Feature-Space Clustering for fMRI Meta-Analysis

Cyril Goutte,^{1*} Lars Kai Hansen,² Matthew G. Liptrot,² and Egill Rostrup³

¹ INRIA Rhône-Alpes, Montbonnot, France

² Department of Mathematical Modeling, Technical University of Denmark, Lyngby, Denmark

³ Danish Research Centre for Magnetic Resonance, Hvidovre, Denmark

Edited by: Karl Friston, Associate Editor

Abstract: Clustering functional magnetic resonance imaging (fMRI) time series has emerged in recent years as a possible alternative to parametric modeling approaches. Most of the work so far has been 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. In particular this allows the checking of the differences and agreements between different methods of analysis. Both approaches are illustrated on a fMRI data set involving visual stimulation, and we show that the feature space clustering approach yields nontrivial results and, in particular, shows interesting differences between individual voxel analysis performed with traditional methods. *Hum. Brain Mapping* 13:165–183, 2001. © 2001 Wiley-Liss, Inc.

Key words: clustering, fMRI, feature extraction, meta analysis, Gaussian mixture model, information criteria

INTRODUCTION

Feature-space clustering

The high temporal resolution of functional magnetic resonance imaging (fMRI) has inspired a host of single-voxel analysis methods. They typically compute some summary statistics characterizing 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, e.g., Bandettini et al., 1993; Baker et al., 1994; Worsley and Friston, 1995; Xiong et al., 1996; and 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, e.g., concerning the stimulus (binary baseline-activation conditions), the haemodynamic response (modeled 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],

Contract grant sponsors: European Union BIOMED II, the Danish Research Councils through the Danish Computational Network (CONNECT) and the THOR Centre for Neuroinformatics; Contract grant number: BMH4-CT97-2775 (BIOMED II).

*Correspondence to: Cyril Goutte, INRIA Rhône-Alpes, ZIRST-655 Avenue de l'Europe, Montbonnot, 38334 Saint Ismier Cedex, France. E-mail: cyril.goutte@inrialpes.fr

Received 13 November 2000; Accepted 13 March 2001

EDA in neuroimaging has been mostly carried out using clustering methods. The most popular method is certainly fuzzy c-means [Baumgartner et al., 1997, 1998; Moser et al., 1997; Golay et al., 1998; Moser 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 have hierarchical approaches to clustering [Goutte et al., 1999; Filzmoser et al., 1999]. Also, using hierarchical arguments together with a nonparametric approach, innovative methods are starting to appear [e.g., Wismüller et al., 1998; Domany, 1999; Goutte, unpublished research].

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, e.g., the stimulus or the underlying haemodynamic response. Clustering on the raw time series is potentially able to isolate cognitive or haemodynamic effects without precisely modeling 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 generalizes 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. The goal of this paper is to illustrate to the reader the potential of the feature-based clustering method through the development of a procedure that highlights the differences obtained on an example data set.

At first, we focus on two simple features, namely the delay and strength of activation 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” modeling toolbox [Hansen et al., 2000]. 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 data set 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 data set was acquired at Hvidovre Hospital on a 1.5 T Magnetom Vision MR scanner. The scanning sequence was a 2D gradient echo EPI (T_2^* weighted) with 66 ms echo time and 50 degrees RF flip angle. The images were acquired with a matrix of 128×128 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 68×82 2D voxel map. The voxel dimension is $1.8 \times 1.8 \times 10$ mm.

The visual paradigm consists of a rest period of 20 sec of darkness using a light fixation dot, followed by 10 sec of full-field checkerboard reversing at 8 Hz, and ending with 20 sec of rest (darkness). In total, 150 images were acquired in 50 sec, 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 data set studied in this article was built by combining these into a single sequence of 1,210 images. However, as the runs were acquired separately, it should be noted that there cannot be any causality between the activation in one run and the signal measured in the next. Note also that because of the haemodynamic delay, the signal measured in activated voxels will be roughly centered within the remaining 40 sec 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

boxcar signal for simple on/off excitation. The complete time series in (x, y, z) is a vector $f(x, y, z) \in \mathbb{R}^T$ 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, 1998; Moser et al., 1997; 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 (i.e., a notion of distance) in the input (i.e., time series) space. However, it is well known that when the dimensionality of the space increases, the notion of distance becomes counterintuitive. This is easily seen by considering what happens when points are randomly distributed uniformly in a multidimensional 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 1.2–1.4].

