An Adaptive Regression Mixture Model for fMRI Cluster Analysis

V. P. Oikonomou and K. Blekas
Department of Computer Science, University of Ioannina
P.O.Box 1186, Ioannina 45110 - GREECE
E-mail: viknmu@gmail.com,kblekas@cs.uoi.gr

***Abstract*—Functional MRI (fMRI) has become one of the most important techniques for studying the human brain in action. A common problem in fMRI analysis is the detection of activated brain regions in response to an experimental task. In this work we propose a novel clustering approach for addressing this issue using an adaptive regression mixture model. The main contribution of our method is the employment of both spatial and sparse properties over the body of the mixture model. Thus, the clustering approach is converted into a Maximum a Posteriori (MAP) estimation approach, where the Expectation-Maximization (EM) algorithm is applied for model training. Special care is also given to estimate the kernel scalar parameter per cluster of the design matrix by presenting a multi-kernel scheme. In addition an incremental training procedure is presented so as to make the approach independent on the initialization of the model parameters. The latter also allows us to introduce an efficient stopping criterion of the process for determining the optimum brain activation area. To assess the effectiveness of our method, we have conducted experiments with simulated and real fMRI data, where we have demonstrated its ability to produce improved performance and functional activation detection capabilities.**

Keywords: fMRI analysis, regression mixture models, EM algorithm, MRF, Sparse modeling

I. Introduction

Functional Magnetic Resonance Imaging (fMRI) is a powerful, non - invasive technique that has been utilized in both research and clinical fields, providing valuable information about the structure and the organization of the human brain. fMRI measures the tiny metabolic changes that takes place in an active part of the brain. It is a common diagnostic method for the behavior of a normal, diseased or injured brain, as well as for assessing the potential risks of surgery or other invasive treatments of the brain [1], [2].

Many methods have been proposed for the analysis of fMRI data coming from fields such as signal processing, machine learning and statistics. They can be divided into two major categories: the model-based and the data driven techniques. The first category consists of approaches which are mainly based on the general linear regression model [3] and its extensions [4], [5], while the data driven techniques includes the principal component analysis (PCA) [6], independent component analysis (ICA) [7], [8] and clustering algorithms [6], [8]–[10]. They mainly differ on the use of the hemodynamic model and the experimental paradigm for detecting the brain response [10], [11].

Clustering is the procedure of dividing a set of unlabeled data into a number of groups (called clusters), in such a way that similar in nature samples to belong to the same cluster, while dissimilar samples to become members of different clusters [12]. Cluster analysis of fMRI data constitutes a very interesting application that has been successfully applied during last years. Most popular clustering methods are the k-means, fuzzy clustering and hierarchical clustering which are applied to either raw data, or features that are extracted from the fMRI signals [9], [10], [13]–[19]. The aim of any clustering method in fMRI is to create a partition of the entire data set into distinct regions where each region consists of voxels with similar temporal behavior. Analysing fMRI data finds some obstacles which are the large size and complexity of raw data, the low contrast-to-noise ratio and the presence of artifacts. Moreover, there are some significant constraints coming from the nature of data, such as spatial correlation among them, which must be taken into account.

In this direction spatially constrained models have been proposed for fMRI data analysis. In [20] a linear regression model has been adopted, where a spatially constrained mixture model was used for the modeling of the Hemodynamic Response Function (HRF). In [21] the fuzzy c-means algorithm in cooperation with a spatial MRF was proposed to cluster the fMRI data. Furthermore, in [22], [23] mixture models with spatial MRFs have been applied on statistical maps to perform the clustering. However, in the above works the clustering procedure was performed indirectly on fMRI time series, either through careful construction of the regression model, or using features extracted from the fMRI time series. Also, the temporal patterns of clusters have not been taken into account. A solution to this is to perform the clustering directly to fMRI time series. A useful tool to achieve that is the mixture model, where each component is a linear regression model. In [24] a mixture of General Linear Regression models (GLMs) was proposed for clustering that takes into account the spatial correlation of time series using a spatial prior based on the Euclidean distances between the positions of time series and cluster centers in 3-D space head model. Recently, in [25] a mixture of linear regression models was also proposed, where spatial correlations among the time series is achieved through Potts models over the hidden variables of the mixture model.

In this work we propose an advanced regression mixture modeling approach for clustering fMRI time series. The main advantage of the proposed method lies on three aspects.

Firstly, we employ a sparse representation of every cluster regression model through the use of an appropriate sparse prior over the regression coefficients [26]. Enforcing sparsity is a fundamental machine learning regularization principle [12], [26] and has been used in fMRI data analysis [4], [27], [28]. Secondly, spatial constraints of fMRI data have been incorporated directly to the body of mixture model using a Markov random field (MRF) prior over the voxel's labels [29], [30], so as to create smoother activation regions. Finally, we present a kernel parameter estimation framework through a multi-kernel scheme over the design matrix of the regression models. In this way we manage to improve the data fitting procedure and to design more compact clusters increasing the quality of the clustering solution.

The task of clustering is then formulated as a Maximum A Posteriori (MAP) estimation problem, where the known Expectation-Maximization (EM) algorithm can be applied for solving it. Since there is a dependence on the initialization, we also present an incremental learning procedure for building the proposed regression mixture model. This happened with a repeated component splitting procedure using a particular stopping criterion. In this way we manage not only to make the learning process independent on the initialization of model parameters, but also to construct a model-order selection criterion for the complexity of the mixture model. We have evaluated the proposed adaptive regression mixture model with both artificial and real fMRI datasets. Comparison has been made using the a regression mixture model with only spatial properties, the Generalized Linear regression Model (GLM) that constitutes a classical model-based approach and the known k-means clustering algorithm. As experiments have shown, the proposed method offers very promising results with an excellent behavior in difficult and noisy environments.

The remainder of this paper is organized as follows: In section II we briefly describe the standard mixture of linear regression models as a platform for clustering fMRI time series. In section III the proposed regression mixture is presented, which considers the spatial and sparse priors, the multi-kernel scheme and the incremental learning strategy. To assess the performance of the proposed methodology we present in section IV numerical experiments with artificial and real datasets, while, in section V we give some concluding remarks.

II. The mixture of linear regression models

Let $\mathbf{Y} = \{\mathbf{y}_1, \mathbf{y}_2, \cdots, \mathbf{y}_N\}$ be a set of N fMRI time series of length T, i.e. each element $\mathbf{y}_n$ is a sequence of values measured at T successive time instances x_l, i.e. $\mathbf{y}_n = \{y_{nl}\}_{l=1,\cdots,T}$. It must be noted that the data analysis in our study has been made on a single slice. During our experiments we have studied the possibility of our method to handle 3D cases with fMRI data by considering independence among different slices. However, working directly in 3D brain images finds a limitation of increasing computational complexity.

Linear regression modeling constitutes an elegant functional description framework for analyzing sequential data. It is described with the following form:

$$\mathbf{y}_n = \mathbf{X}\mathbf{w}_n + \mathbf{e}_n \ , \tag{1}$$

where $\mathbf{w}_n$ is the vector of M (unknown) linear regression coefficients, while $\mathbf{e}_n$ corresponds to the stochastic noise term that is assumed to be zero mean Gaussian with variance σ_n^2, i.e. $\mathbf{e}_n \sim \mathcal{N}(0, \sigma_n^2\mathbf{I})$. Finally, $\mathbf{X}$ is the design matrix where its selection plays an important role for the data analysis. A typical design matrix scheme is the Vandermonde or B-splines matrix dealing with polynomial or splines models, respectively. However, a more powerful strategy is to assume a kernel design matrix using an appropriate kernel basis function over time instances $\{x_l\}_{l=1}^T$. A common choice is to use the Gaussian kernel

$$[X]_{lk} = K(x_l, x_k; \lambda) = \exp\Big(-\frac{(x_l - x_k)^2}{2\lambda}\Big) \ ,$$

where λ is a scalar parameter. Specifying the proper value for this parameter is an important issue that may affect drastically the quality of the fitting procedure. In general, its choice depends on the amount of local variations of data which must be taken into account. In addition, the design matrix may contain information about the experimental paradigm of fMRI experiment. According to this model, the conditional probability density of the observative sequence $\mathbf{y}_n$ is also Gaussian

$$p(\mathbf{y}_n|\theta_n) = \mathcal{N}(\mathbf{X}\mathbf{w}_n, \sigma_n^2\mathbf{I}) \ .$$

where θ_n is the set of model parameters, i.e. $\theta_n = \{\mathbf{w}_n, \sigma_n^2\}$.

