Markov Random Field Modelling of fMRI Data Using a Mean Field EM-algorithm

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

Max-Plank-Institute of Cognitive Neuroscience
Postfach 500 355, D-04303, Leipzig, GERMANY
{svensen,kruggel,cramon}@cns.mpg.de

Abstract. This paper considers the use of the EM-algorithm, combined with mean field theory, for parameter estimation in Markov random field models from unlabelled data. Special attention is given to the theoretical justification for this procedure, based on recent results from the machine learning literature. With these results established, an example is given of the application of this technique for analysis of single trial functional magnetic resonance (fMR) imaging data of the human brain. The resulting model segments fMR images into regions with different 'brain response' characteristics.

1 Introduction

The purpose of this paper is two-fold: first, it reviews the theoretical underpinnings for the use of the EM-algorithm in conjunction with mean field theory for parameter estimation in Markov random field (MRF) models from unlabelled data. Second, it demonstrates the usefulness of this approach by a MRF model for single trial functional magnetic resonance imaging (fMRI) data.

Techniques for learning from unlabelled data are important in the analysis of fMRI data of the human brain, since the data generating mechanism is still far from completely understood. Obvious ethical reasons put limitations on what sort of alternative methods we can use to verify results obtained from fMRI. Other functional brain imaging techniques, which may appear as the obvious answer, suffer exactly the same problem. At the same time, the quantity and quality of fMRI data make automated analysis procedures necessary.

2 Markov Random Fields, Mean Field Theory and the EM-algorithm

In this section, we briefly review Markov random field models, the mean field theory and its connection to the EM-algorithm. Mean field theory is a since long established tool in statistical mechanics and statistical physics. It has also been extensively used in the fields of computer vision and, more recently, machine learning.

E.R. Hancock and M. Pelillo (Eds.): EMMCVPR'99, LNCS 1654, pp. 317–330, 1999.

2.1 Markov Random Field Models

A MRF [24] is a set of N random variables indexed over the vertices, or sites, in an ordered lattice. The typical example is a 2-D image, where the random variables are the labels (e.g. colour) associated with the pixels. The MRF variables are not independent, but are mutually coupled; the key property of MRFs is that the distribution of the random variable associated with a site, n, given the values associated with the sites in a (typically small) *neighbourhood* of n, is independent of the rest of the sites in the MRF. This can be formalised as

$$p(\boldsymbol{x}_n|\boldsymbol{x}_m, n \neq m) = p(\boldsymbol{x}_n|\boldsymbol{x}_m \in \mathcal{N}_n) \ ,$$

where $\boldsymbol{x}_n$ denotes the random variable of site n and $\mathcal{N}_n$ is the set of random variables associated with the sites that are in the neighbourhood of site n.

The distribution over the MRF variables, which is assumed to be strictly positive, can be written as a Gibbs distribution,

$$p(\boldsymbol{x}) = \frac{1}{Z} \exp(-E(\boldsymbol{x})) \tag{1}$$

where $\boldsymbol{x}$ is a NK-dimensional vector formed by concatenating the vectors $\boldsymbol{x}_n$ $(n = 1, \ldots, N)$, E is an *energy function* and Z is a normalisation constant,

$$Z = \sum_{\boldsymbol{x}} \exp(-E(\boldsymbol{x})) \ , \tag{2}$$

where the sum runs over all possible values of $\boldsymbol{x}$. Note that, computing Z, which is known as the *partition function*, is generally tractable only for very small MRFs, since the number of terms in the sum in (2) increase exponentially with the size of the MRF. This is due to the mutual coupling between the MRF variables. Same problem emerges if we want to compute the marginal posterior distribution over any of the individual MRF variables – e.g. for the purpose of parameter fitting – since this requires summing over all remaining variables.

The energy function E defines the properties of the MRF model and can generally be written

$$E(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\Theta}, \boldsymbol{\beta}) = E^{\text{ext}}(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\Theta}) + E^{\text{int}}(\boldsymbol{x}, \boldsymbol{\beta}) \ .$$

E^{ext} denotes the energy (or potential) arising from external influence; in the context of probabilistic image modelling, this typically comes from observed data $\boldsymbol{y}$ via a model determined by parameters $\boldsymbol{\Theta}$, and corresponds to a log-likelihood term. E^{int} denotes the internal energy which, as suggested by the notation, only depends on the MRF variables $\boldsymbol{x}$ and parameter $\boldsymbol{\beta}$, and corresponds to a prior distribution over $\boldsymbol{x}$.

