

Methods for Detecting Functional Classifications in Neuroimaging Data

F. DuBois Bowman,* Rajan Patel, and Chengxing Lu

Department of Biostatistics, Rollins School of Public Health, Emory University, Atlanta, Georgia

Abstract: Data-driven statistical methods are useful for examining the spatial organization of human brain function. Cluster analysis is one approach that aims to identify spatial classifications of temporal brain activity profiles. Numerous clustering algorithms are available, and no one method is optimal for all areas of application because an algorithm's performance depends on specific characteristics of the data. *K*-means and fuzzy clustering are popular for neuroimaging analyses, and select hierarchical procedures also appear in the literature. It is unclear which clustering methods perform best for neuroimaging data. We conduct a simulation study, based on PET neuroimaging data, to evaluate the performances of several clustering algorithms, including a new procedure that builds on the *k*th nearest neighbor method. We also examine three stopping rules that assist in determining the optimal number of clusters. Five hierarchical clustering algorithms perform best in our study, some of which are new to neuroimaging analyses, with Ward's and the beta-flexible methods exhibiting the strongest performances. Furthermore, Ward's and the beta-flexible methods yield the best performances for noisy data, and the popular *K*-means and fuzzy clustering procedures also perform reasonably well. The stopping rules also exhibit good performances for the top five clustering algorithms, and the pseudo- T^2 and pseudo-*F* stopping rules are superior for noisy data. Based on our simulations for both noisy and unscaled PET neuroimaging data, we recommend the combined use of the pseudo-*F* or pseudo- T^2 stopping rule along with either Ward's or the beta-flexible clustering algorithm. *Hum Brain Mapp* 23:109–119, 2004. © 2004 Wiley-Liss, Inc.

Key words: cluster analysis; hierarchical clustering; *F*-score; stopping rules; PET

INTRODUCTION

Functional neuroimaging with positron emission tomography (PET) and functional magnetic resonance imaging (fMRI) measures localized brain activity using correlates of regional cerebral blood flow (rCBF). In vivo neuroimaging enables the investigation of a wide range of aspects of human brain function. Descriptive statistical approaches play

an important role in understanding characteristics of neuroimaging data and often generate hypotheses for analyses that seek to identify task-related changes in brain activity. In particular, methods for classifying anatomical regions that exhibit similar patterns of activity are useful tools for describing the organization and connectivity of human brain function.

Cluster analysis is a collection of data-driven statistical procedures that can assist in identifying spatial classifications of brain activity. Clustering has a long history in the statistical literature and has recent applications in functional neuroimaging, particularly in fMRI. Our analyses focus on the classification of measured brain activity, but cluster analysis is also useful for anatomical segmentation of brain tissues [Zhu and Jiang, 2003]. Neuroimaging applications currently utilize a limited number of existing clustering algorithms, largely selected based on computational facility and speed. Popular methods include the *K*-means approach [Balslev et al., 2002; Goutte et al., 1999; Goutte et al., 2001;

Contract grant sponsor: NIH; Contract grant number: K25-MH65473.

*Correspondence to: Dr. F. DuBois Bowman, Department of Biostatistics, Emory University, 1518 Clifton Road, N.E. Atlanta, GA 30322. E-mail: dbowma3@sph.emory.edu

Received for publication 23 January 2004; Accepted 30 March 2004
DOI: 10.1002/hbm.20050

Published online in Wiley InterScience (www.interscience.wiley.com).

MacQueen, 1967] and fuzzy clustering [Baumgartner et al., 2000; Fadili et al., 2000, 2001; Sommer and Wichert, 2003]. In addition, select hierarchical clustering methods appear including single, complete, and average linkage methods [Cordes et al., 2002; Goutte et al., 1999; Stanberry et al., 2003] as well as a hybrid hierarchical K-means approach [Filzmoser et al., 1999], and dynamical cluster analysis (DCA) [Baune et al., 1999].

Clustering procedures either make classifications based on a pre-specified number of groups, or they require the analyst to select the number of groups from a hierarchy of nested merges or divisions. Hierarchical clustering methods often employ stopping rules as criteria for selecting the number of clusters, and partition-based methods with a fixed number of clusters utilize various optimization schemes. A vast number of methods exist for determining the number of clusters, but relatively few appear in the neuroimaging literature. Milligan and Cooper [1985] provide an empirical comparison of 30 stopping rules for hierarchical clustering algorithms. The most competitive stopping rules in their study were pseudo-F [Calinski and Harabasz, 1974] and pseudo-T² [Duda and Hart, 1973].

