CHAPTER 1

Regression Mixture Modeling for fMRI data analysis

V.P. Oikonomou
School of Business and Economics,
Department of Business Administration,
TEI of Ionian Islands, 31100 Lefkada, Greece
E-mail:viknmu@gmail.com

K. Blekas
Department of Computer Science & Engineering,
University of Ioannina,
45110 Ioannina, Greece
E-mail: kblekas@cs.uoi.gr

Functional magnetic resonance imaging (fMRI) has become a novel technique for studying the human brain and obtaining maps of neuronal activity. An important goal in fMRI studies is to decompose the observed series of brain images in order either to detect activation when a stimulus is presented to the subject, or to identify and characterize underlying brain functional networks when the subject is at rest. In this chapter a model class is presented for addressing this issue that consists of finite mixture of generalized linear regression models. The main building block of the method is the general linear model which constitutes a standard statistical framework for investigating relationships between variables of fMRI data. We extend this into a finite mixture framework that exploits enhanced modeling capabilities by incorporating some innovative *sparse* and *spatial* properties. In addition a weighted multi-kernel scheme is employed dealing with the selection problem of kernel parameters where the weights are estimated during training. The proposed regression mixture model is trained using the maximum a posteriori approach, where the Expectation-Maximization (EM) algorithm is applied for constructing update equations for the model parameters. We provide comparative experimental results in both activation-based and resting state applications that illustrate the ability of the proposed method to produce improved performance and discrimination capabilities.

1. Introduction

Human brain represents the most complex system in the nature. It is the center of the nervous system. This organ of 1.5 Kg and a volume around of $1200cm^3$ is responsible for almost every complex task of a human being. Millions of elementary

components, called neurons, are interconnected to each other creating a complex information processing network. The activity of this network is associated with the mind and gives rise to consciousness. Despite of the rapid scientific progress of last decades, how the brain works remains a mystery. While the brain is protected by the bones of the skull, it is still vulnerable to damage and disease. Also it is susceptible to degenerative disorders, such as Parkinson's disease, multiple sclerosis, and Alzheimer's disease. Understanding the human brain is one of the greatest scientific challenges of the years to come [1, 2].

Brain imaging uses various techniques to produce images of the brain. Electroencephalography (EEG) is the oldest technique for brain imaging and it produces images of the brain by recording the electrical activity along the scalp. Another neuroimaging technique is the Magnetoencephalography (MEG) which records the magnetic fields produced by the electric currents of the brain. While both techniques present excellent temporal resolution, their spatial resolution is a major drawback since they cannot describe the anatomical structures of the brain. Hence, the use of them has decreased after the introduction of anatomical imaging techniques with high spatial resolution such as Magnetic Resonance Imaging (MRI). Positron Emission Tomography (PET) is a functional neuroimaging technique used to examine various tissues of human body. This technique presents very good spatial resolution. However, the time resolution is very bad and this affects the experimental design since only blocked design experiments can be performed. PET is an invasive technique since a radiotracer is injected into the human body.

Today, the most popular technique for functional neuroimaging is the fMRI. It is a noninvasive technique which presents very good spatial resolution while its time resolution is better compared to other similar techniques such as PET, that offers the opportunity to perform more complicated experimental designs. The fMRI analysis is based mostly on the Blood Oxygenation Level Dependent (BOLD) effect, firstly reported in [3]. When a stimulus is applied to a subject, regions of the brain involved in the process are becoming active. As a result the rate of blood flow is increased and more oxygenated blood arrives. Furthermore, the blood contains iron which is a paramagnetic material. In the above metabolic procedure oxygenated and deoxygenated blood are taking part. However, the deoxygenated blood is more paramagnetic than oxygenated. This difference on the magnetic properties between oxygenated and deoxygenated blood is exploited by MRI technology to produce brain images. The increase in blood flow is known as the *hemodynamic response*. For the statistical analysis of the fMRI data two properties of the hemodynamic response are important. First, the hemodynamic response is slow compared to the neuronal activity. Second, it can be treated (or approximated) as a linear time invariant system. The linearity property together with the mathematical operation of convolution constitute the basic tools to construct statistical models and environments such as the SPM [4] and the FSL [5], for studying fMRI applications.

Image acquisition of fMRI constructs a $4-D$ dataset consisting of $3-D$ brain

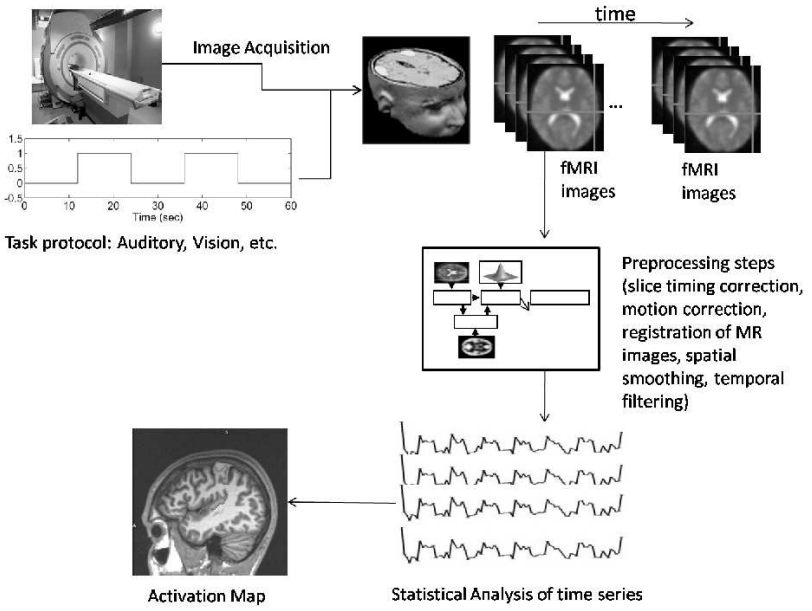


Fig. 1. Overall scheme in fMRI analysis.

volumes that evolve in time. The basic element is called *voxel* and represents a value on a grid in $3-D$. By taking the values of voxels over time we create a set of *time-series*, i.e. sequential type of data measured in successive time instances at uniform intervals. The fMRI data contains various important properties and a careful analysis of them is needed for the subsequent analysis. Temporal correlations between the samples are found due to physiological properties and experimental conditions. This phenomenon depends mostly on how frequently we acquire the images in conjunction with the duration of BOLD effect. Also, spatial correlation can be observed in the data. This is derived from physiological properties, such as the activated brain areas and the connectivity between brain areas, as well as technical considerations, such as the smallest size of brain location in space that we can obtain. In addition, what affects the quality of fMRI data is the presence of noise that is observed in the data. There are two main sources of noise: noise due to the imaging process and noise due to the human subject.