In a traditional fMRI time series clustering setting, the input space is the time series space, whose dimension is equal to the number of images. As a consequence, for experiments involving large numbers of images, clustering using a standard Euclidean distance is carried out in a high-dimensional space and 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 then corresponds to the number of cross-correlation coefficients retained and will still be relatively high ($T = 50$) [see Goutte et al., 1999]. In addition, the inputs are correlated, and the relatively high dimensionality of the input space 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 the dimensionality of the data to cluster 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 preprocess the data using a run-based detrending method. For each run of 121 images, we estimate the linear trend in the data based on the first 31 images and the last 31 images, which we expect have no activation, 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 centered in the acquisition interval. In particular, delayed activations would 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 six images (or 2 sec) off center.

The rationale for the choice of the images we use to estimate the trend is that the first 31 images are recorded before activation occurs and therefore cannot be activated. The last 31 images are recorded at least 10 sec after stimulus cessation, where most of the activation is expected to have been reduced [see, e.g., 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 5–7) 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 zero mean. While the mean of the nonactivated measurements is roughly zero, the presence of activation in a voxel will usually result in a nonzero mean value (Fig. 2). This contrasts with the usual detrending using a global trend, i.e., including the activated images, where the mean is by construction zero (Fig. 2, right plot).

There is a possibility of an artifactual bias in the detrended time series because of the presence of a long postactivation 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 artifactual positive trend. Note that this effect is minor compared to the amplitude of the correction brought by the run-based detrending preprocessing (cf. Fig. 1). Furthermore, it should be noted that the goal of clustering is to identify similar patterns of activations. Although the trends may all be different, similar patterns of activation will have similar (if any) biases, such that this artifact will not interfere with the clustering and the analysis results.

Data reduction

In usual brain imaging experiments, a large portion of the brain shows no stimulus-related activation. 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 $P = 0.01$. To take into account the high temporal resolution of our data and get a meaningful F-statistic, the paradigm is first shifted 7 sec.

The goal of this sieving is merely 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

Figure 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 nonactivated 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.

Figure 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 centering 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 color bar applies to both plots.

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, i.e., where there is a possible activation.

In the context of this study, the use of F-test sieving is seen merely as a data-reduction device. Other techniques could be used with similar effect, e.g., based on the maximum cross-correlation as exemplified by Goutte et al. [1999]. Nevertheless, the reliance of the F test on a prespecified paradigm may be questioned. To answer this concern, note that the use of a simple paradigm makes sense in the context of this experiment, which involves a strong visual activation. It is possible to adapt the data-reduction procedure to more complex experiments by involving multiple sieving operations, for example, nested test involving different paradigms (one for preparation, one for execution, etc.). Note also that sieving may be performed by methods that are less dependent on assumptions about the paradigm (involving, e.g., the mean or variance of the detrended signal).

Finally, it is always possible to a posteriori salvage discarded voxels by assigning them to the clusters resulting from the analysis. This limits the risk of the data reduction procedure discarding useful information and could in principle “recycle” voxels. However, in the analyses presented here we found no use for this particular postprocessing.

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:

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

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, e.g., by Bandettini et al. [1993]. In particular it is neither centered, nor normalized, though in practice we usually normalize the paradigm $p(t)$ (i.e., 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 two 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, and (2) The *delay* of activation is the delay δ for which the cross-correlation function reaches its extremum.

Because of the periodicity in the data, we will only look for the maximum between delays -60 and 60. For

Figure 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 autocorrelation 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.

Figure 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).

noisy and discrete data, the cross-correlation signal is itself noisy and discrete. As a consequence, the 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 (Fig. 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 2D data vector $u(x, y, z)$.

