research
reports

No. 414 March 2000

N. Væver Hartvig and
J. Ledet Jensen

Spatial mixture modelling of fMRI data

department of
theoretical
statistics

university of
aarhus

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
Ny Munkegade, DK-8000 Aarhus C, Denmark

March 28, 2000

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.

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

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 generalised 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.* (1995). 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 reduces the spatial resolution and tends to underestimate the height of activation peaks (Hartvig, 1999).

A quite different approach are multivariate methods, such as principal component analysis (PCA), neural networks or independent component analysis (ICA), where all time-series are modelled simultaneously. Again we will refer to Lange *et al.* (1999) for a detailed list of references. However with these methods, the physical spatial structure is ignored, in the sense that we may permute the order of the voxels in any way, without changing the estimate.

The reason for this lack of spatial models is perhaps two-fold: 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. 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.

In the first section we present the method and the models. Next we have three examples, where we demonstrate how to use the method in practice, and compare the results with the approach in EB and traditional smoothing estimates. We finally discuss our results.

1.1 The model

Let $C_i = \{i^0, i^1, \ldots, i^k\}$ denote a neighbourhood of voxel i, where $i^0 = i$ is the centre and $i^1, \ldots, i^k$ are neighbouring voxels (in a general sense). Let $a = (a^0, a^1, \ldots, a^k)$ be an activation configuration on C_i, meaning that $a^j = 1$ if voxel i^j is activated and 0 if not, for $j = 0, 1, \ldots, k$. We will use notation like A_{C_i} for $(A_{i^0}, A_{i^1}, \ldots, A_{i^k})$. Let finally—for typographical reasons— $x_i^j = x_{ij}$, $j = 0, 1, \ldots, k$.

Imagine that we have specified $P(A_{C_i} = a)$ and the densities $f(x_{C_i}|A_{C_i} =$

a) for all possible values of $a \in \{0,1\}^{k+1}$. We then have that

$$f(x_{C_i}, A_i = a^0) = \sum_{a^j \in \{0,1\}, j=1,\ldots,k} f(x_{C_i}|A_{C_i} = a)P(A_{C_i} = a). \quad (1)$$

The conditional probability we want to use for finding activated regions now becomes

$$P(A_i = 1|x_{C_i}) = \frac{f(x_{C_i}, A_i = 1)}{f(x_{C_i}, A_i = 1) + f(x_{C_i}, A_i = 0)} \quad (2)$$

In all our applications of this method we will assume that given we know which voxels are activated the responses x_i are independent. This means that we have

$$f(x_{C_i}|A_{C_i} = a) = \prod_{j=0}^{k} f(x_i^j|A_{ij} = a^j), \quad (3)$$

and

$$f(x_{C_i}, A_i = a^0) = f(x_i|A_i = a^0) \sum_{a^j \in \{0,1\}, j=1,\ldots,k} \left(\prod_{j=1}^{k} f(x_i^j|A_{ij} = a^j) \right) P(A_{C_i} = a). \quad (4)$$

Then in order to calculate $f(x_{C_i}|A_{C_i} = a)$ we need only specify the two distributions $f(x|0) = f(x|A = 0)$ and $f(x|1) = f(x|A = 1)$.

In EB the image $\{x_i\}$ consists of fundamental power quotients (FPQ), which have respectively a central and a non-central χ^2-distribution under the two activation states. If instead $\{x_i\}$ represents the estimated activity level from a regression analysis, it will be 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. In our Example 3, we find that the range of different activation levels are described well by a Gamma distribution, $(x|A = 1) \sim \Gamma(\lambda, \beta)$.

In some cases there are both positive and negative BOLD effects. We then have three densities $f(x|0)$, $f(x|1)$, and $f(x|-1)$, corresponding to no activation, positive activation and negative activation. We may handle this by running the algorithm twice to find first the voxels with positive activation and next the voxels with negative activation. Let p_+ and p_- be the probabilities of a positive and a negative activation, respectively, and let $p_0 = 1 - p_+ - p_-$. When we run the algorithm to find the positive activation,

say, we must then use the mixture density $\tilde{f}(x|0) = \frac{p_0}{1-p_+} f(x|0) + \frac{p_-}{1-p_+} f(x|-1)$ for the density given that there is no activation. Similarly, when we look for negative activation we use $\tilde{f}(x|0) = \frac{p_0}{1-p_-} f(x|0) + \frac{p_+}{1-p_-} f(x|1)$ for the density given that there is no activation.

Let us next discuss the choice of neighbourhood region C_i. If we are modelling a 2-dimensional slice we may take C_i to be the 3 by 3 square formed by pixel i together with its eight closest neighbours. Alternatively we may extend this to a 5 by 5 square, with 24 neighbours in total. When modelling a 3D volume of scans, it is natural to let the region be a $3\times3\times3$ cube. When the voxels are anisotropic an alternative would be to take C_i to be the square in the slice direction, and only the voxel on top and just below voxel i in the other direction.

Notice that in the case of, for instance, a 5 by 5 neighbourhood, there are $2^{24} = 116777216$ terms in the sum in (2). This is clearly too many to make a direct calculation feasible. Fortunately there exist simple expressions for the sum, for all the models we propose for the marginal probabilities $P(A_{C_i} = a_{C_i})$. This reduces the computation time drastically, and makes the models applicable in practice.

1.2 Models for the marginal probabilities

In this section we give three choices for the marginal probabilities $P(A_{C_i} = a)$.

