigma<1} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(1-a^{\frac{1}{\nu-1}})^2} \frac{1}{\nu} a^{\frac{\nu}{\nu-1}} \\ &= \frac{1}{\sqrt{2\pi}}(1 - a^{\frac{1}{\nu-1}})^{-1} e^{-1/2} \frac{1}{\nu} a^{\frac{\nu}{\nu-1}} \to 0 \text{ as } a \to \infty. \end{aligned}$$

Thus we have that $\{|h(X_\sigma)|, 0 < \sigma < 1\}$ is uniformly integrable.

2) By the argument in the proof of Propostion 2 we have that

$$E\left[e^{\frac{\sqrt{2Z_\nu}}{\sigma}}\right] = E\left[e^{\frac{\sqrt{2}}{\sigma}(\sqrt{Z_\nu} - \sqrt{\frac{\nu}{2}})}\right] e^{\frac{\sqrt{\nu}}{\sigma}} = e^{\frac{1}{4\sigma^2} + \frac{\sqrt{\nu}}{\sigma}}(1 + o(1)),$$

which together with the expression (16) proves 2.

3) follows by using Lebesque's dominated convergence theorem on the expression for $C_-(\nu, \sigma)$ in (17).

4) By splitting the integral in (17) into two, we have

$$\begin{aligned} C_-(\nu, \sigma) &= \int_0^1 x^{\nu-1} e^{-\frac{1}{2\sigma^2}(x-1)^2}\, dx + O(1) = \int_0^1 x^{\nu-1}(e^{-\frac{1}{2\sigma^2}} + O(x))\, dx + O(1) \\ &= \nu^{-1} e^{-\frac{1}{2\sigma^2}} + O(1), \end{aligned}$$

which proves 4) □

Proof (Proposition 3)

All the cases follows directly by inserting the limiting values of $C_-(\nu, \sigma)$ found in Lemma 1 in the expression (9) for $c_4(\nu, \sigma)$. □

Proof (Proposition 4) 1) In the limit as $\sigma \to 0$ we have by Lemma 1

$$c_5(\nu, \sigma) = \left(\frac{\sigma^2(\nu-1)}{\mu-1} \right)^{\nu-1} e^{\frac{1}{2\sigma^2}(\mu^2-1)-(\nu-1)} (1+o(1)).$$

Since $(\mu^2-1)/(2\sigma^2) \to \nu - 1$ and $(\mu-1)/\sigma^2 \to \nu - 1$ as $\sigma \to 0$, we see that $c_5(\nu, \sigma) \to 1$.

2) By Lemma 1 we have the following expression for $c_5(\nu, \sigma)$,

$$c_5(\nu, \sigma) = \sqrt{2\pi} \left(\frac{\sigma(\nu-1)}{e(\mu-1)} \right)^{\nu-1} e^{\frac{\mu^2}{2\sigma^2}} 2^{1-\frac{\nu}{2}} \Gamma(\tfrac{\nu}{2})^{-1} E\left[e^{\frac{\sqrt{2Z}}{\sigma}} \right]^{-1}, \tag{18}$$

where $Z \sim \Gamma(\frac{\nu}{2}, 1)$. The mean value term tends to 1 as $\sigma \to \infty$, and furthermore $(\mu-1)/\sigma \to \sqrt{\nu-1}$ and $\mu^2/2\sigma^2 \to (\nu-1)/2$. Hence we find

$$c_5(\nu, \sigma) \to \sqrt{2\pi} \left(\frac{\nu-1}{e} \right)^{\frac{\nu-1}{2}} 2^{1-\frac{\nu}{2}} \Gamma(\tfrac{\nu}{2})^{-1} \quad \text{as } \sigma \to \infty.$$

Denote the right hand side above $\varphi(\nu)$. Using Stirling's formula for the Gamma function (14) it is now straightforward to verify that $\varphi(\nu) \to 2$ as $\nu \to 1$ and $\varphi(\nu) \to \sqrt{2}$ as $\nu \to \infty$.

3) Using Lemma 1 we find that as $\nu \to \infty$,

$$\begin{aligned} c_5(\nu, \sigma) = \sqrt{2\pi}(\nu-1)^{\frac{\nu-1}{2}} & \left(\frac{\sigma\sqrt{\nu-1}}{\mu-1} \right)^{\nu-1} \\ & \times \exp\left\{ \tfrac{1}{2\sigma^2}(\mu^2-1) - (\nu-1) + \tfrac{1}{4\sigma^2} - \tfrac{\sqrt{\nu}}{\sigma} \right\} 2^{1-\frac{\nu}{2}} \Gamma(\tfrac{\nu}{2})^{-1} (1+o(1)). \end{aligned} \tag{19}$$

Now

$$\frac{\mu-1}{\sigma\sqrt{\nu-1}} = 1 - \frac{1}{2\sigma\sqrt{\nu-1}} + o(1),$$

which implies

$$\left(\frac{\sigma\sqrt{\nu-1}}{\mu-1} \right)^{\nu-1} = e^{\frac{\sqrt{\nu-1}}{2\sigma}} (1+o(1)).$$

By inserting this in (19) and using Stirling's formula for the Gamma function (14), we find that

$$c_5(\nu, \sigma) = \sqrt{2e} \left(\frac{\nu-1}{\nu} \right)^{\frac{\nu-1}{2}} (1+o(1)) = \sqrt{2}(1+o(1)),$$

which proves 3)

4) It is easy to see that $(\mu-1)/(\nu-1) \to \sigma^2$ and $\mu \to 1$ as $\nu \to 1$. Inserting these limits and result 3. from Lemma 1 in the expression (10), we obtain 4) □

References

Abramowitz, M. and Stegun, I.A. (1965) *Handbook of Mathematical Functions.* Applied Mathematics Series. National Bureau of Standards, third edn.

Castillo, J.d. and Puig, P. (1997) Testing departures from Gamma, Rayleigh and truncated normal distributions. *Ann. Inst. Statis. Math.*, **49**, 255–269.

Devroye, L. (1986) *Non-Uniform Random Variate Generation.* New York: Springer-Verlag.

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

Hoffmann-Jørgensen, J. (1994) *Probability with a View toward Statistics*, vol. 1. Chapman & Hall.

Koop, G., Steel, M. and Osiewalski, J. (1995) Posterior analysis of stochastic frontier models using Gibbs sampling. *Computational Statistics*, **10**, 353–373.

Ripley, B.D. (1987) *Stochastic simulation.* Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics. New York: John Wiley & Sons, Inc.

Ritter, C. and Simar, L. (1997) Pitfalls of normal-Gamma stochastic frontier models. *Journal of Productivity Analysis*, **8**, 167–182.

Toranzos, F.I. (1952) An asymmetric bell-shaped frequency curve. *Ann. Math. Statist.*, **23**, 467–469.

ASYMPTOTIC NORMALITY OF THE MAXIMUM LIKELIHOOD ESTIMATOR IN STATE SPACE MODELS

JENS LEDET JENSEN[1] NIELS VÆVER PETERSEN[2]

University of Aarhus[1,2] *and MaPhySto*[1*]

Abstract

State space models is a very general class of time series models capable of modeling dependent observations in a natural and interpretable way. Inference in such models have been studied by Bickel *et al.*, who consider hidden Markov models, which are a special kind of state space models, and prove that the maximum likelihood estimator is asymptotically normal under mild regularity conditions. In this paper we generalize the results of Bickel *et al.* to state space models, where the latent process is a continuous state Markov chain satisfying regularity conditions, which are fulfilled if the latent process takes values in a compact space.

AMS 1991 subject classification. Primary 62F12; secondary 62M09.
Keywords and phrases. State space models, asymptotic normality, maximum likelihood estimation.

1. Introduction. A state space model is a discrete time model for dependent observations $\{Y_k\}$, where the dependence is modelled via an unobserved Markov process $\{X_k\}$ such that, conditionally on $\{X_k\}$ the Y_k's are independent, and the distribution of Y_k depends on X_k only. The unobserved process $\{X_k\}$ is often referred to as the *latent* process. The state space framework encompasses the classical ARMA models, but, more interestingly, non-linear and non-Gaussian models can be formulated in this framework as well.

We will consider inference in state space models by the likelihood method. The likelihood function can not always be calculated explicitly in these models, however, for linear state space models with Gaussian errors the likelihood function can be calculated

[*] Centre for Mathematical Physics and Stochastics, funded by a grant from the Danish National Research Foundation.

by the Kalman filter. There is an extensive literature on Kalman filtering, see for instance West & Harrison (1989) who gives a comprehensive treatment of linear state space models with many examples.

For non-linear state space models and for state space models with non-Gaussian errors the likelihood function can rarely be calculated explicitly. Instead different approximations to the likelihood function have been proposed. Kitagawa & Gersch (1996) discusses an approximation to the likelihood function based on numerical integration techniques, an approach which is also studied in Frühwirth-Schnatter (1994). However, with these techniques the likelihood function can only be approximated to a certain degree of accuracy. Alternatively the likelihood function can be approximated to any degree of accuracy by simulation techniques. This approach is investigated by Durbin & Koopman (1997), Shephard & Pitt (1997) and references therein.

Inference in state space models has mainly been studied in the case of hidden Markov models where the latent process takes values in a finite set. Leroux (1992) proved consistency of the maximum likelihood estimator and Bickel, Ritov & Rydén (1998) proved asymptotic normality. The purpose of this paper is to extend the results of Bickel et al. to cover more general state space models where the latent process is a continuous Markov process. We show that the distributional inequality in Lemma 4 in Bickel et al. is valid in our setup also, under regularity conditions which can be fulfilled if the state space of the latent process is a compact set. The inequality states a mixing property of the latent process, given the observed process, and is the main key to proving asymptotic normality. Having established this mixing result we follow Bickel et al. in their proof of the central limit theorem for the score function and in the proof of the uniform law of large numbers for the observed information.

In Section 2 we state the model and the assumptions we will work under. In Section 3 we state our main results, the central limit theorem for the score function, the uniform law of large numbers for the observed information, and, finally, asymptotic normality of the maximum likelihood estimate. In Section 4 we prove the central limit theorem and in Section 5 we prove the law of large numbers.

2. Notation and assumptions. Let $\{X_k\}$ denote a stationary homogenous Markov chain on the measurable space $(\mathcal{X}, \mathcal{A}, \mu)$. Here $\mathcal{X}$ may be continuous or discrete. A typical setting fulfilling our assumptions below, is where $\mathcal{X}$ is a compact set. Let $\alpha_\theta(x,z)$ denote the transition densities with respect to μ which are parametrized by a parameter $\theta \in \Theta \subseteq \mathbb{R}^d$. Let $\{Y_k\}$ be a sequence of stochastic variables on the measurable space $(\mathcal{Y}, \mathcal{B}, \nu)$ such that given $\{X_k\}$ the Y_k's are independent, and the distribution of Y_i depends through $\{X_k\}$ on X_i only and has density $g_\theta(y_i|x_i)$ wrt. ν. The model can thus be formulated as

$$\begin{aligned} Y_k \,|\, X_k &\sim g_\theta(y_k \,|\, x_k), \\ X_k \,|\, X_{k-1} &\sim \alpha_\theta(x_{k-1}, x_k). \end{aligned}$$

We will let π_θ denote the density wrt. μ of the stationary distribution of X.

We observe values $Y_1, Y_2, \ldots, Y_n$ of the process $\{Y_k\}$ while $\{X_k\}$ remains unobserved, and we wish to estimate θ by the maximum likelihood method. We will let $l_n(\theta)$ denote the log likelihood function based on $Y_1, \ldots, Y_n$. In Section 4 an expression for this function is derived. For the moment we only give the expression for the simultaneous density of $(X_1, \ldots, X_n, Y_1, \ldots, Y_n)$ wrt. $\mu^n \times \nu^n$,

$$p_\theta(x_1, \ldots, x_n, y_1, \ldots, y_n) = \pi_\theta(x_1) g_\theta(y_1 \mid x_1) \prod_{k=2}^{n} \{\alpha_\theta(x_{k-1}, x_k) g_\theta(y_k | x_k)\}. \qquad (1)$$

Above, as everywhere else in this paper, we use the sloppy, but hopefully clear notation $p_\theta(z)$ for the density of a stochastic vector Z with respect to a measure given by the context.

We will let Dg_θ denote the gradient of g_θ wrt. θ and $D^2 g_\theta$ will denote the Hessian, and we will let $\tau_\theta(x) = D \log \pi_\theta(x)$, $\lambda_\theta(x,z) = D \log \alpha_\theta(x,z)$ and $\gamma_\theta(y|x) = D \log g_\theta(y|x)$. The true parameter will be denoted θ_0 and a notation like τ_0 is short for τ_{θ_0}. Throughout the paper X_1^n will denote the vector $(X_1, \ldots, X_n)$ and $\mathbf{c}$ will denote an unspecified finite constant. In the assumptions below we will let $\| \cdot \|$ denote the max-norm of a $d \times d$ maxtrix, $\|A\| = \max_{i,j} |A_{ij}|$.

We will assume that there exists a $\delta > 0$ such that with $B_0 = \{\theta \in \Theta \mid |\theta - \theta_0| < \delta\}$ the following conditions hold.

(A1) There exists a $\sigma > 0$ and an $M < \infty$ such that $\sigma \leq \alpha_\theta(x,z) \leq M$ for all $x, z \in \mathcal{X}$ and all $\theta \in B_0$.

(A2) For all $x, z \in \mathcal{X}$ the maps $\theta \mapsto \alpha_\theta(x,z)$ and $\theta \mapsto \pi_\theta(x)$ are twice continuously differentiable on B_0. Likewise, for all $x \in \mathcal{X}$ and $y \in \mathcal{Y}$ the map $\theta \mapsto g_\theta(y|x)$ is twice continuously differentiable on B_0.

(A3) Define $\rho(y) = \sup_{\theta \in B_0} \sup_{x,z \in \mathcal{X}} g_\theta(y|z)/g_\theta(y|x)$, then

$$\inf_{x \in \mathcal{X}} \int_{\mathcal{Y}} g_0(y|x)/\rho(y)\, \nu(dy) > 0.$$

(A4) (i) $\sup_{\theta \in B_0} \sup_{x,z \in \mathcal{X}} |\lambda_\theta(x,z)| < \infty$ and $\sup_{\theta \in B_0} \sup_{x \in \mathcal{X}} |\tau_\theta(x)| < \infty$.

(ii) $\sup_{\theta \in B_0} \sup_{x,z \in \mathcal{X}} \|D\lambda_\theta(x,z)\| < \infty$ and $\sup_{\theta \in B_0} \sup_{x \in \mathcal{X}} \|D\tau_\theta(x)\| < \infty$.

(iii) Define $\gamma^*(Y_1) = \sup_{\theta \in B_0} \sup_{x \in \mathcal{X}} |\gamma_\theta(Y_1|x)|$ then $\gamma^*(Y_1) \in \mathbb{L}^2(P_0)$ and $\sup_{\theta \in B_0} \sup_{x \in \mathcal{X}} \|D\gamma_\theta(Y_1|x)\| \in \mathbb{L}^1(P_0)$.

(A5) (i) For ν-almost all $y \in \mathcal{Y}$ there exists a function $h_y : \mathcal{X} \to \mathbb{R}_+$ in $\mathbb{L}^1(\mu)$ such that $|g_\theta(y|x)| \leq h_y(x)$ for all $\theta \in B_0$.

(ii) For μ-almost all $x \in \mathcal{X}$ there exist functions $h_x^1 : \mathcal{Y} \to \mathbb{R}_+$ and $h_x^2 : \mathcal{Y} \to \mathbb{R}_+$ in $\mathbb{L}^1(\nu)$ such that $|Dg_\theta(y|x)| \leq h_x^1(y)$ and $\|D^2 g_\theta(y|x)\| \leq h_x^2(y)$ for all $\theta \in B_0$.

(A6) $\theta_0 \in \mathrm{int}(\Theta)$.

Remark 2.1 Note that if $\sup_{x,z\in\mathcal{X}} |\lambda_\theta(x,z)| < \infty$ for a $\theta \in B_0$ and $\sup_{x\in\mathcal{X}} |\tau_\theta(x)| < \infty$ for a $\theta \in B_0$ then A4(ii) implies A4(i). Likewise in A5(ii) the local dominated ν-integrability assumption of $y \mapsto \|D^2 g_\theta(y|x)\|$ for μ-almost all x implies a similar property of $y \mapsto |Dg_\theta(y|x)|$, provided that $|Dg_\theta(y|x)| \in \mathbb{L}^1(P_0)$ for a $\theta \in B_0$.

Remark 2.2 By assumptions A5(i), A1 and A4 the function $x_1^n \mapsto D^i p_\theta(x_1^n, y_1^n)$ is locally dominated μ^n-integrable around θ_0 for ν^n-almost all y_1^n, any $n \in \mathbb{N}$ and $i = 1, 2$. This is seen by noting that by (1) $Dp_\theta(x_1^n, y_1^n)$ consists of a sum of terms like, for instance,

$$\pi_\theta(x_1) g_\theta(y_1|x_1) \prod_{k=2,k\neq j}^{n-1} \{\alpha_\theta(x_{k-1}, x_k) g_\theta(y_k|x_k)\} \alpha_\theta(x_{j-1}, x_j) D g_\theta(y_j|x_j).$$

By A1 and A5(i) this term is absolutely dominated by

$$M^n \prod_{k=1}^{n} h_{y_k}(x_k) |D \log g_\theta(y_j|x_j)| \leq M^n \sup_{\theta\in B_0} \sup_{x\in\mathcal{X}} |\gamma_\theta(y_j|x)| \prod_{k=1}^{n} h_{y_k}(x_k),$$

which for almost all fixed y_1^n is a μ^n-integrable function, by assumption A4(iii). The domination of the second derivative is similar. The local integrability assumption is needed to "interchange integration and differentiation" in some expressions below.

Remark 2.3 The process Y is ergodic under A1. To see this, we observe that Y_k can be described as $Y_k = g(X_k, U_k; \theta_0)$, where $U_1, U_2, \ldots$ are uniformly distributed $U(0,1)$, and independent of X. Now $\{(X_k, U_k)\}$ is a stationary Markov chain, with transition density $p(x_1, u_1 \,|\, x_0, u_0) = \alpha_0(x_0, x_1)$ and hence ergodic by A1. Thus Y is also ergodic.

Remark 2.4 The assumptions A1, A3 and A4 are restrictive and are not fulfilled in a general state space model. A typical example where A1 to A5 are fulfilled is the following. Suppose $\mathcal{X}$ is a compact set in $\mathbb{R}^q$ and μ is the Lebesque measure. If the transition density $\alpha_\theta(x,z)$ and the stationary density $\pi_\theta(x)$ are positive and satisfies A2 and if $\alpha_\theta(x,z)$, $\pi_\theta(x)$ and their first and second derivatives are continuous functions of (θ, x, z) and (θ, x), respectively, then A1, A4(ii) and A4(i) are satisfied. Suppose, furthermore, that $g_\theta(y|x)$ is an exponential family density,

$$g_\theta(y|x) = e^{\phi(x,\theta)\cdot t(y) - \kappa(\phi(x,\theta))}.$$

Here κ denotes the cumulant transform of $t(Y)$ defined on the full parameter space $\Lambda \subseteq \mathbb{R}^k$, and $\phi(x,\theta) \in \Lambda_0$ where Λ_0 is a subset of $\mathrm{int}(\Lambda)$. Suppose that $\phi(x,\theta)$ is twice differentiable wrt. θ and that ϕ and its derivatives are continuous functions of (x,θ), then $\phi(x,\theta) : \mathcal{X} \times \bar{B}_0 \to \Lambda_0$ takes values in a compact set. By continuity of κ we have

$$\begin{aligned} g_\theta(y|x)/g_\theta(y|z) &= \exp\left[\{\phi(x,\theta) - \phi(z,\theta)\} \cdot t(y) - \{\kappa(\phi(x,\theta)) - \kappa(\phi(z,\theta))\}\right] \\ &\leq \mathbf{c}_1 \exp(\mathbf{c}_2 |t(y)|). \end{aligned}$$

Then

$$\begin{aligned}\inf_{x\in\mathcal{X}}\int_{\mathcal{Y}} g_0(y|x)/\rho(y)\,\nu(dy) &\geq \mathbf{c}_1^{-1}\inf_{x\in\mathcal{X}}\int_{\mathcal{Y}} e^{\phi(x,\theta_0)\cdot t(y)-\kappa(\phi(x,\theta_0))}e^{-\mathbf{c}_2|t(y)|}\,\nu(dy)\\ &\geq \mathbf{c}_3\inf_{x\in\mathcal{X}}\int_{\mathcal{Y}} e^{-|\phi(x,\theta_0)|\,|t(y)|}e^{-\mathbf{c}_2|t(y)|}\,\nu(dy)\\ &\geq \mathbf{c}_3\int_{\mathcal{Y}} e^{-\mathbf{c}_4|t(y)|}\,\nu(dy) > 0,\end{aligned}$$

hence A3 is fulfilled. As for A4(iii) we have

$$D\log g_\theta(y|x) = D\{\phi(x,\theta)\cdot t(y)-\kappa(\phi(x,\theta))\} = \frac{\partial\phi(x,\theta)}{\partial\theta^T}\{t(y)-\tau(\phi(x,\theta))\},$$

where $\tau(\phi)=\frac{\partial\kappa(\phi)}{\partial\phi}$ is the mean of $t(Y)$ under P_ϕ and $\frac{\partial\phi(x,\theta)}{\partial\theta^T}$ denotes the $d\times k$ matrix of partial derivatives of ϕ wrt. θ. Then because of compactness of $\mathcal{X}\times\bar{B}_0$ we get

$$\sup_{\theta\in B_0}\sup_{x\in\mathcal{X}}|D\log g_\theta(y|x)| \leq \mathbf{c}_1+\mathbf{c}_2|t(y)|,$$

and hence

$$E_0(\sup_{\theta\in B_0}\sup_{x\in\mathcal{X}}|D\log g_\theta(y|x)|^2) \leq 2\mathbf{c}_1^2+2\mathbf{c}_2^2E_0(|t(Y)|^2) < \infty.$$

The second derivative $D^2\log g_\theta(y|x)$ can be dominated in the same way, and hence A4 follows. Assumption A5(i) follows again by compactness of the parameter space, and finally A5(ii) follows by the continuity of ϕ.

3. Main results. Our main results are stated in this Section. These are a central limit theorem for the score function and a uniform law of large numbers for the observed information. As a consequence of these we find that with a probability that tends to one as n tends to infinity, there exists a (local) maximum point of the likelihood function, which is consistent in probability and asymptotically normal. If especially the maximum likelihood estimator exists and is consistent, it is asymptotically normal.

Let $l_n(\theta)$ denote the log likelihood function based on observations $Y_1,\ldots,Y_n$. Below, $\mathcal{I}_0$ will denote a Fisher information matrix given by

$$\mathcal{I}_0 = E_0(\eta\eta^T),\quad \text{where } \eta=\lim_{n\to\infty}D\log p_0(Y_1\,|\,Y_{-n}^0).$$

This will be formally defined in Section 4, but as the following theorems show it can be thought of as the asymptotic covariance matrix of the score function or the limit of the normed observed information as the number of observations tends to infinity.

Theorem 3.1 *As n tends to infinity $n^{-1/2}Dl_n(\theta_0)\to N(0,\mathcal{I}_0)$, P_0-weakly.*

This Theorem is proved in Section 4.

Theorem 3.2 *Let $\{\theta_n^*\}$ denote any stochastic sequence in Θ such that $\theta_n^* \to \theta_0$ in P_0-probability as $n \to \infty$. Then $n^{-1}D^2 l_n(\theta_n^*) \to -\mathcal{I}_0$ in P_0-probability as $n \to \infty$.*

This Theorem is proved in Section 5. Having established these two results the following result is proved in Jensen (1986) (see also Sweeting (1980).)