A known limiting characteristic of cluster analysis is that various procedures often produce discrepant results. The performance of a method is largely influenced by characteristics of the data. Many algorithms are capable of detecting well-separated clusters, but generally, a given method tends to detect clusters based on certain characteristics such as compactness, shape, size, and heterogeneity. An important, and unexplored, question is “how do the performances of various clustering algorithms compare for neuroimaging applications?” Selecting the number of clusters is challenging for neuroimaging applications due to the lack of biological evidence supporting a specific number of clusters, coupled with the enormous number of possible choices, e.g., ranging from one to the total number of voxels. The comparison of stopping rules by Milligan and Cooper [1985] uses artificial data that were not aligned to any particular biomedical application. Generalizing their findings to neuroimaging applications is questionable, given the potential dependence of the stopping rules’ performances on inherent aspects of the data and the distinct contrast between characteristics of their simulated data and neuroimaging data. Another limitation of the Milligan and Cooper [1985] simulation study is that they pool results across four hierarchical clustering algorithms, but fail to delineate differences in the performances of the stopping rules for select clustering procedures.

Using data based on a PET neuroimaging experiment, we conduct a simulation study to empirically evaluate the relative performances of numerous clustering algorithms. We include a range of algorithms in our study, several of which are new to neuroimaging, and we discuss the relative merits and limitations of various approaches based on our empirical comparison. Among the algorithms that we consider is a new hierarchical clustering procedure that exhibits competitive properties among the other methods. We also per-

form an empirical evaluation of the pseudo-F, pseudo-T², and cubic clustering criterion (CCC) [Sarle, 1983] stopping rules that help guide the selection of the number of clusters. Our analysis not only allows us to discuss the relative performances of the stopping rules, but we also identify specific clustering methods for which designated stopping rules perform well.

MATERIAL AND METHODS

Notation

The simulated data are based on a study examining the neural correlates of drinking alcohol. The experimental data include 10 subjects, indexed by $k = 1, \dots, K$, and we consider four scans for each individual, indexed by $s = 1, \dots, S$. One scan from an individual consists of rCBF measurements $\mathbf{Y}_{ks} = (Y_{ks}(1), \dots, Y_{ks}(V))'$, expressible in our data as a $91 \times 109 \times 91$ rectangular array of voxels, where $V = 902629$. Injections of a low dose of ethanol and a high dose of ethanol precede the second and third scans, respectively, resulting in increasing blood alcohol concentration levels across the scans. Let $\mathbf{Y}_k(\nu) = (Y_{k1}(\nu), \dots, Y_{kS}(\nu))'$ represent serial rCBF measurements for the k th subject at voxel ν , where $\nu = 1, \dots, V$, and the statistic, $\mathbf{T}(\nu) = (T_1(\nu), \dots, T_P(\nu))' = f(\mathbf{Y}_1(\nu), \dots, \mathbf{Y}_K(\nu))$, $P \leq S$, summarizes data across individuals, e.g., the mean scan value across subjects (see the Results section for further details).

Clustering Methods

We provide a brief exposition of several clustering algorithms that we incorporate into our empirical comparison, including hierarchical clustering, K-means, and fuzzy clustering methods. For consistency with conventional nomenclature, we use the phrase K-means, although we classify data into G groups. We present a new hierarchical clustering method that employs a more stringent criterion for merging clusters than single-linkage.

The general objective of neuroimaging clustering is to spatially distinguish collections of voxels into G well-separated clusters that exhibit similar patterns of measured brain activity within groups. K-means and fuzzy clustering procedures require advanced specification of G , while hierarchical methods allow selection of G following completion of the algorithm. Clustering algorithms use measures of distance to determine dissimilarity between voxels (or clusters), and the standard class of metrics for a pair of voxels (v_i, v_j) is

$$d(\mathbf{T}(v_i), \mathbf{T}(v_j)) = [(\mathbf{T}(v_i) - \mathbf{T}(v_j))' \mathbf{B}_{v_i v_j} (\mathbf{T}(v_i) - \mathbf{T}(v_j))]^{1/2}, \quad (1)$$

where $\mathbf{B}_{v_i v_j}$ is an $S \times S$ positive definite symmetric matrix, often selected as the inverse of a covariance matrix or as the identity matrix. We employ a distance metric that can be expressed in the form of equation (1), where the normalizing matrix is the inverse of the covariance matrix of $\mathbf{T}(v_i) - \mathbf{T}(v_j)$.

K-Means

The *K*-means algorithm begins by specifying the number of clusters G , randomly initializing the voxels to the G groups, and calculating the group centroids. The algorithm iteratively reassigns voxels to the cluster with the nearest centroid and recalculates the cluster means and ceases when no reassignments occur. *K*-means is common in neuroimaging applications, in part because of the computational advantages. Computations are fast, the algorithm does not require retention of all distances, and convergence usually occurs quickly. A substantive disadvantage of the *K*-means approach is that it requires advanced specification of G and there is no scientific basis for setting G in most neuroimaging studies.

