Mixtures of General Linear Models for Functional Neuroimaging

Will Penny, Karl Friston

Abstract— **We set out a new general framework for making inferences from neuroimaging data, which includes a standard approach to neuroimaging analysis, Statistical Parametric Mapping (SPM), as a special case. The model offers numerous conceptual and statistical advantages that derive from analysing data at the 'cluster-level' rather than the 'voxel-level' and from explicit modelling of the shape and position of clusters of activation. This provides a natural and principled way to pool data from nearby voxels for parameter and variance-component estimation. The model can also be viewed as performing a spatio-temporal cluster analysis. The parameters of the model are estimated using an Expectation Maximisation (EM) algorithm.**

I. Introduction

We propose a new approach to the analysis of functional neuroimaging data [8]. The approach is based on a family of models called Mixtures of General Linear Models (MGLMs) which include a standard approach to neuroimaging data analysis, Statistical Parametric Mapping (SPM) [10], as a special case. The central tenet of these models is that the fundamental quantities of interest to the neuroimager are the location, shape and temporal signature of *clusters* of voxels showing task-related activity. In these models data are analysed at the 'cluster-level'. This is to be contrasted with established methodologies in which data are analysed at the 'voxel-level'.

Our work is inspired by the notion of 'borrowing strength', described by Genovese as follows [11]. The shape and magnitude of the hemodynamic response and the impact of physiological variations tend to be consistent across localised groups of voxels. These localised groups represent regions with common physiological and/or functional properties. These consistencies induce dependencies among the model parameters associated with different voxels. By identifying these 'dependence neighbourhoods' we can borrow strength in estimating the model parameters. That is, we use data from multiple voxels to estimate common parameters. Genovese suggests that these neighbourhoods are best identified using an adaptive partitioning of the data based on the temporal signal at each voxel. This idea is readily captured using a Markov Random Field (MRF) model and a number of MRF approaches have appeared in the literature. Descombes *et al.* [7], for example, use a spatio-temporal MRF in which both the spatial and temporal smoothness of the hemodynamic response are modelled. Svensen *et al.* [25] use a spatial MRF to cluster images into regions with homogeneous responses and Rajapakse and Piyaratna [22] use a spatial MRF to cluster maps of statistical parameters.

Both authors are with the Wellcome Department of Imaging Neuroscience, University College, London WC1N 3BG. {wpenny, karl}@fil.ion.ucl.ac.uk

The MGLM approach sits outside the MRF framework. The most fundamental difference is that MGLMs model explicitly the positions and shapes of activated regions. This offers the potential of making formal inferences about the location, size and spatial variability of responses. The MGLM approach may also be viewed as a spatio-temporal clustering algorithm and as such, generalises existing cluster-based methods for analysing functional data (see eg. [4]).

In section 2.A we describe Statistical Parametric Mapping, a dominant paradigm for analysing functional imaging data. In section 2.B we describe the generative process underlying MGLM and use it to generate 'fMRI-like' time series. In section 2.C, we show how the parameters of an MGLM can be estimated from real fMRI time series and how inferences about clusters of activation are made. We then describe how the models are initialised and, in section 3, apply the models to two fMRI data sets, a block-design paradigm and an event-related paradigm [17]. The results are compared with those from SPM.

II. Materials and Methods

A. Statistical Parametric Mapping

Statistical Parametric Mapping (SPM) [10] has been adopted by a large contingent of the neuroimaging community and in this sense may be viewed as a standard approach to neuroimaging analysis. SPM is based on a General Linear Model (GLM) operating at each voxel in a functional image. This is termed a 'mass-univariate' approach. This model consists of a design matrix, common to all voxels, and a set of parameter estimates that are voxel-specific. The design matrix contains information about the activation paradigm and possible confounding variables. The parameter estimates indicate the strength of the activations and confounds at each voxel. After basic pre-processing, data are spatially smoothed and GLMs are fitted to each voxel. To detect voxels which are significantly active, a t-statistic is then computed for each voxel. However, because there are so many voxels it is likely that some will appear active by chance. To account for this a correction for multiple comparisons, based on Gaussian Random Field (GRF) theory, is then made. The product of this analysis is a map of a t-statistic showing which voxels are significantly active. Two such maps are shown in Figures 4 and 8. Notice that although the analysis has proceeded at the voxel-level, the end result is a map containing a small number of blobs which constitute *clusters* of voxels showing task-related activity. It is this structure that is exploited in the MGLM model.

B. Generative Model

A key feature of MGLMs is that they are based on a 'generative model'. The model consists of active components and null components. Active components define spatially-localised clusters of activity that are temporally correlated with the activation paradigm and null components define spatially distributed background activity which is temporally uncorrelated with the paradigm. Active components are characterised spatially by a Gaussian with a mean defining the centre of the cluster and a covariance defining its shape and width. Examples of active components are given in Figures 5,6 and 9. They are temporally characterised by a GLM, defining the activation and possible confounds. Time series defined by GLMs are shown as the solid lines in Figures 7 and 10. In this paper we have a single null component, although generally this need not be the case. It is defined spatially by a uniform distribution and temporally by a Gaussian process, with mean and variance that do not vary over time.

The model for how the time series are generated is as follows (see Figure 1). At each voxel, at each time point, a probabilistic decision is made as to which component to draw a sample from. This decision is based on a spatial prior - components nearer to that voxel are more likely to be chosen. A sample is then drawn from the GLM corresponding to the chosen component. In this way, voxel time series consist of a *mixture* of samples from different GLMs at different time points. This mixing process couples the spatial and temporal domains - voxels at the very edge of a 'signal' [active] component nearly always draw 'noise' [null] samples, but occasionally draw signal samples. More signal samples are drawn as we get closer to the local activation centre. In this way, the overall correlation of voxels with the activation paradigm can vary smoothly over the image - as observed empirically. A sample of images from such a generative model is shown in Figures 2 and 3. The spatio-temporal model underlying this process is separable in the sense that the spatial prior is the same at all time points.

Mathematically, the fundamental assumption of our model is that the likelihood of an observation at the ith voxel and the tth time point, y_i^t, is given by the mixture model

$$p(y_i^t|v_i,\theta) = \sum_{k=1}^{K} p(y_i^t|k,\theta)p(k|v_i,\theta) \quad (1)$$

where the spatial location of the ith voxel is $v_i = [x_i, y_i, z_i]$ and the parameters of the model (introduced below) are collectively written as θ. Note that this is a conditional probability, with the fundamental dependence being on spatial location. The first factor on the right of equation 1 is the probabilistic prediction from the kth component and the second factor is the prior probability. The generative model is shown graphically in Figure 1.

