Probabilistic Modeling of Single-Trial fMRI Data

Markus Svensén*, Frithjof Kruggel, and D. Yves von Cramon

***Abstract*—This paper describes a probabilistic framework for modeling single-trial functional magnetic resonance (fMR) images based on a parametric model for the hemodynamic response and Markov random field (MRF) image models. The model is fitted to image data by maximizing a lower bound on the log likelihood. The result is an approximate maximum *a posteriori* estimate of the joint distribution over the model parameters and pixel labels. Examples show how this technique can used to segment two-dimensional (2-D) fMR images, or parts thereof, into regions with different characteristics of their hemodynamic response.**

***Index Terms*—Hemodynamic response, image segmentation, Markov random field, mean field theory.**

I. Introduction

FUNCTIONAL magnetic resonance imaging (fMRI) attempts to detect brain activity by localized, noninvasive measurements of the change in blood oxygenation, the so-called BOLD contrast [1]. Measurements are collected under controlled conditions where subjects are performing specific tasks prompted by some stimulus (e.g., deciding whether a read out sentence is grammatically correct, performing arithmetic calculations, looking at changing scenes, etc.).

There are two dominating design paradigms for fMRI experiments. In block trial design, subjects repeat the task for extended periods of time (blocks of, say, 30 s), interleaved with similarly long periods of inactivity. In contrast, in single-trial design (also called event-related design [2]) the task is performed only once, followed by a period of inactivity before the next trial. This allows the full response evoked by the task to be analyzed and it is this kind of analysis we are interested in this paper.

The analysis of block trial fMRI data has been focused on identifying regions in the brain where the collected measurements show correlation with the task performed, so called activated regions. For single-trial data, we are interested not only in which regions are activated, but also what kind of activity occurs. Both these problems depend on the characterization of the hemodynamic response (HR) [3], which refers to the local change in blood oxygenation as an effect of increased neuronal activity. This change is not immediate but is delayed by 2–6 s from stimulus onset and this delay varies among subjects, experimental conditions, etc. Moreover, the stimulus subjects are exposed to during data collection, which is assumed to trigger the task-related activity, is normally treated as being discrete. Often it is modeled as a binary (box-car) function, i.e., the stimulus is either present or not present. In contrast, the task-related blood oxygenation rises and decays smoothly.

Successful detection of activated regions requires that the HR is taken into account when analyzing the correlation between collected measurements and the stimulus. The simplest approach is to shift the boxcar-function representing the stimuli in time to account for the delay in the HR [4]. A more sophisticated approach is to convolve a function representing the stimuli with a model function for the HR. A number of functions have been proposed for this purpose: Friston *et al.* [5] used a global Poisson-function whose parameter was estimated from the data. Lange and Zeger [6] instead proposed to use the Gamma function and Rajapakse *et al.* [7] the Gaussian function. In both of the latter cases the model function parameters were estimated separately for each pixel averaged over data from many trials in the Fourier domain. Friston *et al.* [8] proposed modeling the HR using a linear combination of fixed global basis functions chosen to span a space of probable HR model functions. A similar model was proposed by Bullmore *et al.* [9]. In both these models, the weight parameters of the linear combination is determined, per pixel, from the data by averaging over trials.

Improving detection of activated pixels has, until recently, been the primary driving force behind the research in modeling the HR. Only with the introduction of single-trial fMRI experiments, the characterization of the HR has become a research interest in its own right. This strand of research has so far concentrated on *post hoc* analysis to the HR in regions found to be highly activated, called regions of interest (ROI's). It is done either by fitting a HR model function to or averaging the time course over a selected set of pixels and/or over trials with identical stimuli [3], [10], [11]. However, it is not obvious how to choose these (sets of) pixels from an ROI. Moreover, the detection of ROI's relies on an initial estimate of the HR, which will obviously affect which regions become ROI's.

In this paper a multivariate probabilistic framework for modeling fMRI data is investigated. It aims to construct a generative model of the process that generated the observed image, by combining a parametric model for the HR with the use of image models based on Markov random fields (MRF's). It is intended to be complementary to existing methods for analysis of fMRI data, offering the following.

- Mechanisms for modeling multiple regions with different HR characteristics. For example, the response typically occurs later in venous areas than in cortical areas.
- A way to incorporate prior expectations into the modeling.
- Simultaneous image segmentation and parameter estimation.
- A framework which can be extended to model more complex structure in the data, e.g., clusters of single trials with

Manuscript received February 9, 1999; revised October 28, 1999. The Associate Editor responsible for coordinating the review of this paper and recommending its publication was X. Hu. *Asterisk indicates corresponding author*.

*M. Svensén, F. Kruggel, and D. Y. von Cramon are with the Max-Planck Institute of Cognitive Neuroscience, Stephanstr. 1A, Leipzig, Germany (e-mail: svensen@cns.mpg.de; kruggel@cns.mpg.de; cramon@cns.mpg.de).

Publisher Item Identifier S 0278-0062(00)01223-4.

similar spatio-temporal characteristics, in a series of repeated trials.

Applying this model to fMRI data results in an image segmented into regions, where pixels within a region have similar characteristics in their HR, which are distinct to those of pixels in adjacent regions.

The model presented in this paper is restricted to single-trial data where the time between subsequent trials is sufficiently long to allow the HR to return to baseline. Moreover it assumes some basic preprocessing of the data, as will be described.

II. Models and Methods

We propose to model an fMR image, by which we mean a set of pixels on a regular lattice with associated time-series of measurements, as an MRF. Each pixel is assumed to belong to one out of K classes, with each class corresponding to a parametric model function for the HR. The time series associated with each pixel contains measurements collected at the corresponding location during a single trial at times $t_1, \cdots, t_D$. Each trial consists of one repetition of the task, followed by a period of rest during which the subject is assumed to be inactive. By choosing an MRF model, we implicitly assume that the spatial distribution of the classes will be locally smooth so that neighboring pixels typically belong to the same class. This may seem counter to the rather complex spatial structure of the brain, but the limited spatial resolution of fMRI as well as partial volume effects will have the effect of smoothing the spatial distribution. Thus, images will consist of one or more spatially homogeneous regions, each region associated with a parametric HR model function.

The MRF model and associated inference procedures that we are going to describe are independent of the dimensionality of the images, but in the given context, only two-dimensional (2-D) or three-dimensional (3-D) images are of interest. We will restrict our attention to 2-D images, as the resolution of our current fMR images is much lower in the between-plane direction than the in-plane direction.

A. Modeling the Hemodynamic Response

The theoretical framework is independent of the choice of HR model function but for efficient model fitting we should choose a function which is continuously differentiable with respect to its parameters. We follow Kruggel and von Cramon [10] and model it using a Gaussian function. There are two principal reasons behind this choice. First, it provides a parsimonious model which agrees well with observed data. Second, the parameters of the model can be given a physiological interpretation.

To formalize, we model the HR as a function of time as

$$h(t) = \eta \exp\left(-\frac{(t-\mu)^2}{\sigma}\right) + o \tag{1}$$

where, as illustrated in Fig. 1:

- μ lag, i.e., the time from the onset of the stimuli to the peak of the HR;
- σ dispersion, which reflects the rise and decay time;
- η gain, or amplitude, of the response;

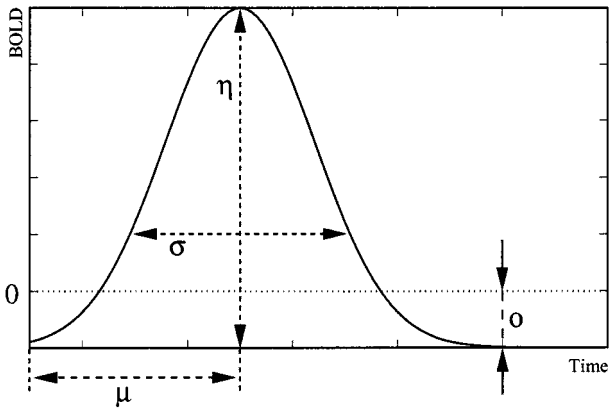

Fig. 1. An illustration of the Gaussian HR model function showing parameters $\mu, \sigma, \eta,$ and o. The dotted horizontal line labeled 0 represents the (arbitrary) baseline level.