Hierarchical Clustering

Hierarchical clustering begins with each voxel representing a separate cluster and proceeds with successive merges or alternatively begins with one cluster and conducts a sequence of nested divisions. For brevity, we focus on agglomerative hierarchical clustering, which is more common in neuroimaging than divisive procedures. Beginning with V clusters and the $V \times V$ distance matrix, with (i, j) th element $d(T(v_i), T(v_j))$, agglomerative methods proceed as follows:

1. Identify the smallest distance and combine the corresponding clusters, reducing the total number of clusters by one after each iteration.
2. Compute an updated distance matrix by removing rows and columns associated with the two merged clusters and adding a row and column containing updated distances between the new cluster and all other clusters. Specific hierarchical clustering algorithms use different updating functions $d(g, g^*)$ to measure the distance between clusters g and g^* . Let Θ_g represent the set of voxels in cluster g , and let V_g denote the number of voxels comprising Θ_g , where $V = \sum_{g=1}^G V_g$. Several hierarchical algorithms are as follows:
 - a. Single Linkage:
 $d(g, g^*) = \min(d(T(v), T(v^*)))$, for all $v \in \Theta_g$ and $v^* \in \Theta_{g^*}$
 - b. Complete Linkage:
 $d(g, g^*) = \max(d(T(v), T(v^*)))$, for all $v \in \Theta_g$ and $v^* \in \Theta_{g^*}$
 - c. Average Linkage:
 $d(g, g^*) = (V_g V_{g^*})^{-1} \sum_{v \in \Theta_g} \sum_{v^* \in \Theta_{g^*}} d(T(v), T(v^*))$
 - d. Centroid Method:
 $d(g, g^*) = d(V_g^{-1} \sum_{v \in \Theta_g} T(v), V_{g^*}^{-1} \sum_{v^* \in \Theta_{g^*}} T(v^*))$
 - e. Median Linkage:
 $d(g, g^*) = \frac{1}{2} [d(g_i, g) + d(g_j, g) - \frac{1}{2} d(g_i, g_j)]$
 - f. Ward's Method:
 $d(g, g^*) = (V_g^{-1} + V_{g^*}^{-1})^{-1} d(V_g^{-1} \sum_{v \in \Theta_g} T(v), V_{g^*}^{-1} \sum_{v^* \in \Theta_{g^*}} T(v^*))$

g. Beta-Flexible:

$$d(g, g^*) = (1 - \beta) \left[\frac{1}{2} (d(g, g_i) + d(g, g_j)) \right] + \beta d(g_i, g_j),$$

where (e) and (f) give updated distances after merging clusters g_i and g_j to form g^* , and $\beta \in [-1, 1]$.

3. Repeat steps until all clusters unite.

Lance and Williams [1967] give a general distance formula that includes all of the hierarchical distances listed above as special cases [Rencher, 2002].

We propose a new hierarchical clustering procedure, called *variable linkage*, which is competitive with other well-performing algorithms. Variable linkage is a k th nearest neighbor clustering procedure [Wong and Lane, 1983], where k increases proportionally with the product of the sizes of the two joining clusters. Thus, we can set α as a fraction in the interval $[(V_g V_{g^*})^{-1}, 1]$ and define $k = \alpha(V_g V_{g^*})$. The distance between two clusters is given by the k th smallest distance among the $V_g V_{g^*}$ voxel-to-voxel distances. Variable linkage approaches single linkage as α approaches its lower bound and complete linkage as α approaches one. We implement variable linkage using $\alpha = 0.15$ in our simulation study.

Fuzzy Clustering

Fuzzy clustering appears frequently in neuroimaging applications and has the distinction of using partial, rather than binary, classifications of voxels to clusters. Let $\mathbf{U}(G \times V)$ represent a matrix that partitions the data into G clusters, with rows of \mathbf{U} indexing clusters and columns indicating voxels. Each element $u_{gv} \in \mathbf{U}$ represents a weight of classification of voxel v into cluster g , where $u_{gv} \in [0, 1]$, $\sum_{g=1}^G u_{gv} = 1$, and $0 < \sum_{v=1}^V u_{gv} \leq V$. Cluster centroids are given by

$$m_p(g) = \frac{\sum_{v=1}^V u_{gv}^z T_p(v)}{\sum_{v=1}^V u_{gv}^z}, \quad p=1, \dots, P, \quad (2)$$

where $z \geq 1$ is a parameter that controls the fuzziness of the clusters. The algorithm computes distances between the g th centroid and the data series from voxel v and then updates \mathbf{U} using

$$u_{gv} = \left[\sum_{g^*=1}^G \left(\frac{d^2(\mathbf{m}(g), T(v))}{d^2(\mathbf{m}(g^*), T(v))} \right)^{1/(z-1)} \right]^{-1}. \quad (3)$$

This procedure characterizes the solution minimizing the function [Bezdek et al., 1984]

$$J_V = \sum_{v=1}^V \sum_{g=1}^G u_{gv}^z d_{gv}^2.$$

The iterations cease when negligible change occurs in U between successive iterations.