In what follows the notation $\mathsf{N}_d(\mu, \Sigma)$ denotes a d-variate Gaussian distribution with mean μ and covariance Σ.

The spatial prior is specified by the likelihood ratio

$$p(k|v_i,\theta) = \frac{p(v_i|k)}{\sum_{k'} p(v_i|k')} \quad (2)$$

where

$$p(v_i|k) = \mathsf{N}_3(m_k, \Sigma_k) \quad (3)$$

and $m_k = [x_k, y_k, z_k]$ is the spatial location of the cluster and Σ_k is its spatial covariance. This says that the probability that voxel v_i belongs to cluster k falls as a Gaussian function of distance from the cluster's centre. The ensuing probability of sampling from k given voxel v_i is this renormalised 'belonging' probability.

In the experiments in this paper we have a single null component with a uniform spatial prior $p(v_i|k) = 1/V$ where V is the number of voxels in the image. Importantly, this means that voxels are by default assigned to the non-activating class. That is, if $p(v_i|k)$ falls below $1/V$ for all of the active components then the voxel is *a priori* assigned to the null component.

For an active component we have

$$p(y_i^t|k,\theta) = \mathsf{N}_1(\hat{y}_k^t, \sigma_k^2) \quad (4)$$

where $\hat{y}_k^t$ is the prediction from the kth GLM at time t. If we let $\hat{Y}_k = [\hat{y}_k^1, \hat{y}_k^2, ..., \hat{y}_k^N]^T$ where N is the number of time-points then

$$\hat{Y}_k = X_k w_k \quad (5)$$

where X_k is the 'design matrix' and w_k are the regression coefficients. This is the same as the usual GLM model used in SPM. The design matrix contains, for example, details of when the various experimental stimuli were given and information about possible confounds (see eg. [8] for more details). For a null component we have

$$p(y_t|k,\theta) = \mathsf{N}_1(\mu_k, \sigma_k^2) \quad (6)$$

where μ_k is the average activity and σ_k^2 is the temporal variance. This can be viewed as a GLM with a single column of 1's in the design matrix and $w_k = \mu_k$.

The parameters of the overall MGLM model are $\theta = \{X_k, w_k, \sigma_k^2, m_k, \Sigma_k\}$. We again stress that we are not analysing the data at the voxel level. That is, we do not have a separate GLM model for each voxel - we have a single GLM model for all voxels in cluster k, and information from all of these voxels is used to estimate the parameters w_k, σ_k^2, m_k and Σ_k (ie. we are borrowing strength).

In this paper we consider the design matrix X_k to be known and to be the same for all k (except for the null component). In the limit that each voxel comprises a cluster $K \rightarrow V$, we then recover the voxel-wise GLM approach that underlies SPM. SPM is, in this sense, a limiting case of the MGLM model.

Figures 2 and 3 show data generated from an MGLM model with a null component ($k = 1$) and two active components ($k = 2, 3$). The active components are Gaussian in shape (see Fig 2) and have a temporal activation given by the regressor at the top of Fig 3. This consists of a boxcar which has been passed through a 'canonical Hemodynamic Response Function' (HRF) comprising two overlaid gamma functions [10]. This captures the temporal aspects of the HRF. In [20], it is surmised that the variability of the magnitude of the HRF from voxel to voxel arises because of differences in the diameter of the local vasculature or the proximity of the voxel to neurally active tissue. This latter component is captured by the Gaussian nature of our spatial prior. The overall MGLM model is therefore capable of capturing both the temporal and spatial aspects of the HRF.

An important feature of MGLM models is that they have far fewer parameters than mass-univariate models. Given p columns in each design matrix and 3-dimensional imaging data we require $p + 10$ parameters per active component (three for m_k, six for Σ_k, p for w_k and 1 for σ_k^2). For, say, $p = 10$ and $K - 1 = 20$ active components we have a total of 400 model parameters. Mass-univariate models, however, require $p + 1$ parameters *per voxel*. Typical sized 3D fMRI images (of dimension 48 by 64 by 64) contain roughly 200,000 voxels giving approximately 2,000,000 parameters. The difference is stark; the MGLM model provides a much more parsimonious representation of the data.

C. Parameter Estimation

The generative model underlying MGLM assumes that the observation noise is independent over voxels and time points. The likelihood of the data under the model is therefore

$$P(D|\theta) = \prod_{i,t} p(y_i^t | v_i, \theta) \tag{7}$$

where $D = \{y_i^t\}$. Note that although the observation noise shows this independence, the deterministic component of the observations, the signal, will show strong regularities both over time, due to the temporal regularity of $\hat{y}_k^t$, and over space, due to the spatial smoothness of the prior probabilities.

An important feature of fMRI time series, however, is that the observation noise is temporally autocorrelated. In the MGLM model, we believe there is no need to take this into account (see discussion).

If we imagine that a given data set has been generated by an MGLM model, then at each voxel and at each time point it will have been decided which component was used to produce that sample. Let us denote this by s_i^t. For example, $s_i^t = 3$ for all t for voxel i at the top of Figure 3 (ie. all samples were generated from the active component $k = 3$). If we were given the variable s_i^t along with each data set then parameter estimation would be easy (the kth GLM, for example, would be inferred by simply fitting it to all data points for which $s_i^t = k$). But of course this variable is not generally available and we must regard it as a *hidden* variable. Fortunately, we can use a general procedure for parameter estimation in models with hidden variables. This is the Expectation-Maximisation (EM) algorithm [6]. In the E-step we compute the probability distribution over hidden variables and in the M-step we maximise the joint log-likelihood of the data and hidden variables under that distribution. EM is a proven method for finding the parameters, θ, which maximise the model likelihood.

An EM algorithm for the MGLM model is derived in the appendix and results in the following update rules. The E-step simply consists of computing the posterior probability of voxel i at time t having been sampled from component k, that is $p(k|y_i^t, v_i)$ which we also write as $\gamma_i^t(k)$. This is given by Bayes rule as

$$\gamma_i^t(k) = \frac{p(y_i^t|k)p(k|v_i)}{\sum_{k'} p(y_i^t|k')p(k'|v_i)} \tag{8}$$

For brevity we have dropped the dependence on the model parameters θ given in equations 2 and 4. We then also compute $\gamma_i(k) = \sum_{t=1}^N \gamma_i^t(k)/N$ and $\gamma_k = \sum_i^V \gamma_i(k)/V$. In the M-step of the EM algorithm the parameters of the spatial and temporal models are updated.