Theorem 3.3 *Assume that $\mathcal{I}_0$ is positive definite. With a P_0-probability that tends to 1 as n tends to infinity, there exists a sequence of local maximum points of the likelihood function $\{\hat\theta_n\}$ such that $\hat\theta_n \to \theta_0$ in P_0 probability, and*

$$\sqrt{n}(\hat\theta_n - \theta_0) \to N(0, \mathcal{I}_0^{-1}), P_0 - \textit{weakly.}$$

If especially the maximum likelihood estimator exists and is consistent in P_0-probability, then this estimator has the same limit distribution.

The proof of the second part of the theorem follows by a Taylor expansion of the likelihood function around θ_0, as the proof of Theorem 1 in Bickel et al. The proof of the first part relies on the assumption that $\mathcal{I}_0$ is positive definite, thus in the limit the likelihood function has negative curvature and hence a local maximum at θ_0.

4. A central limit theorem for the score function. In this Section we prove the central limit theorem for the score function stated in Theorem 3.1. Bickel et al. proved the same result in case where the state space of the latent process is finite. Here we start with some Lemmas which will replace Lemma 4 and 5 in Bickel et al. For notational reasons we will assume that d is equal to one in the rest of this paper. If derivatives are replaced by gradients and second derivatives by Hessians all results are valid for general d.

Lemma 4.1 *Let $J \subseteq \mathbb{Z}$ be an integer set and let $\theta \in B_0$. Conditionally on $Y_J = \{Y_j | j \in J\}$, X constitute an inhomogeneous Markov chain with $p_\theta(X_k|X_{k-1}, Y_J) \geq \omega_k$, where*

$$\omega_k = \begin{cases} \sigma^2/(M\rho(Y_k)) & \textit{if } k \in J \\ \sigma^2/M & \textit{if } k \notin J. \end{cases}$$

The inequality is also true for the reversed chain $\{X_{-k}\}_{k\in\mathbb{Z}}$.

Proof The Markov property is proved by considering $n < k < m$, assuming for simplicity that $n \leq j \leq m$ for all $j \in J$. Then

$$\begin{aligned} p_\theta(X_n^{k-1}, X_{k+1}^m \,|\, X_k, Y_J) &= \pi_\theta(X_n) \prod_{i=n}^{m-1} \alpha_\theta(X_i, X_{i+1}) \prod_{j\in J} g_\theta(Y_j|X_j)/p_\theta(X_k, Y_J) \\ &= h_1(X_n^k, Y_J) h_2(X_k^m, Y_J), \end{aligned}$$

where h_1 and h_2 are functions of (X_n^k, Y_J) and (X_k^m, Y_J), respectively. It follows that X_n^{k-1} and X_{k+1}^m are conditionally independent given (X_k, Y_J).

Suppose $k \in J$. Conditionally on X_{k-1} and X_{k+1}, X_k and $Y_{J\setminus\{k\}}$ are independent by definition of the state space model. Hence the conditional density of X_k given (X_{k-1}, X_{k+1}, Y_J) is

$$\begin{aligned} p_\theta(X_k|X_{k-1}, X_{k+1}, Y_J) &= \frac{\alpha_\theta(X_{k-1}, X_k)\alpha_\theta(X_k, X_{k+1})g_\theta(Y_k|X_k)}{\int_{\mathcal{X}} \alpha_\theta(X_{k-1}, x)\alpha_\theta(x, X_{k+1})g_\theta(Y_k|x)\,\mu(dx)} \\ &\geq \sigma^2 \left(M \int_{\mathcal{X}} \alpha_\theta(X_{k-1}, x) \frac{g_\theta(Y_k|x)}{g_\theta(Y_k|X_k)}\,\mu(dx) \right)^{-1} \\ &\geq \sigma^2/(M\rho(Y_k)). \end{aligned} \tag{2}$$

Integrating the conditional density wrt. $p_\theta(X_{k+1}|X_{k-1}, Y_J)$ gives the stated result. When $k \notin J$ the term $g_\theta(Y_k|X_k)$ vanishes.

The proof of the statement for the reversed chain follows by integrating (2) wrt. $p_\theta(X_{k-1} \mid X_{k+1}, Y_J)$ instead. □

We state the following Lemma for future reference, leaving the proof to the reader.

Lemma 4.2 *Let $(\mathcal{X}, \mathcal{A}, \mu)$ be a measure space and let $h\colon \mathcal{X} \to \mathbb{R}$ be a measurable function on $\mathcal{X}$. Let ν_1 and ν_2 be two measures dominated by μ with $\nu_1(\mathcal{X}) = \nu_2(\mathcal{X})$. Then*

$$|\int_{\mathcal{X}} h\,d\nu_1 - \int_{\mathcal{X}} h\,d\nu_2| \leq \{\sup_{x\in\mathcal{X}} h(x) - \inf_{x\in\mathcal{X}} h(x)\}\{\nu_1(S^+) - \nu_2(S^+)\},$$

where $S^+ = \{\frac{d\nu_1}{d\mu} - \frac{d\nu_2}{d\mu} > 0\}$.

In the next Lemma we will let $\{X_k\}_{k\in\mathbb{Z}}$ denote any inhomogenous Markov chain, that is, $\{X_k\}$ is not necessarily the latent process in the model.

Lemma 4.3 *Let $\{X_k\}_{k\in\mathbb{Z}}$ be a Markov chain with state space $(\mathcal{X}, \mathcal{A}, \mu)$. Assume*

$$\frac{dP_{X_k|X_{k-1}}}{d\mu}(x|z) = p_k(z, x) \geq \delta_k,$$

for all $x, z \in \mathcal{X}$ and $k \in \mathbb{Z}$, where $P_{X_k|X_{k-1}}$ denotes the conditional distribution of X_k given X_{k-1}. Then for any $A \in \mathcal{A}$,

$$\sup_{\xi\in\mathcal{X}} P(X_n \in A|X_0 = \xi) - \inf_{\eta\in\mathcal{X}} P(X_n \in A|X_0 = \eta) \leq \prod_{k=1}^{n}(1 - \delta_k\mu(\mathcal{X})).$$

Proof Let $S_k^+ = \{x \in \mathcal{X} \mid p_k(\xi, x) - p_k(\eta, x) > 0\}$ for fixed ξ and η in $\mathcal{X}$, and let $S_k^- = (S_k^+)^c$. Define $M_A^{(k)} = \sup_{\xi\in\mathcal{X}} P(X_n \in A|X_k = \xi)$ and $m_A^{(k)} = \inf_{\eta\in\mathcal{X}} P(X_n \in$

$A|X_k = \eta)$. Then

$$\begin{aligned}
&M_A^{(k-1)} - m_A^{(k-1)} \\
&= \sup_{\xi,\eta} \left(P(X_n \in A|X_{k-1} = \xi) - P(X_n \in A|X_{k-1} = \eta)\right) \\
&= \sup_{\xi,\eta} \int_{\mathcal{X}} P(X_n \in A|X_k = z)\{p_k(\xi, z) - p_k(\eta, z)\}\, \mu(dz) \\
&\leq \sup_{\xi,\eta}\{P(X_k \in S_k^+|X_{k-1} = \xi) - P(X_k \in S_k^+|X_{k-1} = \eta)\}(M_A^{(k)} - m_A^{(k)}) \\
&= \sup_{\xi,\eta}\{1 - P(X_k \in S_k^-|X_{k-1} = \xi) - P(X_k \in S_k^+|X_{k-1} = \eta)\}(M_A^{(k)} - m_A^{(k)}) \\
&\leq \{1 - \delta_k \mu(\mathcal{X})\}(M_A^{(k)} - m_A^{(k)}),
\end{aligned}$$

where the first inequality follows from Lemma 4.2. The result now follows by induction with $k = n, n-1, \ldots, 1$. (Proof based on Doob (1953, p. 198)) □

We are now ready to prove a result corresponding to Lemma 4 in Bickel et al. Let $\omega(y) = \mu(\mathcal{X})\sigma^2/(M\rho(y))$.

Lemma 4.4 *Let $k < l$ and let $J \subseteq \mathbb{Z}$ such that $\{k, k+1, ..., l-1\} \subseteq J$. Let $Y_J = \{Y_j \mid j \in J\}$ then for any $\theta \in B_0$,*

$$\sup_{A \in \mathcal{A}} \sup_{\xi,\eta \in \mathcal{X}} |P_\theta(X_k \in A \mid Y_J, X_l = \xi) - P_\theta(X_k \in A \mid Y_J, X_l = \eta)| \leq \prod_{i=k}^{l-1}(1 - \omega(Y_i)).$$

Likewise, if $l < k$ and $\{l+1, l+2, \ldots, k\} \subseteq J$ then

$$\sup_{A \in \mathcal{A}} \sup_{\xi,\eta \in \mathcal{X}} |P_\theta(X_k \in A \mid Y_J, X_l = \xi) - P_\theta(X_k \in A \mid Y_J, X_l = \eta)| \leq \prod_{i=l+1}^{k}(1 - \omega(Y_i)).$$

Proof Consider the case $k < l$. Applying Lemma 4.1 on the reversed chain $\{X_{-k}\}_{k \in \mathbb{Z}}$ we get

$$p_\theta(X_i \mid X_{i+1}, Y_J) \geq \sigma^2/(M\rho(Y_i)) = \omega(Y_i)/\mu(\mathcal{X}) \quad \text{for } i = k, \ldots, l-1.$$

Using Lemma 4.3 with $\delta_i = \omega(Y_i)/\mu(\mathcal{X})$ we get the stated result. The proof is similar when $l < k$, applying Lemma 4.1 on the original chain $\{X_k\}_{k \in \mathbb{Z}}$. □

Lemma 4.5 *Let $-m \le -n \le k \le 0$. For any θ in B_0 and any $A, B \in \mathcal{A}$ we have*

$$|P_\theta(X_k \in A \,|\, Y_{-n}^1) - P_\theta(X_k \in A \,|\, Y_{-n}^0)| \le \prod_{i=k}^{0}(1-\omega(Y_i)),$$

$$|P_\theta(X_k \in A, X_{k+1} \in B \,|\, Y_{-n}^1) - P_\theta(X_k \in A, X_{k+1} \in B \,|\, Y_{-n}^0)| \le \prod_{i=k+1}^{0}(1-\omega(Y_i)),$$

$$|P_\theta(X_k \in A \,|\, Y_{-n}^1) - P_\theta(X_k \in A \,|\, Y_{-m}^1)| \le \prod_{i=-n}^{k}(1-\omega(Y_i)),$$

$$|P_\theta(X_k \in A, X_{k+1} \in B \,|\, Y_{-n}^1) - P_\theta(X_k \in A, X_{k+1} \in B \,|\, Y_{-m}^1)| \le \prod_{i=-n}^{k}(1-\omega(Y_i)).$$

The first and second expression hold P_θ-almost surely if n is replaced by ∞. The third and fourth hold P_θ-almost surely if m is replaced by ∞ and for both we can replace Y_{-n}^1 and Y_{-m}^1 by Y_{-n}^0 and Y_{-m}^0, respectively.

Proof The first expression can be evaluated as

$$\begin{aligned}
&|P_\theta(X_k \in A \,|\, Y_{-n}^1) - P_\theta(X_k \in A \,|\, Y_{-n}^0)| \\
&= |\int_{\mathcal{X}} P_\theta(X_k \in A \,|\, Y_{-n}^0, x_1)\{p_\theta(x_1 \,|\, Y_{-n}^1) - p_\theta(x_1 \,|\, Y_{-n}^0)\}\,\mu(dx_1)| \\
&\le \sup_{\xi \in \mathcal{X}} P_\theta(X_k \in A \,|\, Y_{-n}^0, X_1 = \xi) - \inf_{\eta \in \mathcal{X}} P_\theta(X_k \in A \,|\, Y_{-n}^0, X_1 = \eta) \le \prod_{i=k}^{0}(1-\omega(Y_i)),
\end{aligned}$$

where the inequalities follows from Lemma 4.2 and 4.4, respectively. As for the second expression,

$$\begin{aligned}
&|P_\theta(X_k \in A, X_{k+1} \in B \,|\, Y_{-n}^1) - P_\theta(X_k \in A, X_{k+1} \in B \,|\, Y_{-n}^0)| \\
&= |\int_B P_\theta(X_k \in A | x_{k+1}, Y_{-n}^k)\{P_\theta(x_{k+1} \,|\, Y_{-n}^1) - P_\theta(x_{k+1} \,|\, Y_{-n}^0)\}\,\mu(dx_{k+1})| \\
&\le |P_\theta(X_{k+1} \in S^+ \,|\, Y_{-n}^1) - P_\theta(X_{k+1} \in S^+ \,|\, Y_{-n}^0)| \le \prod_{i=k+1}^{0}(1-\omega(Y_i)).
\end{aligned}$$

Here S^+ is a set chosen as in Lemma 4.2 and the second inequality follows from above.

By a martingale convergence theorem by Levy (Hoffmann-Jørgensen 1994, p. 505) we get, for instance, that $P_\theta(X_k \in A \,|\, Y_{-n}^1) \to P_\theta(X_k \in A \,|\, Y_{-\infty}^1)$ P_θ-almost surely as $n \to \infty$. This result shows that we can replace n by ∞ in the inequalities above.

The third expression is proved as the first by conditioning on $X_{-n-1} = x_{-n-1}$ in the integral, and the fourth expression follows from the third by an argument similar to the one used to deduce the second from the first. The arguments are identical when replacing Y_{-n}^1 and Y_{-m}^1 with Y_{-n}^0 and Y_{-m}^0, and the extension to the case $m = \infty$ follows from the martingale convergence argument above. □

Lemma 4.5 corresponds to Lemma 5 in Bickel et al. Having established this result the rest of the proof of the CLT for the score function follows the line of these authors closely. However, we will repeat some of the arguments here since there are some differences due to our latent process being continuous.

We will for notational reasons denote our observations $Y_{-n}, \dots, Y_1$. The score function $Dl(\theta)$ is then given by

$$Dl(\theta) = \sum_{k=-n}^{1} D \log p_\theta(Y_k \mid Y_{-n}^{k-1}),$$

where $p_\theta(Y_k \mid Y_{-n}^{k-1})$ denotes the conditional density of Y_k given Y_{-n}^{k-1} given by

$$D \log p_\theta(Y_k \mid Y_{-n}^{k-1}) = D \log p_\theta(Y_{-n}^{k}) - D \log p_\theta(Y_{-n}^{k-1}).$$

Using assumption A5(i) to interchange integration and differentiation below we find that for any $j = k-1, k$,

$$D \log p_\theta(Y_{-n}^{j}) = E_\theta(D \log p_\theta(Y_{-n}^{j}, X_{-n}^{k}) \mid Y_{-n}^{j}).$$

Hence $D \log p_\theta(Y_k \mid Y_{-n}^{k-1})$ is given by

$$\begin{aligned} &D \log p_\theta(Y_k \mid Y_{-n}^{k-1}) \\ &\quad = E_\theta(D \log p_\theta(Y_{-n}^{k}, X_{-n}^{k}) \mid Y_{-n}^{k}) - E_\theta(D \log p_\theta(Y_{-n}^{k-1}, X_{-n}^{k}) \mid Y_{-n}^{k-1}). \end{aligned} \tag{3}$$

Using the expression for $p_\theta(Y_{-n}^{k}, X_{-n}^{k})$ in (1) we find

$$\begin{aligned} &D \log p_\theta(Y_k \mid Y_{-n}^{k-1}) = \\ &\sum_{i=-n}^{k-1} \left\{ E_\theta(\lambda_\theta(X_i, X_{i+1}) + \gamma_\theta(Y_i|X_i)|Y_{-n}^{k}) - E_\theta(\lambda_\theta(X_i, X_{i+1}) + \gamma_\theta(Y_i|X_i)|Y_{-n}^{k-1}) \right\} \\ &\quad + E_\theta(\tau_\theta(X_{-n}) \mid Y_{-n}^{k}) - E_\theta(\tau_\theta(X_{-n}) \mid Y_{-n}^{k-1}) + E_\theta(\gamma_\theta(Y_k|X_k) \mid Y_{-n}^{k}). \end{aligned} \tag{4}$$

Now, let

$$\begin{aligned} \eta_1 = \sum_{i=-\infty}^{0} &\left\{ E_0(\lambda_0(X_i, X_{i+1}) + \gamma_0(Y_i|X_i)|Y_{-\infty}^{1}) - \right. \\ &\left. E_0(\lambda_0(X_i, X_{i+1}) + \gamma_0(Y_i|X_i)|Y_{-\infty}^{0}) \right\} + E_0(\gamma_0(Y_1|X_1) \mid Y_{-\infty}^{1}). \end{aligned} \tag{5}$$

The infinite sum is absolutely convergent in $\mathbb{L}^2(P_0)$, as will be shown in Lemma 4.6, so η_1 is a well defined variable in $\mathbb{L}^2(P_0)$. Let

$$\mathcal{I}_0 = E_0(\eta_1^2).$$

Letting $|| \cdot ||_2$ denote the $\mathbb{L}^2(P_0)$-norm we have:

Lemma 4.6 *There exists a* $\beta \in [0,1)$ *and a constant* $\mathbf{c}$ *such that*

$$||D\log p_0(Y_1 \mid Y_{-n}^0) - \eta_1||_2 \le \mathbf{c}\beta^n,$$

for all n.

Proof Let

$$Z_k = \lambda_0(X_k, X_{k+1}) + \gamma_0(Y_k|X_k).$$

By splitting the sums in (4) and (5) we can dominate $||D\log p_0(Y_1 \mid Y_{-n}^0) - \eta_1||_2$ by the sum of the following terms:

$$||E_0(\tau_0(X_{-n}) \mid Y_{-n}^1) - E_0(\tau_0(X_{-n}) \mid Y_{-n}^0)||_2, \tag{6}$$

$$||E_0(\gamma_0(Y_1|X_1) \mid Y_{-n}^1) - E_0(\gamma_0(Y_1|X_1) \mid Y_{-\infty}^1)||_2, \tag{7}$$

$$\sum_{k=-[n/2]}^{0} ||E_0(Z_k \mid Y_{-n}^j) - E_0(Z_k \mid Y_{-\infty}^j)||_2, \quad j = 0, 1, \tag{8}$$

$$\sum_{k=-n}^{-[n/2]-1} ||E_0(Z_k \mid Y_{-n}^1) - E_0(Z_k \mid Y_{-n}^0)||_2, \tag{9}$$

$$\sum_{k=-\infty}^{-[n/2]} ||E_0(Z_k \mid Y_{-\infty}^1) - E_0(Z_k \mid Y_{-\infty}^0)||_2, \tag{10}$$

where $[\cdot]$ denotes the integer part. We will show that each of the terms (6)–(10) can be dominated by $\mathbf{c}\beta^n$, where $0 \le \beta < 1$, which proves the Lemma. Furthermore, the domination of (10) shows that the sum in (5) is absolutely convergent as stated earlier.

We will show the domination of (9) and leave the remaining terms to the reader. We will first consider the part of Z_k given by $\gamma_0(Y_k|X_k)$ in (9). By applying Lemma 4.2 and 4.5 we have the following inequality:

$$\begin{aligned}
&\left|E_0(\gamma_0(Y_k|X_k) \mid Y_{-n}^1) - E_0(\gamma_0(Y_k|X_k) \mid Y_{-n}^0)\right| \\
&= \left|\int_{\mathcal{X}} \gamma_0(Y_k|x_k)\{p_0(x_k \mid Y_{-n}^1) - p_0(x_k \mid Y_{-n}^0)\}\, \mu(dx_k)\right| \\
&\le 2\sup_{x\in\mathcal{X}} |\gamma_0(Y_k|x)| \prod_{i=k+1}^{0} (1-\omega(Y_i)).
\end{aligned}$$

Hence the $\mathbb{L}^2$-norm can be dominated as

$$\begin{aligned}
&||E_0(\gamma_0(Y_k|X_k) \mid Y_{-n}^1) - E_0(\gamma_0(Y_k|X_k) \mid Y_{-n}^0)||_2^2 \\
&\le 4E_0\left(E_0\left(\sup_{x\in\mathcal{X}} \gamma_0(Y_k|x)^2 \prod_{i=k+1}^{0} (1-\omega(Y_i))^2 \middle| X_k^0\right)\right) \\
&= 4E_0\left(E_0(\sup_{x\in\mathcal{X}} \gamma_0(Y_k|x)^2|X_k) \prod_{i=k+1}^{0} E_0((1-\omega(Y_i))^2|X_i)\right) \\
&\le \mathbf{c}\beta^{-k},
\end{aligned} \tag{11}$$

where the equality follows by definition of the state space model and where β is given by

$$\begin{aligned}\beta &= \sup_{x\in\mathcal{X}} \int_{\mathcal{Y}} \left(1 - \frac{\mu(\mathcal{X})\sigma^2}{M\rho(y)}\right)^2 g_0(y|x)\,\nu(dy)\\ &\le \sup_{x\in\mathcal{X}} \int_{\mathcal{Y}} \left(1 - \frac{\mu(\mathcal{X})\sigma^2}{M\rho(y)}\right) g_0(y|x)\,\nu(dy)\\ &= 1 - \frac{\mu(\mathcal{X})\sigma^2}{M} \inf_{x\in\mathcal{X}} \int_{\mathcal{Y}} g_0(y|x)/\rho(y)\,\nu(dy) < 1,\end{aligned}$$

by assumption A3. The constant in (11) is finite by assumption A4. For a sum of $\mathbb{L}^2$-norms we get,

$$\sum_{k=-n}^{-[n/2]-1} ||E_0(\gamma_0(Y_k|X_k)\,|\,Y_{-n}^1) - E_0(\gamma_0(Y_k|X_k)\,|\,Y_{-n}^0)||_2$$

$$\le \mathbf{c} \sum_{k=-n}^{-[n/2]-1} \beta^{-k/2} \le \mathbf{c}\beta^{[n/2]/2}.$$

The part of (9) involving $\lambda_0(X_k, X_{k+1})$ can be dominated in a similar way, using A4 and the second inequality in Lemma 4.4. Hence we have proved the claimed domination of (9). □

Lemma 4.6 is the final brick needed to prove Theorem 3.1; it tells us that in the limit the score function function is equivalent to a sum of terms like η_1. These constitute a stationary martingale increment sequence, and hence by a martingal central limit theorem we obtain the stated limit distribution of the score function. The proof is identical to the proof of Lemma 1 in Bickel et al. (p. 1626)

5. A law of large numbers for the observed information. In this Section we will show Theorem 3.2. As in the previous Section we will start with some Lemmas providing inequalities for conditional probabilities. Lemmas 5.1 and 5.3 are multivariate versions of Lemma 4.5.

LEMMA 5.1 *Let $-m \le -n \le k \le l \le 0$, and let $\theta \in B_0$. Then for all $C \in \sigma\{(X_t, Y_t) : t \le l\}$ we have*

$$|P_\theta(C\,|\,Y_{-n}^1) - P_\theta(C\,|\,Y_{-n}^0)| \le \prod_{i=l}^{0}(1 - \omega(Y_i)).$$

Likewise for all $C \in \sigma\{(X_t, Y_t) : t \ge k\}$ and for $j = 0, 1$ we have,

$$|P_\theta(C\,|\,Y_{-n}^j) - P_\theta(C\,|\,Y_{-m}^j)| \le \prod_{i=-n}^{k}(1 - \omega(Y_i)).$$

PROOF Let $C \in \sigma\{(X_t, Y_t), t \le l\}$. Then

$$\begin{aligned}
&|P_\theta(C \,|\, Y^1_{-n}) - P_\theta(C \,|\, Y^0_{-n})| \\
&= |\int_{\mathcal{X}} P_\theta(C \,|\, x_l, Y^l_{-n})\{p_\theta(x_l \,|\, Y^1_{-n}) - p_\theta(x_l \,|\, Y^0_{-n})\}\, \mu(dx_l)| \\
&\le P_\theta(X_l \in S^+ \,|\, Y^1_{-n}) - P_\theta(X_l \in S^+ \,|\, Y^0_{-n}) \le \prod_{i=l}^{0}(1 - \omega(Y_i)),
\end{aligned}$$

where $S^+ = \{x_l \in \mathcal{X} \mid p_\theta(x_l \,|\, Y^1_{-n}) - p_\theta(x_l \,|\, Y^0_{-n}) > 0\}$ is chosen as in Lemma 4.2, and the last inequality follows from 4.5. The second inequality is derived by a similar argument, by conditioning on X_k instead of X_l. □

In the next Lemma $\{X_k\}$ denotes any inhomogenous Markov chain, as in Lemma 4.3.