1.2.1 Model 1

Perhaps the most simple choice is to take

$$P(A_{C_i} = a) = \begin{cases} q_0 & \text{if } a^0 + a^1 + \cdots + a^k = 0 \\ q_1 & \text{if } a^0 + a^1 + \cdots + a^k > 0. \end{cases} \tag{5}$$

Since there are 2^{k+1} values of a 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

$$q_1 = p2^{-k} \quad \text{and} \quad q_0 = 1 - (2 - 2^{-k})p. \tag{6}$$

The above model (5) represents the situation that we neither believe that activated regions consist of single voxels nor that they are very large. To

illustrate this consider an activated region in $\mathbb{Z}^2$ of the form

$$\begin{array}{cccc} & & 1 & \\ & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & \end{array}$$

and let C be a square. Then the region can be hit by the square in 31 positions giving rise to 26 different activated regions inside the square.

We shall be using the equality

$$\sum_{a^j \in \{0,1\}, j=1,\ldots,k} \prod_{j=1}^{k} f(x_i^j|a^j) = \prod_{j=1}^{k} \{f(x_i^j|0) + f(x_i^j|1)\}.$$

Let η denote the above product. We then find

$$\sum_{\substack{a^j \in \{0,1\}, j=1,\ldots,k \\ a^0=0}} \left(\prod_{j=1}^{k} f(x_i^j \mid a^j) \right) P(A_{C_i} = a) = q_1\eta + (q_0 - q_1) \prod_{j=1}^{k} f(x_i^j \mid 0),$$

and from the expressions (2) and (4) we get

$$\begin{aligned} P(A_i = 1|x_{C_i}) &= \frac{q_1 f(x_i|1)\eta}{q_1 f(x_i|1)\eta + f(x_i|0)\{q_1\eta + (q_0 - q_1)\prod_{j=1}^{k} f(x_i^j|0)\}} \\ &= \left\{ 1 + \frac{1}{v_i^0} \left[1 + \left(\frac{q_0}{q_1} - 1 \right) \left(\prod_{j=1}^{k} (1 + v_i^j) \right)^{-1} \right] \right\}^{-1}, \end{aligned} \tag{7}$$

where

$$v_i^j = \frac{f(x_i^j|1)}{f(x_i^j|0)} \quad j = 0, 1, \ldots, k. \tag{8}$$

This formula shows in a direct way the difference to the approach in EB. If all the neighbours are non-activated then (7) will typically be of the order

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

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

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

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

1.2.2 Model 2

Another simple choice of $P(A_{C_i} = a)$ is

$$P(A_{C_i} = a) = \begin{cases} q_0 & \text{if } s = 0 \\ \alpha\gamma^{s-1} & \text{if } s > 0, \end{cases} \tag{9}$$

with

$$s = a^0 + a^1 + \cdots + a^k$$

being the number of activated voxels in C_i. Here $\gamma = 1$ gives back the model 1 in (5), 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.

In this model we have

$$q_0 = 1 - \alpha\frac{(1+\gamma)^{k+1} - 1}{\gamma},$$

and the probability p of a voxel being activated is $p = \alpha(1+\gamma)^k$.

Using equation (4) we find

$$\begin{aligned} f(x_{C_i}, A_i = 1) &= f(x_i|1)\alpha \sum_{a^j \in \{0,1\}, j=1,\ldots,k} \prod_{j=1}^{k} f(x_i^j|a^j)\gamma^{a^j} \\ &= f(x_i|1)\alpha \prod_{j=1}^{k} \{f(x_i^j|0) + f(x_i^j|1)\gamma\}, \end{aligned}$$

and

$$\begin{aligned} &f(x_{C_i}, A_i = 0) \\ &\quad = f(x_i|0)\left\{\alpha\gamma^{-1}\prod_{j=1}^{k}\{f(x_i^j|0) + f(x_i^j|1)\gamma\} + (q_0 - \alpha\gamma^{-1})\prod_{j=1}^{k} f(x_i^j|0)\right\}. \end{aligned}$$

This gives

$$\begin{aligned} &P(A_i = 1|x_{C_i}) \\ &\quad = \left\{1 + \frac{1}{v_i^0}\left[\gamma^{-1} + \frac{1 - \alpha(1+\gamma)^{k+1}/\gamma}{\alpha}\left(\prod_{j=1}^{k}(1+\gamma v_i^j)\right)^{-1}\right]\right\}^{-1}. \end{aligned} \tag{10}$$

1.2.3 Model 3

Finally, we will consider a model of the form (9), but being more symmetric with respect to activated and non-activated voxels. We will consider the model

$$P(A_{C_i} = a) = \begin{cases} q_0 & \text{if } s = 0, \\ \alpha_1 \gamma_1^{s-1} + \alpha_2 \gamma_2^{s-k}, & \text{if } 1 \leq s \leq k \\ q_1 & \text{if } s = k+1 \end{cases} . \tag{11}$$

In this model we have

$$\begin{aligned} 1 &= q_0 + q_1 + \frac{\alpha_1}{\gamma_1}\{(1+\gamma_1)^{k+1} - 1 - \gamma_1^{k+1}\} + \frac{\alpha_2}{\gamma_2^k}\{(1+\gamma_2)^{k+1} - 1 - \gamma_2^{k+1}\}, \\ p &= q_1 + \alpha_1\{(1+\gamma_1)^k - \gamma_1^k\} + \frac{\alpha_2}{\gamma_2^{k-1}}\{(1+\gamma_2)^k - \gamma_2^k\}, \end{aligned} \tag{12}$$

where p is the probability of a voxel being activated. Instead of (10) we find

$$P(A_i = 1 | x_{C_i}) = \left\{ 1 + \frac{1}{v_i^0}\frac{N}{D} \right\}^{-1}, \tag{13}$$