The types of fMRI experiments can be divided into two large groups according to the desired target. In the *activation based* fMRI experiments the human subject is exposed to a series of stimulated events according to the experimental design, which provides a binary vector (stimulus is either present, or not). This vector is combined with the hemodynamic response function, through the convolution operator, to give the BOLD regressor which is very important for the statistical analysis of our data. The second group is the *resting state* type of fMRI experiments where we try to

find connections between various brain areas when the human subject is at rest, i.e. no stimulus is present. Figure 1 illustrates briefly the overall procedure of the fMRI data analysis process in a flow diagram design.

2. An overview of fMRI data analysis

The objective of fMRI data analysis is to detect the weak BOLD signal from the noisy data and determine the activated areas of the brain. It usually consists of two stages: preprocessing and statistical analysis. The first stage contains various techniques that could be made in order to remove artifacts, validate the assumptions of the model and standardize the brain regions across subjects [6, 7]. Among them, the most common preprocessing schemes are: slice timing correction, realignment, coregistration of images, normalization, spatial smoothing and temporal filtering.

In the literature there are many methodologies that have been proposed for the analysis of fMRI data. They can be divided into two major categories: the *model-based* and the *data driven* (or model-free). The term "model" is referred to the process of modeling the hemodynamic response. The model-based approaches are used only for activation based fMRI studies, and are mainly based on the general linear regression model (GLM) [8] and its extensions [9, 10]. At the end of the learning process the statistical activation map is drawn based on $t-$ or $F-$ statistics displaying the activation areas and the importance of each voxel [8]. On the other hand, the data driven methods are applied on both resting state and activation based studies and include the principal component analysis (PCA) [11], independent component analysis (ICA) [12, 13] and clustering algorithms [11, 13–15].

A significant drawback of the GLM is that spatial and temporal properties of fMRI data are not taken into account in its basic scheme. More specifically, autoregressive modeling of noise have been proposed in [9,10] so as to incorporate temporal correlations, while non-stationary models of noise have been presented in [16,17] for the analysis of fMRI time series. Moreover, spatial properties of data are included by usually performing a smoothing with a fixed Gaussian kernel as a preprocessing step [9,18]. Other approaches have been also proposed that elaborate denoising techniques, see for example [9, 19]. Under the Bayesian framework, spatial dependencies have been modeled through Markov Random Field (MRF) priors applied either to temporal and spatial components of the signal, or to the noise process [20]. Also, Gaussian spatial priors have been placed over the regression coefficients, as well as on autoregressive coefficients of the noise process [9].

An important feature of the GLM is the type of the design matrix used which may affect significantly the subsequent statistical analysis. Some typical examples are the Vandermonde or B-splines matrix (dealing with polynomial or spline regression models), while other use some predefined dictionaries (basis functions) derived from transformations, such as Fourier, Wavelets, or Discrete Cosine Transform [10]. Other more advanced techniques apply kernel design matrix constructing from an appropriate parametric kernel function [21, 22]. Alternatively, for the activation

based fMRI studies the design matrix could contain information about the experimental paradigm [8]. Also, regressors related to head motion can be included since remnants from head motion noise could be present in the time series [10]. Finally, following the Bayesian framework, sparse priors over regression coefficients could be introduced so as to determine automatically the design matrix [21–23].

Another family of methods for the fMRI data analysis with special advantages is through clustering techniques. Clustering is the procedure of dividing a set of unlabeled data into a number of groups (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 [24]. Cluster analysis of fMRI data constitutes a very interesting application that has been successfully applied during last years. The target is to create a partition into distinct regions, where each region consists of voxels with similar temporal behavior. Most popular clustering methods use partitioning methodologies such as k-means, fuzzy clustering and hierarchical clustering. They are are applied to either entire raw data, or feature sets which are extracted from the fMRI signals [14, 15, 25–31].

Recently, more advanced approaches have been introduced in order to meet spatial correlation properties of data. In [32] a spatially constrained mixture model has been adopted for capturing the Hemodynamic Response Function (HRF), while in [33] the fuzzy c-means algorithm in cooperation with a spatial MRF was proposed to cluster the fMRI data. Furthermore, a mixture model framework with spatial MRFs applied on statistical maps was described in [19, 34]. However, in the above works the clustering procedure was performed indirectly, either through careful construction of the regression model, or using features extracted from the fMRI time series. Also, temporal patterns of clusters have not been taken into account. A solution to this is to perform the clustering directly to fMRI time series, as for example in [35], where a mixture of GLMs was presented using a spatial prior based on the Euclidean distances between the positions of time series and cluster centers in a 3-D space head model. An alternative solution was given in [36], where spatial correlations among the time series is achieved through Potts models over the hidden variables of the mixture model.

In this chapter we present an advanced regression mixture modeling approach for clustering fMRI time series [22] that incorporates very attractive features to facilitate the analysis of fMRI data. The main contribution of the method lies on three aspects:

- Firstly, it achieves a sparse representation of every regression model (cluster) through the use of an appropriate sparse prior over the regression coefficients [37]. Enforcing sparsity is a fundamental machine learning regularization principle [24,37] and has been used in fMRI data analysis [9,17,23].
- 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 [21], so as to create smoother activation regions.

- Finally, a kernel estimation procedure is established through a multi-kernel scheme over the design matrix of the regression models. In this way we can manage to improve the quality of data fitting and to design more compact clusters.

Training of the proposed regression mixture model is performed by setting a Maximum A Posteriori (MAP) estimation framework and employing the Expectation-Maximization (EM) algorithm [38,39]. Numerous experiments have been conducted using both artificial and real fMRI datasets where we have considered applications on activation based, as well as on resting state fMRI data. Comparison has been made using a regression mixture model with only spatial properties 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.

This chapter is structured as follows. At first we present the basic regression mixture model and then we show how it can be adapted in order to fit the fMRI data and their properties. This is split into descriptions of the priors, the general construction and the MAP likelihood, where we show how the EM algorithm can be used for estimating the model parameters. The experiments section presents several results from functional activation studies of auditory and event-related (foot movement) experiments, as well as from resting state fMRI studies. Comparison has been also made with standard approaches. The chapter finishes with some concluding remarks.

3. Finite mixture of regression models

3.1. *Mixture models*

Mixture models provides a powerful probabilistic modeling tool for data analysis. It has been used in many scientific areas including machine learning, pattern recognition, signal and image analysis and computer vision [24,39]. That makes mixture models so popular and suitable is that they are parametric models of elegant way, yet they are very flexible and easily extensible in estimating any general and complex density and finally, they are capable of accounting for unobserved heterogeneity.