LEMMA 5.2 *Let the setup be as in Lemma 4.3. Let $n \in \mathbb{Z}$ and let Q be the measure on $\mathcal{A} \otimes \mathcal{A}$ defined by,*

$$Q(A \times B) = P(X_0 \in A)P(X_n \in B),$$

for $A, B \in \mathcal{A}$. Then for all $C \in \mathcal{A} \otimes \mathcal{A}$,

$$|P((X_0, X_n) \in C) - Q(C)| \le \prod_{k=1}^{n}(1 - \delta_k \mu(\mathcal{X})).$$

PROOF Let $C_{x_0} = \{x_n \in \mathcal{X} \mid (x_0, x_n) \in C\}$, then

$$\begin{aligned}
&|P((X_0, X_n) \in C) - Q(C)| \\
&\quad = |\int_{\mathcal{X}} \{P(X_n \in C_{x_0} \,|\, X_0 = x_0) - P(X_n \in C_{x_0})\}\, P_{X_0}(dx_0)| \le \prod_{k=1}^{n}(1 - \delta_k \mu(\mathcal{X})).
\end{aligned}$$

Here the last inequality follows from Lemma 4.3 since

$$\begin{aligned}
&|P(X_n \in A \,|\, X_0 = \xi) - P(X_n \in A)| \\
&= |\int_{\mathcal{X}} \{P(X_n \in A \,|\, X_0 = \xi) - P(X_n \in A \,|\, X_0 = \eta)\}\, P_{X_0}(d\eta) \\
&= \sup_{\xi \in \mathcal{X}} P(X_n \in A \,|\, X_0 = \xi) - \inf_{\eta \in \mathcal{X}} P(X_n \in A \,|\, X_0 = \eta) \le \prod_{k=1}^{n}(1 - \delta_k \mu(\mathcal{X})).
\end{aligned}$$

□

LEMMA 5.3 *Let $-m \le -n \le k \le l \le 0$. Let $Q^j_{\theta,-n}$ be the measure on $\mathcal{A} \otimes \mathcal{A}$ defined by*

$$Q^j_{\theta,-n}(A \times B) = P_\theta(X_k \in A \,|\, Y^j_{-n}) P_\theta(X_l \in B \,|\, Y^j_{-n})$$

for $j = 0, 1$ and $A, B \in \mathcal{A}$. Then for all $\theta \in B_0$, for $C \in \mathcal{A} \otimes \mathcal{A}$ and for $j = 0, 1$,

$$|P_\theta((X_k, X_l) \in C \mid Y_{-n}^j) - Q_{\theta,-n}^j(C)| \leq \prod_{i=k}^{l-1}(1 - \omega(Y_i)),$$

$$|Q_{\theta,-n}^1(C) - Q_{\theta,-n}^0(C)| \leq 2\prod_{i=l}^{0}(1 - \omega(Y_i)),$$

$$|Q_{\theta,-n}^j(C) - Q_{\theta,-m}^j(C)| \leq 2\prod_{i=-n}^{k}(1 - \omega(Y_i)).$$

Proof The first inequality follows from Lemma 5.2 and 4.1. To prove the second expression we will let $C_y = \{x \in \mathcal{X} \mid (x, y) \in C\}$, $C'_x = \{y \in \mathcal{X} \mid (x, y) \in C\}$ and proceed as follows,

$$\begin{aligned}
&|Q_{\theta,-n}^1(C) - Q_{\theta,-n}^0(C)| \\
&= |\int_C \{p_\theta(x_k \mid Y_{-n}^1) p_\theta(x_l \mid Y_{-n}^1) - p_\theta(x_k \mid Y_{-n}^0) p_\theta(x_l \mid Y_{-n}^0)\}\, \mu(dx_k)\mu(dx_l)| \\
&\leq |\int_C \{p_\theta(x_k \mid Y_{-n}^1) - p_\theta(x_k \mid Y_{-n}^0)\}\, p_\theta(x_l \mid Y_{-n}^1)\mu(dx_k)\mu(dx_l)| \\
&\qquad + |\int_C \{p_\theta(x_l \mid Y_{-n}^1) - p_\theta(x_l \mid Y_{-n}^0)\}\, p_\theta(x_k \mid Y_{-n}^0)\, \mu(dx_l)\mu(dx_k)| \\
&\leq \int_{\mathcal{X}} |P_\theta(X_k \in C_{x_l} \mid Y_{-n}^1) - P_\theta(X_k \in C_{x_l} \mid Y_{-n}^0)|\, p_\theta(x_l | Y_{-n}^1)\, \mu(dx_l) + \\
&\qquad + \int_{\mathcal{X}} |P_\theta(X_l \in C'_{x_k} \mid Y_{-n}^1) - P_\theta(X_l \in C'_{x_k} \mid Y_{-n}^0)|\, p_\theta(x_k | Y_{-n}^0)\, \mu(dx_k) \\
&\leq \prod_{i=k}^{0}(1 - \omega(Y_i)) + \prod_{i=l}^{0}(1 - \omega(Y_i)) \leq 2\prod_{i=l}^{0}(1 - \omega(Y_i)),
\end{aligned}$$

where the third inequality is given by Lemma 4.5.

The third expression is proved as the second. □

Having established these inequalities we are ready to prove the law of large numbers for the observed information. Using A5(i) to interchange integration and differentiation we find

$$\begin{aligned}
&D^2 \log p_\theta(Y_1 \mid Y_{-n}^0) = D^2 \log p_\theta(Y_{-n}^1) - D^2 \log p_\theta(Y_{-n}^0) \\
&= E_\theta(D^2 \log p_\theta(X_{-n}^1, Y_{-n}^1) \mid Y_{-n}^1) - E_\theta(D^2 \log p_\theta(X_{-n}^1, Y_{-n}^0) \mid Y_{-n}^0) \\
&\quad + \operatorname{var}_\theta(D \log p_\theta(X_{-n}^1, Y_{-n}^1) \mid Y_{-n}^1) - \operatorname{var}_\theta(D \log p_\theta(X_{-n}^1, Y_{-n}^0) \mid Y_{-n}^0)
\end{aligned} \tag{12}$$

Define for notational reasons

$$Z_{\theta,k} = \lambda_\theta(X_k, X_{k+1}) + \gamma_\theta(Y_k | X_k) \text{ and } \dot{Z}_{\theta,k} = D\lambda_\theta(X_k, X_{k+1}) + D\gamma_\theta(Y_k | X_k).$$

Inserting expression (1) in (12) we get,

$$\begin{aligned}
&D^2 \log p_\theta(Y_1 \,|\, Y_{-n}^0) \qquad\qquad (13)\\
&= E_\theta(D\tau_\theta(X_{-n}) \,|\, Y_{-n}^1) - E_\theta(D\tau_\theta(X_{-n}) \,|\, Y_{-n}^0) + E_\theta(D\gamma_\theta(Y_1|X_1) \,|\, Y_{-n}^1)\\
&\quad + \sum_{k=-n}^{0} \{E_\theta(\dot{Z}_{\theta,k} \,|\, Y_{-n}^1) - E_\theta(\dot{Z}_{\theta,k} \,|\, Y_{-n}^0)\}\\
&\quad + \mathrm{var}_\theta(\gamma_\theta(Y_1|X_1) \,|\, Y_{-n}^1) + \mathrm{var}_\theta(\tau_\theta(X_{-n}) \,|\, Y_{-n}^1) - \mathrm{var}_\theta(\tau_\theta(X_{-n}) \,|\, Y_{-n}^0)\\
&\quad + \sum_{k=-n}^{0} \sum_{l=-n}^{0} \left\{\mathrm{cov}_\theta(Z_{\theta,k}, Z_{\theta,l} \,|\, Y_{-n}^1) - \mathrm{cov}_\theta(Z_{\theta,k}, Z_{\theta,l} \,|\, Y_{-n}^0)\right\}\\
&\quad + 2\sum_{k=-n}^{0} \left\{\mathrm{cov}_\theta(\tau_\theta(X_{-n}), Z_{\theta,k} \,|\, Y_{-n}^1) - \mathrm{cov}_\theta(\tau_\theta(X_{-n}), Z_{\theta,k} \,|\, Y_{-n}^0)\right\}\\
&\quad + 2\sum_{k=-n}^{0} \mathrm{cov}_\theta(\gamma_\theta(Y_1|X_1), Z_{\theta,k} \,|\, Y_{-n}^1) + 2\mathrm{cov}_\theta(\gamma_\theta(Y_1|X_1), \tau_\theta(X_{-n}) \,|\, Y_{-n}^1).
\end{aligned}$$

We then have the following convergence result.

Lemma 5.4 *As $m, n \to \infty$,*

$$\left|\left|\sup_{\theta\in B_0} |D^2 \log p_\theta(Y_1 \,|\, Y_{-n}^0) - D^2 \log p_\theta(Y_1 \,|\, Y_{-m}^0)|\right|\right|_1 \to 0,$$

where $||\cdot||_1$ denotes the $\mathbb{L}^1(P_0)$-norm.

This Lemma states that $\{D^2 \log p_\theta(Y_1 \,|\, Y_{-n}^0)\}$ is a uniform Cauchy sequence in $\mathbb{L}^1(P_0)$. This is important because it proves the existence of a limit in $\mathbb{L}^1(P_0)$ of $D^2 \log p_\theta(Y_1 \,|\, Y_{-n}^1)$ as $n \to \infty$ for any $\theta \in B_0$, and not only $\theta = \theta_0$. In the proof we will need the following Lemma.

Lemma 5.5 *Let $-m \le -n \le k \le l \le 0$ and let $Z_{\theta,k}$ be defined as above. Then there exists a $\beta \in [0, 1)$ such that the following inequalities hold for $j = 0, 1$,*

$$\left|\left|\sup_{\theta\in B_0} |\mathrm{cov}_\theta(Z_{\theta,k}, Z_{\theta,l} \,|\, Y_{-n}^1) - \mathrm{cov}_\theta(Z_{\theta,k}, Z_{\theta,l} \,|\, Y_{-n}^0)|\right|\right|_1 \le \mathbf{c}\beta^{-l}, \qquad (14)$$

$$\left|\left|\sup_{\theta\in B_0} |\mathrm{cov}_\theta(Z_{\theta,k}, Z_{\theta,l} \,|\, Y_{-n}^j) - \mathrm{cov}_\theta(Z_{\theta,k}, Z_{\theta,l} \,|\, Y_{-m}^j)|\right|\right|_1 \le \mathbf{c}\beta^{k+n}, \qquad (15)$$

$$\left|\left|\sup_{\theta\in B_0} |\mathrm{cov}_\theta(Z_{\theta,k}, Z_{\theta,l} \,|\, Y_{-n}^j)|\right|\right|_1 \le \mathbf{c}\beta^{l-k}. \qquad (16)$$

Above $Z_{\theta,i}$ may be replaced by $\tau_\theta(X_i)$ or $\gamma_\theta(Y_i \,|\, X_i)$ for $i = k, l$.

Proof Recall that $Z_{\theta,k} = \lambda_\theta(X_k, X_{k+1}) + \gamma_\theta(Y_k \mid X_k)$. Thus the covariance of $Z_{\theta,k}$ and $Z_{\theta,l}$ splits into the sum of four covariance terms involving λ_θ and γ_θ. We will show that $\operatorname{cov}_\theta(\gamma_\theta(Y_k \mid X_k), \gamma_\theta(Y_l \mid X_l) \mid Y)$ satisfies the claimed inequalities. The three remaining terms are similar.

To show the first inequality we will consider the expression

$$\begin{aligned}
&\sup_{\theta\in B_0} |E_\theta\{\gamma_\theta(Y_k \mid X_k)\gamma_\theta(Y_l \mid X_l) \mid Y_{-n}^1\} - E_\theta\{\gamma_\theta(Y_k \mid X_k)\gamma_\theta(Y_l \mid X_l) \mid Y_{-n}^0\}| \\
&= \sup_{\theta\in B_0} | \int_{\mathcal{X}^2} \gamma_\theta(Y_k \mid x_k)\gamma_\theta(Y_l \mid x_l)\{p_\theta(x_k, x_l \mid Y_{-n}^1) - p_\theta(x_k, x_l \mid Y_{-n}^0)\}\, \mu(dx_k)\mu(dx_l)| \\
&\le 2\gamma^*(Y_k)\gamma^*(Y_l) \sup_{\theta\in B_0} |P_\theta((X_k, X_l) \in S^+ \mid Y_{-n}^1) - P_\theta((X_k, X_l) \in S^+ \mid Y_{-n}^0)| \\
&\le 2\gamma^*(Y_k)\gamma^*(Y_l) \prod_{i=l}^{0}(1 - \omega(Y_i)).
\end{aligned}$$

Here $\gamma^*(Y_k) = \sup_{\theta\in B_0}\sup_{x\in\mathcal{X}} |\gamma_\theta(Y_k|x)|$ as defined in assumption A4, and the inequalities follows from Lemma 4.2 and 5.1, respectively. The $\mathbb{L}^1\,(P_0)$-norm of such a term is thus less than

$$\begin{aligned}
&2E_0\left(\gamma^*(Y_k)\gamma^*(Y_l) \prod_{i=l}^{0}(1 - \omega(Y_i))\right) \\
&= 2E_0\left(E_0\left(\gamma^*(Y_k)\gamma^*(Y_l) \prod_{i=l}^{0}(1 - \omega(Y_i))\middle| X_k^0\right)\right) \\
&\le 2E_0\left(E_0(\gamma^*(Y_k)|X_k)E_0(\gamma^*(Y_l)|X_l) \prod_{i=l+1}^{0} E_0(1 - \omega(Y_i)|X_i)\right) \\
&\le 2E_0(\gamma^*(Y_1)^2)\beta^{-l} = \mathbf{c}\beta^{-l},
\end{aligned}$$

where the first inequality follows by definition of the state space model, and where β is given by

$$\beta = \sup_{x\in\mathcal{X}} E_0(1 - \omega(Y_i)|X_i = x) = \sup_{x\in\mathcal{X}} \int_{\mathcal{Y}} \left(1 - \frac{\mu(\mathcal{X})\sigma^2}{M\rho(y)}\right) g_0(y|x)\,\nu(dy) < 1,$$

by assumption A3. Assumption A4 assures that the constant $\mathbf{c}$ above is finite.

The expression

$$\begin{aligned}
\Bigg|\Bigg| &\sup_{\theta\in B_0} |E_\theta\{\gamma_\theta(Y_k \mid X_k) \mid Y_{-n}^1\}E_\theta\{\gamma_\theta(Y_l \mid X_l) \mid Y_{-n}^1\} - \\
&\qquad E_\theta\{\gamma_\theta(Y_k \mid X_k) \mid Y_{-n}^0\}E_\theta\{\gamma_\theta(Y_l \mid X_l) \mid Y_{-n}^0\}| \Bigg|\Bigg|_1
\end{aligned}$$

can be dominated by the same technique, using the second expression in Lemma 5.3. Hence (14) is proved.

The second inequality (15) is proved as (14) using the second expression in Lemma 5.1 and the third expression in Lemma 5.3, respectively. As for (16) we have

$$\begin{aligned}
&\sup_{\theta\in B_0} |\operatorname{cov}_\theta(\gamma_\theta(Y_k \,|\, X_k), \gamma_\theta(Y_l \,|\, X_l) \,|\, Y_{-n}^j)| \\
&= \sup_{\theta\in B_0} \Big| \int_{\mathcal{X}^2} \gamma_\theta(Y_k \,|\, x_k)\gamma_\theta(Y_l \,|\, x_l) \\
&\qquad\qquad \{p_\theta(x_k, x_l \,|\, Y_{-n}^j) - p_\theta(x_k \,|\, Y_{-n}^j) p_\theta(x_l \,|\, Y_{-n}^j)\}\, \mu(dx_k)\mu(dx_l)\Big| \\
&\le 2\gamma^*(Y_k)\gamma^*(Y_l) \prod_{i=k}^{l-1} (1-\omega(Y_i)),
\end{aligned}$$

by the first expression in Lemma 5.3. The claimed domination of the $\mathbb{L}^1(P_0)$-norm of this term is proved as above. □

PROOF (Lemma 5.4.) Considering the expression for $D^2 \log p_\theta(Y_1 \,|\, Y_{-n}^0)$ in (13) we will show that the term

$$\begin{aligned}
&\Bigg\| \sup_{\theta\in B_0} \Bigg| \sum_{k=-m}^{0} \sum_{l=-m}^{0} \{\operatorname{cov}_\theta(Z_{\theta,k}, Z_{\theta,l} \,|\, Y_{-m}^1) - \operatorname{cov}_\theta(Z_{\theta,k}, Z_{\theta,l} \,|\, Y_{-m}^0)\} \\
&\qquad - \sum_{k=-n}^{0} \sum_{l=-n}^{0} \{\operatorname{cov}_\theta(Z_{\theta,k}, Z_{\theta,l} \,|\, Y_{-n}^1) - \operatorname{cov}_\theta(Z_{\theta,k}, Z_{\theta,l} \,|\, Y_{-n}^0)\} \Bigg| \Bigg\|_1 \to 0 \quad (17)
\end{aligned}$$

as $n, m \to \infty$. The remaining terms in (13) can be treated with similar arguments.

Suppose $m > n$. By symmetry of k and l in the sum in (17) it suffices to consider the sum over the region where $k \le l$. This region can be further divided into 5 subregions,

$$\begin{aligned}
D_1 &= \{(k,l) \in \mathbb{Z}^2 \mid -[n/2] \le k \le 0, k \le l \le 0\}, \\
D_2 &= \{(k,l) \in \mathbb{Z}^2 \mid -n \le k \le -[n/2], [k/2] \le l \le 0\}, \\
D_3 &= \{(k,l) \in \mathbb{Z}^2 \mid -m \le k \le -n, [k/2] \le l \le 0\}, \\
D_4 &= \{(k,l) \in \mathbb{Z}^2 \mid -n \le k \le -[n/2], k \le l \le [k/2]\}, \\
D_5 &= \{(k,l) \in \mathbb{Z}^2 \mid -m \le k \le -n, k \le l \le [k/2]\}.
\end{aligned}$$

We will show that the sum over each of these regions tends to zero in $\mathbb{L}^1(P_0)$ as $n, m \to \infty$, hence proving (17). Using (15) we find that

$$\begin{aligned}
&\sum_{(k,l)\in D_1} \Bigg\| \sup_{\theta\in B_0} |\{\operatorname{cov}_\theta(Z_{\theta,k}, Z_{\theta,l} \,|\, Y_{-m}^1) - \operatorname{cov}_\theta(Z_{\theta,k}, Z_{\theta,l} \,|\, Y_{-m}^0)\} \\
&\qquad - \{\operatorname{cov}_\theta(Z_{\theta,k}, Z_{\theta,l} \,|\, Y_{-n}^1) - \operatorname{cov}_\theta(Z_{\theta,k}, Z_{\theta,l} \,|\, Y_{-n}^0)\}| \Bigg\|_1 \le \mathbf{c} \sum_{k=-[n/2]}^{0} \sum_{l=k}^{0} \beta^{k+n}
\end{aligned}$$

Using (16) we find that the corresponding sums over D_2 and D_3 are less than

$$\mathbf{c} \sum_{k=-n}^{-[n/2]} \sum_{l=[k/2]}^{0} \beta^{l-k} \qquad \text{and} \qquad \mathbf{c} \sum_{k=-m}^{-n} \sum_{l=[k/2]}^{0} \beta^{l-k},$$

respectively, and by (14) the sums over D_4 and D_5 are dominated by

$$\mathbf{c} \sum_{k=-n}^{-[n/2]} \sum_{l=k}^{[k/2]} \beta^{-l} \qquad \text{and} \qquad \mathbf{c} \sum_{k=-m}^{-n} \sum_{l=k}^{[k/2]} \beta^{-l},$$

respectively. Since $0 \leq \beta < 1$ these sums all tend to zero as $n, m \to \infty$ and the proof is complete. □

Lemma 5.6 *The map $\theta \mapsto D^2 \log p_\theta(Y_1|Y^0_{-n})$ from B_0 to $\mathbb{L}^1(P_0)$ is continuous.*

Proof Let $\{\theta_m\} \subseteq B_0$ be a sequence such that $\theta_m \to \theta$ as $m \to \infty$. We will show that $E_0\{|D^2 \log p_{\theta_m}(Y_1 \mid Y^0_{-n}) - D^2 \log p_\theta(Y_1 \mid Y^0_{-n})|\} \to 0$, as $m \to \infty$. Considering the expression in (13) we must show that terms like, for instance,

$$E_0\big[|E_{\theta_m}\{\gamma_{\theta_m}(Y_k|X_k)\gamma_{\theta_m}(Y_l|X_l) \mid Y^1_{-n}\} - E_\theta\{\gamma_\theta(Y_k|X_k)\gamma_\theta(Y_l|X_l) \mid Y^1_{-n}\}|\big]$$

tend to zero as $m \to \infty$. The integrand can be evaluated as

$$\begin{aligned}
&|E_{\theta_m}\{\gamma_{\theta_m}(Y_k|X_k)\gamma_{\theta_m}(Y_l|X_l) \mid Y^1_{-n}\} - E_\theta\{\gamma_\theta(Y_k|X_k)\gamma_\theta(Y_l|X_l) \mid Y^1_{-n}\}| \\
&\leq |\int_{\mathcal{X}^2} \gamma_{\theta_m}(Y_k|x_k)\gamma_{\theta_m}(Y_l|x_l)\{p_{\theta_m}(x_k, x_l \mid Y^1_{-n}) - p_\theta(x_k, x_l \mid Y^1_{-n})\}\, \mu(dx_k)\mu(dx_l)| \\
&+ |\int_{\mathcal{X}^2} \{\gamma_{\theta_m}(Y_k|x_k)\gamma_{\theta_m}(Y_l|x_l) - \gamma_\theta(Y_k|x_k)\gamma_\theta(Y_l|x_l)\} p_\theta(x_k, x_l \mid Y^1_{-n})\, \mu(dx_k)\mu(dx_l)|.
\end{aligned}$$

The first term is less than

$$\gamma^*(Y_k)\gamma^*(Y_l) \int_{\mathcal{X}^2} |p_{\theta_m}(x_k, x_l \mid Y^1_{-n}) - p_\theta(x_k, x_l \mid Y^1_{-n})|\, \mu(dx_k)\mu(dx_l) \tag{18}$$

$$= \frac{\gamma^*(Y_k)\gamma^*(Y_l)}{p_{\theta_m}(Y^1_{-n})} \int_{\mathcal{X}^2} |p_{\theta_m}(x_k, x_l, Y^1_{-n}) - p_\theta(x_k, x_l \mid Y^1_{-n}) p_{\theta_m}(Y^1_{-n})|\, \mu(dx_k)\mu(dx_l). \tag{19}$$

The integral tends to zero as $m \to \infty$ as can be seen by considering the simultaneous density

$$p_{\theta_m}(x_k, x_l, Y^1_{-n}) = \int_{\mathcal{X}^n} \pi_{\theta_m}(x_{-n}) g_{\theta_m}(Y_{-n}|x_{-n}) \prod_{i=-n+1}^{1} \{\alpha_{\theta_m}(x_{i-1}, x_i) g_{\theta_m}(Y_i|x_i)\} \prod_{\substack{i=-n \\ i \neq k,l}}^{1} \mu(dx_i). \tag{20}$$

Since the integrand here is continuous and can be dominated by

$$M^{n+2}\prod_{i=-n}^{1} h_{Y_i}(x_i) \in \mathbb{L}^1(\mu^n), \tag{21}$$

by assumption A1 and A5, we have from Lebesgue's dominated convergence theorem that

$$p_{\theta_m}(x_k, x_l, Y_{-n}^1) \to p_\theta(x_k, x_l, Y_{-n}^1) \quad \text{as } m \to \infty.$$

Likewise $p_{\theta_m}(Y_{-n}^1) \to p_\theta(Y_{-n}^1)$ as $m \to \infty$, and hence the integrand in (19) tends to zero. By (21) the integrand can be dominated in $\mathbb{L}^1(\mu^2)$, and therefore (19) tends to zero.

