Geostatistical analysis in clustering fMRI time series

Jun Ye1,*,†, Nicole A. Lazar2 and Yehua Li2

1Department of Medicine, Massachusetts General Hospital, Boston, MA 02114, U.S.A.
2Department of Statistics, The University of Georgia, Athens, GA 30602, U.S.A.

SUMMARY
Clustering of functional magnetic resonance imaging (fMRI) time series—either directly or through characteristic features such as the cross-correlation with the experimental protocol signal—has been extensively used for the identification of active regions in the brain. Both approaches have drawbacks; clustering of the time series themselves may identify voxels with similar temporal behavior that is unrelated to the stimulus, whereas cross-correlation requires knowledge of the stimulus presentation protocol. In this paper we propose the use of autocorrelation structure instead—an idea borrowed from geostatistics; this approach does not suffer from the deficits associated with previous clustering methods. We first formalize the traditional classification methods as three steps: feature extraction, choice of classification metric and choice of classification algorithm. The use of different characteristics to effect the clustering (cross-correlation, autocorrelation, and so forth) relates to the first of these three steps. We then demonstrate the efficacy of autocorrelation clustering on a simple visual task and on resting data. A byproduct of our analysis is the finding that masking prior to clustering, as is commonly done, may degrade the quality of the discovered clusters, and we offer an explanation for this phenomenon. Copyright © 2009 John Wiley & Sons, Ltd.

KEY WORDS: autocorrelation; masking; marginal effect; silhouette values; structural classification

1. INTRODUCTION
The brain is the most interesting but least understood organ in the human body. Even with the rapid development of neuroimaging techniques in recent years, many problems still have not been solved. In particular, the detection of active brain areas remains a challenging problem. Functional magnetic resonance imaging (fMRI) is a relatively new non-invasive technique for studying the workings of the active human brain. During an fMRI experiment, a sequence of brain images is acquired while the subject performs specific cognitive tasks. Changes in the measured signal are

*Correspondence to: Jun Ye, Department of Medicine, Massachusetts General Hospital, Boston, MA 02114, U.S.A.
†E-mail: athensye@yahoo.com

Contract/grant sponsor: NSF; contract/grant number: DMS-0706192

Received 19 June 2008
Accepted 16 April 2009
used to identify and characterize the brain activity that results from the task performance. In recent years, characterizing brain activation by clustering the fMRI time series has gained popularity [1]. Since regions that react to the experimental task may be at several different physical locations, the goal in clustering time series is to partition the brain into clusters, where voxels within each cluster have similar temporal patterns. The underlying assumption is that voxels with similar temporal characteristics belong to the same functional regions of the brain [2, 3].

Owing to the high noise level in the fMRI data, the results of clustering on the raw time series are often unsatisfactory and do not necessarily group data according to the similarity of their pattern of response to the stimulus [1]. Hence, many clustering techniques have focused instead on cross-correlation functions of the fMRI time series. Bandettini et al. [4] considered the correlation between the data time series and a reference function from a ‘seed’ voxel to characterize the temporal response of the brain to the paradigm. Goutte et al. [1, 5] suggested clustering voxels on the basis of the cross-correlation function between the fMRI activation and the experimental protocol signal. (Note that, strictly speaking this is not a correlation in the statistical sense, since the experimental design is fixed, but this is the standard nomenclature in the fMRI literature [6].) Yeo and Ou [3] devised a ‘signal to protocol metric’ by combining the cross-correlation of two fMRI signals and the Euclidean distance between voxels to do the clustering. Friman et al. [7] approached the problem by viewing the correlation between two time series as an objective function to be maximized with respect to some parameters. Clustering on the cross-correlation function instead of the raw time series increases robustness and yields improved performance. However, methods based on cross-correlation depend heavily on the choice of reference function. Note that different types of reference functions are possible, and they result in different analyses [8].

In this paper, a special classification method for fMRI data is introduced. We first summarize the traditional classification methods and formalize them as three steps: feature extraction, choice of classification metric and choice of classification algorithm. In the feature extraction step, we propose the use of the autocorrelation function as the main feature and demonstrate its efficacy. In geostatistics, use of autocovariance (autocorrelation) or variogram to characterize the spatial or temporal structure of the data is called structural analysis. When the autocovariance (autocorrelation) function or, equivalently, the variogram function is used in classification, this is called structural (or textural) classification in geostatistics. This classification method has been widely used in geostatistical areas, e.g. ecology and remote sensing [9–11], but, to date, not in fMRI. The analysis treats a set of data as a sample from the realization of a random process, and stresses the structural features. The correlation structures for different functional regions of the brain are different, and can be used to inform classification. For example, in functional regions that react to the task, the time series generally exhibits idiosyncratic patterns (e.g. periodicity) and the autocorrelation function may be expected to fluctuate accordingly [12–14]. For irrelevant regions the autocorrelation is expected to be flat.

