Mixture Models with Adaptive Spatial Regularisation for Segmentation with an Application to FMRI Data

FMRIB Technical Report TR04MW1
(A related paper has been submitted to TMI IEEE)

Mark W. Woolrich, Timothy E.J. Behrens, Christian F. Beckmann and Stephen M. Smith

Oxford Centre for Functional Magnetic Resonance Imaging of the Brain (FMRIB),
Department of Clinical Neurology, University of Oxford, John Radcliffe Hospital,
Headley Way, Headington, Oxford, UK
Corresponding author is Mark Woolrich: woolrich@fmrib.ox.ac.uk

Abstract

Mixture models are often used in the statistical segmentation of medical images. For example, they can be used for the segmentation of structural images into different matter types or of statistical parametric maps in functional imaging. Non-spatial mixture models segment using models of just the histogram of intensity values. Spatial mixture models have also been developed which augment this histogram information with spatial regularisation using Markov Random Fields. However, these techniques have control parameters, such as the strength of spatial regularisation, which need to be tuned heuristically to particular datasets. We present a novel spatial mixture model within a fully Bayesian framework with the ability to perform fully *adaptive* spatial regularisation using Markov Random Fields. This means that the amount of spatial regularisation does not have to be tuned heuristically but is adaptively determined from the data. We examine the behaviour of this model when applied to artificial data with different spatial characteristics, and to FMRI statistical parametric maps.

1 Introduction

Automatic segmentation in medical imaging is an important tool. It allows for objective classification of regions in images based upon statistical models. Typical tasks include the segmentation of structural images into different matter types and of statistical parametric maps in functional imaging to perform inference.

Statistical segmentation can be achieved by modelling the histogram of the observations as being made up of a mixture of the distributions of the different classes that we want to segment the image into. For example, in brain segmentation tasks the observations are typically intensity levels in a structural brain image. The mixture is then made up of the distributions of the different matter types in the brain (e.g. white matter, grey matter, cerebral spinal fluid) (Wells et al., 1996; Guillemaud and Brady, 1997; Held et al., 1997; Zhang et al., 2001). The task is then to classify areas in the brain to a particular matter type.

Another example is in functional brain imaging, where the observations might be statistical parametric maps (SPMs). The SPMs are typically the result of a temporal linear model of the 4-dimensional functional brain imaging data (e.g. FMRI), and the SPM statistics represent the height (and its uncertainty) of the measured response to a neural stimulation. The mixture is then made up of distributions representing non-activation, activation, and possibly deactivation. The task is then to classify areas in the brain as either activating, deactivating, or not activating. Everitt and Bullmore (1999) considered a non-spatial approach to mixture modelling for SPMs in FMRI.

Spatial mixture models have also been developed to augment this histogram information with spatial regularisation. This is to encode the prior belief that neighbouring voxels in our images are likely to come from the same

class. The spatial mixture modelling implemented in Salli et al. (1999); Zhang et al. (2001); Marroquin et al. (2003) introduced spatial regularisation of the classification labels using discrete Markov Random Field (MRF) priors. Salli et al. (1999) apply spatial mixture modelling to FMRI, and Zhang et al. (2001) apply spatial mixture modelling to structural brain segmentation. Essentially, Salli et al. (1999); Zhang et al. (2001); Marroquin et al. (2003) incorporate a discrete MRF prior on the spatial map of classification labels.

Svensén et al. (2000) use mixture models in a different and interesting way on FMRI data. Instead of having a mixture model on SPMs of the activation height, they effectively have a mixture model on the haemodynamic response function (HRF). This allows segmentation into regions with different characteristics of the HRF. They also use a discrete MRF prior on the spatial map of classification labels.

A discrete MRF prior simply penalises when neighbouring voxels are of a different class. Crucially, the amount of penalisation depends on an MRF control parameter which controls how strong this spatial regularisation is. Marroquin et al. (2003) refer to this as the parameter which controls the granularity of the field, and they discuss how the use of different values for this parameter can affect the resulting segmented field. However, a problem with Salli et al. (1999); Svensén et al. (2000); Zhang et al. (2001); Marroquin et al. (2003)'s use of discrete classification label MRFs is that the normalising constant (or partition function) of the MRF is not known analytically. This makes it very difficult to infer on the MRF control parameter. Consequently, Salli et al. (1999); Svensén et al. (2000); Zhang et al. (2001); Marroquin et al. (2003), effectively have to heuristically tune the MRF control parameter. This is a strong limitation. If we applied the model to a new dataset then the optimal MRF control parameter would need to be heuristically deduced, and would then be open to subjective judgement. It would be favourable to allow the actual data itself to automatically determine the amount of spatial regularisation.

Hartvig and Jensen (2000) also use a spatial mixture model, on FMRI data. Their solution to the problem of not knowing the partition function for discrete MRF priors is to use a different spatial prior altogether. The joint spatial prior over the map of labels is specified indirectly by specifying the marginal prior for the labels in a 3×3 voxel neighbourhood. The advantage is that by choosing marginal priors which depend on summary statistics (such as the number of voxels in the 3×3 voxel neighbourhood of the same class) the posterior probability that a voxel is activated can be calculated analytically. This provides inference which is much quicker than iterative techniques such as ICM, simulated annealing or MCMC.

However, Hartvig and Jensen (2000) themselves make the point that these marginal spatial priors are not as flexible as MRFs. There is also a problem when we come to adaptively determine the global parameters, such as the class distribution parameters and the parameters which control the strength of the spatial regularisation. This is because it is not obvious how to go from the modelling of marginals in 3×3 voxel neighbourhoods to the joint posterior over the entire spatial map, and it is the latter that is required to do inference on the global parameters. As a result Hartvig and Jensen (2000) propose a contrast function to represent the joint posterior so that the global parameters can be determined. However, whilst a sensible choice of contrast function can be made, it is still somewhat arbitrary. Hence the difference between the resulting global parameter estimates and those that could be obtained if the joint posterior was available is unclear. In addition, this means that there is no formal way to assess or include uncertainty in these global parameters.

In this paper we propose an alternative spatial mixture model to Hartvig and Jensen (2000)'s in determining the amount of spatial regularisation. Unlike Hartvig and Jensen (2000) we use a discrete labels Markov Random Field. This paper describes a novel way to do spatial mixture modelling with a MRF with the amount of spatial regularisation determined unambiguously and adaptively from the data.

This is achieved by approximating the discrete labels with a vector of continuous weights. The imposed properties of our weights vector ensures that the new posterior distribution is the same as the posterior distribution when we use discrete labels. Crucially, instead of a discrete MRF prior on the discrete labels, we can now use a continuous Gaussian MRF (or conditionally specified auto-regressive process) prior on parameters related to the continuous weights, for which we *do* know what the normalising constant is.

Subsequently, are able to automatically determine the continuous Gaussian MRF control parameter, allowing us to adaptively determine the amount of spatial regularisation. Heuristic tuning of control parameters is no longer required. All parameters in the model are adaptively determined from the data.

Paper Summary In section 2 we describe the mixture models considered. Firstly, we describe the discrete classification mixture models. We then describe how we can approximate these discrete classification mixture models using the continuous weights mixture models to allow adaptive spatial regularisation. In section 3 we describe the class distributions. We then describe how we infer on the model using Markov Chain Monte Carlo

techniques in section 4. In section 5 we examine the behaviour of this model when applied to artificial data with different spatial characteristics, and finally in section 6 we apply it to FMRI statistical parametric maps.

2 Modelling

We consider a mixture model on a regular spatial lattice of observations y, where y_i is the observation at spatial location i, and $i = 1, \ldots, N$. There are $k = 1, \ldots, K$ distributions/classes in the mixture (e.g. for the FMRI data in section 6 we use $K = 3$ classes, one for activation, a second for de-activation and a third for non-activation). The parameters of the class distributions are represented by the vector $\boldsymbol{\theta} = \{\theta_k : k = 1 \ldots K\}$.

We start by describing discrete classification mixture models. We will then go on to describe how we can approximate these models using the continuous weights mixture model, where we have replaced the discrete labels with vectors of continuous weights. We shall then see how the continuous weights spatial mixture model allows for the adaptive determination of the amount of spatial regularisation.

2.1 Discrete Labels Mixture Model

The spatial map of discrete class labels is $\boldsymbol{x}$, where x_i is the class label at spatial location i. Assuming conditional independence of the likelihood, the full posterior distribution of the unknown parameters given the observed spatial map is:

$$p(\boldsymbol{x} = \boldsymbol{\kappa}, \boldsymbol{\theta}, \boldsymbol{\lambda}|\boldsymbol{y}) \propto \prod_i^N \{p(y_i|x_i = \kappa_i, \theta_{\kappa_i})\} p(\boldsymbol{x} = \boldsymbol{\kappa}|\boldsymbol{\lambda}) p(\boldsymbol{\lambda}) p(\boldsymbol{\theta}) \tag{1}$$

where $\boldsymbol{\kappa}$ is a specific configuration (spatial map) of the class labels, and $\boldsymbol{\lambda}$ are any hyperparameters required to describe the prior on spatial map of class labels $\boldsymbol{x}$.

We consider three different mixture models, which are distinguished by their priors on $\boldsymbol{x}$ ($p(\boldsymbol{x}|\boldsymbol{\lambda})$). These are non-spatial, with and without global class proportion parameters, and spatial mixture models.

2.1.1 Non-spatial without Class Proportions

We assume spatial independence between the classification labels, x_i, at each voxel:

$$p(\boldsymbol{x} = \boldsymbol{\kappa}|\boldsymbol{\lambda}) = \prod_i^N p(x_i = \kappa_i|\boldsymbol{\lambda}) \tag{2}$$

and that the prior on the discrete classification labels, x_i, is a non-informative distribution with each class having equal probability. Using this in equation 1, the posterior becomes:

$$p(\boldsymbol{x} = \boldsymbol{\kappa}, \boldsymbol{\theta}, \boldsymbol{\lambda}|\boldsymbol{y}) \propto \prod_i^N \{p(y_i|x_i = \kappa_i, \theta_{\kappa_i})\} p(\boldsymbol{\theta}) \tag{3}$$

2.1.2 Non-spatial with Class Proportions

Again, we assume spatial independence between the classification labels as in equation 2. However, taking $\boldsymbol{\lambda} = \boldsymbol{\pi} = \{\pi_k : k = 1 \ldots K\}$, where π_k are the adaptive global class proportion parameters, the prior on x_i is now:

$$p(x_i = k|\boldsymbol{\pi}) = \pi_k \tag{4}$$

The global class proportions, π_k, are the relative weighting of each of the distributions in the mixture. The prior on $\boldsymbol{\pi}$ is non-informative (uniform) over the range $0 < \pi_k < 1$, and $\sum_k^K \pi_k = 1$. Using this in equation 1, the posterior becomes:

$$p(\boldsymbol{x} = \boldsymbol{\kappa}, \boldsymbol{\theta}, \boldsymbol{\lambda}|\boldsymbol{y}) \propto \prod_i^N \{\pi_{\kappa_i} p(y_i|x_i = \kappa_i, \theta_{\kappa_i})\} p(\boldsymbol{\theta}) p(\boldsymbol{\pi}) \tag{5}$$

2.1.3 Spatial

The prior on $\boldsymbol{x}$ is now a spatial prior. In this work we assume a discrete MRF (Besag, 1986; Geman and Geman, 1984). Taking $\boldsymbol{\lambda} = \{\phi_x\}$, where ϕ_x is the MRF control parameter, which controls the amount of spatial regularisation. We have:

$$p(\boldsymbol{x} = \boldsymbol{\kappa}|\phi_x) \propto f(\phi_x) exp\left(-\frac{\phi_x}{4}\sum_i \sum_{j \in \mathcal{N}_i} I[x_i \neq x_j]\right) \tag{6}$$

where $\mathcal{N}_i$ is the spatial neighbourhood of i (for this we use 26-connectivity in 3-dimensions), $I[x_i \neq x_j]$ is an indicator function (it is one if $x_i \neq x_j$ and is zero otherwise), and $f(\phi_x)$ is some unknown function of ϕ_x. This prior is identical to the prior used in Salli et al. (1999); Zhang et al. (2001), if the parameter ϕ_x is set to one. Usually, ϕ_x is hand tuned to work well for particular types of dataset. The "best" value for ϕ_x will depend on the amount of, and topography of, the different classes. Marroquin et al. (2003) refer to this as the parameter which controls the granularity of the field, and they discuss how the use of different values for this parameter can affect the resulting segmented field. Indeed, we shall demonstrate later how fixing the amount of spatial regularisation to a single value will perform considerably less well than determining it adaptively from the data.

