Feature space clustering for fMRI meta-analysis

Cyril Goutte, Lars Kai Hansen, Matthew G. Liptrot, Egill Rostrup[1]

Department of Mathematical Modelling, Technical University of Denmark, Lyngby, Denmark, and [1]Danish Research Centre for Magnetic Resonance, Hvidovre, Denmark

September 22, 1999

Draft version 1.0

Running title: Feature-space clustering

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

Clustering fMRI time series has emerged in recent years as a possible alternative to parametric modelling approaches. Most of the work has been so far concerned with clustering raw time series. In this contribution we investigate the applicability of a clustering method applied to *features* extracted from the data. This approach is extremely versatile and encompasses previously published results (Goutte et al., 1999) as special cases. A typical application is in data reduction: as the increase in temporal resolution of fMRI experiments routinely yields fMRI sequences containing several hundreds of images, it is sometimes necessary to invoke feature extraction to reduce the dimensionality of the data space. A second interesting application is in the meta-analysis of fMRI experiment, where features are obtained from a possibly large number of single-voxel analyses. This allows in particular to check the differences and agreements between different methods of analysis. Both approaches are illustrated on a fMRI dataset involving visual stimulation, and we show that the feature space clustering approach yields non-trivial results, and in particular shows interesting differences between individual voxel analysis performed with traditional methods.

FEATURE-SPACE CLUSTERING

The high temporal resolution of fMRI has inspired a host of single-voxel analysis methods. They typically compute some summary statistics characterising the temporal response in each voxel and represent them in the spatial domain by means of a brain map for visual inspection or additional inference (see eg Bandettini et al., 1993; Baker et al., 1994; Worsley and Friston, 1995; Xiong et al., 1996; Lange and Zeger, 1997, for a by no means exhaustive sample). These traditional approaches usually rely on some model and/or assumption about the fMRI acquisition, eg concerning the stimulus (binary baseline-activation conditions), the haemodynamic response (modelled by a filter taken from a parametric family), among others. These techniques can be loosely qualified as inferential analyses.

A second, more recent approach to neuroimaging using fMRI relies on exploratory data analysis (EDA). Apart from principal component analysis (PCA) (Sychra et al., 1994) and recent developments in independent component analysis (ICA) (Mckeown et al., 1998), EDA in neuroimaging has been mostly carried out using clustering methods. The most popular method is certainly fuzzy c-means (Baumgartner et al., 1997; Moser et al., 1997; Baumgartner et al., 1998; Golay et al., 1998; Moser et al., 1999; Somorjai et al., 1999). The standard K-means algorithm has also been used with some success (Ding et al., 1994; Toft et al., 1997; Goutte et al., 1999), as well as hierarchical approaches to clustering (Goutte et al., 1999; Filzmoser et al., 1999). Also using hierarchical arguments together with a non-parametric approach, innovative methods are starting to appear (eg Wismüller et al., 1998; Domany, 1999; Goutte, 1999).

Clustering techniques aim at identifying regions with similar patterns of activation. Clustering is often performed on the raw fMRI time series, with the added benefit that no additional information is necessary regarding eg the stimulus or the underlying haemodynamic response. Clustering on the raw time series is potentially able to isolate cognitive or haemodynamic effects without precisely modelling these beforehand. However, the low signal-to-noise ratio and the increasing size of the acquired fMRI data, hence that of the space in which clustering is performed, lead to practical difficulties, notably with the estimation of pairwise distances. We have previously shown (Toft et al., 1997; Goutte et al., 1999) that clustering on the cross-correlation function instead of the raw time series provides increased robustness.

In this contribution, we explore the possibility of clustering features extracted from the fMRI time series. This generalises the approach of clustering on the cross-correlation function proposed by Goutte et al. (1999) to a more flexible set of features. In particular, it allows the analyst to concentrate on what is felt constitutes the features of interest in a given fMRI experiment. As an example, we focus simply on delay and strength measured on a voxel-by-voxel basis, and show that we can identify regions with significantly different delays and activations.

Another potential of the feature-based approach is the use of clustering as a meta-analysis tool. In that context, we try to cluster regions that lead to similar results using (possibly many) single voxel analyses. This has been to some extent exemplified on finite impulse response (FIR) filters by Purushotham et al. (1999). In this contribution, we use a number of standard analyses performed within the publicly available “Lyngby” modelling toolbox (Hansen et al., 1999). Our analysis, performed on data from a visual stimulation experiment, isolates small homogeneous areas that display striking differences in activations, leading to different, and incompatible, results on a voxel-by-voxel basis.

In the following sections, we first present the dataset, and the methods that will be used in this article. We then present the results of two different feature space clustering analyses. Finally, the discussion section discusses some of the implications of the proposed method, and some of the neuroscientific results obtained.

DATA

The dataset was acquired at Hvidovre Hospital 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 and 50 degrees RF flip angle. The images were acquired with a matrix of 128x128 pixels, with FOV of 230 mm, and 10 mm slice thickness, in a para-axial orientation parallel to the calcarine sulcus. The region of interest (ROI) will be limited to a 68x82 two-dimensional voxel map. The voxel dimension is 1.8x1.8x10 millimetre.

The visual paradigm consists 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 with 20 seconds of rest (darkness). In total, 150 images were acquired in 50 seconds, corresponding to a period of approximately 330 ms per image. The experiment was repeated in 10 separate runs containing 150 images each. In order to reduce saturation effects, the first 29 images were discarded, leaving 121 images for each run.

The dataset studied in this article was built by combining these into a single sequence of 1210 images. However, as the runs were acquired separately, it should be noted that there can not be any causality between the activation in one run and the signal measured in the next. Note also that due to the haemodynamic delay, the signal measured in activated voxels will be roughly centred within the remaining 40 seconds of each run.

METHODS

Motivation

In the following, the measurement in voxel (x, y, z) at time t will be denoted $f(x, y, z, t)$. This signal is the measured response to a pattern of activation represented by the paradigm $p(t)$, usually a square wave or box-car signal for simple on/off excitation. The complete time series in (x, y, z) is a vector $\boldsymbol{f}(x, y, z) \in \mathbb{R}^T$ where T is the number of images, $1 \leq t \leq T$. While for PET experiments the number of images is usually limited, T can easily reach several hundred for fMRI data.

A number of attempts have been made at clustering PET (Ashburner et al., 1996) and especially fMRI data (Baumgartner et al., 1997; Moser et al., 1997; Baumgartner et al., 1998; Golay et al., 1998; Goutte et al., 1999) in the past few years. Most of these use the raw time series as input and all of them cluster based on a metric, ie a notion of distance, in the input (ie time series) space. However, it is well known that when the dimensionality of the space increases, the notion of distance becomes counter-intuitive. This is easily seen by considering what happens when points are randomly distributed uniformly in a multi-dimensional unit hypercube. As the dimension grows, the average absolute distance between two random points grows linearly, while the average distance of a point to the hypercube's boundaries decreases. This means that points tend to lie far away from each other in a shell at the boundary of the space (see also Bishop, 1995, exercises 1.2 to 1.4).

For experiments involving large numbers of images, clustering in input space using a standard Euclidean distance is therefore expected to be difficult. The approach adopted by Goutte et al. (1999) relies on the cross-correlation between the fMRI time series and the excitation paradigm. This focuses on a more relevant feature space, namely the pattern of activation, and reduces the input space significantly. However, the dimension of the input space will still be relatively high ($T = 50$ in (Goutte et al., 1999)), and the inputs are correlated. This high dimensionality contrasts with the fact that we will typically be interested in a small number of intrinsic features such as the strength of activation in each voxel, the probability of activation or the response delay.

A further reduction in dimensionality can be obtained by extracting, either from the signal or from the cross-correlation function, some relevant features describing the quantities of interest. In the following sections, we will consider two such examples. One extracts simple features, the activation strength and the response delay, from the cross-correlation function. The second considers the output of standard analysis tools, such as statistical tests or linear filtering, as features and then implements a "meta-clustering" analysis. In all cases, the features are clustered using a K-means algorithm where the number of clusters is set using some information criterion.

Preprocessing

We pre-process the data using a run-based detrending method. For each run of 121 images, we use the first 31 images and the last 31 images to estimate the linear trend in the data, and subtract this trend from the raw time series. This is justified by the fact that including activated images in the trend estimation potentially yields incorrect estimates. This is especially true when the activation is not centred in the acquisition interval. In particular, delayed activations will lead to an upward bias in the estimate of the trend's slope. In our experiment, taking into account the haemodynamic delay (Bandettini et al., 1993), we would expect the increased blood oxygenated level dependent (BOLD) signal between (approximately) images 50 and 80 of each run. This is around 6 images (or 2 seconds) off-centre.

The rationale for the choice of the images we use to estimate the trend is that the first 31 images are

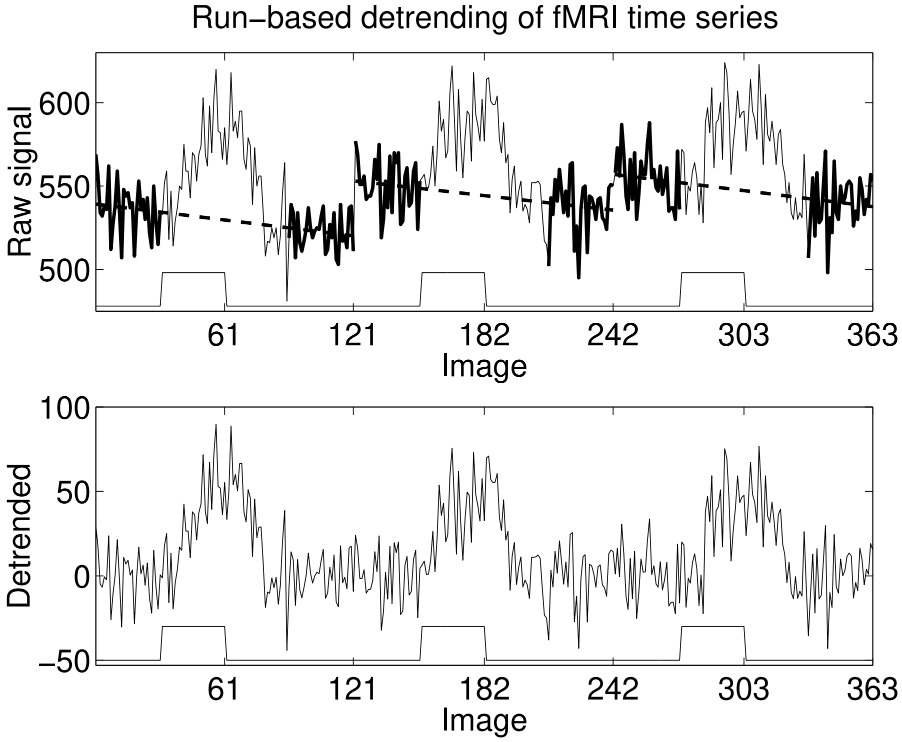

FIG. 1. Run-based detrending of the fMRI time series. Top: three runs of the raw data (solid) and the trend (dashed) estimated for each run using the data indicated by the bold solid line. Bottom: the corresponding detrended data. Note that the non-activated measurements have roughly zero mean, but this is not true of the full time series. The square wave at the bottom of each plot indicates the paradigm.

recorded before activation occurs and therefore cannot be activated. The last 31 images are recorded at least 10 seconds after stimulus cessation, where most of the activation is expected to have been reduced (see eg Buxton et al., 1998). This choice gives us 62 data points for each run from which to estimate the slope and the offset of the trend.