Since the expression in (18) is less than

$$\gamma^*(Y_k)\gamma^*(Y_l)\int_{\mathcal{X}^2}\{p_{\theta_m}(x_k, x_l|Y_{-n}^1) + p_\theta(x_k, x_l|Y_{-n}^1)\}\,\mu(dx_k)\mu(dx_l) \\ = 2\gamma^*(Y_k)\gamma^*(Y_l), \tag{22}$$

it is dominated in $\mathbb{L}^1(P_0)$ and hence tends to zero in $\mathbb{L}^1(P_0)$ as $m \to \infty$.

The second term can be dominated similarly and tends to zero P_0-almost surely, and therefore also in $\mathbb{L}^1(P_0)$, by the continuity of γ_θ. □

Lemma 5.4 and 5.6 show that $\{D^2 \log p_\theta(Y_1|Y_{-n}^0)\}_{n\in\mathbb{N}}$ is a uniform Cauchy sequence of continuous functions in $\mathbb{L}^1(P_0)$, which proves Lemma 10 of Bickel et al. The final Lemma states a usual property of the Fisher information. With this result, the remaining part of the proof of Theorem 3.2 is now identical to the proof of Lemma 2 in Bickel et al. (p. 1633)

Lemma 5.7 *For any n,*

$$E_0\{D^2 \log p_0(Y_1|Y_{-n}^0)\} = -E_0\{[D\log p_0(Y_1|Y_{-n}^0)]^2\}.$$

Proof By (3) and (12) we have

$$\begin{aligned}
&(D\log p_0(Y_1 \mid Y_{-n}^0))^2 + D^2\log p_0(Y_1 \mid Y_{-n}^0) \\
&= 2\Big(E_0(D\log p_0(X_{-n}^1, Y_{-n}^0) \mid Y_{-n}^0)^2 \\
&\qquad - E_0\big(D\log p_0(X_{-n}^1, Y_{-n}^1) \mid Y_{-n}^1\big) E_0\big(D\log p_0(X_{-n}^1, Y_{-n}^0) \mid Y_{-n}^0\big)\Big) \\
&+ E_0((D\log p_0(X_{-n}^1, Y_{-n}^1))^2 \mid Y_{-n}^1) - E_0((D\log p_0(X_{-n}^1, Y_{-n}^0))^2 \mid Y_{-n}^0) \\
&+ E_0(D^2\log p_0(X_{-n}^1, Y_{-n}^1) \mid Y_{-n}^1) - E_0(D^2\log p_0(X_{-n}^1, Y_{-n}^0) \mid Y_{-n}^0).
\end{aligned} \tag{23}$$

The first term has zero mean. This follows by noting from (1) that

$$D\log p_0(X_{-n}^1, Y_{-n}^1) = D\log p_0(X_{-n}^1, Y_{-n}^0) + \gamma_0(Y_1|X_1), \tag{24}$$

thus

$$\begin{aligned}
&E_0\big[E_0\{D\log p_0(X_{-n}^1,Y_{-n}^1)\mid Y_{-n}^1\}E_0\{D\log p_0(X_{-n}^1,Y_{-n}^0)\mid Y_{-n}^0\}\big]\\
&= E_0\big\{D\log p_0(X_{-n}^1,Y_{-n}^1)E_0\{D\log p_0(X_{-n}^1,Y_{-n}^0)\mid Y_{-n}^0\}\big\}\\
&= E_0\big\{D\log p_0(X_{-n}^1,Y_{-n}^0)E_0(D\log p_0(X_{-n}^1,Y_{-n}^0)\mid Y_{-n}^0)\big\}\\
&\quad + E_0\big\{\gamma_0(Y_1|X_1)E_0(D\log p_0(X_{-n}^1,Y_{-n}^0)\mid Y_{-n}^0)\big\}\\
&= E_0\big\{D\log p_0(X_{-n}^1,Y_{-n}^0)E_0(D\log p_0(X_{-n}^1,Y_{-n}^0)\mid Y_{-n}^0)\big\}\\
&\quad + E_0\big\{E_0\big[\gamma_0(Y_1|X_1)|X_1\big]E_0\big[E_0(D\log p_0(X_{-n}^1,Y_{-n}^0)\mid Y_{-n}^0)\mid X_1\big]\big\}\\
&= E_0\big\{D\log p_0(X_{-n}^1,Y_{-n}^0)E_0(D\log p_0(X_{-n}^1,Y_{-n}^0)\mid Y_{-n}^0)\big\},
\end{aligned}$$

where the third equality follows from the conditional independence of Y_{-n}^0 and Y_1 given X_1, and the last equality from the fact that $E_0(\gamma_0(Y_1|X_1)\mid X_1) = 0$ by A5(ii). The mean of the first term in (23) is then

$$\begin{aligned}
&2E_0\big\{E_0(D\log p_0(X_{-n}^1,Y_{-n}^0)\mid Y_{-n}^0)\\
&\qquad\big[E_0(D\log p_0(X_{-n}^1,Y_{-n}^0)\mid Y_{-n}^0) - D\log p_0(X_{-n}^1,Y_{-n}^0)\big]\big\} = 0.
\end{aligned}$$

By (24) the mean of the sum of the two last terms is given by

$$\begin{aligned}
&E_0\big\{D^2\log g_0(Y_1\mid X_1)\big\} + E_0\big\{(D\log g_0(Y_1\mid X_1))^2\big\}\\
&\qquad + 2E_0\big\{\gamma_0(Y_1\mid X_1)D\log p_0(X_{-n}^1,Y_{-n}^0)\big\}.
\end{aligned}$$

The last term is zero, which is seen by conditioning on X_1 and using the argument from above. The sum of the two first terms is zero by assumption A5(ii), which completes the proof. □

References

Bickel, P. J., Ritov, Y. & Rydén, T. (1998), 'Asymptotic normality of the maximum-likelihood estimator for general hidden Markov models', *Ann. Statist.* **26**(4), 1614–1635.

Doob, J. L. (1953), *Stochastic Processes*, John Wiley & Sons, Inc.

Durbin, J. & Koopman, S. J. (1997), 'Monte Carlo maximum likelihood estimation for non-Gaussian state space models', *Biometrika* **84**(3), 669–684.

Frühwirth-Schnatter, S. (1994), 'Applied state space modelling of non-Gaussian time series using integration-based Kalman filtering', *Stat. Comp.* **4**, 259–269.

Hoffmann-Jørgensen, J. (1994), *Probability with a View toward Statistics*, Vol. 1, Chapman & Hall.

Jensen, J. L. (1986), Nogle asymptotiske resultater. University of Aarhus (In Danish).

Kitagawa, G. & Gersch, W. (1996), *Smoothness Priors Analysis of Time Series*, Springer-Verlag, New York.

Leroux, B. G. (1992), 'Maximum-likelihood estimation for hidden Markov models', *Stoch. Proc. Appl.* **40**, 127–143.

Shephard, N. & Pitt, M. K. (1997), 'Likelihood analysis of non-Gaussian measurement time series', *Biometrika* **84**(3), 653–667.

Sweeting, T. (1980), 'Uniform asymptotic normality of the maximum likelihood estimator', *Ann. Statist.* **8**(6), 1375–1381.

West, M. & Harrison, J. (1989), *Bayesian Forecasting and Dynamic Models*, Springer-Verlag, New York.

Dept. of Theoretical Statistics
University of Aarhus
Ny Munkegade
DK-8000 Aarhus C, Denmark

E-mail: jlj@imf.au.dk
vaever@imf.au.dk

Non-linear state space models with applications in functional magnetic resonance imaging

Niels Væver Petersen

April 20, 1998

Abstract

In functional magnetic resonance imaging (fMRI) movies consisting of hundreds of images of the human brain are acquired and analysed to identify regions of the brain, where the intensity changes according to controlled stimuli. Commonly observed phenomena in fMRI time series are trends and fluctuations, possibly caused by movement effects remaining in the images after realignment, physiological effects like changes in blood pressure and cardiopulmonary effects. A widespread method is to model the trend as a linear term and, furthermore, some authors have designed digital filters to reduce the pulsation effects. However, the trend is not neccesarily linear, and eliminating the pulsations by filtering might not be the optimal solution if the cardiac rate varies during the experiment. In this paper we will estimate non-linear trend terms and cardiac rhythms directly from the data by a multidimensional state space model, and model any pixel time series with the trend and fluctuation terms included in the mean value space. In the first part of the paper an introduction to state space models is given and an approximate Kalman filter for non-linear models is developed. In the second part these models are employed in fMRI.

1 Introduction

This paper falls in two parts. The first is an introduction to state space models, with a description of a Kalman filter for non-linear models. The theory is exemplified by an analysis of the Canadian Lynx dataset. In the second part we consider functional magnetic resonance imaging (fMRI) data, and the problem of modelling trends and fluctuations in these data. We employ non-linear state space models and the general Kalman filter to estimate physiological noise components in the images, and include these in a general model for any fMRI time series.

2 State space models

In this part we will give a general introduction to state space models. In Section 2.1 we will formulate a general state space model and motivate the use of this model. In Section 2.2 we will concentrate on linear state space models and derive the Kalman filter and Kalman smoother. In Section 2.3 we will consider non-linear state space models with Gaussian errors and derive an approximation to the Kalman filter. Finally in 2.4 we will discuss estimation of parameters in these models.

2.1 Definition of a state space model

State space models is a class of time series models for discrete time observations, which is becoming increasingly popular these years. The reason for this is that the state space framework provides an intuitive and easily interpretable way of modelling dependencies among stochastic variables, as well as a natural way of incorporating known deterministic covariates. The applications cover all classical areas of time series modeling, such as biological processes and financial time series (Kitagawa & Gersh 1984, West & Harrison 1989), but state space models can be formulated for many other types of data. For instance Jørgensen et al. (1997) consider models for multivariate count data and apply the models to daily counts of emergency room visits for respiratory diseases, and to cucumber yields.

Suppose we observe a sequence of correlated stochastic variables $\{Y_t\}$. In the state space framework the dependency between the Y_t's are modelled via an unobserved Markov process $\{X_t\}$ such that conditionally on $\{X_t\}$ the Y_t's are independent and the distribution of Y_i depends on $\{X_t\}$ through X_i only. Graphically the dependencies are represented as in Figure 1

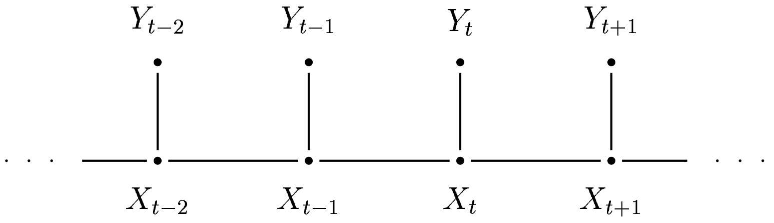


Figure 1: A graphical representation of a state space model

The model can be stated as

$$Y_t \,|\, X_t \sim g_t(y_t \,|\, x_t), \tag{1}$$

$$X_t \,|\, X_{t-1} \sim \alpha_t(x_{t-1}, x_t), \tag{2}$$

for $t = 1, 2, \dots, n$ and

$$X_0 \sim \pi(x_0). \tag{3}$$

Equation (1) is often referred to as the *observation equation* and equation (2) as the *state equation* or *system equation*. Equation (3) is the *initial distribution* or, in a Bayesian terminology, the *initial prior*. Considering the $\{Y_t\}$'s as the output of a dynamical system the process $\{X_t\}$ is often interpreted as the state of the system, explaining the terminology "state space models". The process $\{X_t\}$ is called the *state process*, the *latent proces* or the *regime* of the system.

The state space framework is suitable for many different time series models, where an unobserved process enters directly or indirectly. The unobserved process might be of interest in its own right or it might be a technical tool for formulating a specific correlation structure. As an example of the first, the state space framework can be used to model financial time series, where the variance, or the volatility, changes over time. Here the unobserved process is the volatility, which is of interest since it is a measure of the stability of the market. As an example of the latter, one can formulate the classical ARMA models in the state space framework (West & Harrison 1989, p. 306). In this situation the latent process contains lagged observations of the observed process, which obviously is not of separate interest.

As mentioned above, state space models can be seen as an extension of the Box-Jenkins models. However, while it is difficult to interpret the dependency structure in an ARMA(2,2) model, say, the state space models often provide an intuitive and interpretable framework. The state space formulation not only allows us to model an observed time series satisfactorily, but also provides insight in the process that are generating the observations, which is of course the ultimate goal of all statistical modelling. As an example of this viewpoint, and as a motivation for studying state space models, let us consider the classical Candian lynx data set.

2.1.1 Example: The Canadian lynx data

In this example we will consider the Canadian lynx data set consisting of the number of annual trappings of lynx in the Mackenzie River district of North-West Canada in the years 1821–1934. A plot of the number of trappings can be seen in Figure 2.

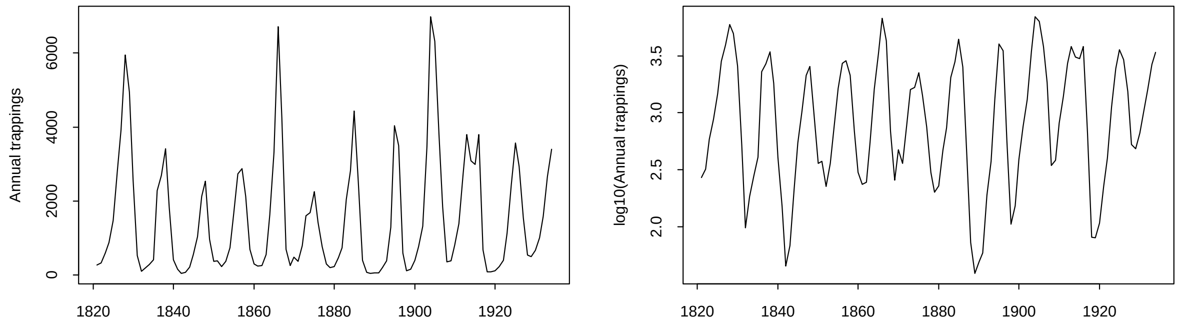


Figure 2: The Candian lynx data set: The number of annual trappings of lynx in the years 1821–1934 in the Canadian Mackenzie River district.

The lynx data is a classical time series data set, which has been analysed several times since Elton & Nicholson (1942) introduced it in the statistical literature. The number of trappings shows a clear periodicity which is often explained by a predator-prey relationship, the prey being the snowshoe rabbit. Our ambition is not to review and compare the different approaches to analysing the lynx data, we will merely use this data set to compare an ARMA approach and a state space approach, and examplify our points of view stated above. An overview of published analyses of the lynx data set is given by Campbell & Walker (1977) with discussions, and Tong (1990).

An example of an autoregressive model for the lynx data is presented in Tong (1977). He considers models for the logarithm of the annual trappings. The use of the log transform is widespread in the published analyses, primarily to reduce the asymmetry of the series. Tong (1990) argues furthermore that his prime interest is the relative population fluctuation rather than the absolute fluctuation. A plot of the log transformed data set is given in Figure 2. Tong considers autoregressive Gaussian models of order p and selects the order by Akaike's information criteria, given by

$$\mathrm{AIC}(p) = n \log \hat{\sigma}^2(p) + 2p, \tag{4}$$

where $\hat{\sigma}^2(p)$ is the estimated innovation variance in the fitted model. The first term is thus minus two times the maximized likelihood function, and the second term is a penalising term. The model yielding the lowest AIC is the best model in terms of a balance between a good fit and parsimony in the number parameters. Tong estimates the autoregressive parameters and the innovation variance by solving the Yule-Walker equations, and finds that the AIC best model is given by $p = 11$. Redoing his calculations with the S-plus procedure `ar` we reach the same conclusion, and find that the fitted AR(11)-model is given by

$$\begin{aligned} Y_t = 2.90 &+ 1.14Y_{t-1} - 0.51Y_{t-2} + 0.21Y_{t-3} \\ &- 0.27Y_{t-4} + 0.11Y_{t-5} - 0.12Y_{t-6} + 0.07Y_{t-7} \\ &- 0.04Y_{t-8} + 0.13Y_{t-9} + 0.19Y_{t-10} - 0.31Y_{t-11} + \nu_t, \end{aligned}$$

where Y_t denotes the $\log_{10}$ transformed number of trappings at year t, and where $\{\nu_t\}$ is a white noise sequence with zero mean and estimated variance 0.0477. The observed residual variance is 0.0367. The fit of the model can be judged in Figure 3.

While Tong's model fits data very well it is not straightforward to interpret. A characteristic feature of the data is the periodic behavior, with a period of roughly 10 years. Assuming the number of trappings is proportional to the size of the population, one might pose questions like "Can we estimate a fluctuation corrected size of the population, and if so, how does this vary through the observed time period?" or "How does the fluctuation period vary through time?". In order to answer questions like these, we need a model that captures the characteristic features of the data set more directly than the autoregressive model does. An obvious approach is to model the series by a deterministic harmonic component plus random noise. Tong (1977) argues, by inspection of the spectral density estimate, that the length of the period appears to vary through time, a feature which a fixed frequency model cannot capture satisfactorily. This leads us to a model where the frequency is allowed to change gradually, which can be formulated as a state space model.

Let Y_t denote the $\log_{10}$ transformed number of trappings in year t as above. The state space model is then

$$\begin{aligned} Y_t &= \mu_t + a_t\phi(2\pi\lambda t + \gamma_t) + \nu_t, & \nu_t &\sim N(0,\sigma^2), & (5) \\ \mu_t &= \mu + \rho_\mu(\mu_{t-1} - \mu) + \omega_t^\mu, & \omega_t^\mu &\sim N(0,\sigma_\mu^2), & \\ a_t &= a + \rho_a(a_{t-1} - a) + \omega_t^a, & \omega_t^a &\sim N(0,\sigma_a^2), & (6) \\ \gamma_t &= \gamma + \rho_\gamma(\gamma_{t-1} - \gamma) + \omega_t^\gamma, & \omega_t^\gamma &\sim N(0,\sigma_\gamma^2), & \end{aligned}$$

for $t = 1, 2, \ldots, n$. The periodic function ϕ was chosen as

$$\phi(x) = \cos(x) + 0.15\cos(2x - 1.5),$$

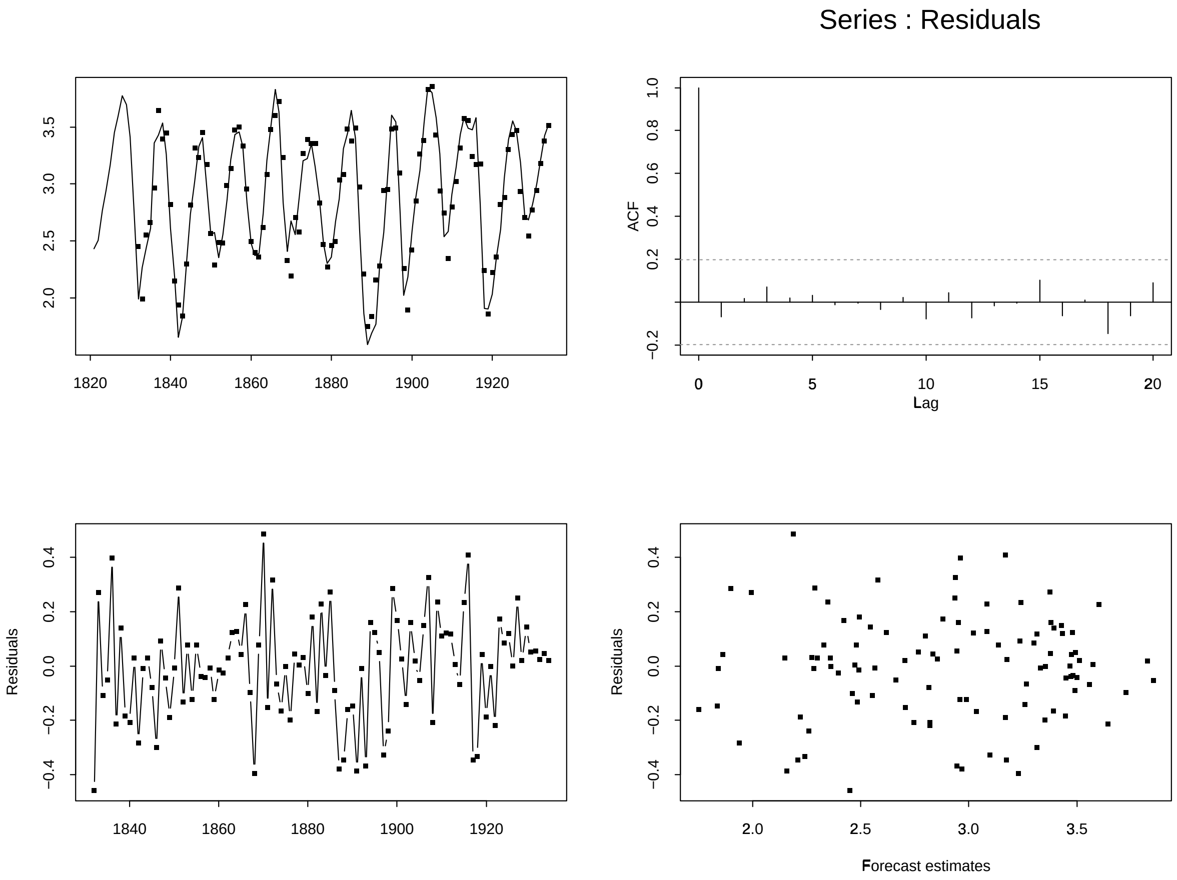


Figure 3: Diagnostic plots of Tong's AR(11)-model. Top left: The log lynx data plotted as a line and one step ahead forecasts marked by points. Top right: The autocovariance function of the residuals. Bottom left: The residuals plotted against time. Bottom right: The residuals plotted against one step ahead forecasts.

by inspection of the shape of individual cycles. The noise processes $\{\nu_t\}$,$\{\omega_t^\mu\}$, $\{\omega_t^a\}$ and $\{\omega_t^\gamma\}$ are *iid* and independent of each other. The process $\{\mu_t\}$ can be interpreted as the mean level of the series, $\{a_t\}$ is the amplitude process and $\{\gamma_t\}$ is the relative phase of the fluctuation at time t. The triple (μ_t, a_t, γ_t) constitute the unobserved latent process.

In the formulation of the model we assume that the underlying frequency λ is fixed, yet by varying the phase γ_t the observed period can vary slightly. One could also formulate a model where the frequency is allowed to change over time, a type of model which we will consider in a later section for modelling fluctuations caused by the heart rhythm in fMRI data. However, in the lynx data the frequency changes are only slight, which is the reason for choosing the model above. We chose to set $\lambda = 0.105$ years, by inspection of the periodogram of the series and by fitting the model with different values of λ. One could alternatively estimate λ by the maximum likelihood method. Notice, however, that the value of λ is not essential to the goodness of fit of the model, since if λ is chosen too small, say, the process γ_t will compensate for this by showing a positive trend.