The hyperprior we use on ϕ_x is a non-informative Gamma distribution:

$$p(\phi_x|\tilde{a}_{\phi_x}, \tilde{b}_{\phi_x}) = Ga(\phi_x; \tilde{a}_{\phi_x}, \tilde{b}_{\phi_x}) \tag{7}$$

Using all of this in equation 1, the posterior becomes:

$$p(\boldsymbol{x} = \boldsymbol{\kappa}, \boldsymbol{\theta}, \boldsymbol{\lambda}|\boldsymbol{y}) \propto \prod_i^N \{p(y_i|x_i = \kappa_i, \theta_{\kappa_i})\} p(\boldsymbol{x} = \boldsymbol{\kappa}|\phi_x) p(\phi_x) p(\boldsymbol{\theta}) \tag{8}$$

Clearly, a fourth mixture model could be considered. This would be a spatial mixture model *with* global class proportions. However, it is far from clear how we would combine the prior on x in equation 4 with that in equation 6. Anyway, as we shall see in the results, we obtain good global histogram fits with the spatial mixture model specified here without including global class proportions.

2.2 Continuous Weights Mixture Model

The discrete labels model does not allow the amount of spatial regularisation to be determined adaptively from the data. Therefore, in this section we will show how we can replace the discrete labels model with a new model that is an approximation to it. This new model *does* allow for the amount of spatial regularisation to be determined adaptively from the data.

The problem with the discrete labels model is that when we introduce spatial priors via the discrete MRF, we have an *unknown* function $f(\phi_x)$ as part of the prior (equation 6). It represents the effect which the MRF control parameter, ϕ_x, has on the normalising constant (also known as the partition function), and is required, up to a proportionality constant, to allow adaptive determination of the MRF control parameter, ϕ_x. The function, $f(\phi_x)$, can not be calculated analytically and computational evaluation is very difficult, requiring summation over all of the possible states of x. As a result, the usual approach to such models (e.g. Salli et al. (1999); Zhang et al. (2001)) is to set ϕ_x to a fixed value based upon prior knowledge from other datasets of what makes a good value for ϕ_x. This means that the amount of spatial regularisation used by the MRF is non-adaptive.

To establish this approach we firstly introduce a vector of continuous class weights to replace the three discrete class models already described, whilst approximating the same posterior distribution. We then proceed to show how this allows adaptive determination of the amount of spatial regularisation in the spatial mixture models.

2.2.1 Non-spatial without Class Proportions

We are going to approximate the distribution in equation 3 by replacing the discrete labels, x_i, with $K \times 1$ continuous weights vectors, $\boldsymbol{w_i}$:

$$p(\boldsymbol{w}, \boldsymbol{\theta}|y) \propto \prod_i^N \sum_{k=1}^K \{w_{ik} p(y_i|x_i = k, \theta_k)\} p(\boldsymbol{w}) p(\boldsymbol{\theta}) \tag{9}$$

where $\boldsymbol{w} = \{\boldsymbol{w_i} : i = 1 \ldots N\}$ and $\boldsymbol{w_i} = \{w_{ik} : k = 1 \ldots K\}$ is the continuous weights vector at voxel i. Equation 9 only approximates equation 3 if we apply certain constraints to the continuous weights vectors. If we choose a prior on the continuous weights vector, $\boldsymbol{w_i}$, with the constraints that $0 < w_{ik} < 1$ and $\sum_k w_{ik} = 1$, then as $p(\boldsymbol{w_i})$ tends to delta functions at $w_{ik} = 0$ and $w_{ik} = 1$, then equation 9 will tend to equation 3. Therefore, to apply these constraints the prior we use is:

$$p(\boldsymbol{w}) = p(\boldsymbol{w}|\tilde{\boldsymbol{w}}, \gamma)p(\tilde{\boldsymbol{w}}) \tag{10}$$

where:

$$\begin{aligned} p(\tilde{\boldsymbol{w}}) &= \prod_{ik}^{N} p(\tilde{\boldsymbol{w}}_{ik}) \\ p(\tilde{\boldsymbol{w}}_{ik}) &= Uniform(\tilde{\boldsymbol{w}}_{\boldsymbol{ik}}; -\infty, +\infty) \end{aligned} \tag{11}$$

and $p(\boldsymbol{w}|\tilde{\boldsymbol{w}}, \gamma) = \prod_i p(\boldsymbol{w_i}|\tilde{\boldsymbol{w}}_{\boldsymbol{i}}, \gamma)$, where crucially $p(\boldsymbol{w_i}|\tilde{\boldsymbol{w}}_{\boldsymbol{i}}, \gamma)$ is a deterministic relationship by which $\boldsymbol{w_i}$ and $\tilde{\boldsymbol{w}}_{\boldsymbol{i}}$ are related by the logistic transform:

$$w_{ik} = \frac{\exp(\tilde{w}_{ik}/\gamma)}{\sum_{k=1}^{K} \exp(\tilde{w}_{ik}/\gamma))} \tag{12}$$

The normalising constant in the logistic transform $\sum_{k=1}^{K} exp(\tilde{w}_{ik}/\gamma)$ ensures that the condition $\sum_{k=1}^{K} w_{ik} = 1$ is met. This expression also ensures that $\tilde{w}_{ik} > \tilde{w}_{jk}$, if and only if $w_{ik} > w_{ij}$. Figure 1 shows how the logistic transform produces an approximation to the delta functions as γ gets smaller. We fix the value of γ to 0.05 whilst bounding $-10 < \tilde{w}_{ik} < 10$, this ensures that we get the desired approximation to delta functions at 0 and 1, whilst ensuring that we can compute $exp(\tilde{w}_{ik}/\gamma)$ without causing overflow.

To summarise, we now have two vectors of continuous weights at each voxel, w_{ik} and $\tilde{w}_{ik}$. $\tilde{w}_{ik}$ are weights which have a prior on them which is uniform on the real line. We then use the logistic transform to deterministically map the weights $\tilde{w}_{ik}$ to w_{ik} at each voxel. Then, w_{ik} are the continuous weights which represent approximations to the discrete labels with delta functions at 0 and 1.

2.2.2 Non-spatial with Class Proportions

We can approximate the distribution in equation 5 by replacing the discrete labels, x_i, with continuous weights vectors, $\boldsymbol{w_i}$:

$$p(\boldsymbol{w}, \boldsymbol{\theta}, \boldsymbol{\pi}|\boldsymbol{y}) \propto \prod_{i}^{N} \sum_{k=1}^{K} \{\pi_k w_{ik} p(y_i|x_i = k, \theta_k)\} p(\boldsymbol{w}) p(\boldsymbol{\theta}) p(\boldsymbol{\pi}) \tag{13}$$

where π_k are the global class proportion parameters defined in equation 4 and represent the proportion of each class in the mixture model, and $p(\boldsymbol{w})$ is given by equations 10–12. Note that when we infer on the posterior, $\boldsymbol{\pi}$ will depend upon $\boldsymbol{w}$ in same way that $\boldsymbol{\pi}$ depends on $\boldsymbol{x}$ in the discrete labels mixture model.

2.2.3 Spatial

We can approximate the distribution in equation 8 by replacing the discrete labels, x_i, with continuous weights vectors, $\boldsymbol{w_i}$:

$$p(\boldsymbol{w}, \boldsymbol{\theta}|\boldsymbol{y}) \propto \prod_{i}^{N} \sum_{k=1}^{K} \{w_{ik} p(y_i|x_i = k, \theta_k)\} p(\boldsymbol{w}) p(\boldsymbol{\theta}) \tag{14}$$

where $p(\boldsymbol{w})$, is given by equation 10 with $p(\boldsymbol{w}|\tilde{\boldsymbol{w}}, \gamma) = \prod_i p(\boldsymbol{w_i}|\tilde{\boldsymbol{w}}_{\boldsymbol{i}}, \gamma)$ where as before $p(\boldsymbol{w_i}|\tilde{\boldsymbol{w}}_{\boldsymbol{i}}, \gamma)$ is specified by a deterministic mapping between $\tilde{\boldsymbol{w}}$ and $\boldsymbol{w}$ (the logistic transform, equation 12).

However, instead of equation 11, $p(\tilde{\boldsymbol{w}})$ is now a continuous Gaussian conditionally specified auto-regressive (CAR) or continuous MRF prior (Cressie, 1993) on each of the K class maps, i.e. $p(\tilde{\boldsymbol{w}}) = \prod_k p(\tilde{\boldsymbol{w}}_{\boldsymbol{k}}|\phi_{\tilde{w}})$ (where $\tilde{\boldsymbol{w}} = \{\tilde{\boldsymbol{w}}_{\boldsymbol{k}} : k = 1 \ldots K\}$ and $\tilde{\boldsymbol{w}}_{\boldsymbol{k}} = \{\tilde{w}_{ik} : i = 1 \ldots N\}$) with:

$$p(\tilde{\boldsymbol{w}}_{\boldsymbol{k}}|\phi_{\tilde{w}}) \sim MVN(\tilde{\boldsymbol{w}}_{\boldsymbol{k}}; \mathbf{0}, (\mathbf{I} - \mathbf{C})^{-1}\mathbf{M}) \tag{15}$$