Stopping Rules

We implement pseudo-F, pseudo-T², and CCC to estimate the number of population clusters. These methods were among the most competitive stopping rules in an empirical comparison of 30 criteria conducted by Milligan and Cooper [1985]. When merging clusters g_i and g_j to form g^* , the pseudo-F and pseudo-T² statistics, for a given number of clusters, are

$$\text{pseudo-F} = \frac{(Q - \sum_{g=1}^G W_g)/(G-1)}{\sum_{g=1}^G W_g/(V-G)} \quad \text{and}$$

$$\text{pseudo-T}^2 = \frac{W_{g^*} - W_{g_i} - W_{g_j}}{(W_{g_i} + W_{g_j})/(V_{g_i} + V_{g_j} - 2)}. \quad (4)$$

where $Q = \sum_{v=1}^V \|T(v) - \bar{T}\|^2$ and $W_{g_i} = \sum_{v \in \Theta_{g_i}} \|T(v) - \bar{T}_{g_i}\|^2$. The distributions of these statistics are difficult to ascertain, particularly for neuroimaging data, but the statistics still provide useful information regarding within and between cluster variability. Under the naive assumptions that the data reflect independent samples from a multivariate normal distribution with a scalar covariance matrix and that voxels are randomly assigned to each cluster, the pseudo-F statistic follows an F distribution with $q(G-1)$ and $q(n-G)$ degrees of freedom, and pseudo-T² follows an F distribution with q and $q(V_{g_i} + V_{g_j} - 2)$ degrees of freedom. CCC compares the square of the observed correlation coefficient (R^2) and an approximation of its expected value, given that the data represent samples from a uniform distribution on a hyperbox and that the clusters have hypercubic shapes. Details for computing CCC are available in Sarle [1983]. Positive values of CCC indicate the possible presence of clusters.

Evaluating Cluster Performance

An important consideration for any clustering application is the quality of the final partition. We outline two procedures to evaluate cluster performance in our simulation study that utilize the defined partition as a benchmark. Bowman and Patel [2004] present a likelihood-based approach that is more useful in practice since it does not rely on a benchmark. For the defined partition, let Ω_{g^*} represent a class of V_{g^*} functionally connected voxels (a true cluster), and V_{g^*} is the number of voxels in both Ω_{g^*} and Θ_{g^*} , where $g^* = 1, \dots, G^*$.

F-score measure

A measure of overlap between a true class and a computed cluster is given by

$$F_1(\Theta_{g^*}, \Omega_{g^*}) = \frac{2\hat{\phi}_1\hat{\phi}_2}{\hat{\phi}_1 + \hat{\phi}_2}, \quad (5)$$

where $\phi_1 = \Pr(v \in \Omega_{g^*} | v \in \Theta_{g^*})$ and $\phi_2 = \Pr(v \in \Theta_{g^*} | v \in \Omega_{g^*})$ [Larsen and Aone, 1999]. We define estimates $\hat{\phi}_1 = V_{g^*}/V_{g^*}$ and $\hat{\phi}_2 = V_{g^*}/V_{g^*}$, representing the predictive ability and the accuracy of a partition, respectively. F_1 represents the harmonic mean of $\hat{\phi}_1$ and $\hat{\phi}_2$, since it is expressible as one over the average of $\hat{\phi}_1^{-1}$ and $\hat{\phi}_2^{-1}$. We define $F_2(\Omega_{g^*}) = \max_{\Theta_{g^*}} \{F_1(\Theta_{g^*}, \Omega_{g^*})\}$, giving the maximum F_1 among clusters at the node selected to terminate the hierarchical tree or the final partition when specifying the number of clusters in advance. This is a variation from the measure presented by Larsen and Aone [1999], which alternatively defines F_2 considering all nodes in a tree. The overall F -score is the weighted sum of the class-specific F_2 scores

$$F = \sum_{g^*=1}^{G^*} \frac{V_{g^*}}{V} F_2(\Omega_{g^*}). \quad (6)$$

A perfect clustering solution yields an F -score equal to one, and F tends toward zero for clustering with poor predictive ability and poor accuracy. Generally, higher F values indicate improved clustering.

Clustering reliability

A reliable partition of the data should produce clusters that are credible, in light of the known classifications. One approach to evaluating reliability, in this sense, is to compute a measure of model uncertainty, given the data, relative to the true model.

We compute the clustering reliability (CR) (λ_i) for a partition \mathcal{R}_i , as

$$\lambda_i = \frac{f(T|\hat{\mathbf{r}}^{(i)}, \mathcal{R}_i)}{f(T|\hat{\mathbf{r}}^{(0)}, \mathcal{R}_0)} (qV)^{-(1/2)[(G-G^*)(q+1)]}, \quad (7)$$

where the index "0" denotes the true classifications, and $f(T|\hat{\mathbf{r}}, \mathcal{R})$ represents the probability density function of T , given a partition and corresponding parameter estimates. A value of λ_i greater than one suggests that the associated algorithm produces a more plausible partition of the data than the true classifications. CR readily extends to probability models other than the multivariate normal distribution. CR approximates a Bayes factor [Carlin and Louis, 2000], which allows model comparisons under a Bayesian framework.