- o offset that defines the minimum level for the HR model function, relative to some baseline level.

The function can be fitted to a set of data by minimizing, e.g., the least-squares fit with respect to the parameters, using standard methods for nonlinear optimization (see, e.g., [12]). Note that, in comparison to the model proposed by Rajapakse *et al.* [7], this model does not involve a convolution with a function representing the stimulus. This is equivalent to representing the stimulus as an impulse at the beginning of each trial, which will be a reasonable approximation as long as the stimulus is short relative to the temporal resolution of the experiment, i.e., the rate at which fMR images are collected. We also assume that the time between two trials, the intertrial interval, is sufficiently long for the HR to return to the offset level. Both these restrictions could potentially be relaxed, given that certain assumptions about the stimulus and the HR are made.

For numerical convenience, σ and η can be expressed using auxiliary variables z_σ and z_η so that

$$\sigma = \exp(z_\sigma) \qquad \text{and} \qquad \eta = \exp(z_\eta). \tag{2}$$

This will ensure that σ and η are always positive. For σ, this is certainly a sensible and numerically necessary restriction, while for η it must be regarded as a simplification. It is known that there are regions in the brain that exhibit a deactivation in response to stimulus. However, since the underlying mechanisms are not well understood we prefer to restrict ourselves to positive responses only. This constraint can be relaxed simply by modeling η as direct parameter. We will use $\boldsymbol{\theta}$ to denote the parameter vector $[\mu, z_\sigma, z_\eta, o]$.

It is well known that fMRI data is subject to noise from a number of sources, e.g., physiological processes such as pulse and breathing, baseline drift, etc. [29]. However, we will assume that preprocessing the data will remove systematic noise components, so that we can treat remaining noise as being approximately spherical Gaussian with variance α^{-1} and uniform over the whole region, i.e., the noise level is the same for all subregions. Empirical results suggest that these are reasonable approximations, although choosing K small will typically result in some residual correlations in the noise. This highlights an important consequence of having a spatially varying versus

spatially constant model. In the latter case, spatial variance in the signal will be regarded as noise, with a resulting spatial covariance [10]. Although a spatially constant noise model almost certainly is an approximation, combined with a spatially varying model for the signal and preprocessed data it can be regarded as sufficient.

For a pixel n which belongs to class k the probability distribution over observable trial vectors $\boldsymbol{y}_n$ can then be written as

$$p(\boldsymbol{y}_n|\boldsymbol{\theta}_k, \alpha) = \left(\frac{\alpha}{2\pi}\right)^{D/2} \exp\left(-\frac{\alpha}{2}\|\boldsymbol{y}_n - \boldsymbol{h}_k\|^2\right) \quad (3)$$

where $\boldsymbol{h}_k$ is the HR model function computed from (1) and (2) for $[t_1, t_2, \cdots, t_D]$ using the parameter vector $\boldsymbol{\theta}_k = [\mu_k, z_{\sigma k}, z_{\eta k}, o_k]$. We assume that pixels are independent given their class labels, so the likelihood for an image or, more generally, a region of N pixels can be written

$$\prod_n^N \left(\frac{\alpha}{2\pi}\right)^{D/2} \sum_k^K \iota_{nk} \exp\left(-\frac{\alpha}{2}\|\boldsymbol{y}_n - \boldsymbol{h}_k\|^2\right) \quad (4)$$

where ι_{nk} is an indicator function that equals one if pixel n belongs to class k and zero otherwise. We can then select the parameter values for the K HR functions that maximize (4). However, we do not know ι_{nk}, and in order to deal with this problem in the given context we will need some more theory, but before coming to that, we consider prior expectations about HR functions.

1) Additional HR Priors: Our choice of model function for the HR forms in itself a fairly strong prior about what we expect to find in the data. Nevertheless, the experimental design normally provides additional knowledge, in the light of which certain values for the parameters of the HR functions seem much more likely than others. We can incorporate this into our model by defining a prior distribution over the space of parameter values. The fact that our chosen HR model function has interpretable parameters facilitates this specification.

Here, we choose a simple independent Gaussian distribution

$$p(\boldsymbol{\Theta}) = \left(\prod_i^{4K} 2\pi V_{\boldsymbol{\Theta}}(i)\right)^{-1/2} \cdot \exp\left(-\frac{1}{2}\left(\boldsymbol{\Theta} - \overline{\boldsymbol{\Theta}}\right)^{\mathrm{T}} V_{\boldsymbol{\Theta}}^{-1}\left(\boldsymbol{\Theta} - \overline{\boldsymbol{\Theta}}\right)\right) \quad (5)$$

where $\overline{\boldsymbol{\Theta}}$ is a $4K$-element vector contain the expected values for μ_k, $x_{\sigma k}$, $x_{\eta k}$ and o_k, $k = 1, \cdots, K$, and $\boldsymbol{V}_{\boldsymbol{\Theta}}$ is a diagonal covariance matrix with the corresponding variances along its diagonal. From (2) we see directly that an expected value for the dispersion $\overline{\sigma}$ corresponds to an expected value for $z_\sigma = \ln \overline{\sigma}$. The expected variance of σ can be translated into a expected variance of z_σ using the linear approximation

$$\mathrm{Var}\,(\sigma) \approx \mathrm{Var}\,(\phi z_\sigma) = \phi^2\,\mathrm{Var}\,(z_\sigma),\ \phi = \left.\frac{d\sigma}{dz_\sigma}\right|_{\overline{z_\sigma}} = \overline{\sigma}.$$

Prior expectations about η can be translated into expectations for z_η in the same way.

We will refer to $\overline{\boldsymbol{\Theta}}$ and $\boldsymbol{V}_{\boldsymbol{\Theta}}$ as hyperparameters, since they control the distribution over the parameter $\boldsymbol{\Theta}$. We assume that the prior and likelihood are independent, so we can multiply them together to obtain a penalized likelihood. The prior will appear as an additional term in the update formula for the parameters of the HR functions which penalizes parameter values that depart significantly from the *a priori* specified $\overline{\boldsymbol{\Theta}}$. $\boldsymbol{V}_{\boldsymbol{\Theta}}$ will control the scaling of this penalty so that when the element along its diagonal are small, the corresponding penalty will be large and vice versa. We can think of $\boldsymbol{V}_{\boldsymbol{\Theta}}$ as expressing our confidence in our estimate $\overline{\boldsymbol{\Theta}}$.

B. Modeling Images Using MRF's

An MRF [14] is a set of random variables indexed over the vertices in an ordered lattice. The typical example is a 2-D image where the random variables are the labels (e.g., color) associated with the pixels. The key property of MRF's is that the distribution of the random variable associated with a pixel n given the values associated with the pixels in a (typically small) neighborhood of n is independent of the rest of the pixels in the image. This can be formalized as

$$p(x_n|x_m, n \neq m) = p(x_n|x_m \in \mathcal{N}_n) \quad (6)$$

where

- x_n class label of pixel n;
- $\mathcal{N}_n$ set of random variables representing the labels for the pixels that are in the neighborhood of pixel n.

In this paper, we will restrict ourselves to discrete distributions over 2-D regular lattices.

The distribution over an MRF can be written as a Gibbs distribution

$$P(\boldsymbol{x}) = \frac{1}{Z} \exp\left(\sum_c U_c(\boldsymbol{x})\right) \quad (7)$$

where

- $\boldsymbol{x}$ vector of class labels for all the pixels in the image;
- U_c *potential function* for *clique* c in the lattice of pixels;
- Z normalization constant.