2.2 The Mean Field Theory

To address the computational difficulties associated with MRF models, a number of approximate methods have been proposed [24]. One popular such method is

the so called *mean field* approximation, from statistical mechanics [7]. This consists of replacing $p(\boldsymbol{x}|\boldsymbol{y}, \boldsymbol{\Theta}, \boldsymbol{\beta})$ with an approximating, computationally tractable, parameterised distribution, $q(\boldsymbol{x}|\boldsymbol{m})$. As has been shown by several authors [3,7,31,38], the mean field approximation can be given a formal justification as providing a computationally tractable bound on quantities of interest (e.g. the partition function). Moreover, we can optimise the variational parameter $\boldsymbol{m}$ by minimising the Kullback-Leibler (KL) divergence between $q(\boldsymbol{x}|\boldsymbol{m})$ and $p(\boldsymbol{x}|\boldsymbol{\Theta}, \boldsymbol{\beta}, \boldsymbol{y})$,

$$D(p\|q) = \sum_{\boldsymbol{x}} q(\boldsymbol{x}|\boldsymbol{m}) \ln \frac{q(\boldsymbol{x}|\boldsymbol{m})}{p(\boldsymbol{x}|\boldsymbol{\Theta}, \boldsymbol{\beta}, \boldsymbol{y})} \quad , \tag{3}$$

which is non-negative for all probability distributions q and p and equals zero only when they are identical. The literature on statistical mechanics [7], MRFs [3,38] and probabilistic graphical models [19,31] provides examples of applying this methodology to different models. Section 3.3 in this paper provides an example for a multi-level logistic MRF model.

2.3 A 'Variational' View of the EM-algorithm

Traditionally, the EM-algorithm [8] is viewed as a two-step algorithm for maximum-likelihood parameter estimation from incomplete data. The first step (the E-step) consists of computing the expectation over the random variables which are missing in the data (e.g. the labels in unlabelled data), $\boldsymbol{x}$, given the observed variables, $\boldsymbol{y}$, and the current set of parameters, $\boldsymbol{\Theta}$. The second step (the M-step) maximises the resulting expected complete log-likelihood function with respect to its adjustable parameters, $\boldsymbol{\Theta}$. However, it can also be seen as a algorithm for minimising the variational free energy from statistical mechanics and statistical physics [35,26], linking it to the mean field theory. From this point of view, it is natural to also consider situations where the exact distribution over the missing variables cannot be computed, but has to be replaced by an approximate distribution. This yields an algorithm which maximises a lower bound of the log-likelihood. The difference between this bound and the true log-likelihood is the KL-divergence between the exact and approximating distributions.

Following Jordan et al. [19], our objective is to maximise the log-likelihood function of the observed data $\ln p(\boldsymbol{y}|\boldsymbol{\Theta})$, with respect to the parameters $\boldsymbol{\Theta}$. We now write

$$\begin{aligned}
\ln p(\boldsymbol{y}|\boldsymbol{\Theta}) &= \ln \sum_{\boldsymbol{x}} p(\boldsymbol{y}, \boldsymbol{x}|\boldsymbol{\Theta}, \boldsymbol{\beta}) \\
&= \ln \sum_{\boldsymbol{x}} q(\boldsymbol{x}|\boldsymbol{m}) \frac{p(\boldsymbol{x}|\boldsymbol{\Theta}, \boldsymbol{\beta}, \boldsymbol{y}) p(\boldsymbol{y}|\boldsymbol{\Theta})}{q(\boldsymbol{x}|\boldsymbol{m})} \\
&\geq \sum_{\boldsymbol{x}} q(\boldsymbol{x}|\boldsymbol{m}) \ln \frac{p(\boldsymbol{x}|\boldsymbol{\Theta}, \boldsymbol{\beta}, \boldsymbol{y}) p(\boldsymbol{y}|\boldsymbol{\Theta})}{q(\boldsymbol{x}|\boldsymbol{m})} \\
&= \sum_{\boldsymbol{x}} q(\boldsymbol{x}|\boldsymbol{m}) \ln p(\boldsymbol{y}|\boldsymbol{\Theta}) - q(\boldsymbol{x}|\boldsymbol{m}) \ln \frac{q(\boldsymbol{x}|\boldsymbol{m})}{p(\boldsymbol{x}|\boldsymbol{\Theta}, \boldsymbol{\beta}, \boldsymbol{y})} \quad ,
\end{aligned} \tag{4}$$

where we have used Jensen's inequality. $q(\boldsymbol{x}|\boldsymbol{m})$ is an arbitrary, non-singular probability distribution, parameterised by the variational parameter $\boldsymbol{m}$. From (4), which apart from a change of sign corresponds to the variational free energy from statistical physics, we see directly that the difference between the two sides is the KL divergence (3) between $q(\boldsymbol{x}|\boldsymbol{m})$ and $p(\boldsymbol{x}|\boldsymbol{\Theta}, \boldsymbol{\beta}, \boldsymbol{y})$. As shown by Neal and Hinton [26], maximisation of (4) with respect to $\boldsymbol{m}$ corresponds to the E-step of the EM-algorithm, whenever $q(\boldsymbol{x}|\boldsymbol{m})$ is rich enough to model $p(\boldsymbol{x}|\boldsymbol{\Theta}, \boldsymbol{\beta}, \boldsymbol{y})$ exactly. When, on the other hand, computational considerations force us to resort to simpler distribution models, we can still be certain that resulting algorithm will increase the lower bound of the log-likelihood function, unless already at (a local) maximum.