Feature extraction for metaclustering

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 techniques:

- standard Student t test (between rest and activation)
- standard Kolmogorov-Smirnov test [Baker et al., 1994]
- correlation with the paradigm, delayed by 7 sec [Bandettini et al., 1993]
- finite impulse response (FIR) filter model, fitted on the fMRI signal [Nielsen et al., 1997]

- Gamma filter model, fitted on the fMRI signal [Lange and Zeger, 1997]

All of those are voxel-based analyses [Hansen et al., 2000] that cover a broad range of techniques actually in use by neuroimaging practitioners. As they are based on different assumptions, they 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. Accordingly, we will concentrate on a seven-dimensional feature space, containing the following seven features: (1) Student t statistic; (2) Kolmogorov-Smirnov statistic d ; (3) correlation coefficient r between the signal and the shifted paradigm; (4) standard deviation of the fitted signal modeled by a FIR filter; (5) delay estimated from the FIR filter; (6) strength parameter from the Lange-Zeger filter; and (7) delay estimated from the Lange-Zeger filter.

Note that features 1–4 and 6 are used to somehow characterize the strength of the fMRI response, while features 5 and 7 estimate the delay in activation. Not only can the different measures of activation strength usually presented on individual activation maps join additional similar features in a single analysis, but they can be combined with features that are harder to present visually, such as the delays, leading in effect to a spatio-temporal analysis [Goutte et al., 1998].

Clustering

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

1. Initialize K clusters C_k , $k = 1 \dots K$, with centers $c_k^{(i)}$, for iteration $i = 0$.

2. Assign each data vector u_j to the cluster C_k with the nearest center $c_k^{(i)}$, based on a distance metric between the cluster center and the data vector, $d(u_j, c_k^{(i)})$.
3. Set new cluster center $c_k^{(i+1)}$ to the average of its members: $c_k^{(i+1)} = 1/|C_k| \sum_{j \in C_k} u_j$

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 $d(u_j, c_k^{(i)})$ is discussed in detail by Goutte et al. [1999].

Initialization

There are usually several stable partitions, depending on the initial configuration. Of the different possible initialization strategies available, we simply use a random subset of the data as initial centers. To limit the effect of the stochastic initialization, we will usually perform several clusterings using different random initializations, 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 initialization procedure is in turn a stochastic process, as the minimum will depend on the particular set of initial centers 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 the Appendix, 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 Completed Likelihood (ICL) of Biernacki et al. [2000]. Using this framework, we will select the number of clusters that yield the highest value for the information criterion. The three criteria given in the Appendix will usually lead to different clustering models. In

particular, AIC is known to overestimate the number of clusters. We will illustrate and discuss this below for our two clustering problems.

Normalization

In some cases, the features, i.e., the Q dimensions of the u_j 's, have very different ranges. In that case, it is advisable to normalize the data to avoid one feature dominating 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.

Normalization typically involves centering by subtracting the mean and dividing by the standard deviation.

Functional MRI data are characterized by a usually limited number of activated voxels and fairly high activation statistics. In order to robustify the centering, we will use the median, and we scale the centered data such that 99% of the sample is placed in the interval $-2.57; 2.57$ (i.e., the 99% confidence interval for a standard normal variable). This takes into account 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.

For easy reference, we will always display the results in the original (unnormalized) domain. This allows the reader to relate to the actual meaning of the features.

RESULTS

Preprocessing

The data set acquired as a sequence of 128×128 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 3,891 voxels, and each time series has 1,210 elements. The run-based detrending is performed according to the description given above.

For 1,210 images, the value of the F statistic corresponding to the 1% threshold is $F_{.99} = 6.67$. To get a meaningful F statistic, the paradigm is first shifted 7 sec to take into account the delay of the haemodynamic response. In the current analysis, 889 voxels out of 3,891 have an F statistic greater than this threshold. They include the primary visual area and lateral areas in the visual cortex, in particular the lateral occipital cortex (Fig. 5).

Figure 5.

Sieving nonactivated 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.