RESULTS

Simulated Data

The simulated data consist of $N = 100$ data sets with 11 clusters and are based on experimental PET data. Here, we focus on an axial slice 20 mm below the anterior commissure. Each of the 10 subjects in the experimental data has a

series of 4 scans for a total of 40 observations. We first fit a general linear model (GLM) to localized brain activity from the experimental data and compute voxel-specific summary statistics $T(\nu) = \hat{\beta}(\nu)$ and corresponding estimates of the covariance matrices $\hat{\sigma}_\nu^2(X'X)^{-1}$ using least-squares estimation. The GLM for rCBF includes cell means for each scan and a covariate adjustment for the global cerebral blood flow (gCBF). We use the K -means algorithm to organize the vectors of voxel-specific summary statistics into 11 clusters. This preliminary cluster analysis employs a distance metric with a normalizing matrix that is interpretable as an inverse covariance matrix under the assumption that every voxel has a distinct variance parameter. We select $G = 11$ as the number of clusters by examining the pseudo- T^2 , pseudo- F , and CCC statistics across the range of 1 to 30 clusters. These criteria reveal a local optimum at or near 11 for several of the clustering algorithms.

We also consider Brodmann's standard anatomical and functional classifications of the human brain [Brodmann, 1909] to corroborate our selection of 11 clusters. There are 15 Brodmann areas present in the anatomical region that we evaluate, and four of these regions are extremely small, collectively containing fewer than 2.5% of the brain voxels. Our empirical assessment and clinical considerations for determining the number of clusters both lead to G within a very small range. We present results here assuming $G = 11$ clusters; however, we examined the sensitivity of our findings to this selection by re-analyzing the data using 15 clusters. The results for $G = 15$ clusters are quite consistent with those given in this report.

We simulated data, for each voxel, from a multivariate normal distribution with the mean vector and covariance matrix matching the corresponding cluster-specific values in the experimental data. Maitra [1997] provides justification for assuming a multivariate normal distribution for PET data, but our analyses also permit other probability models. To facilitate execution of the clustering algorithms, we apply a spherical transformation $Z(\nu) = H\hat{\beta}(\nu)$, where $X'X = H'H$ and H is a $q \times q$ upper triangular matrix, and we perform all cluster analyses on $Z(\nu)$. Asymptotically, $Z(\nu)$ follows a multivariate normal distribution and contains elements that are independent with equal variances σ_g^2 , where $\nu \in \Theta_g$. Using the transformed data, estimation of the $G(q+1)$ parameters $\hat{\tau}_g = (\hat{\mu}_g, \hat{\sigma}_g^2)$, $g = 1, \dots, G$, is quick and involves only routine computations.

We normalize the distance between two voxels using $B_{\nu_1\nu_2} = (\sigma_{\nu_1}^2 + \sigma_{\nu_2}^2)^{-1}\mathbf{I}$, where \mathbf{I} represents an identity matrix, the distance between two cluster centroids by

$$B_{gg^*} = \left(V_g^{-2} \sum_{i \in \Theta_g} \sigma_i^2 + V_{g^*}^{-2} \sum_{j \in \Theta_{g^*}} \sigma_j^2 \right)^{-1} \mathbf{I}, \quad (8)$$

and the distance between a voxel ν and the centroid of cluster g using

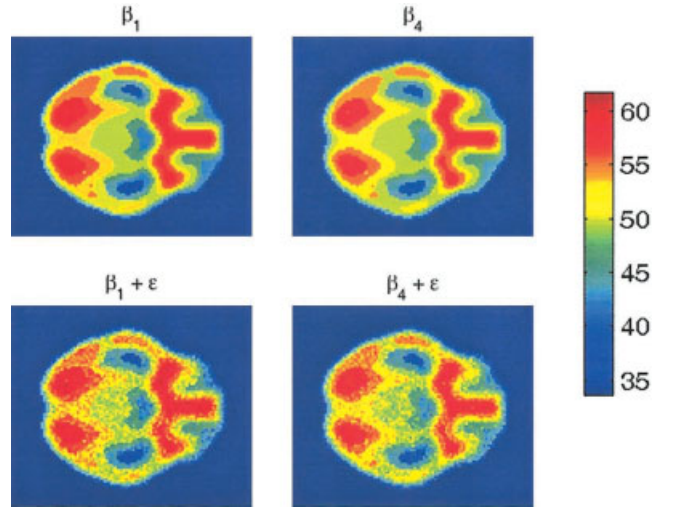


Figure 1.