3 An Application to fMRI Data

Functional magnetic resonance imaging (fMRI) attempts to detect brain activity by localised, non-invasive measurements of the change in blood oxygenation, the so called *BOLD contrast* [27]. This is sensitive to the relative local concentrations of oxygenated hemoglobin (HbO_2) vs. deoxy-hemoglobin and provides an indirect measure of the brain's neuronal activity.

Measurements, in the form of a time-series of images, 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 or not, perform arithmetic calculations, looking at changing scenes, etc.). We only consider fMRI experiments with a single trial (or 'event-related') design, which consist of a series of individual trials. Each trial consist one repetition of the task, followed by a period of rest during which the subject is assumed to be inactive.

When we want to model the fMRI data generating process, there are neurophysiological factors we must take into consideration. The local change in blood oxygenation as an effect of increased neuronal activity, which is called the *hemodynamic response* (HR), is delayed by 2–6 seconds from stimulus onset and dispersed by 2–3.5 seconds. This delay and dispersion vary between subjects, experimental conditions, etc. By contrast, the stimuli 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 modelled as a binary ('box-car') function, i.e. the stimuli is either present or not present.

Traditional analysis of fMRI data essentially amounts to locating so-called *activated* pixels, where the observed measurements shows significant correlation with a function representing the the task. There are different strategies for altering this function to account for the HR, ranging from simply just shifting it in time [2] to convolving it with a HR model function [13,23,29]. The correlation is computed for each pixel individually and the correlation scores are transformed into *Z-scores* [1]. The resulting image of Z-scores, called a *Z-map*, is then thresholded at a level chosen so that the probability of wrongly classifying a pixel as being activated is suitably low (see e.g. [13]).

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 a 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 $t_1, \ldots, t_D$. By choosing a MRF model, we implicitly assume that the spatial distribution of the classes will be locally smooth, so that neighbouring pixels typically belong to the same class. Thus, images will consist of one or more spatially homogeneous regions, each region associated with a parametric HR model function. The model can also be seen as a K-component mixture model [17] in the D-dimensional observable pixel space (i.e. in the temporal domain of the HR), combined with a smoothing MRF prior distribution over pixel classes (in the pixels lattice).

3.1 The Multi-level Logistic MRF Model

To specify the prior pixel class distribution, we use the commonly applied multi-level logistic (MLL) model [12,14], where we specify neighbourhoods such that each pixel only depends on its nearest neighbours (distance equal to one in the lattice of pixels). We represent the MRF variable associated with pixel n as a K-dimensional binary vector, $\boldsymbol{x}_n$. Pixel n belongs to class k if and only if the kth element of $\boldsymbol{x}_n$, denoted x_{nk}, equals 1 and all other elements equals 0. This model contains the binary MRF as a special case ($K = 2$).

We then define the energy function,

$$E^{\text{int}}(\boldsymbol{x}, \beta) = \frac{\beta}{2} \sum_{n}^{N} \sum_{\boldsymbol{x}_m \in \mathcal{N}_n} \boldsymbol{x}_m^{\mathrm{T}} \boldsymbol{U} \boldsymbol{x}_n \ , \tag{5}$$

where $\boldsymbol{U}$ is a $K \times K$ matrix with elements along its diagonal equal to -1 and all other element equal to 1. The scalar β plays the role of a scale parameter for the prior. As β increases, so does the cost for neighbouring pixels from different classes, which in effect forces a smoother image.

3.2 Modelling the Hemodynamic Response

Several model functions for the HR have been proposed [5,13,23,29]; we choose to model the HR using a Gaussian function [22], such that

$$h(t) = \eta \exp\left(-\frac{(t-\mu)^2}{\sigma}\right) + o, \tag{6}$$

where,

μ denotes the *lag*, i.e the time from the onset of the stimuli to the peak of the HR,

σ denotes the *dispersion*, which reflects the rise and decay time,

η denotes the *gain*, or amplitude, of the response, and finally
o denotes an *offset* that defines the minimum level for the HR model function, relative to some baseline level.