where

$$\begin{aligned} N &= \frac{\alpha_1}{\gamma_1}\prod_{j=1}^{k}(1+\gamma_1 v_i^j) + \frac{\alpha_2}{\gamma_2^k}\prod_{j=1}^{k}(1+\gamma_2 v_i^j) + q_0 - (\frac{\alpha_1}{\gamma_1} + \frac{\alpha_2}{\gamma_2^k}), \\ D &= \alpha_1\prod_{j=1}^{k}(1+\gamma_1 v_i^j) + \frac{\alpha_2}{\gamma_2^{k-1}}\prod_{j=1}^{k}(1+\gamma_2 v_i^j) + \{q_1 - (\alpha_1\gamma_1^k + \alpha_2\gamma_2)\}\prod_{j=1}^{k} v_i^j. \end{aligned}$$

1.3 Estimation of parameters

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

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

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

$$\gamma(\phi, \beta) = \sum_{i \in V} \log f(x_{C_i}; \phi, \psi). \tag{14}$$

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 we get

$$\begin{aligned} & f(x_{C_i};\phi,\gamma,\alpha) \\ &= \prod_{j=0}^{k} f(x_i^j|0;\phi)\left\{\frac{\alpha}{\gamma}\prod_{j=0}^{k}(1+\gamma v_i^j(\phi))+1-\frac{\alpha(1+\gamma)^{k+1}}{\gamma}\right\}, \end{aligned} \tag{15}$$

and for model 3 we find

$$\begin{aligned} & f(x_{C_i};\phi,\psi) \\ &= \prod_{j=0}^{k}\{f(x_i^j|0;\phi)\left\{\frac{\alpha_1}{\gamma_1}\prod_{j=0}^{k}(1+\gamma_1 v_i^j(\phi))+\frac{\alpha_2}{\gamma_2^k}\prod_{j=0}^{k}(1+\gamma_2 v_i^j(\phi))\right. \\ &\quad \left.+q_0-(\frac{\alpha_1}{\gamma_1}+\frac{\alpha_2}{\gamma_2^k})+\{q_1-(\alpha_1\gamma_1^k+\alpha_2\gamma_2)\}\prod_{j=0}^{k} v_i^j(\phi)\right\}, \end{aligned} \tag{16}$$

with $\psi=(\alpha_1,\alpha_2,\gamma_1,\gamma_2,q_1)$ and q_0 given by the constraint in (12).

Usually, though, we will take a more simple approach instead of using (14). 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

$$f(x;\phi,p)=(1-p)f(x|0;\phi)+pf(x|1;\phi), \tag{17}$$

or, if we have both positive and negative activation,

$$f(x;\phi,p_+,p_-)=(1-p_+-p_-)f(x|0;\phi)+p_-f(x|-1;\phi)+p_+f(x|1;\phi). \tag{18}$$

We thus maximise the contrast function

$$\gamma_m(\phi,p)=\sum_{i\in V}\log f(x_i;\phi,p) \tag{19}$$

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

When $P(A_{C_i}=a)$ is given by model 2 we still estimate $p=\alpha(1+\gamma)^k$ from (19). 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

$$\begin{aligned} \mathrm{Cov}(X_i,X_j) &= P(A_i=A_j=1)-P(A_i=1)P(A_j=1) \\ &= P(A_i=A_j=1)-p^2. \end{aligned}$$

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

$$P(A_i = A_j = 1) = \sum_{a:a^0=a^1=1} P(A_{C_i} = a) = \alpha\gamma \sum_{a^j\in\{0,1\},j=2,\ldots,k} \gamma^{a^2+\cdots+a^k}$$
$$= \alpha\gamma(1+\gamma)^{k-1} = p\frac{\gamma}{1+\gamma}.$$

We may estimate the covariance by the correlogram (Cressie, 1991)

$$\hat{C}_{j-i} = \frac{1}{N_{j-i}} \sum_{l\in V, l+j-i\in V} (X_l - \bar{X}.)(X_{l+j-i} - \bar{X}.),$$

where V denotes the set of brain voxels, N_{j-i} is the number of terms in the sum, and $\bar{X}.$ is the average of the X_i's. Finally we may estimate γ by

$$\hat{\gamma} = \frac{b}{1-b}, \qquad \text{where } b = \hat{C}_{j-i}\hat{p}^{-1} + \hat{p}.$$

When p is given, model 3 has 4 free parameters which may be estimated from (14).

2 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 power rate (TPR) and level of significance or false power 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.

2.1 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

$$f(x_i \mid A_i = 0) = q^{x_i}(1-q)^{1-x_i}, \quad x_i \in \{0,1\},$$
$$f(x_i \mid A_i = 1) = (1-q)^{x_i}q^{1-x_i}, \quad x_i \in \{0,1\}.$$

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 summarised 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

$$f(x_i \mid A_i = 0) = \frac{1}{\sqrt{2\pi}\tau} e^{-\frac{1}{2\tau^2}x_i^2}, \quad x_i \in \mathbb{R},$$
$$f(x_i \mid A_i = 1) = \frac{1}{\sqrt{2\pi}\tau} e^{-\frac{1}{2\tau^2}(x_i-1)^2}, \quad x_i \in \mathbb{R},$$

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

For each model, the parameters were estimated both by maximising the contrast function (14) and, for model 1 and 2, by the simple estimators described in Section 1.3. 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 $A_i = 1$ given X_{C_i} in each pixel i, 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