A clique is an ordered set of pixels which are all in the neighborhoods of each other. The sum in (7) runs over all the cliques in the lattice, as defined by our choice of neighborhoods. The potential function gives a potential, or cost, for the particular combination of labels in the cliques, given by the corresponding elements in $\boldsymbol{x}$. The normalization constant, Z, also known as the partition function, ensures that the distribution integrates to one and is computed as

$$Z = \sum_{\boldsymbol{x} \in \mathbb{X}} \exp\left(\sum_c U_c(\boldsymbol{x})\right). \quad (8)$$

Computing Z exactly is tractable only for very small images, since it requires summing over the whole set of possible pixel-class combinations, $\mathbb{X}$. Fortunately, as will be seen, computing Z is not necessary for the purpose of model fitting.

We use the commonly applied multilevel logistic model [15], [16], where we specify neighborhoods such that each pixel only depends on its nearest neighbors (distance equal to one in the

lattice of pixels). Thus, our model contains cliques with two pixels, for which we define the potential function

$$U(x_n, x_m) = \begin{cases} \beta/2, & \text{if } x_n = x_m \\ -\beta/2, & \text{otherwise.} \end{cases} \tag{9}$$

Here we have dropped the index c since U is the same for all cliques that we are considering. β plays the role of a scale parameter for the prior. As β increases, so does the cost for neighboring pixels from different classes, which in effect forces a smoother image.

C. Parameter Estimation and Segmentation

We are seeking to jointly estimate the parameters of the K HR model functions, $\boldsymbol{\Theta} = [\boldsymbol{\theta}_1, \cdots, \boldsymbol{\theta}_K]$, the noise parameter α, and a corresponding segmentation of an fMR image (or a part thereof) into regions. We do this by maximizing the likelihood of the image under the constraints imposed by the MRF prior and the prior distribution over the parameters of the HR functions. The hyperparameters $\overline{\boldsymbol{\Theta}}$ and $\boldsymbol{V}_{\boldsymbol{\Theta}}$, and the MRF scale parameter β are considered to be defined *a priori* and are kept fixed.

To formalize, we want to maximize the penalized likelihood function

$$\begin{aligned} &p(\boldsymbol{\Theta}) \prod_n^N p(\boldsymbol{y}_n, \boldsymbol{x}|\boldsymbol{\Theta}, \alpha) \\ &\quad = p(\boldsymbol{\Theta}) \prod_n^N p(\boldsymbol{y}_n|\boldsymbol{\Theta}, \alpha, x_n) p(x_n|\mathcal{N}_n) \end{aligned} \tag{10}$$

with respect to $\boldsymbol{\Theta}$ and α. Here we have used (6) combined with the fact that $\boldsymbol{y}_n$ depends on x_m, $n \neq m$, only through x_n and that the prior distribution over the pixel class labels is independent of $\boldsymbol{\Theta}$ and α. Once we have found the parameter values that maximize (10), we can use these in conjunction with the data to determine the most likely segmentation of the image, as will be described shortly.

Normally, it is more convenient to take the logarithm of (10), to get the penalized log-likelihood function

$$\ln p(\boldsymbol{\Theta}) + \underbrace{\sum_n^N \ln p(\boldsymbol{y}_n|\boldsymbol{\Theta}, \alpha, x_n)}_{\ell} + \sum_n^N \ln p(x_n|\mathcal{N}_n). \tag{11}$$

The last term in this expression is independent of $\boldsymbol{\Theta}$ and α and maximization of $\ln p(\boldsymbol{\Theta})$ with respect to $\boldsymbol{\Theta}$ is straightforward for the prior chosen in (5). However, in the remaining middle term, denoted ℓ, which represents the conditional log-likelihood of the data, the class labels x_n are unknown. The correct way to deal with such unknown random variables is to integrate (sum) them out of the expression of interest, in our case yielding [from (3)]

$$\begin{aligned} \langle \ell \rangle &= \sum_n^N \langle \ln p(\boldsymbol{y}_n|\boldsymbol{\Theta}, \alpha, x_n) \rangle_{x_n} \\ &= \sum_n^N \sum_k^K P_{nk} \ln p(\boldsymbol{y}_n|\boldsymbol{\theta}_k, \alpha) \end{aligned} \tag{12}$$

where $\langle \cdot \rangle$ denotes expectation and

$$P_{nk} = P(x_n = k|\boldsymbol{y}_n, \boldsymbol{\Theta}, \alpha, \mathcal{N}_n) \tag{13}$$

i.e., the posterior probability that the nth pixel belongs to class k, given the data observed at the pixel and the classes of neighboring pixels. Comparing with (4), P_{nk} plays the role of a soft indicator function, replacing the hard ($\{0, 1\}$) indicator function ι_{nk}. Equation (12) can be recognized as the expected complete log-likelihood of the expectation-maximization algorithm (EM) [17] where P_{nk} are the so-called responsibilities.

In order to compute (12), we first need to compute the marginal posterior distributions $p(x_n|\boldsymbol{y}_n, \boldsymbol{\Theta}, \alpha, \mathcal{N}_n)$ and using Bayes' theorem we get

$$p(x_n|\boldsymbol{y}_n, \boldsymbol{\Theta}, \alpha, \mathcal{N}_n) \propto p(\boldsymbol{y}_n|x_n, \boldsymbol{\Theta}, \alpha) p(x_n|\mathcal{N}_n). \tag{14}$$

However, due to the mutual coupling between the MRF variables, computing $p(x_n|\mathcal{N}_n)$ would require integrating over all the remaining $N - 1$ unknown MRF variables: an intractable computational task. However, there exist approximate methods to tackle this problem.

1) The Mean Field Theory: The mean field theory has a long history in statistical physics and statistical mechanics, and has more recently become popular as a method for approximating intractable probability distributions in the fields of MRF image modeling and statistical machine learning (several references are given in [18]). The basic idea is that influence on a pixel n from a neighboring pixel m can be approximated by a mean influence, computed over the distribution for m. In practice, this approximation has been shown to give good results at moderate computational costs. It can be given a formal justification by appealing to variational approximation methods [18].

The problem we are facing is to compute the marginal distribution $p(x_n|\mathcal{N}_n)$ which, using (6), (7), and (9), can be written as

$$p(x_n|\mathcal{N}_n) = \frac{1}{Z_n} \exp\left(\sum_{x_m \in \mathcal{N}_n} 2U(x_n, x_m) \right) \tag{15}$$

where

the factor 2	accounts the dual $U(x_m, x_n)$ cliques;
Z_n	local partition function.

If we for a moment assume that we know $p(x_m|\boldsymbol{y}_m, \boldsymbol{\Theta}, \alpha, \mathcal{N}_m)$ for all $x_m \in \mathcal{N}_n$, we can integrate those variables out of (15), yielding

$$p(x_n|\mathcal{N}_n) = \frac{1}{Z_n^{\mathrm{MF}}} \exp\left(\sum_{x_m \in \mathcal{N}_n} \sum_k^K 2U(x_n, k) P_{mk} \right) \tag{16}$$

where we have again used the notation introduced in (13). Now, we do not know $p(x_m|\boldsymbol{y}_m, \boldsymbol{\Theta}, \alpha, \mathcal{N}_m)$, but (14) and (16) together suggest an iterative scheme. Starting from an initial guess for $p(x_m|\boldsymbol{y}_m, \boldsymbol{\Theta}, \alpha, \mathcal{N}_m)$, we can compute $p(x_n|\mathcal{N}_n)$ from (16), which then can be used in (14) to give a new estimate for $p(x_n|\boldsymbol{y}_n, \boldsymbol{\Theta}, \alpha, \mathcal{N}_n)$. Alternating between these two steps, each step computing the respective distribution over

all MRF variables, will result in an approximate distribution $\hat{p}(x_n|\boldsymbol{y}_n, \boldsymbol{\Theta}, \alpha, \mathcal{N}_n)$, which can be substituted into (12) to give

$$\langle\hat{\ell}\rangle = \sum_n^N \sum_k^K \hat{P}_{nk} \ln p(\boldsymbol{y}_n|\boldsymbol{\theta}_k, \alpha). \tag{17}$$

This can be seen as a lower bound of the expected complete log likelihood. Alternating between computing $\hat{p}(x_n|\boldsymbol{y}_n, \boldsymbol{\Theta}, \alpha, \mathcal{N}_n)$ and maximizing (17) with respect to $\boldsymbol{\Theta}$ and α, we obtain a variational variant of the EM algorithm, which maximizes this lower bound, rather than the log likelihood itself [19], [20]. Note that $\boldsymbol{\Theta}$ and α are kept fixed when computing $\hat{p}(x_n|\boldsymbol{y}_n, \boldsymbol{\Theta}, \alpha, \mathcal{N}_n)$, which in turn is kept fixed as $\boldsymbol{\Theta}$ and α are updated.