Top: The within-cluster mean rCBF values at -20 mm for the first scan (β_1) and the last scan (β_4), revealing the defined clusters. **Bottom:** One simulated data set for the first and last scans.

$$B_{\nu g} = \left(a\sigma_\nu^2 + V_g^{-2} \sum_{i \in \Theta_g} \sigma_i^2 \right)^{-1} \mathbf{I},$$

$$\text{where } \begin{cases} a = 1, & \text{if } \nu \notin \Theta_g \\ a = (1 - 2V_g^{-1}), & \text{if } \nu \in \Theta_g \end{cases}. \quad (9)$$

Each normalizing matrix corresponds to an inverse covariance matrix based on the transformed data. Clustering $Z(\nu)$ is equivalent to clustering the untransformed data with distances normalized by the inverse covariance matrices from the original data scale. Specifically,

$$\begin{aligned} d^2(Z(\nu_i), Z(\nu_j)) &= (Z(\nu_i) - Z(\nu_j))' B_{\nu_i\nu_j} (Z(\nu_i) - Z(\nu_j)) \\ &= (HT(\nu_i) - HT(\nu_j))' B_{\nu_i\nu_j} (HT(\nu_i) - HT(\nu_j)) \\ &= (T(\nu_i) - T(\nu_j))' B_{\nu_i\nu_j}^* (T(\nu_i) - T(\nu_j)), \end{aligned} \quad (10)$$

where $B_{\nu_i\nu_j}^* = B_{\nu_i\nu_j}(X'X)$. Therefore, the transformation simplifies computations while retaining the interpretation of a cluster analysis performed on the original summary statistics.

Figure 1 (top) displays means of the first and last scans (β_1 and β_4) for the 11 defined clusters from an axial slice of the experimental data at -20 mm. Figure 1 (bottom) shows one simulated data set containing voxel-wise summary statistics, i.e., simulated mean rCBF values for scan 1 and scan 4, adjusted for gCBF. The simulated data retain much of the overall structure visible in the defined clusters, amid random noise. The clustering algorithms attempt to extract the structure from the simulated images by identifying groups that exhibit similar blood flow characteristics.

TABLE I. Performance of pseudo-F, CCC, and pseudo-T² stopping rules from the simulation study for selecting the number of clusters*

Stopping criterion	Clustering method	Number of computed clusters				
		≤9	10	11	12	≥13
Pseudo-F	Single	92	3	2	0	3
	Complete	56	9	9	10	16
	Variable	0	0	100	0	0
	Median	4	18	53	17	8
	Centroid	0	0	100	0	0
	Average	0	0	100	0	0
	Ward's	0	0	100	0	0
	Beta-flex	0	0	100	0	0
	K-means	78	19	0	3	0
	Fuzzy K	2	11	56	3	28
CCC	Single	95	1	2	0	2
	Complete	60	10	9	10	11
	Variable	0	0	100	0	0
	Median	17	20	48	15	0
	Centroid	0	0	100	0	0
	Average	0	0	100	0	0
	Ward's	0	0	100	0	0
	Beta-flex	0	0	100	0	0
	K-means	100	0	0	0	0
	Fuzzy K	6	13	53	1	27
Pseudo-T ²	Single	27	15	20	20	18
	Complete	29	23	19	14	15
	Variable	0	0	100	0	0
	Median	5	21	52	17	5
	Centroid	0	0	100	0	0
	Average	0	0	100	0	0
	Ward's	0	0	100	0	0
	Beta-flex	0	0	100	0	0

* The true number of clusters is 11. For each stopping criteria, the table indicates the number of times that the 10 clustering algorithms select the specified number of clusters. Pseudo-T² selects the number of clusters g^* at which a large decrease occurs from $(g^* - 1)$ clusters. Pseudo-F and CCC both select the number of clusters based on the maximum value.

Stopping Rules

Table I presents simulation results for the stopping rules. The three stopping rules perform well in the simulation study for several clustering algorithms, correctly identifying the number of clusters in the true data, with high probability. In particular, variable linkage, centroid linkage, average linkage, Ward's, and the beta-flexible methods accurately detect the correct number of clusters for all three stopping rules, and single and complete linkage perform very poorly for the three criteria. *K*-means is calculable for only pseudo-F and CCC, and it exhibits poor performances using both measures. Pseudo-F and CCC consistently underestimate the true number of clusters in the data for the *K*-means algorithm, selecting $G \leq 9$ as the true number of clusters in all of the simulations for CCC and in 78% of the simulations for pseudo-F. The stopping rules tend to identify the correct number of clusters for fuzzy *K*-means and median linkage, but with much less consistency than the five algorithms that perform best. The prototypical application of partitioning methods such as *K*-means and fuzzy *K*-means assumes that the number

of clusters is known prior to execution of the algorithm. Although the stopping rules are often adopted for partitioning methods in practice, they are primarily intended to select the number of clusters among a nested sequence of cluster merges, perhaps contributing to the superior performances of some of the hierarchical algorithms in our study.