In this study we consider the clustering problem, i.e. the division of the input set of time series Y into K clusters, in such a way that each cluster contains similar in nature elements. This is equivalent of assuming that each cluster has its own regression generative mechanism, as given by a conditional density with parameters $\theta_j = \{\mathbf{w}_j, \sigma_j^2\}$. Mixture modeling provides a powerful platform of establishing the clustering procedure. It is described with the following probability density:

$$f(\mathbf{y}_n|\Theta) = \sum_{j=1}^{K} \pi_j p(\mathbf{y}_n|\theta_j) \ , \tag{2}$$

where π_j are the weights (prior probabilities) of every cluster that satisfy the constraints: $\pi_j \geq 0$ and $\sum_{j=1}^K \pi_j = 1$, while Θ is the set of all mixture model parameters, i.e $\Theta = \{\pi_j, \theta_j\}_{j=1}^K$. Assignment of the data to the K groups is then achieved according to the maximun of the posterior probabilities of component membership:

$$P(j|\mathbf{y}_n, \Theta) = \frac{\pi_j p(\mathbf{y}_n|\theta_j)}{f(\mathbf{y}_n|\Theta)} \ . \tag{3}$$

Based on the above formulation, the task of clustering can be converted into a parameter estimation problem. In this direction the Expectation Maximization (EM) algorithm [31] constitutes an elegant solution for fitting the model and maximizing the log-likelihood function:

$$l(\Theta) = \log p(Y|\Theta) = \sum_{n=1}^{N} \log\Big\{\sum_{j=1}^{K} \pi_j p(\mathbf{y}_n|\theta_j)\Big\}. \tag{4}$$

It consists of two main steps that are applied iteratively. The E-step, where the current posterior probabilities of component membership are calculated $z_{nj} = P(j|\mathbf{y}_n, \Theta)$ (Eq. 3), and the

M-step where the maximization of the expected complete log-likelihood (Q-function) is performed with respect to the model parameters

$$Q(\Theta|\Theta^{(t)}) = \sum_{n=1}^{N}\sum_{j=1}^{K} z_{nj}\Big\{\log\pi_j - \frac{T}{2}\log 2\pi - T\log\sigma_j - \frac{\|\mathbf{y}_n - \mathbf{X}\mathbf{w}_j\|^2}{2\sigma_j^2}\Big\}. \quad (5)$$

The maximization leads to the following update rules:

$$\pi_j = \frac{\sum_{n=1}^{N} z_{nj}}{N}, \quad (6)$$

$$\mathbf{w}_j = \Big(\sum_{n=1}^{N} z_{nj}\mathbf{X}^T\mathbf{X}\Big)^{-1}\mathbf{X}^T\sum_{n=1}^{N}(z_{nj}\mathbf{y}_n), \quad (7)$$

$$\sigma_j^2 = \frac{\sum_{n=1}^{N} z_{nj}\|\mathbf{y}_n - \mathbf{X}\mathbf{w}_j\|^2}{T\sum_{n=1}^{N} z_{nj}}. \quad (8)$$

After the convergence of the EM algorithm, each sequence $\mathbf{y}_n$ is assigned to the cluster with the maximum posterior.

III. Regression mixture modeling with spatial and sparse properties

The above structure of the regression mixture model for clustering fMRI data has some limitations, since it does not capture some important features arisen from the nature of the observations. In particular, the fMRI data are structures that involve spatial properties, where adjacent voxels tend to have similar activity behavior [32]. Furthermore, there are temporal correlations which are derived from neural, physiological and physical sources [5]. These are physical constraints that must be incorporated to the model. This can be carried out by introducing appropriate priors.

A. Basic features of the proposed model

In particular, we can treat the probabilities (voxel labels) π_{nj} of each fMRI sequence $\mathbf{y}_n$ belongs to the j-th cluster as random variables, which additionally satisfy the constraints $\pi_{nj} \geq 0$ and $\sum_{j=1}^{K}\pi_{nj} = 1$. Local characteristics of voxels can be employed through the use of Markov Random Fields (MRFs) that have successfully applied to computer vision applications [29], [33]. More specifically, we assume that the set of voxel labels $\Pi = \{\boldsymbol{\pi}_n\}_{n=1}^{N}$ follows the Gibbs prior distribution with density [29]

$$p(\Pi) = \frac{1}{Z}\exp\{-\sum_{n=1}^{N} V_{N_n}(\Pi)\}\ . \quad (9)$$

The function $V_{N_n}(\Pi)$ denotes the clique potential function around the neighborhood N_n of the n-th voxel taking the following form:

$$V_{N_n}(\Pi) = \sum_{m\in N_n}\sum_{j=1}^{K}\beta_j(\pi_{nj} - \pi_{mj})^2. \quad (10)$$

In our case we consider neighborhood consisted of eight (8) voxels which are horizontally, diagonally and vertically adjacent. We also assume that every cluster has its own regularization parameter β_j. This has the ability to increase the flexibility of model, since it allows different degree of smoothness at each cluster. It is interesting to note here that in our model these regularization parameters belong to the set of the unknown parameters which are estimated during the learning process. Finally, the term Z of Eq. 9 is the normalizing factor that is analogous to $Z \propto \prod_{j=1}^{K}\beta_j^{-N}$.

An important role in using a regression model is how to estimate its order M, i.e. the size of linear regression coefficients $\mathbf{w}_j$. Estimating the proper value of M depends on the shape of data to be fitted, where models of small order may lead to underfitting, while large values of M may become responsible for data overfitting. This may deteriorate significantly the clustering performance. Bayesian regularization framework provides an elegant solution to this problem [12], [26]. It initially assumes a large value of order M. Then, a heavy tailed prior distribution $p(\mathbf{w}_j)$ is imposed upon the regression coefficients that will enforce most of the coefficients to be zero out after training.

The sparsity can be achieved in an hierarchical way [26] by considering first a zero-mean Gaussian distribution over the regression coefficients:

$$p(\mathbf{w}_j|\boldsymbol{\alpha}_j) = \mathcal{N}(\mathbf{w}_j|\mathbf{0}, A_j^{-1}) = \prod_{l=1}^{M}\mathcal{N}(w_{jl}|0,\alpha_{jl}^{-1})\ , \quad (11)$$

where A_j is a diagonal matrix containing the M components of the precision (inverse variance) vector $\boldsymbol{\alpha}_j = (a_{j1},\ldots,a_{jM})$. At a second level, precision can be seen as hyperparameters that follow a Gamma prior distribution:

$$p(\alpha_j) = \prod_{l=1}^{M}\Gamma(\alpha_{jl}|b,c) \propto \prod_{l=1}^{M}\alpha_{jl}^{b-1}\exp^{-c\alpha_{jl}}\ . \quad (12)$$

Note that both Gamma parameters b and c are a priori set to zero so as to achieve uninformative priors. The above two-stage hierarchical sparse prior is actually the Student's-t distribution enforcing most of the values α_{jl} to be large and thus eliminating the effect of the corresponding coefficients w_{jl} by setting to zero. In such way the regression model order for every cluster is automatically selected and overfitting is avoided.

As mentioned before, the construction of the design matrix $\mathbf{X}$ is a crucial part of the regression model. In our case we have adopted a multi-kernel scheme [34], [35] by considering a pool of S kernel matrices $\{\boldsymbol{\Phi}_s\}_{s=1}^{S}$ which varies in their scalar parameter value λ_s. In particular, we assume that the kernel matrix $\mathbf{X}_j$ for the $j-th$ cluster can be written as a linear combination of these S kernel matrices

$$\mathbf{X}_j = \sum_{s=1}^{S} u_{js}\boldsymbol{\Phi}_s\ , \quad (13)$$

where u_{js} are the coefficients of the multi-kernel scheme which are unknown and satisfy the constraints $u_{js} \geq 0$ and $\sum_{s=1}^{S} u_{js} = 1$. These parameters should be estimated during learning in order to construct the kernel design matrix that better suits to every cluster. As experiments have shown, the employance of the proposed multi-kernel scheme has the

ability to improve significantly the performance and the quality of the clustering procedure.

B. MAP Estimation

The incorporation of the above properties leads to a modification of the regression mixture model which is written as:

$$f(\mathbf{y}_n|\Theta) = \sum_{j=1}^{K} \pi_{nj} p(\mathbf{y}_n|\theta_j), \tag{14}$$

where $\Theta = \{\{\pi_{nj}\}_{n=1}^{N}, \theta_j = (\mathbf{w}_j, \boldsymbol{\alpha}_j, \sigma_j^2, \boldsymbol{u}_j, \beta_j)\}_{j=1}^{K}$ is the set of mixture model parameters. The clustering procedure becomes now a Maximum-A-Posteriori (MAP) estimation problem, where the MAP log-likelihood function is given by

$$\begin{aligned} l_{MAP}(\Theta) &= \log p(Y|\Theta) + \log p(\Theta) \\ &= \sum_{n=1}^{N} \log\{\sum_{j=1}^{K} \pi_{nj} p(\mathbf{y}_n|\theta_j)\} + \log p(\Pi) \\ &+ \sum_{j=1}^{K} \big\{ \log p(\mathbf{w}_j|\boldsymbol{\alpha}_j) + \log p(\boldsymbol{\alpha}_j) \big\} \ . \end{aligned} \tag{15}$$

Employing the EM algorithm to MAP estimation requires at each iteration the conditional expectation values z_{nj} of the hidden variables to be computed first (E-step):

$$z_{nj} = P(j|\mathbf{y}_n, \Theta) = \frac{\pi_{nj} p(\mathbf{y}_n|\theta_j)}{f(\mathbf{y}_n|\Theta)} \ . \tag{16}$$

During the M-step the maximization of the complete data MAP log-likelihood (Q-function) expectation is performed:

$$\begin{aligned} Q(\Theta|\Theta^{(t)}) &= \sum_{n=1}^{N}\sum_{j=1}^{K} z_{nj} \Big\{ \log \pi_{nj} - \frac{T}{2}\log 2\pi - \\ &T \log \sigma_j - \frac{\|\mathbf{y}_n - \mathbf{X}_j \mathbf{w}_j\|^2}{2\sigma_j^2} \Big\} - \\ &\sum_{j=1}^{K} \Big\{ - N \log \beta_j + \\ &\beta_j \sum_{n=1}^{N} \sum_{m \in N_n} (\pi_{nj} - \pi_{mj})^2 + \\ &\frac{1}{2} \mathbf{w}_j^T \mathbf{A}_j \mathbf{w}_j - \\ &\sum_{l=1}^{M} [(b-1)\log \alpha_{jl} - c\alpha_{jl}] \Big\} \ . \end{aligned} \tag{17}$$

By setting the partial derivatives of the above Q function with respect to label parameters π_{nj} equal to zero, we obtain the following quadratic equation:

$$\pi_{nj}^2 - < \pi_{nj} > \pi_{nj} - \frac{1}{2\beta_j |N_n|} z_{nj} = 0 \ , \tag{18}$$

where $|N_n|$ is the cardinality of the neighborhood N_n and $< \pi_{nj} >$ is the mean value of the j-th cluster's probabilities of the spatial neighbors of the n-th voxel, i.e. $< \pi_{nj} >= \frac{1}{|N_n|} \sum_{m \in \mathcal{N}_n} \pi_{mj}$. The above quadratic expression has two roots, where we select only the one with the positive sign since it yields $\pi_{nj} \geq 0$:

$$\pi_{nj} = \frac{< \pi_{nj} > + \sqrt{< \pi_{nj} >^2 + \frac{2}{\beta_j |N_n|} z_{nj}}}{2} \ . \tag{19}$$

Note that in the above update rule the neighborhood N_n may contain label parameters π_{mj} that have been either already updated or not. However, these values do not satisfy the constraints $0 \leq \pi_{nj} \leq 1$ and $\sum_{j=1}^{K} \pi_{nj} = 1$, and there is a need to project them on their constraint convex hull. For this purpose, we apply an efficient convex quadratic programming approach presented in [33], that is based on the active-set theory.