Since auto-correlation does not count the order of the raw data matrix over time, two time series having identical auto-correlation structure does not imply that they are coherent or synchronous. This is one advantage of using autocorrelation in clustering instead of raw time series, because the autocorrelation functions automatically take care of the time delays in different time series. We know that the active voxels in the brain communicate to each other when responding to the stimulus, i.e. some voxels might have delayed reaction to the stimulus; some voxels might have early reaction due to experimenter’s anticipation of the stimulus. The autocorrelation function gets rid of the time shifting, prevents us from declaring too many clusters and in general provides more stable clustering results.

Carew et al. [15] describe two conceptual approaches to the modeling of fMRI time series: the general linear model [16, 17] and spline smoothing. In the general linear model approach, the autoregressive function AR(\(p\)) is used for modeling the error term. As indicated by Carew et al. [15], this approach might introduce a bias. In the spline smoothing approach, all the data structures are considered in a spline function. But in spline smoothing approach, the covariance structure of the spline function is fixed, and therefore it does not take account of the ‘true’ temporal structure of the time series from a geostatistical point of view [18–21]. Our work can be considered as a third conceptual rail, which considers the correlation structure of the time series directly. Also, the two previous approaches are termed ‘univariate’ in the fMRI literature. That is, analysis is carried out on an individual voxel basis (for instance, the general linear model is fit to each voxel separately). Our approach inherently falls into the ‘multivariate’ line of fMRI data analysis, since it involves clustering and searching for commonalities in temporal patterns of behavior across voxels.

The rest of the paper is organized as follows: In Section 2, we introduce basic concepts and definitions of our clustering method. In Section 3, we describe the fMRI data sets used to test our methods; application of our method to these fMRI data sets is described in Section 4. Section 5 discusses the proposed method.

2. METHODS

2.1. Clustering analysis

2.1.1. Three steps in clustering. Classification is a procedure in which individual objects are placed into different classes based on their characteristics. When neither the number of the classes nor the classes themselves are known in advance, this is known as ‘cluster analysis’ [22]. In any clustering problem, a good solution depends on three steps: feature extraction, the choice of the clustering metric, and the clustering algorithm [3, 7, 23].

Feature extraction step: Feature extraction is a special form of data reduction, that decreases the resources required to describe a large set of data accurately. In geostatistics, structural clustering uses the autocorrelation or variogram as a feature, thereby incorporating spatial (or temporal) information [21]. Several recent studies from ecology have revealed relations between periodic landscape patterns and periodicity of variograms [12, 14], highlighting that the variogram can be used to detect and describe complicated spatial structures. In geostatistical applications, it is common to work with the variogram, rather than the autocorrelation function [24]. The advantage of the former is that it only compares the average square difference between locations; this is quite useful in geology since one usually doesn’t know the population mean and hence it is assumed to be constant. But it may be a disadvantage in other settings since the variogram does not account explicitly for mean and variance changes. Specifically, in fMRI data, the mean or variance may change over time. Therefore, we choose to work with the autocorrelation instead of the variogram because of the former’s robustness in the case that local means and variances might vary [12, 25].

Definitions: Consider a time series \(\{Z(t) : t \in D\}\), where \(t\) denotes time; \(Z(t)\) denotes the random variable of interest at time \(t\); \(D\) denotes the set of the time points of interest \(t_1, t_2, \ldots, t_N\). We define the lag \(h\) as the separation between two different time points. The time series is second-order stationary if \(E[Z(t_i)] = E[Z(t_i + h)] = \mu\) and \(\text{Cov}[Z(t_i + h), Z(t_j)] = C(h)\), for any \(t_i + h, t_j \in D\), where \(\mu\) is a constant and \(C(\cdot)\) is a positive semi-definite function depending only on \(h\). The autocorrelation function is defined as \(\rho(h) = C(h)/C(0)\).
As mentioned before, different functional regions in the brain have different temporal behavior during the experiment, and their correlation structures are potentially also different. As a result, we can classify voxels by comparing their correlation structures. The reason to use correlation instead of covariance in clustering is that without the standardization, clustering will extract regions with large variances rather than similar temporal patterns [1].