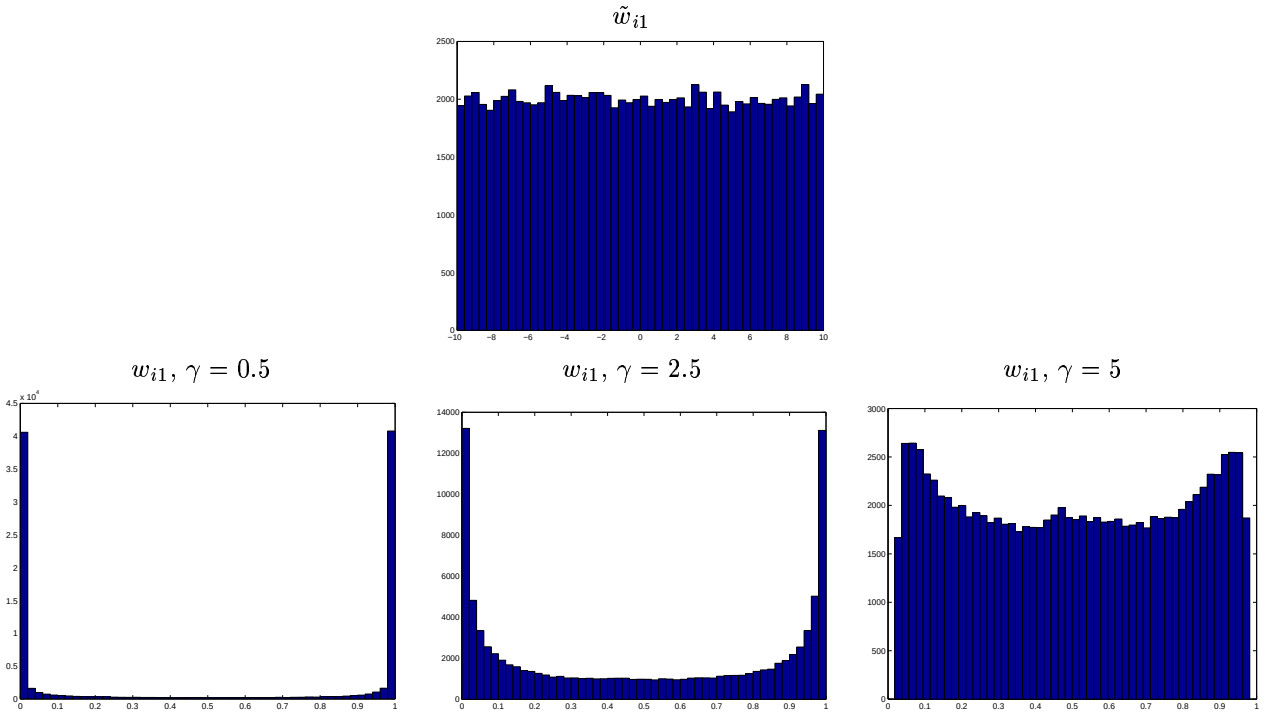

Figure 1: Consider that we have the number of classes as $K = 2$. [top] shows samples from the prior of $\tilde{w}_{i1}$, $\tilde{w}_{i1} \sim Uniform(-10, 10)$ (samples from $\tilde{w}_{i2}$ are similar). [bottom] shows the samples from w_{i1} (w_{i2} is similar), which the samples from the prior of $\tilde{w}_{i1}$ and $\tilde{w}_{i2}$ transform to under the logistic transform with different values of γ (equation 12). Hence, it can be seen how this produces a prior for $\boldsymbol{w_i}$, which approximates the desired delta functions at 0 and 1 as γ gets smaller.

where $\mathbf{C}$ is an $N \times N$ matrix whose (i,j)th element is c_{ij}, $\mathbf{M} = \frac{1}{\phi_{\tilde{w}}}\mathbf{I}$, and $\phi_{\tilde{w}}$ is the MRF control parameter which controls the amount of spatial regularisation. We set $c_{ii} = 0$, $c_{ij} = 1/N_{ij}$ if i and j are spatial neighbours and $c_{ij} = 0$ otherwise (where N_{ij} is the geometric mean of the number of neighbours for voxels i and j) , giving approximately:

$$p(\tilde{\boldsymbol{w}}|\phi_{\tilde{w}}) \propto f(\phi_{\tilde{w}}) \exp\left(\frac{-\phi_{\tilde{w}}}{4} \sum_i \sum_{j \in \mathcal{N}_i} (\tilde{w}_{ik} - \tilde{w}_{jk})^2\right) \tag{16}$$

How does this posterior approximation using the continuous class weights vectors instead of class labels allow us to adaptively determine the amount of spatial regularisation? The answer is that $\tilde{w}_{ik}$ is a continuous random variable ranging effectively between $-\infty$ and $+\infty$. Therefore, unlike $f(\phi_x)$ in equation 6, the normalising constant, $f(\phi_{\tilde{w}})$, in equation 16 is known:

$$f(\phi_{\tilde{w}}) \propto \frac{1}{\phi_{\tilde{w}}^N} \tag{17}$$

and hence we can adaptively determine the MRF control parameter, $\phi_{\tilde{w}}$. To achieve this $\phi_{\tilde{w}}$ becomes a parameter in the model, and equation 14 becomes:

$$p(\boldsymbol{w}, \boldsymbol{\theta}, \phi_{\tilde{w}}|\boldsymbol{y}) \propto \prod_i^N \sum_{k=1}^K \{w_{ik} p(y_i|x_i = k, \theta_k)\} p(\boldsymbol{w}|\phi_{\tilde{w}}) p(\phi_{\tilde{w}}) p(\boldsymbol{\theta}) \tag{18}$$

where the prior $p(\phi_{\tilde{w}})$, is a non-informative conjugate gamma prior:

$$\phi_{\tilde{w}}|\tilde{a}_{\tilde{w}}, \tilde{b}_{\tilde{w}} \sim Ga(\phi_{\tilde{w}}; \tilde{a}_{\tilde{w}}, \tilde{b}_{\tilde{w}}) \tag{19}$$

In summary, we have approximated the discrete labels with vectors of continuous weights which a priori approximate delta functions at 0 and 1. Each vector of continuous weights deterministically corresponds (via the logistic transform) to another vector of continuous weights, which a priori are uniform across the real line. As they are uniform across the real line these vectors of continuous weights can be regularised using a spatial prior (a continuous MRF) for which the amount of spatial regularisation can be determined adaptively.

Figure 2 shows a graphical representation of the discrete labels mixture model (equation 8) and the continuous weights spatial mixture model with adaptive spatial regularisation (equation 14).

3 Class Distributions

We need to specify $p(y_i|x_i = k, \theta_k)$ for all classes. This is application specific. In brain segmentation tasks the observations, y_i, might be intensity levels in a brain image, and the mixture might typically be made up of Gaussians (Zhang et al., 2001). In this paper we intend to use the mixture models proposed for classifying statistical parametric maps (SPMs) in the anlysis of FMRI data.

3.1 Classifying SPMs in FMRI

The mixture models could be used as a second stage to analyze the SPMs resulting from a first stage of a univariate temporal FMRI analysis (Woolrich et al., 2001). This is the same two-stage approach used when using the commonly used random field theory of Worsley et al. (1992) and is also the same approach that Everitt and Bullmore (1999); Hartvig (2000) take with the mixture models they consider.

The result of a univariate temporal analysis is effectively the parameter estimate, a_i, of the correlation with the assumed response at each voxel i. We might consider using a_i as the observations, y_i, for our mixture models. However, this carries over none of the uncertainty in the estimation of a_i from the univariate temporal analysis. Hence, we use instead the normalised version:

$$y_i = \frac{a_i}{std(a_i)} \tag{20}$$

See Woolrich et al. (2001) for one way of estimating a_i and $std(a_i)$ using a univariate temporal analysis.

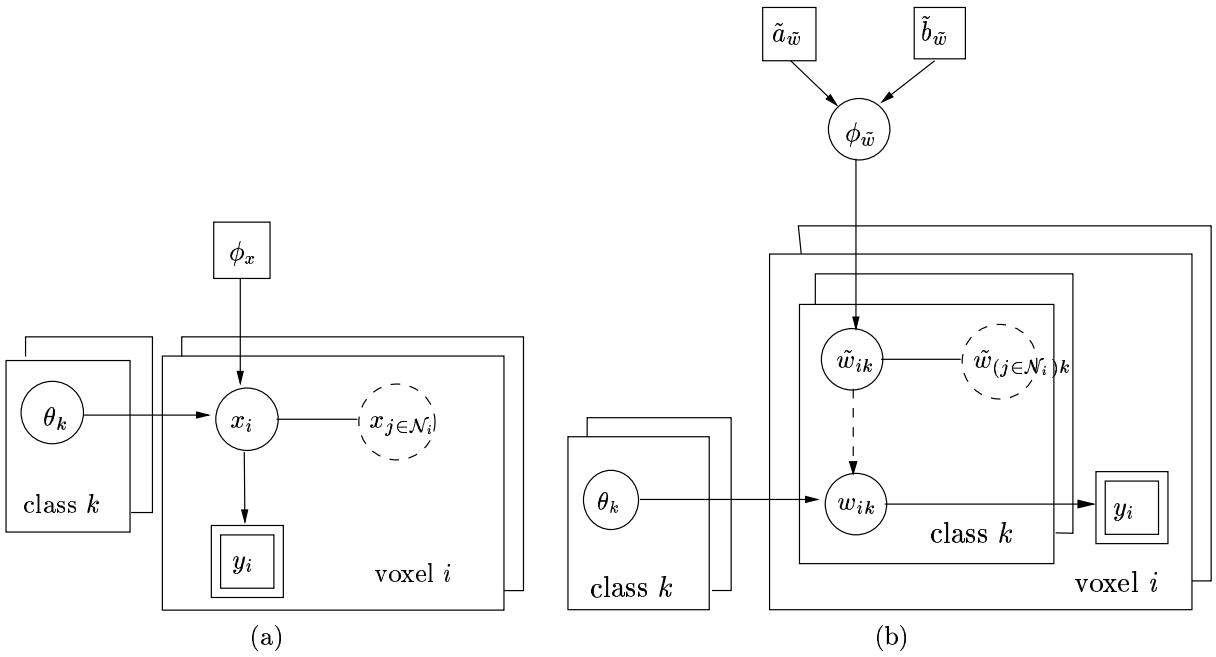

Figure 2: Graphical representation of (a) the discrete labels mixture model (equation 8), and (b) the continuous weights spatial mixture model with adaptive spatial regularisation (equation 14). Each parameter is a node in the graph and direct links correspond to direct dependencies. Solid links are probabilistic dependencies and dashed arrows show deterministic functional relationships. A rectangle denotes fixed quantities, a double rectangle indicates observed data, and circles represent all unknown quantities. Repetitive components are shown as stacked sheets. Dashed circles represent nodes that really correspond to a different stacked sheet, but which are shown on the top stacked sheet for ease of display.

We now need to specify $p(y_i|x_i = k, \theta_k)$ for all classes. For this we propose to use a similar approach to Hartvig and Jensen (2000), who uses three classes: activation, deactivation and non-activation. The non-activating class is modelled as a Normal distribution:

$$y_i|x_i = k_n, \theta_{k_n} \sim N(y_i; \mu_{k_n}, \sigma^2_{k_n}) \tag{21}$$

where k_n represents the label for the non-activating class. To reflect the assumption that the activating class can only have positive values of y_i, we use a Gamma function for the activation component:

$$y_i|x_i = k_a, \theta_{k_a} \sim Ga(y_i; \tilde{a}_{k_a}, \tilde{b}_{k_a}) \tag{22}$$

where k_a represents the label for the activating class. For the deactivating class, we also use a Gamma function:

$$y_i|x_i = k_d, \theta_{k_d} \sim Ga((-y_i); \tilde{a}_{k_d}, \tilde{b}_{k_d}) \tag{23}$$

where k_d represents the label for the deactivating class.

Note that we sometimes find it useful for the interpretation of parameters to reparameterise a Gamma distribution in terms of its mean, μ_k, and variance, σ^2_k. See the appendix for this transformation.