Figure 1 illustrates the detrending process on three runs (runs 5 to 7) recorded in a particular activated voxel (coordinates 64, 32). The measurements that are used to estimate each trend are indicated in bold. Notice that, by construction, the detrended time series do not have 0 mean. While the mean of the non-activated measurements is roughly zero, the presence of activation in a voxel will usually result in a non-zero mean value (figure 2). This contrasts with the usual detrending using a global trend, ie including the activated images, where the mean is by construction zero (figure 2, right plot).

There is a possibility of an artefactual bias in the detrended time series, due to the presence of a long post-activation undershoot. If some of the last 31 images used for estimating the trend contain this undershoot, this will lead to a negative bias in the estimate of the slope. Hence the detrended data could display an artefactual positive trend. Note that this effect is minor compared to the amplitude of the correction brought by the run-based detrending pre-processing (cf. figure 1). Furthermore, it should be noted that the goal of clustering is to identify similar patterns of activations. As similar patterns of activation will have similar (if any) biases, this artefact will not interfere with the analysis results.

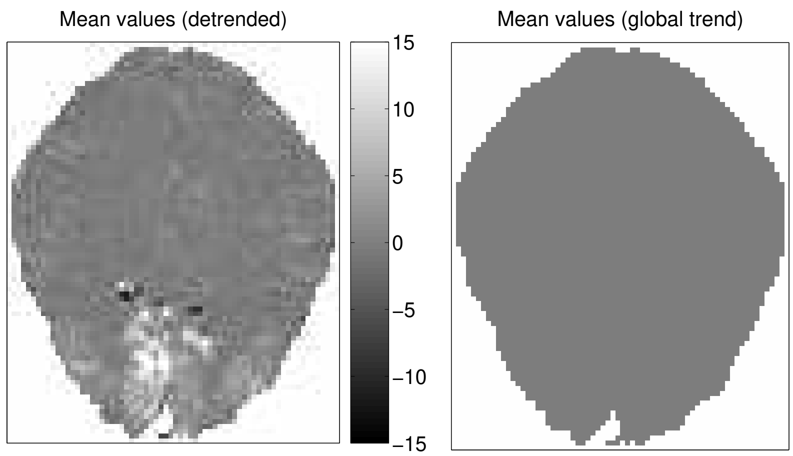

FIG. 2. Map of the mean value of the detrended time series, using the run-based detrending (left) and using a global trend (right). Contrary to the usual practice of centring the time series, which is inherent in the use of a global trend, the run-based detrending preserves some of the information in the mean time series. The colorbar applies to both plots.

Data reduction

In order to discard those voxels that clearly have no overall effect, we perform an omnibus F-test (Holmes and Friston, 1997, sec. 6.4), and threshold the data at $p = 0.01$. In order to take into account the high temporal resolution of our data and get a meaningful F-statistic, the paradigm is first shifted 7 seconds.

The goal of this *sieving* is to concentrate on the possibly activated voxels. As noted by Holmes and Friston (1997), it is extremely unlikely that we can show any specific effect in voxels where there is no overall effect. All the analyses that we consider below, and in particular the clustering, will be performed on the voxels exceeding the $F_{.99}$ threshold. An additional advantage of this sieving is that it will avoid wasting a possibly large number of clusters on uninteresting, noisy data, therefore focusing on the relevant data, ie where there is a possible activation. A similar data-reduction device was used by Goutte et al. (1999), based on the value of the maximum cross-correlation.

Finally, let us note that it is always possible to salvage *a posteriori* discarded voxels, by assigning them to the clusters resulting from the analysis (Somorjai, 1999). This could in principle "recycle" voxels. However, in the analyses presented here we found no use for this particular post-processing.

Feature extraction from cross-correlation

The first feature extraction we will consider uses the cross-correlation function (Toft et al., 1997; Goutte et al., 1998, 1999). The cross-correlation function between the signal $f(x, y, z, t)$ and the paradigm $p(t)$ is defined as:

$$xc(x, y, z, \delta) = \frac{1}{T} \sum_{t=1}^{T} f(x, y, z, t) p(t - \delta) \tag{1}$$

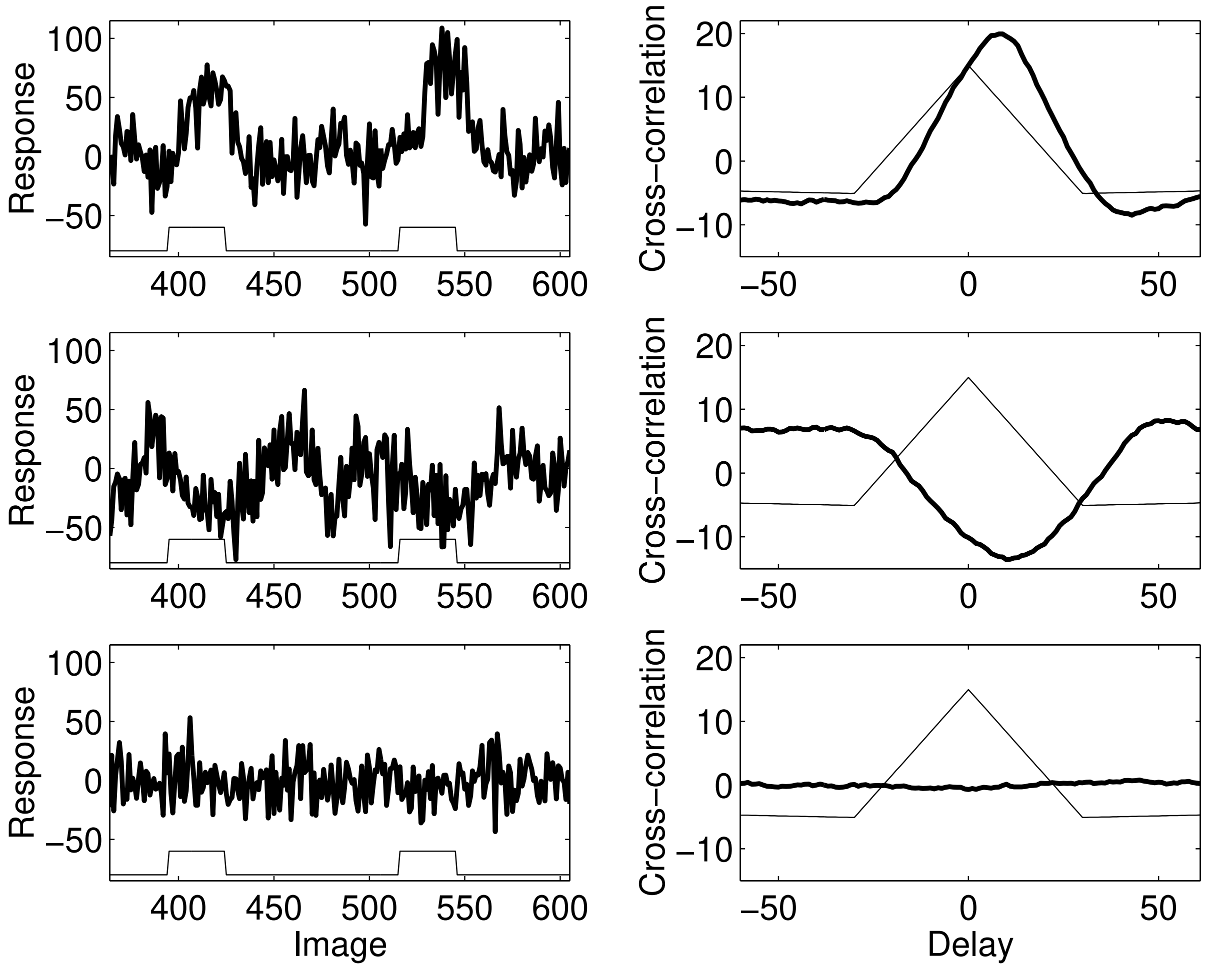

FIG. 3. Functional MRI signal (left) and cross-correlation with the paradigm (right) for three voxels (in bold). Top: positive correlation with the paradigm; middle: negative correlation; bottom: no correlation. The paradigm and its auto-correlation are indicated with a thin solid line on both plots. Notice that the maximum amplitude of the cross-correlation is obtained with a delay of around 15 images.

where for notational convenience we define $p(t) = 0$ for $t \leq 0$ or $t > T$. Note that the cross-correlation function should be distinguished from the correlation used eg by Bandettini et al. (1993). In particular it is neither centred, nor normalised, though in practice we usually normalise the paradigm $p(t)$ (ie zero mean, unit variance). Figure 3 displays the raw signals and the cross-correlation function for three voxels. The first two display strong positive and negative activations (respectively), while the third one is essentially uncorrelated with the stimulus.

From the cross-correlation function, we extract 2 quantities of interest:

1. The *strength* of activation is the (signed) extremum of the cross-correlation function. For the voxels in figure 3, the strengths are 19.58, -13.29 and 0.68 respectively.

2. The *delay* of activation is the delay δ for which the cross-correlation function reaches its extremum.

Due to the periodicity in the data, we will only look for the maximum between delays -60 and 60. For noisy and discrete data, the cross-correlation signal is itself noisy and discrete. As a consequence, the

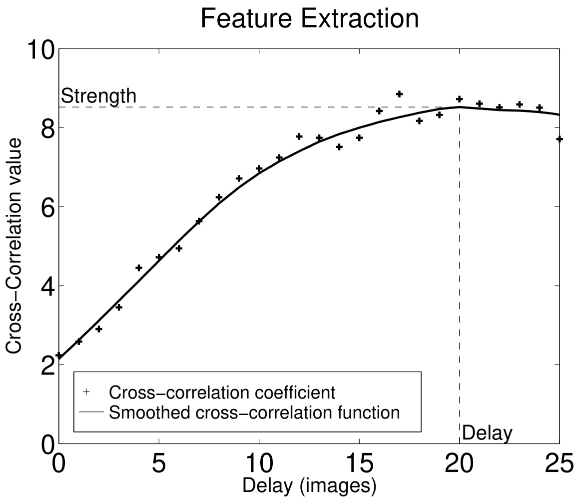

FIG. 4. Smoothing of the cross-correlation values (+) using a kernel smoother and extraction of the features (dashed) from the resulting smoothed curve (solid). Note that the smoothed maximum (at 20 images delay) is different from the maximum of the cross-correlation coefficients (at 17 images delay).

maximum cross-correlation coefficient $xc(x, y, z, \delta_{max})$ could lie away from the maximum suggested by the overall shape of the cross-correlation curve (figure 4). In order to improve the feature extraction, we smooth the cross-correlation function using a standard kernel smoother using an Epanechnikov kernel with a fixed bandwidth (Wand and Jones, 1995, sec. 5.8). The process is illustrated in figure 4. The resulting two features are gathered in a two dimensional data vector $\boldsymbol{u}(x, y, z)$.