The parameters of the temporal model are estimated as follows. If we let $\Gamma_i(k) = diag[\gamma_i^1(k), \gamma_i^2(k), ..., \gamma_i^N(k)]$ be a diagonal matrix with entries being the temporal weights for that voxel and $Y_i = [y_i^1, y_i^2, ..., y_i^N]^T$ be the time series for voxel i then, for cluster k, we can define

$$Y_k = \sum_i \Gamma_i(k) Y_i \tag{9}$$

If we also let $\Gamma(k) = \sum_i \Gamma_i(k)$ then the regression coefficients are estimated as

$$w_k = (X_k^T \Gamma(k) X_k)^{-1} X_k^T Y_k \tag{10}$$

This is equivalent to Iteratively Re-weighted Least Squares (IRLS) [19] but with the addition that the voxel time-series receive different weightings at different locations *and* at different time points. This arises because the mixing process operates at each voxel and at each time point. The observation noise can then be re-estimated using

$$\sigma_k^2 = \frac{VN}{\gamma_k} \sum_t \sum_i \gamma_i^t(k)(y_i^t - \hat{y}_k^t)^2 \tag{11}$$

The means and covariances of the spatial parameters are updated using gradient ascent. To ensure that the covariances remain positive definite we use the decomposition (see eg. [27])

$$\Sigma_k = r_k r_k^T + \lambda_k I \tag{12}$$

with the constraint that $\lambda_k > 0$. This effectively renders the spatial density of active component k an ellipsoid with major and minor axes pointing along the columns of r_k. The parameters are then updated using

$$\begin{aligned} m_k &= m_k + \alpha_1 \frac{dQ_s}{dm_k} \\ r_k &= r_k + \alpha_2 \frac{dQ_s}{dr_k} \\ \lambda_k &= \lambda_k + \alpha_2 \frac{dQ_s}{d\lambda_k} \end{aligned} \tag{13}$$

where Q_s is the 'EM auxiliary function' for the spatial parameters. The function Q_s and the gradients are given in the appendix. Each gradient ascent step is implemented with Brent's line search algorithm (see p. 402 in [21]) which implicitly finds the optimal step size α_i. A small, positive minimal value for λ_k is naturally enforced in the initial bracketing used in Brent's algorithm.

To summarise, the EM algorithm operates as follows. In the E-step the posterior probabilities are updated using Equation 8. In the M-step w_k, σ_k^2, m_k and Σ_k are updated using Equations 10, 11, 12 and 13. The E and M steps are iterated until the proportionate increase in model log-likelihood from one step to the next is less than $1e^{-6}$, an arbitrary convergence criterion.

The main computational overheads of the EM algorithm are in the gradient ascent steps of Equation 13. Within these updates, the main bottleneck is in the evaluations of Q_s in the line search algorithm. This can be speeded up by noting that the Gaussians have only local support; changing m_k and Σ_k will only make a difference to Q_s in a small region. Therefore, by restricting the domain in which Q_s is computed we can greatly reduce the amount of computation required.

The algorithm we have described is, strictly speaking, a Generalised EM algorithm since each M-step does not maximise the auxiliary function but merely increases it (see, for example, [14]). We have also considered the use of a Conditional EM algorithm as described in [16], but found no computational advantage.

D. Inference

The probability that a voxel belongs to an active cluster is given by

$$\gamma_i^a = \sum_{k'} \gamma_i(k') \tag{14}$$

where k' are the active components. An image of γ_i^a constitutes a 'Posterior Probability Map (PPM)'. Three such maps are shown in Figures 5,6 and 9 which superimpose PPMs, thresholded at $h = 0.95$, on structural MRI images. Voxels can be declared active by comparing γ_i^a to some threshold h.

We can also define a likelihood ratio, l_i^a, as the ratio of the likelihood of the data under the active models to the likelihood of the data under the null model. The posterior probabilities and likelihood ratios are related as follows

$$\begin{aligned} l_i^a &= \frac{\gamma_i^a}{1-\gamma_i^a} \\ \gamma_i^a &= \frac{l_i^a}{1+l_i^a} \end{aligned} \tag{15}$$

Now, if we know the prior probability of observing an active voxel, p^a, then the optimum threshold for the likelihood ratio is $(1-p^a)/p^a$ [5]. This then implicitly defines the optimum value for h. For example, in a sensory study we may a priori expect 5% of voxels to activate. This corresponds to $h = 0.95$.

A second quantity of interest is the number of active components. This can, in principle, be found using Bayesian model order selection methods (see for example [23]) but this is beyond the scope of the present paper. Instead we use the following heuristic. We fit a family of MGLM models with increasing K and for each $k = 2..K$ compute a value t_k which is the t-statistic corresponding the inferred values for w_k and σ_k^2. If $p(t > t_k) < 0.001$ then we declare that at least $K-1$ components are active ($k = 1$ is the null component), and proceed to fit an MGLM model with K active components. Otherwise the model order selection process stops.

By noting that the mean activity of voxels in cluster k is given by

$$\bar{Y}_k = \frac{Y_k}{\Gamma(k)} \tag{16}$$

we can extract an 'unsupervised' or a 'semi-supervised' estimate of the temporal activity underlying each cluster. By this we mean that inference could also proceed on the basis of $\bar{Y}_k$ rather than on the basis of parameters from the GLM. This would be in the spirit of cluster-based analyses of fMRI (see eg. [4]).

E. Initialisation

One potential problem with the MGLM model is the possibility that there may be many local maxima in the likelihood landscape. The current optimisation approach does not warrant that the global maximum is found, especially with a large number of mixture components. The empirical work in this paper concedes that only a local maxima will be reached but to guarantee that this is a useful solution the spatial priors are set so that the active components are initially centred on voxels strongly correlated with the activation paradigm. More principled solutions include the use of split and merge heuristics where components are divided or combined and the resulting model kept, depending on whether a probabilistic fitness criterion is met [28]. A fully Bayesian solution to this problem can be implemented using Reversible Jump Markov Chain Monte Carlo (RJ-MCMC) [12]. These approaches can, in principle, be applied to the MGLM model and will be the subject of future work.

The current initialisation method proceeds as follows. We first find the voxel positions of the K largest maxima in the correlation or t-statistic image which are at least 15mm apart. These are used as K 'seed points'. We then fit GLMs to the data at these voxels and so infer w_k and σ_k^2. The mean m_k is set to the seed position and the diagonal terms in the covariance, Σ_k, are set so as to correspond to a Full Width at Half Maximum (FWHM) of 6mm. The initial solutions thus correspond to strong, focal activations. By optimising w_k, m_k and Σ_k we can find weaker or stronger, more or less diffuse activations. The extent to which the MGLM homes in on each is decided by the model likelihood and the EM optimisation process.