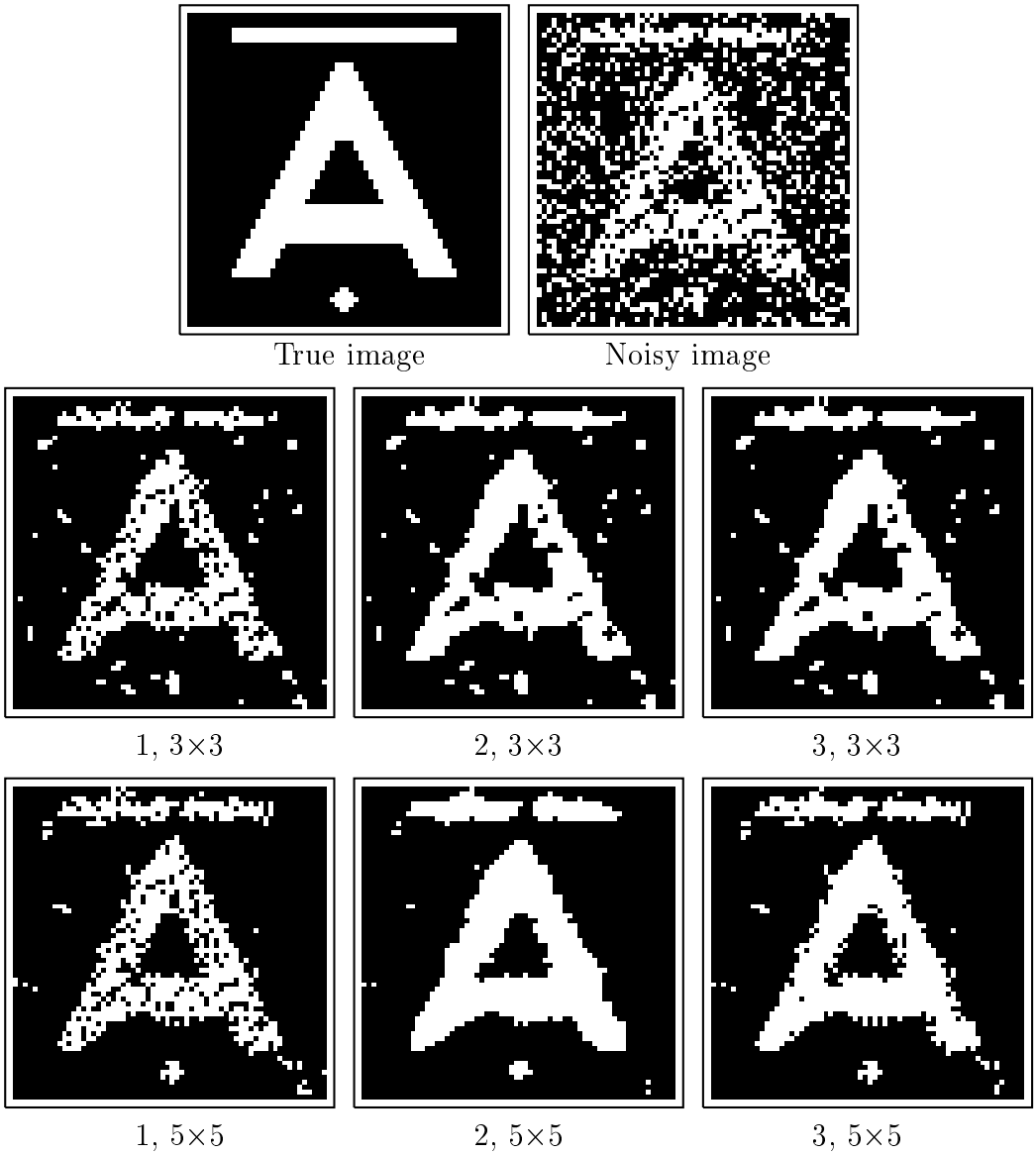

Figure 1: Comparison of spatial mixture models. Top row: True image and degraded version. Middle row: Estimates of the true image based on model 1, 2 and 3 applied to a 3 by 3 pixel region. Bottom row: Same as above, but with the models defined on a 5 by 5 region.

this example. The models 1, 2 and 3 of Section 1.2 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.

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 generalised for instance to multicolour settings.

2.2 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_i^2, 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 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.

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

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 $t = 1, \ldots, T_0$, $T_0 = 96$. 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

$$X_i = \frac{\hat{b}_i}{\sqrt{s_i^2 / SSD_x}} \quad i = 1, \ldots, V,$$

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 1.1 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 sub-datasets. 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.