2) Mean Field Annealing: A potential problem with the method discussed so far, which plagues most forms of nonlinear optimization problems, is that of local maxima. When the data are noisy, only slightly different initial values for the parameters can result in very different maxima after optimization and deciding whether a found maxima is a good one may be difficult.

This is the motivation for the use of annealing techniques in nonlinear optimization problems [21]. The aim is to make it less likely that the maximization gets stuck early in poor local maxima. The way to achieve this is to initially smooth the function that we are optimizingand then gradually, as the optimization proceeds, remove the smoothing. It has been used successfully with MRF models for restoration of images [16], in particular, fMR images [22] and, in combination with mean field theory, anatomical magnetic resonance images [23].

In our case, the smoothing is imposed by introducing a temperature parameter $T \geq 1$ for the marginal posterior distributions in (14), via (3) and (16), yielding

$$\hat{P}(x_n = k|\boldsymbol{y}_n, \boldsymbol{\Theta}, \alpha, \mathcal{N}_n) \propto \exp\left(\frac{1}{T}\left(-\frac{\alpha}{2}\|\boldsymbol{y}_n - \boldsymbol{h}_k\|^2 + \sum_{x_m \in \mathcal{N}_n} \sum_{k'}^K 2U(k,k')\hat{P}_{mk'}\right)\right). \tag{18}$$

T is initially set to a large value and is then gradually decreased during optimization, until its value reaches one. Note that we keep α fixed at an initial value as long as $T > 1$, so the annealing should be followed by further optimization, where $T = 1$ and both $\boldsymbol{\Theta}$ and α are being optimized to converge to their final values.

3) Putting the Pieces Together: We now summarize the complete sequence of steps required for parameter estimation and segmentation, as shown in Algorithm 1. This algorithm maximizes a lower bound of the penalized log likelihood (11) and will converge to a set of parameters, $\boldsymbol{\Theta}^*$ and α^*, which can be regarded as an approximate maximum *a posteriori* estimate of the distribution over HR parameters. Given these parameters, we can use the mean field theory to compute an approximate, conditional distribution over class labels for all the pixels in the image.

The maximization with respect to $\boldsymbol{\Theta}$ and α proceeds in two substeps. First, we maximize with respect to $\boldsymbol{\Theta}$, by solving

$$\frac{\partial(\langle\hat{\ell}\rangle + \ln p(\boldsymbol{\theta}_k))}{\partial\boldsymbol{\theta}_k} = 0 \tag{19}$$

for $k = 1, \cdots, K$ using numerical optimization. Second, by differentiating (17) with respect to α, using (3) we obtain an update formula in closed form as

$$\alpha = \frac{ND}{\sum_n^N \sum_k^K \hat{P}_{nk}\|\boldsymbol{y}_n - \boldsymbol{h}_k\|^2} \tag{20}$$

where $\boldsymbol{h}_k$ are computed using the updated parameters $\boldsymbol{\theta}_k$.

The mean field computation and the overall fitting procedure, are iterated until the penalized log likelihood (11) no longer changes significantly. From empirical experience, ten iterations seem to be sufficient for the mean field computation. Failure to reach absolute convergence simply means that the bound on the likelihood will be less tight. For the full fitting procedure, we have so far found 40 iterations (or less) to be sufficient.

Once the fitting procedure has converged, we can use the mean-field theory to compute the approximate distribution $\hat{p}(x_n|\boldsymbol{y}_n, \boldsymbol{\Theta}^*, \alpha^*, \mathcal{N}_n)$. This distribution can be interpreted as a soft segmentation of the image, which also expresses the uncertainty in the classification of the pixels. If desired, an absolute segmentation can be obtained simply by choosing the most likely class for each pixel.

Algorithm 1 Estimation of HR parameters and image segmentation

Given: A data set $\{\boldsymbol{y}_n\}$, $n = 1, \cdots, N$; values for β, $\overline{\boldsymbol{\Theta}}$ and $\boldsymbol{V}_{\boldsymbol{\Theta}}$; initial values for $\boldsymbol{\Theta}$, α and T.
Compute $p(\boldsymbol{y}_n|x_n, \boldsymbol{\Theta}, \alpha)$ for all $n = 1, \cdots, N$ and $x_n = 1, \cdots, K$, using (3).
Initialize $\hat{p}(x_n|\boldsymbol{y}_n, \boldsymbol{\Theta}, \alpha, \mathcal{N}_n)$ using $p(\boldsymbol{y}_n|x_n, \boldsymbol{\Theta}, \alpha)$, i.e., ignoring the influences from neighboring pixels.
repeat
  repeat
    for $n = 1, \cdots, N$ **do**
      Update $\hat{p}(x_n|\boldsymbol{y}_n, \boldsymbol{\Theta}, \alpha, \mathcal{N}_n)$ using (18).
    end for
  until convergence
  Update $\boldsymbol{\Theta}$ using (19).
  if $T = 1$ **then**
    Update α using (20).
  else
    Decrease T suitably.
  end if
  Recompute $p(\boldsymbol{y}_n|x_n, \boldsymbol{\Theta}, \alpha)$ for all $n = 1, \cdots, N$ and $x_n = 1, \cdots, K$, using (3).
until convergence
repeat
  for $n = 1, \cdots, N$ **do**

```
    Compute p̂(x_n|y_n, Θ*, α*, N_n) using (18).
  end for
until convergence
As an optional step, segment the image:
```

$$\forall n\colon \tilde{x}_n = \operatorname*{argmax}_{x_n} \hat{p}(x_n|\boldsymbol{y}_n, \boldsymbol{\Theta}^*, \alpha^*, \mathcal{N}_n).$$

III. Examples

In this section we look at three examples illustrating the proposed method. The first example uses realistic synthetic data consisting of artificial HR signals modulated on noise signals from a real fMRI experiment. The second example uses real data from an fMRI experiment with a sentence comprehension task, and demonstrates how the proposed model can be used to segment whole images, relating this to results from conventional fMRI analysis techniques. Using the same data set, we then focus on a region in the images and demonstrate how it can be divided into subregions with different HR characteristics. Finally, we apply the model to data from another fMRI experiment with a task switching paradigm, comparing sets of trials with different stimulus conditions. Before coming to the actual experiments, we first describe the preprocessing of the data, which was carried out in the same manner for all experiments.

A. Preprocessing

The signal recorded during an fMRI experiment does not only consist of the hemodynamic response but is rather a complex mixture of several sources with different spatial and temporal behavior. An example of such a source, which is temporally constant but spatially varying, is the brain being imaged whose physical structure in terms of white and grey matter and cerebral spinal fluid causes spatial variations in the recorded signal. Other sources that have already been mentioned are physiological sources, such as pulse, breathing, and movements by the subject during signal acquisition, and system sources, such as baseline drift and ghost images. These typically vary both spatially and temporally.

Taking the approach of generative modeling, there are essentially two ways we can deal with such uninteresting sources. Either, we can try to explicitly account for them in our model or try to filter them out of the data before trying to model the sources we are actually interested in.

