On Clustering fMRI Time Series

Cyril Goutte, Peter Toft, Egill Rostrup[†],

Finn Å. Nielsen and Lars Kai Hansen

Department of Mathematical Modelling, Building 321,

Technical University of Denmark

cg,pto,fn,lkhansen@imm.dtu.dk

[†] The Danish Research Centre for Magnetic Resonance,

Hvidovre, Denmark, egillr@magnet.drcmr.dk

June 26, 1998

Running title: Clustering fMRI time series.

Address for correspondence:

Cyril Goutte

Department of Mathematical Modelling, Building 321,

Technical University of Denmark, Denmark

Tel: +45 4525 3921

Fax: +45 4587 2599

E-mail: cg@imm.dtu.dk

ABSTRACT

Analysis of fMRI time series is often performed by extracting one or more parameters for the individual voxel. Methods based e.g. on various statistical tests are then used to yield parameters corresponding to probability of activation or activation strength. However, these methods do not indicate whether sets of voxels are activated in a similar way, or activated in different ways. Typically, delays between two activated signals are not identified. In this article, we use clustering methods to detect similarities in activation between voxels. We employ a novel metric which measures the similarity between the activation stimulus and the fMRI signal. We present two different clustering algorithms and use them to identify regions of similar activations in an fMRI experiment involving a visual stimulus.

INTRODUCTION

In the recent years many contributions have addressed the analysis of fMRI time series. A large number of models and techniques from signal processing and statistics have been applied to fMRI analysis. Several flavours of statistical tests have been used (Xiong et al., 1996). The t-test implemented in SPM (Worsley and Friston, 1995), derived from the well-known general linear model (McCullagh and Nelder, 1989), and the non-parametric Kolmogorov-Smirnov test (Baker et al., 1994) are the most widespread examples. The correlation between the fMRI signal and the activation paradigm has also been used in different contexts (Bandettini et al., 1993; Golay et al., 1997), while linear filters, like the finite input response (FIR) filter, are slowly emerging as a possible alternative (Lange and Zeger, 1997; Nielsen et al., 1997). The above methods focus solely (at least in a first stage) on estimating either the probability or the strength of activation on a voxel by voxel basis.

In this contribution we consider an alternative approach. We assume that the pattern of activation actually has a structure, and can be divided into a few types of similar activations. To each of these types corresponds a cluster of similarly activated voxel, the centre of which represents the “typical” time series for these voxels. Subsequently, cluster centres can be analysed with regard to descriptive parameters such as activation strength and delay. Clustering techniques provide additional information, namely the cluster assignments, ie labels for each of the voxels according to their similarity. It is therefore possible to isolate zones with similar activation, as well as to see whether two given voxels have similar behaviour.

Clustering methods have been previously used in neuroimaging for similar purposes (Baumgartner et al., 1997, 1998; McIntyre et al., 1996; Moser et al., 1997; Scarth et al., 1996). These contributions performed a clustering directly on the fMRI time series, using the fuzzy K-means algorithm (see Davé and Krishnapuram, 1997, for a general review). Due to the high noise level in fMRI experiments, the results of clustering on the raw time series is often unsatisfactory and does not necessarily group data according to the similarity of their pattern of response to the stimulus. This consideration has led Golay et al. (1997) and Toft et al. (1997), in two independant abstracts for the Human Brain Conference, to consider a metric based on the correlation between stimulus and time series.

Toft et al. (1997) illustrated the stability problems due to the high noise level in the raw data, and suggested to cluster voxels on the basis of the cross-correlation function, yielding improved performance and noise reduction.

The aim of this contribution is to focus on the application of clustering to fMRI time series using two different algorithms. The well-known K-means algorithm is a simple method with a fast convergence, but also a number of limitations based on its underlying parametric assumptions. As an alternative, we present a hierarchical method which addresses a number of these limitations by providing a different outlook on the clustering problem. We provide the theoretical basis for both techniques, suggest a simple stochastic procedure to choose the initial set of cluster centres in the K-means method, and discuss the issue of the number of clusters. In this study, the emphasis is on *exploratory*, rather than *inferential*, data analysis; however, inferences can be drawn from the clustering results and we provide some ways to do so. In order to illustrate these ideas, a number of experiments are performed on a set of fMRI images obtained from a visual experiment. This contribution extends our previous results and provides additional tools and methods for clustering fMRI time series.

Let us finally note that clustering provides a general tool to perform post-processing with a number of methods. It can be applied, among other possibilities, on low-dimensional features extracted from the original data (Goutte et al., 1998b), statistical tests results or FIR coefficients after a linear filtering.

In the following section, we present the dataset used in this study, introduce the necessary concepts and methods and insist on the role of the metric. We then present the results obtained with both clustering algorithms in different configurations. In particular we use the hierarchical method to provide a heuristic to choose the number of clusters. Finally, the discussion section addresses the neuroscientific aspects of this work and discusses some statistical issues.

MATERIALS AND METHODS

Dataset

The experiments discussed below will be performed on a dataset acquired at Hvidovre Hospital in Denmark on a 1.5 T Magnetom Vision MR scanner. The scanning sequence was a 2D gradient echo EPI (T2* weighted) with 66 ms echo time. The RF flip angle was set to 50 degrees, and a scan target was a matrix of 128x128 pixels, with FOV of 230 mm, and the slice thickness was 10 mm. Images were obtained in a para-axial orientation parallel to the calcarine sulcus. The region of interest will be limited to a 71×91 pixels map.

The visual paradigm consisted of a rest period of 20 seconds of darkness (using a light fixation dot), followed by 10 seconds of full-field checker board reversing at 8 Hz, and ending by 20 seconds of darkness. A total of 150 images was obtained in a run of 50 seconds, corresponding to approximately 330 ms between the images. 10 separate runs containing 150 images each were completed. For computational reasons, the dataset used in this article was built by using 3 of these runs. Furthermore, the first and last 25 scans in each run where left out, so that the assembled data consists of a total of 300 time samples (3 runs of 100 images) for each voxel.

Analytical tools

Let us first introduce a number of useful quantities. Let $\{\boldsymbol{z}_j\}$ be a set of N vectors from $I\!R^P$, eg the fMRI time series in each of N voxel—in which case P is the number of images. Let us consider K clusters, represented by their cluster centre $\boldsymbol{c}_k \in I\!R^P$, with $1 \leq k \leq K$. Each cluster C_k is a set of indexes from $\{1, \ldots N\}$. The clusters are a partition of the data so that each vector $\boldsymbol{z}_j$ belongs to exactly one cluster. Clustering consists of assigning each vector $\boldsymbol{z}_j$ to a cluster C_k. The within-class (or intra-class) inertia of the resulting partition is:

$$\mathcal{I}_W = \frac{1}{N} \sum_{k=1}^{K} \sum_{j \in C_k} d^2\left(\boldsymbol{z}_j, \boldsymbol{c}_k\right) \tag{1}$$

and the between-class (or inter-class) inertia is:

$$\mathcal{I}_B = \frac{1}{N} \sum_{k=1}^{K} |C_k| d^2\left(\boldsymbol{c}_k, \overline{\boldsymbol{c}}\right) \tag{2}$$

where $d^2(\boldsymbol{a}, \boldsymbol{b})$ is the squared distance between vectors $\boldsymbol{a}$ and $\boldsymbol{b}$, $|C_k|$ is the number of elements in cluster C_k, and $\overline{\boldsymbol{c}} = \sum_{k=1}^{K} \frac{|C_k|}{N} \boldsymbol{c}_k$, is the weighted average of the cluster centres. Intuitively, $\mathcal{I}_W$ is the average squared distance from a point to its cluster centre, while $\mathcal{I}_B$ is the average squared distance from a cluster centre to the centre of gravity. A commendable goal in clustering would thus be to minimise the within-class inertia in order to have homogeneous clusters, while maximising the between-class inertia so that these clusters are as different as possible.