For the regression model parameters $\{\mathbf{w}_j, \boldsymbol{\alpha}_j, \sigma_j^2, \beta_j\}$ the update rules can be obtained as

$$\begin{aligned} \mathbf{w}_j &= \Big[\Big(\sum_{n=1}^{N} z_{nj} \Big) \frac{1}{\sigma_j^2} \mathbf{X}_j^T \mathbf{X}_j + \mathbf{A}_j \Big]^{-1} \cdot \\ &\frac{1}{\sigma_j^2} \mathbf{X}_j^T \Big(\sum_{n=1}^{N} z_{nj} \mathbf{y}_n \Big), \end{aligned} \tag{20}$$

$$\alpha_{jl} = \frac{1 + 2c}{w_{jl}^2 + 2b}, \tag{21}$$

$$\sigma_j^2 = \frac{\sum_{n=1}^{N} z_{nj} \|\mathbf{y}_n - \mathbf{X}_j \mathbf{w}_j\|^2}{T \sum_{n=1}^{N} z_{nj}}, \tag{22}$$

$$\beta_j = \frac{N}{\sum_{n=1}^{N} \sum_{m \in N_n} (\pi_{nj} - \pi_{mj})^2}. \tag{23}$$

Finally, the weights u_{js} of the multi-kernel scheme are adjusted by solving the following minimization problem, where we have considered only the part of likelihood function that involves $\mathbf{u}_j$:

$$\begin{aligned} &\min_{\mathbf{u}_j} \sum_{n=1}^{N} z_{nj} \parallel \mathbf{y}_n - \sum_{s=1}^{S} u_{js} \boldsymbol{\Phi}_s \mathbf{w}_j \parallel^2 = \\ &\min_{\mathbf{u}_j} \sum_{n=1}^{N} z_{nj} \parallel \mathbf{y}_n - \mathcal{X}_j \mathbf{u}_j \parallel^2 = \\ &\min_{\mathbf{u}_j} \Big\{ \mathbf{u}_j^T \mathcal{X}_j^T \mathcal{X}_j \mathbf{u}_j - 2\mathbf{u}_j^T \mathcal{X}_j^T \frac{\sum_{n=1}^{N} z_{nj} \mathbf{y}_n}{\sum_{n=1}^{N} z_{nj}} \Big\}, \\ &\text{s.t.} \ \sum_{s=1}^{S} u_{js} = 1 \text{ and } u_{js} \geq 0 \ . \end{aligned} \tag{24}$$

In the above formulation, the matrix $\mathcal{X}_j$ has S columns calculated by $\boldsymbol{\Phi}_s \mathbf{w}_j$, i.e. $\mathcal{X}_j = [\boldsymbol{\Phi}_1 \mathbf{w}_j \ \ \boldsymbol{\Phi}_2 \mathbf{w}_j \ \cdots \ \boldsymbol{\Phi}_S \mathbf{w}_j]$. The minimization problem described in Eq. 24 is a typical constrained linear least-squared problem that can be easily solved again with the active-set theory [36].

At the end of the learning process the activation map of the brain is constructed with the following manner: Initially, we select the cluster h that best match with the BOLD signal $\boldsymbol{\xi}$ (which is known before the data analysis) among the K mixture components. This is done according to the

Pearson correlation measurement (cosine similarity) between the estimated mean curve $\boldsymbol{\mu}_j = \mathbf{X}_j \mathbf{w}_j$ of each cluster with the BOLD signal $\boldsymbol{\xi}$, i.e.

$$h = \arg\max_{j=1}^{K} \frac{\boldsymbol{\mu}_j^T \boldsymbol{\xi}}{|\boldsymbol{\mu}_j||\boldsymbol{\xi}|} \ . \tag{25}$$

Then, the voxels that belong to cluster h determine the brain activation region, while the rest voxels (that belong to all other $K-1$ clusters) correspond to the non-activation region. In this way we create a binary image with activated and non-activated pixels. Alternatively, we can obtain a scalar activation map as follows: First we select a color for showing the brain activation (e.g. white). Each cluster is then drawn with this color weighted by the ratio between its correlation measurement (with the BOLD signal) and the correlation of the cluster h (Eq. 25), that has been estimated as the brain activation region. In this way, we create an activation map with K levels of activation.

A necessary observation must be made here about the computational cost of our method. Indeed, the computational effort goes in the estimation of the label parameters $\{\pi_{nj}\}$ in the M-step, since they must be examined sequentially according to the neighborhood system. Also, a constrained optimization problem is required to be solved for each voxel. Although, it is typically expensive to account for the data spatial dependencies during inference/learning, this is necessary in order to capture these properties of data. In addition, it provides with a flexibility of the proposed method to find optimum solutions due to the optimization procedure (it projects the unique unconstrained maximum onto the constraint boundary) which implies better clustering performance. However, it is our intension to work further on reducing the complexity (possibly) by adopting some approximations [37]. Finally, it must be noted that a way for speeding up the convergence of the method is to repeatedly execute the updating procedure for the label parameters in the same M-step. As experiments have shown, this causes to require less number of EM iterations.

C. Incremental mixture learning

A drawback of the EM algorithm is its sensitivity to the initialization of the model parameters due to its local nature. Improper initialization may lead to poor local maxima of the log-likelihood that sequentially affects the quality of the clustering solution. The most commonly used initialization strategy follows a random selection of K sequences among the input set Y. Then, the regression coefficients $\mathbf{w}_j$ are initialized according to the least-square fit rule, while the other model parameters are calculated using Eqs. 21-24. Finally, the log-likelihood value is calculated after performing one step of the EM algorithm. Several trials (e.g. 100) of such one-EM-step procedure are made and finally the solution with the maximum log-likelihood value is selected for initializing the model parameters.

In our study we have adopted a more advanced methodology based on an incremental strategy that has been successfully applied in Gaussian mixture models [38]–[40]. Our approach iteratively adds a new component to the mixture by performing a component splitting procedure. Initially, we start with a model having a single linear regression model. Let now assume that we have already constructed a mixture model f_k with k linear regression components, i.e.

$$f_k(\mathbf{y}_n|\Theta_k) = \sum_{j=1}^{k} \pi_j p(\mathbf{y}_n|\theta_j) \ . \tag{26}$$

Then we select next a cluster j^* for splitting among the k components. This is done by finding the mean curve $\boldsymbol{\mu}_j$ that is more similar with the BOLD signal $\boldsymbol{\xi}$, according to the cosine similarity function. Intuitively thinking, the splitting procedure can be seen as a pruning mechanism that is repeated until found the cluster that best describes the BOLD effect. Thus, we can use this relation as a stopping criterion, and also as a model order selection rule.

Let $f(\mathbf{y}_n|\Theta_k^{-j^*})$ be the mixture without the j^*-th component. A new component $k+1$ is generated and the resulting mixture after the split operation takes the following form:

$$\begin{aligned} f(\mathbf{y}_n|\Theta_{k+1}) &= f(\mathbf{y}_n|\Theta_k^{-j^*}) + \pi_{n,j^*}{}^{new} p(\mathbf{y}_n|\theta_{j^*}) \\ &+ \pi_{n,k+1} p(\mathbf{y}_n|\theta_{k+1}) \end{aligned} \tag{27}$$

For initializing the parameters of the new added regression model we follow the next strategy:

- Among the time series that currently belong to the selected cluster j^*, find a small percentage (referred to as r) of the less probable cases (outliers) and calculate their mean value $\overline{y_*}$.
- Initialize the regression coefficients $\mathbf{w}_{k+1}$ according to the least-square fit rule over $\overline{y_*}$, as well as its regularization parameter and noise variance (Eqs. 21 and 22, respectively). The kernel weights of the design matrix are set as $u_{k+1,s} = 1/S, \ \forall \ s = 1, \ldots, S$.
- Also, the label parameters are initialized as

$$\pi_{n,k+1} = \pi_{n,j^*}{}^{new} = \frac{\pi_{n,j^*}{}^{old}}{2}$$

The EM algorithm can be applied next for estimating the parameters Θ_{k+1} of the new mixture model. For terminating the procedure we have used the criterion of the percentage of the correlation (with the BOLD signal) increase between two successive steps. When this percentage becomes very small the incremental training process is terminated. In this case the mixture increment from Θ_k to Θ_{k+1} does not offer any significant improvement to the correlation criterion, and thus the best found cluster h (Eq. 25) from the previous step is the final solution. The algorithmic description of the proposed incremental strategy for training the adaptive regression mixture model is given in **Algorithm** 1.

