Spatial Mixture Modeling of fMRI Data

Niels Væver Hartvig[1]* **and Jens Ledet Jensen**[1,2]

[1] *Department of Mathematical Sciences, University of Aarhus, Aarhus, Denmark*
[2] *MaPhySto,*† *University of Aarhus, Aarhus, Denmark*

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 nonactivated 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 using 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 nonspatial 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. *Hum. Brain Mapping 11:233–248, 2000.* © 2000 Wiley-Liss, Inc.

Key words: functional magnetic resonance imaging; spatial model; mixture model; image analysis

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 that account for correlated noise. The latter encompasses generalized linear models and time series models. A few articles that 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 article such as Lange et al. [1999] for the long list of references that should be cited in this context.

The spatial properties of the data are rarely modeled 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 articles (see, e.g., Worsley et al., 1996). Smoothing the data spatially is in fact equivalent to using a nonparametric model for the spatial activation pattern, assuming only smoothness of the latter [Müller, 1988]. It should hence be viewed as an estimation procedure that is optimal in this model, but there is no general statistical reason for smoothing. On the contrary, smoothing may produce

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

*Correspondence to: Niels Væver Hartvig, Department of Mathematical Sciences, University of Aarhus, Ny Munkegade, DK-8000 Aarhus C, Denmark. E-mail: vaever@imf.au.dk

Received for publication 3 April 2000; accepted 11 July 2000
Published online xx October 2000

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 suprathreshold clusters. This was first suggested by Poline and Mazoyer [1993] and later 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 nonparametric 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 as 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 because 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. Specifically, the activation is modeled 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 require a lot of user interaction. The latter makes them less suitable for routine use. In this article 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 × 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 nonactivated 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 have 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.

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.

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 $A^1, \ldots, A^k$ be a vector of indicators for activation in the k neighbours. We index the As by a superscript to avoid confusion with the usual voxel subscript. Finally $A_C = (A, A^1, \ldots, A^k)$ 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 its 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 noncentral 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,

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

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 $3 \times 3 \times 3$ 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,

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

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

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.

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

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

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,

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

where

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

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 nonactivated 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.

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.

Model 3

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

Estimation of parameters

For estimation purposes, we will now study the whole volume of voxels, v, 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

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

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

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

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

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

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 $(x|A =$

0) $\sim N(0, \sigma^2)$, and $(x|A = 1) \sim N(1, \sigma^2)$. Then the covariance of x_i and x_j is given by

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

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

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

Notice that the two expressions above do 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,

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

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 $(x|A = 0) \sim N(0, \sigma^2)$ and $(x|A = 1) \sim N(1, \sigma^2)$. When more generally $(x|A = 1) \sim N(\mu, \sigma^2)$ where $\mu > 0$, we obtain the setup above by scaling x by μ^{-1}. When the distribution of x is not normal, we need 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

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

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 Examples 2 and 3, they are quite robust to the choice of model.

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 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 number of active voxels classified as active, divided by the total number of active voxels. The FPR is estimated as the number of nonactive voxels that are classified as active, divided by the total number of nonactive voxels.