For a large class of distance $d(\cdot, \cdot)$, the inertia of each cluster (inner sum in equation 1) is minimised when the cluster centre is the average of all cluster members: $\boldsymbol{c}_k = \dfrac{1}{|C_k|} \sum\limits_{j \in C_k} \boldsymbol{z}_j$. Under these conditions, the average cluster centre is also the average of the data, ie the centre of gravity: $\overline{\boldsymbol{c}} = \overline{\boldsymbol{z}}$. $\mathcal{I}_W$ and $\mathcal{I}_B$ thus become the intra-class and inter-class *variances*. According to Huygens' formula, the sum of within- and between-class variances is constant and equal to the total data variance, regardless of the number of clusters or their compositions. Thus minimising $\mathcal{I}_W$ or maximising $\mathcal{I}_B$ is equivalent. Accordingly, the within-class inertia alone provides a possible way of assessing the quality of a partition of K clusters, but it does in no way make it possible to compare two partitions with different numbers of clusters. In particular, the within-class inertia of the optimal partition with K clusters is always higher than that of the optimal partition with $K+1$ clusters. Furthermore, it can be noticed that $\mathcal{I}_W$ is globally minimised by the trivial partition of N clusters containing one point each.

K-means

The above considerations provide a natural introduction to one of the most widely used clustering techniques: the K-means algorithm (MacQueen, 1967; Hartigan and Wong, 1979). For a given number K of clusters, the within-class inertia is iteratively minimised by assigning data to the nearest center and recalculating each centre as the average of its members (minimising eq. 1):

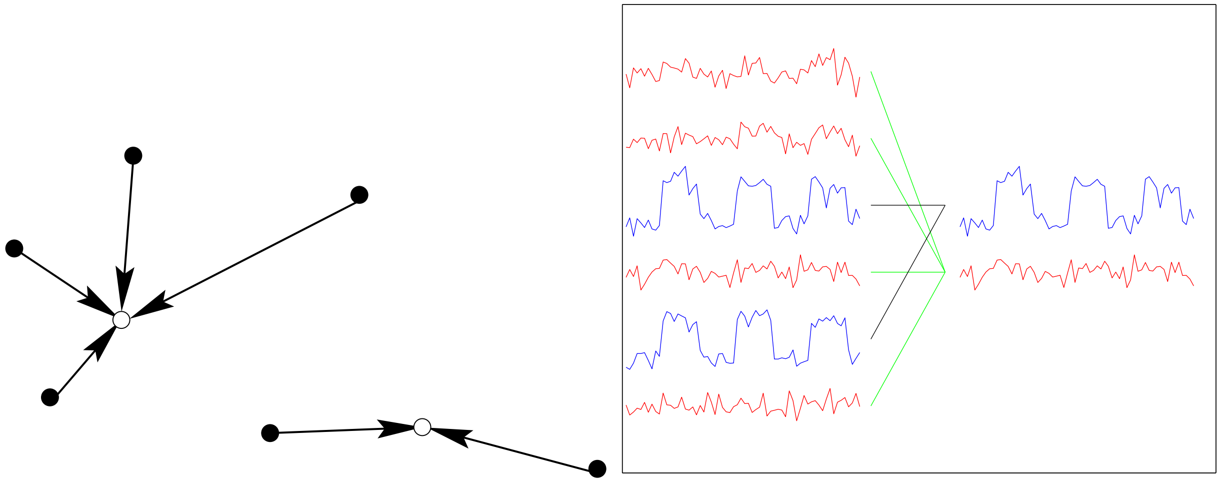

Figure 1: Left: two-dimensional projection of the assignment of data (black circles) to the closest centre (white circles). Right: its implication for the corresponding time series. The six data vectors (left hand side) are assigned to two cluster centres (right hand side).

1. Initialise K clusters $k = 1 \ldots K$, with centres $\boldsymbol{c}_k{}^{(0)}$. Iteration $i = 0$.

2. Assign each data vector $\boldsymbol{z}_j$ to the cluster C_k with the nearest centre $\boldsymbol{c}_k{}^{(i)}$, based on a distance metric between the cluster centre and the data vector, $d\left(\boldsymbol{z}_j, \boldsymbol{c}_k{}^{(i)}\right)$.

3. Set new cluster centre $\boldsymbol{c}_k{}^{(i+1)}$ to the average of its members: $\boldsymbol{c}_k{}^{(i+1)} = \frac{1}{|C_k|} \sum_{j \in C_k} \boldsymbol{z}_j$

4. Increment i and go to step 2 until the partition is stable.

Both steps 2 and 3 decrease the within-class inertia, so that the algorithm converges in a finite number of steps. The convergence is usually very fast (Bottou and Bengio, 1995) and the algorithm requires to store and consider only $K \times N$ distances between the data and the centres. For fMRI clustering, each data vector $\boldsymbol{z}_j$ could be the time series measured in voxel j. The cluster centre $\boldsymbol{c}_k$ would then also be a time series, representing the "typical" response for this group of voxels. Figure 1 shows a typical K-means clustering step, and its implication for fMRI time series.

Note that the results are very dependent on a number of factors. The algorithm relies on the parametric assumption that the data distribution is a mixture of K identical components. The first implication is that the metric implemented by the distance $d(\cdot, \cdot)$ has a large influence on the result. More important, the number K of clusters must be specified in advance. When the chosen number is not reflected in the data, the results might end up being essentially meaningless. Lastly, K-means is a non-deterministic algorithm and the resulting partition depends on the initial clusters assignment (step 1 above). A useful heuristic is to use several random assignment and select the best result according to some criteria, eg the intra-class inertia.

Hierarchical clustering

The hierarchical algorithm addresses a number of limitations of the K-means method by adopting a different outlook. Biologists, for example, cluster data using taxonomic hierarchies. Plants or animals are grouped in species, which are in turn grouped in genera, then families, orders, classes and finally phyla. Each level of the taxonomy gathers several members of the previous level. Hierarchical methods (see Ripley (1996), section 9.3 for a general introduction) proceed from this idea. They iteratively join clusters that are the most similar into a larger structure. The result is usually presented in a tree-like structure, the dendrogram, which shows which groups have been joined at which level of similarity. This circumvents one of the main drawbacks of the K-means algorithm, as we do not need to specify the number of clusters in advance: the hierarchical scheme provides different partitions obtained by cutting the tree at different levels. These are only locally optimal, in the sense that each K-cluster partition is the best possible starting from the $K+1$ groups in the previous level, but not necessarily the best possible K-cluster partition starting from the initial data. Furthermore the process is entirely deterministic.

In the following algorithm, known as the group-average agglomerative method and attributed to Ward (1963), we start with one cluster per data vector. The two closest points/clusters are joined into one cluster, resulting in $N-1$ clusters: $N-2$ containing one vector, and one containing two data points. The same operation is carried out with the $N-1$ resulting centres, and so on:

1. Initialise by assigning one cluster of unit weight $w_j = 1$ to each data vector $\boldsymbol{z}_j$. Calculate the squared dissimilarities $\delta_{i,j} = \frac{1}{N} d^2 \left(\boldsymbol{x}_i, \boldsymbol{x}_j\right)$ between clusters $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$.