The hyperpriors $p(\boldsymbol{\theta})$ to be used on the component parameters are non-informative, disperse priors. However, we do place a restriction on the mode of the activation and deactivation Gamma classes. The mode of the activation class is constrained to be greater than the mean of the non-activation class, and the mode of the deactivation class is constrained to be less than the mean of the non-activation class. This encodes a sensible prior belief about the expected shape of the activation and deactivation distributions with respect to the non-activation. The mode of a Gamma distribution is given in equation 27.

4 Inference

To explore the differences that certain aspects of the models make, there are three different mixture models which we want to be able to infer upon in this paper, these are:

- Model 1. Non-spatial with global class proportions (equation 9)
- Model 2. Spatial without adaptive spatial regularisation with fixed spatial smoothness (equation 14 with fixed $\phi_{\tilde{w}} = 1$)
- Model 3. Spatial with adaptive spatial regularisation (equation 18)

To achieve this we need to obtain the marginal posterior distribution for the weights, w_{ik}, given the observed data, $\boldsymbol{y}$. Whilst it is possible to perform approximations to the distribution, it is difficult to assess the effect of these approximations on the inference performed. Therefore, the approach taken instead is to use Markov Chain Monte Carlo (MCMC) sampling from the full joint posterior distribution (see Gilks et al. (1996) and Gamerman (1997) for texts on MCMC). This also automatically provides us with samples from the marginal posterior distribution for the weights.

We are able to use Gibbs sampling for the adaptive MRF smoothness parameter, $\phi_{\tilde{w}}$. See the appendix for the required full conditional distribution of $\phi_{\tilde{w}}$.

For all other parameters (i.e. the continuous weights and class distribution parameters) we use single-component Metropolis-Hastings jumps (i.e. we propose separate jumps for each of the parameters in turn). The updates are detailed in the appendix. We use separate Normal proposal distributions for each parameter, with the mean fixed on the current value, and with a scale parameter σ_p for the p^{th} parameter that is updated every 30 jumps. At the j^{th} update σ_p is updated according to:

$$\sigma_p^{j+1} = \sigma_p^j S \frac{(1 + A + R)}{(1 + R)} \tag{24}$$

where A and R are the number of accepted and rejected jumps since the last σ_p update respectively, S is the desired rejection rate, which we fix at 0.5.

We require a good initialisation of the parameters in the model purely to reduce the required burn-in of the MCMC chains (the burn-in is the part of the MCMC chain which is used to ensure that the chain has converged

to be sampling from the true distribution). To initialise we use the non-spatial class labels with class proportions model (equation 5) along with the class distributions specified in section 3. The joint maximum a posterior over class labels, x_i, and distribution parameters, $\boldsymbol{\theta}$, can be obtained for this model using the Expectation-Maximisation (EM) algorithm (Beckmann et al., 2003).

As a result of this good initialisation, we use a burn-in of 1000 jumps, followed by 1000 further jumps of which every 2nd is sampled. Observation of the chains with different initial conditions confirmed that a burn-in of 1000 jumps was sufficient.

5 Artificial data

5.1 Methods

For all but the Gaussian activation artificial dataset (figure 6) we use artificial data generated by sampling from the mixture component distributions described in the last section, with the data deactivation parameters set to $\mu_{k_d} = -3$ and $\sigma^2_{k_d} = 3$, and the activation parameters set to $\mu_{k_a} = 4$ and $\sigma^2_{k_a} = 3$. For all artificial datasets (including the Gaussian activation artificial dataset), the non-activation parameters were $\mu_{k_n} = 0$ and $\sigma^2_{k_n} = 1$.

Unlike the other artificial datasets, the Gaussian activation artificial dataset (figure 6) is not generated fully from the model. This is because the activation and deactivation were not sampled from the mixture component distributions, but were instead modelled as 2-D spatial Gaussians added to non-activation. This dataset is included to see how the model deals with data not generated from the model. The 2-D spatial Gaussians used have a maximum peak of 4 for the activation and -3 for the deactivation, and a diagonal covariance matrix equal to $\sigma^2\mathbf{I}$, where $\sigma \approx 3.7 voxels$ for both activation and deactivation. These 2-D spatial Gaussians were added to samples from the non-activation distribution

In all cases we generate a 2-dimensional dataset with 100×100 pixels. The different spatial patterns of the five different artificial datasets are shown in the top right of figures 3—7.

The models we infer upon are the three models described in section 4. Model 2 is interesting as it is similar to those proposed in Salli et al. (1999); Zhang et al. (2001), in that the spatial smoothness is non-adaptive. Instead the MRF spatial smoothness parameter is arbitrarily fixed at a value of $\phi_{\tilde{w}} = 1$. This will act as a comparison for Model 3 where the spatial regularisation is adaptively determined.

5.2 Results

Figures 3—7 show the results of inferring on the three different continuous weights mixture models on the five different artificial datasets. The spatial maps in the figures are unthresholded marginal posterior means of w_{ik}, i.e. $E_{w_{ik}|y}(w_{ik})$, for all three classes of deactivation, non-activation and activation.

We can compute the effective global class proportions based upon the classifications, i.e.:

$$\tilde{\pi}_k = \frac{\sum_{i \in k} \bar{w}_{ik}}{\sum_{ik} \bar{w}_{ik}} \tag{25}$$

where the weights $\bar{w}_{ik}$ are the mean marginal posterior weights. These effective global class proportions can be combined with the mean marginal posterior class distribution parameters to give a histogram fit. These are shown in figures 3—7(a).

The box plot in figure 8 shows the marginal posterior distributions of the MRF smoothness parameter $\phi_{\tilde{w}}$ for the different artificial datasets. The value of $\phi_{\tilde{w}}$ clearly varies a lot from dataset to dataset emphasising the need for adaptive determination of the spatial smoothness.

It is interesting to compare model 2's performance with that of model 3. Model 3 is the same as model 2, except that in model 3 the spatial regularisation parameter is adaptive and in model 2 it is fixed. Model 2 works well on some datasets, for example figures 6 and 7. This is because the arbitrarily chosen value of $\phi_{\tilde{w}} = 1$ is close to the adaptively determined values of $\phi_{\tilde{w}}$ for those datasets, which can be seen in figure 8 (note that $log(1) = 0$). For the other datasets it works less well. For example, it overblurs the edges in figure 3 and overblurs the activation and deactivation to the point of removing much of it in figures 4 and 5.

It is worth emphasising that model 3 also works well on the no activation dataset (figure 7). This is a dataset where there is in fact only one class present (the non-activation class), and yet we fit the mixture model assuming

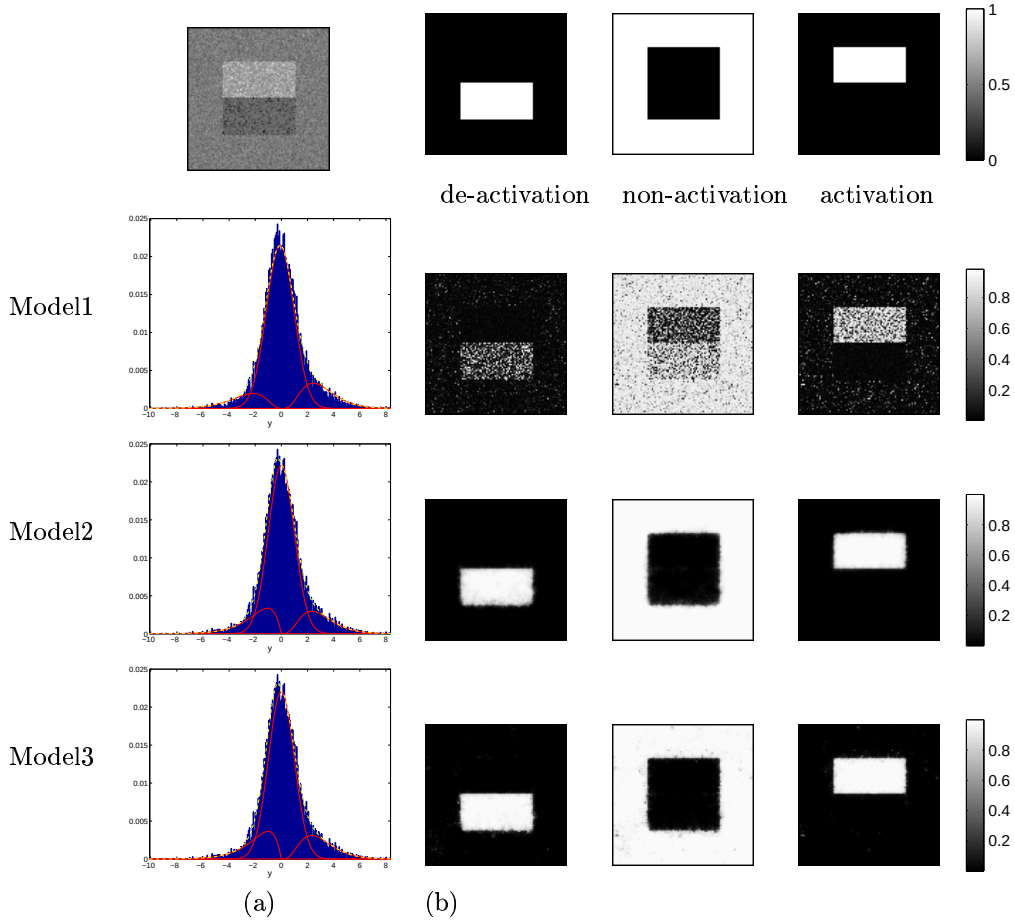

Figure 3: Results for the large activation artificial dataset. Top left is the actual data, y_i. (a) shows the histograms of y_i along with the fit for the different mixture models (the red lines show the individual class distributions and the dashed yellow line shows the overall fit). (b) Spatial maps of unthresholded w_{ik} for [left] deactivation [middle] non-activation [right] activation.

three classes. Despite this, model 3 forces all probabilities to well less than 0.5 for the activation and deactivation classes.

6 FMRI data

6.1 Methods

We use an audio-visual dataset taken using echo planar images (EPI) acquired using a 3 Tesla system with TR=3 seconds, time to echo (TE) = 30ms, in-plane resolution 4mm and slice thickness 7mm. The first 4 scans were removed and the data was motion corrected using MCFLIRT (Jenkinson et al., 2002) and high-pass filtered as described in Woolrich et al. (2001). The data was *not* spatially smoothed. The visual stimulus was a reversing checkerboard boxcar stimulus (30 seconds on, 30 seconds off). The auditory stimulus was also a boxcar stimulus (45 seconds on. 45 seconds off). We infer upon the same three models we used on the artificial datasets.