There are two main approaches put forth in the geostatistical literature for estimating the correlation function from the data [9]. The first approach is based on the nonparametric or empirical covariance estimator described below; the second approach is parametric, i.e. fit a pre-existing covariance model to the empirical autocovariance and obtain the model parameters. Usually the modeling approach is considered to be more efficient than the empirical one, as the covariance feature is described by a small number of parameters (usually three or four suffice). However the modeling approach relies on the restrictive assumption that the covariance belongs to a specific parametric family. No such parametric models have been proposed and proven to be valid for fMRI data (Note this is totally different from the autoregressive function in the general linear model where many models have been tested). By contrast, the empirical approach is easy to use and very popular in geostatistics [9]. We consider the empirical approach here.

The empirical autocovariance \( \hat{C}(h) \) can be calculated as follows [26]:

\[
\hat{C}(h) = \frac{1}{N(h)} \sum_{(t_i, t_j) | \|t_i - t_j\| = h} Z(t_i) \cdot Z(t_j) - \hat{\mu}_h^2 \tag{1}
\]

where \( N(h) \) is the number of data pairs, and \( \hat{\mu}_h = 1/N(h) \sum_{(t_i | \|t_i - t_j\| = h} Z(t_i) \), i.e. the average of the \( Z(\cdot) \) values for points within a lag \( h \) of the reference point. The autocorrelation \( \hat{\rho}(h) \) is the autocovariance standardized by the standard deviations:

\[
\hat{\rho}(h) = \frac{\hat{C}(h)}{\hat{\sigma}_h^2} \tag{2}
\]

where \( \hat{\sigma}_h^2 = 1/N(h) \sum_{(t_i | \|t_i - t_j\| = h} Z(t_i)^2 - \hat{\mu}_h^2 \), i.e. the variance of the \( Z(\cdot) \) values for points within a lag \( h \) of the reference point.

In the above calculations, the summation is over only the \( N(h) \) pairs of observations that are separated by lag \( h \). Clearly, \( N(h) \) decreases as \( h \) increases. Since a relatively large \( N(h) \) is associated with a relatively small standard error [27], Journel and Huijbregts [28] suggest that the number of data pairs \( N(h) \) in each class should be at least 30, and this rule of thumb has been widely accepted.

The definitions of autocovariance and autocorrelation can be extended to describing the relations between several variables, i.e. cross-covariance and cross-correlation. For details see Isaaks and Srivastava [26].

**Clustering metric step**: The clustering metric is usually defined as the distance between two objects in multi-dimensional space or time. Two metrics are commonly used: one is the Euclidean distance between objects, called ‘Euclidean’; the other is one minus the sample correlation between objects, called ‘correlation’ [29]. The ‘Euclidean’ method can only compare the similarity of average profile levels among the voxels. We are also interested in the similarity of profile shapes, since voxels with similar patterns should belong to the same functional regions of the brain. The ‘correlation’ metric is more appropriate for this task [29, 30]. We further discuss the connections between the two metrics in Appendix A.
Clustering algorithm step: The k means algorithm is common in neuroimaging applications because of its computational advantages: computations are fast, the algorithm does not require retention of all distances, and convergence occurs quickly [31]. For a given number of clusters k, it iteratively minimizes the within-class variance by assigning data to the nearest center and recalculating each center [5]. Hence, we choose k means algorithm here. However, to use this algorithm we need to choose the number of clusters in advance; silhouette values [32] are used to determine k.

Silhouette values, first introduced by Rousseeuw [32], are commonly used to judge the clustering results. Each cluster is represented by a silhouette with a value between −1 and 1, showing which objects lie well within the cluster and which objects have probably been misclassified. The entire clustering is displayed by plotting all silhouettes into a single diagram, allowing the user to compare the quality of the clusters [33]. The average silhouette width for the entire data set is used to select the number of clusters k, by choosing k so that the average silhouette width is highest [33].

3. DESCRIPTION OF fMRI DATA

An fMRI experiment yields a sequence of 3-dimensional images of the subject’s brain. Each image comprises measurements of the MR signal over a grid of small regular volume elements called voxels. The analysis of eye movements (saccades) has long been used in neurology, since lesions in different brain structures may result in deficits in eye movement control [34]. When subjects are instructed before a visual stimulus to look at the stimulus, this is a prosaccade; when subjects are instructed to perform eye movements in the opposite direction from the location of a stimulus that appears in their peripheral vision, this is called an antisaccade [35]. Prosaccade and antisaccade experiments are commonly used in detecting deficits related to brain lesions [36].