2. Join the least dissimilar clusters A and B into a new cluster $A \cup B$ of weight $w_{A\cup B} = w_A + w_B$.

3. For all clusters C different from A or B, update the dissimilarities by the formula:

$$\delta_{C,A\cup B} = \frac{(w_A + w_C)\,\delta_{A,C} + (w_B + w_C)\,\delta_{B,C} + w_C\delta_{A,B}}{w_A + w_B + w_C}$$

4. Iterate: go to step 2 until there is only one cluster left.

The computational burden lies in the calculation of the dissimilarities in step 1. The algorithm requires to calculate, store and consider an order $N \times N$ dissimilarities. This is much more demanding than the K-means algorithm for small values of K. Note however that once the original $N \times N$ matrix of dissimilarities is obtained (step 1 above), the update formula from step 3 makes each iteration very fast. Furthermore, we obtain all partitions, for K varying from 1 to N in only one pass. Despite a lesser demand for each individual clustering attempts, estimating several partitions from $K = 1$ to $K = K_{max}$ clusters with K-means using the random initialisation heuristics could turn out to be computationally comparable to hierarchical clustering.

The Metric

Both clustering algorithms above rely on the use of a metric, ie a definition of distances between two points in P-dimensional space. The resulting partition is potentially highly dependent on the particular choice of metric. A fairly broad class of metrics can be obtained by defining the generalised distance (Mahalanobis, 1936) between two vectors $\boldsymbol{a}$ and $\boldsymbol{b}$ in $I\!R^P$ as:

$$d^2(\boldsymbol{a}, \boldsymbol{b}) = (\boldsymbol{a} - \boldsymbol{b})^\top \mathbf{D}(\boldsymbol{a} - \boldsymbol{b}) \qquad (3)$$

where $\mathbf{D}$ is a $P \times P$ symmetric positive definite matrix that uniquely defines the metric. For $\mathbf{D} = \mathbf{I}_P$ (the identity matrix), we have the standard Euclidean distance. If $\mathbf{D}$ is a diagonal matrix with positive elements on the diagonal, we have a *scaling* metric. When the diagonal contains the inverse variance of the data on each coordinate, this will be equivalent to using Euclidean distance on the normalised data. For other choices of symmetric positive definite distance matrix, there exists a matrix $\boldsymbol{T}$ such that $\boldsymbol{D} = \boldsymbol{T}^\top\boldsymbol{T}$. This means that the corresponding metric is equivalent to a Euclidean distance after a linear data transformation given by $\boldsymbol{T}$. $\boldsymbol{\Sigma}$, the $P \times P$ covariance matrix of the data, leads to Euclidean or scaling distance in the principal component axes. Indeed, let us write the eigenvalue decomposition $\mathbf{\Sigma} = \mathbf{U}^\top \mathbf{\Lambda} \mathbf{U}$. The generalised metric using $\boldsymbol{D} = \mathbf{\Sigma}^{-1}$ becomes:

$$d^2(\boldsymbol{a}, \boldsymbol{b}) = (\boldsymbol{U}\boldsymbol{a} - \boldsymbol{U}\boldsymbol{b})^\top \mathbf{\Lambda}^{-1} (\boldsymbol{U}\boldsymbol{a} - \boldsymbol{U}\boldsymbol{b}) \tag{4}$$

By editing the diagonal elements of $\mathbf{\Lambda}$, we obtain a number of interesting metrics like the Euclidean or scaling distance in any principal subspace projection.

Alternatively, equation 3 allows us to perform an implicit linear filtering of the data. Let us write the filtered data as $\boldsymbol{X} = \boldsymbol{F}\boldsymbol{Z}$, where $\boldsymbol{Z} = [\boldsymbol{z}_1, \ldots \boldsymbol{z}_N]$ is a $P \times N$ matrix containing the original data, $\boldsymbol{F}$ is a matrix of filter coefficients, and $\boldsymbol{X} = [\boldsymbol{x}_1, \ldots \boldsymbol{x}_N]$ is the matrix containing filtered data. It is equivalent to use a Euclidean distance on the filtered data or to use a generalised metric (3) on the original data with $\boldsymbol{D} = \boldsymbol{F}^\top \boldsymbol{F}$. The pre-processing presented below is a typical example of such use.

Pre-processing

Some previous attempts at clustering fMRI time series (Baumgartner et al., 1997, 1998; McIntyre et al., 1996; Moser et al., 1997; Scarth et al., 1996) use the raw time series measurement as input. A limitations of this approach is the potentially high dimensional space—especially for fast sampling rates. Using all 10 experiments from our dataset, the resulting fMRI time series would belong to a 1500-dimensional space. A second problem is the high noise level, which leads to stability problems and the risk of clustering on the noise rather than on the activation. An associated concern is that we are actually interested in the similarity in temporal activation, especially in connection with the stimulus (Golay et al., 1997; Toft et al., 1997). This has led Toft et al. (1997) to propose clustering on the cross-correlation function between the fMRI activation and the paradigm. For voxel j, $\boldsymbol{y}_j$ denotes the measured fMRI time series, and $\boldsymbol{p}$ is the activation stimulus, common to all j, usually, but not limited to, a square wave ("box-car model"). The cross-correlation function is defined as:

$$x_j(t) = \frac{1}{P} \sum_{u=1}^{P} y_j(u) p(u - t) \tag{5}$$

where we force $p(i) = 0$ for $i < 1$ or $i > P$. Equation 5 is known as the *biased* estimator. The cross-correlation function often has a periodic structure, so that it is possible to truncate $\boldsymbol{x}_j$, retaining a limited interval centred on 0. Note that the cross-correlation function is a linear filter, and (5) can be expressed as:

$$\boldsymbol{X} = \frac{1}{P} \begin{bmatrix} p(T{+}1) & \cdots & p(N) & 0 & \cdots & 0 \\ \vdots & & & \ddots & \ddots & \vdots \\ p(2) & p(3) & \cdots & & p(N) & 0 \\ p(1) & p(2) & p(3) & \cdots & p(N{-}1) & p(N) \\ 0 & p(1) & p(2) & \cdots & p(N{-}2) & p(N{-}1) \\ \vdots & \ddots & \ddots & & & \vdots \\ 0 & \cdots & 0 & p(1) & \cdots & P(N{-}T) \end{bmatrix} \boldsymbol{Y} = \boldsymbol{F}\boldsymbol{Y} \tag{6}$$

where we have retained only those coefficients for which t lies between $-T$ and T. Accordingly, clustering on the cross-correlation function can be viewed as the use of an alternative metric. Furthermore, T is of the order of the stimulus period, so that the resulting vector space has much lower dimension than the original time-series. Finally, note that the cross-correlation *function* is different from the cross-correlation *coefficient* used eg by Bandettini et al. (1993) and Golay et al. (1997).