Feature extraction for meta-clustering

The idea of feature-space clustering can be applied in a straightforward manner as a meta-analysis tool. In particular, we will be interested in evaluating similarities and differences in the results provided by several kinds of standard single-voxel analyses. In the following experiments, we focus on the following features:

- t-statistic from a standard Student t-test,
- KS statistic from a standard Kolmogorov-Smirnov test (Baker et al., 1994),
- correlation with the paradigm, delayed by 7 seconds (Bandettini et al., 1993),
- standard deviation of the fitted signal modelled with a finite impulse response (FIR) filter (Nielsen et al., 1997),
- delay estimated from the fitted FIR filter (Nielsen et al., 1997),
- strength parameter from a Gamma-filter fitted on the signal (Lange and Zeger, 1997),
- delay estimated from the fitted Gamma-filter (Lange and Zeger, 1997).

All features come from different voxel-based analyses (Hansen et al., 1999) and are likely to yield individually different results (Lange et al., 1999). We expect however some strong correlations, as most of the areas that come up as activated using one type of analysis will most likely be also identified as activated using a second method. The motivation for meta-clustering can be thought of as a natural way of performing meta-analysis. An investigator interested in comparing methods would typically produce activation maps for the different analyses, and find some kind of consensus by identifying regions that "activate in the same way" on the different maps. The proposed meta-clustering method is a way of automating this process. Furthermore, the different measures of activation strength usually presented on activation maps can be combined with additional features that are harder to present visually, such as the delays, leading to a spatio-temporal analysis (Goutte et al., 1998).

Clustering

In this paper, we use the standard and simple K-means algorithm, as presented eg by Tou and Gonzalez (1974), Ripley (1996) or Goutte et al. (1999). For a given number K of clusters, the algorithm iteratively minimises the within-class variance by assigning data to the nearest centre and recalculating each centre as the average of its members. For simplicity, the Q-dimensional data vectors $\boldsymbol{u}_j$ will be indexed with $j \in \{1, \ldots N\}$, such that for all (x, y, z), there exist a unique j such that $\boldsymbol{u}(x, y, z) = \boldsymbol{u}_j$. The K-means algorithm iterates the following steps:

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

2. Assign each data vector $\boldsymbol{u}_j$ 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, $d\left(\boldsymbol{u}_j, \boldsymbol{c}_k{}^{(i)}\right)$.

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

until a stable partition is found (this is achieved in a finite number of iterations and usually quite fast, cf Bottou and Bengio (1995)). The role of the distance metric $d\left(\boldsymbol{u}_j, \boldsymbol{c}_k{}^{(i)}\right)$ is discussed in detail by Goutte et al. (1999).

Initialisation: There are usually several stable partitions, depending on the initial configuration. Among possible initialisation strategies, we simply use a random subset of the data as initial centres. In order to limit the effect of the stochastic initialisation, we will usually perform several clusterings using different random initialisations, and choose the final partition that has the lowest within-class variance (Goutte et al., 1999). This is usually no problem as the K-means algorithm itself is extremely fast. Note that this multiple initialisation procedure is in turn a stochastic process, as the minimum will depend on the particular set of initial centres chosen. This problem is addressed below by deriving empirical error-bars on the resulting configurations.

Number of clusters: One of the key problems in clustering is to decide on the number of clusters. Goutte et al. (1999) propose a heuristic argument based on the use of hierarchical clustering, while other contributions on clustering fMRI time series seem to mostly ignore this issue. In this contribution, we adopt a model-based approach and assume that the data comes from an underlying mixture of Gaussian distributions, with equal isotropic covariance matrices. In that context, we can invoke several information criteria. A more detailed description of the underlying model and information criteria is left to appendix A, where we give the expression of the Information Criterion proposed by Akaike (1974) (AIC), the Bayesian Information Criterion (BIC) proposed by Schwartz (1978) and the Integrated Classification Likelihood (ICL) of Biernacki et al. (1998). Using this framework, we will select the number of clusters that yield the highest value for the information criterion. The three criteria given in appendix A will usually lead to different clustering models. In particular, AIC is known to over-estimate the number of clusters. We will illustrate and discuss this below for our two clustering problems.

Normalisation In some cases, the features, ie the dimensions of the $\boldsymbol{u}_j$'s have very different ranges. In that case, it is advisable to normalise the data, in order to avoid for example that one feature dominates the clustering process. Indeed, if one feature accounts for most of the variance, the algorithm will tend to cluster this dimension and disregard the remaining features.

Normalisation typically involves centring by subtracting the mean and dividing by the standard deviation. Functional MRI data are characterised by the fact that there is usually a limited number of activated voxels and the activation statistics can be fairly high. In order to robustify the centring we will use the median, and we scale the centred data such that 99% of the sample is placed in the interval $[-2.57; 2.57]$. This takes into account the fact that typically less than 5% of the data will have values exceeding the activation threshold. As noted by Goutte et al. (1999), this is equivalent to using a scaling metric instead of the standard Euclidean distance.

Note that for easy reference, we will always display the results in the original (ie unnormalised) domain. This allows the reader to relate to the actual meaning of the features.

RESULTS

Preprocessing

The dataset acquired as a sequence of 128x128 voxels images was first processed to identify the location of the brain in the image, using an automated thresholding method in conjunction with standard filters. The masked brain contains 3891 voxels, and each time series has 1210 elements. The run-based detrending is performed according to the description given above.

For 1210 images, the value of the F-statistic corresponding to the 1% threshold is $F_{.99} = 6.67$. In

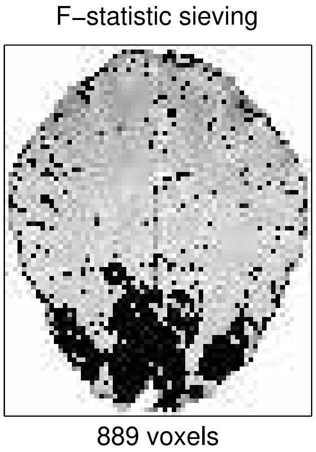

FIG. 5. Sieving non-activated voxels using an omnibus F-test. The 889 voxels that pass the test are indicated in black over the anatomic background. The analyses presented below will be limited to these voxels.

order to get a meaningful F-statistic, the paradigm is first shifted 7 seconds to take into account the delay of the haemodynamic response. In the current analysis, 889 voxels out of 3891 have a F-statistic greater than this threshold. They include the primary visual area, as well as lateral areas in the visual cortex, in particular the lateral occipital cortex (figure 5).

Clustering on delay and strength

The cross-correlation and feature extraction produce 3891 points in the two-dimensional delay-strength space. In addition to the sieving based on the F-statistic, we discard the points for which the delay estimated on the cross-correlation signal is either negative or larger than 10 seconds, as these delays would have no meaningful explanation. Negative delays, for example, can and do appear by chance for noisy voxels. Though the F-thresholding used as a data reduction device is supposed to discard noisy voxels, it is well known that due to the large number of the voxels, some will appear as false positive after our sieving. The selection of the voxels with correct delay retains 851 voxels, on which we apply clustering.

We apply K-means (without normalisation) for an increasing number of clusters and calculate the values of the 3 information criteria mentioned above. The model is considered optimal with respect to a given criterion when this criterion reaches its maximum. In order to estimate the variation due to the stochastic initialisation of the K-means algorithm (cf. methods section), we derive empirical error-bars on the criteria by replicating the multiple initialisation procedure. Figure 6 displays the AIC, BIC and ICL criteria and their error bars. Notice how the error bars grow for larger cluster numbers, as the number of possible final configurations grows. The optimal cluster number is selected by finding the maximum point, taking the error bars into account. In these conditions, we see in figure 6 that ICL selects 4 clusters, BIC is maximum for 8 clusters, and AIC chooses 18 clusters. This illustrates nicely a key difference between the

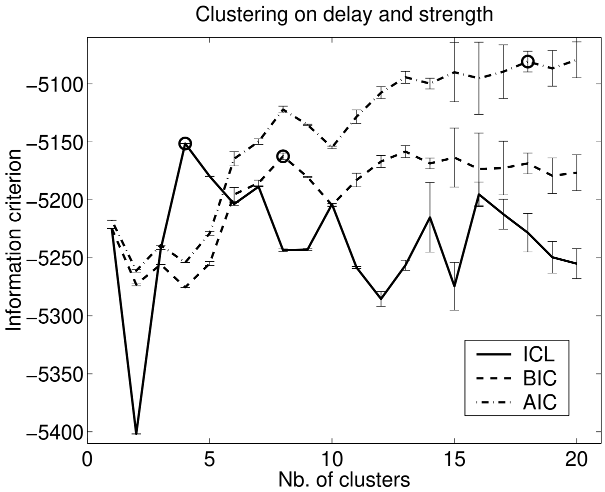

FIG. 6. The three information criteria calculated on the two-dimensional delay-strength data: AIC (dash-dotted), BIC (dashed) and ICL (solid). The error bars indicate 2 standard deviation. The maxima, taking the error bars into account are obtained for 18, 8 and 4 clusters respectively (indicated by circles). Notice how much more ICL penalises large numbers of clusters (cf. equation 8).

three criteria. The BIC is known to asymptotically estimate the "true" model structure (Schwartz, 1978). On the other hand, AIC is targeted towards the minimisation of the generalisation error, and tends to overestimate the model size, as exemplified by the expressions of the penalty terms in (6) and (7). Note, however, that the model studied here (equation 4) is not very flexible: it is necessarily composed of equally weighted isotropic Gaussian distributions. In a case like ours where the data and the clusters are highly non-Gaussian and unbalanced, the Gaussian mixture will try to model each non-Gaussian cluster by a number of smaller isotropic Gaussian components, and larger clusters by a large number of components. Even though BIC will asymptotically give the "right" number of components for *modelling the density* with Gaussian components, the resulting mixture will not necessarily make sense in a clustering context. This is what ICL takes into account by considering not only the fit to the data, but also the resulting classification of each data point into the identified mixture components (Biernacki et al., 1998; Goutte, 1999).

We will first investigate the most parsimonious model, chosen by ICL, using 4 clusters. Figure 7 displays the 4 clusters obtained by K-means, displayed with different markers. As clustering is invariant with respect to permutation of the cluster labels, we will arbitrarily order the resulting clusters according to the strength of their centre. Most of the noisy voxels end up in the third cluster, and are indicated by dots. One cluster has negative activation as well as noise (o). Interestingly, the two remaining clusters display two different levels of positive activation. The least activated of these two clusters also contains a certain amount of noise. Baumgartner et al. (1997), in an analysis involving photic stimulation, also reported a graded pattern of positive response, and attributed the few very activated voxels to medium to large veins running perpendicular to the slice. The extent of cluster 1 and its spatial location (figure 8) suggests that this is not the case here and that this cluster reflects a large, non-artefactual positive

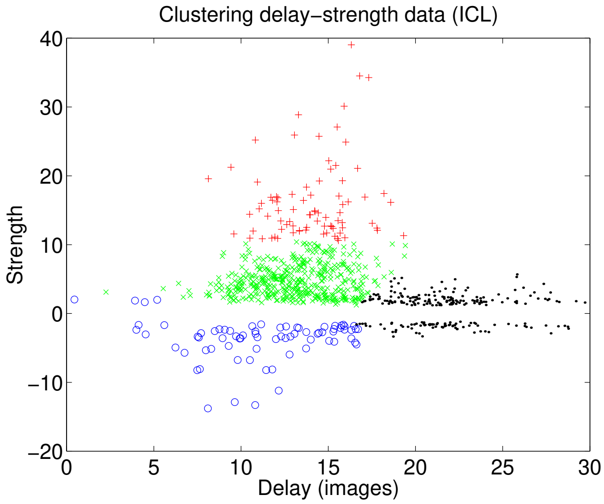

FIG. 7. The 851 voxels (after F-test sieving) described by their delay and strength, partitioned in 4 clusters identified by K-means (maximum of the ICL). One cluster (.) seems to be just noise. Another one (o) contains negatively activated voxels as well as a number of noisy voxels. The last two clusters (x and +) contain positively activated voxels, with two levels of activation, and also some noise (x cluster). Note that the gap around 0 strength is due to the sieving done with the F-test.

activation.