A mixture model of order K is a linear combination of K probability density parametric functions $p(\boldsymbol{y}|\boldsymbol{\theta}_j)$ of different sources and it is formulated as:

$$p(\boldsymbol{y}|\Theta) = \sum_{j=1}^{K} \pi_j p(\boldsymbol{y}|\boldsymbol{\theta}_j) \ . \tag{1}$$

The parameters π_j are the mixing weights satisfying the constraints:

$$0 \leq \pi_j \leq 1 \text{ and } \sum_{j=1}^{K} \pi_j = 1 \ , \tag{2}$$

while $\Theta = \{\pi_j, \theta_j\}_{j=1}^K$ is the set of model parameters which are unknown and must be estimated. According to this model, each observation is generated by first selecting a source j based on the probabilities $\{\pi_j\}$ and then by performing sampling based on the corresponding distribution with parameters θ_j. Having found the parameters Θ, the posterior probabilities that an observation $\boldsymbol{y}$ belongs to the j-th component can be calculated:

$$P(j|\boldsymbol{y}) = \frac{\pi_j p(\boldsymbol{y}|\theta_j)}{\sum_{k=1}^K \pi_k p(\boldsymbol{y}|\theta_k)} \tag{3}$$

Then, an observation belongs to the component j^* with the largest posterior value, i.e. $P(j^*|y) > P(j|y)\ \forall j \neq j^*$.

Let us assume that we have a data set of N samples $Y = \{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_N\}$ which are independent. The estimation of the mixture model parameters Θ can be made by maximizing the log-likelihood function:

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

The Expectation-Maximization (EM) [38] algorithm provides a useful framework for solving likelihood estimation problems. It uses a data augmentation scheme and is a general estimation method in the presence of missing data. In the case of finite mixture models the component memberships play the role of missing data. EM iteratively performs two main steps. During the *E-step* the expectation of hidden variables are calculated based on the current estimation of the model parameters:

$$z_{nj} = P(j|\boldsymbol{y}_n) = \frac{\pi_j p(\boldsymbol{y}_n|\theta_j)}{\sum_{k=1}^K \pi_k p(\boldsymbol{y}_n|\theta_k)}\ . \tag{5}$$

At the *M-step* the maximization of the complete data log-likelihood function (Q-function) is performed:

$$Q(\Theta|\Theta^{(t)}) = \sum_{n=1}^N \sum_{j=1}^K z_{nj}\{\log \pi_j + \log p(\boldsymbol{y}_n|\theta_j)\} \tag{6}$$

This leads to obtaining new estimates of the mixture weights:

$$\pi_j = \frac{\sum_{n=1}^N z_{nj}}{N}\ , \tag{7}$$

as well as of the model components parameters $\theta_j^{(t+1)}$. The received update rules depend on the type of component density functions. In the case of multivariate Gaussian mixture models for example, i.e. $p(\boldsymbol{y}|\theta_j) = N(\boldsymbol{y}; \mu_j, \Sigma_j)$, these rules become [24, 39]:

$$\boldsymbol{\mu}_j = \frac{\sum_{n=1}^N z_{nj}\boldsymbol{y}_n}{\sum_{n=1}^N z_{nj}} \tag{8}$$

$$\boldsymbol{\Sigma}_j = \frac{\sum_{n=1}^N z_{nj}(\boldsymbol{y}_n - \boldsymbol{\mu}_j)(\boldsymbol{y}_n - \boldsymbol{\mu}_j)^T}{\sum_{n=1}^N z_{nj}} \tag{9}$$

The E- and M- steps are alternated repeatedly until some specified convergence criterion is achieved.

3.2. *Regression Mixture Modeling*

In the case of fMRI data analysis, we are dealing with *time-series* type of data which are sequences of values measured at T successive time instances x_l, i.e. $\boldsymbol{y}_n = \{y_{nl}\}_{l=1}^{T}$. Linear regression modeling constitutes an elegant functional description framework for analyzing sequential data. It is described with the following form:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{w} + \boldsymbol{e}_n \ , \tag{10}$$

where $\boldsymbol{w}$ is the vector of M (unknown) linear regression coefficients. The $\boldsymbol{e}_n$ is an additive error term (T dimensional vector) that is assumed to be zero mean Gaussian with a spherical covariance $\boldsymbol{e}_n \sim \mathcal{N}(0, \sigma^2\boldsymbol{I})$, i.e. errors are not correlated.

For constructing the design matrix $\boldsymbol{X}$ several approaches can be employed. A common choice is to use Vandermonde or B-splines matrix in cases where we assume polynomial or splines regression models, respectively. Another option is to assume a kernel design matrix using an appropriate kernel basis function over time instances $\{x_l\}_{l=1}^{T}$, with the RBF kernel function to be the most commonly used:

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

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.

Following the Eq. 10 it is obvious that, given the set of regression model parameters $\theta = \{\boldsymbol{w}, \sigma^2\}$, the conditional probability density of time-series $\boldsymbol{y}_n$ is also Gaussian, i.e.

$$p(\boldsymbol{y}_n|\theta) = \mathcal{N}(\boldsymbol{X}\boldsymbol{w}, \sigma^2\boldsymbol{I}) \ .$$

Regression mixture models [39] provides a natural framework for fitting a given set of sequential data $Y = \{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n\}$. They allow for simultaneously modeling heterogeneous regression functions by training a mixture of distinct distributions where each one corresponds to a latent class. Obviously, this is equivalent to the task of time-series clustering, i.e. the division of the set Y into K clusters, in such a way that each cluster contains similar in nature elements. Therefore each cluster has its own regression generative mechanism, as given by a conditional density with parameters $\theta_j = \{\boldsymbol{w}_j, \sigma_j^2\}$, $j = 1, \ldots, K$.

The EM algorithm can be then applied in order to train regression mixture models. That differs from the basic scheme described previously is the expected

complete log-likelihood Q-function which takes the following form:

$$Q(\Theta|\Theta^{(t)}) = \sum_{n=1}^{N}\sum_{j=1}^{K} z_{nj}\Big\{\log\pi_j - \frac{T}{2}\log 2\pi - T\log\sigma_j - \frac{\|\boldsymbol{y}_n - \boldsymbol{X}\boldsymbol{w}_j\|^2}{2\sigma_j^2}\Big\}, \quad (11)$$

as well as the update rules of the regression component parameters θ_j which are

$$\boldsymbol{w}_j = \Big(\sum_{n=1}^{N} z_{nj}\boldsymbol{X}^T\boldsymbol{X}\Big)^{-1}\boldsymbol{X}^T\sum_{n=1}^{N}(z_{nj}\boldsymbol{y}_n), \quad (12)$$

$$\sigma_j^2 = \frac{\sum_{n=1}^{N} z_{nj}\|\boldsymbol{y}_n - \boldsymbol{X}\boldsymbol{w}_j\|^2}{T\sum_{n=1}^{N} z_{nj}}. \quad (13)$$

After the convergence of the EM algorithm, each sequence $\boldsymbol{y}_n$ is assigned to the cluster with the maximum posterior probability $P(j|\boldsymbol{y}_n)$ (similar to Eq. 3).