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

$$\sigma = \exp(z_\sigma) \qquad \text{and} \qquad \eta = \exp(z_\eta). \tag{7}$$

This will ensure that σ and η are always positive. In case of η, this is actually a simplification, since there is evidence for localised *de*activation in response to stimuli. We denote the parameters $\boldsymbol{\Theta} = [\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K]$, where $\boldsymbol{\theta}_k = [\mu_k, z_{\sigma k}, z_{\eta k}, o_k]$.

We combine the K HR model functions with an isotropic Gaussian noise process with variance α^{-1}, common to all HR model functions. For a pixel n, which belongs to class k, the probability distribution for the D-dimensional observable trial vector $\boldsymbol{y}_n$ can then be written as

$$p(\boldsymbol{y}_n|\boldsymbol{\theta}_k, \alpha) = \left(\frac{\alpha}{2\pi}\right)^{D/2} \exp\left(-\frac{\alpha}{2}\|\boldsymbol{y}_n - \boldsymbol{h}_k\|^2\right) \tag{8}$$

where $\boldsymbol{h}_k$ is a D-dimensional vector corresponding to the HR model function, computed from (6) and (7) at times $t_1, \ldots, t_D$, using the parameter vector $\boldsymbol{\theta}_k$. Note that, this model implicitly assumes that any two random vectors $\boldsymbol{y}_n$ and $\boldsymbol{y}_m$, $n \neq m$, are independent given the classes of the corresponding pixels.

From the negative logarithm of (8), we can derive the external energy for the MRF model

$$E^{\text{ext}}(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\Theta}, \alpha) = \sum_{n,k}^{N,K} \frac{\alpha}{2}\|\boldsymbol{y}_n - \boldsymbol{h}_k\|^2 x_{nk} \ , \tag{9}$$

where $\sum_{n,k}^{N,K} = \sum_n^N \sum_k^K$; this abbreviated notation will be used throughout the rest of this paper. Recall that x_{nk} is 1 if and only if pixel n belongs to class k and 0 otherwise. The term arising from the normalisation factor, $(\alpha/2\pi)^{D/2}$, has been dropped as it does depend on $\boldsymbol{x}$.

A Prior for the HR Parameters. Given our knowledge about neurology and fMRI in general and the experimental design in particular, we have certain a-priori beliefs about what can be considered reasonable values of the HR parameters. We can express beliefs by specifying a prior distribution over the HR parameters. Here, we choose a simple independent Gaussian distribution,

$$p(\boldsymbol{\Theta}) = \left(\prod_i^{4K} 2\pi V_{\boldsymbol{\Theta}}(i,i)\right)^{-1/2} \exp\left(-\frac{1}{2}\left(\boldsymbol{\Theta} - \overline{\boldsymbol{\Theta}}\right)^{\mathrm{T}} \boldsymbol{V}_{\boldsymbol{\Theta}}^{-1}\left(\boldsymbol{\Theta} - \overline{\boldsymbol{\Theta}}\right)\right), \tag{10}$$

where $\overline{\boldsymbol{\Theta}}$ is a $4K$-element vector containing the expected values for μ_k, $z_{\sigma k}$, $z_{\eta k}$ and o_k, $k = 1, \ldots, K$, and $\boldsymbol{V}_{\boldsymbol{\Theta}}$ is a diagonal covariance matrix with the corresponding variances along its diagonal.

3.3 Mean Field Equations for the Multi-level Logistic Model

Combining (9) with (5), we get the energy function

$$E = \sum_{n,k}^{N,K} E_{nk} x_{nk} \ , \tag{11}$$

where

$$E_{nk} = \frac{\alpha}{2} \|\boldsymbol{y}_n - \boldsymbol{h}_k\|^2 + \frac{\beta}{2} \sum_{\boldsymbol{x}_m \in \mathcal{N}_n} \boldsymbol{x}_m^{\mathrm{T}} \boldsymbol{U}_k \ , \tag{12}$$

where in turn $\boldsymbol{U}_k$ denotes the kth column of $\boldsymbol{U}$. From (1), we can write the corresponding distribution over the MRF as

$$p(\boldsymbol{x}|\boldsymbol{y}, \boldsymbol{\Theta}, \alpha, \beta) = \frac{1}{Z} \exp\left(- \sum_{n,k}^{N,K} E_{nk} x_{nk} \right) \ . \tag{13}$$

For the mean field approximation, we choose q to be a simple independent multinomial distribution, where each lattice variable, $\boldsymbol{x}_n$, has its own variational parameter, $\boldsymbol{m}_n$,

$$q(\boldsymbol{x}|\boldsymbol{m}) = \prod_{n,k}^{N,K} m_{nk}^{x_{nk}}. \tag{14}$$

$\boldsymbol{m}_n$ is a K-dimensional vector whose elements are all positive and sum to 1; the kth element of $\boldsymbol{m}_n$, m_{nk}, represents the probability that pixel n belongs to class k. $\boldsymbol{m}$ denotes the concatenation of $\boldsymbol{m}_n$, $n = 1, \ldots, N$.