Clustering on delay and strength

The cross-correlation and feature extraction produce 889 points in the 2D 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 sec, 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 because of the large number of 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 normalization) for an increasing number of clusters and calculate the values of the three information criteria mentioned above. The model is considered optimal with respect to a given criterion when this criterion reaches its maximum. To estimate the variation caused by the stochastic initialization of the K-means algorithm (cf. Methods section), we derive empirical error-bars on the criteria by replicating the multiple initialization 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 four clusters, BIC is maximum for eight clusters, and AIC chooses 18 clusters. This illustrates nicely a key difference between the three crite-

Figure 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 two standard deviations. 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. Eq. 8).

ria. The BIC is known to asymptotically estimate the “true” model structure [Schwartz, 1978]. On the other hand, AIC is targeted toward the minimization of the generalization error, and tends to overestimate the model size, as exemplified by the expressions of the penalty terms in Equations (6) and (7) in the appendix. Note, however, that the model studied here (Eq. 4 in Appendix) 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 modeling 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., 2000; Goutte, unpublished research].

We will first investigate the most parsimonious model, chosen by ICL, using four clusters. Figure 7 displays the four 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 center. 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 (○). 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 (Fig. 8) suggests that this is not the case here and that this cluster reflects a large, nonartifactual positive activation.

Figure 8 shows the spatial location of the four clusters in the slice. The first cluster (+ in Fig. 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 artifacts from the F filtering. The negative activation in cluster 4 seems to be located at the periphery of the primary visual area, but there are a certain amount of noisy voxels, corresponding to the data at the extreme left of Figure 7. Notice that the area

Figure 7.

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

Figure 8.

The four clusters obtained from K-means for the optimal ICL (cf. Fig. 7), on top of the anatomical background. From left to right, the corresponding markers in Figure 7 are +, ×, ·, and ○.

delimited by cluster 2 (primary and part of the supplementary visual 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 asymmetry 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 runs 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 two standard deviations are indicated in dashed lines (“model error bar”), and the error bars taking into account the noise variance (“data error bars”) are 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 overshoot of similar *relative* amplitude is also observed in cluster 2, though it fails to display any significance,

because of the smaller *absolute* amplitude. Note that in cluster 3 also, the average activation seems to rise significantly above zero (roughly between images 60 and 80). However, the amplitude of the signal is extremely small, and this can be attributed to an artifact 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 sec for clusters 1, 2, and 4, respectively. The delay from stimulus cessation to return below 10% of the maximum activation is 6.8, 6.4, and 6.5 sec, respectively.

Using the structure that maximizes BIC ($K = \text{eight}$ clusters), the results are only slightly different. Figure 10 shows that the main differences are more noisy clusters, and positive activation spread differently. The two most positively activated clusters (indicated by ‘+’ and ‘×’ on Fig. 10) overlap almost exactly with cluster 1 from Figure 8, while the third (diamonds on Fig. 10) contains mostly the positive activation part of a cluster that was earlier composed of both activation and noise (‘×’ on Fig. 8). Accordingly, out of the eight resulting clusters, we will investigate only the four that display some kind of activation (three with positive activation and one with negative). The cluster containing the negative activation (cluster 8, ‘○’ in Fig. 10) again contains voxels located at the periphery of the primary visual area (Fig. 11). The first three clusters in Figure 11 contain voxels with decreasing positive activation. Note that cluster 3 (diamonds in Fig. 10) shows lateral activation in two smaller areas that could correspond to V5.

Figure 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). [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

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 significance of the activation appears quite clearly in all plots. The three clusters with positive response (cluster 1–3 on Figs. 11 and 12) display a graded pattern of positive activation. The postactivation 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 because the clusters contain less voxels than above, hence larger variance and error bars.

The clusters resulting from AIC will not be investigated as we have seen theoretically and empirically that AIC overestimates the right amount of clusters.

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 approx-

imated by Gaussian distributions. The delays for the negative activation clearly seems to be 1–1.5 sec shorter on average than the delays measured for the positive activation [Buxton et al., 1998]. We assess the significance of this difference using a nonparametric 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).

Meta clustering

The seven features are extracted using the “Lyngby” modeling toolbox [Hansen et al., 1999]. As the features have very different ranges (e.g., between 0 and 0.5 for the correlation, and between -25 and 55 for the Gamma filter strength), we normalize the data as explained in the Methods section before clustering. All

Figure 10.

The 851 voxels (after F-test sieving and discarding incorrect delays) described by their delay and strength, partitioned in eight clusters identified by K-means (maximum of BIC). Four clusters contain noisy voxels (·): one (○) contains negatively activated voxels as before; three clusters (×, +, and diamond) contain positively activated voxels, with decreasing levels of activation.

the results presented below will be shown in the original, nonnormalized domain, to relate to their actual meaning (e.g., delays in images, t statistics, etc).