Clustering Algorithms

We consider 10 clustering algorithms and compute classifications consisting of 9 to 13 clusters, for a total of 50 partitions (i.e., the number of algorithms times the number of computed clusters). The results of the empirical comparisons appear in Table II and Figure 2. Table II presents the 15 partitions with the highest F-scores among the 50 candidates, and Figure 2 provides a comparison of all classifications with 11 clusters. Overall, the best classifications, according to F-score, correspond to Ward's, beta-flexible ($\beta = -0.5$), variable linkage ($\alpha = 0.15$), and centroid linkage methods with 11 clusters. The top 11 methods in Table II all

TABLE II. Top 15 classifications according to F-score*

Algorithm	Number of clusters	F-score	log(CR)
1. Ward's method	11	0.99877 (0.0008)	44022.75 (199.58)
2. Beta-flexible	11	0.99860 (0.0008)	44008.47 (197.25)
3. Variable linkage	11	0.99845 (0.0008)	43999.97 (199.51)
4. Centroid linkage	11	0.99837 (0.0008)	43997.77 (196.17)
5. Variable linkage	12	0.99835 (0.0008)	43994.68 (201.67)
6. Average linkage	11	0.99830 (0.0009)	43997.57 (199.74)
7. Centroid linkage	12	0.99827 (0.0008)	44000.70 (196.79)
8. Variable linkage	13	0.99824 (0.0008)	44000.90 (203.48)
9. Average linkage	12	0.99820 (0.0009)	44000.38 (195.89)
10. Centroid linkage	13	0.99817 (0.0008)	43999.58 (203.50)
11. Average linkage	13	0.99806 (0.0009)	44000.55 (199.08)
12. Median linkage	13	0.96632 (0.0422)	41285.57 (3577.12)
13. Ward's method	12	0.96044 (0.0081)	44414.20 (204.37)
14. Beta-flexible	12	0.96024 (0.0093)	44386.65 (202.67)
15. Median linkage	12	0.95995 (0.0445)	40710.99 (3789.96)

* The table displays F-score (s.d.) and the corresponding clustering reliability (CR) (s.d.) values. The true number of clusters in the simulated data is 11.

appear to perform very well and exhibit similar F-scores and CR scores.

Directed by the selection of 11 clusters as the optimal number, the F-scores reveal marked differences between the

algorithms' performances (see Fig. 2). Along with the top four classifications overall, each of which contains 11 clusters, average linkage also performs extremely well. These classifications all have associated F-scores that exceed 0.998. In addition, median linkage and fuzzy *K*-means produce reasonably accurate classifications with respective F-scores 0.943 and 0.877, reflecting roughly 6 and 12% decreases from the highest possible F-score. Our graphic depiction of the fuzzy *K*-means classification for one data set (Fig. 3) suggests that F-scores in the range of median linkage and fuzzy *K*-means retain much of the visual quality displayed by the top-performing methods. The popular *K*-means method, single linkage, and complete linkage poorly classify measured brain activity into well-separated clusters. CR yields similar results to F-score. The only notable exception between the two performance measures, applied to the classifications with 11 groups, is that CR reverses the rankings of median linkage and fuzzy *K*-means.

Figure 3 illustrates differences between the algorithms in detecting functional classifications of measured brain activity for one simulated data set. The colors represent the within-cluster mean rCBF across scans, characterizing the level of brain activity in each cluster. The six images in the top panel of Figure 3 display the final partitions of six algorithms with 11 clusters, along with the defined clusters. Visually, the classifications that perform well in our analysis reveal few differences, which have little bearing on substantive conclusions. However, *K*-means lacks some of the detail of the other methods, e.g., in the cerebellum, and it reflects more noise than the algorithms that exhibit better performances. Single linkage essentially fails to recover any of the structure from the defined clusters. The low mean F-score for *K*-means with 11 clusters may relate to the algorithm's

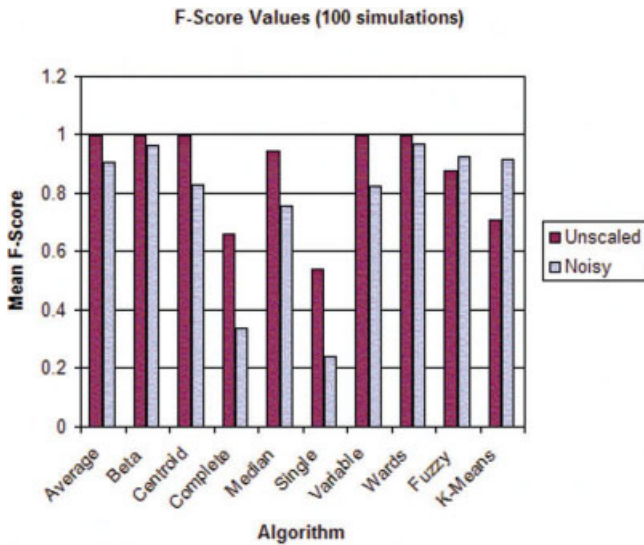


Figure 2.