In Section 2.3 we will explain how one can calculate an approximation to the likelihood function, estimate one step ahead forecasts and estimate the state of the latent process at different timepoints given the observations $Y_1, Y_2, \ldots, Y_n$. Later we will also explain in more detail how the estimation of the parameters is performed. We will skip the technical details in this introductionary example, and refer the reader to Section 2.3 and 2.4.

For this model we fixed the three correlation parameters at 0.95, and estimated the remaining parameters by minimizing the residual variance numerically. The reason for fixing the correlation parameters was to stabilize the maximization procedure. Correlation parameters are generally not very well estimated, and in the formulation of the model we implicitly assume that the latent processes are smoothly varying series, rather than "wiggly" less correlated series. One can thus consider the fixed values of the correlation parameters as a part of the model formulation.

The fitted model is as follows:

$$\begin{aligned}
Y_t &= \mu_t + a_t\phi(2\pi\lambda t + \gamma_t) + \nu_t, & \nu_t &\sim N(0, 0.0938^2), \\
\mu_t &= 2.96 + \rho(\mu_{t-1} - 2.96) + \omega_t^\mu, & \omega_t^\mu &\sim N(0, 0.110^2), \\
a_t &= 0.886 + \rho(a_{t-1} - 0.886) + \omega_t^a, & \omega_t^a &\sim N(0, 0.0171^2), \\
\gamma_t &= 1.69 + \rho(\gamma_{t-1} - 1.69) + \omega_t^\gamma, & \omega_t^\gamma &\sim N(0, 0.429^2),
\end{aligned} \tag{7}$$

where $\rho = 0.95$. The observed residual variance is 0.0478. The fit of the model can be judged in Figure 4. In Figure 5 are plots of estimates of the latent processes with 95% confidence limits.

It is clear from the plots, that the fit of the state space model is not as good as that of Tongs model. When comparing the two models, however, one must consider that we have estimated 8 parameters (including λ) to fit our model, while Tong has estimated 13. One might calculate the AIC for our model and use this to compare the models. Yet our main point here is not to compare the models by a single number, but to illustrate the main quality of the state space model: By fitting the model we also estimate the latent processes and thereby gain information about characterstic features of the data set. Returning to the questions posed above, we can refer to Figure 5, in which the course of the latent processes can be studied. We can see how the estimated fluctuation corrected size of the population varies through time, and how the amplitude of the fluctuations varies. Perhaps surprisingly, the model shows no coherence between the mean size of the population and the amplitude of the fluctuations. The plot of the phase shows, for instance, that the fluctuation period around 1890 is longer than the neighbouring periods around 1880 and 1900.

The example illustrates the contrast between an intuitive and natural model formulation as given by the state space model, and the more subtle formulation of the ARMA model. When formulating the state space model, one can use the plots in Figure 5 as diagnostic plots. If the models fits poorly one can examine the latent processes and evaluate if any shows an unintended course, using this as a guideline for improving the model. If, for instance, the process $\{\mu_t\}$ in the example above contained fluctuations corresponding to those of the observed series, the model for the periodicity as given by (a_t, γ_t) would not be satisfactory.

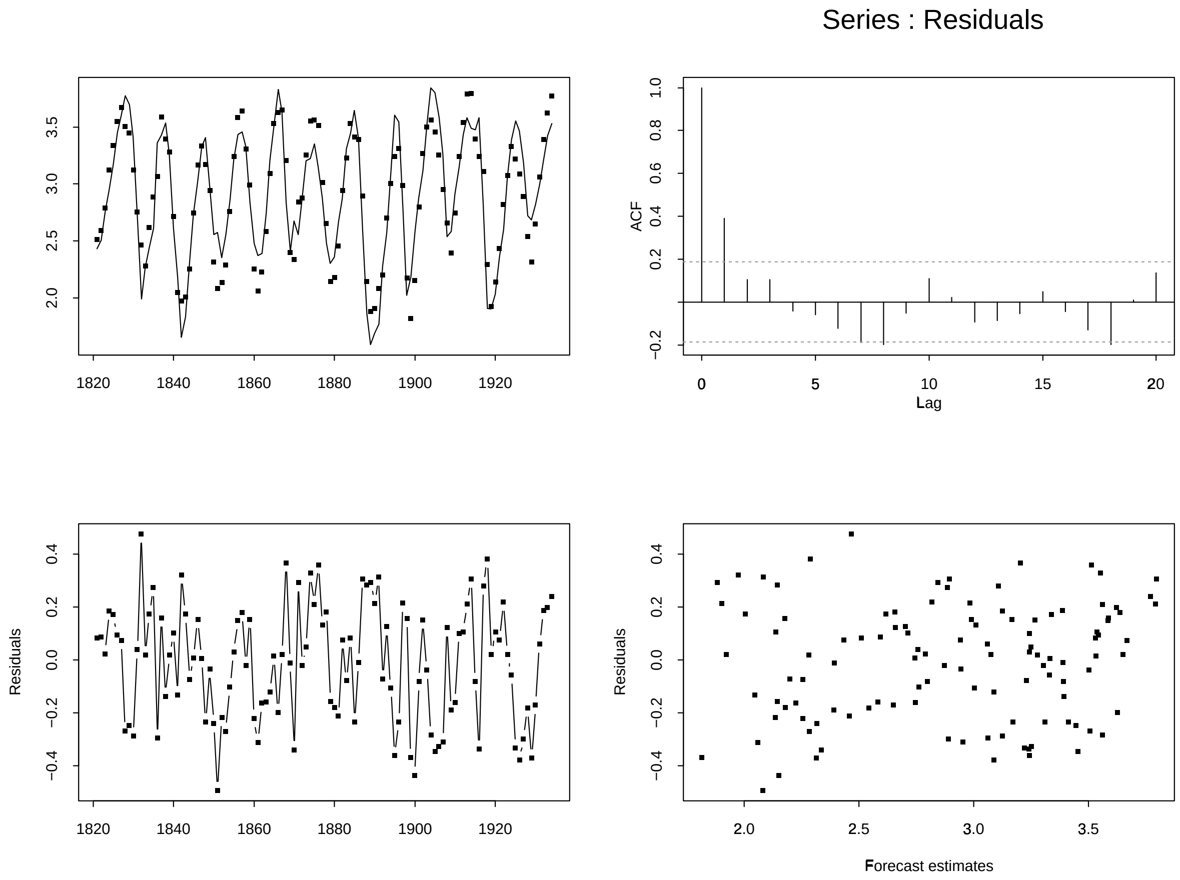


Figure 4: Diagnostic plots of the state space model. Top left: The log lynx data plotted as a line and one step ahead forecasts marked by points. Top right: The autocovariance function of the residuals. Bottom left: The residuals plotted against time. Bottom right: The residuals plotted against one step ahead forecasts.

2.2 The Kalman filter

Having introduced and motivated the use of state space models, we will now briefly introduce linear Gaussian models and the Kalman filter. The linear state space model with Gaussian errors can be formulated as

$$Y_t = F_t X_t + \nu_t, \qquad \nu_t \sim N_k(0, V_t), \tag{8}$$

$$X_t = G_t X_{t-1} + \omega_t, \qquad \omega_t \sim N_d(0, W_t), \tag{9}$$

for $t = 1, 2, \dots, n$ and $X_0 \sim N_d(m_0, C_0)$. Here F_t and G_t are known $k \times d$ and $d \times d$ matrices respectively, and the error sequences $\{\nu_t\}$ and $\{\omega_t\}$ are independent and mutually independent. Linear state space models are studied in West & Harrison (1989) who gives many examples.

Having observed $Y_1, Y_2, \dots, Y_n$ the model is often used as a basis for forecasting values of Y_t for $t > n$ or estimate the current or previous values of the latent process X_t, $t \leq n$, the latter is denoted *smoothing*. Tools for this task are the Kalman filter and the Kalman smoother, a set of recursive equations stating conditional distributions of the processes. The Kalman filter is stated in Theorem 2.1 and the Kalman smoother

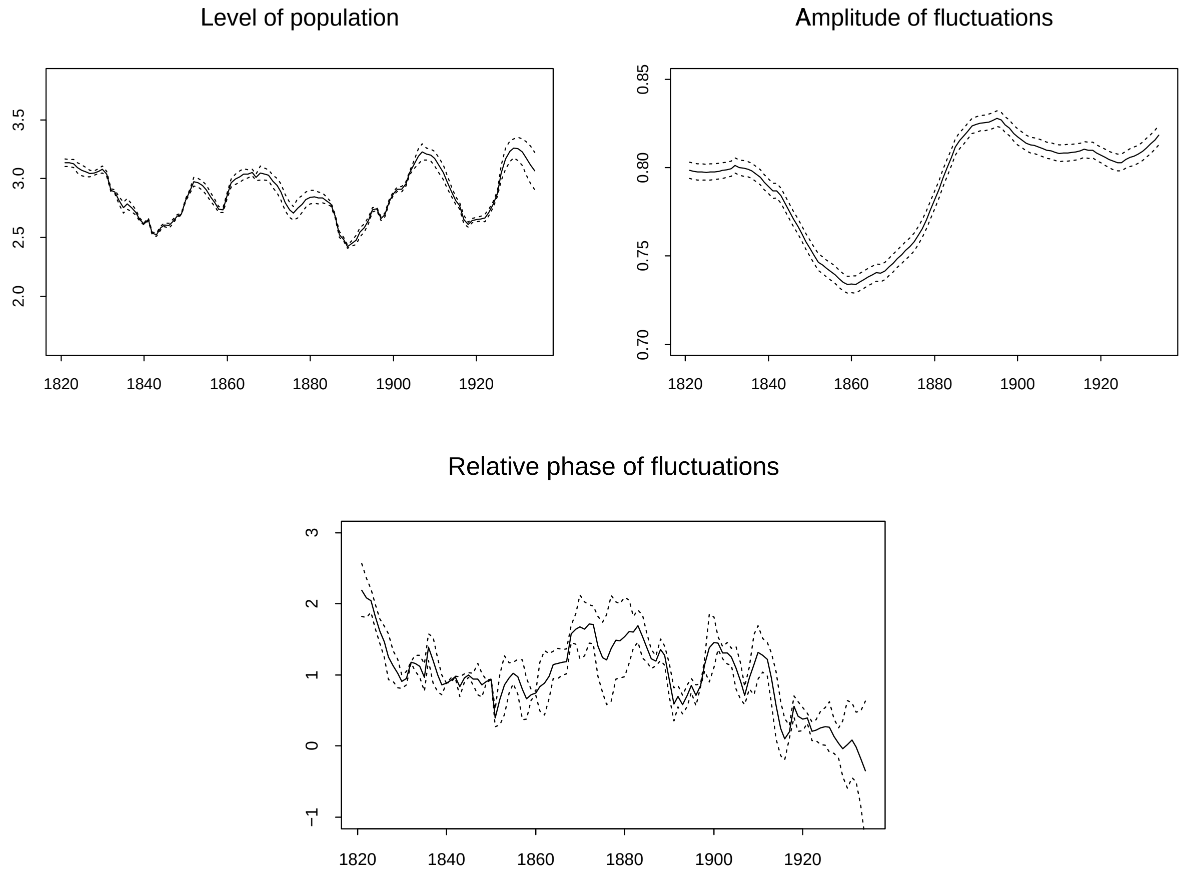


Figure 5: Estimates of the latent processes with 95% confidence limits, as given by the Kalman smoother. Top left: $\{\mu_t\}$, top right: $\{a_t\}$, bottom: $\{\gamma_t\}$.

in Theorem 2.2.

Theorem 2.1 (The Kalman filter) *Let $D_t = \sigma(Y_1, Y_2, \ldots, Y_t)$ be the information available at time t. Suppose that*

$$X_{t-1} \,|\, D_{t-1} \sim N(m_{t-1}, C_{t-1}).$$

If (8) *and* (9) *holds then*

$$\begin{aligned}
X_t \,|\, D_{t-1} &\sim N(a_t, R_t), & a_t &= G_t m_{t-1}, \\
& & R_t &= G_t C_{t-1} G_t' + W_t \\
Y_t \,|\, D_{t-1} &\sim N(f_t, Q_t) & f_t &= F_t a_t, \\
& & Q_t &= F_t R_t F_t' + V_t, \\
X_t \,|\, D_t &\sim N(m_t, C_t) & m_t &= a_t + A_t(Y_t - f_t), \\
& & C_t &= R_t - A_t Q_t A_t' \\
& & A_t &= R_t F_t' Q_t^{-1}.
\end{aligned}$$

Proof Assume that $X_{t-1} \,|\, D_{t-1} \sim N(m_{t-1}, C_{t-1})$. By (8) and (9) we get that

$$\begin{pmatrix} X_t \\ Y_t \end{pmatrix} |D_{t-1} \sim N\left(\begin{pmatrix} a_t \\ f_t \end{pmatrix}, \begin{pmatrix} R_t & R_t F_t' \\ F_t R_t & Q_t \end{pmatrix}\right),$$

which gives the two first expressions. Conditioning on Y_t above, we get

$$X_t \mid (D_{t-1}, Y_t) \sim N(a_t + R_t F_t' Q_t^{-1}(Y_t - f_t), R_t - R_t F_t' Q_t^{-1} F_t R_t) \sim N(m_t, C_t),$$

and by noting that $X_t \mid D_t \sim X_t \mid (D_{t-1}, Y_t)$ we get the last expression. □

Notice that the posterior mean m_t is an adjustment of the prior mean a_t by the error $Y_t - f_t$ scaled by A_t. The matrix A_t is denoted the *Kalman gain*. It is a measure of the influence, that the current observation Y_t has on the updating of the prior distribution $X_t \mid D_{t-1}$ to the posterior $X_t \mid D_t$. It is also worth noting, that the variance C_t does not depend on the observation Y_t. Hence if the model is constant in the sence that F_t, G_t, V_t and W_t does not depend on t, the variance of the latent process will converge to a steady level (West & Harrison 1989, Theorem 5.1). This is not the case for the non-linear models as can be seen in Figure 5.

The model can be extended to the case where the noise terms ν_t and ω_t have non-zero means. The formulas above are then changed to $a_t = G_t m_{t-1} + E(\omega_t)$ and $f_t = F_t a_t + E(\nu_t)$.

Theorem 2.2 (The Kalman smoother) *Let the notation be as in Theorem 2.1. Let $\tilde{a}_n = m_n$, $\tilde{R}_n = C_n$, and define recursively*

$$\tilde{a}_t = m_t + B_t(\tilde{a}_{t+1} - a_{t+1})$$
$$\tilde{R}_t = C_t + B_t(\tilde{R}_{t+1} - R_{t+1})B_t',$$

for $t = 1, 2, \ldots, n-1$, where $B_t = C_t G_{t+1}' R_{t+1}^{-1}$. If (9) holds and if $X_t \mid D_t \sim N(m_t, C_t)$ for $t = 1, 2, \ldots, n$, then $X_t \mid D_n \sim N(\tilde{a}_t, \tilde{R}_t)$ for $t = 1, 2, \ldots, n$.

Proof Clearly the result is true for $t = n$. Let $t < n$ and suppose that $X_{t+1} \mid D_n \sim N(\tilde{a}_{t+1}, \tilde{R}_{t+1})$. By assumption $X_t \mid D_t \sim N(m_t, C_t)$, thus by (9) we have

$$\begin{pmatrix} X_t \\ X_{t+1} \end{pmatrix} \mid D_t \sim N\left(\begin{pmatrix} m_t \\ a_{t+1} \end{pmatrix}, \begin{pmatrix} C_t & C_t G_{t+1}' \\ G_{t+1} C_t & R_{t+1} \end{pmatrix} \right).$$

By definition of the state space model $X_t \mid (X_{t+1}, D_n) \sim X_t \mid (X_{t+1}, D_t)$ and hence

$$\begin{aligned} X_t \mid X_{t+1}, D_n &\sim N(m_t + C_t G_{t+1}' R_{t+1}^{-1}(X_{t+1} - a_{t+1}), C_t - C_t G_{t+1}' R_{t+1}^{-1} G_{t+1} C_t) \\ &\sim N(m_t + B_t(X_{t+1} - a_{t+1}), C_t - B_t R_{t+1} B_t'). \end{aligned}$$

Since $X_{t+1} \mid D_n \sim N(\tilde{a}_{t+1}, \tilde{R}_{t+1})$ by the induction assumption, we then get

$$X_t | D_n \sim N(m_t + B_t(\tilde{a}_{t+1} - a_{t+1}), C_t + B_t(\tilde{R}_{t+1} - R_{t+1})B_t'),$$

and the proof is complete. □

Note that the result is only indirectly dependent on the form of (8). As long as the assumption $X_t \mid D_t \sim N(m_t, D_t)$ is fulfilled, the observation equation does not enter the expression for the smoothing density.

We can define filter residuals by $R_t = Y_t - f_t$. By Theorem 2.1, $R_t = Y_t - E(Y_t \mid D_{t-1})$ and hence $\{R_t\}$ will be an uncorrelated sequence with zero mean. Furthermore, $\{R_t\}$ will be Gaussian since $\{Y_t\}$ is. The residuals can be used for diagnostic plots to examine the fit of the model.

2.3 An approximative Kalman filter for non-linear models

In this section we will consider an approximative Kalman filter for non-linear models with Gaussian errors. The approximation to the filter is based on numerical techniques. We will consider models of the form

$$
\begin{aligned}
Y_t &= F_t(X_{t,2})X_{t,1} + h_t(X_{t,2}) + \nu_t & \nu_t &\sim N_k(0, V_t), & (10)\\
X_t &= G_t X_{t-1} + \omega_t & \omega_t &\sim N_d(b_t, W_t), & (11)
\end{aligned}
$$

for $t = 1, 2, \ldots, n$ and $X_0 \sim N_d(m_0, C_0)$. Here $Y_t \in \mathbb{R}^k$, $X_t \in \mathbb{R}^d$ and $X_t = (X'_{t,1}, X'_{t,2})'$, where $X_{t,i} \in \mathbb{R}^{d_i}$, $i = 1, 2$. The matrices F_t and G_t have dimensions $k \times d_1$ and $d \times d$ respectively, $h_t(X_{t,2})$ is a vector in $\mathbb{R}^k$ and b_t is a vector in $\mathbb{R}^d$. The noise sequences $\{\nu_t\}$ and $\{\omega_t\}$ are independent series and mutually independent of each other.

In the model formulation we assume that even though the latent process enters non-linearly in the observation equation, the model is linear conditionally on a part of the latent process, namely $X_{t,2}$. In the examples below the dimension of $X_{t,2}$, d_2, is one.

2.3.1 Example 1: The lynx model

Consider the model for the lynx data, introduced in Section 2.1,

$$
\begin{aligned}
Y_t &= \mu_t + a_t\phi(2\pi\lambda t + \gamma_t) + \nu_t, & \nu_t &\sim N(0, \sigma^2),\\
\mu_t &= \mu + \rho_\mu(\mu_{t-1} - \mu) + \omega_t^\mu, & \omega_t^\mu &\sim N(0, \sigma_\mu^2),\\
a_t &= a + \rho_a(a_{t-1} - a) + \omega_t^a, & \omega_t^a &\sim N(0, \sigma_a^2),\\
\gamma_t &= \gamma + \rho_\gamma(\gamma_{t-1} - \gamma) + \omega_t^\gamma, & \omega_t^\gamma &\sim N(0, \sigma_\gamma^2),
\end{aligned}
$$

for $t = 1, 2, \ldots, n$. In the notation above the model is given by:

$$
\begin{gathered}
X_t = (\mu_t, a_t, \gamma_t)', \qquad X_{t,1} = (\mu_t, a_t)', \qquad X_{t,2} = \gamma_t,\\
F_t(\gamma_t) = (1, \phi(2\pi\lambda t + \gamma_t)), \qquad h_t(\gamma_t) = 0,
\end{gathered}
$$

and

$$
G_t = \begin{pmatrix} \rho_\mu & 0 & 0\\ 0 & \rho_a & 0\\ 0 & 0 & \rho_\gamma \end{pmatrix}, \qquad b_t = \begin{pmatrix} \mu(1-\rho_\mu)\\ a(1-\rho_a)\\ \gamma(1-\rho_\gamma) \end{pmatrix}, \qquad W_t = \begin{pmatrix} \sigma_\mu^2 & 0 & 0\\ 0 & \sigma_a^2 & 0\\ 0 & 0 & \sigma_\gamma^2 \end{pmatrix}.
$$

□

2.3.2 Example 2: Dynamic frequency model

Consider a model for k time series. Common to all series is a fluctuation term with a frequency that varies in time. Added to this is a trend term for each time series given by a random walk. The model is given by

$$
\begin{aligned}
Y_t &= \mu_t + a\cos(v_t) + \nu_t, & \nu_t &\sim N_k(0, \Sigma),\\
\mu_t &= \mu_{t-1} + \omega_t^\mu, & \omega_t^\mu &\sim N_k(0, \sigma_\mu^2 I_k),\\
d_t &= d + \rho(d_{t-1} - d) + \omega_t^d & \omega_t^d &\sim N(0, \sigma_d^2)\\
v_t &= v_{t-1} + d_t,
\end{aligned}
$$

for $t = 1, 2, \ldots, n$. Here $Y_t \in \mathbb{R}^k$, $\mu_t = (\mu_{t,1}, \mu_{t,2}, \ldots, \mu_{t,k})'$ is the vector of trend terms and $a = (a_1, \ldots, a_k)'$ is a vector of amplitudes. $\{d_t\}$ can be interpreted as the frequency process and v_t represents the phase at time t. This model will be applied in Section 3.3 as a model for fluctuations in fMRI images caused by the pulse rhythm.

Formulating the model in the standard notation we get:

$$X_t = (\mu_{t,1}, \ldots, \mu_{t,k}, d_t, v_t)', \qquad X_{t,1} = (\mu_{t,1}, \ldots, \mu_{t,k}, d_t)', \qquad X_{t,2} = v_t,$$

$$F_t(v_t) = \begin{pmatrix} 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & 0 \end{pmatrix}, \qquad h_t(v_t) = \begin{pmatrix} a_1 \cos(v_t) \\ a_2 \cos(v_t) \\ \vdots \\ a_k \cos(v_t) \end{pmatrix}, \qquad b_t = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ d(1-\rho_d) \\ d(1-\rho_d) \end{pmatrix}$$

and

$$G_t = \begin{pmatrix} 1 & 0 & \cdots & 0 & 0 & 0 \\ 0 & 1 & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & 0 & 0 \\ 0 & 0 & \cdots & 0 & \rho_d & 0 \\ 0 & 0 & \cdots & 0 & \rho_d & 1 \end{pmatrix}, \qquad W_t = \begin{pmatrix} \sigma_\mu^2 & 0 & \cdots & 0 & 0 & 0 \\ 0 & \sigma_\mu^2 & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & \sigma_\mu^2 & 0 & 0 \\ 0 & 0 & \cdots & 0 & \sigma_d^2 & \sigma_d^2 \\ 0 & 0 & \cdots & 0 & \sigma_d^2 & \sigma_d^2 \end{pmatrix}.$$

□

A crucial assumption in the derivation of the Kalman filter in Theorem 2.1 is the linearity of the model, and it is not possible to construct exact updating equations for the models considered in this section. Instead we will consider approximate Kalman filtering. At each time point t we will approximate the posterior density $X_t \mid D_t$ by a normal density, where the approximation is calculated by numerical techniques. This approach is also studied by Frühwirth-Schnatter (1994). However, this author considers non-normal observation densities $p(y_t \mid x_t) = p(y_t \mid \lambda_t)$ which depend on the latent process through the linear predictor $\lambda_t = H_t X_t$ only. The latter has the same dimension as y_t. We restrict ourselves to Gaussian errors, but allow the observation densities to depend on the latent process in more complicated ways.