We have chosen a bandpass filtering approach, since this allows us to use our knowledge about the HR and the experiment at hand to filter out whatever signal that cannot reasonably originate from the stimulus induced activation. At the same time, we note that since our knowledge about the HR is limited and we do not want to risk to filter out any signal of interest, the width of an accordingly chosen bandpass filter will almost certainly let some uninteresting signals slip through. Averaging over trials will retain the trial-periodic signals, while further suppressing random signals and signals with different periodicity.

In all of the following experiments data were preprocessed as follows: the signals recorded at each pixels were filtered through a low-pass filter with a cut-off frequency of $2/3$ times the frequency of the stimulation. The resulting signal was subtracted from the original signal, resulting in a high-pass filtering. Note that this removes the anatomical image from the signal, which is important since our model is trying to capture spatial variations in the HR, which otherwise would be confounded with spatial variations in the anatomical image. As a second step the high-pass filtered signal was again passed through a low-pass filter, with a cut-off frequency of three times the frequency of the stimulation. This will suppress signals from physiological sources, such as breathing and pulse.

B. Synthetic Data

The synthetic data set was produced using time courses from a real fMRI experiment recorded at a patch of 10×10 pixels where it is assumed that no stimulus-induced activation occurs. The time courses consisted of 912 samples for each pixel, which were treated as 76 trials of 12 samples each. This corresponded to the trial length of the experiment from which they were taken. These noise time courses were then added to synthetically generated HR responses, in the shape of Gaussian functions with a periodicity of 12 time steps. The 10×10 patch was divided in three regions, where each region had a distinct, associated HR response, as illustrated in Fig. 2. Three different data sets were produced in this fashion, corresponding to three different signal to noise ratios (SNR's), by varying the amplitude of the added noise. The amplitudes were chosen so that the SNR in the resulting trial-averaged data was 2.0, 4.0, and 8.0, corresponding to SNR's in the raw signals of 0.20, 0.41, and 0.81, respectively.

The model trained on this data had three classes, which shared a common prior for the parameters, given in Table I. They were initialized with random samples from a distribution with the same means as this prior, but with the standard deviation parameters divided by ten. The resulting initial model was then fitted to the image data using Algorithm 1. The results for different SNR are shown in Fig. 3. We note the following.

- The accuracy improves with the SNR, both in terms of the segmentation and the estimated HR models.
- For lower SNR, the confusion in the segmentation is greater between the two classes which have a similar HR (grey and black).
- The spatial distribution of the pixel classes is fairly smooth in all three examples.

This is all in line with what we would expect. To investigate whether the MRF prior has any effect on these results, we also reordered the pixels within the image randomly and repeated the experiment. In all three cases the performance was poorer. In fact, for the data sets with SNR's of 2.0 and 4.0 it was no longer possible to confidently determine the correspondence between the estimated and true HR classes.

C. Sentence Comprehension Data

In this example we use data from an fMRI experiment designed for investigating the neuronal correlates of sentence comprehension in the brain [25]. Subjects had to decide whether an aurally presented sentence contained a syntactical violation or not. Each trial started with a sentence being read out, which lasted 2.3–4.5 s. fMR images were collected every 2 s, with 12 images being recorded in each trial. We note that this repetition

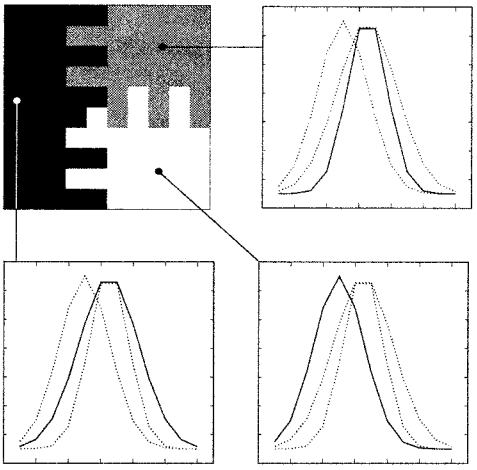

Fig. 2. Illustration of the generation of synthetic data. The 10×10 image patch (top left) was divided into three regions, black, grey, and white, each with an associated HR response. These are shown in the surrounding plots where the HR response of the associated region is plotted with a solid line and the responses of the other regions are plotted with dotted lines. Note that the grey and black regions are closer to each other than to the white region in terms of the HR response time course.

time may result in aliasing of physiological noise [29]. A subset of 72 trials were selected and were preprocessed as described in Section III-A.

As a first experiment, data from the selected trials were averaged and the resulting data were used to fit a model with 2 HR functions ($K = 2$) and a constant background function intended to explain regions where no task-related activity occur. This constant function has a single parameter, namely its value, whose maximum likelihood update is given by averaging over the time course obtained from a weighted average over pixels, where the weights are given by $\hat{P}(x_n = \text{background}|\boldsymbol{y}_n, \boldsymbol{\Theta}, \alpha, \mathcal{N}_n)$. The HR functions shared a common prior given in Table I and β was set to 1.0. The fitting procedure started with 20 iterations during which T was decreased from 10 to 1 and α was held fixed at 1.0, followed by another 20 iterations where $T = 1$ and α was allowed to adapt.

The left image in Fig. 4 shows a Z-map (see, e.g., [5]) for the chosen data set, overlaid on the functional mask. It is based on correlation with a time shifted boxcar-function thresholded at a Z score of 2.75 with a significance value of 0.05. Note that only pixels with positive activation are shown as we do not consider deactivated regions. The right image in Fig. 4 shows the corresponding soft segmentation of the functional mask obtained from our model where pixels have been color coded according to their class probabilities. Red and green correspond to the HR model functions plotted in Fig. 5 while blue corresponds to the background function. Note that although the model is generally fairly confident, there are pixels, especially around borders between regions, which are not clearly classified. They are, e.g., purple rather than blue or red. The number of pixels not classified with certainty will increase somewhat if we remove the MRF prior and base the classification only on the likelihood scores, but the overall picture will remain the same.

As can be seen, the two HR functions take on different roles, one explaining regions with a relatively strong and slightly earlier response and corresponding roughly to pixels with strong activation (high Z scores); the other explaining a weaker and slightly later response and including pixels with lower activation.

As a second example using this data set, we looked at data from a selected spatial region, formed by the red pixels inside the rectangular white border in the right image in Fig. 4. In this region, which corresponds roughly to the left superior temporal gyrus (STG) and Heschl's gyrus (HG), a strong HR was found in the averaged data. Now we apply the model to data from individual trials separately. We want to investigate the HR characteristics over this region and how they vary over trials. The model used was identical to that described in the previous example, with the exception that it was fitted to data from a predefined region, a single trial at the time.

Fig. 6 shows a segmentation and corresponding time courses for one trial. This kind of segmentation appeared in 20 of the 72 trials. This result suggests that: 1) regions of activated pixels may not be homogeneous, but rather have subregions with different characteristics in their HR and 2) during a set of M trials, it may be that P distinct patterns of spatio-temporal activity are observed, where $1 < P \ll M$. Averaging over trials may hide this!

What gives rise to these different patterns of activity in this experiment is not known. A plausible neuroanatomical explanation is given by assigning the red region to the primary auditory cortex (HG), while the green region, which is located anteriorly and more laterally on the STG, corresponds to the secondary auditory cortex. Although, at the neuronal level, the time differences between these areas are presumably small (in the order of 100 ms), the corresponding differences at the metabolic time scale might be greater. A different explanation is of a vascular nature. The early response might arise from a mere cortically weighted area, while the late response might correspond to a venous area.