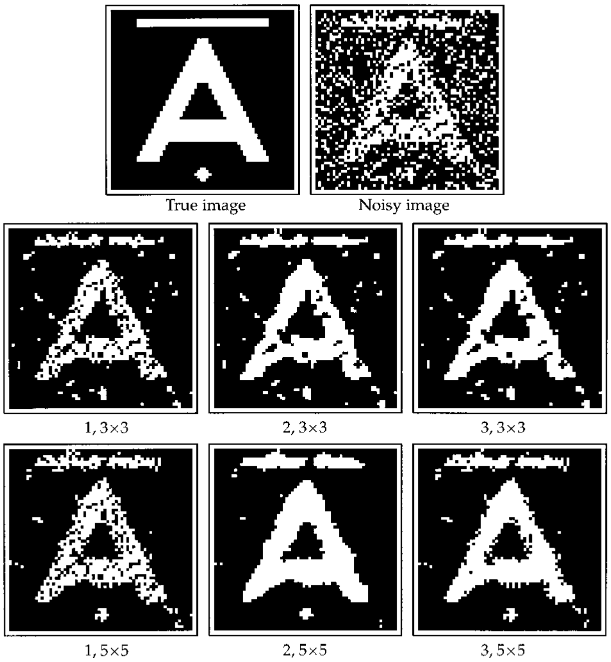

Figure 1.
Comparison of spatial mixture models. Top row: Image I and degraded version. Middle row: Estimates of the true image based on model 1, 2, and 3 applied to a 3 × 3 pixel region. Bottom row: Same as above, but with the models defined on a 5 × 5 region.

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

$$f(x|A = 0) = q^x(1 - q)^{1-x}, \quad x \in \{0, 1\},$$

$$f(x|A = 1) = (1 - q)^x q^{1-x}, \quad x \in \{0, 1\}.$$

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 summarized in Table I and some of the image estimates are displayed in Figure 1.

The second image (Image II) is the binary image displayed in Figure 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

$$f(x|A = 0) = \frac{1}{\sqrt{2\pi\tau}} e^{-x^2/2\tau^2}, \quad x \in \mathbb{R},$$

$$f(x|A = 1) = \frac{1}{\sqrt{2\pi\tau}} e^{-(x-1)^2/2\tau^2}, \quad x \in \mathbb{R},$$

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

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 earlier. As 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.

TABLE I. Estimated classification errors for the three models and the ICM and MAP estimates*

	Classification error	
Model	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)

* 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 Figure 4a in Besag [1986], degraded with Gaussian noise. All figures are in percent; standard errors of estimates are given in parentheses.

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 I. The first column lists the models used in this example. The models 1, 2, and 3 were applied, respectively, defined on a 3 × 3 pixel region and on a 5 × 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.

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 × 5 region models are superior in this setting, which is not surprising as the true images are quite regular with large patches of either black or white. We might suspect that the 3 × 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: First, 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. Second, the MAP and ICM procedures depend on a smoothing parameter that, 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.

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 article for a full description of the data, but will repeat the basic properties here. A region of 24 × 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 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. [1999], 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 article, 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 specifically, the estimate of the activation level is given by

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

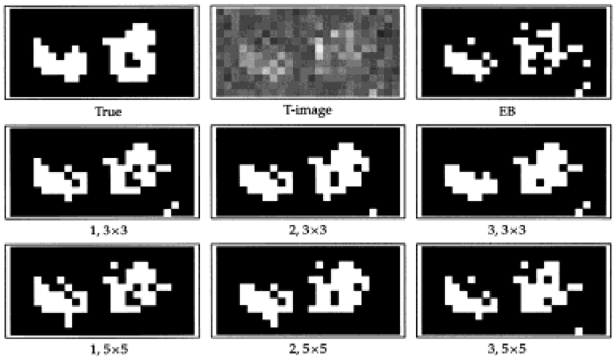

Figure 2.
Activation images for the first of four subsets of the synthetic dataset. Top left and middle: True binary activation image and observed *t*-statistics image. The remaining are thresholded posterior probability images for the different models. EB: Everitt and Bullmore's mixture model. 1, 2, 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.

$$\hat{b}_i = \frac{1}{SSD_x}\sum_{t=1}^{T_0} Z_{it}(x_t - \bar{x}.), \quad SSD_x = \sum_{t=1}^{T_0}(x_t - \bar{x}.)^2,$$

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 the Overview of the Mixture Model section with

$$p(x|A = 0) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}, \quad x \in \mathbb{R},$$

$$p(x|A = 1) = \frac{1}{\sqrt{2\pi}} e^{-(x-\mu)^2/2}, \quad x \in \mathbb{R}.$$

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 subdata sets. 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.

In Table II 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

TABLE II. Comparison of models for the synthetic fMRI data in Figure 2*

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)

* 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.

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.