Figure 8 shows the spatial location of the four clusters in the slice. The first cluster (+ in figure 7) contains the largest positive activation. Not surprisingly, it is located in the primary visual area. The second cluster contains a large amount of activation in V1, as well as evidence of activation in the lateral area, and in particular would cover V5. Cluster 3 is clearly noise apart from two small lateral areas in the visual cortex that could be artefacts from the F-filtering. The negative activation in cluster 4 seems to be located at the periphery of the primary visual area, but there is a certain amount of noisy voxels, corresponding to the data at the extreme left of figure 7. Notice that the area delimited by cluster 2 (primary and part of the supplementary motor cortex) seems reasonably symmetric, as expected. However, the strong activation detected in cluster 1 looks very asymmetric, with a dominance on the left side of V1. This very likely reflects an assymetry in the slice position.

We investigate the responses in each of the four clusters by estimating the average response to the stimulus for a typical run. This is done by first averaging the fMRI signal in each cluster, forming the average cluster response. Then the average obtained for all 10 run are overlaid, such that for each image index between 1 and 121, we have 10 possible values (one per run). In figure 9, each individual point represents the average cluster response, for a given image in one of the 10 runs. Accordingly, for each image (between 1 and 121), there are 10 values (10 data points), one for each run. This data is smoothed using a one-dimensional Gaussian Process regression model (Williams, 1998), allowing for error bars on the activation estimates and an estimate of the noise. In figure 9, the estimate of the average activation is indicated by a solid line. Error bars consisting of 2 standard deviations are indicated in dashed lines ("model error bar"), and the error bars taking into account the noise variance ("data error bars") are

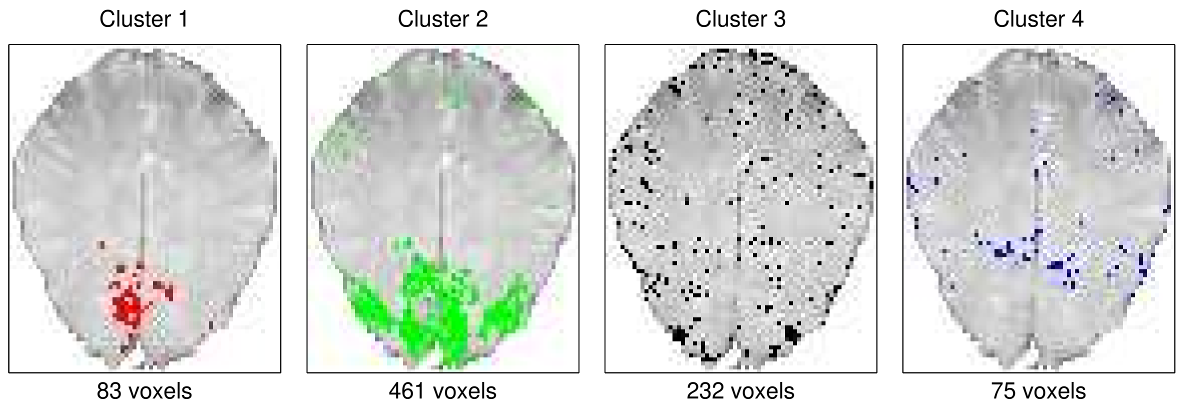

FIG. 8. The four clusters obtained from K-means for the optimal ICL (cf. figure 7), on top of the anatomical background. From left to right, the corresponding markers in figure 7 are $+$, $\times$, $\cdot$ and $\circ$.

indicated by dotted lines. The square wave paradigm is plotted in solid, with dots indicating the scans time stamps. Clearly, the data in clusters 1, 2 and 4 display very significant activations of various strength. The significance is assessed by the fact that the confidence interval for the average activation (dashed lines) lies far away from zero. Furthermore, there is a significant undershoot in cluster 1 around images 85 to 105. An under- or over-shoot of similar *relative* amplitude is also observed in clusters 2, though it fails to display any significance, due to the smaller *absolute* amplitude. Note that in cluster 3 also, the average activation seems to rise significantly above 0 (roughly between images 60 and 80). However, the amplitude of the signal is extremely small, and this can be attributed to an artefact of our sieving method. By construction, all voxels retained by the F-statistic thresholding seemingly display some kind of activation that is consistent with the stimulus. Therefore, their average tends to display a consistent correlation with the stimulus. Based on the smoothed curves, the delay from stimulus onset to 90% of the maximum activation, can be estimated at 6.5, 6.2 and 6.8 seconds for clusters 1, 2 and 4 (respectively). The delay from stimulus cessation to return under 10% of the maximum activation is 6.8, 6.4 and 6.5 seconds (respectively).

Using the structure that maximises BIC ($K = 8$ clusters), the results are only slightly different. Figure 10 shows that the main differences lie in the fact that there are more noisy clusters, and the positive activation is spread in 3 clusters rather than one. Note however that the first two clusters (indicated by '+' and 'x' on figure 10) overlap almost exactly with cluster 1 from figure 8. Accordingly, out of the 8 resulting clusters, we will investigate only the 4 that display some kind of activation. The cluster containing the negative activation (cluster 8, 'o' in figure 10) again contains voxels located at the periphery of the primary visual area (figure 11). The first three clusters in figure 11 contain voxels with decreasing positive activation. Note that cluster 3 (diamonds in figure 10) shows lateral activation in two smaller areas that could correspond to V5.

Again, we investigate the average run activation in each of the four clusters presented in figure 11. Figure 12 presents the average activation, with error bars, smoothed in the same way as above. The

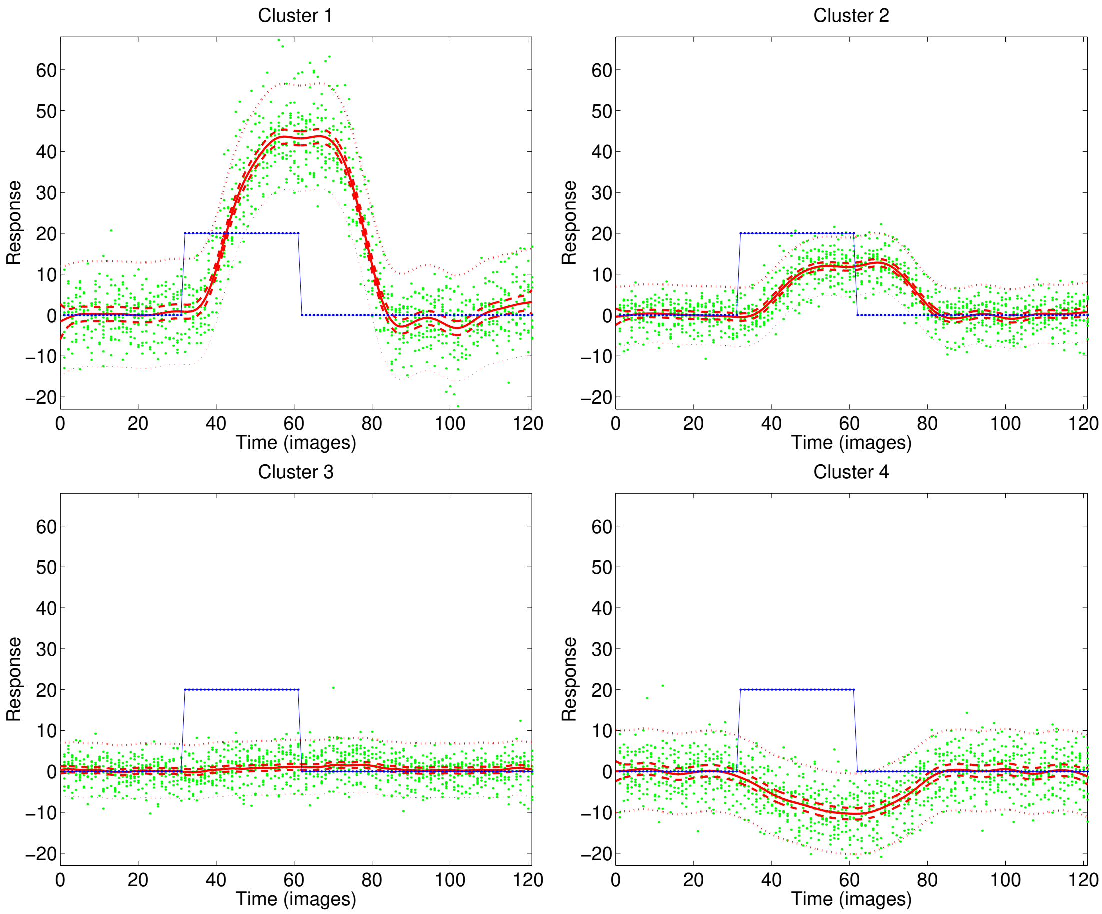

FIG. 9. Average run-response in the four clusters obtained from K-means (maximum of ICL), on the delay-strength data. Dots: average values, over all voxels in a cluster, for all 10 runs; Solid: smoothed value; Dashed: error bars on the smooth; Dotted: error bars on the smooth, plus noise; Solid with dots: paradigm. Note that the y-scale is similar in all plots (10 units represent roughly a 2% relative increase).

significance of the activation appears quite clearly in all plots. The three clusters with positive response (cluster 1 to 3 on figures 11 and 12) display a graded pattern of positive activation. The post-activation undershoot observed in all three clusters around 85 to 105 images is only slightly significant in the two most activated clusters. This weak significance is due to the fact that the clusters contain less voxels than above, hence larger variance and error bars.

Finally, we investigate the difference in the delay measured for the different clusters. In figure 13, the distributions of the delays in each cluster are approximated by Gaussian distributions. The delays for the negative activation clearly seems to be 1 to 1.5 seconds shorter on average that the delays measured for the positive activation (Buxton et al., 1998). We assess the significance of this difference using a non-parametric Kolmogorov-Smirnov test. In all cases, the difference is found to be highly significant, with p-values well under 10^{-4} (a t-test using the Gaussian assumption provides similar results).

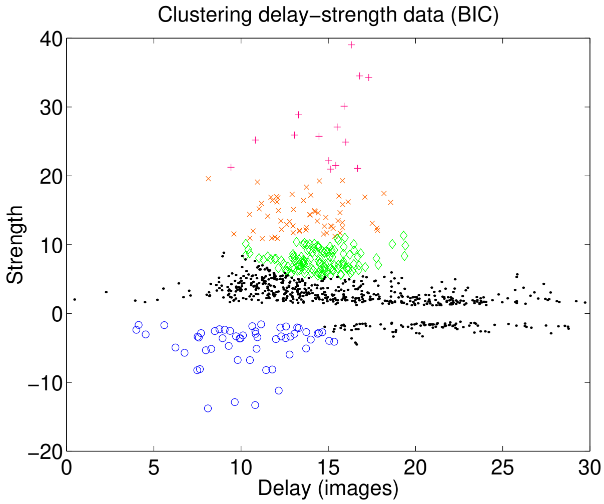

FIG. 10. The 851 voxels described by their delay and strength, partitioned in 8 clusters identified by K-means (maximum of BIC). Four clusters contain noisy voxels (.); one (o) contains negatively activated voxels as before; three clusters (x, +, and diamond) contain positively activated voxels, with decreasing levels of activation.