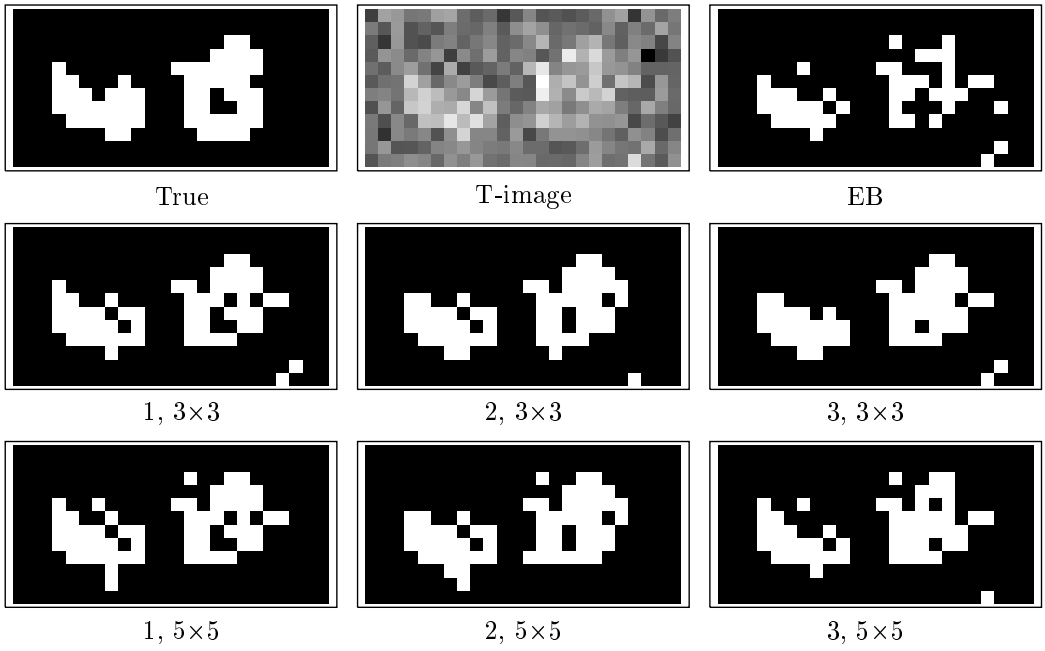

Figure 2: Activation images for the first of four subsets of the synthetic data-set. Top left and middle: True binary activation image and observed t-statistics image. The remaining are thresholded posterior probability images for the different models. EB=Everitt and Bullmore's mixture model. 1, 2 and 3: Models 1, 2 and 3 defined on a 3 by 3 region or a 5 by 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 in Figure 3. 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.*, 1995). 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.

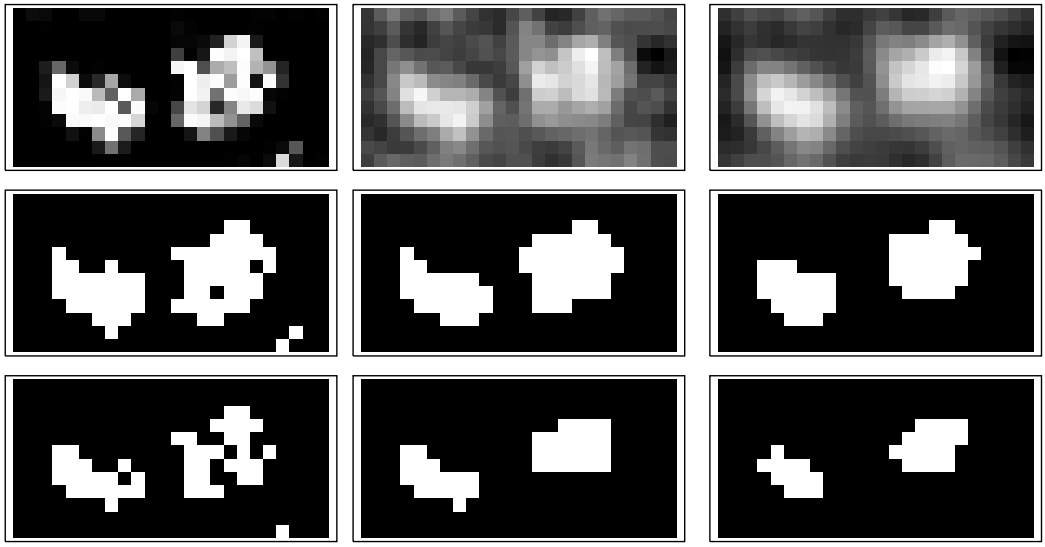

Figure 3: Activation images for the first of four subsets of the synthetic data-set. From left to right: Model 2 defined on a 3 by 3 region and the non-parametric model with FWHM 2 and 3 respectively. Top row: Original activation images. Below: Images thresholded at empirical FPR 5% (middle) and 1% (bottom).

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)

2.3 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 minimising 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

$$f(x) = p_0 f_N(x; 0, \sigma^2) + p_- f_\Gamma(-x; \lambda_-, \beta_-) + p_+ f_\Gamma(x; \lambda_+, \beta_+), \qquad (20)$$

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 maximising 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 (20) 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.

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 1.2. 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.

3 Discussion

The proposed mixture model accounts to some extent for the spatial structure of the underlying activation pattern. We found that the 3 different models worked almost equally well on synthetic and real fMRI data. In fact we tested 2 more advanced models also (see the Appendix), but they gave similar results. 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 true activation field. The parameters may easily be estimated directly from the data.

We found significant improvements compared to the non-spatial mixture model. A 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 maximise 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