F. Data Sets

We use two fMRI data sets [1]. Both were acquired on a 2T VISION system (Siemens, Erlangen, Germany) which produces T2*-weighted transverse Echo-Planar Images (EPIs) with Blood Oxygen Level Dependent (BOLD) contrast. The first was recorded during an auditory stimulation task. This consisted of bi-syllabic words (eg. 'motor', 'robust') being presented at a rate of 60 per minute. The data set is made up of six blocks of auditory stimulation alternated with 6 blocks of rest, each block lasting 30 seconds (this block structure is reflected in the time series in Figure 7). Whole-brain fMRI images were acquired every seven seconds using 30 transverse slices.

The second data set was recorded during an experiment concerned with the processing of images of faces [15]. This was an event-related study in which greyscale images of faces were presented for 500ms, replacing a baseline of an oval chequerboard which was present throughout the interstimulus interval (ISI). The ISI followed a stochastic distribution with a minimal interval of 4.5s. In this paper we focus on only a subset of this data concerned with the differential activation of voxels subsequent to the presentation of face trials versus baseline trials. Differentially activated areas will be involved in face processing rather than the processing of images per se. Whole brain EPIs consisting of 24 transverse slices were acquired with an effective repetition time of two seconds.

All functional images were realigned to the first functional image using a six-parameter rigid-body transformation [9]. Functional images were then spatially normalised to a standard Echo-Planar Image (EPI) template using a nonlinear warping method [1]. The images were then scaled to remove global effects using proportional scaling.

We then created two different sets of images. For analysis using SPM, the images were spatially smoothed using a Gaussian with FWHM=6mm (this is necessary to ensure that Gaussian Random Field Theory is not too conservative [8]). Importantly, however, images to be analysed using the MGLM model were *not* spatially smoothed. Instead, they were processsed so that each voxel had zero mean and unit variance. This is necessary as the signal components will only be driven to areas containing signals if the residual error in the signal areas is less than the error in the noise areas, otherwise the overall model would not have a higher likelihood. Note that this normalization does not affect the t-values in a conventional SPM t-map.

For both data sets we focus on single slices. For the auditory data we chose a transverse slice at $z = 10$mm and for the face data a transverse slice at $z = -24$mm (these positions are given in Talairach co-ordinates [26]).

For the auditory data, the design matrices, for both SPM and MGLM, contained a column of 1's and a variable indicating the experimental condition. This variable consisted of a boxcar, with 1's indicating the presence of an auditory stimulus and 0's indicating its absence, convolved with a

[1] These data sets and a full description of the experiments and data pre-processing are available from http://www.fil.ion.ucl.ac.uk/spm/data

hemodynamic response function [10], a standard way of modelling the hemodynamic response. For the face data, the design matrices contained a column of 1's and 4 other columns indicating when the face images were presented. There were 4 such columns rather than one as there were two types of image, famous and non-famous, and each face was presented twice. Modelling the response in this way, rather than with a single variable, results in a more accurate model fit. It also allows for the investigation of repetition effects [15], although this is not explored in the current paper.

III. Results

The results of a standard SPM analysis are given in Figure 4. The results are displayed in the form of a t-statistic image overlaid on a structural fMRI scan from that subject. The plot shows diffuse bilateral activation of primary auditory cortex. We then applied a series of MGLM models to the data with increasing K. Our model order selection heuristic (see the section on inference) stopped at a model with 3 active components; two covering the left activation and one covering the right. The corresponding Posterior Probability Map (PPM) is shown in Figure 6. We also show the MGLM model with two active components in Figure 5. The PPMs have been thresholded at $h = 0.95$.

The PPMs and SPMs are in general agreement, with the PPMs being somewhat more conservative. This is, however, an artifact due to the choice of thresholds for which the images are plotted ie. SPMs corrected at $p < 0.01$ rather than $p < 0.05$ show a similarly conservative pattern.

Figure 7 shows the corresponding time course of activations for the right active component, the block structure reflecting the block-like nature of the stimulus and the peaks and troughs reflecting the hemodynamic over- and undershoot. The consistent excessive undershoot in later blocks shows that the fit of the GLM could be improved by adding regressors to the design matrix (eg. a time effect).

As mentioned earlier, the MGLM model is a much more economic model of functional activation than is the mass-univariate approach underlying SPM. For this data set, the 3-component MGLM model has 24 parameters whereas the SPM model has 15,010 parameters. If one were interested in finding efficient codes for storing the data then MGLM would offer a considerable advantage. We note that if this were truly the case, then a finessed characterisation of the null component would be appropriate (see discussion).

The result of a standard SPM analysis of the event-related study is given in Figure 8. This shows bilateral activation of fusiform cortex and earlier visual areas. We then applied a series of MGLM models to the data with increasing K. Our model order selection heuristic stopped at a model with 2 active components, one covering the left activation and one covering the right. The Posterior Probability Map (PPM), thresholded at $h = .95$, is shown in Figure 9. Again, the PPM and SPM are in general agreement, with the more conservative nature of the PPM being attributable to differences in thresholding and possibly smoothing.

Figure 10 shows the corresponding time course of activations for the active component in the right hemisphere. The solid line shows the estimated response from the GLM, $\hat{Y}_k$, and the dotted line shows the unsupervised estimate of temporal activity $\bar{Y}_k$. This is the quantity of interest in cluster-based analysis [4]. By comparing $\bar{Y}_k$'s from different clusters, inferences can be made, albeit informally, about differential delays in hemodynamic response. We note, however, that this can also be achieved by including the temporal derivates of the canonical HRF in the design matrix and making formal inferences about the corresponding regression coefficient. The middle time series in Figure 10 show how the estimated GLM responses can be improved by including more regressors. We used a model with 19 regressors (as in [15]) and this led to a 20% reduction in the fitted error.

For this data set, the MGLM model has 45 parameters whereas the SPM model has 20,034. Again, a great saving.

IV. Discussion

We have proposed a new approach to the analysis of functional neuroimaging data. The central tenet of these models is that the fundamental quantities of interest to the neuroimager are the location, shape and temporal signature of *clusters* of voxels showing task-related activity. SPM is a special case of our model, recovered when the number of clusters equals the number of voxels and all active clusters have the same design matrix.

For each cluster of activation we have a single representative time series. This means that the MGLM model may be particularly helpful in the analysis of effective connectivity [3] that examines the interactions among different brain regions. Previously, the requisite time series have been derived by defining a region of interest using, for example, a sphere of arbitrary size and then finding the principal eigen-time-series [3]. MGLM offers a much more precise and principled approach.