The saccade data of a single subject used here consist of 30 axial slices of size 64×64, taken over 156 time points, with images every 2.5 s, that is, \((x, y, z, t) = 64 \times 64 \times 30 \times 156\). A block design alternating antisaccade and prosaccade tasks was performed. The first two conditions (Anti 1; Pro 1) were set to allow for a 5 s delay in the hemodynamic response. The conditions during the 156 time points were thus: Anti, 1; Pro, 1; Pro, 12; Anti, 12; Pro, 12; ...; Anti, 12; Pro, 10. Pre-processing steps performed on this data set included removal of spatial outliers, correction of head motion, outlier correction in image space, Gaussian filter smoothing with a radius of 2 voxels, removal of linear pixel-wise trends, removal of linear drifts over time for each voxel.

We also analyzed a resting data set, again for a single subject (different from the subject of the first data set), which contains three axial slices of size 64×64, taken over 1498 time points, with images every 2 s, that is, \((x, y, z, t) = 64 \times 64 \times 3 \times 1498\). This is a long-range resting data set. The data were minimally preprocessed.

Masking the brain is a popular dimension reduction technique. When masking, the data image is first processed to identify the location of the brain; thereafter, a suitable threshold method is used to remove all the voxels outside the brain [1, 5, 8, 23, 31]. For the saccade data, we will use both masked and unmasked data in clustering, and compare the results. We choose a single slice for the purpose of demonstration; in this slice, there were 630 voxels (out of 4096) left after masking. For the long-range resting data, we only have masked data leaving us with 1096 voxels out of the original 4096 at the chosen slice.

4. DATA ANALYSIS

4.1. Clustering procedure for saccade data

We first perform clustering analysis on the saccade data, hoping to find the regions in the brain reacting to the task. We focus on the fourth axial slice of the brain, which for this subject is expected by the researcher who provided the data to have the most activity.

In the data preprocessing step, linear drifts over time have already been removed for each voxel. Hence the time series of each voxel is globally stationary and the assumption of the geostatistical analysis is satisfied. We apply clustering based on autocorrelation, and compare the results with those based on cross-correlation [1].

Data reduction: The number of voxels in even a single slice of data still poses a challenge to many clustering techniques. Clustering methods usually do not work well for ill-balanced data. ‘Ill-balanced’ means that the numbers of observations belonging to the different classes are widely disparate, e.g. if most observations belong to one class, then all observations might be put into a single cluster, even if there are different patterns in reality. In fMRI data the population of ‘activated’ (i.e. stimulus-related) voxels is much smaller than the total population of voxels [1, 37, 38]. Hence, a data reduction step is advisable before the three clustering steps to avoid all voxels being assigned to a single cluster.

Since the goal of the screening is to reduce the large amount of non-stimulus-related voxels that could seriously affect the robustness and sensitivity of the clustering results [1, 37], we use a simple two-sample t test procedure comparing prosaccade to antisaccade to screen out the probable non-active voxels [1, 37, 38]. A suitably generous threshold (|t| > 2) is applied to create an image showing those regions of the brain with moderate to strong task-related activation. After thresholding, 345 of 4096 voxels are retained in the unmasked data and 167 of 630 voxels are retained in the masked data; only these are subject to clustering.

Feature extraction: For the 345 most active voxels in the unmasked data, we use the empirical autocorrelations at the 156 time points as the main feature [39]. For the kth voxel, \( k = 1, 2, \ldots, 345 \), define the time series \( \{Z_k(t); \ t = 1, \ldots, 156\} \), with empirical autocovariance and autocorrelation \( \hat{C}_k(h), \hat{\rho}_k(h) \) as in equations (1) and (2). Every voxel is represented by the new vector \( \hat{\mathbf{p}}_k = (\hat{\rho}_k(0), \hat{\rho}_k(1), \ldots, \hat{\rho}_k(155)) \) (Figure 1, graph (3)) instead of the original time series \( Z_k(t) \) (Figure 1, graph (1)). As a comparison, we also give the stimulus series \( Y(t) \) (Figure 1, graph (2)) and calculate its autocorrelation (Figure 1, graph (4)).

Owing to a lack of data, \( \hat{\rho}(h) \) becomes more variable when \( h \) is large. In our analysis, we thus consider \( h \) from 0 up to a maximum lag \( H \). The results in the sequel are based on \( H = 99 \). However, the results are not very sensitive to the particular choice of \( H \). We discuss this further in Section 5.