6.2 Results

Figures 9 and 10 show the results of inferring on the three different continuous weights mixture models we have described on the visual paradigm and audio paradigm SPMs respectively. The spatial maps in the figures are

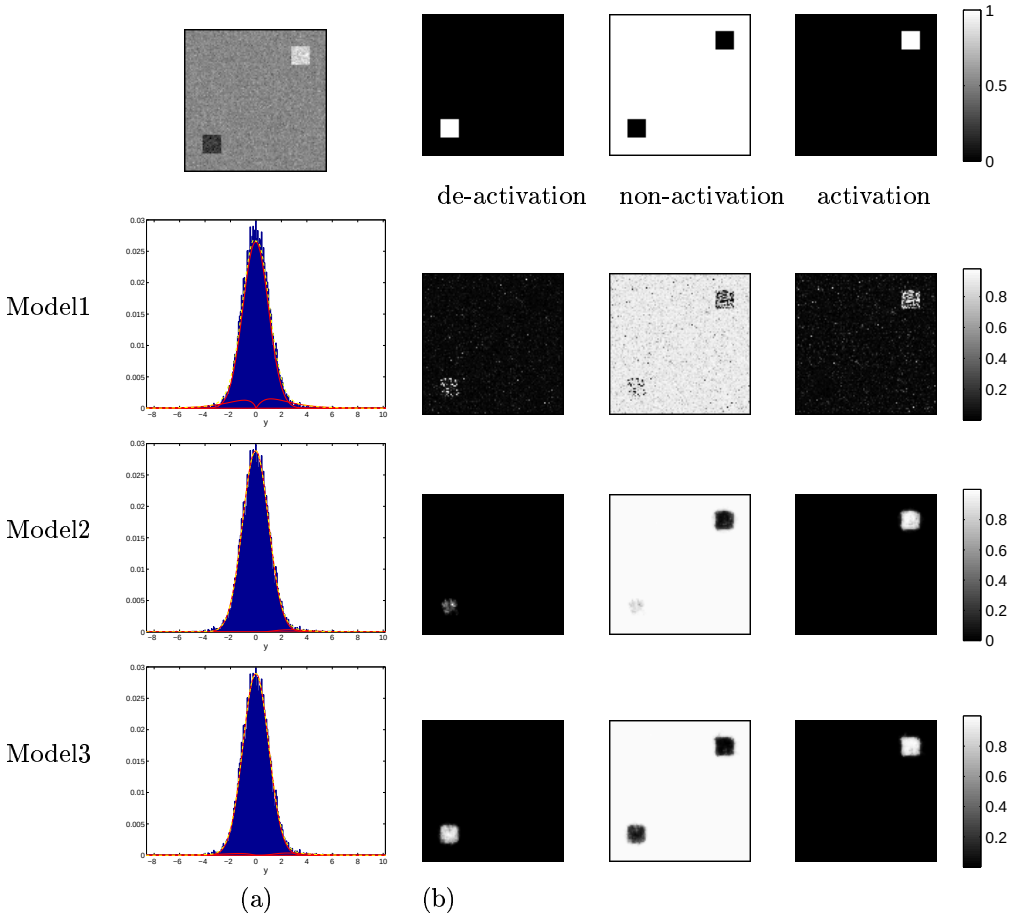

Figure 4: Results for the small activation artificial dataset. Top left is the actual data, y_i. (a) shows the histograms of y_i along with the fit for the different mixture models (the red lines show the individual class distributions and the dashed yellow line shows the overall fit). (b) Spatial maps of unthresholded w_{ik} for [left] deactivation [middle] non-activation [right] activation.

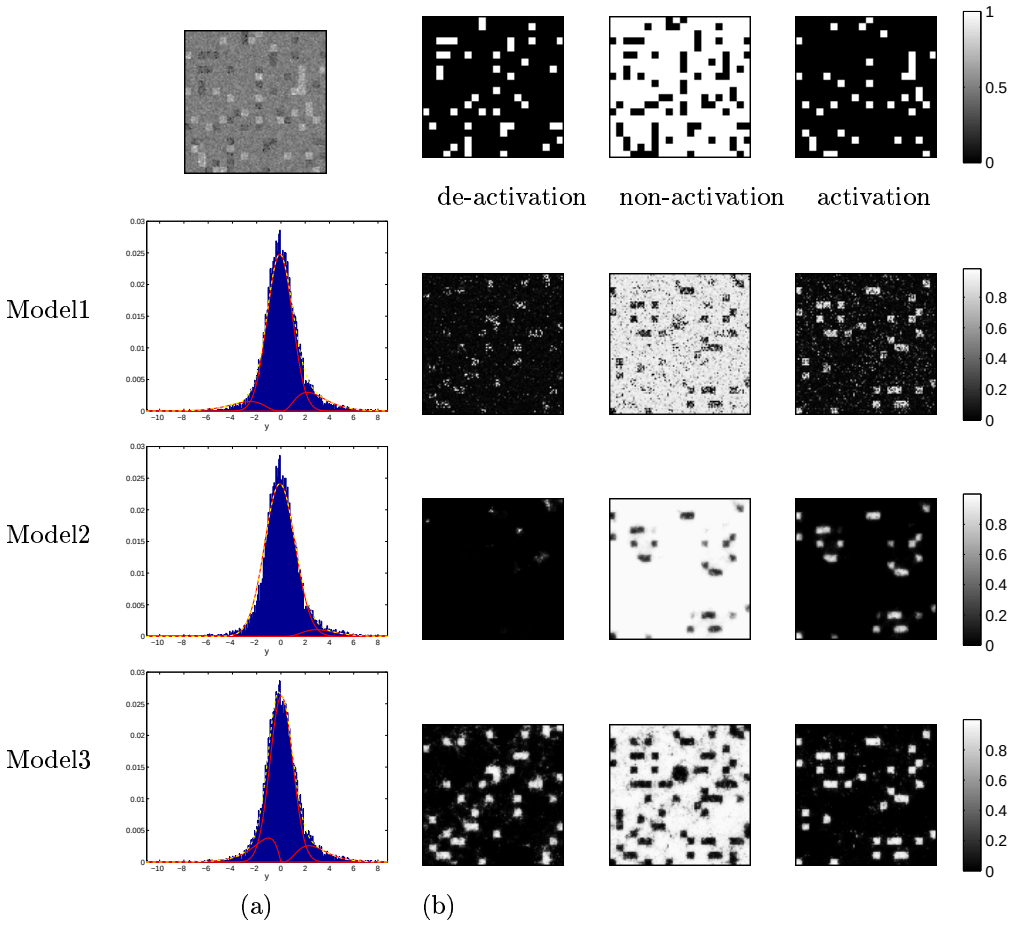

Figure 5: Results for the random checker activation artificial dataset. Top left is the actual data, y_i. (a) shows the histograms of y_i along with the fit for the different mixture models (the red lines show the individual class distributions and the dashed yellow line shows the overall fit). (b) Spatial maps of unthresholded w_{ik} for [left] deactivation [middle] non-activation [right] activation.

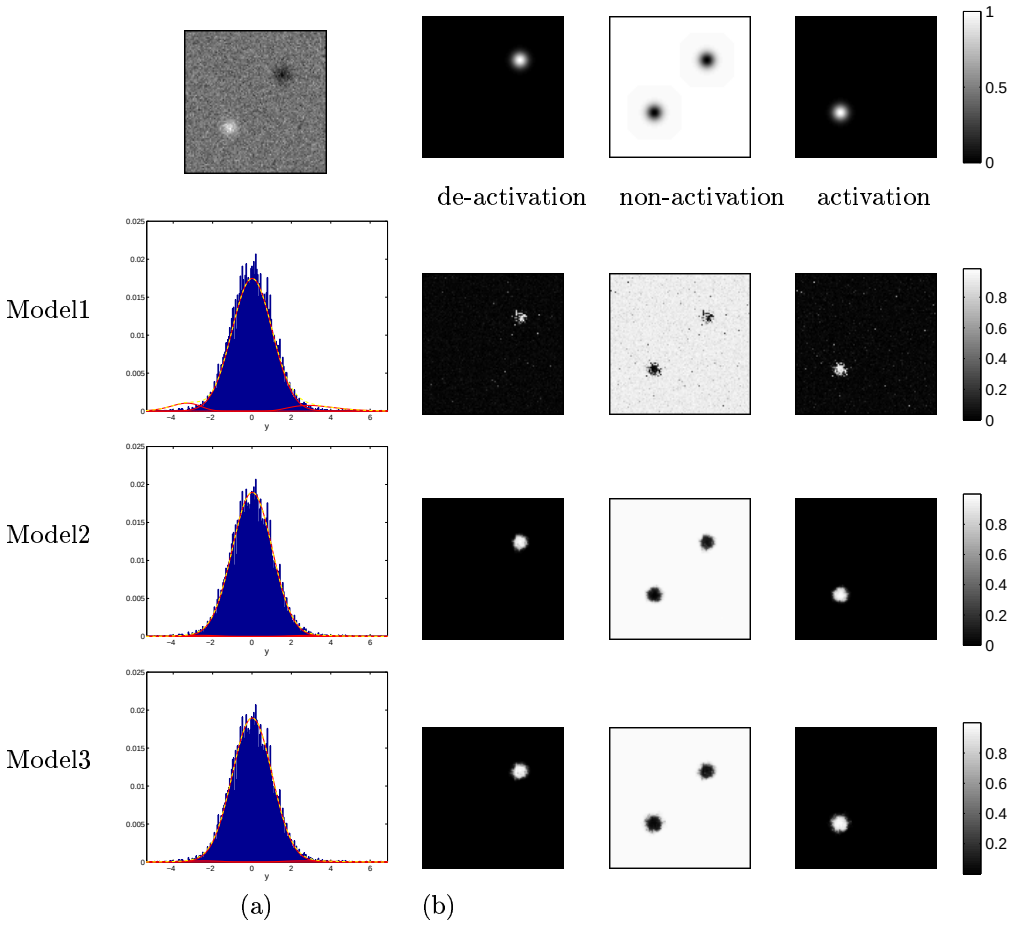

Figure 6: Results for the Gaussian activation artificial dataset. Top left is the actual data, y_i. (a) shows the histograms of y_i along with the fit for the different mixture models (the red lines show the individual class distributions and the dashed yellow line shows the overall fit). (b) Spatial maps of unthresholded w_{ik} for [left] deactivation [middle] non-activation [right] activation.

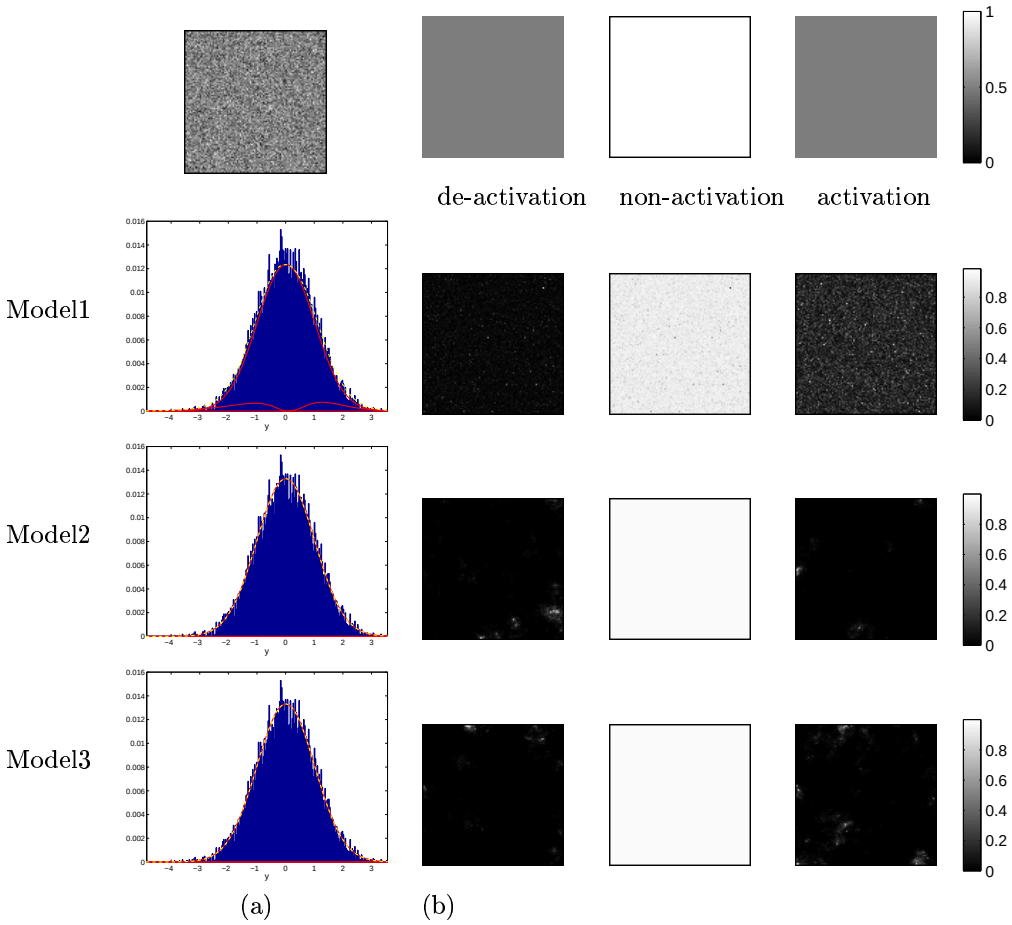

Figure 7: Results for the no activation artificial dataset. Top left is the actual data, y_i. (a) shows the histograms of y_i along with the fit for the different mixture models (the red lines show the individual class distributions and the dashed yellow line shows the overall fit). (b) Spatial maps of unthresholded w_{ik} for [left] deactivation [middle] non-activation [right] activation.