4. Regression mixture analysis of fMRI time-series

The application of the basic ML-based scheme of regression mixture models to the task of fMRI data analysis has some limitations due to its weakness to capture some important features arisen from the nature of these observations. In particular, the fMRI data are structures that involve spatial properties, where adjacent voxels tend to have similar activity behavior [40]. Furthermore, there are temporal correlations which are derived from neural, physiological and physical sources [10]. These are physical constraints that must be incorporated to the model.

4.1. *General construction*

A significant advantage of Bayesian estimation is its flexibility of incorporating appropriate priors and its full characterization of the posterior. Bayesian modeling also enable us to model the uncertainty of the hyperparameters so as the final performance to be more robust. In such a way we can eliminate the phenomenon of data overfitting found in the ML case. Three are the main building blocks for constructing a maximum a-posteriori (MAP) approach which offers a more advanced solution: *sparseness, spatial, and multi-kernel*.

4.1.1. *Sparse modeling*

An important issue when using a regression model is how to estimate its order M, i.e. the size of linear regression coefficients $\boldsymbol{w}_j$. 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 [24, 37]. It initially assumes a large value of order M. Then, a heavy tailed prior distribution $p(\boldsymbol{w}_j)$ is imposed upon the regression coefficients that will enforce most of the coefficients

to be zero out after training. This has been successfully employed in the Relevance Vector Machine model [37].

More specifically, the prior is defined in an hierarchical way by considering first a zero-mean Gaussian distribution over the regression coefficients:

$$p(\boldsymbol{w}_j|\boldsymbol{\alpha}_j) = \mathcal{N}(\boldsymbol{w}_j|\mathbf{0}, A_j^{-1}) = \prod_{l=1}^{M} \mathcal{N}(w_{jl}|0, \alpha_{jl}^{-1}) \ , \tag{14}$$

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:

$$p(\alpha_j) = \prod_{l=1}^{M} \Gamma(\alpha_{jl}|b, c) \propto \prod_{l=1}^{M} \alpha_{jl}^{b-1} \exp^{-c\alpha_{jl}} \ . \tag{15}$$

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.

4.1.2. *Spatial regularization*

A common approach to achieve spatial correlations between voxels is to apply a spatial Gaussian filter to smooth the signal prior to statistical analysis. This is used for instance in Statistical Parametric Mapping (SPM) [4]. However, this can lead to overlay blurred results, where effects with small spatial extend can be lost and detected regions may extend beyond their actual boundaries. A more advanced approach to spatial regularization is through the use of Markov Random Field (MRF) prior [41] which models the conditional dependence of the signals in neighboring voxels.

MRFs have been successfully applied to computer vision applications [41, 42]. Conventional use of MRFs requires the set of sites of the random field as the image voxels, with the neighborhood structure given by a regular lattice. More specifically, we can treat the probabilities (voxel labels) π_{nj} of each fMRI sequence $\boldsymbol{y}_n$ belongs to the j-th cluster (mixture component) as random variables, which also satisfy the constraints $\pi_{nj} \geq 0$ and $\sum_{j=1}^{K} \pi_{nj} = 1$. We assume that the set of voxel labels $\Pi = \{\boldsymbol{\pi}_n\}_{n=1}^{N}$ follows the Gibbs prior distribution with density [41]

$$p(\Pi) = \frac{1}{Z} \exp\{-\sum_{n=1}^{N} V_{N_n}(\Pi)\} \ . \tag{16}$$

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:

$$V_{N_n}(\Pi) = \sum_{m \in N_n} \sum_{j=1}^{K} \beta_j (\pi_{nj} - \pi_{mj})^2. \tag{17}$$

In our case we consider neighbourhood 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 this framework the regularization parameters β_j belong to the set of the unknown parameters and thus can be estimated during the learning process. Finally, the term Z of Eq. 16 is the normalizing factor that is analogous to $Z \propto \prod_{j=1}^{K} \beta_j^{-N}$.

An alternative methodology on using a capable MRF prior to leverage spatial correlations in brain maps is through a recent non-parametric scheme shown in [43]. In particular, an appropriate class-specific Gibbs potential function has been proposed of the following form:

$$\vartheta_{nj} = \sum_{m \in N_n} z_{nj} z_{mj} \ , \tag{18}$$

that gives the influence of the neighborhood to the decision process. This function acts as a smooth filter to the estimated posteriors and it works like a voting system, where the majority cluster-label among its closest neighbors is assigned to every sequence. Then, probabilities of voxels' labels are given according to a *softmax* function:

$$\pi_{nj} \propto \frac{e^{\vartheta_{nj}}}{\sum_{k=1}^{K} e^{\vartheta_{nk}}} \ . \tag{19}$$

4.1.3. *Multi-kernel scheme*

As mentioned before, the construction of the design matrix $\boldsymbol{X}$ is a crucial part of the regression model and may be significantly affected by the parameter value of the desired kernel function. This problem can be solved by adopting a multi-kernel scheme [44,45]. In particular, we assume a pool of S kernel matrices $\{\boldsymbol{\Phi}_s\}_{s=1}^{S}$, each one having its own scalar parameter value λ_s. Thus the composite kernel matrix $\boldsymbol{X}_j$ for the $j-th$ cluster can be written as a linear combination of S kernel matrices Φ_s:

$$\boldsymbol{X}_j = \sum_{s=1}^{S} u_{js} \boldsymbol{\Phi}_s \ , \tag{20}$$

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 use of the proposed multi-kernel scheme has the ability to significantly improve the performance and the quality of the data fitting procedure.