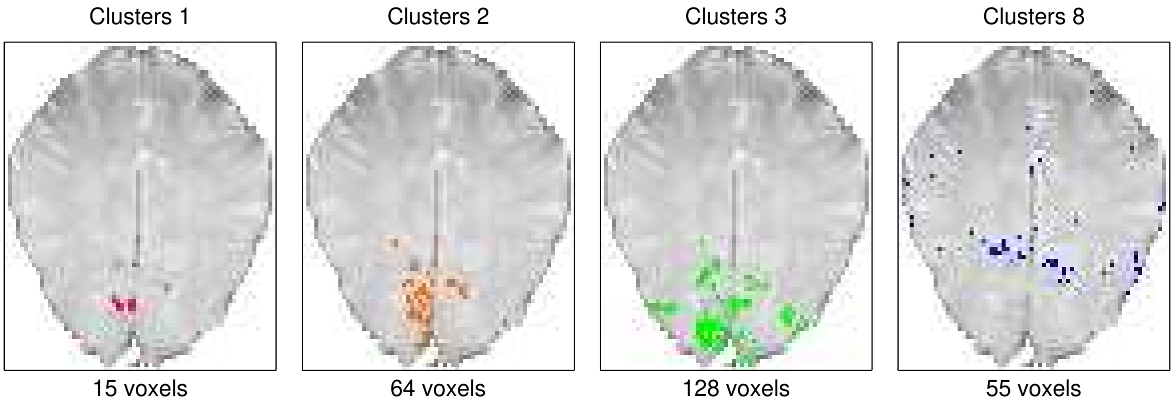

FIG. 11. Four of the 8 clusters obtained from K-means for the optimal BIC (cf. figure 10), on top of the anatomical background. From left to right, the corresponding markers are $+$, $\times$, $\diamond$ and $\circ$. Notice the similarity between cluster 1 in figure 8 and the union of clusters 1 and 2 here.

Meta clustering

The 7 features are extracted using the "Lyngby" modelling toolbox (Hansen et al., 1999). As the features have very different ranges (eg between 0 and 0.5 for the correlation, and between -25 and 55 for the Gamma filter strength), we normalise the data as explained in the methods section before clustering. All the results presented below will be shown in the original, non normalised domain, in order to be able to relate to their actual meaning (eg delays in images, t-statistics, etc).

On the 889 resulting points in a 7-dimensional feature space, we apply K-means with increasing number of clusters up to $K = 60$. The ICL criterion has a first maximum at $K = 26$ and a second one for $K = 35$, while AIC and BIC pick larger amounts of clusters (figure 14). In particular, AIC does not seem to reach a maximum before $K = 60$. Note that even 26 is a rather large number of clusters. However it is a

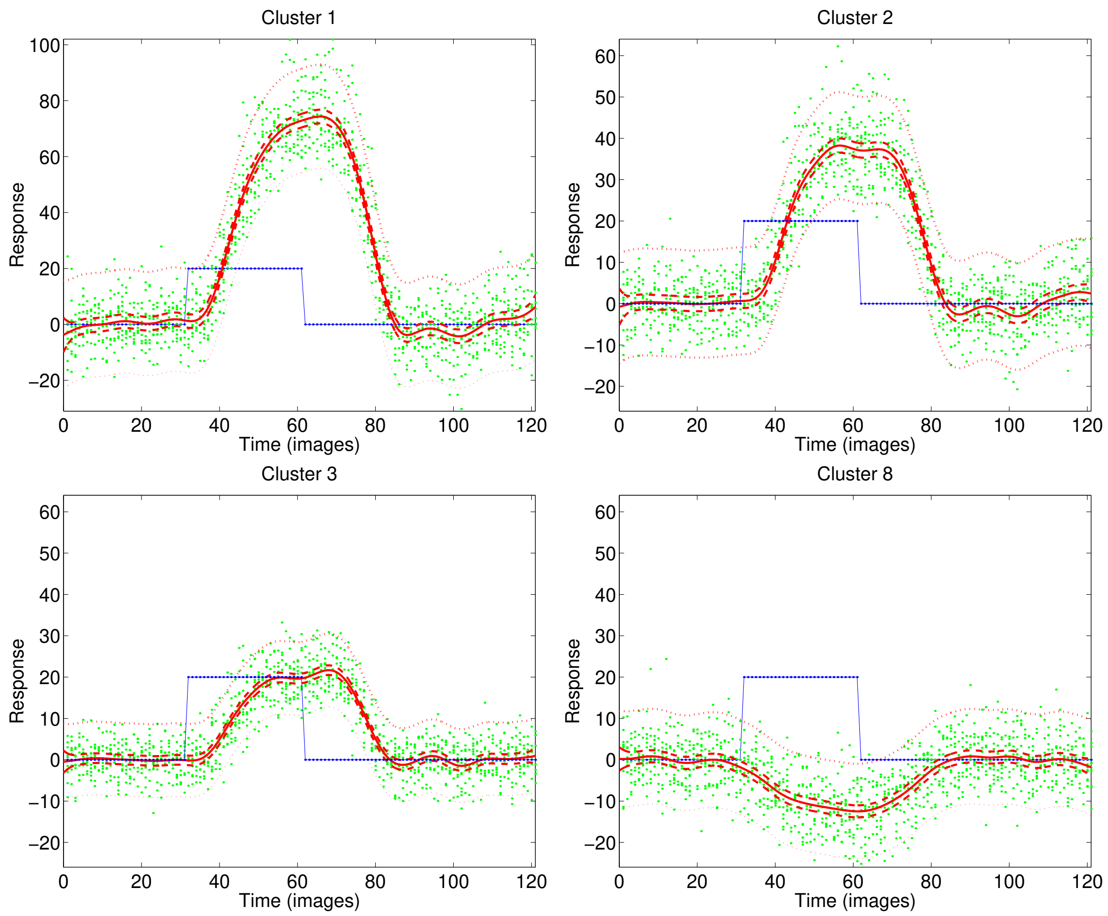

FIG. 12. Average run-response in the four clusters displayed in figure 11, obtained on the delay-strength data. Dots: average values over all voxels in the cluster, for all 10 run; Solid: smoothed value; Dashed: error bars on the smooth; Dotted: error bars on the smooth, plus noise; Solid with dots: paradigm. Note that all plots except for cluster 1 (top left) have the same y-scale.

requirement of the K-means algorithm that all clusters should be of the same size and weight, and as such most of the clusters obtained will in practice cover non- or weakly activated voxels. This is illustrated on figure 15, where we show the data together with the 26 cluster centres. As labeling is arbitrary, we will use the t-statistic as a reference and label the clusters according to the t-statistic of the cluster centre, ie the first dimension in the feature space. For example, the cluster with largest t is labelled 1, and the cluster with the most negative t is labelled 26 (figure 15). In the following we consider only the case of $K = 26$ clusters, for two main reasons: 1) even though the ICL has a maximum in $K = 35$, it lies within the error bars of the the first extremum, in $K = 26$; 2) investigation of the clustering results for $K = 35$ shows that the difference lies mainly in the partition of the non-activated voxels. In the clusters displaying possible (negative or positive) activation, there is little difference, such that the following analysis and conclusion apply.

Figure 15 shows that only a few clusters cover the activated data, which by construction correspond to extreme values for most of the features. Only cluster 26 contains negatively activated data (leftmost in the top plots), and 6 clusters contain the most positive activation. It is already clear from the plots

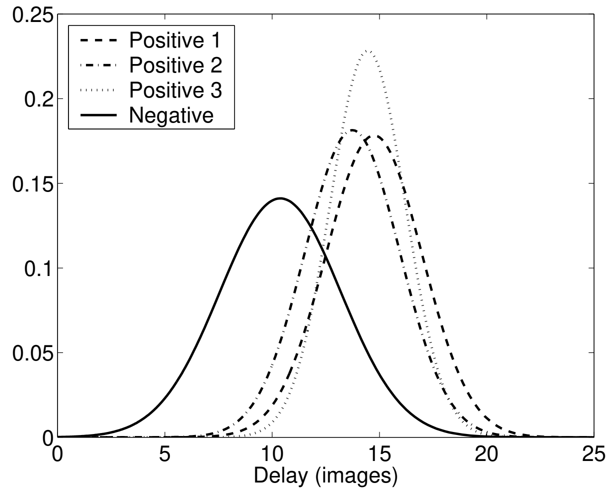

FIG. 13. Gaussian approximations of the distributions of delays in the four clusters displayed in figure 11. Clusters 1 to 3 are labelled 'Positive 1' to 'Positive 3' in the legend, and cluster 8 is labelled 'Negative'. The delay for the negative activation seems to be 1 to 1.5 seconds (3 to 5 images) shorter than the delay for positive activation. This difference is highly significant.

that all features do not carry the same information. In particular, the ordering of the clusters according to the different features can differ markedly. Obviously, it is impossible to differentiate the cluster with negative activation (cluster 26) from clusters with positive activation using features with positive values such as the Kolmogorov-Smirnov statistics or the standard deviation of the fitted FIR filter (features 2 and 4). It is also interesting to notice that the clusters with the largest positive activation (clusters 1 to 4) are not distributed in the same way along all features. Their very different behaviour, according to eg the FIR standard deviation or Gamma filter strength, is not reflected in the Kolmogorov-Smirnov statistic or the t-statistics, for which eg clusters 1 and 2 are nearly identical. Another interesting aspect is that cluster 11, which seems fairly nondescript according to the t and Kolmogorov-Smirnov statistics, has a much different status according to the correlation, standard deviation of the FIR filter and Gamma filter strength (see in particular the top right subplot in figure 15)

Ordering of clusters: It is interesting to analyse the clustering results from the point of view of the different features. If the features were totally redundant, they would produce the same ordering of the clusters: the cluster with the largest t-statistic would also have the largest correlation, the largest strength of the fitted Gamma filter, and so on. However, we expect that this is not the case. Different activation-related features like the t-statistics or the strength of the Gamma filter are based on different assumptions and focus on different aspects of the fMRI signal. Hence clusters will display different patterns of activation according to different features, as noticed in figure 15.

Table 1 illustrates this by showing the order of the different clusters along the 5 activation-based features, ie excluding the delays. As the t-statistic was used to order the resulting clusters, the first column can be used as a reference. Notice that the assessment of cluster 11 along the first two features

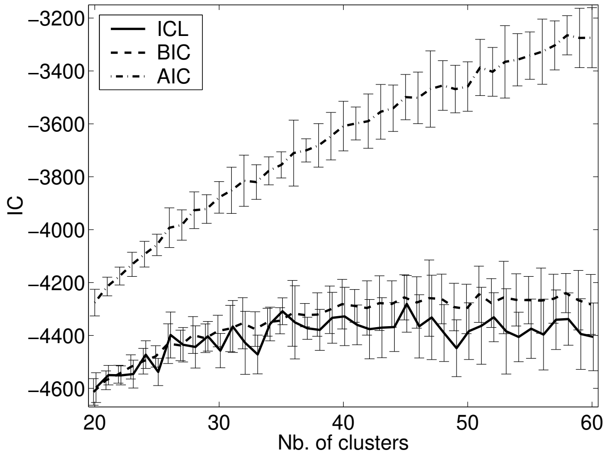

FIG. 14. The three information criteria, calculated on the 7 features for increasing numbers of clusters: AIC (dash-dotted), BIC (dashed) and ICL (solid). To increase readability, the BIC and ICL error bars have been shifted slightly. The first maximum is obtained for $K = 26$ for ICL, much larger values for AIC and BIC. Notice how much more ICL penalises large numbers of clusters.