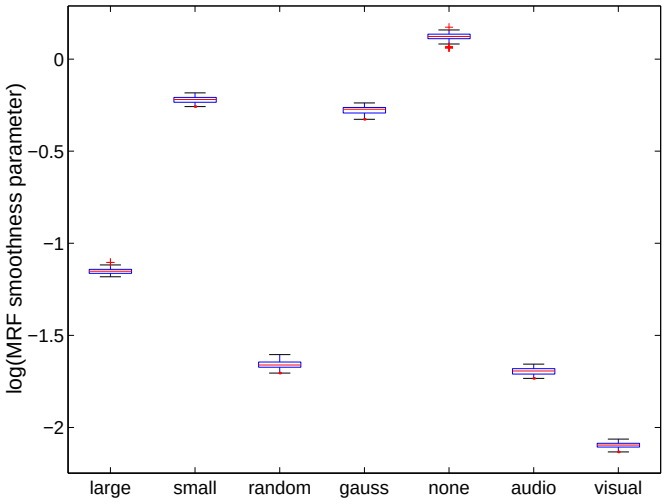

Figure 8: Box plot showing the marginal posterior distributions of the MRF smoothness parameter $\phi_{\tilde{w}}$ for the different datasets. The value of $\phi_{\tilde{w}}$ clearly varies a lot from dataset to dataset emphasising the need for adaptive determination of the spatial smoothness.

unthresholded marginal posterior means of w_{ik}, i.e. $E_{w_{ik}|y}(w_{ik})$ for all three classes of deactivation, non-activation and activation. Interestingly, the audio dataset shows a large amount of deactivation. If we qualitatively consider the data y shown in the top left of figure 10 then the spatial pattern of deactivation would seem to be strongly supported. Model 2, with the spatial smoothness parameter set to $\phi_{\tilde{w}} = 1$ imposes too much spatial smoothness. Indeed, if we look at the adaptively determined spatial smoothness in the box plot of figure 8, then the MRF smoothness parameter, $\phi_{\tilde{w}}$, for the visual and audio datasets is far less than the fixed value of $\phi_{\tilde{w}} = 1$ used for model 2.

Figures 11 and 12 show maps of the weights for the activation class, w_{ik_a}, thresholded to leave only those voxels with $p(w_{ik_a}|y) > 0.5$ for the visual paradigm and audio paradigm SPMs respectively. The choice of threshold on $p(w_{ik_a}|y)$ is, as with any thresholding, a decision that needs to be made by the experimenter. However, this is the *only* time that any value in the inference of the model has to be chosen. Indeed, even when we choose the threshold for $p(w_{ik_a}|y)$, there is a natural choice to make. That is we can choose the threshold of 0.5 which gives us an equal loss function where the chance of a false positive is equal to the chance of a false negative.

7 Discussion

It is worth considering the different sources of individual class broadening (variance) in the image intensity histogram. One source is intrinsic within-class variation in the underlying signal, for example, FMRI, where the activation strength does vary within an "activated" cluster. Secondly, broadening is caused by image noise, both high frequency noise and also, for example, in the case of structural brain images, low frequency noise or "bias field". A third cause is limitations of the model, for example, in the case of the models described in this paper, partial volume effect (PVE), where a voxel in reality contains a mixture of different classes. In Santago and Gage (1995); Ruan et al. (2000) spatial neighbourhood information is used to aid classification into pure voxels and PVE voxels. A very similar approach to PVE models is that of fuzzy mixture segmentation (Caillol et al., 1997). The approach in our paper, using the language of Caillol et al. (1997) is a "hard segmentation" in that the intention is to find posterior pdfs on discrete classification labels. The fuzzy approach of Caillol et al. (1997), like PVE modelling, consider voxels as being mixtures of classes rather than discretely belonging to one class or another. In PVE models (or fuzzy mixture models), the resulting models have weights vectors at each voxel to specify the mixture of the different classes. However, this should *not* be confused with the weights vectors used in this paper, which by virtue of the logistic transform are not producing mixtures of classes at a voxel but instead an *approximation* to discrete labels (to perform a hard segmentation).

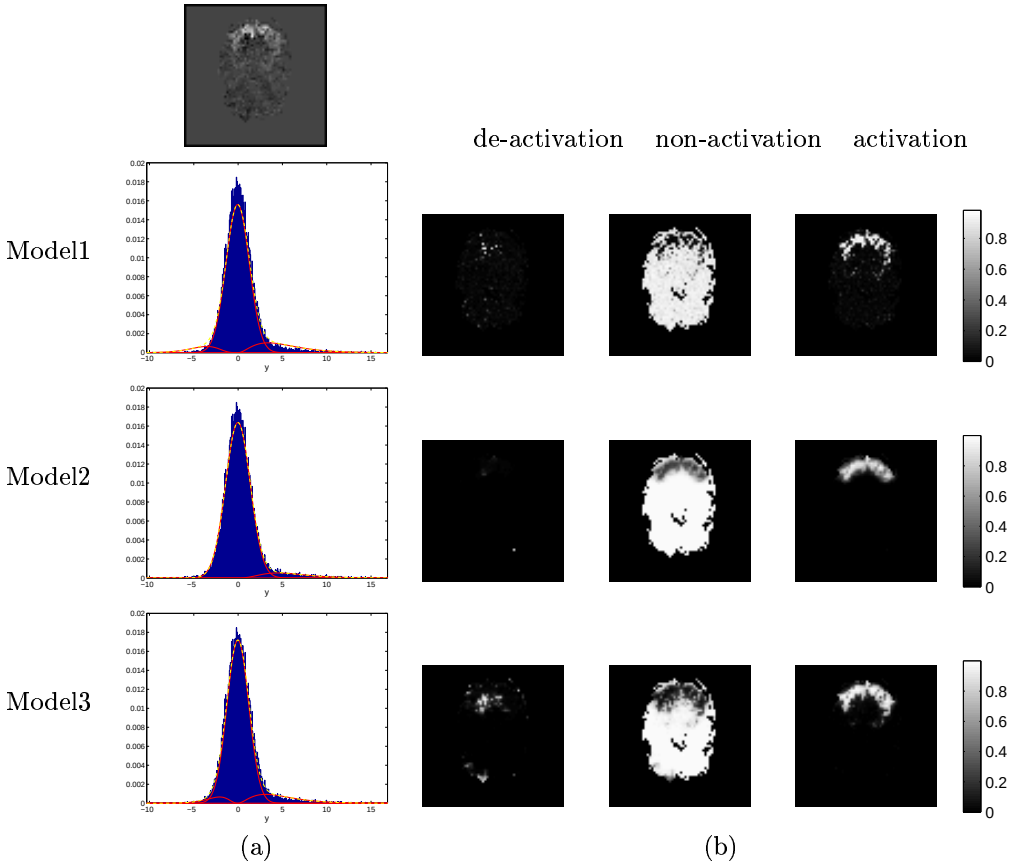

Figure 9: Results for the visual activation from the audio-visual dataset. Top left is the actual data, y_i. (a) shows the histograms of y_i along with the fit for the different mixture models (the red lines show the individual class distributions and the dashed yellow line shows the overall fit). (b) Spatial maps of unthresholded w_{ik} for [left] deactivation [middle] non-activation [right] activation.

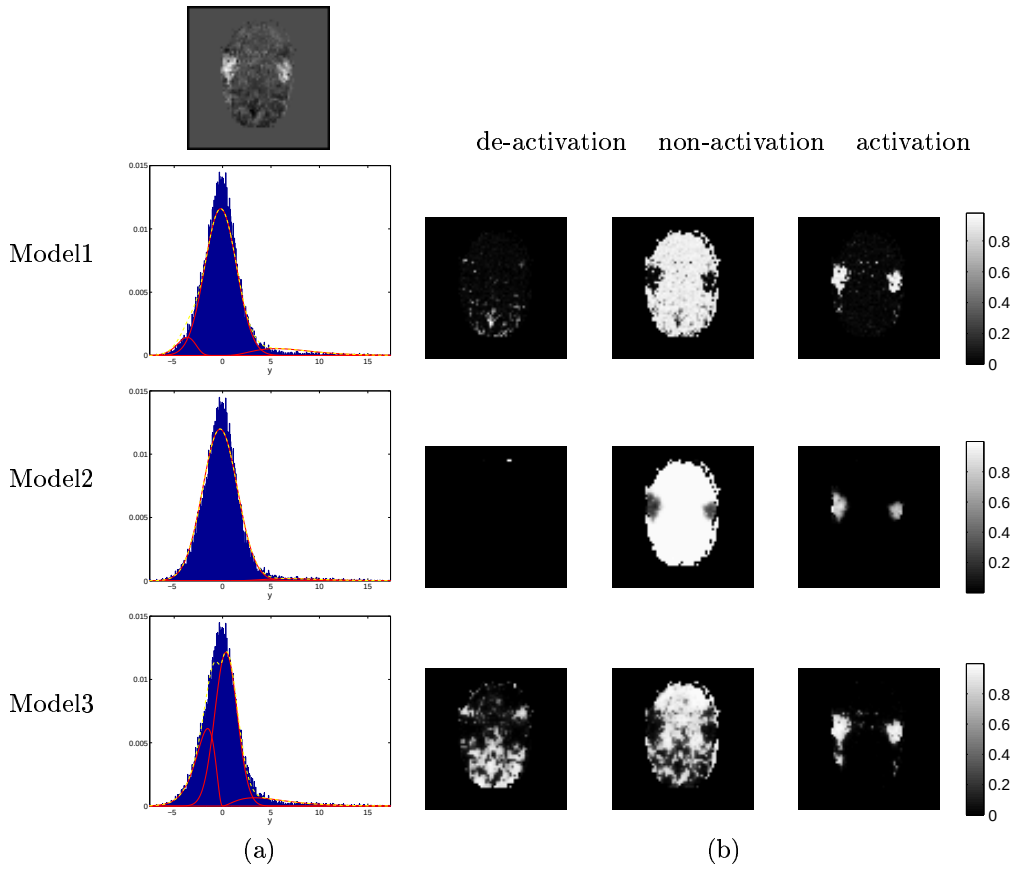

Figure 10: Results for the audio activation from the audio-visual dataset. Top left is the actual data, y_i. (a) shows the histograms of y_i along with the fit for the different mixture models (the red lines show the individual class distributions and the dashed yellow line shows the overall fit). (b) Spatial maps of unthresholded w_{ik} for [left] deactivation [middle] non-activation [right] activation.