A Two Stage Strategy

Most fMRI experiments provide a wealth of data. Though fMRI time series are measured in numerous voxels, only very few of them are activated. This poses a problem for clustering because the underlying groups are ill-balanced. For example, K-means might have difficulties isolating possibly activated clusters and spread the clusters over the non-activated voxels instead. An additional concern is the computational cost, which grows as the square of the number of data vectors for our hierarchical method. In order to reduce the amount of data, we propose a *two-stage* strategy in which we first use a loose statistical test to discard voxels that are almost surely non-activated, then cluster the remaining data. A possible strategy would be to use a simple F-test (Holmes and Friston, 1997, section 6.3) or other statistical tests along the same lines, and threshold at a given level. It should be noted that the traditional use of statistical testing in neuroimaging puts the emphasis on the type I error, or risk of false positives. In the context of our study, this thresholding is used solely as a data reduction device. We will consequently be more

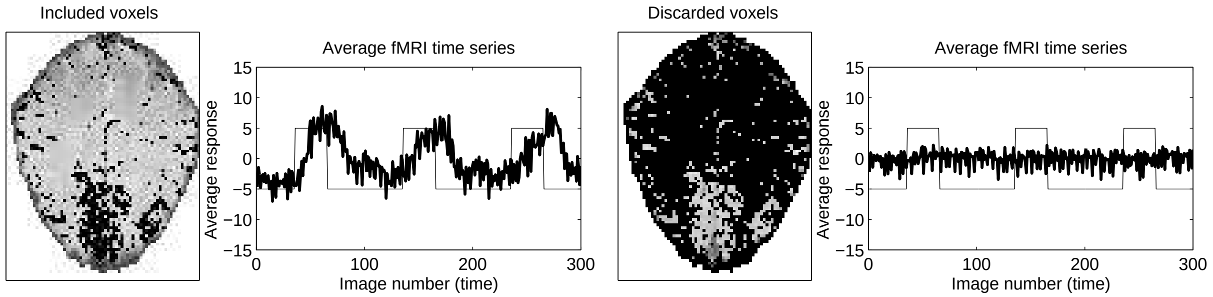

Figure 2: Brain map and average activation for the thresholded and discarded voxels. Left: the 696 voxels selected after thresholding the maximum of the cross-correlation function are indicated in black on top of the anatomical reference (average value of the fMRI time series in each voxel). Right: the discarded voxels, indicated in black, cover 84% of the slice. fMRI plots: average fMRI time series for the voxels indicated in black on the corresponding brain map.

interested in lowering the type II error, so that we minimise the risk of discarding possibly activated voxels.

As the cross-correlation function forms the basis of the clustering method, we will also use it to reduce the data in this two-stage strategy. We consider the extreme value of the cross-correlation function as the statistic of interest, and the null hypothesis that brain activation is only Gaussian noise, uncorrelated with the stimulus and with variance σ^2. According to (6), the cross-correlation coefficients $x_j(t)$ will have a multivariate Gaussian distribution, with covariance $\boldsymbol{F}\boldsymbol{F}^\top/\sigma^2$. To our knowledge, there is no simple expression giving the distribution of the maximum coordinate of vectors sampled from a general multivariate Gaussian. However, it is easy to sample from such a distribution[1] and obtain a Monte-Carlo estimate of the p-value associated with the maximum cross-correlation coefficient measured in a given voxel (Goutte et al., 1998a).

In the experiments presented below, we use a low cross-correlation threshold in order to minimise the risk of discarding activated voxels. After thresholding, we retain 696 voxels out of 4391, ie 16%. Figure 2 shows the selected voxels, marked in black. For anatomical reference, the background represents the mean fMRI activation, averaged over time for each voxel. The corresponding time series, averaged over all selected voxels, are plotted on the right of each brain map.

[1] Note that σ^2 shall be estimated from the data, and the resulting statistic will have a multivariate t-distribution. However, when the number of images is moderately high, the Gaussian approximation will hold.

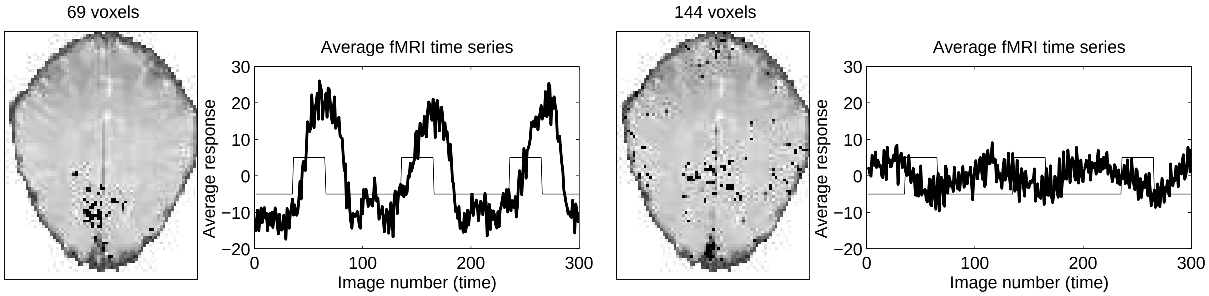

Figure 3: Two of the three clusters obtained in our first K-means experiment ($K = 3$). Brain maps: cluster members indicated in black on top of the anatomical reference. fMRI plots: average fMRI time series in the corresponding voxels in thick black line, paradigm (stimulus) plotted as a reference in thin black.

RESULTS

K-means

We first use K-means clustering on the thresholded data using 3 clusters. The motivation is to try to isolate two clusters of activated voxels with different types of activation, while leaving a cluster for non-activated or weakly activated voxels. Each data vector contains elements $x_j(-24)$ to $x_j(25)$ with the corresponding 50 values of the cross-correlation function between the fMRI time series and the activation paradigm. We use the stochastic initialisation procedure described above with 100 random initial configurations. The resulting partitions turn out to be very similar, with within-cluster variances between 30.15 and 30.17 (standard deviation 0.007), and only 8 distinct configurations.

Figure 3 presents the results for the best partition, ie the lowest within-cluster variance. One cluster (left) contains 69 voxels, located mostly in the visual cortex. The average time series in these voxels shows that their response is highly correlated with the paradigm. The delay, defined as the location of the largest absolute value of the cross-correlation function, is around 15 images or 5 seconds. The second cluster (right) contains 144 members. Though a number of voxels from this group are distributed across the slice, a majority of them are located in two areas: the neighbourhood of the visual cortex, close to members of the previous cluster, and the *sinus sagittalis* (bottom). Interestingly, the average fMRI time series in this second cluster suggests a negative correlation with the paradigm (right plot). However, the relatively modest level of the correlation suggests that this average

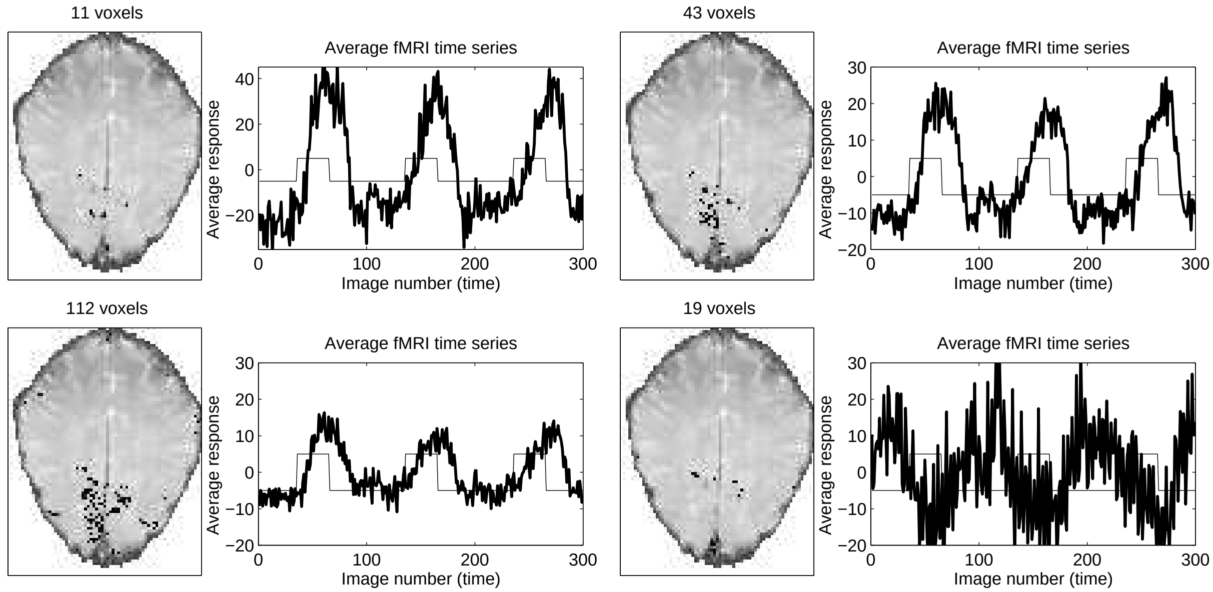

Figure 4: Four of the seven clusters obtained in our second K-means experiment ($K = 7$). Brain maps: cluster members indicated in black on top of the anatomical reference. fMRI plots: average fMRI time series in the corresponding voxels in thick black line, paradigm (stimulus) plotted as a reference in thin black.

effect might not be significant against the null hypothesis of no activation. The third and final cluster contains the remaining, weakly-correlated voxels (not plotted).