The model we have proposed is similar in spirit to the Stochastic Geometry Model (SGM) of Hartvig [13]. This models the activations as a sum of Gaussians of unknown location and scale whose parameters are estimated using Bayesian methods. The generative models underyling MGLM and SGM are, however, very different. The most fundamental difference is that the SGM Gaussians reflect the magnitude of activations whereas the MGLM Gaussians reflect the likelihood that a voxel belongs to a cluster.

We also note similarities with the Nonlinear Spatio-Temporal (NST) model proposed by Solo *et al* [24]. A common feature is that information is pooled across voxels using an adaptive spatial kernel. Again, the MGLM and NST generative models are very different. For example, in NST pooling is used to estimate noise parameters rather than signal parameters.

An important feature of fMRI time series is that the observation noise is temporally autocorrelated. In the mass-univariate approach it is necessary to take this correlation into account as, to neglect it, would severely bias the subsequent inferences. Essentially, instead of the degrees of freedom being equal to the length of each time series it is much less. For MGLM models, however, the degrees of freedom in each temporal model is equal to the length of the time series times the number of voxels belonging to that cluster (because we have borrowed strength). As this is so large, any reduction due to temporal autocorrelation is likely to make little difference to the subsequent inferences. Those not persuaded by this argument could alter the generative model underlying MGLM so that the mixing process takes place at each voxel, rather than at each voxel and at each time point. Standard time series models which allow for temporal autocorrelation such as GLMs with autoregressive error terms could then be implemented. We note that since the amount of autocorrelation is dependent on space, the resulting cluster shapes may be different. However, this is likely to be a subtle effect because, on average, Weighted Least Squares parameter estimates (which take into account error autocorrelation) are identical to Ordinary Least Squares estimates (which do not).

On a more critical note, the amount of computation required to estimate the parameters in a MGLM model is an order of magnitude greater than that for the mass-univariate approach, taking several minutes per slice instead of several seconds. This is, however, an attribute shared by other spatio-temporal models [7], [24], [13], and appears to be the price we pay for more parsimonious yet informed characterisations of fMRI data.

We also note that the MGLM model is closely related to cluster-analysis methods. An important difference between the PPMs from the MGLM model and the maps of spatial activation produced by cluster analysis, however, is that the PPMs have blobs whereas the cluster maps have speckles (see eg. [4]). This is because, for a voxel to belong to a cluster in the MGLM model it must have an appropriate time series *and* be in the appropriate position. In essence, MGLM performs a semi-supervised spatio-temporal cluster analysis.

The model we have proposed could be usefully enhanced by greater use of prior information. For example, instead of having a single prior for the null class, being a uniform density over the whole brain, we envisage the use of tissue-specific priors describing the spatial distribution of white matter and cerebro-spinal fluid would make a useful contribution. This would increase the probability of functional activations being identified in gray matter. This is in the spirit of previous work in the area by Kiebel and Friston [18].

Acknowledgments

We thank John Ashburner and Stefan Kiebel for advice on fMRI processing and Rik Henson for advice on the face processing data. We thank all members of the FIL methods group for their help in improving this work, especially Stefan Kiebel. Both authors are supported by the Wellcome Trust.

References

[1] J. Ashburner and K. J. Friston. Nonlinear spatial normalization using basis functions. *Human Brain Mapping*, 7(4):254–266, 1999.

[2] C. M. Bishop. *Neural Networks for Pattern Recognition*. Oxford University Press, Oxford, 1995.

[3] C. Buchel and K.J. Friston. Modulation of connectivity in visual pathways by attention: Cortical interactions evaluated with structural equation modelling and fMRI. *Cerebral Cortex*, 7:768–778, 1997.

[4] E. Rostrup F.A. Nielsen C. Goutte, P. Toft and L.K. Hansen. On Clustering fMRI Time Series. *NeuroImage*, 9:298–310, 1999.

[5] D.R. Cox and D.V. Hinkley. *Theoretical Statistics*. Chapman and Hall, 1974.

[6] A.P. Dempster, N.M. Laird, and D.B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society B*, 39:1–38, 1977.

[7] X. Descombes, F. Kruggel, and D.Y. von Cramon. fMRI Signal Restoration Using a Spatio-Temporal Markov Random Field Preserving Transitions. *NeuroImage*, 8:340–349, 1998.

[8] R.S.J. Frackowiak, K.J. Friston, C.D. Frith, R.J. Dolan, and J.C. Mazziotta, editors. *Human Brain Function*. Academic Press USA, 1997.

[9] K. J. Friston, J. Ashburner, C. D. Frith, J.-B. Poline, J. D. Heather, and R. S. J. Frackowiak. Spatial registration and normalization of images. *Human Brain Mapping*, 2:165–189, 1995.

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

[11] C.R. Genovese. Functional magnetic resonance imaging and spatio-temporal inference. In A.P. Dawid J.M. Bernardo, J.O. Berger and A.F.M. Smith, editors, *Bayesian Statistics*, volume 6, pages 255–274, 1999.

[12] P.J. Green. Reversible jump Markov chain Monte Carlo computation and bayesian model determination. *Biometrika*, 82:711–732, 1995.

[13] N.V. Hartvig. A stochastic geometry model for fMRI data. Technical Report 410, Department of Theoretical Statistics, University of Aarhus, 1999.

[14] T. Hebert and R. Leahy. A generalized EM algorithm for 3-D Bayesian Reconstruction from Poisson Data Using Gibbs Priors. *IEEE Transactions on Medical Imaging*, 8(2):194–202, 1989.

[15] R.N.A. Henson, T. Shallice, M.L. Gorno-Tempini, and R.J. Dolan. Face repetition effects in implicit and explicit memory tests as measured by fMRI. *Cerebral Cortex*, 12:178–186, 2002.

[16] T. Jebara and A. Pentland. Maximum conditional likelihood via bound maximization and the CEM algorithm. In *Neural*

Information Processing Systems 11 (NIPS '98), Denver, CO, December, 1998.

[17] O. Josephs, R. Turner, and K.J. Friston. Event-related fMRI. *Human Brain Mapping*, 5:243–248, 1997.

[18] Stefan J. Kiebel, Rainer Goebel, and Karl J. Friston. Anatomically informed basis functions. *NeuroImage*, 11(6):656–667, 2000.

[19] P. McCullagh and J.A. Nelder. *Generalised Linear Models.* Chapman and Hall, 1983.