IV. Experimental results

The performance of the proposed method was studied in cases with simulated and real fMRI data. We have studied both versions of our method: the random initialization (SSRM) and the incremental learning version (iSSRM). In all experiments, the multi-kernel scheme for the design matrix was constructed using a mixture of $S = 10$ design matrices $\boldsymbol{\Phi}_s$ by considering 10 different values for the scalar parameter λ_s, varying from

Algorithm 1 Incremental mixture learning

1: Input : set of N fMRI time-series $\mathbf{Y} = \{\mathbf{y}_1, \mathbf{y}_2, \cdots, \mathbf{y}_N\}$, BOLD signal $\boldsymbol{\xi}$ and S kernel parameter values λ_s.
2: Start with $k = 1$, and calculate correlation c_1.
3: **repeat**
4: Select cluster j^* for splitting.
5: Initialize parameters θ_{k+1} of the new added component $k+1$.
6: $\Theta_{k+1} = \Theta_k \cup \theta_{k+1}$. Apply EM algorithm to the new mixture $f_{k+1}(\Theta_{k+1})$.
7: Calculate the maximum correlation among the $k+1$ components mean curves with the BOLD signal $\boldsymbol{\xi}$,

$$c_{k+1} = \arg\max_{j=1}^{k+1} \frac{\boldsymbol{\mu}_j^T \boldsymbol{\xi}}{|\boldsymbol{\mu}_j||\boldsymbol{\xi}|} \ .$$

8: $k = k + 1$
9: **until** $k = K_{max}$ or $\frac{c_k - c_{k-1}}{c_{k-1}} < \epsilon$
10: Output: Binary activation map

0.1 to 2 with step 0.2. Also, in these matrices a column is added which contains the BOLD signal. Note that the time instances x_l were normalized before to $[0, 1]$. We have compared our method with the following approaches:

- SRM: the regression mixture model with only spatial properties (non-sparse regression modeling).
- GLM: the generalized linear model which is a known model-based approach [3] that performs a voxel by voxel analysis and a t-statistic decision process. In all experiments the threshold for the p-values was set to $p = 0.01$, since it gave the best performance. For constructing the design matrix of the GLM we have followed a common approach [41] that uses the BOLD signal, as well as the Discrete Cosine Transform (DCT) basis functions to capture the slow varying components of the fMRI time series (we have selected the first ten) .
- k-means: a well known vector-based clustering method.

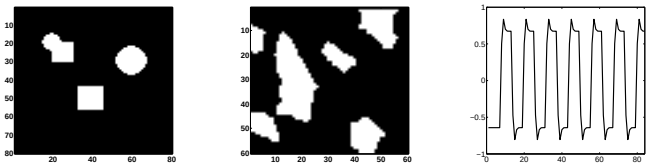
All experiments have been performed using Matlab (Mathworks, Inc.) on a laptop PC with CPU Intel Dual Core at 1.60 GHz and 2 GB RAM.

A. Experiments using simulated fMRI data

During the experiments with simulated fMRI data, we have created 3-D datasets of time series using linear regression models with known design matrix and regression coefficients. In these time series, we have added white Gaussian noise of various SNR levels (as defined between the BOLD signal and the white gaussian noise component of the model). Note that for the calculation of the SNR level we have used the next formula:

$$SNR = 10 \log_{10} \left(\frac{\mathbf{s}^T \mathbf{s}}{N\sigma^2} \right) \ ,$$

where σ^2 is the variance of the noise and $\mathbf{s}$ is the BOLD signal. The spatial correlation between the time series is achieved through the regression coefficients. The spatial patterns, that we have used, are drawn in Fig. 1a (dataset1) and Fig. 1b (dataset2). The BOLD signal, which is used to model the neural activity, is shown in Fig. 1c. Also, in these time series we have added a slow varying component to model the drift in the fMRI time series. This is done by using a linear regression model where the regressors are the first ten basis vector of DCT basis and the regression coefficients are sampled from a $\mathcal{N}(0, 1)$. The size of both datasets were 80x80x84 (dataset1) and 60x60x84 (dataset2). Finally, for each SNR level we studied the performance of the comparative methods by executing 50 Monte Carlo simulations, where we took the statistics of those runs (mean and variance).



(a) (b) (c)

Fig. 1. Two spatial patterns (a), (b) and the BOLD signal (c) used in our experiments with simulated data

To measure the quality of each clustering approach, we have used two evaluation criteria:

- the performance, calculated as the percentage of correctly classified time series, after labeling the cluster h (Eq .25) as the brain activation area and the rest voxels as the non-activation region, and
- the normalized mutual information (NMI), which is an information theoretic measure based on the mutual information between the true (Ω) and the estimated ($\mathcal{C}$) labeling normalized by their entropies:

$$NMI(\Omega, \mathcal{C}) = \frac{I(\Omega, \mathcal{C})}{(H(\Omega) + H(\mathcal{C}))/2} \ , \qquad (28)$$

where

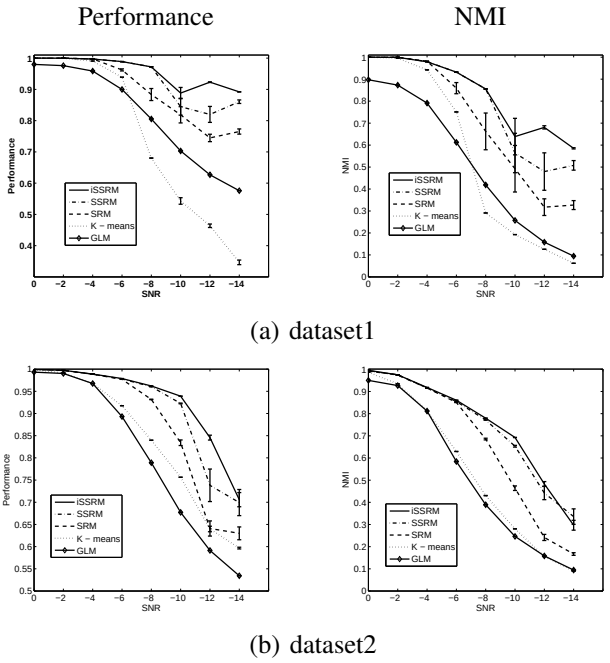
$$I(\Omega, \mathcal{C}) = \sum_k \sum_j P(\omega_k, c_j) \log \frac{P(\omega_k, c_j)}{P(\omega_k) P(c_j)} \ ,$$

$$H(\Omega) = -\sum_k P(\omega_k) \log P(\omega_k),$$

$$H(\mathcal{C}) = -\sum_k P(c_k) \log P(c_k) \ .$$

The quantities $P(\omega_k)$, $P(c_j)$ and $P(\omega_k, c_j)$ are the probabilities of a sequence being in class ω_k, cluster c_j and in their intersection, respectively, and are computed based on set cardinalities (frequencies).

Figure 2 shows the comparative results for both simulated datasets of Fig. 1. The superiority of the iSSRM is obvious based on two evaluation criteria, especially in small SNR values (noisy problems). Between two versions of our approach (incremental vs. random sampling), we observe that the iSSRM gave better results with much less variability. Comparison with the SRM method that holds only the spatial properties, has shown a significant improvement in terms of both evaluation criteria. This proves the usefulness of the



(a) dataset1

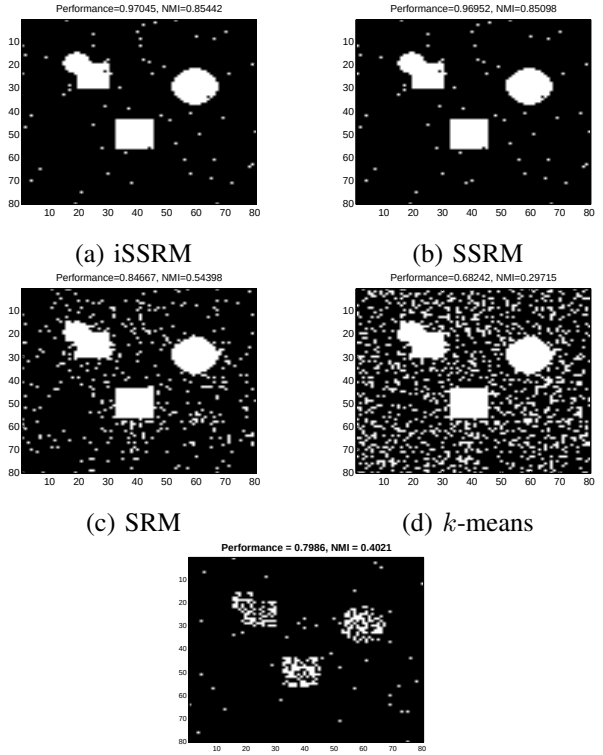
(b) dataset2

Fig. 2. Comparative results in the case of two datasets of Fig. 1. Error bars for the two evaluation criteria are shown in terms of several SNR values.

sparse term to modeling procedure. Among all regression mixture models, the GLM approach gave the worst results. This is in agreement with our belief that the voxel by voxel analysis as made by the GLM, is not able to correctly capture the structure of data [14]. Another drawback of the GLM approach is its great sensitivity to the choice of threshold on the t-statistic.

Two examples of the activation maps as estimated by each method are shown in Figs. 3,4 in the case of $SNR = -8$ dB for the two datasets, respectively. Clearly, our method had better discrimination ability and achieved to discover more accurately the original spatial pattern, while at the same time reduced significantly the false negative activation cases. This statement can be further justified by calculating the False Positive Rate (FPR) and True Positive Rate (TPR) (Table I) from all methods. Especially for the GLM approach we also present in Fig. 5 the ROC curve that plots TPR vs. FPR, where we also present (in parentheses) the performance for some characteristic cases. According to the results there is a great sensitivity of GLM to the choice of threshold. For example, in Fig. 2 (b) the performance of the GLM was 0.77 in the case of -8 dB that corresponds to threshold $p = 0.01$. Although, there are other p-values with better performance (e.g. 0.05 with performance 0.89), as shown in ROC curve of Fig. 5, the obtained performance curve over all SNR levels was worst that this presented in Fig. 2 (b). As mentioned before, we have tested many p-values and we have shown that the value of 0.01 had better overall performance.