A second experiment is performed involving 7 clusters. The stochastic initialisation heuristic is used again with 100 random initial conditions. The resulting partitions are more varied than in the 3-cluster case, with 60 distinct configurations. The minimum and maximum within-cluster variances are 15.67 and 18.14 respectively, with a mean of 16.45 and a standard deviation of 0.42. Figure 4 presents four of the seven clusters in the best partition. The first three (top row and bottom left) are positively correlated with the paradigm and are displayed here in decreasing order of their maximum cross-correlation. Notice that the average response strength in the first cluster (top left, 11 voxels) is almost three times higher than that of the third cluster (bottom left, 112 voxels). It is also sharper and with a slightly shorter delay compared to the second and third clusters. This difference in delay is naturally accounted for by the cross-correlation metric. The three positive clusters are located mainly the visual cortex. In addition, some of the less activated voxels cover two lateral areas that could correspond to visual area V5.

The fourth cluster in figure 4 (bottom right) contains 19 voxels with two noticeable features. They are anti-correlated with the stimulus, like the voxels gathered in the second cluster in figure 3, though with a larger cross-correlation (hence a more significant effect). Secondly, the fMRI signal contains a high frequency component with a period of around 4 images. Due to the high sampling rate used to collect this dataset, this corresponds to a frequency slightly lower than 1Hz which turns out to reflect the heart beat. This is supported by the fact that this cluster contains voxels that cover the *sinus sagittalis*, located at the back of the brain (bottom of the slice, see also figure 3). The rest of the thresholded voxels are weakly correlated with the stimulus and are distributed in the three remaining clusters (not shown).

Hierarchical Clustering

As noted above, the use of K-means poses a crucial problem: how many clusters should we consider? The choice of three clusters could be justified by our attempt to identify two zones with different activation patterns. But what if there are more such patterns (eg short, medium and long term delays), or conversely only one? Furthermore, in our second experiment, there is no real rationale behind the choice of $K = 7$. Hierarchical clustering provides an answer to these questions and a principled way to decide on the number of clusters that provide a good balance between the number of classes and their homogeneity. Let us apply Ward's hierarchical clustering method presented above to the 696 voxels obtained after thresholding. Each data vector contains 50 values, $x_j(-24)$ to $x_j(25)$, of the cross-correlation between the fMRI time series and the activation paradigm.

In one deterministic pass, the hierarchical algorithm provides a dendrogram (Ripley, 1996, p. 320), ie a binary tree representing the way each cluster is composed of clusters obtained in previous steps. The tree can be cut at several levels in order to obtain an arbitrary number of clusters. Figure 5 displays the resulting cluster centres when the tree is cut at different levels, corresponding to from 7 down to 2 clusters. This figure gives an interesting insight into the way hierarchical clustering operates. In each plot, each curve is a cluster centre, representing the "typical" cross-correlation function of the voxels in the associated cluster. The dotted line is the cross-correlation of the paradigm with itself or auto-correlation. It allows us to assess the delay in the voxel responses. Notice that when we go from one plot to the next, two curves (eg the two middle curves in the bottom left

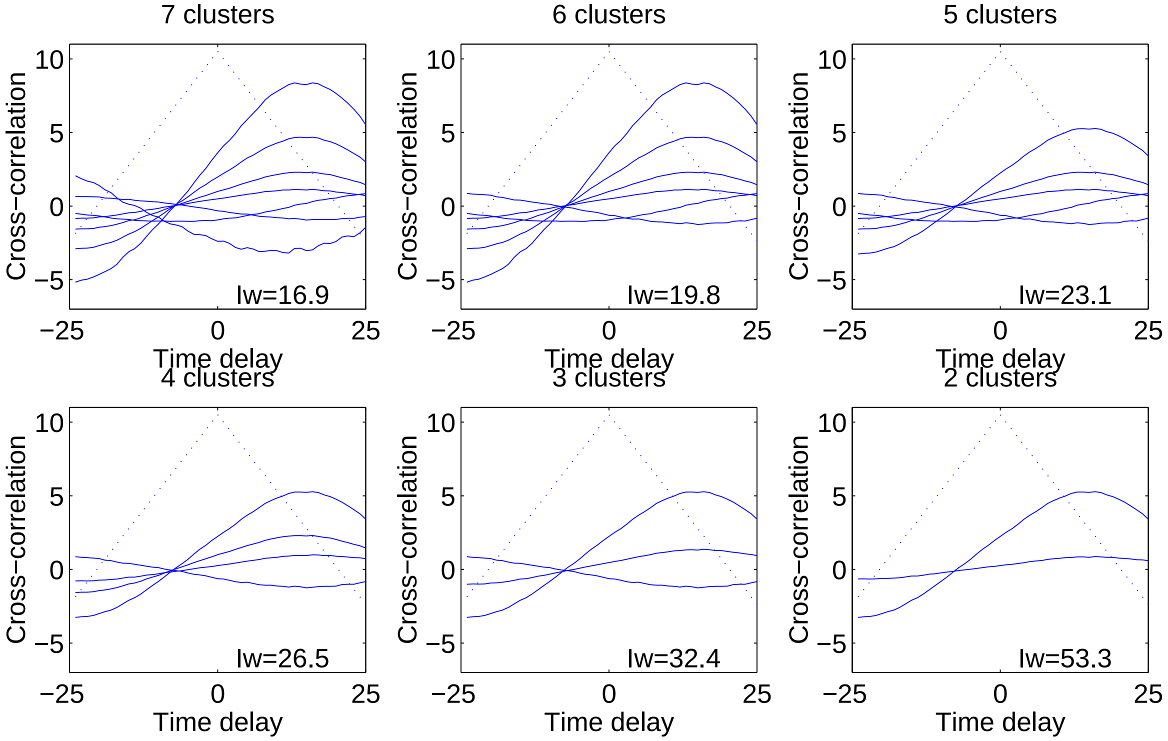

Figure 5: The cluster centres selected by hierarchical clustering, from 7 to 2 clusters. In each plot, the solid lines are the cross-correlation functions of the cluster centre, which are also the average of the cross-correlation functions of the cluster members. The dotted line is the auto-correlation function of the paradigm, suitably scaled, which allows to assess the delay for each cluster. In most cases the cross-correlation functions show that the associated voxels display one of three different effects: no activation, positive activation and negative activation. The within-class inertia is indicated in the lower right corner of each plot.

plot) are replaced by one (the middle curve in the bottom centre plot). This reflects the fact that each step of the algorithm joins two previously obtained clusters (represented by two cluster centres, ie two curves, on figure 5) into a new cluster, while the rest of the groups remain unchanged.