[20] R.S. Menon, S. Ogawa, X. Hu, J.P. Strupp, P. Anderson, and K. Ugurbil. Bold based functional MRI at 4 tesla includes a capillary bed contribution: Echo planar imaging correlates with previous optical imaging using intrinsic signals. *Magnetic Resonance in Medicine*, 33:453–459, 1995.

[21] W. H. Press, S.A. Teukolsky, W.T. Vetterling, and B.V.P. Flannery. *Numerical Recipes in C.* Cambridge, 1992.

[22] J.C. Rajapakse and J. Piyaratna. Bayesian approach to segmentation of statistical parametric maps. *IEEE transactions on Biomedical Engineering*, 48(10):1186–1194, 2001.

[23] S.J. Roberts, D. Husmeier, I. Rezek, and W. Penny. Bayesian approaches to Gaussian mixture modelling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(11):1133–1142, 1998.

[24] V. Solo, P. Purdon, R. Weiskoff, and E. Brown. A Signal Estimation Approach to Functional MRI. *IEEE Transactions on Medical Imaging*, 20(1):26–35, 2001.

[25] M. Svensen, F. Kruggel, and D. Yves von Cramon. Probabilistic modeling of single trial fMRI data. *IEEE Transactions on Medical Imaging*, 19(1):25–35, 2000.

[26] J. Talairach and P. Tournoux. *Coplanar stereotaxic atlas of the human brain.* Thieme Medical, New York, 1988.

[27] M. E. Tipping and C. M. Bishop. Probabilistic Principal Component Analysis. Technical report, Neural Computing Research Group, Aston University, 1997.

[28] N. Ueda, R. Nakano, Z. Ghahramani, and G.E. Hinton. SMEM algorithm for mixture models. In *Advances in Neural Information Processing Systems 11.* MIT Press, 1999.

Appendix

I. EM Algorithm

The log-likelihood of the data is given by

$$L = \sum_i \sum_t \log p(y_i^t|v_i) \tag{17}$$

The likelihood can be maximised by maximising an EM auxiliary function, Q. Maximising Q provably maximises the model likelihood, as shown in [6]. For models with hidden variables this Q function has a standard form; it is the log of the joint probability of observed and hidden variables averaged over the posterior distribution of hidden variables. For the MGLM model we have

$$Q = \left\langle \sum_k \sum_i \sum_t \log p(y_i^t, k|v_i) \right\rangle \tag{18}$$

where the angled brackets denote expectation over the distribution $p(k|y_i^t, v_i)$. For brevity we write $\gamma_i^t(k) = p(k|y_i^t, v_i)$ which can be expressed as

$$\gamma_i^t(k) = \frac{p(y_i^t, k|v_i)}{p(y_i^t|v_i)} \tag{19}$$

and re-written in terms of the temporal and spatial probabilities (see Figure 1 and equation 1)

$$\gamma_i^t(k) = \frac{p(y_i^t|k)p(k|v_i)}{\sum_{k'} p(y_i^t|k')p(k'|v_i)} \tag{20}$$

The E-step of the EM algorithm simply consists of computing this distribution. We also compute $\gamma_i(k) = \sum_{t=1}^N \gamma_i^t(k)/N$ and $\gamma_k = \sum_i^V \gamma_i(k)/V$.

The joint distribution in equation 18 is given by

$$p(y_i^t, k|v_i) = p(y_i^t|k)p(k|v_i) \tag{21}$$

Hence its expectation over $\gamma_i^t(k)$ is

$$\begin{aligned} Q &= \sum_k \sum_i \sum_t \gamma_i^t(k) \log p(y_i^t|k) \\ &+ \sum_k \sum_i \sum_t \gamma_i^t(k) \log p(k|v_i) \end{aligned} \tag{22}$$

which can be written in terms of a temporal term (first) and a spatial term (second)

$$Q = Q_t + NQ_s \tag{23}$$

The update rules are derived by finding the turning points of the above function.

A. Spatial model

The likelihood is given by

$$p(v_i|k) = \frac{1}{(2\pi)^{d/2}|\Sigma_k|^{1/2}} \exp\left(-\frac{1}{2}(v_i - m_k)^T \Sigma_k^{-1}(v_i - m_k)\right) \tag{24}$$

and

$$p(k|v_i) = \frac{p(v_i|k)}{\sum_{k'} p(v_i|k')} \tag{25}$$

We can re-write the above equation in terms of the *softmax* function

$$g_i(k) = \frac{\exp[a_i(k)]}{\sum_{k'} \exp[a_i(k')]} \tag{26}$$

where

$$a_i(k) = \frac{-d}{2}\log(2\pi) - \frac{1}{2}\log|\Sigma_k| - \frac{1}{2}(v_i - m_k)^T \Sigma_k^{-1}(v_i - m_k) \tag{27}$$

We have

$$Q_s = \sum_k \sum_i \gamma_i(k) \log g_i(k) \tag{28}$$

We can then use the standard result (see eg. p.237-240 in [2])

$$\frac{dQ_s}{da_i(k)} = \gamma_i(k) - g_i(k) \tag{29}$$

and combine it with

$$\frac{da_i(k)}{dm_k} = \Sigma_k^{-1}(v_i - m_k) \tag{30}$$

and

$$\frac{da_i(k)}{d\Sigma_k^{-1}} = -\frac{1}{2}(v_i - m_k)(v_i - m_k)^T - \frac{1}{2}\Sigma_k \tag{31}$$

to get $\frac{dQ_s}{dm_k}$ and $\frac{dQ_s}{d\Sigma_k}$. These gradients can then be used in a line search to find updates for m_k and Σ_k. Whilst this is straightforward for the mean it does not ensure positive definiteness for Σ_k. We therefore decompose the spatial covariance using $\Sigma_k = r_k r_k^T + \lambda_k I$ and use gradient-based line searches to optimise r_k and λ_k. The required gradients can be derived using the chain rule or estimated using central differences (see eg. p. 146 in [2]).