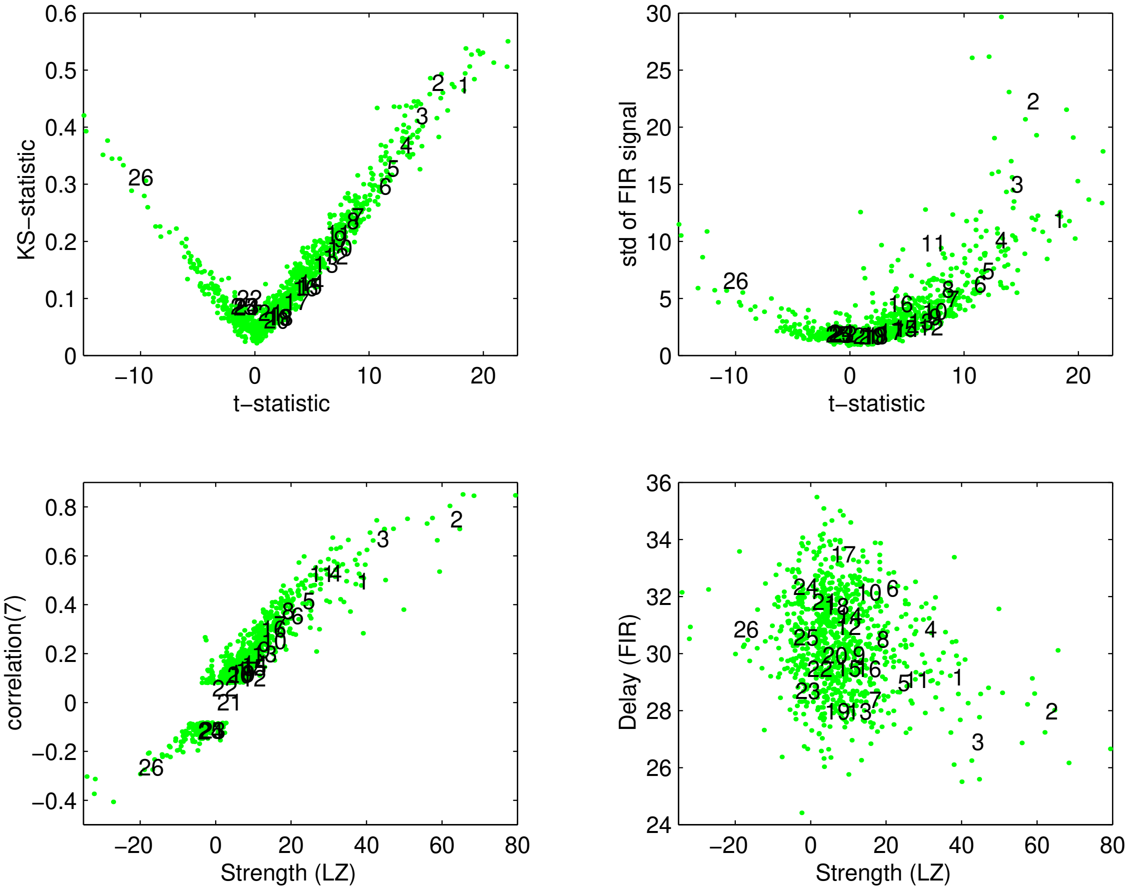

FIG. 15. Results of the meta-clustering. Dots indicate data point, numbers show cluster centres. A choice of features are plotted against each other. Note that there is clearly one cluster with negative activation (26) and several clusters with positive activations, which sometimes seem to overlap on those 2D plots (especially left plots).

t-statistic	Kolmogorov-Smirnov statistic	correlation at 7 seconds	standard deviation of FIR signal	Strength of Lange-Zeger filter
1	2	2	2	2
2	1	3	3	3
3	3	4	1	1
4	4	11	4	4
5	5	1	11	11
6	26	5	5	5
7	6	8	26	6
8	7	6	6	8
...	...	...	...	...
25	18	24	18	24
26	20	26	20	26

Table 1. Ordering of clusters along the 5 activation-based features, from largest value (top) to lowest (bottom). The t-statistic is chosen as reference. Notice the markedly different ordering, especially for cluster 1 and cluster 11 in the last 3 columns. The high rank of cluster 26 in columns 2 and 4 is due to the fact that these features are always positive.

(first two columns) and along features 3, 4 and 6 (last three columns) are totally different.

Spatial location: The voxels belonging to each of the clusters displaying the strongest activations are located in the visual cortex and primarily V1. Figure 16 shows the location of clusters 1 to 4, as well as clusters 11 and 26. In addition, sizeable portions of what appears to be area V5 are contained in clusters 7 to 10. Note that even though the physical location of the clusters is not taken into account in the clustering process, voxels with similar patterns of activation seem to be localised in the neighbourhood of each other. A notable exception is cluster 11, which seems to be distributed over a fairly large area, with no clear grouped localisation. Cluster 26 shows voxels with negative response distributed exclusively in the periphery of visual areas. In that case, the large number of clusters adopted here leads to a more precise location than in the previous (2D) analysis, with notably less artefactual voxels. Interestingly, cluster 2 is almost identical to cluster 1 in the 8-cluster analysis above (maximising the BIC). As this cluster contained the 15 most activated voxels according to the cross-correlation analysis, it is somewhat surprising that it would come only second according to the t-statistic (cf. table 1). An individual analysis of the estimated time course in each voxel provides an explanation to this phenomenon.

Response patterns: The typical run response in each cluster is estimated using the same smoothing method as above. Figure 17 presents the results in 4 clusters which are ordered differently according to table 1. Clusters 2 and 3 (left plots in figure 17) show very similar patterns, with different activation

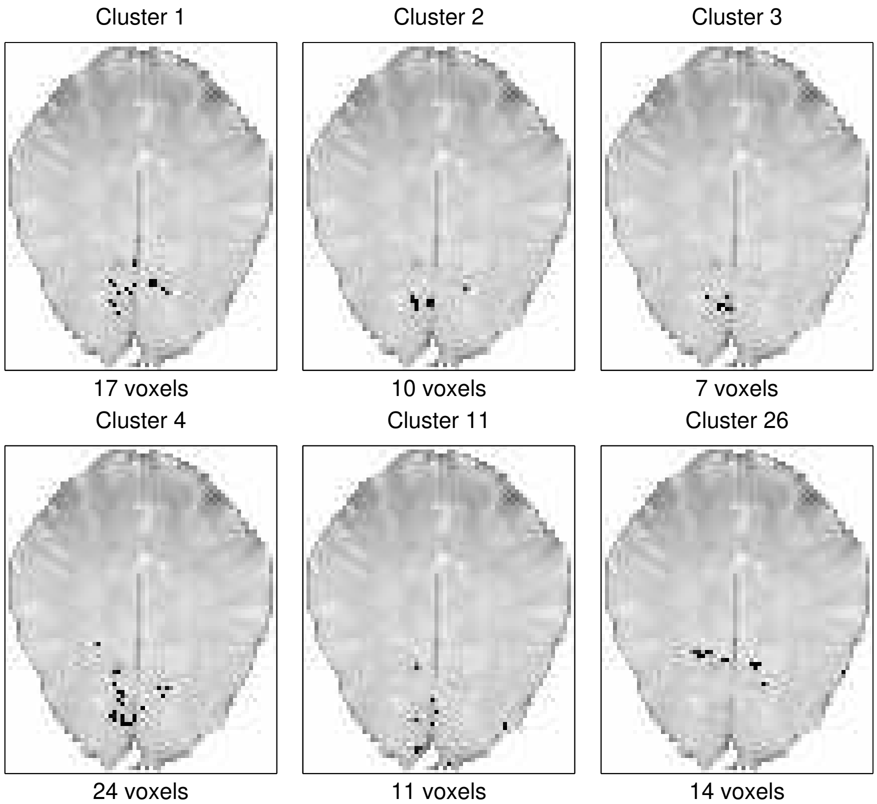

FIG. 16. Spatial location of the voxels from 6 clusters obtained from meta-clustering ($K = 26$, maximum of the ICL), cf. figure 15.

levels. This is not surprising as their ordering seem to be very consistent along all features (cf. figure 15 and table 1). More interesting is the difference between clusters 1 and 2 (or 3). The average signal in cluster 1 seems to rise more sharply and faster that in the remaining clusters. This does undoubtedly favour the t-statistic, which relies heavily on the square wave design, and does not properly take into account possible delays in the activation. Accordingly, the faster signal rise in cluster 1 compensates for the lower activation, compared to cluster 2.

On the contrary, cluster 11 displays a much delayed, and softer rise in signal. Features obtained from filters that model the activation delay properly do correctly place cluster 11 on par with clusters that have similar activation strength (clusters 4 and 5). Similarly, the correlation calculated with a suitably delayed signal ranks cluster 11 properly. On the contrary, the reliance of the t-statistic and Kolmogorov-Smirnov statistics on the square wave design, as mentioned above, is badly hurt by the large delay, which places more than half the activation after stimulus cessation.