As expected, the within-class inertia increases as the number of classes decreases. Note that for 7 clusters, it is close to the average value of the K-means results (estimated at 16.45 from our 100 random initialisations). On the other hand, the 3 cluster partition is sizeably worse than any equivalent partition obtained with K-means. This is due to the increasing constraints on the partition introduced by the algorithm. While K-means gathers points in clusters with virtually no constraints, the hierarchical method is forced to join clusters that were obtained in the previous steps of the algorithm. Figure 6 plots the within-class inertia calculated for 1 to 20 clusters. The curvature, ie second derivative

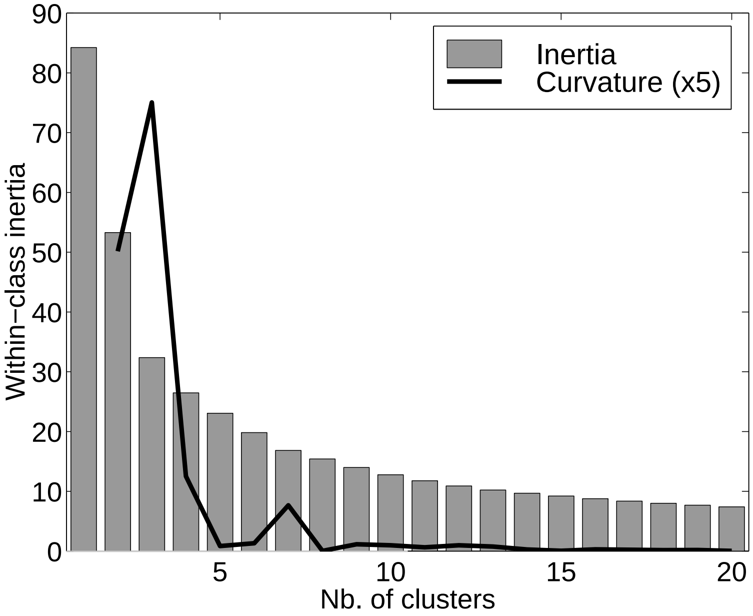

Figure 6: Within-class inertia (or variance) for the partitions generated by the hierarchical clustering algorithm, from 1 to 20 clusters (bars). The estimated curvature is displayed as a solid line, and shows two clear peaks for $K = 3$ and $K = 7$ clusters.

of the curve represents the way the increase in inertia evolves. High curvatures mean that joining two clusters at the corresponding level provoked a sharp change in inertia, or that the homogeneity of the associated clusters have changed drastically. The curvature is estimated using the central difference approximation, and plotted together with the inertia in figure 6. Two peaks appear clearly for 3 and 7 clusters. This indicates that the 6 (resp. 2) clusters configuration is much less homogeneous than the 7 (resp. 3) clusters partition. Accordingly, we will analyse the resulting groups for $K = 3$ and $K = 7$. Note that while the choice of clusters in the previous section was motivated by an arbitrary, *a priori* choice, inspection of the inertia gives us a convenient heuristic to estimate which cluster numbers we should concentrate on.

The binary tree or dendrogram generated by the hierarchical clustering algorithm can be cut at a level corresponding to $\mathcal{I}_W = 32.4$ in order to produce 3 clusters. Figure 7 displays 2 of these, which roughly correspond to positive and negative correlations with the paradigm. Comparison with figure 3 shows that the groups formed by both clustering methods are highly consistent. Note that the positively correlated voxels (left) are located in the visual cortex as before. Compared to K-means, the hierarchical algorithm seems to have gathered less voxels in both presented cluster, at the cost of a small increase (6%) in

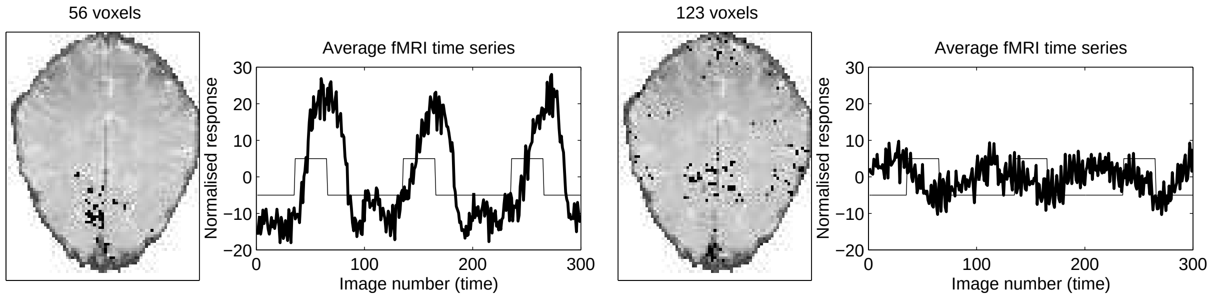

Figure 7: Two of the three clusters obtained by hierarchical clustering at level $\mathcal{I}_W = 32.4$. Brain maps: cluster members indicated in black on top of the anatomical reference. fMRI plots: average fMRI time series in the corresponding voxels in thick black line, paradigm (stimulus) plotted as a reference in thin black.

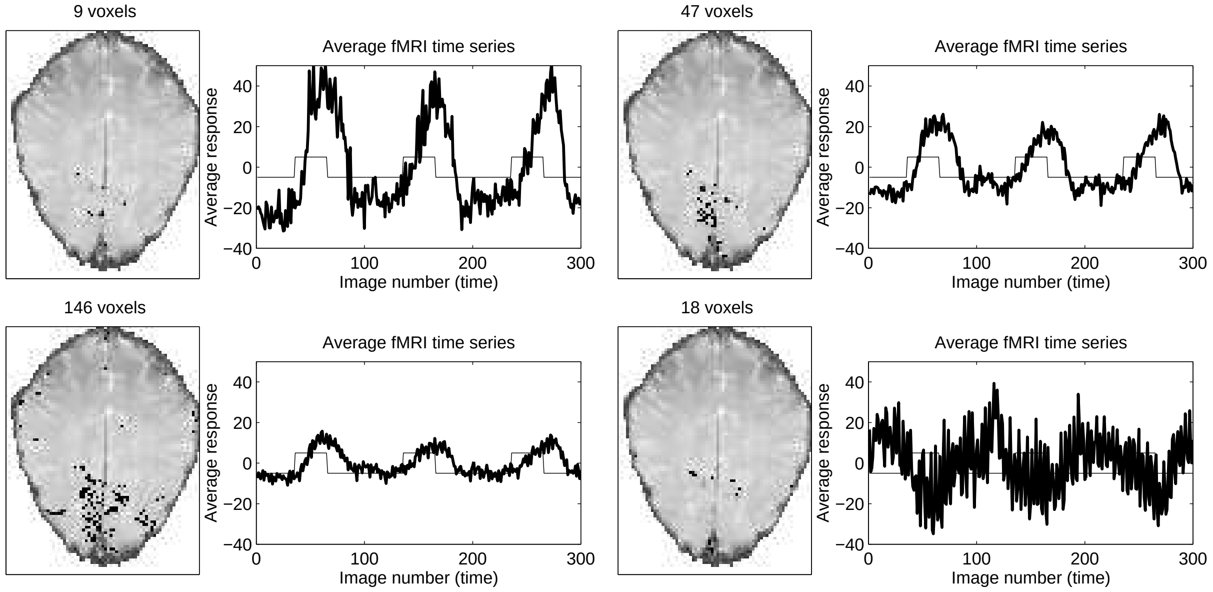

Figure 8: Four of the seven clusters obtained by hierarchical clustering at level $\mathcal{I}_W = 16.9$. Brain maps: cluster members are indicated in black on top of the anatomical reference. fMRI plots: average fMRI time series in the corresponding voxels in thick black line, paradigm (stimulus) plotted as a reference in thin black.

within-class variance.