4.2. *Estimation of model parameters*

After defining the sparse and sparse priors together with the multi-kernel scheme, we are now ready to describe the estimation process of the model parameters. The incorporation of the above properties leads to a modification of the regression mixture model which is written as:

$$f(\boldsymbol{y}_n|\Theta) = \sum_{j=1}^{K} \pi_{nj} p(\boldsymbol{y}_n|\theta_j), \tag{21}$$

where $\Theta = \{\{\pi_{nj}\}_{n=1}^{N}, \theta_j = (\boldsymbol{w}_j, \boldsymbol{\alpha}_j, \sigma_j^2, \boldsymbol{u}_j, \beta_j)\}_{j=1}^{K}$ 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

$$\begin{aligned} l_{MAP}(\Theta) &= \log p(Y|\Theta) + \log p(\Theta) = \\ &= \sum_{n=1}^{N} \log f(\boldsymbol{y}_n|\Theta) + \log p(\Pi) + \sum_{j=1}^{K} \Big\{ \log p(\boldsymbol{w}_j|\boldsymbol{\alpha}_j) + \log p(\boldsymbol{\alpha}_j) \Big\} \ . \end{aligned} \tag{22}$$

The EM algorithm can then be applied for MAP-estimating the model parameters. Likewise, it requires at each iteration the conditional expectation values z_{nj} of the hidden variables to be computed first (E-step):

$$z_{nj} = P(j|\boldsymbol{y}_n, \Theta) = \frac{\pi_{nj} p(\boldsymbol{y}_n|\theta_j)}{f(\boldsymbol{y}_n|\Theta)} \ . \tag{23}$$

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

$$\begin{aligned} Q(\Theta|\Theta^{(t)}) = \sum_{n=1}^{N}\sum_{j=1}^{K} z_{nj} \Big\{ \log \pi_{nj} - \frac{T}{2}\log 2\pi - T\log \sigma_j - \frac{\|\boldsymbol{y}_n - \boldsymbol{X}_j \boldsymbol{w}_j\|^2}{2\sigma_j^2} \Big\} - \\ \sum_{j=1}^{K} \Big\{ -N\log\beta_j + \beta_j \sum_{n=1}^{N} \sum_{m\in N_n} (\pi_{nj} - \pi_{mj})^2 + \frac{1}{2}\boldsymbol{w}_j^T \boldsymbol{A}_j \boldsymbol{w}_j - \sum_{l=1}^{M} [(b-1)\log\alpha_{jl} - c\alpha_{jl}] \Big\} \ . \end{aligned} \tag{24}$$

By setting the partial derivatives of the above Q function with respect to all model parameters we can obtain the update rules. For the regression model parameters

$\{\boldsymbol{w}_j, \sigma_j^2, \boldsymbol{\alpha}_j, \beta_j\}$ we can easily obtain the next equations:

$$\boldsymbol{w}_j = \Big[\Big(\sum_{n=1}^{N} z_{nj}\Big)\frac{1}{\sigma_j^2}\boldsymbol{X}_j^T\boldsymbol{X}_j + \boldsymbol{A}_j\Big]^{-1} \cdot \frac{1}{\sigma_j^2}\boldsymbol{X}_j^T\Big(\sum_{n=1}^{N} z_{nj}\boldsymbol{y}_n\Big), \tag{25}$$

$$\sigma_j^2 = \frac{\sum_{n=1}^{N} z_{nj}\|\boldsymbol{y}_n - \boldsymbol{X}_j\boldsymbol{w}_j\|^2}{T\sum_{n=1}^{N} z_{nj}}, \tag{26}$$

$$\alpha_{jl} = \frac{1+2c}{w_{jl}^2 + 2b}, \tag{27}$$

$$\beta_j = \frac{N}{\sum_{n=1}^{N}\sum_{m\in N_n}(\pi_{nj} - \pi_{mj})^2}. \tag{28}$$

In the case of the label parameters π_{nj} we obtain the following quadratic equation:

$$\pi_{nj}^2 - < \pi_{nj} > \pi_{nj} - \frac{1}{2\beta_j|N_n|}z_{nj} = 0\ , \tag{29}$$

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 $\pi_{nj} \geq 0$:

$$\pi_{nj} = \frac{< \pi_{nj} > + \sqrt{< \pi_{nj} >^2 + \frac{2}{\beta_j|N_n|}z_{nj}}}{2}\ . \tag{30}$$

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 [42], that is based on the active-set theory [46].

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 $\boldsymbol{u}_j$:

$$\min_{\boldsymbol{u}_j}\sum_{n=1}^{N} z_{nj}\parallel \boldsymbol{y}_n - \sum_{s=1}^{S} u_{js}\boldsymbol{\Phi}_s\boldsymbol{w}_j \parallel^2 = \min_{\boldsymbol{u}_j}\sum_{n=1}^{N} z_{nj}\parallel \boldsymbol{y}_n - \mathcal{X}_j\boldsymbol{u}_j \parallel^2 =$$

$$\min_{\boldsymbol{u}_j}\Big\{\boldsymbol{u}_j^T\mathcal{X}_j^T\mathcal{X}_j\boldsymbol{u}_j - 2\boldsymbol{u}_j^T\mathcal{X}_j^T\frac{\sum_{n=1}^{N} z_{nj}\boldsymbol{y}_n}{\sum_{n=1}^{N} z_{nj}}\Big\}\ ,\ \text{s.t.}\ \sum_{s=1}^{S} u_{js} = 1\ \text{and}\ u_{js} \geq 0\ . \tag{31}$$

In the above formulation, the matrix $\mathcal{X}_j$ has S columns calculated by $\boldsymbol{\Phi}_s\boldsymbol{w}_j$, i.e. $\mathcal{X}_j = [\boldsymbol{\Phi}_1\boldsymbol{w}_j\ \ \boldsymbol{\Phi}_2\boldsymbol{w}_j\ \cdots\ \boldsymbol{\Phi}_S\boldsymbol{w}_j]$. The minimization problem described in Eq. 31 is a

typical constrained linear least-squared problem that can be solved again with the active-set theory [46].

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 $\boldsymbol{\mu}_j = \boldsymbol{X}_j \boldsymbol{w}_j$ of each cluster with the BOLD signal $\boldsymbol{\xi}$, i.e.

$$h = \arg\max_{j=1}^{K} \frac{\boldsymbol{\mu}_j^T \boldsymbol{\xi}}{|\boldsymbol{\mu}_j||\boldsymbol{\xi}|} \ . \tag{32}$$

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.

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. A common practice is to initialize mixture model parameters by randomly selecting K input time-series and to perform only a few EM steps. Several trials of such procedure can be made and finally the solution with the maximum log-likelihood value can be selected for the initialization.

A more advanced approach has been proposed in [22] that follows an incremental strategy for building the regression mixture model. Starting with a mixture model with one regression component, the learning methodology adds a new component to the mixture based one a *component splitting* procedure. In activation based fMRI data analysis this is done by selecting a cluster for splitting based on their similarity with the BOLD signal. A detailed description can be found in [22]. It must be noted that an obvious advantage of the incremental learning scheme is that of simultaneously offering solutions for the intermediate models with $k = \{1, \ldots, K\}$ components. This can be seen very convenient for introducing model order selection criteria and terminating the evolution of learning: stop training when the insertion of a new component does not offer any significant improvement of the (penalized) likelihood function.