After clustering, we aim to locate a region or regions that are potentially activated in response to the eye movement task. Since the empirical autocorrelation structures of the found clusters are noisy, it might be difficult to compare them directly with the autocorrelation of the stimulus trail. Hence, we use Fourier transform to convert the autocorrelations (of both clusters and stimulus trail) to their frequency components. If one of the clusters has similar amplitude in the frequency domain to that of the stimulus trail, we posit that this cluster is the one that is responding to the stimulus.

4.1.1. Results of autocorrelation method. In this subsection, we discuss clustering results using autocorrelation as the feature.
Unmasked data: All 345 voxels that passed the initial $t$ test screening are used for clustering. We use average silhouette value to choose the number of clusters; according to the criterion, the appropriate number of clusters is four. After clustering, we mask the average brain image over time and discard all the voxels outside the brain. There are 167 voxels inside the brain of the 345 retained voxels (Figure 2).

Cluster 1 has 111 voxels, of which 88 are inside the brain (Figure 2, graphs (5) and (6)); empirical autocorrelations derived from the time points clearly exhibit strong peaks and troughs in the shape of waves, showing periods of time correlation (Figure 3, graph (4)), which likely result from variations of the blocks in the design paradigm [12, 14]. The signals of these voxels are quite strong (Figure 3, graph (3)), indicating that the activation is probably due to ‘true’ brain activity. The regions identified in cluster 1 (Figure 2, graphs (5) and (6)) are also confirmed to be the active regions by another lab.

Cluster 2 has 80 voxels, of which 46 are inside the brain (Figure 2, graphs (7) and (8)). The empirical autocorrelations of these voxels also exhibit some peaks and troughs, but they do not match up with the stimulus sequence as well as those in cluster 1 (Figures 3, graph (6)). By graphs (7) and (8) in Figure 2, we can see that the voxels in this cluster are around the brain. It seems that the ‘activation’ of this cluster is due to uncorrected head motion, according to the hypothesis of the researcher who supplied the data.
Cluster 3 has 62 voxels, the correlation is around zero. There are no obvious patterns in the signal sequence; the autocorrelation shows some overlapping cyclic patterns that do not correspond to the experimental design (Figure 3, graphs (7), (8)) but may be due to machine artifacts. As shown in graphs (9) and (10) in Figure 2, 33 voxels of this cluster are inside the brain and 29 voxels are outside the brain. We conclude that this cluster contains noise voxels that failed to be screened out by the \( t \) test. Cluster 4 has 92 voxels, the mean of the correlations is almost zero (Figure 3, graph (10)). By looking at graphs (11) and (12) in Figure 2, all the voxels are outside the brain, hence are clearly noise.

To provide further confirmation of these conclusions, we use Fourier transforms to convert the autocorrelations of the stimulus trial (Figure 4, graph (1)) and masked clusters 1, 2 and 3 (Figure 4, graphs (2), (3), (4)) to their frequency components. The dominant frequency in the spectrum of cluster 1 matches remarkably well with that of the stimulus; hence, we conclude that this cluster
Figure 3. Time patterns for saccade data: (1) is the mean of the 345 retained voxels; (3), (5), (7), (9) are the means of the four clusters; (2) is the mean correlation of the 345 voxels; and (4), (6), (8), (10) are the mean correlations of the four clusters. The mean correlation of cluster 1 shows clear peaks and troughs in the shape of waves; the mean correlation of cluster 2 also shows some peaks and troughs in the shape of waves, but they do not match up with the stimulus sequence as well as those in cluster 1; the mean correlations of cluster 3 show some overlapping cyclic patterns that do not correspond to the experimental design. Cluster 4 shows no patterns.

represents voxels that are reacting to the stimulus, as described above. On the other hand, the dominant frequency in the spectrum of cluster 2 is lower than that of the stimulus, therefore can only be explained by ‘head motion’: the dominant frequency in the spectrum of cluster 3 is much higher than that of the stimulus, suggesting it is due to some unknown noises. Our results show consistent conclusions we got before.

**Masked data:** Here the 167 retained voxels in the masked fMRI data are clustered directly. Since in unmasked data, all the voxels in cluster 4 are outside the brain, for comparing with the unmasked data fairly, we look here at the results when $k = 3$. Cluster 1 has 60 voxels and the empirical autocorrelations show strong periods of time correlations, indicating ‘activation’; cluster 2 has
Figure 4. (1) is the spectrum in autocorrelation of stimulus and (2), (3), (4) are the spectra in autocorrelations of clusters 1, 2 and 3 after masking. Since (1) and (2) have the most similar frequencies, cluster 1 is directly in reaction to the stimulus.