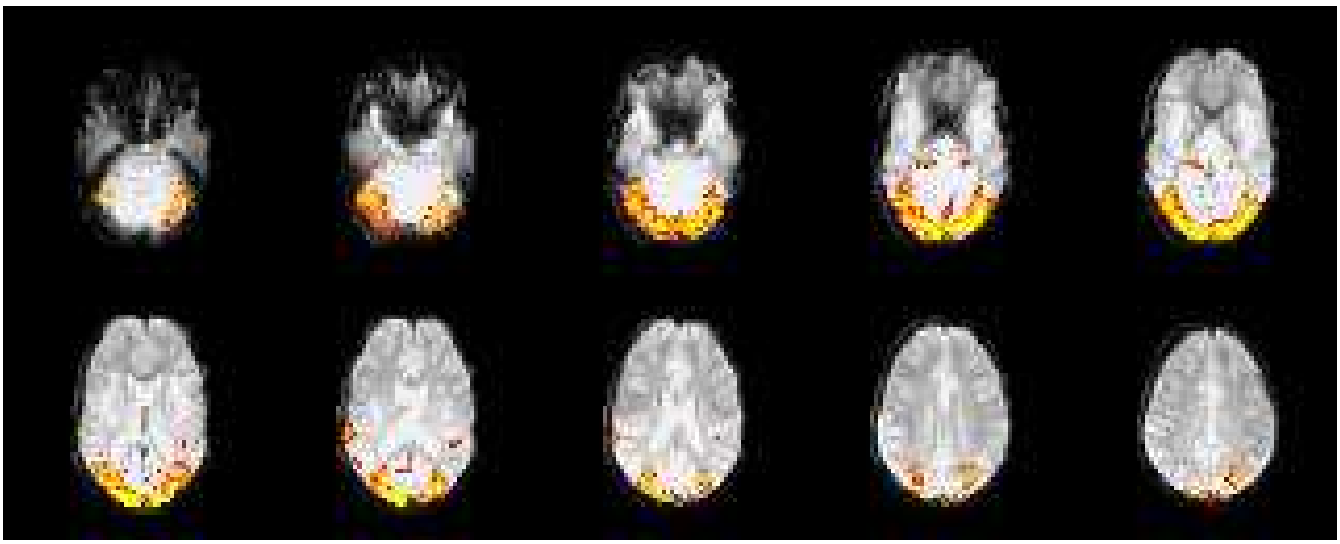

Figure 11: Map of w_{ik_a} thresholded to leave only those voxels with $p(w_{ik_a}|y) > 0.5$. This is for the visual activation from the audio-visual dataset using the SPM as the observations.

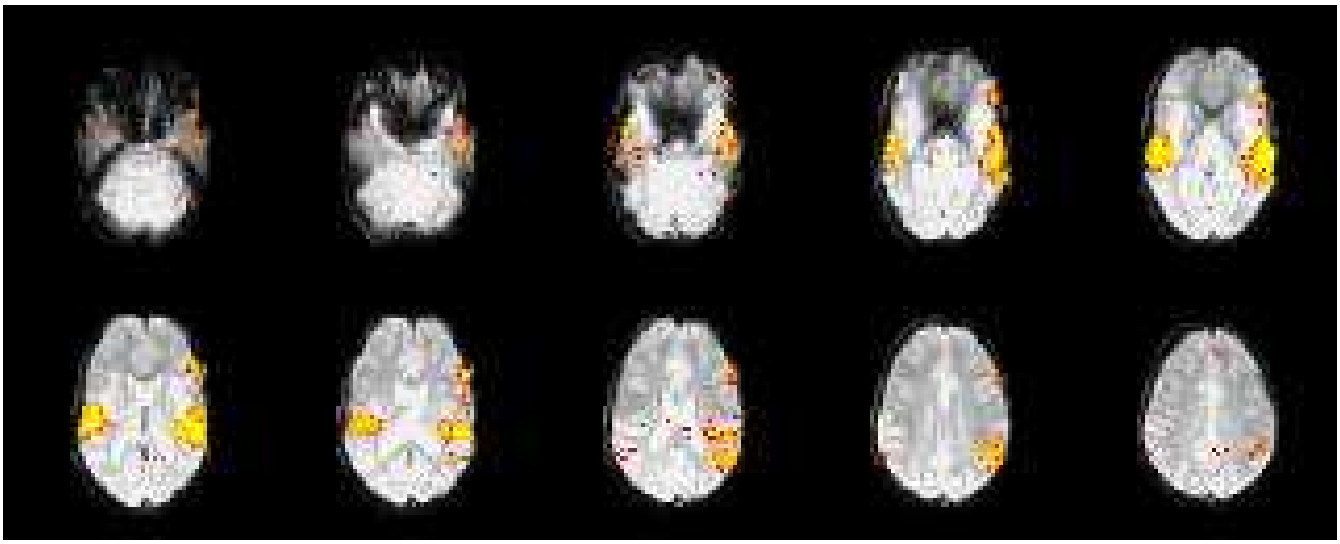

Figure 12: Map of w_{ik_a} thresholded to leave only those voxels with $p(w_{ik_a}|y) > 0.5$. This is for the audio activation from the audio-visual dataset using the SPM as the observations.

Nevertheless, it might be of interest to extend the models described in this paper to include PVE estimation. Note, however, that whilst this makes sense conceptually for situations where the original idealised class distribution is a delta-function (e.g., structural image segmentation), it is more complex for situations where real class intensities vary within-class (e.g., FMRI activation map classification). This is because there is now an ambiguity between varying PVE fractions and real within-class signal variations; in fact, when considering the underlying causes of variation in activation intensity in FMRI, the two issues (varying activation strength and PVE) may well be physically/physiologically the same thing, in which case it may make no sense to attempt separate PVE (or fuzzy) modelling.

Commonly in FMRI, people use null hypothesis testing to label voxels, or clusters of voxels, as being "active" if they reject the null hypothesis at a given false positive rate (FPR) (Friston et al., 1995). This depends on knowing the null distribution (or non-activation distribution) for relevant statistics under the null hypothesis.

In contrast, spatial mixture modelling provides us with a way of estimating the "activating" and "non-activating" distributions from the data itself. One advantage of estimating the "non-activating" distribution (in particular) from the data is that it allows some adjustment for modelling assumption violations. An example of the value of explicit separate models for background and activation distributions is that common problems in time-series modelling (such as imperfect autocorrelation modelling or the presence of structured noise) can cause the "null" distribution to be both shifted and scaled away from its assumed form, invalidating null-hypothesis testing. However, it is worth noting that the limitation of spatial mixture modelling for FMRI (or any other application) inference is the validity of the distributional assumptions. Whilst the adaptivity of the parameters of the Gaussian distribution we use to model the "non-activating" class can protect against small modelling assumption violations, this will not protect us against all modelling assumption violations.

There is also a well-known problem in null hypothesis testing of FMRI. The problem is that if enough observations are made, then every voxel in the brain will reject the null hypothesis (Friston et al., 2002). This is because in practice *no* voxels will show completely no reponse to the stimulus, if only due to modelling inadequacies such as unmodelled stimulus correlated motion or the point spread function of the scanner. By doing mixture modelling we can overcome this by instead of asking the question "Is the activation zero or not?", we ask the question "Is the activation bigger than the overall background level of "activation"?".

There is also the question of inference flexiblity. Because we have both the "activating" and "non-activating" distributions we can calculate the probability of a voxel being "activating" *and* the probability of a voxel being "non-activating". This provides us with far more inference flexiblity compared with null hypothesis testing. We can still look to control the FPR by thresholding using the probability of a voxel being "non-activating". But now we could also look to control the true postive rate (TPR) by thresholding using the probability of a voxel being

"activating". Controlling the TPR may be of real importance when using FMRI for pre-surgery planning. Usually, however, we would use a third option of comparing a voxel's probabilities of being in each of the classes. In this case there is a natural choice of threshold of assigning a voxel to the class with highest probability of membership.

There are other limitations of current FMRI null hypothesis techniques. A major issue is the incorporation of spatial information. The most popular method for this is via Gaussian Random Field Theory (Worsley et al., 1992; Friston et al., 1994). Gaussian Random Field Theory provides null distributions for clusters of voxels in SPMs. The problem is that to form clusters an arbitrary threshold has to be chosen (there is no objective way of setting this threshold, and its choice has a huge impact on the final thresholded results). Also, to meet the assumptions of Gaussian Random Field Theory, preprocessing spatial smoothing needs to be performed, the amount of which is another arbitrary choice. This is in stark contrast to the approach described in this paper for which there are *no* parameter choices to be made, apart from the final inference threshold.

In Fernandez and Green (2002) mixture models are used for disease mapping, in which the data is a count (the number of observed cases of disease) in each region $i = 1, \ldots, N$ that constitutes a geographical partition of the area of interest. Compellingly, they allow for the adaptive determination of the number of mixtures in the data via the use of reversible jumps MCMC sampling (Green, 1995). However, in tissue-type segmentation and FMRI activation classification the number of classes is normally known. Indeed, even in the absence of an assumed class, classifications were shown to be forced to zero using our method on artificial data when no activation or deactivation is supported by the data.

Fernandez and Green (2002) use a model related to our discrete labels non-spatial mixture model with class proportions (equation 5). However, they actually integrate out the discrete labels as they are not interested in classification. Instead, they are effectively looking to fit a different mixture at each voxel. This would mean allowing the class proportion parameters, $\boldsymbol{\pi}$, *and* the distribution parameters, $\boldsymbol{\theta}$, to vary locally. However, they restrict themselves to only allowing the class proportion parameters, $\boldsymbol{\pi}$, to vary locally (i.e. π_i). Sanjay-Gopal and Hebert (1998) and Marroquin et al. (2003) use similar models, but they have to hand-tune the parameter controlling the amount of spatial regularisation of these class proportion parameters. In contrast, Fernandez and Green (2002) adaptively determine the amount of spatial regularisation. To spatially regularise using a Gaussian MRF they need to map the local class proportion parameters π_i ($0 < \pi_{ik} < 1$, $\sum_k \pi_{ik} = 1$) to random variables which are uniform from $-\infty$ to $+\infty$. Hence, as we do in this paper, they also use the logistic transform. However, to maintain the interpretation of class proportion parameters they need to use a large γ in equation 12 (they require uniformity between 0 and 1 for π_i, whereas we required delta functions at 0 and 1 for $\boldsymbol{w_i}$, hence we used a small value of γ).