B. Temporal model

We let $\Gamma_i(k) = diag[\gamma_i^1(k), \gamma_i^2(k), ..., \gamma_i^N(k)]$ be a diagonal matrix with entries being the temporal weights for that voxel and $Y(i) = [y_i^1, y_i^2, ..., y_i^N]^T$ be the time series for voxel i. For the kth component we have

$$\begin{aligned} Q_t(k) &= \sum_i Y^T(i)\Gamma_i(k)Y(i) \\ &- 2\sum_i w_k^T X_k^T \Gamma_i(k) Y(i) + \sum_i w_k^T X_k^T \Gamma_i(k) X_k w_k \end{aligned} \tag{32}$$

For the regression coefficients we have

$$\frac{dQ_t(k)}{dw_k} = -2\sum_v X_k^T \Gamma_k(i) Y(i) + 2X_k^T \sum_i \Gamma_i(k) X_k w_k \tag{33}$$

Letting

$$\begin{aligned} Y_k &= \sum_i \Gamma_i(k) Y(i) \\ \Gamma_k &= \sum_i \Gamma_i(k) \end{aligned} \tag{34}$$

we have

$$\frac{dQ_t(k)}{dw_k} = -2X_k^T Y_k + 2X_k^T \Gamma_k X_k w_k \tag{35}$$

which has a turning point at

$$w_k = (X_k^T \Gamma_k X_k)^{-1} X_k^T Y_k \tag{36}$$

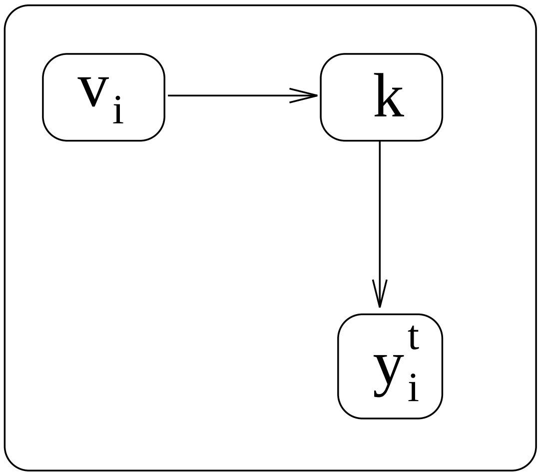

Fig. 1. *The generative model underlying MGLM. Voxel i is chosen deterministically. For time point t we choose component k with probability $p(k|v_i)$. A data point is then selected with probability $p(y_i^t|k)$. This is then repeated for all voxels and all time points. The probabilistic dependence means that we can write $p(y_i^t, k|v_i) = p(y_i^t|k)p(k|v_i)$. Summing over k gives equation 1.*

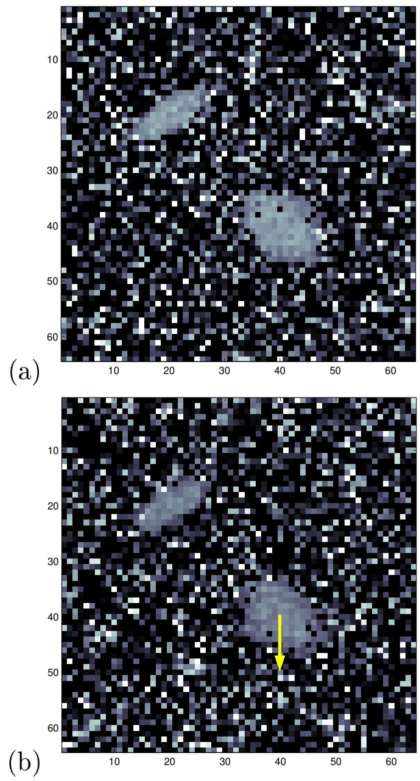

Fig. 2. *Images from generative model at times (a) $t = 8$ and (b) $t = 9$. This model comprises a null component ($k = 1$) and two active components (top left, $k = 2$; bottom right, $k = 3$). Note that the shape of the active regions is consistent between scans, but it is not identical. This is due to the mixing process operating at each time point. The arrow in figure (b) indicates the voxels whose time series are plotted in Figure 3.*

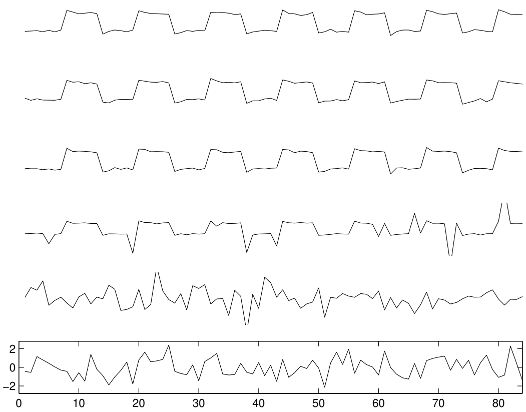

Fig. 3. *Time series from generative model. Going down the page, we see time series from voxels at positions indicated by the arrow in Figure 2, that is, leaving the centre of the bottom right cluster at* $v = [40, 40]$ *and going down to* $v = [40, 50]$*. Time series are plotted every two voxels. As we leave the cluster the time series gradually become more noisy.*

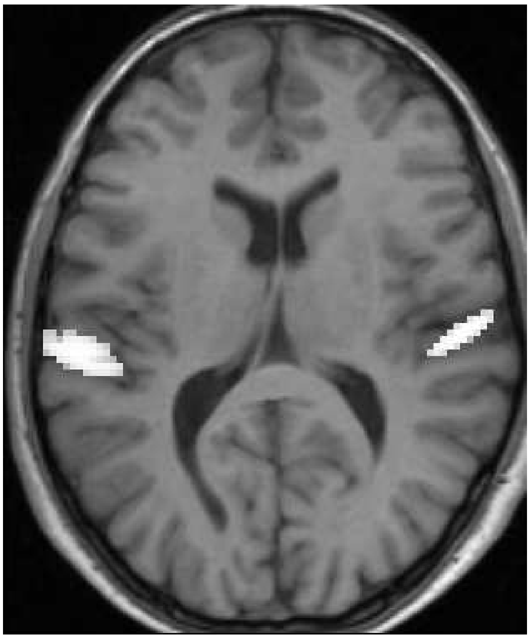

Fig. 5. **Auditory Data** *Posterior Probability Map (PPM) from an MGLM model with two active components. The bright pixels correspond to voxels for which* $\gamma_i^a > 0.95$*.*

Fig. 4. **Auditory Data** *Statistical Parametric Map (SPM) showing active voxels (*$p \leq 0.05$*, corrected for multiple comparisons). The bright pixels correspond to the SPM t-values and are scaled so that t=5.23 (this corresponds to* $p = 0.05$*, corrected) is gray and the maximum t-value is white.*