On the 889 resulting points in a seven-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 (Fig. 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 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 centers. 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 center, i.e., the first dimension in the feature space. For example, the cluster with largest t is labeled 1, and the cluster with the most negative t is labeled 26 (Fig. 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 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 nonactivated 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 six clusters contain the most positive activation. It is already clear from the plots 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

Figure 11.

Four of the eight clusters obtained from K-means for the optimal BIC (cf. Fig. 10), on top of the anatomical background. From left to right, the corresponding markers are +, ×, ◇, and ○. Notice the similarity between cluster 1 in Figure 8 and the union of clusters 1 and 2 here.

Figure 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 runs; 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. [Color figure can be viewed in the online issue, which is available at www.interscience.com.]

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 behavior, according to the FIR standard deviation or Gamma filter strength, is not reflected in the Kolmogorov-Smirnov statistic or the t statistics, for which clusters 1 and 2 are nearly identical (upper left plot of Fig. 15). 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 Fig. 15).

Ordering of clusters

It is interesting to analyze 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 I illustrates this by showing the order of the different clusters along the five activation-based features, i.e., 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 (first two

Figure 13.

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

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 1–4, as well as clusters 11 and 26. In addition, sizable portions of what appears to be area V5 are contained in clusters 7–10. 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 localized in the neighborhood of each other. A notable exception is cluster 11, which seems to be distributed over a fairly large area, with no clear grouped localization. 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 artifactual voxels. Interestingly, cluster 2 is almost identical to cluster 1 in the eight-cluster analysis above (maximizing 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 I). 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 four clusters that are ordered differently according to Table I. Clusters 2 and 3 (right plots in Fig. 17) show very similar patterns, with different activation levels. This is not surprising as their ordering seem to be very consistent along all features (cf. Fig. 15 and Table I). 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 (almost as soon as the stimulus onset) than in the remaining clusters, and it reaches a maximum early, around images 50–55. This does undoubtedly favor 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.

Conversely, cluster 11 displays a much-delayed and softer rise in signal (as late as image 40). Features obtained from filters that model the activation delay properly do correctly place cluster 11 on a 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. Conversely, the large delay, placing more than half the activation after stimulus cessation, badly damages the results of the t statistic and Kolmogorov-Smirnov statistics, which rely heavily on a square wave design.

Figure 14.

The three information criteria, calculated on the seven 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 penalizes large numbers of clusters.

Figure 15.

Results of the metaclustering. Dots indicate data point, numbers show cluster centers. 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). [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

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 num-

TABLE I. Ordering of clusters along the five activation-based features, from largest value (top) to lowest (bottom)*

<i>t</i> statistic	Kolmogorov-Smirnov statistic	Correlation at 7 sec	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

* The *t* statistic is chosen as reference. Notice the markedly different orderings, especially for cluster 1 and cluster 11 in the last three columns. The high rank of cluster 26 in columns 2 and 4 is because these features are always positive.

Figure 16.

Spatial location of the voxels from six clusters obtained from meta-clustering ($K = 26$, maximum of the ICL), cf. Figure 15.

ber of clusters, the picture offered by our metaclustering scheme is very different, as shown on Figure 18. Cluster 26 has a much faster onset response, which seems consistent with our findings in the previous analysis [see also Goutte et al., 1998]. The voxels from cluster 1 react 1–2 sec faster than voxels in clusters 2–4. 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 cessation of the stimulus.

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

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 (Fig. 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 five 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, e.g., spotting voxels that react strongly according to some analyses, while being fairly non-descript 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.

Although the features exemplified above require a predefined paradigm and are therefore dependent on assumptions on the activation pattern, it should be noted that feature-space clustering generalizes rather than restricts the use of clustering for analyzing fMRI experiments. Indeed, the traditional use of clustering is trivially obtained by using the original data as “features.” Other features with weak hypotheses could come, for example, from a wavelet transform of the time series. In general, there is a tradeoff between obtaining assumption-free results and working with rich features, usually requiring modeling assumptions. This simply reflects the commonsense idea that it is easier to find something when you know what you are looking for.