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 × 3 voxels gives the best result for this data. If the activated areas were larger than these, the 5 × 5 model might be more suitable; however, this activation pattern seems reasonably representative for real data, and hence we recommend the 3 × 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. Models 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 × 3 neighbourhood is preferable in this situation: The statistical power is more than 90% at a significance level of 5%, and the misclassification is reduced by more than 40% compared to the simple mixture model.

We will compare the performance of model 2 with a nonparametric model, where the activation is estimated by smoothing the data spatially with 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 as 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.

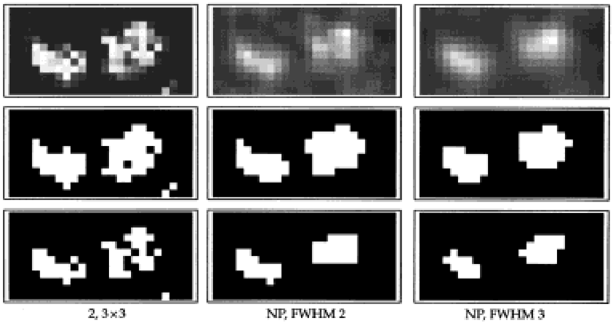

Figure 3.
Activation images for the first of four subsets of the synthetic dataset. From left to right: Model 2 defined on a 3 × 3 region and the nonparametric model with FWHM 2 and 3 voxels, respectively. Top row: Original activation images. Below: Images thresholded at empirical FPR 5% (middle) and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

TABLE III. Estimates of TPR for nonparametric activation images in Figure 3*

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)

* Thresholded at a FPR of 5% and at 1%, respectively. All figures are in percent. Standard errors of estimates are given in parentheses.

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 nonparametric 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 nonparametric model are given in Table III. By comparing this with Table II, 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.

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. Five oblique slices were acquired in axial-coronal direction through the visual cortex. The stimulus was a 7 Hz 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.