73 voxels and the empirical autocorrelations show weak periods of time correlations, indicating head motion; cluster 3 has 34 voxels and the empirical autocorrelation shows some overlapping cyclic patterns that do not correspond to the experimental design, indicating noise or scanner effects (not shown).
Figure 5. Comparison between unmasked and masked methods. The numbers of voxels in ‘activation’, ‘head motion’ and ‘noise’ are 88, 46, 33, respectively, in the unmasked method (graphs (1), (4), (7)), and 60, 73, 34, respectively, in the masked method (graphs (2), (5), (8)). Both methods can extract the 33 voxels of noise (graph (9)), but the unmasked method prefers to assign more voxels to ‘activation’ (graph (3)), while the masked method prefers to attribute more voxels to ‘head motion’ (graph (6)).

The results are very similar to the unmasked data except the numbers of voxels in each cluster change (Figure 5). Both methods extract the 33 voxels of noise inside the brain (Figure 5, graph (9)), but the unmasked method prefers to attribute more voxels to ‘activation’ (Figure 5, graph (3)), whereas the masked method prefers to assign more voxels to ‘head motion’ (Figure 5, graph (6)). By comparing the two ‘activation’ clusters with the maps from another lab, the results from the unmasked method are closer to what has been found by other researchers. Because the saccade data have been preprocessed to correct for head motion, the number of voxels in the ‘head motion’ cluster should be small compared with the number of voxels in the ‘activation’ cluster. This is true for the unmasked data, but not for the masked data. Figure 6 shows the mean correlations of the 60 voxels in the masked method from Figure 5, graph (2), the 46 voxels in the unmasked method from Figure 5, graph (4) and the 28 voxels from Figure 5, graph (3). Clearly, the degree of similarity between graphs (1) and (3) is higher than that between graphs (2) and (3). Hence, the 28 voxels are closer to the ‘activation’ cluster.
This phenomenon can be explained by the so-called ‘marginal effect’. In the calculation of a characteristic of a specific region, the lack of sufficient data outside this region may result in biases or other errors, which is called the ‘marginal effect’ or ‘edge effect’. In order to reduce the impact of the marginal effect on a region, an easy solution is to calculate first on an extended region, then remove the extension after calculation to get the actual results. This idea can be referred to in our clustering analysis. By looking at the unmasked ‘head motion’ cluster (Figure 2, graph (7)), we observe that almost all of the 80 voxels are around the edges of the brain. The 34 voxels outside the brain have a strong ‘marginal effect’ with the 46 voxels that are just inside the brain (Figure 2, graph (8)). Masking the brain before clustering removes all the voxels outside the brain, including the above 34 voxels; therefore, the 46 voxels may group with some other voxels inside the brain (Figure 5, graphs (3) and (6)). Hence, to reduce the ‘marginal effect’, it is preferred to use unmasked data in clustering especially when the data have head motion evident around the edges of the brain.

4.1.2. Results of cross-correlation method. We also use the cross-correlation between the 345 retained voxels and the stimulus as the main feature for clustering. Again for fair comparison, we look at results when $k=4$. There are 72, 106, 109, 58 voxels in the four clusters and 50, 50, 66, 1 voxels inside the brain, respectively. Using the maps in Figure 3 as our standard, the maps in
Figure 7. Maps of the brain for unmasked saccade data, cross-correlation method: (2) and (1) are maps of the 345 retained voxels and overlaid on the original brain. (4) and (3) are maps of the 167 voxels after masking and overlaid on the masked brain; (5) and (6) are maps of cluster 1 before and after masking; (7) and (8) are maps of cluster 2 before and after masking; (9) and (10) are maps of cluster 3 before and after masking; and (11) and (12) are maps of cluster 4 before and after masking. There are 72, 106, 109, 58 voxels in the four clusters before masking, and 50, 50, 66, 1 voxels left after masking.