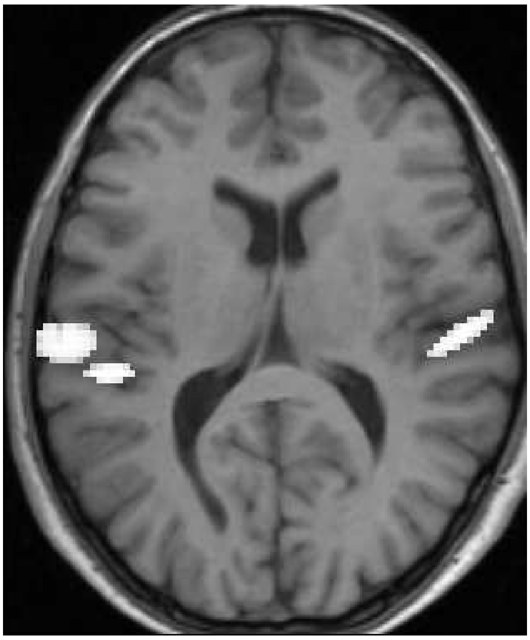

Fig. 6. **Auditory Data** *Posterior Probability Map (PPM) from an MGLM model with three active components.*

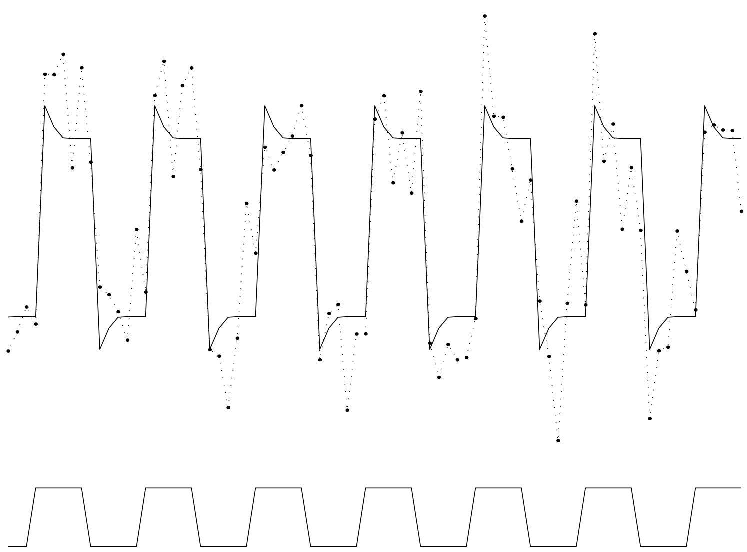

Fig. 7. **Auditory Data** *Time series for the right-hemisphere component of the MGLM model. The solid lines show the responses estimated under the General Linear Model,* $\hat{Y}_k$*, and the dotted lines show the mean voxel actvity,* $\bar{Y}_k$*. The boxcar at the bottom is high when the auditory stimulus was present (this demarcates a 30-second period).*

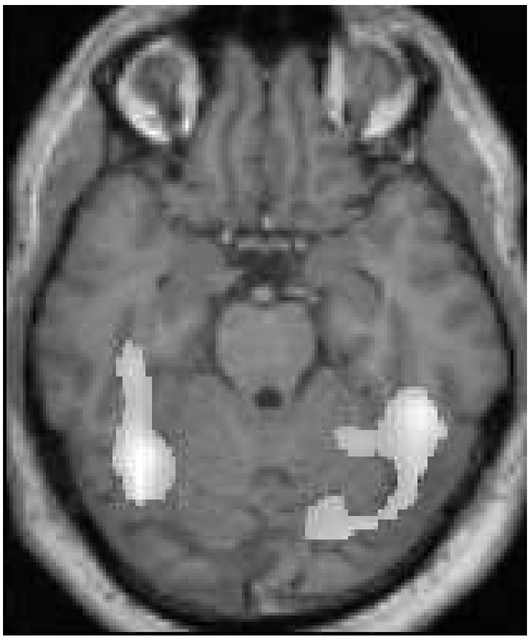

Fig. 8. **Face Data** *Statistical Parametric Map (SPM) showing active voxels (*$p \leq 0.05$*, corrected for multiple comparisons).*

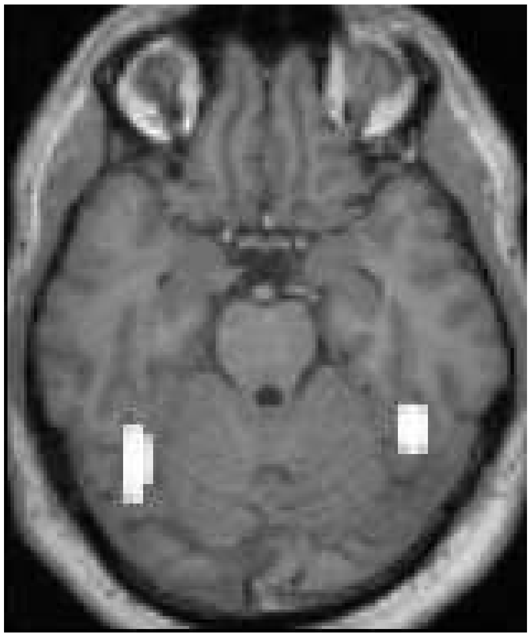

Fig. 9. **Face Data** *Posterior Probability Maps (PPM) from MGLM model with two active components. The single active component model consisted of just the left-hemisphere activation.*

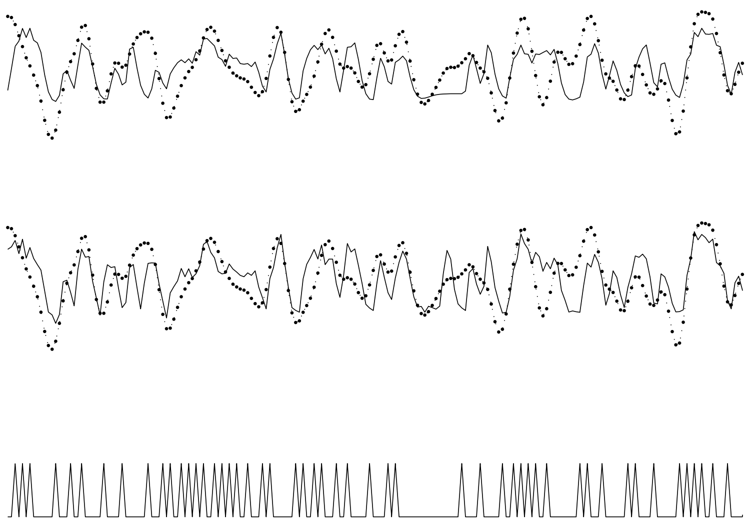

Fig. 10. **Face Data** *Time series for the right-hemisphere component of the MGLM model. A four-hundred second period is shown. The top time series show the estimated responses (solid line),* $\hat{Y}_k$*, and the mean voxel activity (dotted line),* $\bar{Y}_k$*, for an MGLM model with 5 regressors in each design matrix. The middle time series show how the estimated responses are improved when using 19 regressors in each design matrix. The spikes at the bottom indicate when the face images were presented.*