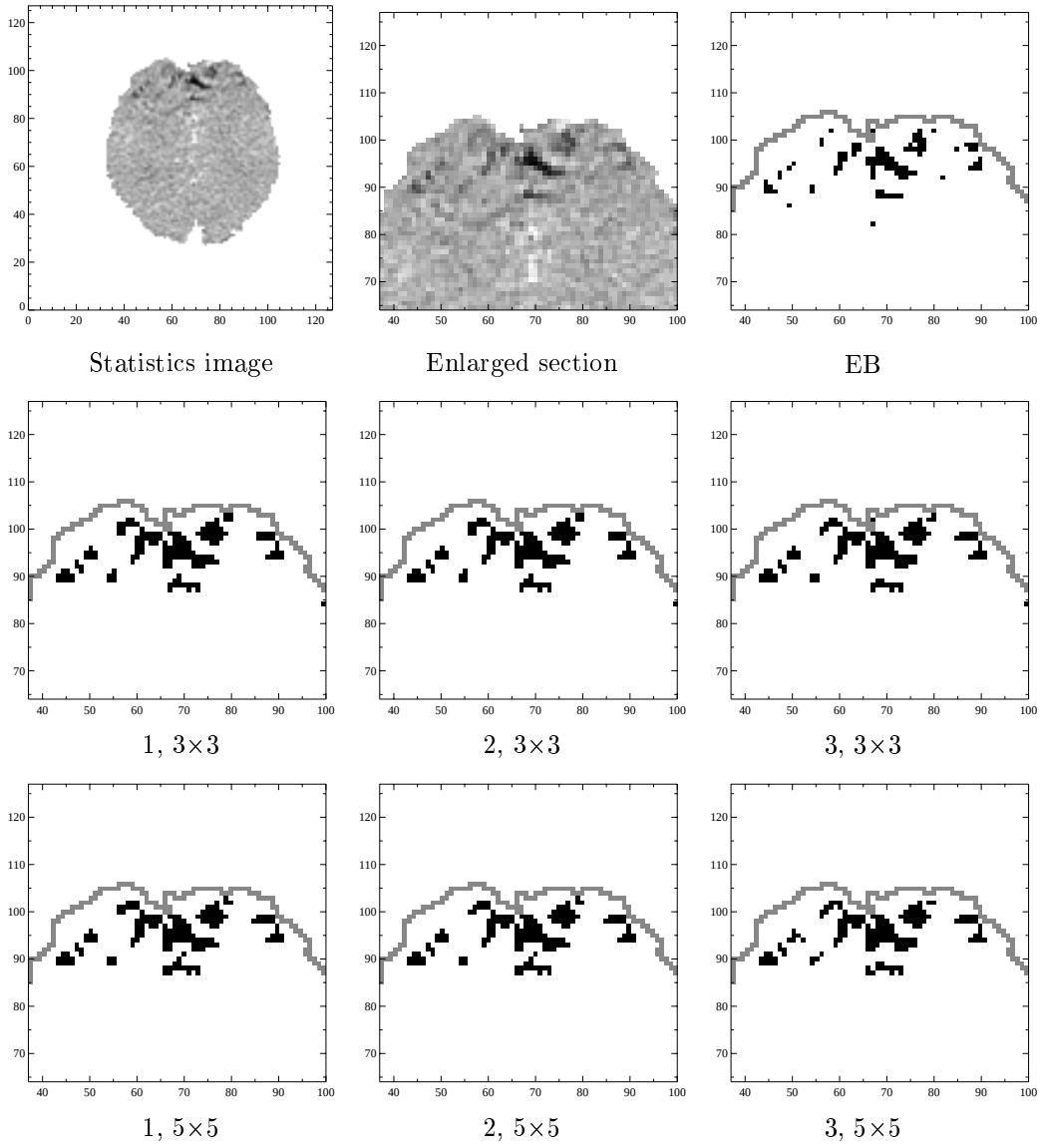

Figure 4: Comparison of estimated activation patterns for the different mixture models. Top left and middle: Raw image of t-statistics and an enlarged section of this. The remaining panels are posterior probability images thresholded at 0.5. Top right: non-spatial mixture model. Middle and last row: Models 1, 2 and 3 defined on respectively a 3×3 voxel region (middle row) and 5×5 voxel region (last row.)

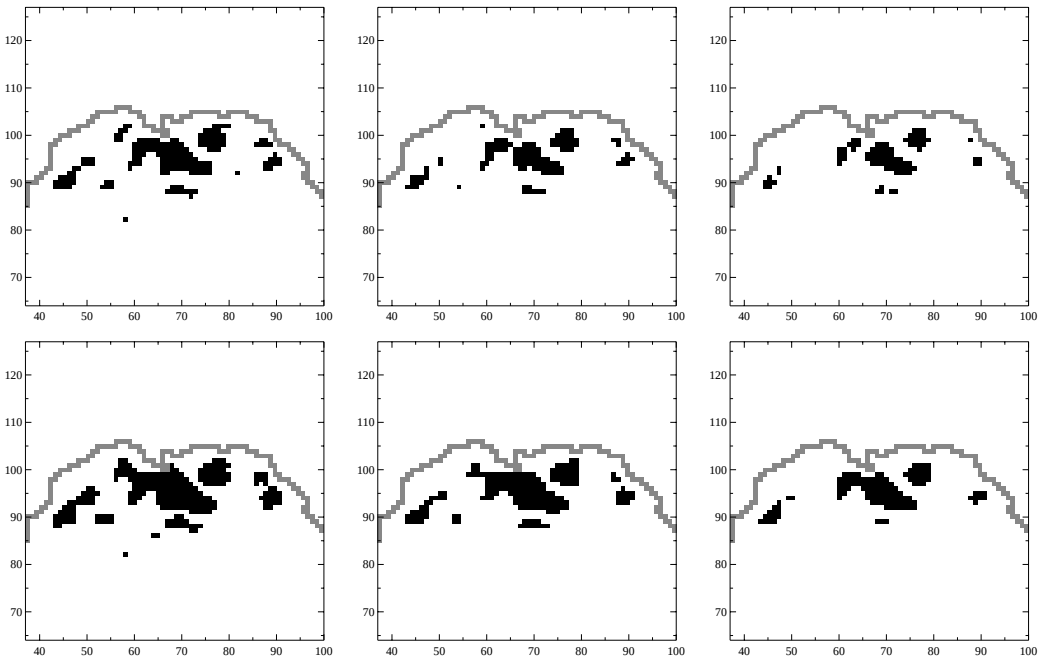

Figure 5: Images of t-statistics based on data smoothed spatially with a Gaussian kernel of FHWM 2 voxels (top row) and 3 voxels (bottom row). The images are thresholded at 5.0 (left), 6.0 (middle) and 7.0 (right).