We have also studied the effect of the proposed multi-kernel scheme to the quality of clustering solution. To do this we have compared the proposed multi-kernel iSSRM with the best found single-kernel iSSRM, referred to as iSSRM*. The latter was established by finding the regression mixture having a

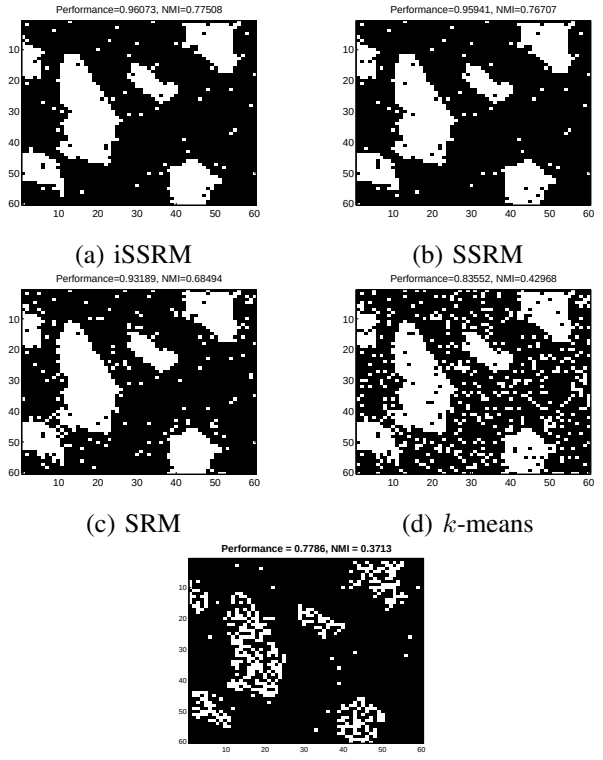


(a) iSSRM (b) SSRM

(c) SRM (d) k-means

(e) GLM

Fig. 3. Spatial patterns for dataset1 as estimated by all methods in the case of -8 dB.



(a) iSSRM (b) SSRM

(c) SRM (d) k-means

(e) GLM

Fig. 4. Spatial patterns for dataset2 as estimated by all methods in the case of -8 dB.

TABLE I
THE TRUE POSITIVE RATES (TPR) AND FALSE POSITIVE RATES (FPR) AS CALCULATED BY ALL METHODS OBTAINED FROM THE RESULTS OF FIG. 4.

	FPR	TPR
iSSRM	0.02	0.96
SSRM	0.02	0.95
GLM	0.01	0.49
SRM	0.06	0.97
K-means	0.16	0.94

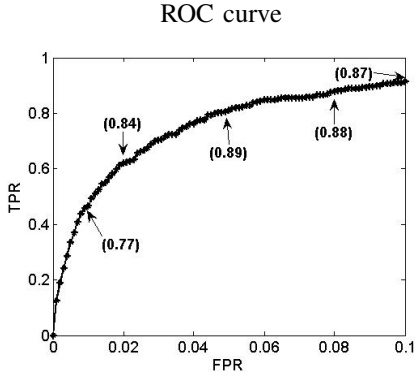


Fig. 5. ROC curve of the GLM method in the case of SNR=-8dB (Fig. 4 (e)).

single common design matrix with the best kernel parameter λ^*, among the $S = 10$ design matrices. In Table II we present the results for the two compared methods (iSSRM and iSSRM*) in the case of dataset1 (Fig. 1(a)). As we can see our approach compares favorably in high SNR values. However, in noisy environments the iSSRM method has shown better performance in terms of both evaluation criteria, where the difference is significant in some cases. An important advantage of the proposed multi-kernel scheme is that it improves the fitting data process, since it produces better solutions with more adaptive clusters having its own regression coefficients and design matrix (not common).

Furthermore, we have studied the behavior of our method using other type of noise (non-Gaussian) obtained from real fMRI data. More specifically, we have selected some slices from a real fMRI experiment (auditory - it will be described later) and we have applied them to the SPM package for detecting the regions with no activation (background). Then, from the corresponding time-series we had set up a set of noise samples and we have used them for determining the noise term in the linear Equation 1 by random selection.

TABLE II
THE EFFECT OF MUTLI - KERNEL SCHEME TO OUR METHOD.

	Performance		*NMI*	
SNR	iSSRM	iSSRM* (λ^*)	iSSRM	iSSRM* (λ^*)
0	1.0000	1.0000 (0.1)	1.0000	1.0000 (0.1)
-2	0.9997	0.9998 (1.1)	0.9972	0.9981 (1.1)
-4	0.9982	0.9983 (0.1)	0.9857	0.9864 (0.1)
-6	0.9895	0.9895 (1.3)	0.9358	0.9360 (1.3)
-8	0.9719	0.9656 (1.9)	0.8561	0.8360 (1.9)
-10	0.9491	0.8910 (0.1)	0.7682	0.6357 (0.1)
-12	0.9372	0.9247 (0.3)	0.7133	0.6726 (0.1)
-14	0.9083	0.8891 (0.1)	0.6226	0.5609 (0.1)

TABLE III
COMPARATIVE RESULTS USING NOISE FROM REAL DATA. MEAN VALUES AND STANDARD DEVIATIONS ARE SHOWN OBTAINED FROM 30 TRIALS

	Dataset2	
	Performance	NMI
iSSRM	0.9783 $\pm$ 0.0031	0.8610 $\pm$ 0.0156
SSRM	0.9772 $\pm$ 0.0175	0.8665 $\pm$ 0.0644
SRM	0.7541 $\pm$ 0.1649	0.4330 $\pm$ 0.3048
GLM	0.9464 $\pm$ 0.0042	0.7558 $\pm$ 0.0137

Table III presents the comparative results of all methods after performing 30 Monte Carlo runs, where we took better results with the proposed approach. This constitutes the flexibility of our method to attain its performance in non-Gaussian noise. However, assuming alternative noise distribution (e.g. Student's-t) could be an interesting direction for future work.

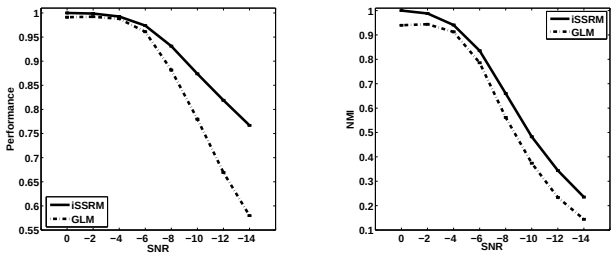
Finally, we have tested the behavior of our method when applying spatial smoothing as a preprocessing step. Experiments have been made using the simulated fMRI dataset of Fig. 1 (b) (dataset2). In particular, we have employed a spatially stationary Gaussian filter for various values of full width at half maximum (FWHM), while comparison has been made with the GLM approach. The depicted results are shown in Fig. 6 for three characteristic FWHM values. As we can see by comparing curves of Fig. 6 (b),(c) with those of Fig. 2 (b), the smoothing process improves the performance of both methods in large noisy environments. However, the GLM fails to accurately discover the true activation areas in cases with high SNR values. This behavior can be explained by the fact that large smoothing blurs activations, leading to a biased estimate of the location of activation peaks [4]. And this is more obvious in low noisy environments, while in cases with high noise this oversmoothing effect is eliminated. On the other hand, the iSSRM method is more consistent incorporating an adaptive smoothing procedure that can better preserve the shape of the active regions. Also, it is interesting to note that our method did not show any significant sensitivity to the choice of FWHM value (we took similar results with greater FWHM values). Note that we have performed spatially smoothing during the experiments with real fMRI data as will be presented later.

B. Experiments using real fMRI data

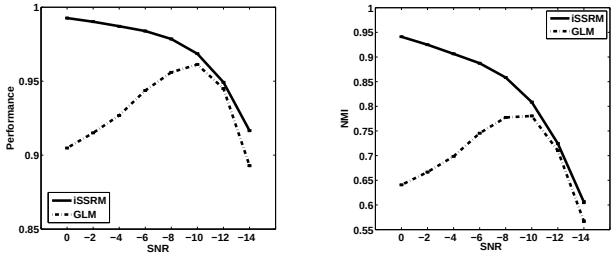
We have made additional experiments using real fMRI data. In our study, we have selected three datasets: a block-designed auditory paradigm, an event-related foot movement paradigm and a block-designed hand movement paradigm. In these experiments, we have followed the standard preprocessing steps of the SPM package, i.e. realignment, segmentation, normalization and spatial smoothing. Data are then scaled by using the global mean value of all time series as a factor. Finally, every time series was high pass filtered using the standard methodology of the SPM package, where the default cut off frequency was 0.008 Hz. The BOLD signals for all experiments are shown in Fig. 7.

At first we have studied a block-designed fMRI dataset[1]

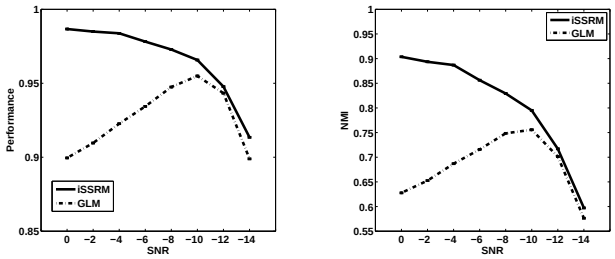
[1]It was downloaded from the SPM web page http://www.fil.ion.ucl.ac.uk/spm/



(a) FWHM=1

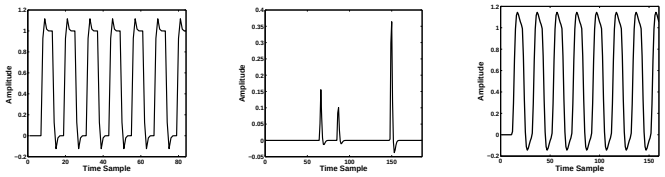


(b) FWHM=3



(c) FWHM=6

Fig. 6. Comparative results in the case of dataset2 where the time series have been spatially smoothed.