Assume at time $t-1$ the posterior distribution is given by

$$X_{t-1} \mid D_{t-1} \sim N(m_{t-1}, C_{t-1}),$$

where $D_{t-1} = \sigma(Y_1, \ldots, Y_{t-1})$. Since the state equation (11) is linear, the prior distribution is

$$X_t \mid D_{t-1} \sim N(a_t, R_t),$$

where $a_t = G_t m_{t-1} + b_t$ and $R_t = G_t C_{t-1} G_t' + W_t$, and thus the joint distribution of (X_t, Y_t) is

$$\begin{aligned} p(X_t, Y_t \mid D_{t-1}) \propto \exp \Bigg[& -\frac{1}{2}(X_t - a_t)' R_t^{-1} (X_t - a_t) \\ & - \frac{1}{2}\{Y_t - F_t(X_{t,2}) X_{t,1} - h_t(X_{t,2})\}' V_t^{-1} \{Y_t - F_t(X_{t,2}) X_{t,1} - h_t(X_{t,2})\} \Bigg]. \quad (12) \end{aligned}$$

Because of the non-linearity in (10) the distributions of $Y_t \mid D_{t-1}$ and $X_t \mid D_t$ are not normal. However, we will approximate the latter with a normal distribution,

$$X_t \mid D_t \sim N(m_t, C_t),$$

where $m_t = E(X_t \mid D_t)$ and $C_t = \operatorname{var}(X_t \mid D_t)$. The conditional moments of X_t can be calculated by numerical integration. At a first glance this seems like a computationally infeasible task, unless the dimension of the state vector is small. Even two-dimensional numerical integration can be a rather large computational burden. However, we will exploit the structure of the model in (10) where only $X_{t,2}$ enters non-linearly in the observation equation. The dimension of the integral can then be reduced to the dimension of $X_{t,2}$, which is one in our examples.

The following Lemma is the basis for this approach.

Lemma 2.1 *Assume that*

$$X_{t-1} \mid D_{t-1} \sim N(m_{t-1}, C_{t-1}). \tag{13}$$

Let $a_t = G_t m_{t-1} + b_t$ *and* $R_t = G_t C_{t-1} G_t' + W_t$. *Let* R_t *be divided as*

$$R_t = \begin{pmatrix} R_{t,11} & R_{t,12} \\ R_{t,21} & R_{t,22} \end{pmatrix},$$

where $R_{t,ij}$ *is* $d_i \times d_j$, $i, j = 1, 2$, *and let* a_t *be divided as* $a_t = (a_{t,1}', a_{t,2}')'$ *where* $a_{t,i} \in \mathbb{R}^{d_i}$. *Then*

$$\begin{aligned} X_{t,1} \mid D_{t-1}, X_{t,2} &\sim N(a_t^*(X_{t,2}), R_t^*), &(14)\\ Y_t \mid D_{t-1}, X_{t,2} &\sim N(f_t^*(X_{t,2}), Q_t^*(X_{t,2})), &(15)\\ X_{t,1} \mid D_t, X_{t,2} &\sim N(m_t^*(X_{t,2}), C_t^*(X_{t,2})), &(16) \end{aligned}$$

where

$$\begin{aligned} a_t^*(x_2) &= a_{t,1} + R_{t,12} R_{t,22}^{-1}(x_2 - a_{t,2}), & R_t^* &= R_{t,11} - R_{t,12} R_{t,22}^{-1} R_{t,21},\\ f_t^*(x_2) &= F_t(x_2) a_t^*(x_2) + h_t(x_2), & Q_t^*(x_2) &= F_t^*(x_2) R_t^* F_t^*(x_2)' + V_t,\\ m_t^*(x_2) &= a_t^*(x_2) + A_t^*(x_2)(Y_t - f_t^*(x_2)), & C_t^*(x_2) &= R_t^* - A_t^*(x_2) Q_t^*(x_2) A_t^*(x_2)',\\ A_t^*(x_2) &= R_t^* F_t^*(x_2)' Q_t^*(x_2)^{-1}. \end{aligned}$$

Furthermore, the density of $X_{t,2} \mid D_t$ *is given by*

$$\begin{aligned} &p(x_{t,2} \mid D_t)\\ &\quad \propto |C_t^*(x_{t,2})|^{1/2} \exp\Big[-\frac{1}{2}\{(m_t^*(x_{t,2})', x_{t,2}') - a_t'\} R_t^{-1} \{(m_t^*(x_{t,2})', x_{t,2}')' - a_t\}\\ &\quad - \frac{1}{2}\{Y_t - F_t(x_{t,2}) m_t^*(x_{t,2}) - h_t(x_{t,2})\}' V_t^{-1} \{Y_t - F_t(x_{t,2}) m_t^*(x_{t,2}) - h_t(x_{t,2})\}\Big]. \end{aligned} \tag{17}$$

Remark Note that if $F_t(X_{t,2})$ does not depend on $X_{t,2}$ then neither does Q_t^*, A_t^* or C_t^*. When performing a numerical integration wrt. $X_{t,2}$ we then only need to calculate these expressions once. This is the case in Example 2.

Proof By (13) we get $X_t \mid D_{t-1} \sim N(a_t, R_t)$, and hence

$$X_{t,1} \mid D_{t-1}, X_{t,2} \sim N(a_{t,1} + R_{t,12}R_{t,22}^{-1}(X_{t,2} - a_{t,2}), R_{t,11} - R_{t,12}R_{t,22}^{-1}R_{t,21}),$$

which proves (14). Conditionally on $X_{t,2}$ the model is linear, and the proof of (15) and (16) are then identical to the derivation of the Kalman filter in Theorem 2.1.

To prove (17) we consider the density $p(x_t \mid D_t)$ which is given by

$$\begin{aligned} p(x_t \mid D_t) &\propto p(Y_t, x_t \mid D_{t-1}) = p(x_t \mid D_{t-1})p(Y_t \mid x_t) \\ &\propto \exp\Bigg[-\frac{1}{2}\{x_t - a_t\}' R_t^{-1}\{x_t - a_t\} \\ &\quad - \frac{1}{2}\{Y_t - F_t(x_{t,2})x_{t,1} - h_t(x_{t,2})\}' V_t^{-1}\{Y_t - F_t(x_{t,2})x_{t,1} - h_t(x_{t,2})\}\Bigg]. \end{aligned}$$

For any $x_{t,1}$ we have that

$$p(x_t \mid D_t) = p(x_{t,1} \mid x_{t,2}, D_t)p(x_{t,2} \mid D_t).$$

The expression in (17) now follows by setting $x_{t,1} = m_t^*(x_{t,2})$ above and noting that by (16) the first term on the right hand side then is proportional to $|C_t^*(x_{t,2})|^{-1/2}$. □

Lemma 2.1 provides the following formulas for conditional moments:

$$\begin{aligned} E(X_{t,1} \mid D_t) &= \int m_t^*(x_{t,2})\, p(x_{t,2} \mid D_t)\, dx_{t,2}, \\ E(X_{t,2} \mid D_t) &= \int x_{t,2}\, p(x_{t,2} \mid D_t)\, dx_{t,2}, \\ E(X_{t,1}X_{t,1}' \mid D_t) &= \int \{C_t^*(x_{t,2}) + m_t^*(x_{t,2})m_t^*(x_{t,2})'\}\, p(x_{t,2} \mid D_t)\, dx_{t,2}, \\ E(X_{t,1}X_{t,2}' \mid D_t) &= \int m_t^*(x_{t,2})x_{t,2}'\, p(x_{t,2} \mid D_t)\, dx_{t,2}, \\ E(X_{t,2}X_{t,2}' \mid D_t) &= \int x_{t,2}x_{t,2}'\, p(x_{t,2} \mid D_t)\, dx_{t,2}, \\ E(Y_t \mid D_{t-1}) &= \int f_t^*(x_{t,2})\, p(x_{t,2} \mid D_{t-1})\, dx_{t,2}, \\ E(Y_tY_t' \mid D_{t-1}) &= \int \{Q_t^*(x_{t,2}) + f_t^*(x_{t,2})f_t^*(x_{t,2})'\}\, p(x_{t,2} \mid D_{t-1})\, dx_{t,2}, \end{aligned}$$

and the conditional distribution of Y_t given D_{t-1} can be calculated by

$$\begin{aligned} &p(Y_t \mid D_{t-1}) = \\ &\quad \int (2\pi)^{-k/2}|Q_t^*(x_{t,2})|^{-1/2} \exp\left[-\frac{1}{2}\{Y_t - f_t^*(x_{t,2})\}' Q_t^*(x_{t,2})^{-1}\{Y_t - f_t^*(x_{t,2})\}\right] \\ &\qquad\qquad\qquad\qquad p(x_{t,2} \mid D_{t-1})\, dx_{t,2}. \end{aligned}$$

In practice the integration is performed numerically. As in the linear model setup, these formulas provide:

- An approximative updating formula,
$$X_t \mid D_t \sim N(m_t, C_t),$$
where $m_t = E(X_t \mid D_t)$ and $C_t = \mathrm{var}(X_t \mid D_t)$.
- A forecast estimate given by $f_t = E(Y_t \mid D_{t-1})$.
- Filtering residuals, $R_t = Y_t - E(Y_t \mid D_{t-1})$. $\{R_t\}$ will be an approximately uncorrelated sequence, though not Gaussian.
- A contribution to the likelihood function, $p(Y_t \mid D_{t-1})$. Recall that the likelihood function is given by $p(Y_1, \ldots, Y_n) = \prod_{t=1}^n p(Y_t \mid D_{t-1})$.
- A contribution to the residual variance, $||Y_t - f_t||^2$. The residual variance is given by $\sum_{t=1}^n ||Y_t - f_t||^2$.

2.3.3 Numerical integration

A word about the implementation of the numerical integration is in order here. We have chosen a straightforward weighted sum approximation to the integral, however, the main problem is selecting the domain of the sum since the mean and variance of $X_{t,2}|D_t$ are not known explicitly. We have chosen to use the prior distribution

$$X_{t,2} \mid D_{t-1} \sim N(a_{t,2}, R_{t,22})$$

as a starting point, assuming the posterior distribution resembles this. We then select a range of values $x_{t,2}^{(i)} = a_{t,2} + i\Delta$, for $i \in J_t$. Here Δ is a fixed step length, and J_t is an index set. The integral of a function $g(x_{t,2})$ is then calculated by

$$\int g(x_{t,2}) p(x_{t,2} \mid D_t)\, dx_{t,2} \simeq \frac{\sum_{i \in J_t} g(x_{t,2}^{(i)}) \tilde{p}(x_{t,2}^{(i)} \mid D_t)}{\sum_{i \in J_t} \tilde{p}(x_{t,2}^{(i)} \mid D_t)},$$

where $\tilde{p}(x_{t,2} \mid D_t)$ is the expression in (17).

Initially the index set J_t is chosen according to the prior distribution such that $J_t = \{-K_t, \ldots, K_t\}$, where $K_t = [3R_{t,22}^{1/2}/\Delta] + 1$. Here $[\cdot]$ denotes the integer part. The values $x_{t,2}^{(K_t+1)}, x_{t,2}^{(K_t+2)}, \ldots$, are considered sequentially and if

$$p(x_{t,2}^{(K_t+j)} \mid D_t) > \varepsilon p(a_{t,2} \mid D_t), \tag{18}$$

for some small $\varepsilon > 0$, then $K_t + j$ is added to J_t. When a value $K_t + j$ is reached such that (18) does not hold, no more values are considered. The same procedure is performed with $-K_t - j$, $j \geq 1$. In our implementation we have chosen $\varepsilon = 10^{-10}$.

If $F_t(x_{t,2})$ does not depend on $x_{t,2}$ the computational burden can be reduced considerably. The expressions in Lemma 2.1 are then given by

$$\begin{array}{rlrl} a_t^*(x_{t,2}^{(i)}) &= a_{t,1} + R_{t,12} R_{t,22}^{-1} \Delta i, & R_t^* &= R_{t,11} - R_{t,12} R_{t,22}^{-1} R_{t,21}, \\ f_t^*(x_{t,2}^{(i)}) &= F_t a_{t,1} + F_t R_{t,12} R_{t,22}^{-1} \Delta i + h_t(x_{t,2}^{(i)}), & Q_t^* &= F_t^* R_t^* F_t^{*\prime} + V_t, \\ m_t^*(x_2) &= a_t^*(x_2) + A_t^*(Y_t - f_t^*(x_2)), & C_t^* &= R_t^* - A_t^* Q_t^* A_t^{*\prime}, \\ A_t^* &= R_t^* F_t^{*\prime} Q_t^{*-1}. & & \end{array}$$

All matrices which do not dependent on $x_{t,2}$, like $F_t R_{t,12} R_{t,22}^{-1} \Delta$ and A_t^* for instance, need only be calculated once.

The step length must be chosen according to the model in consideration, that is according to an average value of $\text{var}(X_{t,2} \mid D_t)$. Obviously this is a trade-off between computational speed and accuracy of the numerical approximation. For the models in example 1 and 2, $X_{t,2}$ is an angular value in radians, whose posterior distributions are essentially concentrated on an interval of length 2π. Here we have chosen $\Delta = 2\pi/15$.

2.3.4 Approximative smoothing

As remarked after Theorem 2.2, non-linearity of the observation equation does not affect the expression for the smoothing densities. By construction of the numerical integration Kalman filter we have the approximation $X_t \mid D_t \sim N(m_t, C_t)$ for $t = 1, 2, \ldots, n$, and hence the approximative smoothing densities

$$X_t \mid D_n \sim N(\tilde{a}_t, \tilde{R}_t), \qquad \text{for } t = 1, 2, \ldots, n,$$

follows directly from Theorem 2.2. Here $\tilde{a}_t$ and $\tilde{R}_t$ are given as in the theorem.

2.4 Estimation

Unknown parameters may enter the model via the matrices F_t and G_t, via the variances V_t and W_t, via b_t or via the function h_t. Finally the initial moments of the latent process m_0 and C_0 are typically also unspecified.

The initial moments can generally not be estimated by a consistent estimator in the classical likelihood sense. The reason for this is that unless we specify a model with a long-range dependency structure in the latent process, the influence of the initial prior $N(m_0, C_0)$ on the observations Y_t decreases as t increases. Suppose for instance, that the process (X_t, Y_t) is stationary and ergodic. By the theory of Markov processes the process wil converge to an equilibrium distribution independently of its initial distribution. Hence in the limit as n tends to infinity we will observe values from the equilibrium distribution, which will not provide additional information on the initial moments.

On the other hand this also implies that correct specification of the initial moments is not essential to the fit of the model. If the latent process is stationary one can chose m_0 and C_0 as the mean and variance of the process. If the process is not stationary or if the moments are difficult to calculate, one can estimate m_0 and C_0 by applying the Kalman filter on the reversed process. The final prior distribution $X_0 \mid (Y_1, \ldots, Y_n) \sim N(a_0, R_0)$ can then be used as the initial prior.

The remaining parameters may be estimated numerically by an optimisation function, such as the likelihood function $p(Y_1, \ldots, Y_n) = \prod_{t=1}^{n} p(Y_t \mid D_{t-1})$, or the residual variance $\sum_{t=1}^{n} ||Y_t - f_t||^2$ where $f_t = E(Y_t \mid D_{t-1})$. We have chosen to use the residual variance for the following reason: Suppose the model does not fit the data very well, either because we apply the Kalman filter with very "strange" parameter values, as inevitably occurs at some stage in a numerical maximization algorithm, or because the model is not correct. When the observed value Y_t differs a lot from the expected value f_t, the variance of the latent process will increase. At the next iteration, the conditional variance of the observation $\text{var}(Y_{t+1} \mid D_t)$ is thus also increased. In this way the Kalman filter might enter a state, where large differences between Y_t and f_t are reflected in an inflation of the variances rather than a proper adjustment of the mean of the latent process. By plotting the forecast estimates with the observed time series, it is quite obvious that one step ahead observations are estimated quite poorly by the Kalman forecasts. However, this is not necessarily reflected in the same degree by the likelihood function, because the large variances tend to compensate for the lack of fit. This problem does not exist for linear models, where the variance of the latent process does not depend on the observations. We regard it as a problem caused by the sequential approximations made at each iteration in the numerical integration Kalman filter, and one could approximate the likelihood function to any degree of accuracy by simulation techniques, as developed by Durbin & Koopman (1997) and references therein, by which the problem would likely be solved. This is yet to be investigated further. As an alternative we use the residual variance as a measure of the fit of the model. The variance $\mathrm{var}(Y_t \mid D_{t-1})$ does not enter this expression and hence this procedure is more stable to the instabilites of the numerical integration Kalman filter.

In order to reduce the computational burden of estimating many parameters numerically, or as a method for providing good initial values for the estimation algorithm, it is worth considering *ad hoc* methods to estimate parameters. In the Canadian Lynx model formulated in example 1 we have employed the *complex demodulation* technique (Bloomfield 1976). The model is given by

$$Y_t = \mu_t + a_t\phi(2\pi\lambda t + \gamma_t) + \nu_t,$$

where $\{\mu_t\}$, $\{a_t\}$ and $\{\gamma_t\}$ are smooth processes and $\{\nu_t\}$ is a noise term. Assume for simplicity that $\phi(x) = \cos(x)$. Complex demodulation is a non-parametric method for estimating $\{\mu_t\}$, $\{a_t\}$ and $\{\gamma_t\}$. Consider Z_t given by

$$\begin{aligned} Z_t &= Y_t \exp(-i2\pi\lambda t) \\ &= \frac{a_t}{2}\exp(i\gamma_t) + \frac{a_t}{2}\exp(-i(4\pi\lambda t + \gamma_t)) + \mu_t \exp(-i2\pi\lambda t) + \nu_t \exp(-i2\pi\lambda t). \end{aligned}$$

The first term on the right hand side can be regarded as a low-frequency component of $\{Z_t\}$, since a_t and γ_t are smoothly varying processes. The second and third terms are roughly periodic with frequency $4\pi\lambda$ and $2\pi\lambda$, respectively, and the last is a noise term containing mainly high-frequency noise—any low frequency components of the noise term can not be distinguished from the first term. The problem is now to separate the low-frequency component from the higher frequency components, which is a classical problem in signal analysis. The solution is to use a filter, that is a set of weights $\{w_{-r}, w_{-r+1}, \dots, w_0, w_1, \dots, w_r\}$, and consider

$$\tilde{Z}_t = \sum_{s=-r}^{r} w_s Z_{t-s}$$

instead of Z_t. One can design the filter so that low frequency components of Z_t are almost unaffected by the above convolution, but higher frequencies of Z_t are cancelled out. In this situation the ideal filter should eliminate all frequencies above $2\pi\lambda/2$, say, and let lower frequencies pass unaffected. For a general discussion on construction of filters see Bloomfield (1976). When the filter is applied we assume that $\tilde{Z}_t \simeq$

$a_t \exp(i\gamma_t)/2$ and extract a_t and γ_t by $a_t = 2|\tilde{Z}_t|$ and $\gamma_t = \arg(\tilde{Z}_t)$. In the same manner one can extract the mean process $\{\mu_t\}$.

Performing the above procedure on the Canadian lynx data we can filter out the amplitude and phase of the periodic component with frequency λ and the mean value component. Below are listed estimates of the mean and innovation standard deviation of these processes, based on the AR(1) model in (6) with $\rho = 0.95$.

	Mean	Std. Dev.
μ_t	2.87	0.034
a_t	0.67	0.026
γ_t	1.75	0.052

Table 1: Complex demodulation estimates for the parameters of the latent processes in the Canadian Lynx model in (6). For each process, the mean and the standard deviation of the innovations are listed, the latter calculated with $\rho = 0.95$.

The mean values correspond well to the estimated values in (7). The standard deviation estimates differ somewhat, which is not surprising, since these are more sensitive to the choice of the filter applied in the complex demodulation. Yet, the variance estimates are sensible initial values for an estimation procedure.

Notice, that in the above procedure, we actually perform a local Fourier transform at each time point t. More specifically, $\tilde{Z}_t$ is given by

$$\tilde{Z}_t = \sum_{s=t-r}^{t+r} w_{t-s} Y_s \exp(-i2\pi\lambda s)$$

which is a discrete Fourier transform of the weighted series $\{w_{t-s}Y_s\}_s$. Hence, intuitively we perform a weighted fit of a sinusoidal mean value structure with frequency $2\pi\lambda$ to $\{Y_t\}$. Pursuing this a bit further, we can use the same technique to fit an arbitrary mean value structure to the data. Consider the general model in (10),

$$Y_t = F_t(X_{t,2})X_{t,1} + h_t(X_{t,2}) + \nu_t,$$

and suppose that the latent process X_t is smooth. Non-parametric estimates $\hat{X}$ of the latent process can then be obtained by for each time point t minimizing

$$\sum_{s=t-r}^{t+r} ||w_{t-s}(Y_s - F_s(x_2)x_1 - h_s(x_2))||^2$$

with respect to $x = (x_1, x_2)$, and letting $\hat{X}_t$ be the minimumpoint. This method can be regarded as an extension of the complex demodulation technique, though with the general mean value structure, we loose the interpretation of the w_t's as a filter, and hence also loose an optimal criteria for choosing them. For fixed x_2 the minimization is linear in x_1 and hence the minimization problem is only of dimension d_2.

3 Modelling physiological noise in fMRI series

The second part of this paper is an application of non-linear state space models in functional magnetic resonance imaging. Section 3.1 is a general introduction to the subject, stating the aims of this paper. In Section 3.2 we will briefly introduce the data used in this study and in Section 3.3 we will propose a model for fMRI time series and investigate its utility to model physiological fluctuations in fMRI data. Finally in Section 3.4 we discuss the results and topics for future research.

3.1 fMRI time series

Functional magnetic resonance imaging (fMRI) is a technique where fast MR scanners are used to map the neurofunctional centres of the human brain. One of the first published fMRI experiments is Kwong et al. (1992), today, only few years later, fMRI is the most important modality in functional brain imaging. It is superior to positron emission tomography (PET) in several ways; the temporal resolution of fMRI is much better than that of PET, the spatial resolution is often better, and perhaps most importantly, with fMRI the subjects are not exposed to radiactive tracers as is the case with PET. For a short introduction to the principles of fMRI see Cohen & Bookheimer (1994). Lange (1996) gives an overview from a statistical point of view.

The basis for neurofunctional mapping with fMRI are changes in blood flow and blood oxygenation resulting from neuronal activity. Though these processes are not yet fully understood (Buxton et al. 1997) the essense of how fMRI is thought to work is the following. When neurones are activated an increase of deoxyhaemoglobin is detected in blood vessels surrounding the neurones. This is due both to an increase in blood flow and a change in the relative amount of deoxyhaemoglobin and oxyhaemoglobin. The result is a signal increase in certain types of MR images, this is known as the Blood Oxygen Level Dependence or BOLD effect. In a typical fMRI experiment the subject is exposed to an external stimulation while a sequence of MR scans of the brain is acquired. The stimulation could be a flashing light in the eye, an auditory stimulation like music or spoken words, an induced pain, a motor stimulation where the subject is instructed to tap his fingers, an odour stimulation etc. The purpose of the analysis of the images is then to localize areas in the brain where the intensity changes according to the stimulus, i.e. areas that are activated by the stimulus. A widely used experimental design is periodic stimulation, where the stimulus are presented in alternating blocks of on and off.

The data obtained by fMRI consists of a sequence of images, typically around 100, acquired with an inter-image time of less than a few seconds. Each image is represented as a matrix of two byte intensity values, usually of dimension 64 × 64 or 128 × 128. Each pixel in the sequence of images is thus a one-dimensional time series of the intensity values in a specific position in the brain. Three-dimensional information is obtained by interpolation of several two-dimensional slices of the brain.