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 emphasise 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. Also we suspect that thresholding in the mixture setting is more robust 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.

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. One may extent the methodology to correlated noise by estimating the spatial correlation first, and incorporating this in the expression for $f(x_{C_i}|A_{C_i} = a)$. Clearly the computations get more complicated then.

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.

4 Conclusion

We have formulated a simple mixture model for fMRI data which captures some 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.

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 unthresholded 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.

A Complicated marginal models

In the simple models in Section 1.2 for the marginal probabilities of the activity A_{C_i}, the probability depends only on the number of activated voxels. A more realistic model would favour connected components corresponding to cutting a part of a convex boundary. We will describe two such models in the two dimensional situation, where C_i is a square with nine pixels. We have postponed the models to this appendix, since we found that they did not yield improved estimates in our examples. Still the models, or the ideas used for constructing them, might be useful in other contexts, which is the justification for this appendix.

The first model is designed to capture situations where the activated region consists of a number of small, widely separated, convex regions, whereas the second model captures situations with long straight line boundaries between activated and non-activated regions.

We use the notation from (8) for v_i^j and as before let $s = a^0 + a^1 + \cdots + a^8$. For $j > 8$ we let $v_i^j = v_i^{j-8}$. Furthermore, we use the following numbering

$j = 1, \ldots, 8$:

1	2	3
8		4
7	6	5

of the eight boundary pixels. We will consider models of the form

$$P(A_{C_i} = a) = \begin{cases} q_0 & s = 0 \\ q_1(1 + \delta\xi(a)) & s > 0, \end{cases} \tag{21}$$

where $\xi(a)$ is non-zero for certain configurations a only.

In the first model for ξ we imagine that there are four possible shapes for activated subregions. The four shapes are

$a)$

	1	
1	1	1
	1	

$b)$

1	1	1
1	1	1
1	1	1

$c)$

	1	1	
1	1	1	1
1	1	1	1
	1	1	

$d)$

		1			
1	1	1	1	1	
1	1	1	1	1	
1	1	1	1	1	1
1	1	1	1	1	
1	1	1	1	1	
		1			

We let the frequencies of these four regions be inversely proportional to the area of the convex hull of the region. Thus the frequencies are proportional to $w = (\frac{1}{7}, \frac{1}{9}, \frac{1}{14}, \frac{1}{37})$. Most of the cases where $\xi(a) > 0$ can be described by having a consecutive sequence of activated pixels along the boundary of the square. Let this sequence be $a^I, a^{I+1}, \ldots, a^{I+l-1}$. If $a^0 = 0$ and $I \in \{1, 3, 5, 7\}$ we let $\xi(a) = \alpha_l^1$ and if $I \in \{2, 4, 6, 8\}$ we let $\xi(a) = \alpha_l^2$. If $a^0 = 1$ the corresponding values are β_l^1 and β_l^2, except for $l = 8$ where $\xi(a) = \gamma_3$. Furthermore, $\xi(a) = \gamma_2$ for the case where $s = 5$ and $a^0 = a^2 = a^4 = a^6 = a^8 = 1$ and $\xi(a) = \gamma_1$ when $s = 7$ with $a^I = a^{I+2} = 0$, $I \in \{1, 3, 5, 7\}$. We then find

$$\begin{aligned}
&\alpha_1^1 = 2w_1 + w_2 + 2w_3 + 3w_4 \quad \alpha_2^1 = w_2 + w_3 + w_4 \quad \alpha_3^1 = w_2 \quad \alpha_4^1 = w_4, \\
&\alpha_1^2 = w_1 + w_4 \quad \alpha_2^2 = \alpha_2^1 \quad \alpha_3^2 = w_1 + w_3 \quad \alpha_4^2 = \alpha_4^1, \\
&\beta_3^1 = w_1 + w_4 \quad \beta_4^1 = w_3 \quad \beta_6^1 = w_4, \\
&\beta_3^2 = w_2 + w_4 \quad \beta_4^2 = \beta_4^1 \quad \beta_5^2 = w_2 \quad \beta_6^2 = \beta_6^1 \quad \beta_7^2 = w_3, \\
&\gamma_1 = w_4 \quad \gamma_2 = w_1 \quad \gamma_3 = w_2 + 9w_4.
\end{aligned}$$

With these definitions the sum $\sum_{a:a^0=0}\prod_{j=1}^{8} f(x_i^j|a^j)\xi(a)$ when the centre pixel is zero becomes $\prod_{j=1}^{8} f(x_i^j|0)$ times

$$\begin{aligned} T_0 &= \sum_{r\in\{1,3,5,7\}} \big\{ v_i^r(\alpha_1^1 + v_i^{r+1}(\alpha_2^1 + v_i^{r+2}(\alpha_3^1 + \alpha_4^1 v_i^{r+3}))) \\ &\quad + v_i^{r+1}(\alpha_1^2 + v_i^{r+2}(\alpha_2^2 + v_i^{r+3}(\alpha_3^2 + \alpha_4^2 v_i^{r+4})))\big\}, \end{aligned} \tag{22}$$

and the sum $\sum_{a:a^0=1}\prod_{j=1}^{8} f(x_i^j|a^j)\xi(a)$ when the centre pixel is one becomes $\prod_{j=1}^{8} f(x_i^j|0)$ times

$$\begin{aligned} T_1 &= \sum_{r\in\{1,3,5,7\}} \Big\{ v_i^r v_i^{r+1} v_i^{r+2}(\beta_3^1 + v_i^{r+3}(\beta_4^1 + \beta_6^1 v_i^{r+4} v_i^{r+5})) \\ &\quad + v_i^{r+1} v_i^{r+2} v_i^{r+3}(\beta_3^2 + v_i^{r+4}(\beta_4^2 + v_i^{r+5}(\beta_5^2 + v_i^{r+6}(\beta_6^2 + \beta_7^2 v_i^{r+7})))) \\ &\quad + \gamma_1 v_i^{r+1} \prod_{j=r+3}^{r+7} v_i^j \Big\} + \gamma_2 v_i^2 v_i^4 v_i^6 v_i^8 + \gamma_3 \prod_{j=1}^{8} v_i^j. \end{aligned} \tag{23}$$