D. Task Switching Data

This experiment was designed to investigate the effects of task switching, as detected by fMRI [26]. In each trial, subjects received a visual stimulus in the form of a $+$ or a $-$, which was shown in either red or green, and had to respond by pressing one of two buttons: right or left. The green stimulus represented the repetition task, such that the subject should press the left button for $+$ and the right button for $-$. In the switch task, where the stimulus was shown in red, the responses were swapped so that the subject should press the right button for $+$ and the left button for $-$. The whole experiment consisted of 125 repetition trials and 20 randomly interspersed switch trials.

Each trial lasted 16 s, with an fMR images collected every two seconds. We selected data from a single slice, consisting of the 20 switch trials and 20 randomly chosen repetition trials, excluding trials that followed immediately after a switch trial. The data was preprocessed as described above and averaged separately over switch and repetition trials. Two models were fitted separately to the two data sets. The results are shown in Figs. 7 and 8. In Fig. 7, pixels in the functional mask have been assigned to the HR class with highest posterior probability, as color coded in Fig. 8, and overlayed on the anatomical image. Those pixels that were assigned to the background have been

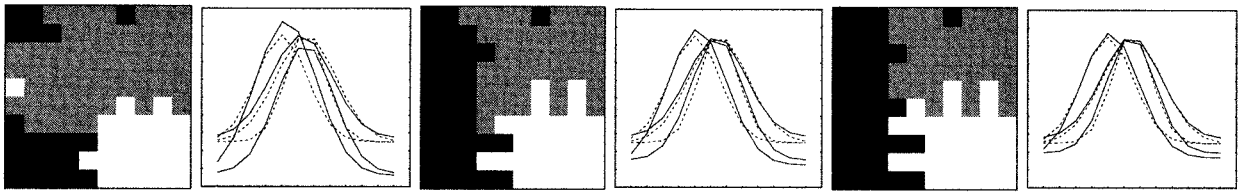

Fig. 3. Results for the synthetic data sets for different SNR's. Left to right, the images shows the MAP classification of the pixels and the plots show the corresponding estimated HR functions (solid) with their targets (dotted) for SNR's 2.0, 4.0, and 8.0. The scale of the plots is identical on all three rows. The correspondence between the estimated and true HR classes was determined by inspections of the plots.

TABLE I
HYPERPARAMETERS FOR THE PRIOR DISTRIBUTIONS OVER HR FUNCTION PARAMETERS IMPOSED IN THE EXAMPLES. μ AND z_σ ARE MEASURED IN (LOG) TIME STEPS, WHILE z_η AND o ARE MEASURED (LOG) RELATIVE TO A NORMALIZED BOLD RESPONSE

Sec.	μ	V_μ	z_σ	V_{z_σ}	z_η	V_{z_η}	o	V_o
III-B	6	3	$\ln 4$	$2/4^2$	$\ln 1$	$1/1^2$	0	1
III-C	6	3	$\ln 4$	$2/4^2$	$\ln 4$	$3/4^2$	0	1
III-D	3	2	$\ln 3$	$2/3^2$	$\ln 4$	$3/4^2$	0	1

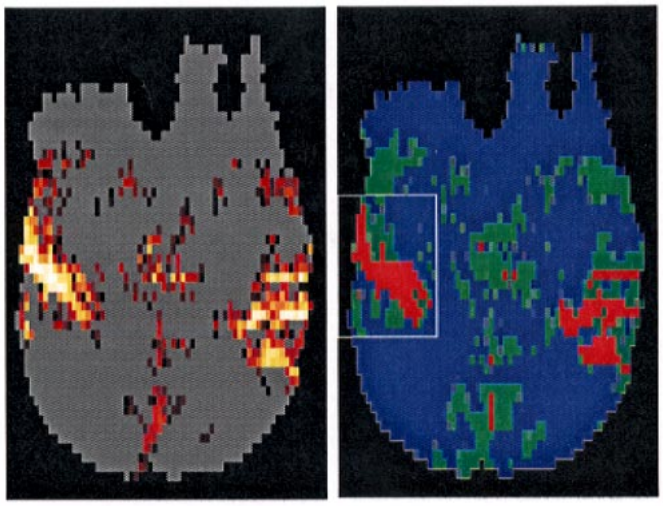

Fig. 4. The left image shows a correlation based Z map overlaid on the functional mask (grey), for the data described in Section III-C. The right image shows a corresponding soft segmentation obtained using our proposed method. In the Z map, pixels have been colored according to their Z score, with dark red being low and white being high. In the right image, pixels have been color coded according to their class probabilities, with the corresponding HR model functions plotted in Fig. 5 the blue class corresponds to the background function. The white rectangular boundary in the right image corresponds to the area shown top right in Fig. 6.

left "transparent," showing the underlying anatomical greyscale image. Note that the response to the switch task is stronger and comes slightly later, relative to the response to the repetition task (solid and dashed lines, respectively, in Fig. 8) which is to be expected. There are fewer pixels assigned to the background in the switch task response but, at the same time, we note that the gain of the dominating class is fairly low for both the switch and repetition trials.

IV. DISCUSSION

This paper describes a multivariate probabilistic method for modeling fMRI data and gives examples of how it can be used.

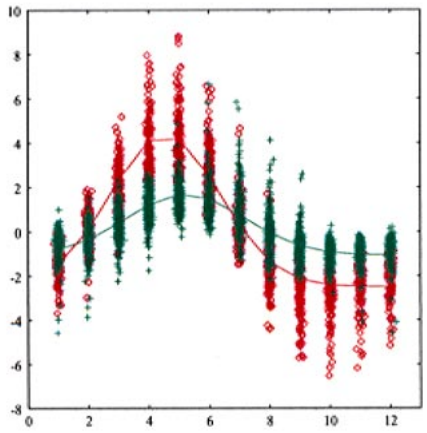

Fig. 5. The HR functions corresponding to the segmentation shown in the right image of Fig. 4. The HR functions are plotted together with the data points they have been assigned by absolute segmentation, i.e., for each pixel, the most likely class has been chosen. In the plot, the x position of the data points corresponding to time steps has been slightly perturbed by adding a small amount of Gaussian noise in order to give a better view of the data.

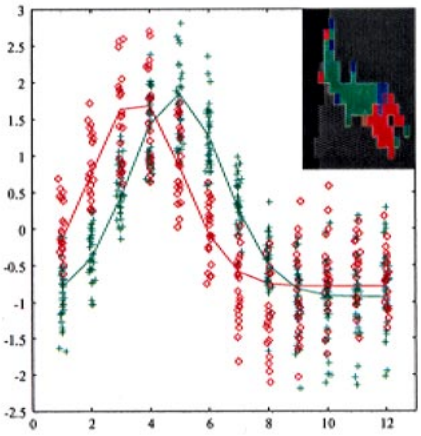

Fig. 6. The top-right inset corresponds to the area inside the rectangular white border drawn at the left side of the right image in Fig. 4. The segmentation shown in this inset is color coded as in Fig. 4, with the corresponding HR functions shown in the plot, using the same plotting scheme as in Fig. 5.

It segments (regions of) fMR images into regions with distinct characteristics in their HR and provides, for each region, a parametric description of its HR characteristics which can be readily understood by physicians and psychologists.

The estimated HR model functions could, if desired, be used in conjunction with, e.g., a linear regression model for detection of activated regions.

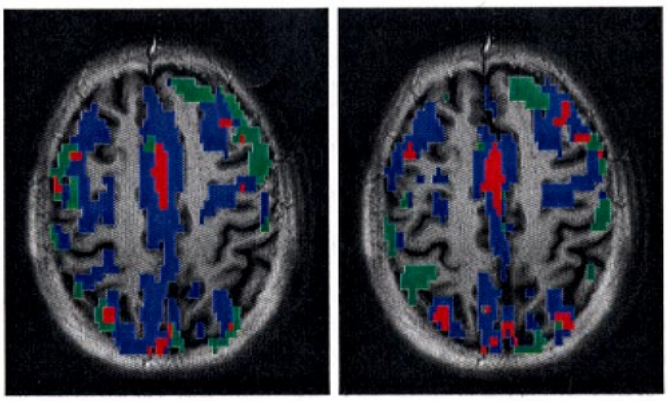

Fig. 7. Segmentations of the functional mask, for the task switching data described in Section III-D, overlayed on the anatomical image. The image to the left shows the segmentation obtained for the switch condition whereas the right image corresponds to the repetition condition. The color coding of the pixels correspond to the color coding of the HR functions shown in Fig. 8.