5. Experiments

The proposed regression mixture model (called as SSRM) has been evaluated using a variety of artificial datasets and real fMRI data. In all experiments for constructing the multi-kernel scheme, we calculated first the total variance of samples, λ. Next, we used a set of $S = 10$ RBF kernel functions, where each one had a scalar parameter $\lambda_s = k_s \lambda$, where $k_s = [0.1, 0.2, \ldots, 1.0]$ (level of percentage). It must be noted that during the activation-based experiments another column has been added to

the design matrix which describes the BOLD signal. Note that the time instances x_l were normalized before to $[0, 1]$. Finally, the linear weights of the multi-kernel scheme were in all cases initialized equally to $u_{js} = 1/S$. Comparison has been made using the SRM method which is a regression mixture model with only spatial properties (and without sparse properties), and the k-means which is a well known vector-based clustering approach. An extended experimental study can be found in [22, 43], that present additional comparative results with the standard GLM model [8] and various types of noise.

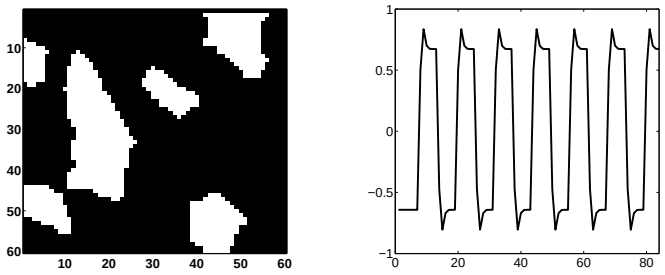
5.1. *Activation-based fMRI experiments*

The goal of this series of experiments is to discover the brain activation areas when the human subject is exposed to a stimulus. At first we have studied the performance of the proposed method using synthetic data, where the ground truth of activation is known. Additional experiments were made using real fMRI datasets taken from block design auditory and event-related foot movement studies.

5.1.1. *Experiments with artificial datasets*

During the experiments with simulated fMRI data, we have created $3 - D$ set of time series using linear regression models with known design matrix and regression coefficients. We have also added white Gaussian noise of various SNR levels according to the formula: $SNR = 10 \log_{10} \left(\frac{\boldsymbol{s}^T \boldsymbol{s}}{N\sigma^2} \right)$, where σ^2 is the noise variance and $\boldsymbol{s}$ is the BOLD signal. The spatial correlation among time series is achieved through the regression coefficients. Figure 2(a) represents the spatial patterns used, while the BOLD signal used to model the neural activity is shown in Fig. 2(b). 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 by the standard normal distribution $\mathcal{N}(0, 1)$. The size of the obtained dataset was $80 \times 80 \times 84$. 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 the depicted results (mean and variance). To measure the quality of each clustering approach, we have used two evaluation criteria: the accuracy performance (percentage of correct classifying data) and the Normalized Mutual Information (NMI) [22].

Figure 3 shows the comparative results for our simulated dataset of Fig. 2. The superiority of the SSRM is obvious based on two evaluation criteria, especially in small SNR values (noisy data). 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 sparse term to modeling procedure. An example of the activation maps as estimated by each method is shown in Figs. 4 in the case of $SNR = -8$ dB. Clearly, our method had better discrimination



(a) (b)

Fig. 2. (a) Spatial patterns and (b) the BOLD signal used in experiments with simulated data

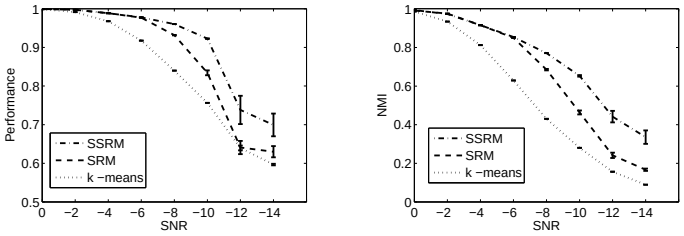
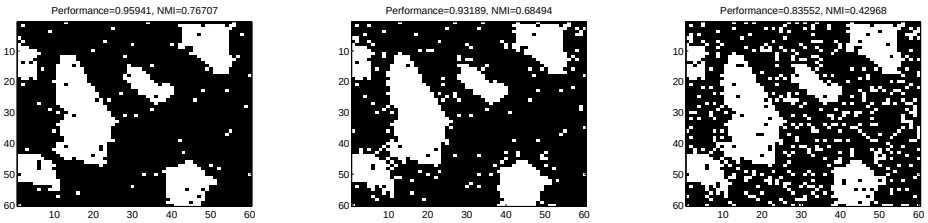


Fig. 3. Comparative results for our dataset of Fig. 2. Error bars for the two evaluation criteria are shown in terms of several SNR values.



(a) SSRM (b) SRM (c) k-means

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

ability and achieved to discover more accurately the original spatial pattern, while at the same time reduced significantly the false negative activation cases. A more comprehensive experimental analysis can be found on [22].

5.1.2. *Experiments using real fMRI data*

We have made additional experiments using real fMRI data. In our study, we have selected a dataset with a block-designed auditory paradigm. In this experiment, we

have followed the standard preprocessing steps of the SPM package. The BOLD signals for the experiment is shown in Fig. 2b. This dataset was downloaded from the SPM webpage [a] and it was 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). Experiments were made with the slice 29 of this dataset, which contains a number of $N = 5118$ time series. In this series of experiments we have employed the incremental learning strategy of the proposed method SSRM [22] which provided us with the proper number of clusters. We have found a number of $K = 5$ cluster and then we have used this value in order to run the other two approaches, SRM and k-means. Figure 5 represents the comparative results of all clustering methods giving the resulting position of the activation area inside the brain. Note that the activated areas are overlaid on grayscale T1 weighted anatomical images. All methods have detected the auditory cortex as the brain activation area. However, the SSRM methods have clearly detected only three distinct areas of activation, while the rest two approaches have additionally detected other small activated islands that may bring difficulties in the decision making process.