At a level corresponding to $\mathcal{I}_W = 16.9$, the hierarchical clustering algorithm yields 7 clusters. Four of these are displayed on figure 8, where we have kept the same indicative ordering as for the corresponding K-means results (figure 4). Note that the first two clusters in figure 8 (9 and 47 voxels, top row) overlap exactly with the first cluster in figure 7 (56 voxels, left). As noted above, this is a consequence of the hierarchical nature of the method. The two most activated out of 7 clusters have been joined into one at level $\mathcal{I}_W = 23.1$, as shown in figure 5. Similarly, the cluster showing a large negative correlation

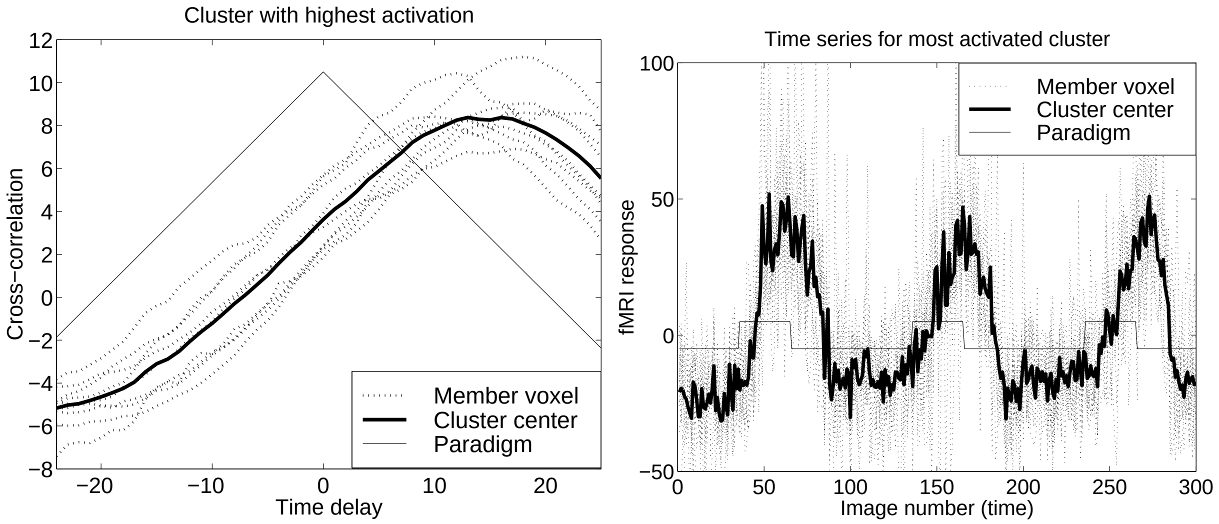

Figure 9: Members of the most activated cluster from figure 8. Left: cross-correlation functions of the cluster members (dotted) and cluster centre (broad solid) compared to the paradigm (thin solid). Right: corresponding raw time series.

with the stimulus in figure 8 (18 voxels, bottom right) was merged with another group to form a larger cluster displayed in figure 7 (right). The composition of the 4 clusters displayed in figures 4 and 8 differ little. The most activated voxels are located in the visual cortex, while area V5 seems to be present again in the moderately activated cluster (bottom left). The negatively activated clusters in figures 4 and 8 (bottom right) correspond for all but one voxel, and cover in particular the *sinus sagittalis.* Though the general similarity is quite good, the 7-cluster partition obtained by hierarchical clustering results in a 7% higher within-class variance.

Let us finally investigate the composition of the cluster with the largest positive activation. Figure 9 displays the cross-correlation function of the 9 voxels in the top left cluster from figure 8, together with the corresponding fMRI time series. The cross-correlation functions of the cluster members appear quite homogeneous, with a strong positive activation, and delays between 10 and 18 images (ie 3.5 to 6 seconds). Taken individually, the raw time series display a clear activation pattern. However, the rather high noise level makes the time series difficult to compare. On the other hand, the cluster centre benefits from the averaging and shows a large activation, with a rather low noise level.

DISCUSSION

Neuroscientific aspects

The analyses presented above are mainly meant as an illustration of the proposed statistical methods. However, they lead to a number of neuroscientific comments and perspectives which we will now address. Figures 3 to 8 show that some clusters seem to differ only in the activation strength of their members. This suggests that there is no clear separation of activated and non-activated voxels, but rather a continuum of activations. Some areas might of course be activated with different strength, eg the primary visual cortex and area V5 in our experiments. However, a more likely cause of the graded response that we observe would be the partial voluming in the visual cortex. The composition of the voxels, in particular with regard to vascular components, potentially influences the local concentration in deoxyhaemoglobin, and therefore the intensity of the signal. Another possibility would be the presence of capillaries instead of veins in the voxel of interest, modifying the blood-oxygenation level-dependent (BOLD) signal.

The resulting cluster partitions in our experiments display an interesting feature: the presence of negative correlation with the stimulus. Let us first emphasise that this negative correlation can not be due to a positive effect with a long delay. Indeed, the dataset was formed from several distinct runs, such that there is no causality between a response and the previous stimulus. As noted above, the voxels showing negative activation contain a high-frequency component representing cardiac rhythm and some of them cover the *sinus sagittalis*. This suggests a link between negative correlation and presence of blood vessels. However, though other voxels might also contain such vessels, there is no direct explanation for the negative correlation, as a specific deactivation would correspond to an increase in *de*oxyhaemoglobin. On the other hand, the hypothesis of a movement-related artifact seems unlikely. A specific movement could indeed locally modify the proportion of capillaries and veins, leading to an increased signal in one voxel, and a corresponding decrease in the other. However, in order to be detected as a positive or negative correlation, this movement would have to be somehow related to the stimulus, and such a stimulus-related movement has not been isolated. Finally, let us mention the possibility of the presence of an inverse BOLD signal in response to an activation. To our knowledge, this effect has so far only been observed in infants (Born et al., 1996).

Statistical aspects

Let us first insist again on the fact that these experiments are of an exploratory, rather than inferential, nature. We have given guidelines as to how significance levels can be estimated for each cross-correlation function, but we consider that the main objective of this work is to explore the data in order to identify interesting differences in activation. A challenging application of the clustering results is the formulation of hypotheses that are more interesting than the standard null hypothesis. Indeed, a limitation of the standard use of statistical testing is that it estimates the probability of lack of activation rather than the extent and probability of actual activation.

We have insisted on the crucial choice of the number of clusters. Choosing the optimal number is a typical capacity control problem, and few principled approaches have been proposed (Hansen and Larsen, 1996). Some alternatives address this problem, eg the classical Isodata algorithm (Tou and Gonzalez, 1974, p. 97) is a popular method relying on K-means and a set of clever heuristics. Unfortunately, many methods proposed in the literature tend to be unreliable (as shown by Moore (1989) for *Adaptive Resonance Theory*). In that respect, the two methods presented here offer a complementary behaviour. Though fast and powerful, K-means requires the number of clusters to be set *a priori*. On the other hand, the hierarchical clustering algorithm makes it possible to choose this number according to the evolution of the within-class variance. It automatically provides partitions for each number of clusters, but these are not as homogeneous as K-means'. This number of cluster/homogeneity dilemma suggests the combination of both methods to exploit their attractive features.

Many other clustering algorithms exist. The most popular in the neuroimaging community is probably the fuzzy K-means method (Baumgartner et al., 1998, and references therein). However, note that Davé and Krishnapuram (1997) have shown that a large number of fuzzy clustering methods are essentially equivalent to traditional techniques in robust statistics. Other robust methods can be derived directly from the K-means algorithm by simply using a different way of updating the cluster centres. Let us just mention the K-medians or K-medoids described eg by Ripley (1996).