Substituting (13) and (14) into the quotient in (3), performing some elementary algebra and then averaging with respect to $q(\boldsymbol{x}|\boldsymbol{m})$, we get

$$\sum_{n,k}^{N,K} [m_{nk} \ln m_{nk} + m_{nk} E_{nk}^{'}] + \ln Z \ ,$$

where $E_{nk}^{'}$ is identical to E_{nk} in (12), except that $\boldsymbol{x}_m$ has been replaced by $\boldsymbol{m}_m$. Taking the derivative of this with respect to m_{nk}, using Lagrange multipliers, ζ_n, to ensure that $\sum_k m_{nk} = 1$ for all n (see e.g. [10]), we get

$$\ln m_{nk} + 1 + E_{nk}^{''} + \zeta_n \ ,$$

where $E_{nk}^{''}$ is identical to $E_{nk}^{'}$, except that the factor $\beta/2$ has been replaced by β as a consequence of neighbourhood symmetry. Setting these to zero, we can solve for ζ_n, using that $\sum_k m_{nk} = 1$, and subsequently for m_{nk}, yielding

$$m_{nk} = \frac{\exp(-E_{nk}^{''})}{\sum_{k'} \exp(-E_{nk'}^{''})} \ . \tag{15}$$

which are the mean field equations for the MLL model which can be solved iteratively for a fixed point solution. An alternative derivation, drawing on analogies to statistical mechanics, can be found by Zhang [36].

At the moment, it is not established under which conditions these equations converge; Zhang [37] analysed the convergence for an Ising model equivalent with a binary MRF, which was found to converge under certain conditions. In practice, convergence does not appear to be a problem – the parameter $\boldsymbol{m}$ settles rapidly, and failure to reach absolute convergence simply means that our bound on the log-likelihood will be less tight.

3.4 Parameter Estimation

Until now, we have implicitly assumed that the all the parameters are known. This is typically not the case, but given the theory in Sect. 2, we can use the EM-algorithm to estimate parameters of interest. In the E-step, we compute the mean field approximation (15) to the posterior distribution over the MRF variables. In the following M-step, we maximise the resulting expected complete log-likelihood with respect to the parameters.

Here we restrict ourselves to maximisation with respect to $\boldsymbol{\Theta}$ and α. The hyperparameters for the prior distribution over HR parameters are set using knowledge about the experimental design and general HR characteristics. β is set by experimenting; experience so far suggest that the final result is not very sensitive to the exact choice of β, which was also reported in [36].

We derive our objective function from a hypothetical log-likelihood function, where the class labels, $\boldsymbol{x}_n$, are known. As we assume that the observations at different pixels are independent given the corresponding class labels, we get the penalised log-likelihood function from (8) and (10) as

$$\ln p(\boldsymbol{\Theta}) + \frac{ND}{2} \ln \alpha - \frac{\alpha}{2} \sum_{n,k}^{N,K} x_{nk} \|\boldsymbol{y}_n - \boldsymbol{h}_k\|^2 \ ,$$

where we have omitted terms which do not depend on $\boldsymbol{\Theta}$ or α. We obtain the corresponding expected complete penalised log-likelihood function simply by replacing the x_{nk} by their correponding mean-field expectations, m_{nk}, computed from (15).

Maximisation of the resulting objective function with respect to $\boldsymbol{\Theta}$ is done by numerical optimisation[1]. For α, we get an update formula in closed form

$$\alpha = \frac{ND}{\sum_{n,k}^{N,K} m_{nk} \|\boldsymbol{y}_n - \tilde{\boldsymbol{h}}_k\|^2} \ ,$$

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

[1] We use the function `fsolve` from the software package Octave [11] for this purpose.

Mean Field Annealing. The parameter estimation problem is fairly difficult optimisation problem, and empirical evidence suggest that there are many poor local optima where the optimisation procedure can get stuck. To reduce the risk of this, we employ a simulated annealing scheme [6,20], multiplying (11) by an inverse temperature factor $(1/T)$, $T \geq 1$. Setting $T > 1$ will smooth the (approximate) posterior distribution, which in effect will smooth out shallow local optima. Thus, optimisation in the high-T regime (say $T = 10$ for data normalised to zero mean and unit variance) will hopefully find a global optimum which then can be tracked by re-estimating the parameters, $\boldsymbol{\Theta}$, as T is being decreased to 1. Note that, as long as $T > 1$, α is kept fixed to 1; this annealing phase is then followed by further optimisation where both $\boldsymbol{\Theta}$ and α are adapted. Annealing approaches have been used successfully with MRF models for restoration of e.g. fMRI images [9] and, in combination with mean field theory, anatomical ('non-functional') magnetic resonance images [33].