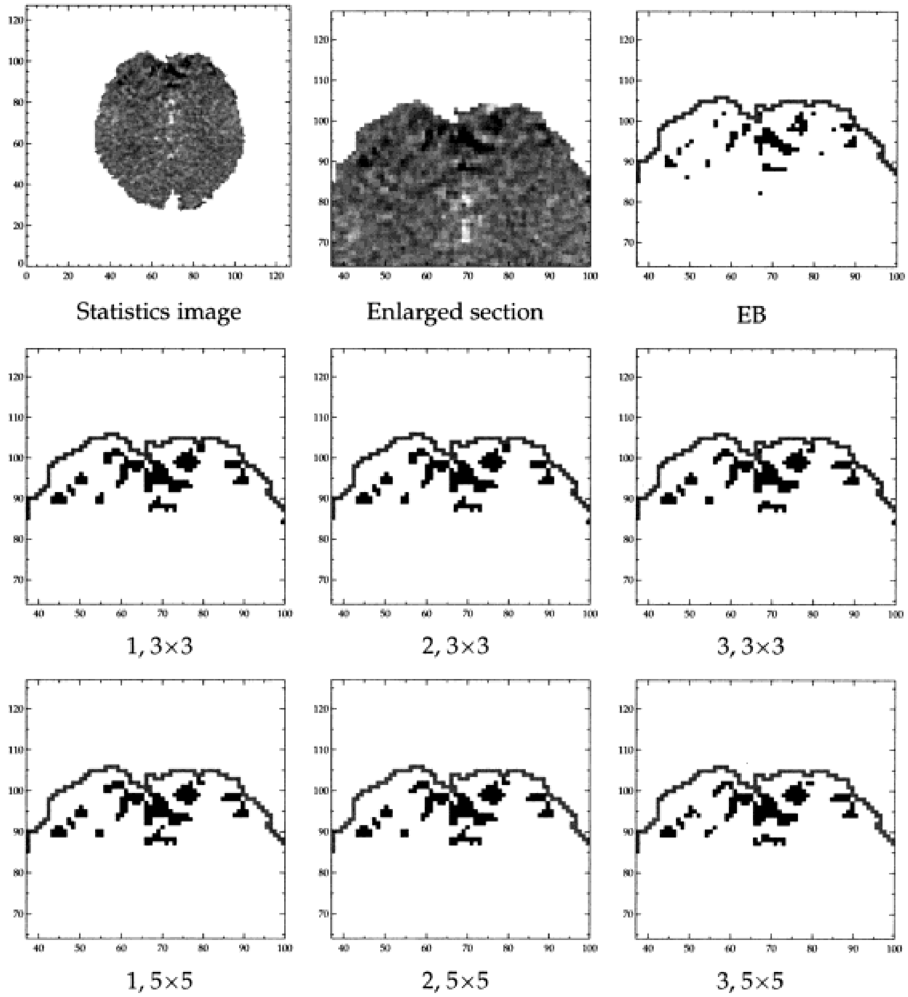

Figure 4.
Comparison of estimated activation patterns for the different mixture models of the visual stimulation data. Top left and middle: Raw image of *t*-statistics and an enlarged section of this. The remaining panels are posterior probability images thresholded at 0.5. Top right: nonspatial mixture 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).

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 4,389 statistics and found that a mixture of three components fitted well to this. Two of these were Gamma distributions, modeling respectively positive and negative BOLD effects, and one was a Normal distribution modeling the noise. The fitted density was

$$f(x) = p_0 f_N(x; 0, \sigma^2) + p_- f_\Gamma(-x; \lambda_-, \beta_-) + p_+ f_\Gamma(x; \lambda_+, \beta_+), \quad (15)$$

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

TABLE IV. Parameter estimates for the distribution (15) of $\{x_i\}$

$\hat{\sigma} = 1.5160$	$\hat{p}_+ = 0.0502$	$\hat{p}_- = 0.0081$
	$\hat{\lambda}_+ = 6.2349$	$\hat{\lambda}_- = 56.923$
	$\hat{\beta}_+ = 0.9433$	$\hat{\beta}_- = 10.253$

$$E(X|X > 0) = \frac{\sum_{i=1}^{V} x_i 1(x_i > 0)}{\sum_{i=1}^{V} 1(x_i > 0)},$$

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 IV.

We are only interested in detecting positive activation in this example. Therefore we write $f(x)$ as

$$f(x) = (1 - p_+)f(x|A = 0) + p_+f(x|A = 1),$$

where

$$f(x|A = 0) = \frac{p_0}{p_0 + p_-} f_N(x; 0, \sigma^2) + \frac{p_-}{p_0 + p_-} f_\Gamma(-x; \lambda_-, \beta_-)$$

is the null distribution and

$$f(x|A = 1) = f_\Gamma(x; \lambda_+, \beta_+)$$

is the distribution of x, given that the voxel is positively activated.

The setup is hence as given previously, 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 because 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 nonspatial mixture model (EB), and for models 1, 2, and 3. 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 more accurate. 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 most accurate. 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.

DISCUSSION

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 not active and when it is. Typically the nonactivated distribution is known, this is the usual null distribution of the SPM. The activation distribution might either be a simple noncentral 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