(a) (b) (c)

Fig. 7. BOLD signals for the (a) auditory, (b) motor event and (c) motor block experiments.

based on an auditory processing task as executed by a healthy volunteer. Its functional images consisted of $\mathcal{M} = 68$ slices ($79 \times 95 \times 68$, $2mm \times 2mm \times 2mm$ voxels). For the auditory data, the design matrix of GLM contains a columns of 1's and the BOLD signal as indicated in [24]. Experiments were made with the slice 29 of this dataset, which contains a number of $N = 5118$ time series. First, we have applied the iSSRM algorithm for estimating the proper number of clusters ($K = 5$) and use this value for executing the other two approaches. Note that the required computing time was around 1 min. Figure 8 represents the comparative results of all clustering methods giving the resulting position of the activation area inside the brain. Note that the activated areas

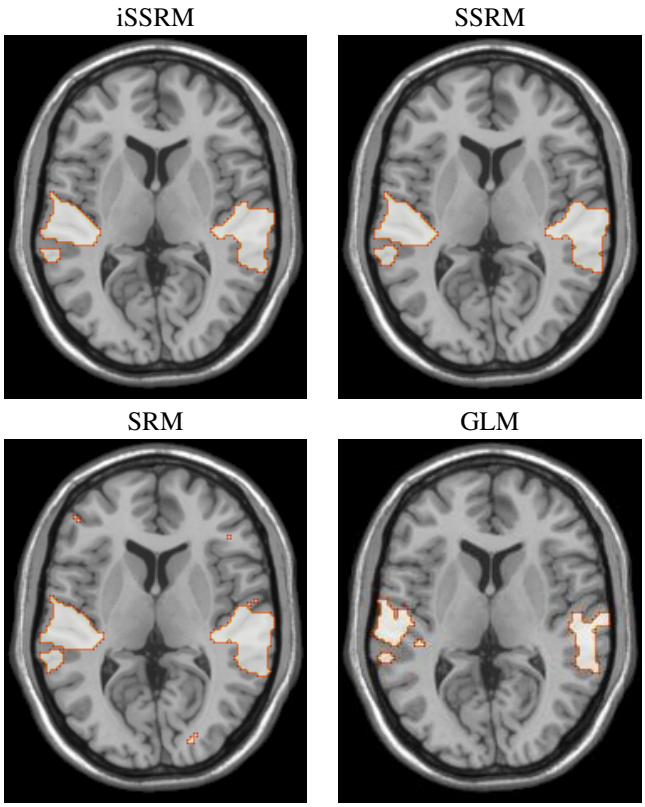


Fig. 8. The binary activation map as estimated by each method in the case of the auditory experiment.

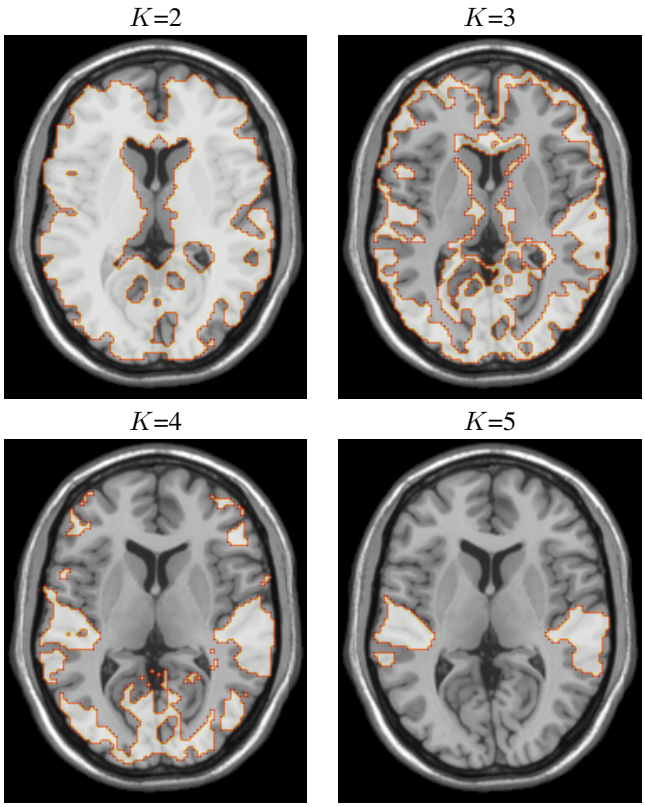


Fig. 9. The progress of incremental training procedure in the auditory experiment. The splitting stops when found $K = 5$ clusters according to the correlation criterion.

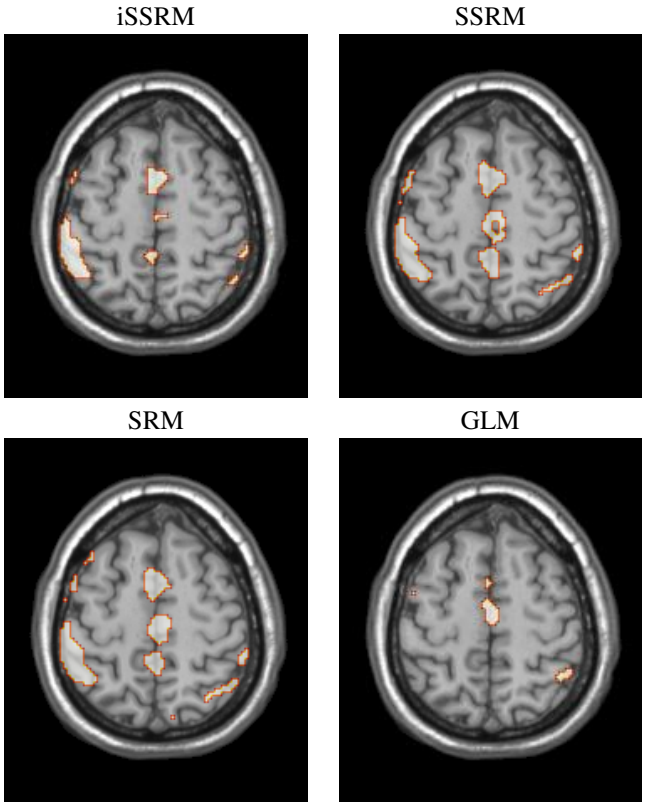


Fig. 10. Estimated motor activated areas of comparative methods in white overlaid on greyscale T1 weighted anatomical images.

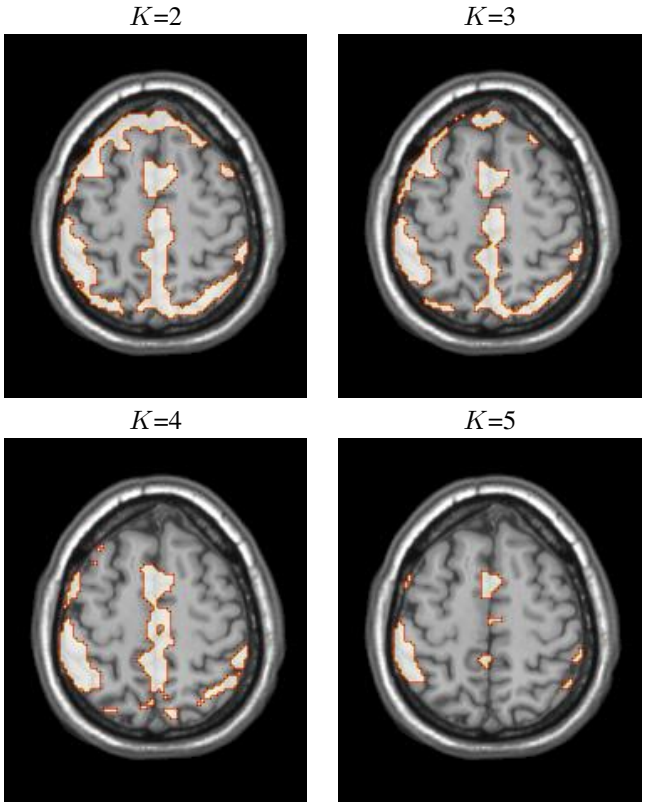


Fig. 11. The progress of incremental training procedure in the motor event experiment. The splitting stops when found $K = 5$ clusters according to the correlation criterion.

are overlaid on greyscale T1 weighted anatomical images. All methods have detected the auditory cortex as the brain activation area. Both iSSRM and SSRM methods have clearly detected three distinct areas of activation. These can be also seen in the maps produced by the rest two approaches, SRM and GLM, where in addition they include other small activated islands that contain only a few voxels. It is also obvious that the three regression mixture models with spatial properties produce larger and smoother activation areas in comparison with the GLM. The progress of incremental training procedure of regression mixture model is also illustrated in Fig. 9. Starting with a large activation area when $K = 2$, the proposed method reaches the expected area of the brain when $K = 5$. For larger value of K ($K = 6$) the increase of the correlation measurement with the BOLD signal is not significant and thus the algorithm terminates.

In the event-related foot-movement experiment we analyzed fMRI data consisted of images acquired from the University Hospital of Ioannina, Greece [42]. The imaging protocol consisted of the following: 1) a T1-weighted high-resolution (0.86x0.86x1 mm) 3D spoiled gradient echo sequence (TR/TE, 25/4.6 ms), which was used for structural imaging; 2) A single-shot multisection gradient EPI was used for BOLD functional images (TR/TE= 3000/50 ms; flip angle= 40; matrix= 64x64; section thickness= 5 mm; gap= 0 mm). Each fMRI session consisted of 160 scans and lasted 480 seconds. At the beginning of each session, 4 dummy scans were acquired to allow equilibration of magnetization. The head of the subject was restrained by using cushions to minimize motion artifacts, and he was advised to keep his eyes closed during the examination so as to minimize potential visual stimuli. During this experiment the subject, suffering from the restless legs syndrome, performed random and spontaneous limb movements evoked by sensory leg uneasiness. These movements were used to create the indicator vector in our modeling that was convolved next with the hemodynamic response function (HRF) in order to provide the BOLD signal. Experiments were made with the slice 54 of this dataset, which contains a number of $N = 2644$ time series. In this case the required computational time of our method was around two minutes[2].