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 ISO-DATA [Tou and Gonzalez, 1974, sec. 3.3.6]. This approach can create problems if, e.g., 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 modeled the K-means algorithms as a particular instance of learning a mixture model (Appendix). 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 favor large numbers of clusters in high-density areas. In the context of fMRI data, this

Figure 17.

Response in four clusters obtained with metaclustering on seven features ($K = 26$, maximum of the ICL). Notice the markedly different responses in clusters 1 and 11 (right) compared to the more “typical” response in clusters 2 and 3 (left). In particular, the signal in cluster 1 rises almost from stimulus onset and reaches a

maximum as soon as images 50–55. In contrast, the other clusters and especially cluster 11 have a more delayed rise (starting from image 40 for cluster 11) and are slower to reach their maximum (around image 70). [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

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, and (2) it forces the algorithm to consider each dimension in space equally. This second point can be partially addressed by a normalization procedure, or using an appropriate metric [Goutte et al., 1999].

More flexible mixture models can be considered. For example, Biernacki et al. [2000] 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 Maximization (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. How-

ever, the description and use of these general model is well beyond the scope of this paper.

Average activation and significance

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 (Fig. 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. Although the error bars on the estimate provide a potentially useful way of assessing the significance of a hypothesized activation, we have refrained from putting too much emphasis on this aspect. Indeed, by construction, the clustering process provides voxels that are expected to have homoge-

Figure 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 six clusters obtained with metaclustering on seven 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.

neous 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 voxels 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, e.g., parametric models of the haemodynamic response. Note also that the experiment analyzed here was performed on several subjects, as well as several times on the same subject. This provides an opportunity 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 (e.g., Menon et al. [1995]; Hu et al. [1997]; Yacoub et al. [1999]; Hathout et al. [1999]) and has been observed in several situations, 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 favorable 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 maximizes an information criterion, over multiple stochastic initial configurations. Using this approach, we obtain a total “plug-in” analysis, reducing the dependency upon the initial conditions 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 show 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.

ACKNOWLEDGMENTS

We thank Carl Edward Rasmussen for help using his Gaussian Process smoothing software, and the anonymous reviewers for constructive criticism.