A typical analysis of fMRI data is a marginal time series analysis. Each pixel in the images is considered after turn, and the intensity values in a single pixel in the sequence of scans, is considered as a one-dimensional time series, $\{Y_t\}$. The time series is modelled as a signal with noise added,

$$Y_t = a\varphi_t + \varepsilon_t.$$

Here φ_t is a *response function*, which is a model for the possible intensity changes due to neuronal activity at the location of the pixel, a is the amplitude of the activation signal and ε_t is a noise component. The term "noise" is to be interpreted very generally here, meaning the mean value structure of the series and the stochastic variation. A measure of the signal amplitude in the time series is calculated, this is either an estimate of a or a test statistic for the hypothesis $a = 0$, and this is used to judge wether the pixel represents neurons that are activated by the presented stimulus or not.

Two main problems arises with this approach: 1) How should the model for ϕ_t and ε_t be formulated? and 2) How should we determine if a pixel is significantly activated? Even if the theoretical distribution of the test statistic in each pixel is known, the second question is not trivial. By considering individual pixels, rather than the entire image, we effectively perform thousands of tests for the same null hypothesis, and hundreds of pixels will be classified as activated by chance if we perform a test at a level of 5%. A further complexity is the spatial correlation between the pixels, which makes correction of the significance level difficult. Current approaches to addressing this problem is to model the field of test statistics as a random field under the null hypothesis, and use probabilistic results on geometry of the field to detect significant peaks or clusters in the image (Worsley 1995). Alternatively a smaller group of pixels rather than the entire brain might be tested for activation, where the selection of the pixels is guided by neurological hypotheses (Lange & Zeger 1997).

In this paper we will address the first question. The response function has received much attention in the literature. Generally the signal reflects dynamic changes of blood oxygenation and blood volume and as mentioned above, these effects are not fully understood at the moment. In the simplest models the response function is a square wave function which is one when the stimulus is presented and zero elsewhere. Due to an inevitable delay (in the order of 4-8 sec.) from the onset of neuronal acitivity to the maximal haemodynamic response, this is, however, not a satisfactory model. Lee et al. (1995) estimated the delay by modelling the reponse as a linear combination of a sine and a cosine term, with period equal to the stimulation period. Bullmore et al. (1996) developed this model further by including the second and third harmonic terms in the response function. The response is thus modelled as a "smooth" periodic function, given by a truncated Fourier series. Bandettini et al. (1993) modelled the shape of the response empirically as the observed response from one or two activated pixels in the actual data set. A more specific approach is Friston et al. (1994), Friston et al. (1995), Worsley & Friston (1995) and Lange & Zeger (1997). These authors assume an additive response function, such that the response from a prolonged period of stimulation is the sum of point response functions. In Friston et al. (1994) the model for the point response function is a Poisson density function with estimated mean 7.69 sec. The authors noted that this function is very similar to a Gaussian density with mean and variance equal to that of the Poisson density, and hence a Gaussian shaped point response function was used in Friston et al. (1995) and Worsley & Friston (1995). Lange & Zeger (1997) assumed a Gamma density shape of the point response, and estimated the parameters of the density separately in each pixel.

The noise term $\{\varepsilon_t\}$ is a model for everything but the signal, i.e. a model for an unactivated time series. Most fMRI time series exhibit a drift, which is typically modelled as a linear trend term in the mean of $\{\varepsilon_t\}$. The trend arises from slight patient motion during the experiment, instrumental instability and perhaps from several physiological sources, for instance changes in blood pressure level. Though the images are temporally aligned, artifacts from subject motion can seldom be completely removed. The choice of a linear trend term seems somewhat *ad hoc* and more general trend models have been proposed, for instance in Holmes et al. (1997) who model the trend as a linear combination of low frequency trigonometric functions.

The remaining variation is often modelled by a stationary Gaussian process. Examples are given in Bullmore et al. (1996), where a first order autoregressive process is fitted to selected time series, and Lange & Zeger (1997), who model the variation as a general stationary Gaussian process. However, a large part of the variation is physiological noise due to cardiac and respiratory processes, which could be more specifically modelled through the mean value of the series than by formulating a general covariance structure. Some authors have designed digital filters to reduce these pulsations. Biswal et al. (1996) recorded cardiac and respiratory rhythms externally and designed band-reject Gaussian filters to reduce the physiological fluctuations, and Buonocore & Maddock (1997) estimated the pulsation rhythms directly from the images and designed Wiener filters to reduce the fluctuations. A filtering approach, however, has the disadvantage that cardiac and respiratory rates are not necessarily constant during the experiment. In that case the physiological noise components contain a wide range of frequencies. Buonocore & Maddock (1997) recognizes this by allowing up to 64 different frequencies in the cardiac and respiratory frequency ranges.

In this paper we propose a new model for the noise part $\{\varepsilon_t\}$ of the time series. We wish to formulate a model, that a) incorporates more flexible trend structures, than just linear terms, and b) models physiological fluctuations more intuitively and specificly than the filter approach. Most current approaches to removing trend and physiological fluctuations are based on filters with a high degree of freedom. Since filtering introduces correlation in the time series, there is a delicate balance between reducing degrees of freedom, and removing unwanted frequencies. This is acknowledged by Buonocore & Maddock, who notes that in some situations filtering seems to be worse than doing no filtering.

Instead of applying very flexible filters we wish to estimate the fluctuation and trend pattern specificly from a few pixels in the images, and use this as a model for any time series. Our main tenet is that any trend term should be visible in pixels in the large vessels outside the brain itself, for instance in sinus sagittalis, which is a large vein circumferencing the brain in the mid-sagittal plane. Both motion artifacts and artifacts caused by physiological processes are visible in these vessels. Hence we will use time series outside the brain as templates for modelling the noise in time series representing neurones. We will accomplish this by formulating a multidimensional model for a group of pixels in sinus sagittalis and use this model to extract trend and fluctuation terms represented in this group of series. Finally these noise templates will be included in a model for any time series in the images.

3.2 The data

The data considered in this study were obtained in a visual stimulation experiment. A GE Signa System 1.5 T scanner was used and the data were obtained by an EPI sequence. In total 90 sets of scans were acquired in each experiment, each scan consisting of 5 oblique slices through the brain, with an inter-image time (repetition time) of 2 sec. Each image is represented as a 128 × 128 matrix of 2 byte intensity values, where roughly 5000 pixels correspond to cerebral tissue. The physical dimension of each pixel in the images is 1.9 mm × 1.9 mm × 5 mm. Two experiments were performed:

1. A visual stimulation experiment, where the subject was exposed to a 7 Hz flashing light in the right eye. The stimulation was presented in alternating blocks of on and off, each block of length 10 images, starting and ending with an off-block. Thus the length of one on-off cycle, or the stimulation period, was 40 sec and 4 periods were presented.

2. A baseline experiment, where the images were obtained under the same conditions as in the first experiment, only in this session no stimulation was presented. The data can thus be regarded as a noise data set, and is used to verify models for the noise part of the time series.

To correct for head movements the images were aligned sequentially. This was done by minizing the $\mathbb{L}^2$ distance between each individual image and a reference image, under all rotations and translations, and then resampling the transformed image onto a grid of pixels. Due to differences in magnetization of the tissue in the first scans compared to the rest, the first three images in each series were removed, and we thus only considered 87 images.

3.3 A state space model for noise in fMRI time series

We will consider k template pixels, represented by the k-dimensional time series $Y_t, t = 1, 2, \ldots, n$. These pixels should be located in veins outside the brain or in the ventricles, so that they contain no activation, yet are still affected by movement artifacts and physiological processes such as cardiac fluctuations.

We will assume that each pixel contains a trend term, overlayed with a fluctuation of random frequency, caused either by the cardiac og respiratory rhythm. We will model the trend terms as random walks, and the frequency as a first order autoregressive process. This is the model introduced in example 2 in section 2.3, with the slight generalisation here that we allow the fluctuations in different time series to differ in phase.

$$
\begin{aligned}
Y_t &= \mu_t + a\cos(v_t) + b\sin(v_t) + \nu_t, & \nu_t &\sim N_k(0, \Sigma),\\
\mu_t &= \mu_{t-1} + \omega_t^{\mu}, & \omega_t^{\mu} &\sim N_k(0, \sigma_{\mu}^2 I_k),\\
\delta_t &= \delta + \rho(\delta_{t-1} - \delta) + \omega_t^{\delta} & \omega_t^{\delta} &\sim N(0, \sigma_{\delta}^2)\\
v_t &= v_{t-1} + \delta_t,
\end{aligned}
$$

for $t = 1, 2, \ldots, n$. Here $Y_t \in \mathbb{R}^k$, $\mu_t = (\mu_{t,1}, \mu_{t,2}, \ldots, \mu_{t,k})'$ is the vector of trend terms and $a = (a_1, \ldots, a_k)'$ and $b = (b_1, \ldots, b_k)'$ are vectors describing the amplitudes and relative phases. The term $\{\delta_t\}$ can be interpreted as the frequency process and v_t represents the phase at time t. By constraining b_1 to zero, the phase of the fluctuations in $\{Y_{1,t}\}$ is given by v_t.

The inference in this model is based on the non-linear Kalman filter described in Section 2.3. Assuming $\Sigma = \sigma^2 I_k$, there are $2k+4$ parameters in the model, which can be estimated numerically by minimizing the residual variance as described earlier. The Kalman smoother can then be applied to the fitted model to extract estimates of the k trend terms, and of the fluctuation pattern.

To investigate the structure of pure noise time series, we considered the baseline data set, and applied the state space model to six pixels, three located in the posterior sinus sagittalis in the neck, and three located in the anterior part at the forehead. The pixels were chosen by visual inspection of the images. In Fig. 6 is a plot of the corresponding time series. As can be seen from this plot there is a strong periodic effect in the series. This is caused by the cardiac rhythm, which is aliased to a lower frequency than the pulse rate due to the relatively long period of 2 sec. between two consecutive images. The estimated parameters can be seen in table 1. In Fig. 6 is a plot of the smoothed mean value estimates with the observed series. As can be seen from the plot, the fit of the model might be improved—the model for the periodic component does not meet the observed periodicity completely. One could improve the model by including a more flexible periodic function than the cosine terms, however, we will direct our attention to the fact, that the frequency of the periodic component is estimated satisfactorily and this is the main purpose for fitting the model. A model with more parameters would necessitate a computationally more demanding and less robust estimation procedure and since the proposed model satisfies the purpose of estimating latent trends and fluctuations, we will refrain from extending it. A plot of the estimated latent processes is given in Fig. 7. We will denote the estimated trends $\hat{\mu} = (\hat{\mu}_1, \ldots, \hat{\mu}_k)$ and the estimated fluctuation phase $\hat{v}$.

a_1	99.74			σ_μ	3.85
a_2	116.21	b_2	-25.96	δ	11.02
a_3	87.97	b_3	-52.93	ρ	0.88
a_4	-47.16	b_4	26.95	σ_δ	0.0227
a_5	-69.38	b_5	50.86	σ	16.16
a_6	46.68	b_6	-19.28		

Table 2: Estimated parameters in the state space model

We will now incorporate the estimated latent processes in a model for any time series X in the image,

$$X = D\beta + \varepsilon, \qquad \text{where } \varepsilon \sim N_n(0, \Sigma). \tag{19}$$

Here the design matrix D is of dimension $n \times d$, where $d = k + 13$. The columns of D are given by the following terms.

- The k estimated trends $\hat{\mu}_1, \ldots, \hat{\mu}_k$ and a constant term, $(1, 1, \ldots, 1)' \in \mathbb{R}^n$.

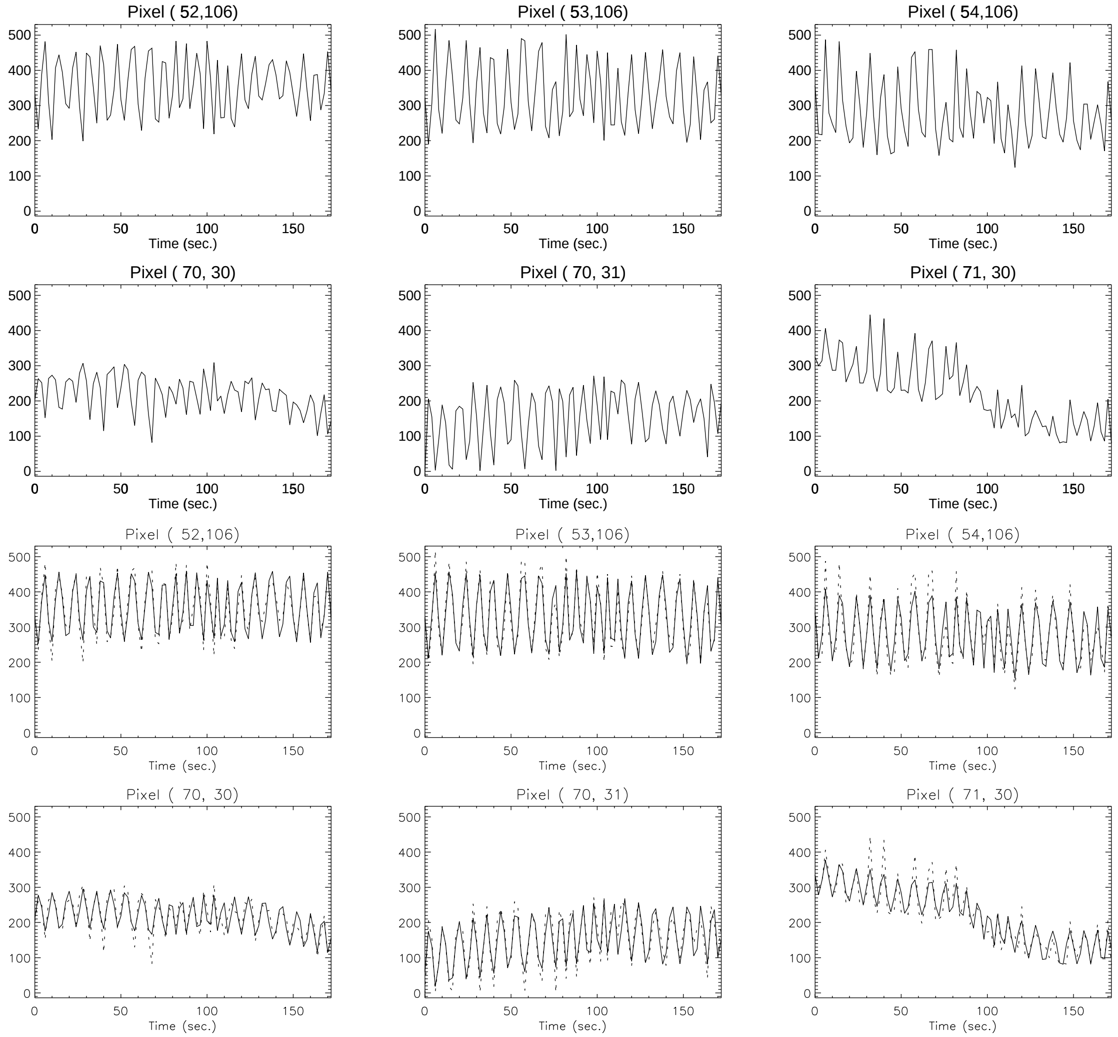


Figure 6: Top: Plot of the six template time series, used to estimate cardiac rate and trend terms. All pixels are located in sinus sagittalis. Bottom: Plot of the noise series (dotted line) with smoothed mean values (solid line).

- Six cardiac fluctuation terms of the form

  $$\cos(\hat{v}_t), \sin(\hat{v}_t), \cos(2\hat{v}_t), \sin(2\hat{v}_t), \cos(3\hat{v}_t), \sin(3\hat{v}_t),$$

  for $t = 1, .., n$.

- Six terms comprising the response function

  $$\cos(2\pi t/T), \sin(2\pi t/T), \cos(2\pi 2t/T), \sin(2\pi 2t/T), \cos(2\pi 3t/T), \sin(2\pi 3t/T),$$

  for $t = 1, \ldots, n$, where T is the stimulation period.

The column space of D is deliberately chosen quite large, since a wide range of mean value structures are observed in fMRI time series, dependent on the physical location of the pixel. As can be seen in Fig. 6, we can not expect even neighbouring pixels to show

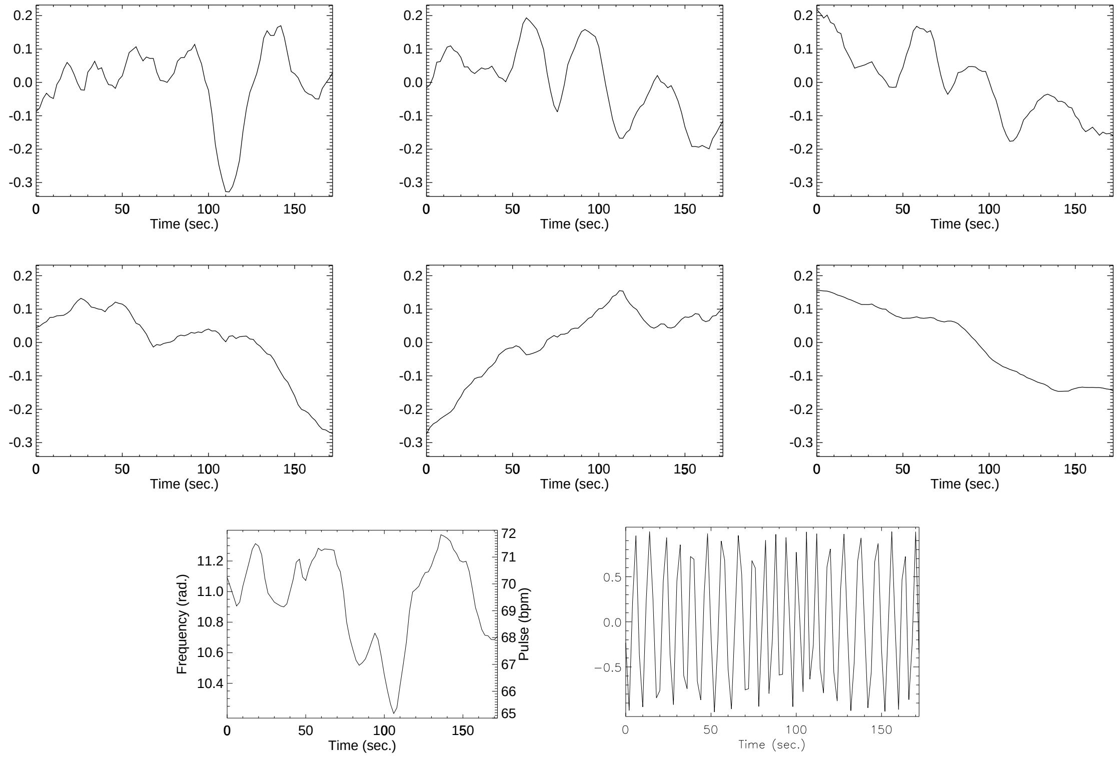


Figure 7: Plots of smoothed latent processes based on the series in Fig. 6. Top: Plots of the six estimated trend terms. Bottom: Plots of the estimated cardiac rate, and the corresponding fluctuation pattern expected in the series.

similar trend patterns, since artifacts from movement and physiological processes may enter the series differently, depending on the surrounding tissue. This is the reason for including several trend terms in the mean. Likewise the shape of cardiac fluctuations may vary from one series to another, and thus a flexible fluctuation structure is also included, through the six sine and cosine terms.

To investigate the qualities of the model, we excluded the six activation terms, and fitted the model to the baseline data set, containing only noise series. Initially we chose $\Sigma = \sigma^2 I_n$ and estimated the parameters by ordinary least squares. While it is necessary to formulate a very large mean value space in order to model every fMRI time series, the dimensionality of the model can be greatly reduced once a specific series is considered. In order to do so, we sequentially excluded column vectors of the design matrix by the principle of minimizing Akaike's information criteria (Akaike 1974). For a general model $\mathcal{P} = \{P_\theta, \theta \in \Theta\}$ where $\Theta \subseteq \mathbb{R}^d$ the criteria is given by

$$AIC(\mathcal{P}) = -2 \log L(\hat{\theta}) + 2d,$$

where $L(\hat{\theta})$ is the maximized likelihood function. By selecting the minimum AIC model we obtain a model that fits data well, yet are parsimoniously parametrized. The reduction procedure can equivalently be regarded as a sequence of likelihood ratio tests for parameters equal to zero, accepting the reduction with a quite conservative level of significance. In the 5353 time series corresponding to cerebral tissue, the dimensionality

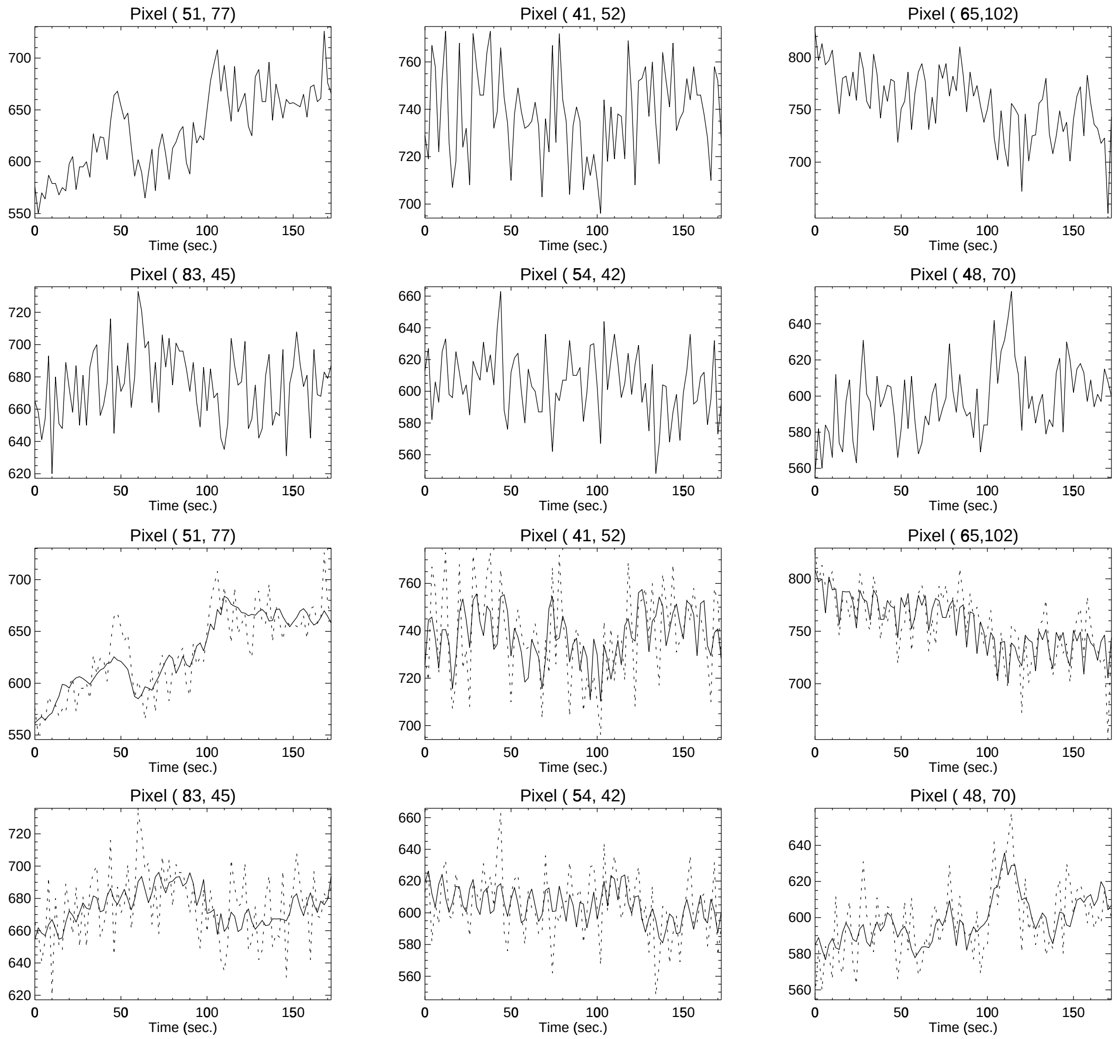


Figure 8: Plots of six different time series in the baseline data set. Top: Plots of the six series. Botton: Plots of the series (dotted line) with the fitted mean value (solid line), as given by the model in (19)

of the mean value space was on average reduced to 4.8. None of the six proposed trend terms seemed superfluous since each was present in around 2200 reduced models. The fit of the model in six time series located in different regions of the images can be judged in Fig. 8.