Figure 10 presents the comparative results in this dataset overlaid on greyscale T1 weighted anatomical images. Note that, in the case of the GLM method the design matrix has two extra columns about the time and the dispersion derivatives. As expected, all methods have detected the primary and the supplementary motor areas of the brain as the activation cluster. Although there is no ground truth for the fMRI data on individual cases the motor system in general is well studied and described in the literature. Contrary to the GLM methodology, our approach gives more activated areas closer to the established motor circuitry and therefore the results are more reasonable at least in this case. In particular, the GLM does not consider the premotor region (large island down

[2]The difference on computing time between the two real experiments is mainly due to larger design matrix on event-related foot-movement experiment

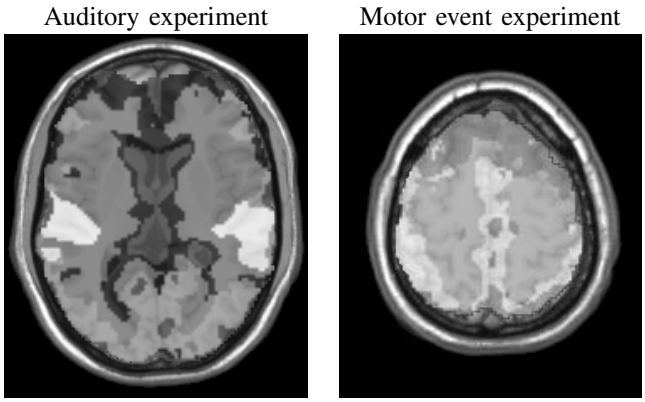


Fig. 12. The obtained scalar activation map in the case of the auditory and motor experiments used in our study.

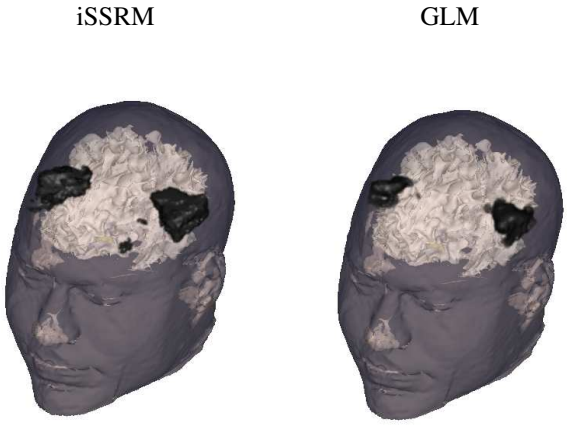


Fig. 13. The 3D head activation maps as estimated by the proposed iSSRM and the GLM approaches.

left) as activation area. More studies on large groups with many different cognitive function are needed to prove these preliminary results. Also, Fig. 11 illustrates the progress of the incremental training of iSSRM.

The above study with real data allows us to make a useful observation. In general, the size of the activation area is much smaller in comparison with the rest non-activated area. This fact may significantly deteriorate the performance of clustering. To face this problem, most approaches usually perform a preprocessing step [14], [43] where the number of voxels is reduced and the cluster size imbalance effect is eliminated. In our case this is not needed due to the proposed incremental learning scheme that repeatedly performs a splitting procedure.

Figure 12 illustrates the scalar activation maps estimated by our method for the two real fMRI experiments. Each map shows the image segmentation result according to the level of activation per cluster, where the white-color cluster is the most activated. According to these results, it turns to be clearer the decision about the brain activation area in the case of the auditory, since the white-color cluster (that best fit the BOLD signal) differs significantly from the rest. On the other hand, in the case of motor event experiment there is also another cluster (with bright grey-color) which is significantly similar with the BOLD signal and surrounds the white-color cluster (activated). This is not far from the physiological properties of the experiments since this second in order most activated cluster covers parts of motor cortex area. We believe that this scalar activation map may be proved a valuable tool for the decision making procedure especially in some difficult cases as they provide with more useful information.

Furthermore, we have studied the capability of our method to construct the 3D activation model. In particular we have applied our method independently to all available slices (68) of the auditory experiment. The resulting activation maps are fed to the 3D Slicer toolkit [44] where it has produced the 3D head model that contains the activation areas. Figure 13 illustrates the resulting 3D models of the proposed iSSRM and the GLM approaches. As shown both methods have detected a significant activation on the temporal lobe. However our method have detected an extra activated region into the frontal lobe which is expected to auditory experiments.

Finally, we made experiments on a group of subjects performed hand movement. This fMRI dataset was acquired at the University Hospital of Ioannina, with a clinical magnet 1.5T (Philips Intera) equipped with fast gradients. The imaging protocol consisted of: i) Axial T1-Weighted: a 3-dimensional sequence which provided a high-resolution reference for alignment between subjects. We used a 3D-Volume Magnetization-Prepared Rapid Acquisition with Gradient Echoes (MP-RAGE). Parameters: TR=25 msec; TE=4,6 msec; flip angle=30; slice thickness=0,8 mm; acquisition matrix =256 ; reconstruction matrix =256. ii) Multilevel BOLD fMRI using a hand task: We used a gradient echo EPI. Parameters: TR= 2000 ms; TE= 50 ms; flip angle=40; acquisition matrix =64; reconstruction matrix =128; slice thickness =5 mm; gap=0; 160 dynamics. Four dummy scans were performed in the beginning of fMRI to stabilize the magnetization level and allow us to keep all the images. This paradigm consisted of alternate action and resting blocks, 20 sec each (block design). During the action epoch the subjects flexed and released continuously at 0.5 Hz rate the fingers of their right hand in unison. The cues for action or rest was announced through headphones using the commands start and stop. The paradigm lasted 320s of 8 action and 8 rest blocks, and we collected 20 images/slice at TR=2s. We restricted motion artifacts by using foam rubber pads and strap across the forehead. Eyes were kept closed, at all times. More specifically, the data from four subjects were used to test our method. We applied the iSSRM method in each subject and the produced activation maps are shown in Fig. 14. In these experiments the slice 54 of each individual dataset was used. In all subjects we consistently found brain activation at expected areas comprising the cortical motor network (i.e. motor, premotor and supplementary motor regions). Note that the variability of motor activation among the subjects in Fig. 14, is a well known feature of brain activation in general and it is manifested even among successive studies of the same subject [45]. Deducing statistical inferences about the average activation of the whole group although is not a demanding task requires a second level group-analysis which lies beyond the scope of the present study.

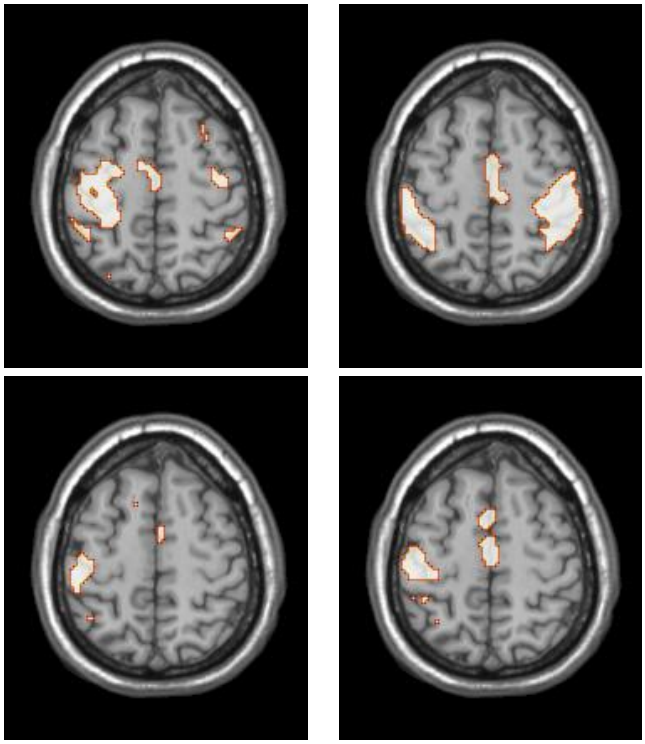


Fig. 14. Results of the application of our method to a group of subjects performed hand movement. These are the activation maps of four such subjects.

V. Conclusions

In this work, a clustering technique, based on probabilistic mixture modelling, is presented for the analysis of fMRI data. More specifically, our method was used to cluster the fMRI time series into two groups, activated and non activated. This is achieved by introducing a new mixture of linear regression models with sparse and spatial properties. Sparse priors are placed on the weights of each linear regression model helping us to deal with the problem of model order selection. Also, spatial priors are used on the mixing coefficients to take into account the spatial correlation between the voxels. This is achieved by using a Gibbs distribution. Furthermore, to avoid sensitivity of the design matrix to the choice of kernel matrix, we have used a kernel composite design matrix constructed as linear combination of Gaussian kernel matrices with different scaling parameter. Finally, an incremental strategy was proposed to find the number of clusters as well as to avoid the sensitivity of the EM algorithm on the initialization.

Our future research study will be focused on examining the appropriateness of other types of sparse priors [46] and alternative Gibbs potential functions, as well as to study different strategies for estimating kernel parameters of the design matrix. Another direction is to apply our method to other fMRI related problems such the study of functional connectivity [47] and the analysis of resting state fMRI data [17]. During the experiments we have shown that a preprocessing step of spatial smoothing may enhance considerably the performance of the proposed method. Therefore, studying the effect of other preprocessing techniques such as those described in [48], constitutes an interesting issue for future work. Finally, a possible extension is to use alternative stopping criteria in the incremental training scheme, especially when the experimental paradigm is not given [48].

References

[1] P. Jezzard, P. M. Matthews, and S. M. Smith, *Functional MRI: An Introduction to Methods*. Oxford University Press, USA, 2001.

[2] R. Frackowiak, J. Ashburner, W. Penny, S. Zeki, K. Friston, C. Frith, R. Dolan, and C. Price, *Human Brain Function, Second Edition*. Elsevier Science, USA, 2004.

[3] K. J. Friston, "Analysis of fMRI time series revisited," *Neuroimage*, vol. 2, pp. 45–53, 1995.

[4] G. Flandin and W. Penny, "Bayesian fMRI data analysis with sparse spatial basis function priors," *NeuroImage*, vol. 34, pp. 1108–1125, 2007.