REFERENCES

- Akaike H (1974): A new look at the statistical model identification. *IEEE Trans Automatic Control* 19:716–723.
- Ashburner J, Haslam J, Taylor C, Cunningham VJ, Jones T (1996): A cluster analysis approach for the characterization of dynamic PET data. In: Myers R, Cunningham V, Bailey D, Jones T, editors. *Quantification of brain function using PET*. San Diego: Academic Press. p 301–306.
- Baker J, Weisskoff R, Stem C, Kennedy D, Jiang A, Kwong K, Kolodny L, Davis T, Boxerman J, Buchbinder B, Wedeen V, Belliveau J, Rosen B (1994): Statistical assessment of functional MRI signal change. In: *Proceedings of the Second Annual Meeting of the Society of Magnetic Resonance in Medicine*. p 626.
- Bandettini PA, Jesmanowicz A, Wong EC, Hyde JS (1993): Processing strategies for time-course data sets in functional MRI of the human brain. *Magn Reson Med* 30:161–173.
- Baumgartner R, Scarth G, Teichtmeister C, Somorjai R, Moser E (1997): Fuzzy clustering of gradient-echo functional MRI in the human visual cortex. Part I: Reproducibility. *J Magn Reson Imaging* 7:1094–1101.
- Baumgartner R, Windischberger C, Moser E (1998): Quantification in functional magnetic resonance imaging: Fuzzy clustering vs. correlation analysis. *Magn Reson Imaging* 16:115–125.
- Benali H, Di Paola M, Burnod Y, Péligrini M, Buvat I, Garnero L, Lehericy S, Di Paola R (1999): A multivariate model for estimating non stationary hemodynamic response functions in event-related fMRI. In: Rosen et al. (1999). p S5.
- Biernacki C, Celeux G, Govaert G (2000): Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE Trans Pattern Analysis Machine Intelligence* 22:719–725.
- Bishop CM (1995): *Neural networks for pattern recognition*. Oxford: Clarendon Press.
- Bottou L, Bengio Y (1995): Convergence properties of the K-means algorithm. In: Tesauro G, Touretzky TS, Leen TK, editors. *Advances in neural information processing systems*. Vol. 7. Cambridge: MIT Press.
- Buxton RB, Wong EC, Frank LR (1998): Dynamics of blood flow and oxygenation changes during brain activation: the balloon model. *Magn Reson Med* 39:855–864.
- Celeux G, Govaert G (1992): A classification EM algorithm for clustering and two stochastic versions. *Comput Stat Data Analysis* 14:315–332.
- Ding X, Tkach J, Ruggieri P, 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*. p 630.
- Domany E (1999): Superparamagnetic clustering of data—the definitive solution of an ill-posed problem. *Physica A* 263:158–169.
- Filzmoser P, Baumgartner R, Moser E (1999): A hierarchical clustering method for analyzing functional MR images. *Magn Reson Imaging* 17:817–826.
- Friberg L, Gjedde A, Holm S, Lassen NA, Nowak M, editors (1997): *Proceedings of the Third International Conference on Functional Mapping of the Human Brain*. *Neuroimage* 5(4):Part 2 of 4.
- Golay X, Kollias S, Stoll G, Meier D, Valavanis A, Boesiger P (1998): A new correlation-based fuzzy logic clustering algorithm for fMRI. *Magn Reson Med* 40:249–260.
- Goutte C, Nielsen FÅ, Svarer C, Rostrup E, Hansen LK (1998): Space-time analysis of fMRI by feature space clustering. In: Paus T, Gjedde A, Evans A, editors. *Proceedings of the Fourth International Conference on Functional Mapping of the Human Brain*, published in *Neuroimage* 7:S610.
- Goutte C, Toft P, Rostrup E, Nielsen FÅ, Hansen LK (1999): On clustering fMRI time series. *Neuroimage* 9:298–310.
- Hansen LK, Nielsen FÅ, Liptrot MG, Goutte C, Strother SC, Lange N, Gade A, Rottenberg DA, Paulson OB (2000): *Lyngby 2.0—a modeler's Matlab toolbox for spatio-temporal analysis of functional neuroimages*. In: Fox PT, Lancaster JL, editors. *Proceedings of the Sixth International Conference on Functional Mapping of the Human Brain*, published in *Neuroimage* 11:S917.
- Hathout GM, Varjavand B, Gopi RK (1999): The early response in fMRI: a modeling approach. *Magn Reson Med* 41:550–554.
- Holmes AP, Friston KJ (1997): Statistical models and experimental design. SPM course notes, chapter 3.
- Hu X, Le TH, Ugurbil K (1997): Evaluation of the early response in fMRI in individual subjects using short stimulus duration. *Magn Reson Med* 37:877–884.
- Lange N, Strother SC, Anderson JR, Nielsen FÅ, Holmes A, Kolenda T, Savoy R, Hansen L (1999): Plurality and resemblance in fMRI data analysis. *Neuroimage* 10:282–303.
- Lange N, Zeger SL (1997): Non-linear Fourier time series analysis for human brain mapping by functional magnetic resonance imaging. *J R Stat Soc C App Stat* 46:1–30.
- Mckeown MJ, Makeig S, Brown GG, Jung T-P, Kindermann SS, Bell AJ, Sejnowski TJ (1998): Analysis of fMRI data by blind separation into independent spatial components. *Hum Brain Mapp* 6:160–188.
- Menon RS, Ogawa S, Hu X, Strupp JP, Anderson P, 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. *Magn Reson Med* 33:453–459.
- Moser E, Baumgartner R, Barth M, Windischberger C (1999): Explorative signal processing in functional MR imaging. *Int J Imaging Systems Tech* 10:166–176.
- Moser E, Diemling M, Baumgartner R (1997): Fuzzy clustering of gradient-echo functional MRI in the human visual cortex. Part II: quantification. *J Magn Reson Imaging* 7:1102–1108.
- Nielsen FÅ, Hansen LK, Toft P, Goutte C, Lange N, Strother SC, Mørch N, Svarer C, Savoy R, Rosen B, Rostrup E, Barr P (1997): Comparison of two convolution models for fMRI time series. In: Friberg et al. (1997). p S473.
- Purushotham A, Nielsen FÅ, Hansen LK, 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). p S51.
- Raichle M (1999): Closing lecture at the Fifth International Conference on Functional Mapping of the Human Brain. *Düsseldorf*, June 22–26, 1999.
- Ripley BD (1996): *Pattern recognition and neural networks*. Cambridge: Cambridge University Press.
- Rosen BR, Seitz R, Volkmann J, editors (1999): *Proceedings of the Fifth International Conference on Functional Mapping of the Human Brain*. *Neuroimage* 9(6):Part 2 of 2.
- Schwartz G (1978): Estimating the dimension of a model. *Ann Stat* 6:461–464.
- Sychra JJ, Bandettini PA, Bhattacharya N, Lin Q (1994): Synthetic images by subspace transforms. I. Principal components images and related filters. *Med Phys* 21:193–201.