3.5 Example

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. The experiment employed a single trial design where each trial had a length of 24 seconds. Each trial started with a sentence being read out, which lasted 2.3–4.5 seconds. fMR images with a spatial resolution of 64×128 pixels were collected every 2 seconds, so the trial vector for each pixel consists of 12 measurements. In total, there were 76 trials, although the first 4 were not used. The data were pre-processed to correct for subject movements, remove baseline trends and filter out physiological and system noise [21].

For this example, data from the 72 selected trials were averaged, to improve the signal to noise ratio. The resulting averaged data were used to train a model with 2 HR functions 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 the time-averaged response at individual pixels, averaged over the posterior distribution over pixel classes. The HR functions shared a common prior given in table 1 and β was set to 1. The fitting procedure started with 20 iterations during which T was decreased linearly 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.

Table 1. Hyperparameters for the prior distribution over HR parameters used in the example described in Sect. 3.5. μ and z_σ are measured in (log) time steps, while z_η and o are measured (log) relative to a normalised BOLD response

μ	V_μ	z_σ	V_{z_σ}	z_η	V_{z_η}	o	V_o
6	3	$\ln 2$	$2/2^2$	$\ln 4$	$3/4^2$	0	1

The left image in Fig. 1 shows the resulting segmentation of the functional mask obtained from our model; pixels have been assigned to the class with highest posterior class probability, computed using the mean field approximation, after the parameters had converged. Figure 2 shows the corresponding HR functions. Different types of filtering in the pre-processing [21] will cause marginal variations in these results, but the overall picture will remain the same. The right image of Fig. 1 shows a Z-map for the same data set, based on correlation with a shifted 'box-car' function, overlaid on the functional mask (see e.g. [13]); note that, only pixels with positive activation are shown, as we do not consider deactivated regions.

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

4 Discussion

In this paper, we have reviewed the use of the EM-algorithm combined with mean field theory for parameter estimation from unlabelled data in MRF models, and the theoretical justification for this, based on results from the machine learning literature. Furthermore, we have shown an application of this procedure for analysis of fMRI data – a learning problem of inherent unsupervised nature.

It should be clear that the overall framework is independent of the choice of HR model function, and thus other variants could be considered. Similarly, we could consider the use of a more elaborate noise model; Kruggel and von Cramon [22] discuss the use of an autoregressive (AR(1)) noise model in the spatial domain.

A limitation of the work we have presented in this paper is the remaining number of free parameters. β is currently set by experimenting. Deriving a method for updating β in the light of observed data is difficult, since the partition function for the MRF prior depends on β. Zhang [36] suggested using a mean field approximation also for the partition function, but as pointed out by Jordan et al. [19], this result in an update equation based on two different bounds, which theoretically may *decrease* the log-likelihood of the data given β. An alternative approach would be to use Monte Carlo sampling methods for the parameter fitting. Such an approach would be computationally demanding; a potential remedy could be to estimate $\boldsymbol{\Theta}$ and α using mean field theory, and use Monte Carlo methods only for the updates of β, which need not be updated every iteration of the EM-algorithm. The number of HR components, K, is currently set by the user, based on empirical evidence, prior knowledge and interpretability. It would clearly be desirable to be able to estimate K from the data, but such estimation would face the same difficulties as the estimation of β, since comparing models with different values for K requires computing the corresponding partition functions. Nevertheless, methods based on minimum de-

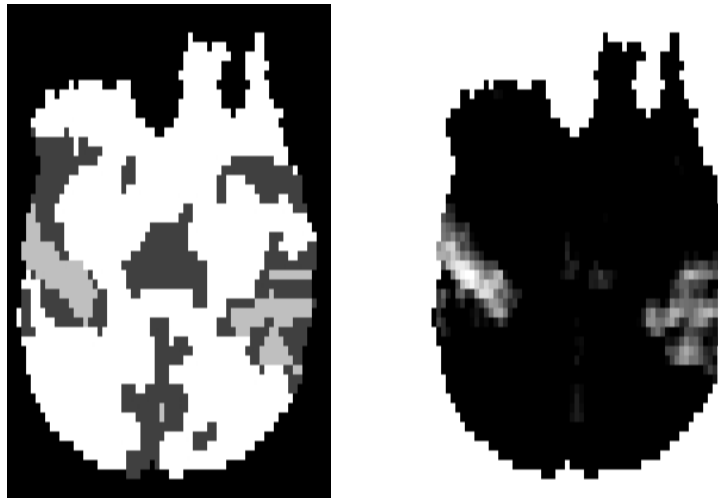

Fig. 1. A segmentation obtained using proposed method (left) and a corresponding correlation based Z-map (right), for the data described in Sect. 3.5. In the left image, pixels in the functional mask have been classified according to their maximum posterior class probabilities; the corresponding HR model functions plotted in Fig. 2; the dominating white class corresponds to the background function. In the Z-map, which is overlaid on the functional mask, pixels have been shaded according to their Z-score, where brighter pixels indicate higher activation