Figure 7 show that clusters 1 and 3 are not as well classified. The active voxels in the anterior and posterior of the brain are mixed up with noise inside the brain (Figure 7, graphs (6) and (10)). The intention of the cross-correlation method is to try to use the characteristics of delay and habituation between the stimulus and the response time series to create different partitions of the brain. The clustering metric is used to count the total number of relevant matches between the partitions. Perhaps differences in delay and habituation are not strong enough to be distinguished by a clustering method. Hence, in cross-correlation the clustering algorithm may just classify the voxels inside the brain by their voxel values, and not by their different properties. Because of this, researchers sometimes smooth the cross-correlation function [5] to improve its precision and to get better results. This makes a simple case overly complicated, and furthermore is unnecessary, as shown by our autocorrelation results.
4.2. Classification for masked long-range resting data

For the masked long-range resting data set, there should be no task-related differences among the three slices since the subject performed no task. Hence we examine the first slice as an example. Since there is no task-related activation over time, the 2-sample t-test is not suitable here as a screening device and all 1096 voxels are used in clustering. This is a long-range data with maximum lag distance 1498, and we pick an effective lag distance of \( h = 1000 \) to reduce the measurement errors. We use the ‘correlation’ metric again and the mean of silhouette values indicates that the best choice for the number of clusters is \( k = 2 \) (not shown). Cluster 1 has 646 voxels (Figure 7, graph (1)); cluster 2 has 450 voxels (Figure 8, graph (2)). There are no clear patterns in the two clusters. Neither cluster shows peaks or troughs in time (Figure 8), indicating that there are no task-related or systematic activations in the resting data. This is what we would expect. Since both clusters are spatially dispersed, it is not entirely clear what the method is picking out. Comparing graphs (4) and (5) in Figure 9, it appears that cluster 2 may be gathering together voxels that are affected by slightly head motion.

5. DISCUSSION AND CONCLUSION

5.1. Different comparisons

Comparison between autocorrelation and cross-correlation method: Clustering on the cross-correlation function instead of the raw time series may provide increased robustness. This type of clustering is an example of an hypothesis-driven analysis [8]. A drawback of hypothesis-driven analysis is that it depends heavily on having prior knowledge of the reference function. Sometimes it is difficult to know this in advance, and sometimes such a reference waveform does not exist. For example, in an experiment that involves having the subject watch a film clip, there is not a clear ‘time course’ with which to cross-correlate. Even in a study with a well-defined stimulus waveform, the subject may not follow the instructions correctly during the experiment and it is not clear how this will affect the cross-correlation analysis. Furthermore, we might not know the expected response to the task for a particular region because the study is exploratory. One can easily envision other scenarios where it might be difficult or undesirable to work with a predetermined
Figure 9. Time patterns of resting data: (1), (2) are the means of the two clusters; (3) is the mean of all 1096 voxels; (4), (5) are the mean covariances of the two clusters; and (6) is the mean covariance of all 1096 voxels. Neither cluster exhibits obvious peaks and troughs in time, indicating that there are no systematic activations in the resting data.

reference function. The autocorrelation method does not rely on such an external standard, and hence can be considered as a type of data-driven analysis [8]. It offers further improvement over the cross-correlation method, which itself improves on clustering of the raw time series. Using the autocorrelation function as a main feature in clustering makes fewer assumptions. In particular, as we have seen, we need not consider the hypothesized relation between the stimulus and the response time series (including concerns about lags in the onset of the hemodynamic response). Our results indicate that this approach is superior in identifying active regions of the brain for task
data, and can be used for resting data to examine functional similarities across the brain during ambient thought.

Comparison between saccade data and resting data: The autocorrelations of the clusters in the saccade data exhibit peaks and troughs in the shape of waves in two of the clusters, showing periods of time correlations. This periodicity is clearer than that shown by the raw voxel time courses and reflects changes in the experimental condition. By contrast, the autocorrelations of clusters for the resting data do not show any clear patterns over time. These results demonstrate the good performance of autocorrelation in a variety of situations and its superiority over using the voxel time courses alone.

Comparison between unmasked method and masked method: It is common to mask the brain as the first step in an analysis [1, 5, 8, 31], since masking can eliminate a large number of ‘useless’ voxels outside the brain and make the further analysis computationally more convenient. But masking the brain before clustering may be problematic, as it ignores the effects of voxels around the edges of the brain. These voxels may affect the clustering results especially for data with head motion.

5.2. Discussion of different techniques

2-sample t-test for dimension reduction: Using the 2-sample $t$-test for dimension reduction is a popular technique in the fMRI data analysis. It is simple, convenient and generally effective in extracting the most important voxels. But it also has some clear disadvantages. First, it requires a contrast between experimental conditions, so it does not work for resting data since there are no such conditions to compare nor is it truly suitable for event-related designs. Second, delay of the responses is not always considered, and so the calculation of the 2-sample $t$ statistic may not be accurate. Sparse principal component analysis [40] is a new technique that can combine dimension reduction and feature extraction together in clustering, and which may overcome the disadvantages of the 2-sample $t$-test. We explore this approach in a companion paper [41].