The difference in activation patterns is further investigated by calculating, on the smoothed response, the delay from stimulus onset to 90% of the maximum signal change during activation and from stimulus cessation to return within 10% of the baseline. Calculated on the large clusters obtained in the previous cross-correlation analysis, these delays were fairly similar across clusters. With the increased resolution brought by the large number of clusters, the picture offered by our meta-clustering scheme is very different, as shown on figure 18. Cluster 26 has a much faster onset response, which seems consistent with our

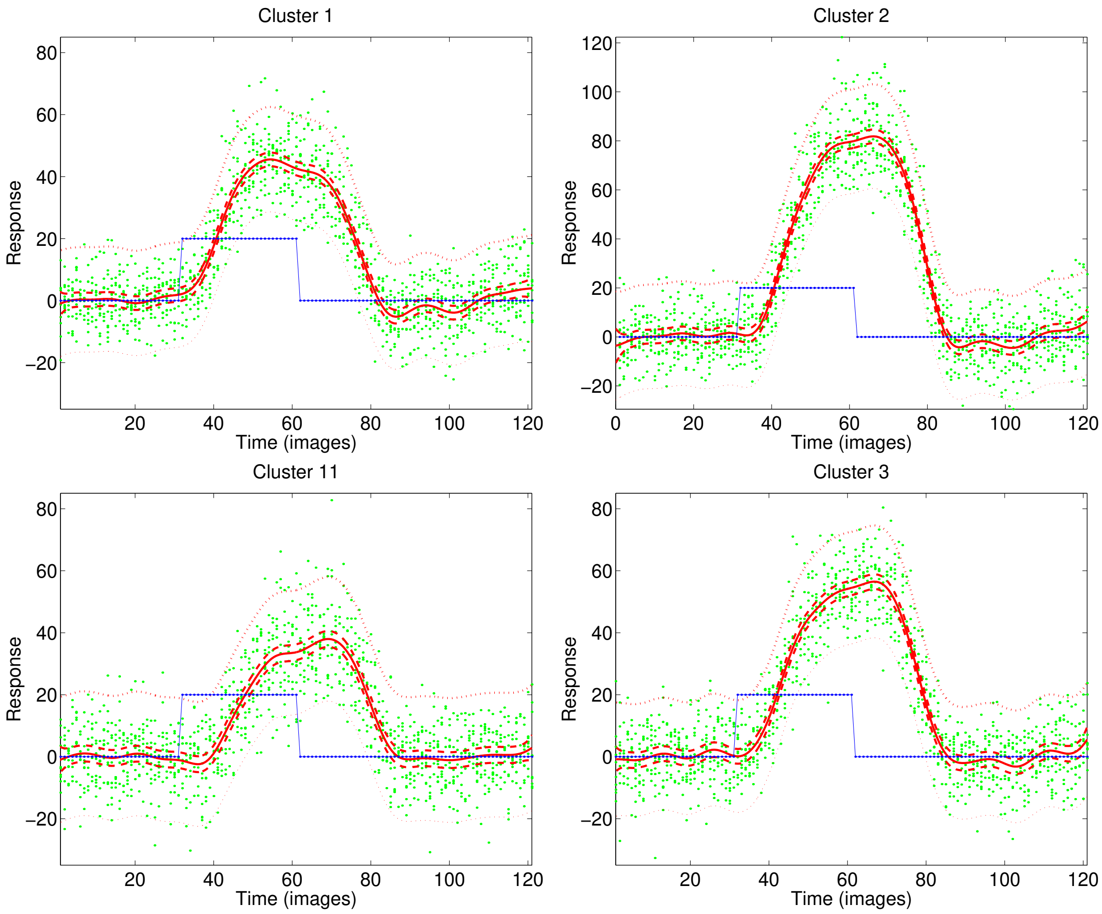

FIG. 17. Response in four clusters obtained with meta-clustering on 7 features ($K = 26$, maximum of the ICL). Notice the markedly different response in clusters 1 and 11 (right), compared to the more "typical" response in clusters 2 and 3 (left). In particular, cluster 1 seems to have a sharper and faster rise, while cluster 11 has a slower and more delayed rise.

findings in the previous analysis (see also Goutte et al., 1998). The voxels from cluster 1 react 1 to 2 seconds faster than voxels in clusters 2 to 4. This gives a quantitative measure of the difference that we noticed above from figure 17. Finally, cluster 11 has clearly a much slower response time, both to onset and offset of the stimulus.

Note that the differences outlined above do only come out thanks to the feature extraction and meta-analysis. Even though we can see that eg clusters 11 and 3 are clearly different (cf. figures 17 and 18), the actual distance, in time series space (ie the squared difference between the activation patterns) is insufficient to differentiate between these clusters using clustering on the raw data.

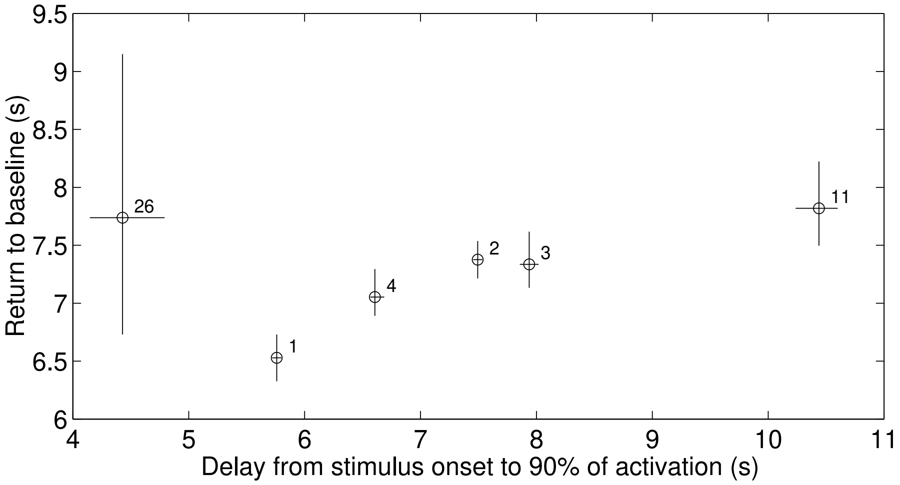

FIG. 18. Delays (with error bars) between stimulus onset and 90% of activation (x-axis) and between stimulus cessation and return within 10% of the baseline (y-axis), for 6 clusters obtained with meta-clustering on 7 features (numbers on the plot). Note that cluster 11 has a notably delayed response, while clusters 26 and 1 have a relatively faster onset response.

DISCUSSION

Feature extraction

There is an almost endless supply of possible features to be used in conjunction with clustering.

One possibility is to use the feature-space as a dimensionality reduction device, and as a way to focus on features that are relevant to the investigator. Our first example used the strength and delay extracted from cross-correlation, and showed for example that we can isolate clusters with significantly different delays in different areas of the brain. This is particularly the case for the voxels displaying a delayed negative correlation with the stimulus, for which the delay is significantly shorter than that of other activated clusters (figure 13).

Another possibility is to use clustering as a meta-analysis tool. In that context, the features are the results of previous analyses performed on the data. This is exemplified by our second example, where we cluster the results obtained from 5 different analyses. Other work along the same lines has focused on clustering the coefficients of a linear FIR filter (Purushotham et al., 1999). The meta-clustering analysis has been able to spot marked differences in the pattern of activations, eg spotting voxels that react strongly according to some analyses, while being fairly nondescript according to others (cluster 11 in the analysis above). With a traditional visual inspection, this means that this cluster would be highlighted in some activation maps, and much less on others.

Clustering

Most of the methods used so far to cluster fMRI data use either a user-defined number of cluster, or some kind of heuristic split-and-merge strategy, *à la* ISODATA (Tou and Gonzalez, 1974, section 3.3.6). This approach can create problems if eg the number of clusters or the parameters used by the heuristics are insufficiently documented. In addition, these user-defined parameters usually require expertise or at least prior information, from the user, making them unsuitable for a totally *plug and play* analysis, especially by users who are not necessarily statistically literate.

We have modelled the K-means algorithms as a particular instance of learning a mixture model (appendix A). There are a number of restrictions in this model. In particular the Gaussian distributions that constitute the mixture are restricted to have isotropic covariances and equal weights. Both of these restrictions can have a serious influence on the results. For example, the equal weighting will favour large numbers of clusters in high density areas. In the context of fMRI data, this means that a lot of clusters will map the relatively small area in space corresponding to noise or lack of activation. The isotropic covariance has two side-effects: 1) all clusters are constrained to have identical shapes 2) it forces the algorithm to consider each dimension in space equally. This second point can be partially addressed by a normalisation procedure, or using an appropriate metric (Goutte et al., 1999).

More flexible mixture models can be considered. For example, Biernacki et al. (1998) detail 28 different models, from the less flexible (corresponding to K-means) to a model with totally flexible weights and covariances. These models can be trained using the same techniques, Expectation Maximisation (EM) and Classification EM (CEM), that we mention in the appendix. Furthermore, the structure can be tuned using the same information criteria, AIC, BIC or ICL, that we have used here. However, the description and use of these general model is well beyond the scope of this paper.

Average activation and significance

In order to estimate the average time course in each voxel, we have used a smoothing method that provides error bars on the smoothed estimate and on the data (figures 9, 12 and 17). The distribution of the data points inside and outside the largest error bars suggest that the assumptions underlying the smoothing process seem reasonable. Though the error bars on the estimate provide a potentially useful way of assessing the significance of an hypothesised activation, we have refrained from putting to much emphasis on this aspect. Indeed the clustering process by construction provides voxels which are expected to have homogeneous patterns of activation, even though the actual time course was never directly taken into account during clustering. The fact that the smoothing occurs within clusters of homogeneous voxels will tend to yield optimistic error bars. The use of this smoothing procedure in conjunction with clustering therefore needs to be investigated further.

Negative activation

In the clustering analyses presented above, we have identified a cluster containing voxel that display a negative pattern of activation. This is consistent with previous studies performed on similar data (Goutte et al., 1998, 1999). Note that this phenomenon has been observed by several authors, and a few explanations have been proposed (Benali et al., 1999; Raichle, 1999). In the context of this study, it seems to us that this negative activation occurs in conjunction with the presence of large venous vessels. This is corroborated by the fact that the spectrum of the signal measured in these voxels contains a large contribution of a frequency that corresponds to the subject's cardiac rhythm. However, we would expect that the signal measured in veins should show a *delayed* activation, not an *anticipated* one.

The presence of negative activation is an interesting open issue for further investigation using eg parametric models of the hemodynamic response. Note also that the experiment analysed here was performed on several subjects, as well as performed several time on the same subject. This gives the possibility to check the reproducibility of this effect both in the same subject and in the general population. We expect that further analysis of this data will allow us to reach a better understanding of this kind of negative activation.

Initial dip

The initial negative response or initial dip has attracted much interest recently (eg Menon et al. (1995); Hu et al. (1997); Yacoub et al. (1999); Hathout et al. (1999)) and has been observed in several situation, even for low field intensity (Yacoub and Hu, 1999). The high temporal resolution of our data and the relatively high signal-to-noise ratio (for fMRI experiments) obtained using visual stimulation potentially provide interesting data for studying this effect. Despite these favourable conditions, we have been unable to observe this effect, at least at cluster level. This does not rule out the presence of an initial dip in some individual voxels. However, this effect, if it is at all present in our data, was not strong enough to provoke the appearance of a cluster of voxels with consistent initial dip. Here again, further analysis using more subjects, and possibly higher field intensity might reveal additional structure in the data.

CONCLUSIONS

Cluster analysis in feature space provides an original scheme for mapping the spatio-temporal distribution of focal neural activation. It can be used either as an analysis to focus on relevant features of the fMRI sequence, or as a meta-analysis tool to explore consensus and disagreement between previous single-voxel analyses. The difficult problem of choosing the relevant number of clusters is addressed in a principled way. We model the K-means clustering process using a mixture model approach, and select the structure that maximises an information criterion, over multiple stochastic initial configurations. Using this approach, we obtain a totally "plug-in" analysis, reducing the dependency on the initial condition and user-dependent choices. An additional advantage is that the results can easily be replicated.

In this contribution, we have shown how feature space clustering can be applied to a short analysis of the delay-strength distribution of fMRI time series, as well as to meta-clustering of the results of common voxel-based analyses obtained from a public domain fMRI toolbox. The examples shows that the approach is versatile and can be applied in different contexts (analysis as well as meta-analysis) on fMRI data. The results show that we are able to identify effects that would not be available using standard clustering on the raw fMRI time series, or not completely apparent using individual single-voxel analyses.

FUNDING ACKNOWLEDGEMENTS:

The authors thank Carl Edward Rasmussen for help using his Gaussian Process smoothing software. This research was supported by the EU, through a BIOMED II grant no. BMH4-CT97-2775, and the Danish Research Councils through the Danish Computational Neural Network Center (CONNECT) and the THOR Center for Neuroinformatics.