- Toft P, Hansen LK, Nielsen FÅ, Goutte C, Strother S, Lange N, Mørch N, Svarer C, Paulson OB, Savoy R, Rosen B, Rostrup E, Born P (1997): On clustering of fMRI time series. In: Friberg et al. (1997). p S456.
- Tou JT, Gonzalez RC (1974): Pattern recognition principles. In: Applied mathematics and computation. No. 7. Reading, MA: Addison-Wesley.
- Wand M, Jones M. (1995): Kernel smoothing. Monographs on statistics and applied probability. No. 60. London: Chapman & Hall.
- Williams CKI (1998): Prediction with Gaussian processes: from linear regression to linear prediction and beyond. In: Jordan MI, editor. Learning and inference in graphical models. Cambridge, MA: MIT Press.
- Wismüller A, Dersch DR, Lipinski B, Hahn K, 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, Ziemke T, editors. Proceedings of the eighth international conference on artificial neural networks. Perspectives in neural computing. London: Springer. p 857–862.
- Worsley K, Friston K (1995): Analysis of fMRI time-series revisited—again. *Neuroimage* 2:173–181.
- Xiong J, Gao J-H, Lancaster JL, Fox PT (1996): Assessment and optimization of functional MRI analyses. *Hum Brain Mapp* 4:153–167.
- Yacoub E, Hu X (1999): Detection of the early negative response in fMRI at 1.5 tesla. *Magn Reson Med* 41:1088–1092.
- Yacoub E, Le TH, Ugurbil K, Hu X (1999): Further evaluation of the initial negative response in functional magnetic resonance imaging. *Magn Reson Med* 41:436–441.

APPENDIX

Information criteria for k-means

From a statistical point of view, the K-means algorithm is a particular instance of the Classification Expectation Maximization 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 $u_j \in \mathbb{R}^Q$, is expressed as:

$$P(u_j|M, \sigma^2) = \frac{1}{K} \sum_{k=1}^K \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2} \frac{(u_j - \mu_k)^2}{\sigma^2}\right] \quad (2)$$

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

$$\sigma^2 = \frac{1}{N} \sum_{k=1}^K \sum_{j \in C_k} \|u_j - c_k\|^2 \quad (3)$$

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

$$P(\mathcal{D}|M, \sigma^2) = \prod_j P(u_j|M, \sigma^2) \quad (4)$$

Alternatively, the classification likelihood is calculated by assigning each point u_j to the mixture component k_j that has the highest probability:

$$u_j \sim \mathcal{N}(u_{k_j}, \sigma^2)$$

$$P(u_j|M, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2} \frac{(u_j - \mu_{k_j})^2}{\sigma^2}\right] \quad (5)$$

where k_j is the cluster to which u_j is assigned. The joint classification likelihood for the entire data set is then calculated like in (4).

Within this probabilistic framework, we can see the K-means algorithm as an attempt at maximizing 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 to estimate the optimal number of clusters:

- Akaike's first information criterion [Akaike, 1974]:

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

- Schwartz' Bayesian information criterion [Schwartz, 1978]:

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

- The integrated completed likelihood [Biernacki et al., 2000]:

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

where in the above criteria, Q is the dimension (i.e., 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.