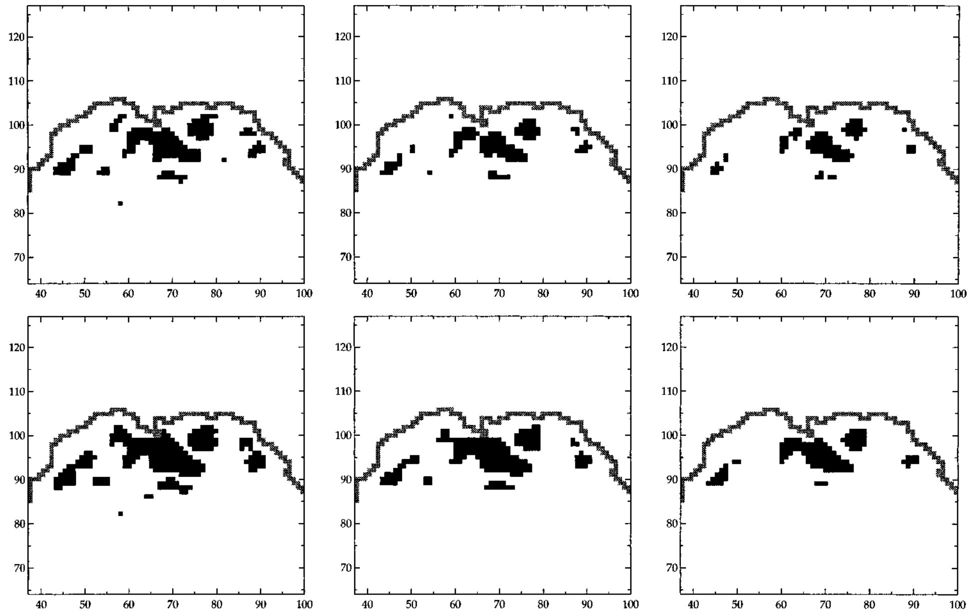

Figure 5.
Images of *t*-statistics based on the visual stimulation data smoothed spatially with a Gaussian kernel of FHWM 2 voxels (top row) and 3 voxels (bottom row). The images are thresholded at 5.0 (left), 6.0 (middle), and 7.0 (right).

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 two 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 be estimated directly from the data.

When modeling only 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.

Comparison with existing methods

The methodology extends that of EB, who proposed a nonspatial 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 misclassification 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 nonparametric 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 that 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 that are less biased and more efficient. Clearly, our model is not "correct," but we would like to emphasize the difference between parametric and nonparametric modeling. Furthermore, the choice of the smoothing parameter, i.e., the FWHM of the kernel, is always a critical point in nonparametric 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 modeling 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 nonactivated, 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, as 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 nonnegligible 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 presurgical 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 article, 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 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 $P(A_C = a_C)$ nonparametrically 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 $(x|A = 0) \sim N(0, \sigma^2)$ and $(x|A = 1) \sim N(1, \sigma^2)$, where our setup is completely general.

We have assumed throughout the article 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 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 $f(x_C|a_C)$. 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 article. 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 3×3 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.

CONCLUSION

We have formulated a simple mixture model for fMRI data that 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 unthresholded probability map, which shows a clear distinction between estimated activation and baseline.

ACKNOWLEDGMENT

We thank Hans Stødkilde-Jørgensen at the MR Research Centre, Aarhus University Hospital, for sharing data and insight. We also thank the anonymous reviewers for valuable suggestions.

REFERENCES

Bandettini PA, Jesmanowicz A, Wong EC, Hyde JS (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 SC, 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.

Bullmore ET, Suckling J, Overmeyer S, Rabe-Hesketh S, Taylor E, Brammer MJ (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 NAC (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, von Cramon DY (1998): fMRI signal restoration using a spatio-temporal Markov random field preserving transitions. Neuroimage 8:340–349.

Everitt BS, Bullmore ET (1999): Mixture model mapping of brain activation in functional magnetic resonance images. Hum Brain Mapp 7:1–14.

Forman SD, Cohen JD, Fitzgerald M, Eddy WF, Mintun MA, Noll DC (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 KJ, Jezzard P, Turner R (1994): The analysis of functional MRI time-series. Hum Brain Mapp 1:153–171.

Gilks WR, Richardson S, Spiegelhalter DJ (eds.) (1996): Markov chain Monte Carlo in practice. London: Chapman & Hall.

Greig DM, Porteous BT, Seheult AH (1989): Exact maximum a posteriori estimation for binary images. J R Statist Soc Ser B 51:271–279.

Hartvig N. Væver (2000): Parametric Modelling of Functional Magnetic Resonance Imaging Data. Ph.D. Thesis. Department of Theoretical Statistics, University of Aarhus. http://www.imf.au.dk.

Heyde CC (1997): Quasi-likelihood and its application. New York: Springer-Verlag.

Lange N, Zeger SL (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 SC, Anderson JR, Nielsen FA, Holmes AP, Kolenda T, Savoy R, Hansen LK (1999): Plurality and resemblance in fMRI data analysis. NeuroImage 10:282–303.