We examined the independence assumption of the errors by considering the residuals $R = X - D\hat{\beta}$, where $\hat{\beta}$ is the maximum likelihood estimate of β. If the model is appropriate, the residuals should be almost uncorrelated. This can be verified by the Box-Pierce-Ljung test statistic (Box et al. 1994). The statistic is given by

$$\tilde{Q} = n(n+2)\sum_{k=1}^{K} r_k^2(R)/(n-k),$$

where $r_k(R)$ denotes the lag k sample autocorrelation of the residuals. If the residual process is white noise, $\tilde{Q}$ will be asymptotically $\chi^2(K)$ distributed as n and K tends

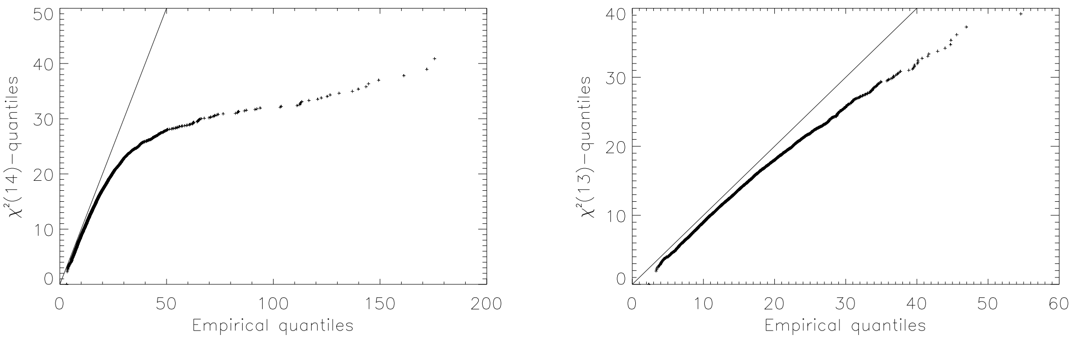


Figure 9: Plot of the theoretical χ^2-quantiles vs. the observed quantiles of the Box-Pierce-Ljung statistic in the 5353 time series. Left: Model with independent errors. Right: Model with AR(1) errors.

to infinity and K/n tends to zero. If more generally the residuals are calculated by fitting an ARMA(p,q) model, the asymptotic distribution of $\tilde{Q}$ is a $\chi^2(K-p-q)$ distribution. We calculated the statistic in every pixel with $K = 14$. Fig. 9 displays a plot of the theoretical $\chi^2(14)$ quantiles vs. the empirical quantiles of $\tilde{Q}$. The plot shows that the observed values of $\tilde{Q}$ are generally much larger than expected, which indicates autocorrelation in the errors. We therefore assumed an AR(1) model for the error terms, that is

$$\varepsilon_t = \rho\varepsilon_{t-1} + \nu_t, \quad \nu_t \sim N(0,\sigma^2),$$

for $t = 1, 2, \dots, n$, where $\{\nu_t\}$ is a white noise process. The model was re-fitted with this covariance structure, the estimation now based on the conditional likelihood obtained from the conditional density of $X_2, \dots, X_n$ given X_1. The maximum likelihood estimates of β, σ^2 and ρ can be obtained by iteratively estimating (β, σ^2) by generalized least squares, using the current value of ρ to determine the covariance structure, and estimating ρ from the obtained residuals (Bloomfield 1991). In practice the generalized least squares estimates can be calculated by considering the transformed variables,

$$X_t^* = X_t - \rho X_{t-1}, \qquad \text{and} \qquad D_t^* = D_t - \rho D_{t-1}, \qquad \text{for } t = 2, 3, \dots, n,$$

where D_t is the t'th row of the design matrix D. The model is then

$$X^* = D^*\beta + \nu, \qquad \text{where } \nu \sim N_{n-1}(0, \sigma^2 I_{n-1}),$$

and $\hat{\beta}$ can be calculated by ordinary least squares. The correlation is then estimated by

$$\hat{\rho} = \sum_{t=2}^{n} R_t R_{t-1} / \sum_{t=1}^{n-1} R_t^2,$$

where R are the residuals, $R = X - D\hat{\beta}$. Given an estimate $\hat{\rho}$ of ρ, the transformed variables X^* also provide the basis for calculating approximately uncorrelated residuals, namely $R^* = X^* - D^*\hat{\beta}$. We found that the iterative procedure converged fast; for most series four iterations were sufficient. To reduce computational complexity we therefore chose to perform four iterations in all series, rather than iterating until convergence. It should be noted also that in order to simplify the computations, the reduction of each model by the minimum AIC procedure was performed under the assumption of independent errors, even in autoregressive model. This is not of great concern, however: Due to positive correlation in the series, the variance of the parameter estimates tend to be underestimated in the *iid*-model, and hence the reduction tests are generally more conservative in the *iid*-case than in the AR(1)-case. Hence columns of the design matrix eliminated under the *iid* model would generally also be eliminated in the autoregressive model.

We refitted the model by the described procedure and performed Box-Pierce-Ljung tests for independence in the R^* series. The corresponding qq-plot is shown in Fig. 9. The plot shows that we have reduced the correlation in the residuals significantly by modelling the errors as an AR(1) process. Yet, the observed values of the statistic are generally slightly larger than expected, indicating that the correlation structure is not completely desribed by the AR(1) model. However, since the overall picture is not disturbing we will aovid introducing further complexity in the model.

In Fig. 10 are diagnostic plots of the R^* residuals of the six time series shown in Fig. 8. The plot of normal quantiles vs. empirical quantiles does not conflict with the normality assumption, and the plot of the sample autocorrelation functions confirms the impression from the Box-Pierce-Ljung tests that the residuals are mainly uncorrelated.

From the diagnostic plots, the proposed model seems to give a reasonable description of the noise in fMRI time series. To compare the model with alternative approaches, we considered models proposed by Bullmore et al. (1996) and Holmes et al. (1997). The authors of the former paper suggest a linear trend model where the remaining variation is described by an AR(1) process, that is

$$X_t = \alpha + \beta t + \varepsilon_t, \qquad \text{where} \qquad \varepsilon_t = \rho\varepsilon_{t-1} + \nu_t, \quad \nu_t \sim N(0, \sigma^2), \tag{20}$$

where $\{\nu_t\}$ is a white noise process. These authors proposed to model the activation term by a linear combination of sine and cosine terms of the fundamental stimulation frequency and the second and third harmonics, a model which we have adopted in our formulation (19). The authors carefully examined the fit of this model, with activation terms included, in 156 fMRI time series. They reached the conclusion, that the model gives an adequate description of observed data.

Holmes et al. (1997) proposed to model the trend by a linear combination of low frequency cosine terms,

$$X_t = \alpha + \beta t + \sum_{k=1}^{K} \gamma_k \cos(k\pi t/n) + \varepsilon_t,$$

where K is chosen such that the period of the cosine terms are well above that of the experimental paradigm. These authors account for the correlation of the errors by a different procedure than the AR(1) model applied in this paper, however, to be able to compare their trend model with ours, we will consider the model

$$X_t = \alpha + \beta t + \sum_{k=1}^{K} \{\gamma_k \cos(k\pi t/n) + \delta_k \sin(k\pi t/n)\} + \varepsilon_t, \tag{21}$$

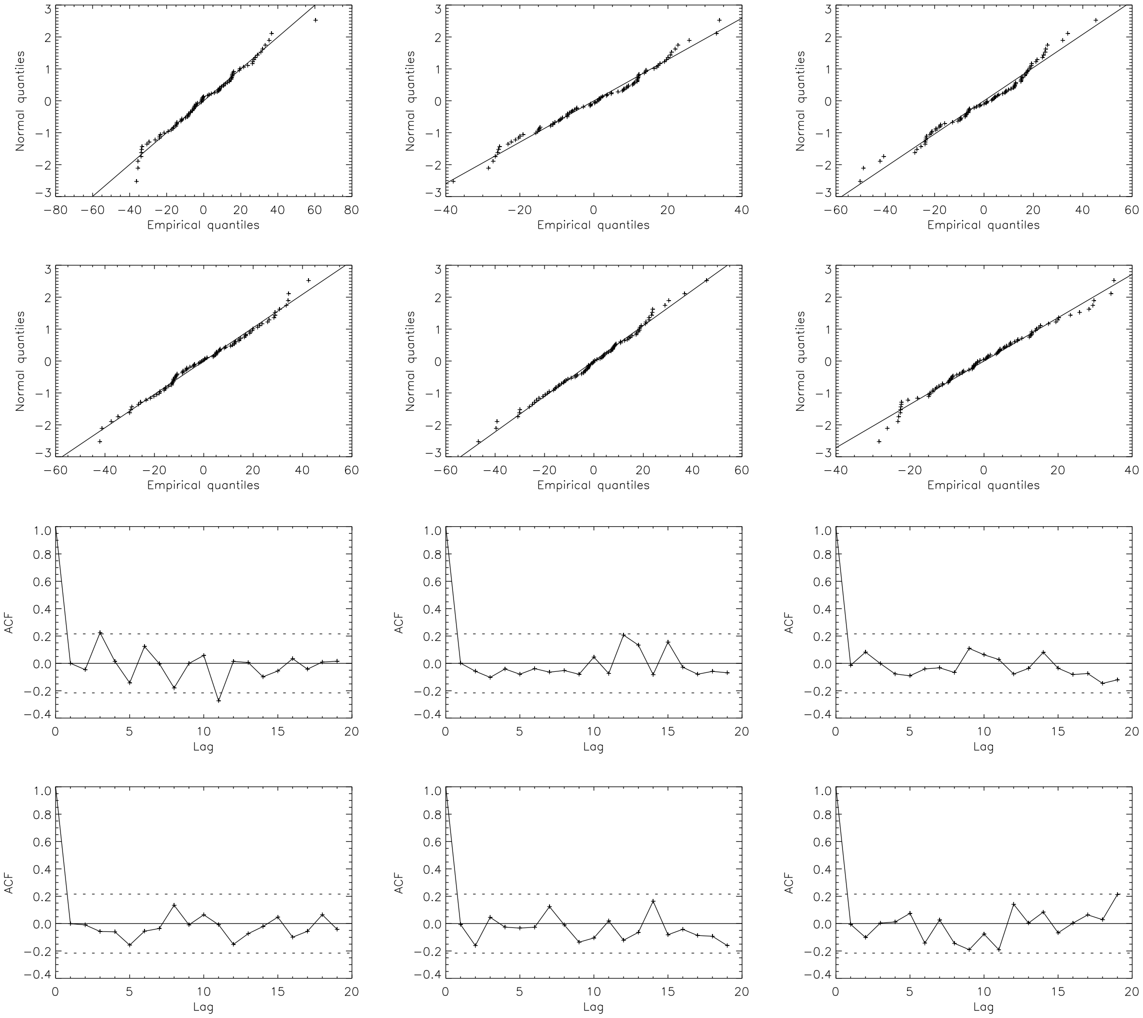


Figure 10: Diagnostic plots of residuals of the six time series displayed in Fig. 8. Top: Plots of normal quantiles vs. empirical quantiles. Bottom: Plots of sample autocorrelation functions.

where $\{\varepsilon_t\}$ is modelled as an AR(1) process. Here we have included sine terms as well to accomodate different phases of the low frequency variations. This trend model acts as a high-pass filter, eliminating components with frequencies $k/2n$ for $k = 1, \dots, K$ from the residuals. The parameter K was chosen to be 4, giving a minimal fluctuation period of 43 images, roughly twice the stimulation period of 20 images in the corresponding activation data.

We fitted these two models to the baseline fMRI data set, by the same procedure as described earlier, including the sequential elimination procedure of non-significant column vectors of the design matrix. For the Bullmore model the average dimension of the mean value space was reduced to 1.6 and for the high-pass filter model the average was reduced to 4.5. We compared the models with the proposed model by comparing the AIC in each pixel. In Fig. 11 are plots of the 5353 AIC values of our model versus each of the two alternative models. As can be seen the proposed model is the AIC

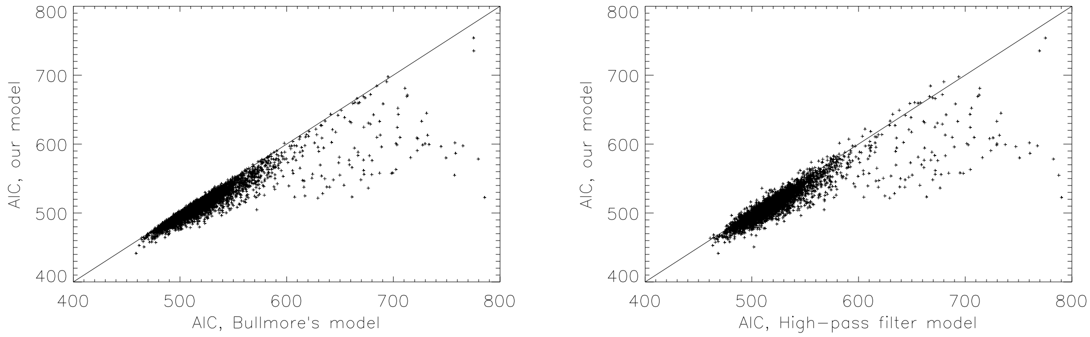


Figure 11: Comparison of the AIC values of the model in (19) and two alternative models. Left: The linear trend model in (20). Right: The high-pass filter model in (21).

best model in the majority of pixels, despite the fact that this model is generally of higher dimension than the alternative models. The Bullmore model had a higher AIC value in 4917 time series out of the 5353 series, for the high-pass model the number was 4725 series. We conclude from these plots, that even though we generally need more parameters to describe the mean value of the series, our model gives a better description of the trends and fluctuations in the series than the alternatives.

3.4 Discussion

We have established a method for estimating trends and pulsation artifacts in fMRI time series. This was performed by employing a non-linear Kalman smoother to few selected pixels located in veins outside the brain. A linear model was then proposed for any fMRI time series where the estimated trend and pulsation terms enter in the mean value structure. By applying the model to a baseline data set without neuronal activation, we have demonstrated that the proposed model describes the baseline noise satisfactorily, and by comparing AIC values we found that our model is superior to two other trend models commonly employed. This is not surprising since our model provides flexible trend terms which are effectively fitted to the specific data set in consideration. By using pixels outside the brain as noise templates we can estimate noise components in a flexible way, yet once estimated, the noise artifacts are included in a restrcitive way through few mean value vectors, ensuring that the noise is parsimoneously parametrized in each individual time series. Hence we obtain a general noise model with only a slight reduction of degrees of freedom in individual series.

Included in the noise components are possibly global activation related intensity changes, for instance activation related movement artifacts. By estimating these artifacts and including them in the noise terms, we effectively reduce the significance of detected activation in cerebral tissue. This is, however, regarded as an advantage, since the activation related intensity changes included in the noise does not result from neuronal activity, and should hence be regarded as artifacts. Any intensity changes corresponding to these artifacts should not falsely improve significance of observed neuronal activation, but rather reduce the confidence of observed activation to reflect the difficulty of seperating true activation and stimulus induced artifacts.

The proposed model gives an intuitive and interpretable way of including commonly observed artifacts in the analysis of fMRI data. For the model to be useful, we still need to develop a test statistic for the hypothesis: "The series contains no activation term". An obvious approach is to consider the likelihood ratio test for $\beta_a = 0$ where β_a is the vector of coefficients of the 6 activation terms in (19). Explorative studies of the baseline data suggest that the distribution of $-2\log Q$, where Q is the likelihood ratio statistic, is well approximated by a scaled $\chi^2(6)$ distribution. Hence we are of the opinion that a Bartlett corrected $-2\log Q$-statistic would be a appropriate. This is the subject of further research.

It would also be of interest to consider the spatial correlation in the images, which could be studied through the residuals from the fitted model. This dependency could be accounted for either by formulating a multidimensional model corresponding to (19) for a group of pixels, or by modelling the spatial correlation of the marginal teststatistics, $-2\log Q$. This, as well, is the subject of further research.

The pixels selected as noise templates are chosen by visual inspection of the individual time series. All noise templates have been selected because of their strong cardiac pulsations. We have not developed criteria for selecting the pixels automaticly, neither have we investigated the optimal number of noise templates to include in the model. It is not unlikely that an objective selection procedure could be developed, for instance based on the amplitude of the cardiac fluctuation term in selected pixels. Alternatively the pixels could be selected manually by identifying larger veins on an anatomical image of the scanning plane, by pointing with the mouse. This would give a more robust selection procedure, where only pixels corresponding to vessels were selected, and pixels corresponding to cerebral tisuue or partly to background noise were avoided. This manual selection tool would also be useful in the development stage, for investigating the effect that various pixels has on the model, by inclusion as noise templates. As for the optimal number of templates, we have found that none of the six estimated trend components were superfluous and it is likely that the model could be further improved by including more trend terms. There is, however, a balance between improving the linear model and maintaining simplicity on the state space model. The estimation procedure in the state space model requires a non-neglectible amount of computer time, and for the analysis to be applicable in practice, it is desirable to keep a sparse dimension in the state space model. It is yet to be investigated if the inclusion of more pixels could improve the model to an extend that would justify the increase in computer time.

We have not included respiratory artifacts in the proposed model. This could be achieved by a similar procedure as the inclusion of cardiac effects, however respiration did not seem to be significant in this data set. It is our experience that pulsations due to respiration is visible in the ventricles and pixels from this area could be used as templates of the respiratory rhythm. This extension should be investigated further on alternative data sets.

We are somewhat concerned, that the repetition time (TR) may effect the performance of the model. By selecting a TR of 1 sec., say, we would record an image roughly once in every pulse period, which means that the observed pulse effect would be almost constant, rather than showing a fluctuation pattern. This would make the estimation of a periodic term in the state space model somewhat unstable, since it would be difficult to distinguish the fluctuation term form the trend. In this situation the cardiac effect could be totally excluded from the model, perhaps automaticly if the models falls short of identifying the effect. The problem is also relevant for the respiratory effect, and since the two effects have periods of typically 1 sec. and 5 sec., respectively, it is not obvious how to determine optimal repetition times in order to identify both physiological pulsations. Furthermore the repetition time is usually restricted by the scanner equipment and parameter settings such as the number of image slices and the resolution of the images.

The fluctuation pattern may be recorded externally and included in the state space model as covariates. This would likely resolve some of our concerns regarding the repetition time. This can be done, though is not a trivial task. Electronic measurements are impeded by the strong magnetic field in the scanner and the measured frequency series need to be carefully registrered with the image sequence. Since, furthermore, the majority of fMRI data sets do not have these covariates attached, the development of models which are not dependent on seperate cardiopulmonary measurements seems very relevant.

References

Akaike, H. (1974), ‘A new look at the statistical model identification’, *IEEE Trans. Automat. Contr.* **AC-19**(6), 716–723.

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

Biswal, B., DeYoe, E. A. & Hyde, J. S. (1996), ‘Reduction of physiological fluctuations in fMRI using digital filters’, *Magn. Reson. Med.* **35**, 107–113.

Bloomfield, P. (1976), *Fourier Analysis of Time Series: An Introduction*, John Wiley & Sons.

Bloomfield, P. (1991), Time series methods, *in* D. V. Hinkley, N. Reid & E. Snell, eds, ‘Statistical Theory and Modelling. In honour of Sir David Cox, FRS’, Chapman and Hall.

Box, G. E. P., Jenkins, G. M. & Reinsel, G. C. (1994), *Time Series Analysis: Forecasting and Control*, 3rd edn, Prentice-Hall, Inc.

Bullmore, E., Brammer, M., Williams, S. C., Rabe-Hesketh, S., Janot, N., David, A., Mellers, J., Howard, R. & Sham, P. (1996), ‘Statistical methods of estimation and inference for functional MR image analysis’, *Magn. Reson. Med.* **35**, 261–277.

Buonocore, M. H. & Maddock, R. J. (1997), ‘Noise suppression digital filter for functional magnetic resonance imaging based on image reference data’, *Magn. Reson. Med.* **38**(3), 456–469.

Buxton, R. B., Wong, E. C. & Frank, L. R. (1997), 'Dynamics of perfusion and deoxyhemoglobin changes during brain activation', *NeuroImage* **5**(4), S32.

Campbell, M. J. & Walker, A. M. (1977), 'A survey of statistical work on the Mackenzie River series of annual Canadian lynx trappings for the years 1821–1934 and a new analysis', *J. R. Statist. Soc.* **140**(A), 411–431.

Cohen, M. S. & Bookheimer, S. Y. (1994), 'Localization of brain function using magnetic resonance imaging', *Trends in neurosciences* **17**(7), 268–277.

Durbin, J. & Koopman, S. J. (1997), 'Monte Carlo maximum likelihood estimation for non-Gaussian state space models', *Biometrika* **84**(3), 669–684.

Elton, C. & Nicholson, M. (1942), 'The ten-year cycle in numbers of the lynx in Canada.', *J. Anim. Ecol.* **11**, 215–244.

Friston, K. J., Holmes, A. P., Poline, J.-B., Grasby, P. J., Williams, S. C. R., Frackowiak, R. S. J. & Turner, R. (1995), 'Analysis of fMRI time-series revisited', *NeuroImage* **2**, 45–53.

Friston, K. J., Jezzard, P. & Turner, R. (1994), 'The analysis of functional MRI time-series', *Human Brain Mapping* **1**, 153–171.

Frühwirth-Schnatter, S. (1994), 'Applied state space modelling of non-Gaussian time series using integration-based Kalman filtering', *Stat. Comp.* **4**, 259–269.

Holmes, A. P., Josephs, O., Büchel, C. & Friston, K. J. (1997), 'Statistical modelling of low-frequency confounds in fMRI', *NeuroImage* **5**(4), S480.

Jørgensen, B., Lundbye-Christensen, S., Song, P. X.-K. & Sun, L. (1997), A state space model for multivariate longitudinal count data, Research report, Department of Mathematics, Aalborg University.

Kitagawa, G. & Gersh, W. (1984), 'A smoothness priors-state space modeling of time series with trend and seasonality', *J. Amer. Statist. Assoc.* **79**, 378–389.

Kwong, K. K., Belliveau, J. W., Chesler, D. A., Goldberg, I. E. et al. (1992), 'Dynamic magnetic resonance imaging of human brain activity during primary sensory stimulation', *Proc. Natl. Acad. Sci. USA* **89**, 5675–5679.

Lange, N. (1996), 'Tutorial in biostatistics. Statistical approaches to human brain mapping by functional magnetic resonance imaging', *Statistics in Medicine* **15**, 389–428.

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

Lee, A. T., Glover, G. H. & Meyer, C. H. (1995), 'Discrimination of large venous vessels in time-course spiral blood-oxygen-level-dependent magnetic-resonance functional neuroimaging', *Magn. Reson. Med.* **33**, 745–754.

Tong, H. (1977), 'Some comments on the Canadian lynx data', *J. R. Statist. Soc.* **140**(A), 432–436.

Tong, H. (1990), *Non-linear Time Series*, Oxford Science Publications.

West, M. & Harrison, J. (1989), *Bayesian Forecasting and Dynamic Models*, Springer-Verlag, New York.

Worsley, K. J. (1995), 'Estimating the number of peaks in a random field using the Hadwiger characteristic of excursion sets, with applications to medical images', *The Annals of Statistics* **23**(2), 640–669.

Worsley, K. J. & Friston, K. J. (1995), 'Analysis of fMRI time-series revisited — again', *NeuroImage* **2**, 173–181.