Also we have the two equations

$$\begin{aligned} 1 &= q_0 + q_1\{511 + \delta(21w_1 + 25w_2 + 32w_3 + 61w_4)\}, \\ p &= q_1\{256 + \delta(5w_1 + 9w_2 + 12w_3 + 29w_4)\}, \end{aligned}$$

where p is the probability of a pixel being activated. Finally, we find

$$P(A_i = 1|x_{C_i}) = \left(1 + \frac{1}{v_i^0}\frac{q_0 + q_1\left(\prod_{j=1}^{8}(1+v_i^j) - 1 + \delta T_0\right)}{q_1\left(\prod_{j=1}^{8}(1+v_i^j) + \delta T_1\right)}\right)^{-1}. \tag{24}$$

In the second model for ξ we imagine that we have long straight line boundaries separating activated and non-activated regions. In a square with nine pixels we see 8 different lines (horizontal, slopes $\frac{1}{2}$, 1 and 2, vertical and slopes -2, -1 and $-\frac{1}{2}$). We can give different weights to the different lines, Here, though, we will imagine that the different lines have the same length. Since a line with length l and slope 1 has horizontal length $l/\sqrt{2}$ and similar a line with slope 2 has horizontal length $l/\sqrt{5}$ we find the following values

for the α and β parameters used in the model above.

$$\alpha_1^1 = \frac{2}{\sqrt{5}} + \frac{1}{\sqrt{2}} \quad \alpha_2^1 = \frac{1}{\sqrt{5}} \quad \alpha_3^1 = 1 \quad \alpha_4^1 = \frac{1}{\sqrt{5}},$$
$$\alpha_2^2 = \alpha_2^1 \quad \alpha_3^2 = \frac{1}{\sqrt{2}} \quad \alpha_4^2 = \alpha_4^1,$$
$$\beta_4^1 = \frac{1}{\sqrt{5}} \quad \beta_5^1 = \frac{1}{\sqrt{2}} \quad \beta_6^1 = \frac{1}{\sqrt{5}},$$
$$\beta_4^2 = \beta_4^1 \quad \beta_5^2 = 1 \quad \beta_6^2 = \beta_6^1 \quad \beta_7^2 = \frac{2}{\sqrt{5}} + \frac{1}{\sqrt{2}}.$$

Furthermore, we introduce a parameter γ by letting $\xi(a) = \gamma$ when $s = 9$. Then we find instead of (22) and (23) the expressions

$$\begin{aligned} T_0 &= \sum_{r\in\{1,3,5,7\}} \big\{ v_i^r(\alpha_1^1 + v_i^{r+1}(\alpha_2^1 + v_i^{r+2}(\alpha_3^1 + \alpha_4^1 v_i^{r+3}))) \\ &\quad + v_i^{r+1}v_i^{r+2}(\alpha_2^2 + v_i^{r+3}(\alpha_3^2 + \alpha_4^2 v_i^{r+4}))\big\}, \end{aligned}$$

and

$$\begin{aligned} T_1 &= \sum_{r\in\{1,3,5,7\}} \big\{ v_i^r v_i^{r+1} v_i^{r+2} v_i^{r+3}(\beta_4^1 + v_i^{r+4}(\beta_5^1 + \beta_6^1 v_i^{r+5})) \\ &\quad + v_i^{r+1}v_i^{r+2}v_i^{r+3}v_i^{r+4}(\beta_4^2 + v_i^{r+5}(\beta_5^2 + v_i^{r+6}(\beta_6^2 + \beta_7^2 v_i^{r+7})))\big\} + \gamma \prod_{j=1}^{8} v_i^j. \end{aligned}$$

Also we have the two equations

$$\begin{aligned} 1 &= q_0 + q_1\{511 + \delta[8(1 + \sqrt{2} + \frac{6}{\sqrt{5}}) + \gamma]\}, \\ p &= q_1\{256 + \delta[4(1 + \sqrt{2} + \frac{6}{\sqrt{5}}) + \gamma]\}, \end{aligned}$$

The probability $P(A_i = 1|x_{C_i})$ is still given by (24)

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.R., Rabe-Hesketh, S. *et al.* (1996) Statistical methods of estimation and inference for functional MR image analysis. *Magn. Reson. Med.*, **35**, 261–277.

Cressie, N. (1991) *Statistics for spatial data*. Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics. New York: John Wiley & Sons, Inc.

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.

Greig, D., Porteous, B. and Seheult, A. (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. Tech. Rep. 410, University of Aarhus. *Submitted for publication*.

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., Anderson, J. *et al.* (1999) Plurality and resemblance in fMRI data analysis. *NeuroImage*, **10**, 282–303.

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

Rosenfeld, A. and Kak, A.C. (1982) *Digital Picture Processing*, vol. 2. Academic Press, Orlando.

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. *et al.* (1995) A unified statistical approach for determining significant signals in images of cerebral activation. *Human Brain Mapping*, **4**, 58–73.