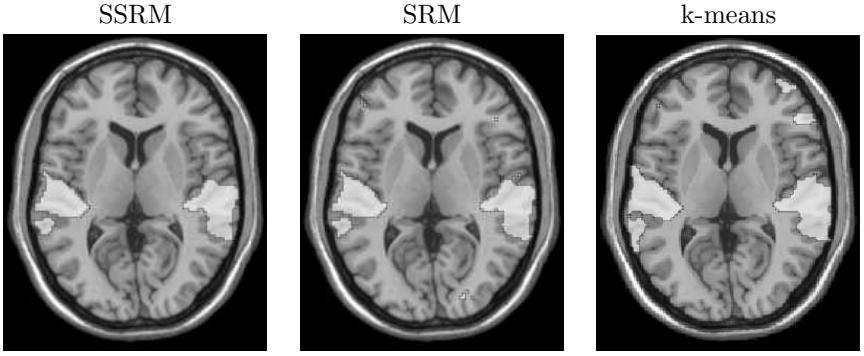
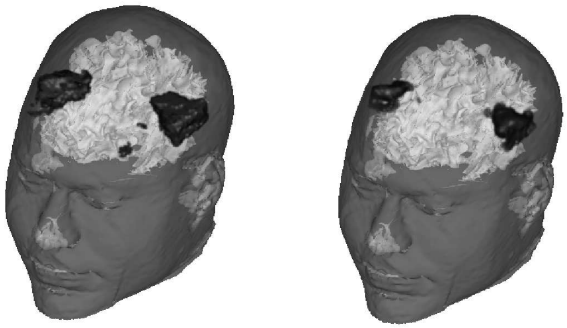


Fig. 5. The binary activation map as estimated by each method in the case of the auditory experiment.

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 [47] that sequentially produces the 3D head model with the activation areas. Figure 6 illustrates the resulting 3D models of our method and the standard GLM approach [8]. Obviously, 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.

In the event-related foot-movement experiment we analyzed fMRI data consisted

[a] http://www.fil.ion.ucl.ac.uk/spm/



(a) (b)

Fig. 6. The 3D head activation maps as estimated by (a) the proposed method SSRM and (b) the standard GLM.

of images acquired from the University Hospital of Ioannina, Greece [48]. Details about the protocol that was followed for constructing the fMRI data can be found in [22]. Experiments were made with the slice 54 of this dataset, which contains a number of $N = 2644$ time series. Figure 7 presents the comparative results in this dataset overlaid on greyscale T1 weighted anatomical images. 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. The proposed regression mixture model gives more activated areas closer to the established motor circuitry and therefore the results are more reasonable (at least in this case).

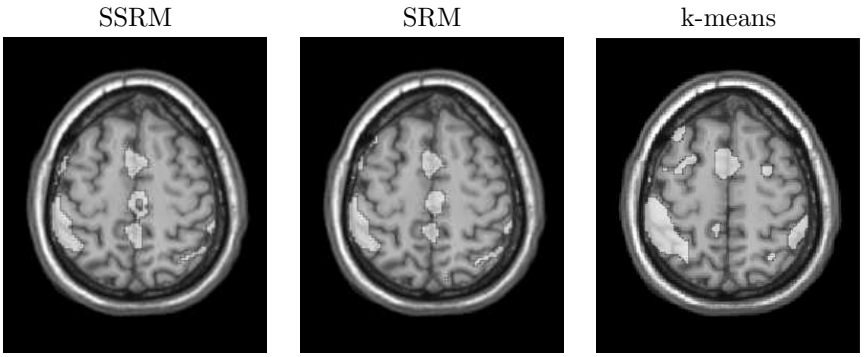


Fig. 7. Estimated motor activated areas of comparative methods in white overlaid on greyscale T1 weighted anatomical images.

5.2. *Resting state fMRI experiments*

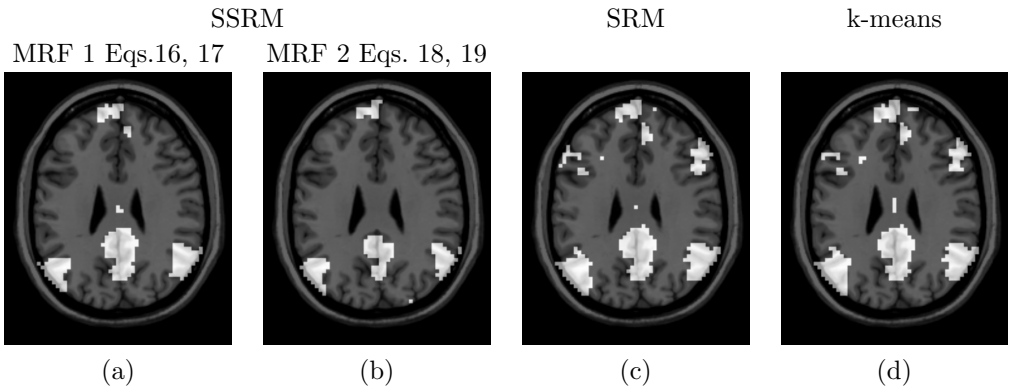


Fig. 8. Default Mode Network estimation from resting state fMRI data using two versions of SSMR (a) and (b) that differ in the type of MRF-based spatial prior, (c) the SRM and (d) the k-means approaches

We have also made experiments with resting state fMRI data obtained from the Neuroimaging Informatics Tools and Resources Clearinghouse (NITRC) [b] under the project name: NYU CSC TestRetest. A detailed description of the dataset can be found in [49]. In our experiments we have selected a subject from this dataset, where we have used the slice 34 (of 52). Two versions of the proposed SSRM method were studied that differ on the spatial prior: (a) SSRM with the exponential potential function of Eqs. 16, 17 and (b) SSRM with the softmax MRF prior of Eqs. 18, 19. The task in these series of experiments was to estimate the Default Mode Network (DMN), which is a resting state network that consists of precuneus, medial frontal, inferior parietal cortical regions and medial temporal lobe. The DMN is expected to be presented in almost every fMRI experiment.

The depicted brain images are shown in Fig. 8 that illustrate the DMN cluster as estimated by each clustering approach. What is interesting to observe is that, although all methods are able to properly identify the DMN, the other two methods, SRM and k-means, tend to overestimate it and construct small islands (almost uniformly the brain) not belonging to DMN. Clearly, the proposed SSRM method seems to have better discrimination ability to discover more accurately the network. Both versions do not show any significant difference, with the second version (b) to appear to be more consistent producing slightly smoother regions. However, a more systematic comparative study is required to evaluate effectiveness of them.

[b] http://www.nitrc.org

6. Conclusions

In this chapter we have presented a regression mixture modeling framework for the analysis of fMRI data. This model class is very flexible and can embody prior knowledge of the nature of data. The key aspect of the proposed technique lies on the superior sparse regression performance to model data of latent classes, as well as the ability to evoke responses which are spatially homogeneous and locally contiguous. It also includes a multi-kernel scheme for composing the kernel matrix of each component that offers better fitting capabilities. Therefore, the proposed method manages to incorporate significant physiological properties of human brain and to tackle important issues that they are possible to deteriorate the performance of fMRI data analysis. As compared to standard approaches, the sparse and spatial regularization procedures of the method have been shown to increase the robustness of detection and to result in inferences with higher sensitivity.