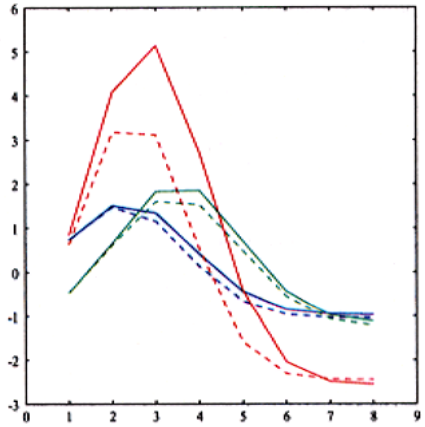

Fig. 8. The HR functions corresponding to the segmentations in Fig. 7. The solid lines correspond to the switch condition (left in Fig. 7), while the dashed lines correspond to the repetition condition (right in Fig. 7).

Since, the overall method is independent of the choice of HR model functions, alternative functions could be considered. We could, for example, consider modeling deactivation. Extending the spatial model to more than two dimensions is simply a matter of implementation.

A. Generalizing the Stimulus Model

In this paper we have restricted ourselves to stimulus modeled as a spike at the beginning of the trial, thereby circumventing the need for a convolution model. However, we could consider introducing a convolution model, making use of the convenient form of the Fourier transform of the Gaussian function [27], which yields an approximate closed form update for the HR parameters. This would also allow the model to be used in a block-trial setting. However, this approach also requires averaging over trials and, thus, the stimulus model must be identical for all trials (or blocks).

Another current restriction on the stimulus model is the intertrial time, which is required to be sufficiently long to allow the HR to return to baseline. Also, this restriction could be relaxed using a convolution model, as long as the HR is assumed to remain constant during all the trials and add linearly [3].

B. Generalizing the Noise Model

At the moment, we assume that preprocessing the data suffices to remove physiological noise in the data, e.g., arising from the pulse and breathing of the subjects. However, we could consider including such sources as part of the model. Petersen *et al.* [28] propose modeling the effects from pulse and breathing using parametric functions. Biswal *et al.* [29] recorded heart and respiratory rates during fMRI scanning and used the resulting signal to construct band-block filters to filter it out of the fMRI data. However, such recordings could also be used to aid modeling of such signals in fMRI data, taking account of them that way.

Alternatively, we could incorporate a more elaborate noise model, treating uninteresting signals in the data as temporally correlated noise. An example of this approach is given by Bullmore *et al.* [9], who suggest using a first order autoregressive [AR(1)] noise model to account for temporal structure in the data that cannot be explained by the stimulus.

In the spatial domain, we currently assume that pixels are independent given their class. In other words, we assume that segmenting an fMR image into regions with distinct HR characteristics accounts for all the spatial correlations in the image. However, Kruggel and von Cramon [10] reported spatial autocorrelations in the residuals over regions of pixels (chosen by hand) whose time courses were modeled using a common HR model function. They addressed this by incorporating an AR(1) noise model in the spatial domain and this approach could also be considered using our method.

A complementary issue is whether the noise model should be spatially dependent or stationary. For example, fluctuations due to pulse to have a larger amplitude in the vicinity of larger vessels [13]. We also could consider allowing the noise level to vary over the spatial domain, but care must taken to control the degrees of freedom in the model. In the simplest case, we could let each HR function and eventual background function(s) have their own noise levels. A more general model could use a discrete or continuous MRF for the noise level. Conceptually, this would be fairly straightforward but would require a more complex implementation, which would also be more demanding in terms of computation.

C. Choosing K and β

An open issue in our current model is the setting of user defined parameters. In the examples described in Section III, the number of HR (and background) functions, the prior for the HR function parameters, and the scale parameter for the MRF, β, were all chosen *a priori* and kept fixed. The hyperparameters specifying the prior over the parameters of the HR functions should be selected based on knowledge about the experiment, neurophysiological knowledge, etc.

The number of HR functions K can strongly influence the final fitted model. If we again consider the synthetic data discussed in Section III-B and fit models with varying K this influence is clearly demonstrated. Fig. 9 shows the segmentations obtained with $K = 2$ and $K = 6$ for SNR's 2.0 and 8.0. They

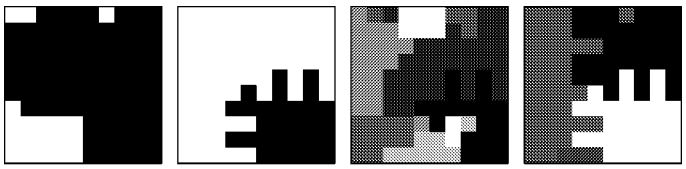

Fig. 9. Segmentations obtained for the synthetic data described in Section III-B, varying the SNR and the number of HR functions.

also demonstrate the influence of the SNR. When the number of HR functions in the model is smaller than the number that are actually generated the data, the model cannot possibly be fully accurate. As long as the SNR is sufficient, the model will provide as close a fit as possible (second left in Fig. 9). However, at low SNR's and with limited degrees of freedom, the model may miss the underlying structure in its attempts to explain as much as possible of the variance in the data (left). When, on the other hand, there are spare degrees of freedom in the model, i.e., when K is higher than the actual number of HR functions that generated the data, these may be used to fit artifactual structure in the data that arise only from noise. Also, in this case, the impact depends on the SNR, as illustrated by the bottom row images in Fig. 9.

The problem of choosing K is a essentially a problem of choosing the appropriate model complexity. A too simple model will not be able to adequately model the true signals, while an overly complex model instead runs the risk of fitting to noise artifacts rather than the true signals. Several approaches have been proposed for inferring K from the data [30]. However, most rely on the evaluation of the full penalized log likelihood and, in our case, the MRF model used in the spatial domain unfortunately causes problems. As mentioned in Section II-B, we generally cannot compute the partition function of the MRF prior, Z [see (8)]. This causes no problems for fitting the parameters of the HR functions in Section II-C, since Z is independent of these parameters. However, it will cause a problem if we want to compare models with different values for K, as Z does depend on K. We could consider a mean field approximation also of Z but this will mean we end up comparing bounds, not actual objective function scores. Alternatively, we can choose to ignore the MRF prior for the estimation of K and form an objective function based on the (unpenalized) log likelihood and a penalty term that increases with the degrees of freedom in the model [30], although such a method may underestimate K, since it ignores the constraints imposed by the MRF prior.

The same problem will arise if we try to derive a scheme to set β, which controls the smoothness of the spatial distribution. Zhang [31] proposed an approximate update formula but reported that the segmentation results were only marginally improved compared to those obtained by setting β empirically by hand, while the computational cost was considerable.

We conclude that, for the time being, K and β are best chosen by experimenting, where the experimenter can also take his or her knowledge about physiology, the fMRI experiment, etc., into account. The difficulties in fully incorporating such expert knowledge into the model means that automatic selection schemes inevitably will be inferior to human guided ones.

D. Future Research

The first example, Section III-C, used data which was averaged over all trials from an experiment, whereas the second example used data from individual trials, without considering the data from any preceding or subsequent trials. Given the results from this second example, a natural next step is to extend the method to model data from multiple trials, going an intermediate way between averaging and individual modeling of single trials. The first step would be a mixture model, containing several instances of the model discussed in this paper, but still many fewer than there are trials. In such a mixture model each instance would explain a set of trials with similar spatio-temporal characteristics, which may or may not correspond to different patterns in the stimulus.