[5] W. Penny, N. Trujillo-Barreto, and K. Friston, "Bayesian fMRI time series analysis with spatial priors," *NeuroImage*, vol. 24, pp. 350–362, Jan. 2005.

[6] R. Baumgartner, L. Ryner, W. Richter, R. Summers, M. Jarmasz, and R. Somorjai, "Comparison of two exploratory data analysis methods for fMRI: fuzzy clustering vs. principal component analysis," *Magnetic Resonance Imaging*, vol. 18, no. 1, pp. 89 – 94, 2000.

[7] M. J. Mckeown, S. Makeig, G. G. Brown, T. P. Jung, S. S. Kindermann, A. J. Bell, and T. J. Sejnowski, "Analysis of fMRI data by blind separation into independent spatial components," *Human Brain Mapping*, vol. 6, no. 3, pp. 160–188, 1998.

[8] A. Meyer-Baese, A. Wismueller, and O. Lange, "Comparison of two exploratory data analysis methods for fMRI: unsupervised clustering versus independent component analysis," *IEEE Transactions on Information Technology in Biomedicine,*, vol. 8, pp. 387 –398, Sept. 2004.

[9] A. Meyer-Base, A. Saalbach, O. Lange, and A. Wismller, "Unsupervised clustering of fMRI and MRI time series," *Biomedical Signal Processing and Control*, vol. 2, no. 4, pp. 295 – 310, 2007.

[10] C. G. Laberge, A. Adler, I. Cameron, T. Nguyen, and M. Hogan, "A Bayesian Hierarchical Correlation Model for fMRI Cluster Analysis," *IEEE Transactions on Biomedical Engineering,*, at press.

[11] M. D'Esposito, L. Y. Deouell, and A. Gazzaley, "Alterations in the BOLD fMRI signal with ageing and disease: a challenge for neuroimaging," *Nature Reviews Neuroscience*, vol. 4, pp. 863–872, Nov. 2003.

[12] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2006.

[13] R. Baumgartner, C. Windischberger, and E. Moser, "Quantification in Functional Magnetic Resonance Imaging: Fuzzy Clustering vs. Correlation Analysis," *Magnetic Resonance Imaging*, vol. 16, no. 2, pp. 115 – 125, 1998.

[14] C. Goutte, P. Toft, E. Rostrup, F. . Nielsen, and L. K. Hansen, "On Clustering fMRI Time Series," *NeuroImage*, vol. 9, no. 3, pp. 298 – 310, 1999.

[15] A. Wismüller, O. Lange, D. R. Dersch, G. L. Leinsinger, K. Hahn, B. Pütz, and D. Auer, "Cluster Analysis of Biomedical Image Time-Series," *Int. J. Comput. Vision*, vol. 46, no. 2, pp. 103–128, 2002.

[16] F. G. Meyer and J. Chinrungrueng, "Spatiotemporal clustering of fMRI time series in the spectral domain," *Medical Image Analysis*, vol. 9, no. 1, pp. 51 – 68, 2005.

[17] A. Mezer, Y. Yovel, O. Pasternak, T. Gorfine, and Y. Assaf, "Cluster analysis of resting-state fMRI time series," *NeuroImage*, vol. 45, no. 4, pp. 1117 – 1125, 2009.

[18] C. Windischberger, M. Barth, C. Lamm, L. Schroeder, H. Bauer, R. C. Gur, and E. Moser, "Fuzzy cluster analysis of high-field functional MRI data," *Artificial Intelligence in Medicine*, vol. 29, no. 3, pp. 203 – 223, 2003.

[19] A. Wismüller, A. Meyer-Base, O. Lange, D. Auer, M. F. Reiser, and D. Sumners, "Model-free functional MRI analysis based on unsupervised clustering," *Journal of Biomedical Informatics*, vol. 37, no. 1, pp. 10 – 18, 2004.

[20] T. Vincent, L. Risser, and P. Ciuciu, "Spatially adaptive mixture modeling for analysis of FMRI time series," *IEEE Transactions on Medical Imaging*, vol. 29, no. 4, pp. 1059 – 1074, 2010.

[21] L. He and I. R. Greenshields, "An MRF spatial fuzzy clustering method for fMRI SPMs," *Biomedical Signal Processing and Control*, vol. 3, no. 4, pp. 327 – 333, 2008.

[22] M. Woolrich, T. Behrens, C. Beckmann, and S. Smith, "Mixture models with adaptive spatial regularization for segmentation with an application to FMRI data," *IEEE Transactions on Medical Imaging*, vol. 24, pp. 1–11, 2005.

[23] N. Hartvig and J. Jensen, "Spatial mixture modeling of fMRI data," *Human Brain Mapping*, vol. 11, no. 4, 2000.

[24] W. Penny and K. Friston, "Mixtures of general linear models for functional neuroimaging," *IEEE Transactions on Medical Imaging*, vol. 22, pp. 504 –514, April 2003.

[25] J. Xia, F. Liang, and Y. M. Wang, "On Clustering fMRI Using Potts and Mixture Regression Models," in *31st Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pp. 4795–4798, 2009.

[26] M. E. Tipping, "Sparse Bayesian Learning and the Relevance Vector Machine," *Journal of Machine Learning Research*, vol. 1, pp. 211–244, 2001.

[27] H. Luo and S. Puthusserypady, "A sparse Bayesian method for determination of flexible design matrix for fMRI data analysis," *Circuits and Systems I: Regular Papers, IEEE Transactions on*, vol. 52, pp. 2699–2706, Dec. 2005.

[28] V. Oikonomou, , E. Tripoliti, and D. Fotiadis, "Bayesian Methods for fMRI Time-Series Analysis Using a Nonstationary Model for the Noise," *IEEE Transactions on Information Technology in Biomedicine*, vol. 14, pp. 664–674, May 2010 2010.

[29] S. Geman and D. Geman, "Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 6, pp. 721–741, 1984.

[30] V. Oikonomou, K. Blekas, and L. Astrakas, "A sparse and spatially constrained generative regression model for fMRI data analysis," *IEEE Transactions on Biomedical Engineering*, accepted.

[31] A. Dempster, L. A., and R. D., "Maximum likelihood from incomplete data via the EM algorithm," *Journal of Royal Statistical Society, Series B*, vol. 39, pp. 1–38, 1977.

[32] L. Harrison, W. Penny, J. Daunizeau, and K. Friston, "Diffusion-based spatial priors for functional magnetic resonance images," *NeuroImage*, vol. 41, pp. 408–423, 2008.

[33] K. Blekas, A. Likas, N. P. Galatsanos, and I. E. Lagaris, "A Spatially-Constrained Mixture Model for Image Segmentation," *IEEE Transactions on Neural Networks* , vol. 16, pp. 494–498, 2005.

[34] S. Gunn and J. Kandola, "Structural modelling with sparse kernels," *Machine Learning*, vol. 48, pp. 137–163, 2002.

[35] M. Girolami and S. Rogers, "Hierarchic Bayesian models for kernel learning," in *ICML '05: Proceedings of the 22nd international conference on Machine learning*, (New York, NY, USA), pp. 241–248, ACM, 2005.

[36] J. Nocedal and S. J. Wright, *Numerical Optimization*. Springer-Verlag New York, Inc., 1999.

[37] A. Diplaros, N. Vlassis, and T. Gevers, "A spatially constrained generative model and an em algorithm for image segmentation," *IEEE Trans. on Neural Networks*, vol. 18, no. 3, pp. 798–808, 2007.

[38] J. Li and A. Barron, "Mixture density estimation," in *Advances in Neural Information Processing Systems*, vol. 12, pp. 279–285, The MIT Press, 2000.

[39] N. Ueda, R. Nakano, Z. Ghahramani, and G. Hinton, "SMEM algorithm for mixture models," *Neural Computation*, vol. 12, no. 9, pp. 2109–2128, 2000.

[40] N. Vlassis and A. Likas, "A greedy EM algorithm for Gaussian mixture learning," *Neural Processing Letters*, vol. 15, pp. 77–87, 2001.

[41] T. E. Lund, K. H. Madsen, K. Sidaros, W.-L. Luo, and T. E. Nichols, "Non-white noise in fmri: Does modelling have an impact?," *NeuroImage*, vol. 29, no. 1, pp. 54 – 66, 2006.

[42] L. Astrakas, S. Konitsiotis, P. Margariti, S. Tsouli, L. Tzarouhi, and M. I. Argyropoulou, "T2 relaxometry and fMRI of the brain in lateonset restless legs syndrome," *Neurology*, vol. 71, no. 12, pp. 911–916, 2008.

[43] M. Fadili, S. Ruan, D. Bloyet, and B. Mazoyer, "A Multistep Unsupervised Fuzzy Clustering Analysis of fMRI Time Series," *Human Brain Mapping*, vol. 10, pp. 160–178, 2000.

[44] S. Pieper, M. Halle, and R. Kikinis, "3D SLICER," *IEEE International Symposium on Biomedical Imaging ISBI 2004*, pp. 632–635, 2004.

[45] D. J. McGonigle, A. M. Howseman, B. S. Athwal, K. J. Friston, R. S. J. Frackowiak, and A. P. Holmes, "Variability in fMRI: An Examination of Intersession Differences," *NeuroImage*, vol. 11, no. 6, pp. 708–734, 2000.

[46] M. Seeger, "Bayesian inference and optimal design for the sparse linear model," *Journal of Machine Learning Research*, vol. 9, pp. 759–813, 2008.

[47] A. Venkataraman, K. V. Dijk, R. L. Buckner, and P. Golland, "Exploring functional connectivity in fmri via clustering," *Acoustics, Speech, and Signal Processing, IEEE International Conference on*, pp. 441–444, 2009.

[48] S.-C. Ngan, X. Hu, and P.-L. Khong, "Investigating the enhancement of template-free activation detection of event-related fmri data using wavelet shrinkage and figures of merit," *Artificial Intelligence in Medicine*, vol. 51, no. 3, pp. 187 – 198, 2011.