An interesting extension of the work in this paper would be to use Fernandez and Green (2002)'s non-stationary mixtures, whilst maintaining the spatially regularised classification of this paper. Non-stationary mixtures would allow local adaptation to spatial variations in the class distributions of intensity values such as can be observed particularly in structural brain MR images. The approach could have a non-stationary locally varying mixture (spatially regularised π_i and θ_i) with spatially regularised classification, $\boldsymbol{w_i}$, with the strengths of spatial regularisation determined adaptively and separately.

On a 2GHz Intel PC the technique takes approximately 30mins on a full volume. Future work needs to be done on alternative inference techniques to MCMC, which could speed up the inference on the model specified in this paper considerably. Broadly, there are two possiblities: (1) a point estimation (maximum a posterior) approach such as ICM or simulated annealing, (2) full posterior approximation approaches such as Laplace or Variational Bayes.

8 Conclusions

In this paper we have proposed a spatial mixture model which automatically determines the amount of spatial regularisation. This is achieved by using a Gaussian MRF prior on a vector of continuous weights, instead of using the standard approach of a discrete MRF prior on discrete labels, whilst approximating the same posterior. With the continuous Gaussian MRF prior we know the normalising constant. Subsequently, are able to automatically determine the continuous Gaussian MRF control parameter, allowing us to adaptively determine the amount of spatial regularisation. All parameters in the model are adaptively determined from the data and heuristic tuning of control parameters is no longer required.

We applied the mixture models to artificial data and demonstrated the usefulness and effectiveness of the adaptive determination of the amount of spatial regularisation. We also applied the mixture model to statistical

parametric maps in FMRI.

Acknowledgements The authors would like to acknowledge support from the UK MRC/EPSRC and GSK. Thanks must also go to Dr Mark Jenkinson, Dr Yongyue Zhang and Professor Sir Michael Brady for their valuable input.

9 Appendix

9.1 Gamma Distribution

x has a two-parameter gamma distribution, denoted by $Ga(a,b)$, with parameters a and b, if its density is given by:

$$f_{Ga}(x;a,b) = \frac{b^a}{\Gamma(a)} x^{a-1} e^{-bx} \tag{26}$$

where $\Gamma(a)$ is the Gamma function. Note, that a gamma distribution has $mean = a/b$ and $variance = a/b^2$. The mode of the Gamma distribution is given by:

$$mode = \frac{1}{b}(a-1) \tag{27}$$

9.2 MCMC

The three models we want to infer upon are described in section 4. As described in section 4 we infer the marginal posterior distribution for the weights, w_{ik}, by sampling from the full joint posterior distribution, i.e.: $p(\boldsymbol{w}, \boldsymbol{\theta}, \boldsymbol{\pi}|\boldsymbol{y})$ for model 1 (equation 9), $p(\boldsymbol{w}, \boldsymbol{\theta}|\boldsymbol{y})$ for model 2 (equation 14) or $p(\boldsymbol{w}, \boldsymbol{\theta}, \phi_{\tilde{w}}|\boldsymbol{y})$ for model 3 (equation 18).

To sample these full joint posterior distributions we use Markov Chain Monte Carlo (MCMC) techniques. See Gilks et al. (1996) or Gamerman (1997) for an introduction to MCMC sampling. A sample from the joint posterior distributions can be obtained by cycling through and sampling a subset of the parameters at a time. We now describe how we sample from each subset of parameters.

9.2.1 Control parameter, $\phi_{\tilde{w}}$

This is only required in model 3. $\phi_{\tilde{w}}$ has a full conditional which can be sampled from. Hence for $\phi_{\tilde{w}}$ we can employ Gibbs sampling. The full conditional for $\phi_{\tilde{w}}$ is given by:

$$\phi_{\tilde{w}}|. \quad \sim \quad Ga\left(\frac{N}{2} + \tilde{a}_{\tilde{w}}, \frac{1}{4}\sum_i \sum_{j \in \mathcal{N}_i} (\tilde{w}_{pi} - \tilde{w}_{pj})^2 + \tilde{b}_{\tilde{w}}\right) \tag{28}$$

$\tilde{a}_{\tilde{w}}$ is set to 10^{-4}, and $\tilde{b}_{\tilde{w}}$ is set to 10^{-4} to give a disperse prior.

9.2.2 Class distribution parameters, $\boldsymbol{\theta}$

This is sampled from using Metropolis-Hastings. With Metropolis-Hastings, a parameter change is proposed and then accepted or rejected according to the standard Metropolis-Hastings rule. This requires that we recalculate the terms in the joint posterior that change when we change $\boldsymbol{\theta}$. These terms are p($\boldsymbol{\theta}$) for all three models plus for model 1:

$$\prod_i^N \sum_{k=1}^K \{\pi_k w_{ik} p(y_i|x_i = k, \theta_k)\} \tag{29}$$

or for models 2 and 3:

$$\prod_i^N \sum_{k=1}^K \{w_{ik} p(y_i|x_i = k, \theta_k)\} \tag{30}$$

9.2.3 Global class proportion parameters, $\boldsymbol{\pi}$

This is only required for model 1. This is sampled from using Metropolis-Hastings. The following terms are recalculated for each parameter change:

$$\prod_i^N \sum_{k=1}^K \{\pi_k w_{ik} p(y_i|x_i = k, \theta_k)\}$$
$$p(\boldsymbol{\pi}) \tag{31}$$

9.2.4 Continuous weights parameters, $\boldsymbol{w_i}$

This is sampled from using Metropolis-Hastings. We loop through all voxels sampling from the continuous weights vector a voxel at a time. Recall that $p(\boldsymbol{w_i}|\tilde{\boldsymbol{w}}_{\boldsymbol{i}}, \gamma)$ is the deterministic logistic transform in equation 12 at voxel i. The logistic transform can only take us from a vector $\tilde{\boldsymbol{w}}_{\boldsymbol{i}}$ to a vector $\boldsymbol{w_i}$. Hence, in practice we propose jumps on $\tilde{\boldsymbol{w}}_{\boldsymbol{i}}$ and then calculate the vector $\boldsymbol{w_i}$ this gives us. Then, for model 1 we need to recalculate:

$$\sum_{k=1}^{K}\{\pi_k w_{ik} p(y_i|x_i = k, \theta_k)\} \tag{32}$$

or for models 2 and 3:

$$\sum_{k=1}^{K}\{w_{ik} p(y_i|x_i = k, \theta_k)\} \tag{33}$$

Plus for models 2 and 3 we also need to update the parts of the MRF prior that change when we change $\boldsymbol{w_i}$. These are:

$$\prod_k exp\left(-\frac{\phi_{\tilde{w}}}{2}\sum_{j\in\mathcal{N}_i}(\tilde{w}_{ik} - \tilde{w}_{jk})^2\right) \tag{34}$$

where $\mathcal{N}_i$ is the set of voxels in the spatial neighbourhood of voxel i.

References

Beckmann, C., Woolrich, M., and Smith, S. (2003). Gaussian / Gamma mixture modelling of ICA/GLM spatial maps. In *Ninth Int. Conf. on Functional Mapping of the Human Brain*.

Besag, J. (1986). On the statistical analysis of dirty pictures. *Journal of the Royal Statistical Society*, 3:259–302.

Caillol, H., Pieczynski, W., and Hillion, A. (1997). Estimation of Fuzzy Gaussian Mixture and Unsupervised Statistical Image Segmentation. *IEEE Trans. on Medical Imaging*, 6(3):425–440.

Cressie, N. (1993). *Statistics for Spatial Data*. Wiley, New York.

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

Fernandez, C. and Green, P. (2002). Modelling spatially correlated data via mixtures: a bayesian approach. *Journal of the Royal Statistical Society Series B*, 64(4):805–826.

Friston, K., Holmes, A., Worsley, K., Poline, J.-B., Frith, C., and Frackowiak, R. (1995). Statistical parametric maps in functional imaging: A general linear approach. *Human Brain Mapping*, 2:189–210.

Friston, K., Worsley, K., Frackowiak, R., Mazziotta, J., and Evans, A. (1994). Assessing the significance of focal activations using their spatial extent. *Human Brain Mapping*, 1:214–220.

Friston, K. J., Penny, W., Phillips, C., Kiebel, S., Hinton, G., and Ashburner, J. (2002). Classical and Bayesian inference in neuroimaging: Theory. *NeuroImage*, 16:465–483.

Gamerman, D. (1997). *Markov Chain Monte Carlo*. Chapman and Hall, London.

Geman, S. and Geman, D. (1984). Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 6(6):721–741.

Gilks, W., Richardson, S., and Spiegalhalter, D. (1996). *Markov Chain Monte Carlo in Practice*. Chapman and Hall, London.

Green, P. (1995). Reversible jump Markov Chain Monte Carlo computation and bayesian model determination. *Biometrika*, 82:711–732.

Guillemaud, R. and Brady, M. (1997). Estimating the bias field of MR images. *IEEE Trans. on Medical Imaging*, 16(3):238–251.

Hartvig, N. (2000). A stochastic geometry model for fMRI data. Technical Report 410, Department of Theoretical Statistics, University of Aarhus.

Hartvig, N. and Jensen, J. (2000). Spatial mixture modelling of fMRI data. *Human Brain Mapping*, 11(4):233–248.

Held, K., Kops, E., Krause, B., Wells, W., Kikinis, R., and Müller-Gärtner, H.-W. (1997). Markov random field segmentation of brain MR images. *IEEE Trans. on Medical Imaging*, 16:878–886.

Jenkinson, M., Bannister, P., Brady, J., and Smith, S. (2002). Improved optimisation for the robust and accurate linear registration and motion correction of brain images. *NeuroImage*, 17(2):825–841.

Marroquin, J., Arce, E., and Botello, S. (2003). Hidden markov measure field models for image segmentation. Accepted in special edition of IEEE TPAMI on Energy Minimization Methods in Computer Vision and Pattern Recognition. *IEEE Trans. on Pattern Analysis and Machine Intelligence*.

Ruan, S., Jaggi, C., Xue, J., Fadili, J., and Bloyet, D. (2000). Brain tissue classification of magnetic resonance images using partial volume modeling. *IEEE Trans. on Medical Imaging*, 19(12):1179–1187.

Salli, E., Visa, A., Aronen, H., Korvenoja, A., and Katila, T. (1999). Statistical segmentation of FMRI activations using contextual clustering. In *Medical Image Computing and Computer-Assisted Intervention*, pages 481–488.

Sanjay-Gopal, S. and Hebert, T. (1998). Bayesian pixel classification using spatially variant finite mixtures and the generalized EM algorithm. *IEEE Trans. on Image Processing*, 7(7):1014–1028.

Santago, P. and Gage, H. (1995). Statistical models of partial volume effect. *IEEE Trans. on Medical Imaging*, 11(4):1531–1540.

Svensén, M., Kruggel, F., and von Cramon, D. (2000). Probabilistic modeling of single-trial fmri data. *IEEE Trans. on Medical Imaging*, 19(1).

Wells, W., Grimson, W., Kikinis, R., and Jolesz, F. (1996). Adaptive segmentation of MRI data. *IEEE Trans. on Medical Imaging*, 15(4):429–442.

Woolrich, M., Ripley, B., Brady, J., and Smith, S. (2001). Temporal autocorrelation in univariate linear modelling of FMRI data. *NeuroImage*, 14(6):1370–1386.

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

Zhang, Y., Brady, M., and Smith, S. (2001). Segmentation of brain MR images through a hidden Markov random field model and the expectation maximization algorithm. *IEEE Trans. on Medical Imaging*, 20(1):45–57.