Use of empirical autocorrelation: Using empirical autocorrelation in the fMRI data analysis is attractive since it captures important characteristics of the voxels over time and does not require prior knowledge of the reference function. But there are some implementation issues. For example, when the method of moments is used to estimate the autocorrelation, the standard error of $\hat{\rho}(h)$ increases as lag $h$ increases, because we have fewer data pairs for large $h$. Hence the autocorrelation needs to be truncated at an effective lag. On the other hand, we do not want to lose important structure by truncating too much. In practice, it is recommended to choose the number of data pairs to be no less than 30 or to truncate the lag distance at half of the maximum distance [28, 42]. We try different effective lag values and find that the clustering results are not very sensitive, for a reasonable range of effective lags. Table I lists average silhouette values and the number of misclassified voxels inside the brain for different effective lag values. To keep the balance between retaining more data information and reducing measurement error, an effective lag distance of 99 appears to be optimal in this case.

6. CONCLUSIONS

In the analysis of brain imaging data, using the autocorrelation function offers an important advantage over existing cross-correlation approaches. Unlike the conventional cross-correlation
Table I. Average silhouette values and the number of misclassified voxels inside the brain at different lags.

<table>
<thead>
<tr>
<th>Effective lag</th>
<th>Silhouette value</th>
<th>Misclassified voxels</th>
</tr>
</thead>
<tbody>
<tr>
<td>91</td>
<td>0.2747</td>
<td>4</td>
</tr>
<tr>
<td>93</td>
<td>0.2710</td>
<td>3</td>
</tr>
<tr>
<td>95</td>
<td>0.2623</td>
<td>3</td>
</tr>
<tr>
<td>97</td>
<td>0.2630</td>
<td>6</td>
</tr>
<tr>
<td>99</td>
<td>0.2682</td>
<td>1</td>
</tr>
<tr>
<td>101</td>
<td>0.2680</td>
<td>3</td>
</tr>
<tr>
<td>103</td>
<td>0.2628</td>
<td>2</td>
</tr>
<tr>
<td>105</td>
<td>0.2584</td>
<td>3</td>
</tr>
</tbody>
</table>

To keep the balance between retaining more data information and reducing measurement error, an effective lag distance of 99 is optimal.

methods, the proposed method does not require prior knowledge about the reference function, and does characterize the important features of voxel changes in time. The analysis also provides evidence that masking the brain may affect the clustering results. Although many researchers often choose masking the brain as the first step to reduce the dimension, we show that this is not necessarily effective and sometimes results in less convincing results, especially when the data have ‘head motion’ evident around the edges of brain.

APPENDIX A

Given an $(n \times p)$ matrix $X$, where $n$ is the number of observations (time points) and $p$ is the number of variables (locations), define $\mu_j$ and $\sigma_j^2$ as the mean and variance of the $j$th vector $x_j = (x_{1j}, \ldots, x_{nj})^T$, where $j = 1, \ldots, p$; and define a new vector $\bar{x}_j = (\bar{x}_{1j}, \ldots, \bar{x}_{nj})^T$, where $\bar{x}_{ij} = (x_{ij} - \mu_j) / (\sqrt{n}\sigma_j)$, $i = 1, \ldots, n$. Then $\bar{x}_j$ is normalized since $\bar{x}_j^T \bar{x}_j = \sum_{i=1}^{n} \bar{x}_{ij}^2 = 1$. Note that $d^2(\bar{x}_j, \bar{x}_k) = (\bar{x}_j - \bar{x}_k)^2 = 2(1 - \bar{x}_j^T \bar{x}_k) = 2d_{corr}^2(x_j, x_k)$ for $j, k = 1, \ldots, n$. The square of the ‘correlation’ metric is an alternative square of the normalized ‘Euclidean’ metric. Hence, the ‘correlation’ metric can extract the similarity of profile shapes in clustering.

ACKNOWLEDGEMENTS

We wish to express our appreciation to Dr Rebecca L. McNamee of the University of Pittsburgh for kindly providing the saccade data set and her conscientious assistance with physiological interpretations. We also would like to thank Dr Nathan Yanasak, now at the Medical College of Georgia, for his assistance with the resting data acquisition. Nicole A. Lazar was partly supported by NSF grant DMS-0706192. We thank the referees of an earlier version of this paper, who gave helpful advice on clarifying and explaining our ideas.

REFERENCES


31. Hallett PE. Primary and secondary saccades to goals defined by instructions. *Vision Research* 1978; **18**:1279–1296.