ACKNOWLEDGMENT

The authors would like to thank Dr. C. Williams, University of Edinburgh, U.K., and Drs. M. Pélégrini and H. Benali, INSERM, U494, Paris, France, for their useful comments and suggestions on earlier versions of this paper. They also would like to thank the anonymous reviewers.

REFERENCES

[1] S. Ogawa, T. M. Lee, A. R. Kay, and D. W. Tank, "Brain magnetic resonance imaging with contrast dependent blood oxygenation," in *Proc. Nat. Acad. Sci.*, vol. 87, USA, 1990, pp. 9868–9872.
[2] O. Joseph, K. J. Friston, and R. Turner, "Event-related fMRI," *Human Brain Mapping*, vol. 5, pp. 243–248, 1997.
[3] R. L. Buckner, "Event-related fMRI and the hemodynamic response," *Human Brain Mapping*, vol. 6, pp. 373–377, 1998.
[4] P. A. Bandettini, A. Jesmanowicz, E. C. Wong, and J. S. Hyde, "Processing strategies for time-course data sets in functional MRI of the human brain," *Magnetic Resonan. Med.*, vol. 30, pp. 161–173, 1993.
[5] K. J. Friston, P. Jezzard, and R. Turner, "Analysis of functional MRI time-series," *Human Brain Mapping*, vol. 1, pp. 153–171, 1994.
[6] N. Lange and S. L. Zeger, "Non-linear time-series analysis for human brain mapping by magnetic resonance imaging," *J. R. Stat. Soc. C*, vol. 46, pp. 1–29, 1997.
[7] J. C. Rajapakse, F. Kruggel, J. M. Maisog, and D. Y. von Cramon, "Modeling hemodynamic response for analysis of functional MRI time-series," *Human Brain Mapping*, vol. 6, pp. 283–300, 1998.
[8] K. J. Friston, P. Fletcher, O. Joseph, A. Holmes, M. D. Rugg, and R. Turner, "Event-related responses in fMRI: Characterising differential responses," *NeuroImage*, vol. 7, pp. 30–40, 1998.
[9] E. Bullmore, M. Brammer, S. C. R. Williams, S. Rabe-Hesketh, N. Janot, A. David, J. Mellers, R. Howard, and P. Sham, "Statistical methods of estimation and inference for functional MR image analysis," *Magnetic Resonan. Med.*, vol. 35, pp. 261–277, 1996.
[10] F. Kruggel and D. Y. von Cramon, "Modeling the hemodynamic response in single trial fMRI experiments," *Magnetic Resonan. Med.*, vol. 42, pp. 787–797, 1999.
[11] D. L. Schacter, R. L. Buckner, W. Koutstaal, A. M Dale, and B. R. Rosen, "Late onset of anterior prefrontal activity during true and false recognition: An event-related fmri study," *NeuroImage*, vol. 6, pp. 259–269, 1997.
[12] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, *Numerical Recipes in C: The Art of Scientific Computing*, 2nd ed. New York, NY: Cambridge Univ. Press, 1992.
[13] L. R. Frank, R. B. Buxton, and E. C. Wong, "Probabilistic analysis of functional magnetic imaging data," *Magnetic Resonan. Med.*, vol. 39, pp. 132–148, 1998.

[14] S. Z. Li, *Markov Random Field Modeling in Computer Vision*. Tokyo, Japan: Springer-Verlag, 1995.

[15] H. Elliot, H. Derin, R. Christi, and D. Geman, "Application of the Gibbs distribution to image segmentation," in *Proc. Int. Conf. Acoustics, Speech, Signal Processing*. San Diego, 1984, pp. 32.5.1–32.5.4.

[16] S. Geman and D. Geman, "Stochastic relaxation, Gibbs distributions, and the Bayesian segmentation of images," *IEEE Trans. Pattern Anal. Machine Intell.,*, vol. 6, pp. 721–741, Nov., 1984.

[17] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. R. Stat. Soc.*, ser. B, vol. 39, no. 1, pp. 1–38, 1977.

[18] M. Svénsen, F. Kruggel, and D. Y. von Cramon, "Markov random field modeling of fMRI data using a mean field EM-algorithm," in *Proceedings Energy Minimization Methods in Computer Vision and Pattern Recognition*. ser. Lecture Notes in Computer Science, E. R. Hancock and M. Pelillo, Eds. Berlin, Germany: Springer-Verlag, 1999, pp. 317–330.

[19] M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. K. Saul, "An introduction to variational methods for graphical models," in *Learning in Graphical Models*. ser. Adaptive Computation and Machine Learning, M. I. Jordan, Ed. Cambridge, MA: MIT , 1998.

[20] R. M. Neal and G. E. Hinton, "A view of the EM algorithm that justifies incremental, sparse and other variants," in *Learning in Graphical Models*. ser. Adaptive Computation and Machine Learning, M. I. Jordan, Ed. Cambridge, MA: MIT , 1998.

[21] S. Kirkpatrick , C. D. , and M. P. Vecchi Gellatt Jr., "Optimization by simulated annealing," *Science*, vol. 220, pp. 671–680, 1983.

[22] X. Descombes, F. Kruggel, and D. Y. von Cramon, "fMRI signal restoration using an edge preserving spatio-temporal Markov random field," *NeuroImage*, vol. 8, pp. 340–348, 1998.

[23] W. Snyder, A. Logenthiran, P. Santago, K. Link, G. Bilbro, and S. Rajala, "Segmentation of magnetic resonance images using mean field annealing," *Image Vision Comput.*, vol. 10, no. 6, pp. 218–226, 1992.

[24] J. W. Eaton *et al.*, "GNU Octave, version 2.0,", http://www.che.wisc.edu/octave/, 1998.

[25] M. Meyer, A. D. Friederici, D. Y. von Cramon, F. Kruggel, and C. J. Wiggins, "Auditory sentence comprehension: Different BOLD patterns modulated by task demands as revealed by a 'single-trial' fMRI-study," *NeuroImage*, vol. 7, no. 4, p. S181, 1998.

[26] A. Dove, S. Pollmann, T. Schubert, C. J. Wiggins, and D. Y. von Cramon, "Prefrontal cortex activation in task switching: An event-related fMRI study," Cognitive Brain Research, 1999, to be published.

[27] M. Pélégrini, S. Zysset, F. Kruggel, and D. Y. von Cramon, "Analysis of fMRI time-series with multiple regression and hemodynamic modeling: An event related memory experiment," Human Brain Mapping, 1998, submitted for publication.

[28] N. V. Petersen, J. L. Jensen, J. Burchhardt, and H. Stødkilde-Jørgensen, "State space models for physiological noise in fMRI time series," *NeuroImage*, vol. 7, p. S592, 1998.

[29] B. Biswal, E. A. DeYoe, and J. S. Hyde, "Reduction of physiological fluctuations in fMRI using digital filters," *Magnetic Resonan. Med.*, vol. 35, pp. 107–113, 1996.

[30] M. A. T. Figueiredo, J. M. N. Leitao, and A. K. Jain, "On fitting mixture models," in *Proceedings Energy Minimization Methods in Computer Vision and Pattern Recognition*. ser. Lecture Notes in Computer Science, E. R. Hancock and M. Pelillo, Eds. Berlin, Germany: Springer-Verlag, 1999, pp. 54–69.

[31] J. Zhang, "The mean field theory in EM procedures for Markov random fields," *IEEE Trans. Signal Processing*, vol. 40, pp. 2570–2583, Oct. 1992.

[32] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proc. IEEE*, vol. 77, no. 2, pp. 257–285, 1989.

[33] E. R. Hancock and M. Pelillo, Eds., *Proceedings Energy Minimization Methods in Computer Vision and Pattern Recognition*. ser. Lecture Notes in Computer Science. Berlin, Germany: Springer-Verlag, 1999.

[34] M. I. Jordan, Ed., *Learning in Graphical Models*. ser. Adaptive Computation and Machine Learning. Cambridge, MA: MIT , 1998.