Finally, we have noticed above that the resulting clusters seem to spread over a continuum of activations. This raises the question of whether the obtained groups are providing any useful information, or merely partitioning a continuous distribution of activations. The sharp changes in within-cluster inertia show that there is indeed some homogeneity in the clusters we have presented. Furthermore, note that not only the activation, but also the delay vary across clusters, as shown by Goutte et al. (1998b).

CONCLUSION

This contribution addresses the problem of clustering fMRI time series in groups of voxels with similar activations. We present and analyse two clustering algorithms and demonstrate their use on fMRI data acquired during a visual experiment. The main contributions of this work are: the use of the cross-correlation function as a feature space, rather than the raw fMRI time series; the introduction of a flexible metric definition linking both spaces and allowing different preprocessing strategies (filtering, PCA, etc.); the use of the hierarchical clustering algorithm in conjunction with the classical K-means method. We present results underlining the complementarity between both techniques, and showing that clustering can effectively identify regions of *similar* activations.

REFERENCES

Baker, J., Weisskoff, R., Stem, C., Kennedy, D., Jiang, A., Kwong, K., Kolodny, L., Davis, T., Boxerman, J., Buchbinder, B., Wedeen, V., Belliveau, J., and Rosen, B. (1994). Statistical assessment of functional MRI signal change. In *Proceedings of the 2nd Annual Meeting of the Society of Magnetic Resonance*, page 626.

Bandettini, P. A., Jesmanowicz, A., Wong, E. C., and Hyde, J. S. (1993). Processing strategies for time-course data sets in functional MRI of the human brain. *Magnetic Resonance in Medicine*, 30(2):161–173.

Baumgartner, R., Scarth, G., Teichtmeister, C., Somorjai, R., and Moser, E. (1997). Fuzzy clustering of gradient-echo functional MRI in the human visual cortex. Part I: reproducibility. *Journal of Magnetic Resonance Imaging*, 7(6):1094–101. see also Moser et al. (1997).

Baumgartner, R., Windischberger, C., and Moser, E. (1998). Quantification in functional magnetic resonance imaging: fuzzy clustering vs. correlation analysis. *Magnetic Resonance Imaging*, 16(2):115–25.

Belliveau, J., Fox, P., Kennedy, D., Rosen, B., and Ungerleider, L., editors (1996). *Second International Conference on Functional Mapping of the Human Brain*, *Neuroimage* 3(3), part 2. Academic Press.

Born, P., Rostrup, E., Leth, H., Peitersen, B., and Lou, H. C. (1996). Changes of visually induced cortical activation patterns during development. *The Lancet*, 347(9000):543.

Bottou, L. and Bengio, Y. (1995). Convergence properties of the K-means algorithm. In *Advances in Neural Information Processing Systems*, volume 7. MIT Press.

Davé, R. N. and Krishnapuram, R. (1997). Robust clustering methods: A unified view. *IEEE Transactions on Fuzzy Systems*, 5(2):270–293.

Friberg, L., Gjedde, A., Holm, S., Lassen, N. A., and Nowak, M., editors (1997). *Third International Conference on Functional Mapping of the Human Brain*, *Neuroimage* 3(3), part 2. Academic Press.

Golay, X., Kollias, S., Meier, D., Valavanis, A., and Boesiger, P. (1997). Fuzzy membership vs. probability in cross correlation based fuzzy clustering of fMRI data. In Friberg et al. (1997), page S481.

Goutte, C., Hansen, L. K., and Larsen, J. (1998a). Monte-carlo assessement of cross-correlation significance. unpublished.

Goutte, C., Nielsen, F. Å., Svarer, C., Rostrup, E., and Hansen, L. K. (1998b). Space-time analysis of fMRI by feature space clustering. In Paus, T., Gjedde, A., and Evans, A., editors, *Fourth International Conference on Functional Mapping of the Human Brain, Neuroimage* 3(3), part 2. Academic Press. `http://eivind.imm.dtu.dk`.

Hansen, L. K. and Larsen, J. (1996). Unsupervised learning and generalization. In *Proceedings of the IEEE International Conference on Neural Networks.* `http://eivind.imm.dtu.dk`.

Hartigan, J. A. and Wong, M. A. (1979). Algorithm AS136. A K-means algorithm. *Applied Statistics*, 28:100–108.

Holmes, A. P. and Friston, K. J. (1997). *Statistical Models and Experimental Design.* SPM course notes, chapter 3.

Lange, N. and Zeger, S. L. (1997). Non-linear Fourier time series analysis for human brain mapping by functional magnetic resonance imaging. *Journal of the Royal Statistical Society, Series C, Applied Statistics*, 46(1).

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In Le Cam, L. M. and Neyman, J., editors, *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 281–297, Berkeley, CA. University of California Press.

Mahalanobis, P. C. (1936). On generalized distance in statistics. *Proceedings of the National Inst. Sci. (India)*, 12:49–55.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models.* Number 37 in Monographs on statistics and applied probability. Chapman & Hall, London, 2nd edition.

McIntyre, M., Wennerberg, A., Somorjai, R., and Scarth, G. (1996). Activation and deactivation in functional brain images. In Belliveau et al. (1996), page S82.

Moore, B. (1989). ART 1 and pattern clustering. In Touretzky, D., Hinto, G., and Sejnowski, T., editors, *Proceedings of the 1988 Connectionist Models Summer School*, pages 174–185, San Mateo, CA. Morgan Kaufmann.

Moser, E., Diemling, M., and Baumgartner, R. (1997). Fuzzy clustering of gradient-echo functional MRI in the human visual cortex. Part II: quantification. *Journal of Magnetic Resonance Imaging*, 7(6):1102–8. see also Baumgartner et al. (1997).

Nielsen, F. Å., Hansen, L. K., Toft, P., Goutte, C., Mørch, N., Svarer, C., Savoy, R., Rosen, B., Rostrup, E., and Born, P. (1997). Comparison of two convolution models for fMRI time series. In Friberg et al. (1997), page S473. `http://eivind.imm.dtu.dk`.

Ripley, B. D. (1996). *Pattern Recognition and Neural Networks*. Cambridge University Press.

Scarth, G., Wennerberg, A., Somorjai, R., Hindmarsh, T., and McIntyre, M. (1996). The utility of fuzzy clustering in identifying diverse activations in fMRI. In Belliveau et al. (1996), page S89.

Toft, P., Hansen, L. K., Nielsen, F. Å., Goutte, C., Strother, S., Lange, N., Mørch, N., Svarer, C., Paulson, O. B., Savoy, R., Rosen, B., Rostrup, E., and Born, P. (1997). On clustering of fMRI time series. In Friberg et al. (1997), page S456. `http://eivind.imm.dtu.dk`.

Tou, J. T. and Gonzalez, R. C. (1974). *Pattern Recognition Principles*. Number 7 in Applied Mathematics and Computation. Addison-Wesley, Reading, Massachusetts.

Ward, J. H. J. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58:236–244.

Worsley, K. and Friston, K. (1995). Analysis of fMRI time-series revisited — again. *Neuroimage*, 2:173–181.

Xiong, J., Gao, J.-H., Lancaster, J. L., and Fox, P. T. (1996). Assessment and optimization of functional MRI analyses. *Human Brain Mapping*, 4(3):153–167.

ACKNOWLEDGEMENTS

This project was partly funded by The Danish Research Councils, the Human Brain Project P20 MH57180 “Spatial and Temporal Patterns in Functional Neuroimaging”, and the EU BIOMED II grant BMH4-CT97-2775 “Visual object recognition”. The authors wish to thank the anonymous reviewers for constructive comments and criticism.