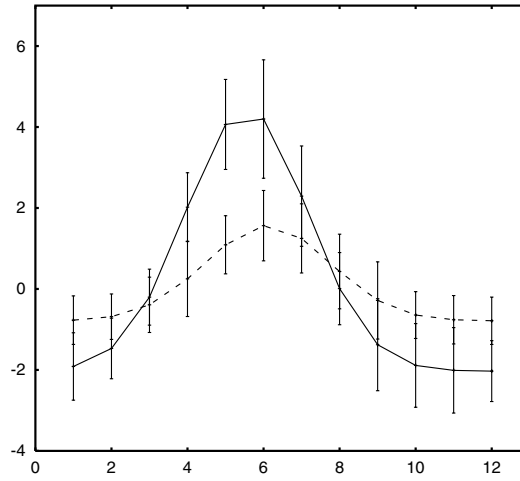

Fig. 2. The HR functions corresponding to the segmentation shown in Fig. 1. The solid line corresponds to the light grey pixels while the dashed line corresponds to dark grey pixels. The error bars corresponds to 1 standard deviation of the data from the mean given by the curve

scription length theory [30] or maximum entropy principles [34], combined with approximate methods for computing the partition function, could be considered.

The idea of deriving mean field equations by minimising the KL-divergence given a choice of approximating distribution raises the question whether other approximating distributions can be found that gives a tighter bound and remains computationally tractable. Jordan et al. [19] gives several such examples examples in the context of learning in graphical models, some of which could potentially be applied to MRF models.

An obvious limitation of the mean field approximation discussed in Sect. 3.3 is that it is unimodal, i.e. the spatial distribution of pixel class labels is centred around a single configuration. This might be a a reasonable approximation when modelling averaged data from a single experiment, as in Sect. 3.5, but if we want to investigate between-trial variance within one experiment or even the (dis)similarities between trials from different experiments, it is clearly insufficient. Jaakkola and Jordan [16] proposed the use of a mixture of fully factorised mean field distributions, and Bishop et al. [4] empirically demonstrated the usefulness of this approach in the context of sigmoid belief networks. A future direction of research will be to extend the approach presented in this paper to the use of such mixture distributions, and investigate the usefulness of this for the purpose of fMRI data modelling.

For fMRI data, it is also natural to consider modelling structure in the time domain, since an experiment consists a sequence of images corresponding to the sequence of trials. Theory for such a model could be built on the existing theory for hidden Markov models (HMM) [28], which has recently been subject to substantial development in the context of graphical models and machine learning [15,19,32]. This strand of research will be pursued as an extension of the mixture model discussed in the previous paragraph.

References

1. M. Abramowitz and I. A. Stegun. *Handbook of mathematical functions with formulas, graphs and mathematical tables, 10th printing.* National Bureau of Standards Applied Mathematics Series 55, Washington, 1972. 320
2. 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 Resonance in Medicine*, 30:161–173, 1993. 320
3. G. L. Bilbro, W. E. Snyder, and R. C. Mann. Mean-field approximation minimizes relative entropy. *Journal of the Optical Society of America A*, 8(2):290–294, 1991. 319
4. C. M. Bishop, N. D. Lawrence, T. S. Jaakkola, and M. I. Jordan. Approximating posterior distributions in belief networks using mixtures. In M. I. Jordan, M. J. Kearns, and S. A. Solla, editors, *Advances in Neural Information Processing Systems*, volume 10. MIT Press, 1998. 327
5. 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 Resonance in Medicine*, 35:261–277, 1996. 321