Further extensions of the finite mixtures are possible for the regression case. Instead of using GLMs as component specific models, generalized additive models can be used which allow to relax some assumptions we have made. Another future research direction is to examine the possibility of applying alternative sparse priors, as well as to assume Student's-t type of distribution for modeling the noise (instead of Gaussian) so as to achieve more robust inference and handle outlying observations [24]. Finally, another possibility is to extend our model to $3-D$ cases and to group analysis applications.

References

1. "Human brain project." https://www.humanbrainproject.eu/, 2012.
2. "Brain initiative." http://www.nih.gov/science/brain/, 2013.
3. S. Ogawa, T. Lee, A. Kay, and D. Tank, "Brain magnetic resonance imaging with contrast dependent on blood oxygenation," *Proceedings of the National Academy of Sciences (USA)*, vol. 87, pp. 9868 – 9872, 1990.
4. K. Friston, "Statistical parametric mapping." http://www.fil.ion.ucl.ac.uk/spm/, 2009.
5. M. Jenkinson, C. Beckmann, T. Behrens, M. Woolrich, and S. Smith, "FSL," *NeuroImage*, vol. 62, pp. 782–90, 2012.
6. N. A. Lazar, *The Statistical Analysis of Functional MRI data*. Springer, 2008.
7. R. A. Poldrack, J. A. Mumford, and T. E. Nichols, *Handbook of Functional MRI Data Analysis*. Campridge University Press, 2011.
8. K. J. Friston, "Analysis of fMRI time series revisited," *Neuroimage*, vol. 2, pp. 45–53, 1995.
9. G. Flandin and W. Penny, "Bayesian fMRI data analysis with sparse spatial basis function priors," *NeuroImage*, vol. 34, pp. 1108–1125, 2007.
10. W. Penny, N. Trujillo-Barreto, and K. Friston, "Bayesian fMRI time series analysis with spatial priors," *NeuroImage*, vol. 24, pp. 350–362, Jan. 2005.
11. 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.

12. 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.
13. 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.
14. 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.
15. 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,*, vol. 57, pp. 1967 – 1976, 2011.
16. H. Luo and S. Puthusserypady, "fMRI data analysis with nonstationary noise models: A bayesian approach," *IEEE Transactions on Biomedical Engineering*, vol. 54, pp. 1621–1630, Sept. 2007.
17. 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.
18. 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.
19. N. Hartvig and J. Jensen, "Spatial mixture modeling of fMRI data," *Human Brain Mapping*, vol. 11, no. 4, pp. 233–248, 2000.
20. M. Woolrich, M. Jenkinson, J. Brady, and S. Smith, "Fully bayesian spatio-temporal modeling of fMRI data," *IEEE Transactions on Medical Imaging*, vol. 23, pp. 213–231, Feb. 2004.
21. V. Oikonomou, K. Blekas, and L. Astrakas, "A sparse and spatially constrained generative regression model for fMRI data analysis," *IEEE Transactions on Biomedical Engineering*, vol. 59, pp. 58–67, 2012.
22. V. P. Oikonomou and K. Blekas, "An Adaptive Regression Mixture Model for fMRI Cluster Analysis," *IEEE Transactions on Medical Imaging*, vol. 32, pp. 649–660, April 2013.
23. H. Luo and S. Puthusserypady, "A sparse Bayesian method for determination of flexible design matrix for fMRI data analysis," *IEEE Trans. on Circuits and Systems I: Regular Papers*, vol. 52, pp. 2699–2706, 2005.
24. C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2006.
25. R. Baumgartner, C. Windischberger, and E. Moser, "Quantification in Functional Magnetic Resonance Imaging: Fuzzy Clustering vs. Correlation Analysis," *Magn. Reson. Imaging*, vol. 16, pp. 115 – 125, 1998.
26. 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.
27. 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.
28. 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.
29. 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.
30. 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," *Artif. Intel. in*

Medicine, vol. 29, no. 3, pp. 203 – 223, 2003.
31. 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.
32. 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.
33. 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.
34. M. Woolrich, T. Behrens, C. Beckmann, and S. Smith, "Mixture models with adaptive spatial regularization for segmentation with an application to FMRI data," *IEEE Trans. on Med. Imaging*, vol. 24, pp. 1–11, 2005.
35. W. Penny and K. Friston, "Mixtures of general linear models for functional neuroimaging," *IEEE Trans. on Med. Imaging*, vol. 22, pp. 504 –514, 2003.
36. 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 EMBS*, pp. 4795–4798, 2009.
37. M. E. Tipping, "Sparse Bayesian Learning and the Relevance Vector Machine," *Journal of Mach. Learn. Research*, vol. 1, pp. 211–244, 2001.
38. 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.
39. G. M. McLachlan and D. Peel, *Finite Mixture Models*. New York: Wiley, 2001.
40. 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.
41. 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.
42. 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.
43. V. Oikonomou, K. Blekas, and L. Astrakas, "Resting state fmri analysis using a spatial regression mixture model," in *Bioinformatics and Bioengineering (BIBE), 2013 IEEE 13th International Conference on*, pp. 1–4, Nov 2013.
44. S. Gunn and J. Kandola, "Structural modelling with sparse kernels," *Machine Learning*, vol. 48, pp. 137–163, 2002.
45. M. Girolami and S. Rogers, "Hierarchic Bayesian models for kernel learning," in *ICML '05: Proceedings of the 22nd Intern. Conf. on Machine Learning*, (New York, NY, USA), pp. 241–248, ACM, 2005.
46. J. Nocedal and S. J. Wright, *Numerical Optimization*. Springer-Verlag New York, Inc., 1999.
47. S. Pieper, M. Halle, and R. Kikinis, "3D SLICER," *IEEE International Symposium on Biomedical Imaging ISBI 2004*, pp. 632–635, 2004.
48. 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.
49. Z. Shehzad, A. M. C. Kelly, P. T. Reiss, D. G. Gee, K. Gotimer, L. Q. Uddin, S. H. Lee, D. S. Margulies, A. K. Roy, B. B. Biswal, E. Petkova, F. X. Castellanos, and M. P. Milham, "The resting brain: Unconstrained yet reliable," *Cerebral Cortex*, vol. 19, no. 10, pp. 2209–2229, 2009.