Liu AK, Belliveau JW, Dale AM (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, Zamar RH (1994): Binary-image restoration. Can J Statist 22:335–355.

Müller HG (1988): Nonparametric regression analysis of longitudinal data. Lecture notes in statistics. Berlin: Springer-Verlag.

Poline JB, Mazoyer BM (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 JB, Worsley KJ, Evans AC, Friston KJ (1997): Combining spatial extent and peak intensity to test for activations in functional imaging. NeuroImage 5:83–96.

Press WH, Teukolsky SA, Vetterling WT, Flannery BP (1992): Numerical recipes in C, 2nd ed. Cambridge: Cambridge University Press.

Rosenfeld A, Kak AC (1982): Digital picture processing, vol. 2. Orlando: Academic Press.

Worsley KJ (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 KJ, Friston KJ (1995): Analysis of fMRI time-series revisited—again. Neuroimage 2:173–181.

Worsley KJ, Marrett S, Neelin P, Vandal AC, Friston KJ, Evans AC (1996): A unified statistical approach for determining significant signals in images of cerebral activation. Hum Brain Mapp 4:58–73.

APPENDIX

We derive the formulas for the posterior probability that a voxel is activated in model 2 and model 3 in the following.

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

$$P(A = 1 | x_C) = \left\{ 1 + \frac{1}{v}\left[\gamma^{-1} + \frac{1 - \alpha(1+\gamma)^{k+1}/\gamma}{\alpha} \left(\prod_{j=1}^{k} (1 + \gamma v^j) \right)^{-1} \right] \right\}^{-1}. \quad (16)$$

Model 3

In this model we have

$$1 = q_0 + q_1 + \frac{\alpha_1}{\gamma_1}\{(1+\gamma_1)^{k+1} - 1 - \gamma_1^{k+1}\}$$

$$+\frac{\alpha_2}{\gamma_2^k}\{(1+\gamma_2)^{k+1} - 1 - \gamma_2^{k+1}\},$$

$$p = q_1 + \alpha_1\{(1+\gamma_1)^k - \gamma_1^k\} + \frac{\alpha_2}{\gamma_2^{k-1}}\{(1+\gamma_2)^k - \gamma_2^k\}, \tag{17}$$

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

$$N = \frac{\alpha_1}{\gamma_1}\prod_{j=1}^{k}(1+\gamma_1 v^j)$$

$$+\frac{\alpha_2}{\gamma_2^k}\prod_{j=1}^{k}(1+\gamma_2 v^j) + q_0 - \left(\frac{\alpha_1}{\gamma_1} + \frac{\alpha_2}{\gamma_2^k}\right),$$

$$D = \alpha_1\prod_{j=1}^{k}(1+\gamma_1 v^j)$$

$$+\frac{\alpha_2}{\gamma_2^{k-1}}\prod_{j=1}^{k}(1+\gamma_2 v^j) + \{q_1 - (\alpha_1\gamma_1^k + \alpha_2\gamma_2)\}\prod_{j=1}^{k} v^j.$$

For model 3 the marginal density of x_{C_i}, used in the constrast function (8), is

$$f(x_{C_i}; \phi, \psi) = \prod_{j=0}^{k} f(x_i^j|0; \phi)$$

$$\times\left\{\frac{\alpha_1}{\gamma_1}\prod_{j=0}^{k}(1+\gamma_1 v_i^j(\phi)) + \frac{\alpha_2}{\gamma_2^k}\prod_{j=0}^{k}(1+\gamma_2 v_i^j(\phi))\right.$$

$$\left. + q_0 - \left(\frac{\alpha_1}{\gamma_1} + \frac{\alpha_2}{\gamma_2^k}\right) + \{q_1 - (\alpha_1\gamma_1^k + \alpha_2\gamma_2)\}\prod_{j=0}^{k} v_i^j(\phi)\right\}, \tag{19}$$

with $\psi = (\alpha_1, \alpha_2, \gamma_1, \gamma_2, q_1)$ and q_0 given by the constraint in (17).