6. V. Cerný. A thermodynamical approach to the travelling salesman problem: An efficient simulation algorithm. *Journal of Optimization Theory and Applications*, 45:41–51, 1985. 325
7. D. Chandler. *Introduction to Modern Statistical Mechanics*. Oxford University Press, New York, 1987. 319
8. A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Statistical Methodology, Journals of the Royal Statistical Society, Series B*, 39(1):1–38, 1977. 319
9. X. Descombes, F. Kruggel, and D. Y. von Cramon. fMRI signal restoration using an edge preserving spatio-temporal Markov random field. *NeuroImage*, 8:340–348, 1998. 325
10. L. C. W. Dixon. *Nonlinear Optimisation*. English Universities Press, London, 1972. 323
11. J. W. Eaton et al. GNU Octave. Available on the Internet at URL: `http://www.che.wisc.edu/octave/`, 1998. version 2.0. 324
12. H. Elliot, H. Derin, R. Christi, and D. Geman. Application of the Gibbs distribution to image segmentation. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, pages 32.5.1–32.5.4, San Diego, 1984. IEEE. 321
13. K. J. Friston, P. Jezzard, and R. Turner. Analysis of functional MRI time-series. *Human Brain Mapping*, 1:153–171, 1994. 320, 321, 326
14. S. Geman and D. Geman. Stochastic relaxation, Gibbs distributions, and the Bayesian segmentation of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6(6):721–741, 1984. 321
15. Z. Gharamani and M. I. Jordan. Factorial hidden Markov models. *Machine Learning*, 29:245–273, 1997. 328
16. T. S. Jaakkola and M. I. Jordan. Improving the mean field approximation via the use of mixture distributions. In Jordan [18]. Proceedings of the NATO Advanced Study Institute, Erice, Italy, 1996. 327
17. A. Jepson and M. Black. Mixture models for image representation. Technical report, Department of Computer Science, University of Toronto, March 1996. PRECARN ARK Project Technical Report ARK96-PUB-54. 321
18. M. I. Jordan, editor. *Learning in Graphical Models*. Adaptive Computation and Machine Learning. MIT Press, 1998. Proceedings of the NATO Advanced Study Institute, Erice, Italy, 1996. 329, 330
19. M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. K. Saul. An introduction to variational methods for graphical models. In Jordan [18]. Proceedings of the NATO Advanced Study Institute, Erice, Italy, 1996. 319, 326, 327, 328
20. S. Kirkpatrick and C. D. and M. P. Vecchi Gellatt, Jr. Optimization by simulated annealing. *Science*, 220:671–680, 1983. 325
21. F. Kruggel, X. Descombes, and D. Y. von Cramon. Preprocessing of fMR data sets. In *Workshop on Biomedical Image Analysis*, pages 211–220, Santa Barbera, 1998. IEEE Computing Society. 325, 326
22. F. Kruggel and D. Y. von Cramon. Modelling the hemodynamic response in single trial fMRI experiments. *Magnetic Resonance in Medicine*, 1998. Under review. 321, 326
23. N. Lange and S. L. Zeger. Non-linear time-series analysis for human brain mapping by magnetic resonance imaging. *Applied Statistics, Journals of the Royal Statistical Society, Series C*, 46:1–29, 1997. 320, 321
24. S. Z. Li. *Markov Random Field Modeling in Computer Vision*. Springer-Verlag, Tokyo, 1995. 318

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*, 7(4):S181, 1998. 325
26. R. M. Neal and G. E. Hinton. A view of the EM algorithm that justifies incremental, sparse and other variants. In Jordan [18]. Proceedings of the NATO Advanced Study Institute, Erice, Italy, 1996. 319, 320
27. S. Ogawa, T. M. Lee, A. R. Kay, and D. W. Tank. Brain magnetic resonance imaging with contrast dependent blood oxygenation. *Proceedings of the National Acadademy of Sciences, USA*, 87:9868–9872, 1990. 320
28. L. R. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–285, 1989. 328
29. 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*, 6:283–300, 1998. 320, 321
30. J. Rissanen. Modelling by shortest data description. *Automatica*, 14:465–471, 1978. 327
31. L. K. Saul and M. I. Jordan. Exploiting tractable substructures in intractable networks. In D. S. Touretzky, M. C. Mozer, and M. E. Hasselmo, editors, *Advances in Neural Information Processing Systems*, volume 8, pages 486–492. MIT Press, 1996. 319
32. P. Smyth, D. Heckerman, and M. I. Jordan. Probabilistic independence networks for hidden Markov probability models. *Neural Computation*, 9:227–270, 1997. 328
33. W. Snyder, A. Logenthiran, P. Santago, K. Link, G. Bilbro, and S. Rajala. Segmentation of magnetic resonance images using mean field annealing. *Image and Vision Computing*, 10(6):218–226, 1992. 325
34. N. Wu. *The Maximum Entropy Method*. Springer Series in Information Sciences. Springer-Verlag, 1997. 327
35. A. L. Yuille, P. Stolorz, and J. Utans. Statistical physics mixtures of distributions, and the EM algorithm. *Neural Computation*, 6:334–340, 1994. 319
36. J. Zhang. The mean field theory in EM procedures for Markov random fields. *IEEE Transactions on Signal Processing*, 40(10):2570–2583, 1992. 324, 326
37. J. Zhang. The convergence of mean field procedures for MRF's. *IEEE Transactions on Image Processing*, 5(12):1662–1665, 1995. 324
38. J. Zhang. The application of the Gibbs-Bogoliubov-Feynman inequality in mean field calculations for Markov random fields. *IEEE Transactions on Image Processing*, 5(7):1208–1214, 1996. 319