APPENDIX A: INFORMATION CRITERIA FOR K-MEANS

From a statistical point of view, the K-means algorithm is a particular instance of the Classification Expectation Maximisation algorithm (see Celeux and Govaert, 1992, for definition and theoretical results) for a Gaussian mixture model with equal mixture weights and equal isotropic variances. The underlying density, for a data point $\boldsymbol{u}_j \in I\!R^Q$, is expressed as:

$$P(\boldsymbol{u}_j|\boldsymbol{M}, \sigma^2) = \frac{1}{K} \sum_{k=1}^{K} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2}\frac{(\boldsymbol{u}_j - \boldsymbol{\mu}_k)^2}{\sigma^2}\right] \tag{2}$$

a mixture of K equally weighted Gaussian distributions with mean $\boldsymbol{\mu}_k$ and common variances σ^2. In our notation, $\boldsymbol{M}$ contains all the means $\boldsymbol{\mu}_k$, $1 \leq k \leq K$. For K-means, we take the mean of each Gaussian to be the centre of the resulting clusters, $\boldsymbol{\mu}_k = \boldsymbol{c}_k$ and the variance σ^2 to be the within-cluster variance:

$$\sigma^2 = \frac{1}{N} \sum_{k=1}^{K} \sum_{j \in C_k} \|\boldsymbol{u}_j - \boldsymbol{c}_k\|^2 \tag{3}$$

where N is the number of vectors we are clustering. Using (4), we can calculate the likelihood of the entire dataset $\mathcal{D} = \{\boldsymbol{u}_j\}$:

$$P(\mathcal{D}|\boldsymbol{M}, \sigma^2) = \prod_j P(\boldsymbol{u}_j|\boldsymbol{M}, \sigma^2) \tag{4}$$

Alternatively, the classification likelihood is calculated by assigning each point $\boldsymbol{u}_j$ to the mixture component k_j that has the highest probability:

$$\boldsymbol{u}_j \sim \mathcal{N}(\boldsymbol{\mu}_{k_j}, \sigma^2) \qquad P(\boldsymbol{u}_j|\boldsymbol{M}, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2}\frac{\left(\boldsymbol{u}_j - \boldsymbol{\mu}_{k_j}\right)^2}{\sigma^2}\right] \tag{5}$$

where k_j is the cluster to which $\boldsymbol{u}_j$ is assigned. The joint classification likelihood for the entire dataset is then calculated like in (4).

Within this probabilistic framework, we can see the K-means algorithm as an attempt at maximising the joint negative log-likelihood (or classification log-likelihood) of the data. This gives a precise probabilistic justification to the algorithm presented above. In addition, we can invoke several information criteria in order to estimate the optimal number of clusters:

- Akaike's first information criterion (Akaike, 1974):

$$AIC = \ln P(\mathcal{D}|\boldsymbol{M}, \sigma^2) - (KQ + 1) \tag{6}$$

- Schwartz' Bayesian information criterion (Schwartz, 1978):

$$BIC = \ln P(\mathcal{D}|\boldsymbol{M}, \sigma^2) - \frac{KQ+1}{2} \ln N \tag{7}$$

- The integrated classification likelihood (Biernacki et al., 1998):

$$ICL = \ln P(\mathcal{D}|\boldsymbol{M}, \sigma^2) - \frac{KQ+1}{2} \ln N - \sum_{i=1}^{N} \ln\left(i + \frac{K+2}{2}\right) + \sum_{k=1}^{K}\sum_{j=1}^{N_k} \ln\left(j + \frac{3}{2}\right) \tag{8}$$

where in the above criteria, Q is the dimension, ie the number of features, N is the total number of data, and N_k is the number of data in cluster C_k such that, $\sum_k N_k = N$. Note that for ICL, we use the classification likelihood (5), not the density likelihood (2) to calculate the criterion. In all three cases, we will select the number of clusters K, which yields the highest value of the information criterion.

REFERENCES

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723.

Ashburner, J., Haslam, J., Taylor, C., Cunningham, V. J., and Jones, T. (1996). A cluster analysis approach for the characterization of dynamic PET data. In Myers, R., Cunningham, V., Bailey, D., and Jones, T., editors, *Quantification of Brain Function Using PET*, pages 301–306, San Diego. Academic Press.

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):1094–1101.

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.

Benali, H., Di Paola, M., Burnod, Y., Pélégrini, M., Buvat, I., Garnero, L., Lehericy, S., and Di Paola, R. (1999). A multivariate model for estimating non stationary hemodynamic response functions in event-related fMRI. In Rosen et al. (1999), page S5.

Biernacki, C., Celeux, G., and Govaert, G. (1998). Assessing a mixture model for clustering with the integrated classification likelihood. Rapport de Recherche 3521, INRIA.

Bishop, C. M. (1995). *Neural Networks for pattern recognition*. Clarendon Press, Oxford.

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

Buxton, R. B., Wong, E. C., and Frank, L. R. (1998). Dynamics of blood flow and oxygenation changes during brain activation: The ballon model. *Magnetic Resonance in Medicine*, 39(6):855–864.

Celeux, G. and Govaert, G. (1992). A Classification EM algorithm for clustering and two stochastic versions. *Computational Statistics and Data Analysis*, 14:315–332.

Ding, X., Tkach, J., Ruggieri, P., and Masaryk, T. (1994). Analysis of the time-course functional MRI data with clustering method without use of reference signal. In *Proceedings of the Second Annual Meeting of the International Society for Magnetic Resonance in Medicine*, page 630.

Domany, E. (1999). Superparamagnetic clustering of data — the definitive solution of an ill-posed problem. *Physica A*, 263(1-4):158–169.

Filzmoser, P., Baumgartner, R., and Moser, E. (1999). A hierarchical clustering method for analyzing functional MR images. *Magnetic Resonance Imaging*, 17(6):817–826.

Friberg, L., Gjedde, A., Holm, S., Lassen, N. A., and Nowak, M., editors (1997). *Proceedings of the Third International Conference on Functional Mapping of the Human Brain*, number 5:4, part 2 of 4 in NeuroImage. Academic Press.

Golay, X., Kollias, S., Stoll, G., Meier, D., Valavanis, A., and Boesiger, P. (1998). A new correlation-based fuzzy logic clustering algorithm for fMRI. *Magnetic Resonance in Medicine*, 40(2):249–260.

Goutte, C. (1999). Non parametric clustering with an adaptive metric, with application to signal detection in fMRI time series. submitted manuscript.

Goutte, C., Nielsen, F. Å., Svarer, C., Rostrup, E., and Hansen, L. K. (1998). Space-time analysis of fMRI by feature space clustering. In Paus, T., Gjedde, A., and Evans, A., editors, *Proceedings of the Fourth International Conference on Functional Mapping of the Human Brain*, number 7:4, part 2 of 3 in NeuroImage, page S610. Academic Press.

Goutte, C., Toft, P., Rostrup, E., Nielsen, F. Å., and Hansen, L. K. (1999). On clustering fMRI time series. *NeuroImage*, 9(3):298–310.

Hansen, L. K., Nielsen, F. Å., Toft, P., Liptrot, M. G., Goutte, C., Strother, S. C., Lange, N., Gade, A., Rottenberg, D. A., and Paulson, O. B. (1999). "lyngby" — a modeler's Matlab toolbox for spatio-temporal analysis of functional neuroimages. In Rosen et al. (1999), page S241.

Hathout, G. M., Varjavand, B., and Gopi, R. K. (1999). The early response in fMRI: A modeling approach. *Magnetic Resonance in Medicine*, 41(3):550–554.

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

Hu, X., Le, T. H., and Ugurbil, K. (1997). Evaluation of the early response in fmri in individual subjects using short stimulus duration. *Magnetic Resonance in Medicine*, 37(6):877–884.

Lange, N., Strother, S. C., Anderson, J. R., Nielsen, F. Å., Holmes, A., Kolenda, T., Savoy, R., and Hansen, L. (1999). Plurality and resemblance in fMRI data analysis. *NeuroImage*. to appear.

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):1–30.

Mckeown, M. J., Makeig, S., Brown, G. G., Jung, T.-P., Kindermann, S. S., Bell, A. J., and Sejnowski, T. J. (1998). Analysis of fMRI data by blind separation into independent spatial components. *Human Brain Mapping*, 6(3):160–188.

Menon, R. S., Ogawa, S., Hu, X., Strupp, J. P., Anderson, P., and Ugurbil, K. (1995). BOLD based functional MRI at 4 Tesla includes a capillary bed contribution: echo-planar imaging correlates with previous optical imaging using intrinsic signals. *Magnetic Resonance in Medicine*, 33(3):453–459.

Moser, E., Baumgartner, R., Barth, M., and Windischberger, C. (1999). Explorative signal processing in functional MR imaging. *International Journal of Imaging Systems Technology*, 10(2):166–176.

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.

Nielsen, F. Å., Hansen, L. K., Toft, P., Goutte, C., Lange, N., Strother, S. C., Mørch, N., Svarer, C., Savoy, R., Rosen, B., Rostrup, E., and Peter, B. (1997). Comparison of two convolution models for fMRI time series. In Friberg et al. (1997), page S473.

Purushotham, A., Nielsen, F. Å., Hansen, L. K., and Kim, S. (1999). Separation of motor preparation and execution regions using meta-K-means clustering on fMRI single trial data. In Rosen et al. (1999), page S51.

Raichle, M. (1999). Closing lecture at the Fifth International Conference on Functional Mapping of the Human Brain.

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

Rosen, B. R., Seitz, R., and Volkmann, J., editors (1999). *Proceedings of the Fifth International Conference on Functional Mapping of the Human Brain*, number 9:6, part 2 of 2 in NeuroImage. Academic Press.

Schwartz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464.

Somorjai, R. (1999). Personnal communication.

Somorjai, R. L., Jarmasz, M., and Baumgartner, R. (1999). Exploratory Data Analysis of functional MRI data — comparing the influence of different preprocessing techniques and distance metrics on fuzzy clustering analysis. *submitted.*

Sychra, J. J., Bandettini, P. A., Bhattacharya, N., and Lin, Q. (1994). Synthetic images by subspace transforms. I. principal components images and related filters. *Medical Physics*, 21(2):193–201.

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.

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

Wand, M. and Jones, M. (1995). *Kernel Smoothing.* Number 60 in Monographs on Statistics and Applied Probability. Chapman & Hall, London.

Williams, C. K. I. (1998). Prediction with Gaussian processes: From linear regression to linear prediction and beyond. In Jordan, M. I., editor, *Learning and Inference in Graphical Models.*

Wismüller, A., Dersch, D. R., Lipinski, B., Hahn, K., and Auer, D. (1998). A neural network approach to functional MRI pattern analysis — clustering of time-series by hierarchical vector quantization. In

Niklasson, L., Bodén, M., and Ziemke, T., editors, *Proceedings of the 8th International Conference on Artificial Neural Networks*, Perspectives in Neural Computing, pages 857–862, London. Springer.

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.

Yacoub, E. and Hu, X. (1999). Detection of the early negative response in fMRI at 1.5 tesla. *Magnetic Resonance in Medicine*, 41(6):1088–1092.

Yacoub, E., Le, T. H., Ugurbil, K., and Hu, X. (1999). Further evaluation of the initial negative response in functional magnetic resonance imaging. *Magnetic Resonance in Medicine*, 41(3):436–441.