The graph displays the quality (F-score) of the classifications with 11 clusters for 10 algorithms. The unscaled case (purple) is based on 100 simulated data sets with error variances that match estimates from an experimental PET study of ethanol. The noisy data case (gray) uses 100 simulated data sets with 125% increases in the error variances from the experimental data. Larger values of F-score represent higher quality classifications.

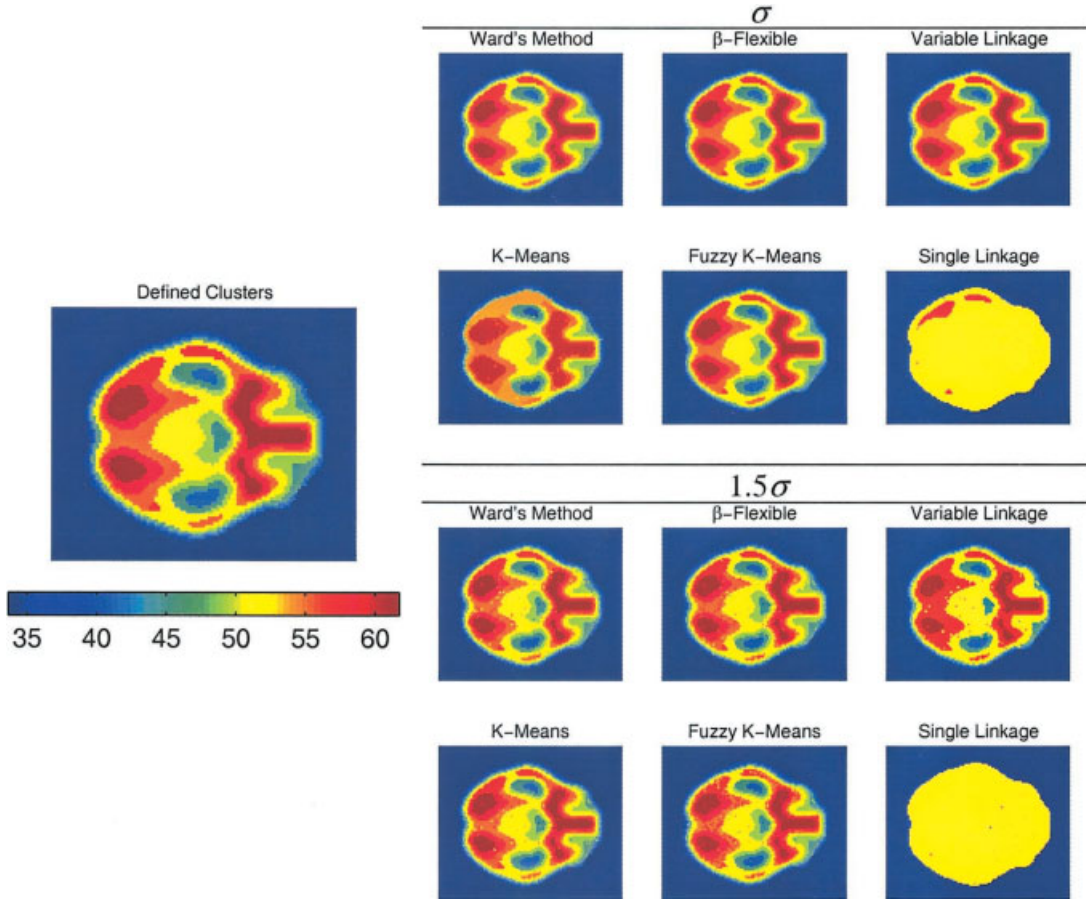


Figure 3. Final classifications, consisting of 11 clusters, for one simulated data set with unscaled error variances (**top**) and for one data set with inflated error variances (**bottom**). The defined clusters appear on the left. The colors represent the mean rCBF (across all scans) within each cluster.

tendency to underestimate the number of true clusters for these data, as illustrated in Table I. In fact, the largest F-score for *K*-means (0.80) occurs for $G = 7$ clusters.

Figure 4 displays mean profiles for the 11 clusters produced by the single linkage algorithm and Ward’s method, for the same simulated data set depicted in Figure 3. Ideally, clusters should consist of similar voxel-specific activity profiles within groups and should have distinct (well-separated) profiles between groups. As illustrated in Figure 4, the mean profiles from the single linkage algorithm reveal poor separation between clusters. In contrast, most of the clusters given by Ward’s method have well-separated mean profiles. Our simulation analysis groups voxel-specific profiles of summary statistics, and we display results for an entire axial slice. In practice, interest generally focuses on functional activity within particular brain regions or within a small subset of brain voxels that satisfy some initial criteria, and an analysis may target linear combinations, e.g., contrasts, of the rCBF profiles. Figure 4 highlights the mean trend in the most active cluster produced by Ward’s method, which exhibits a gradual decrease in activity over time.

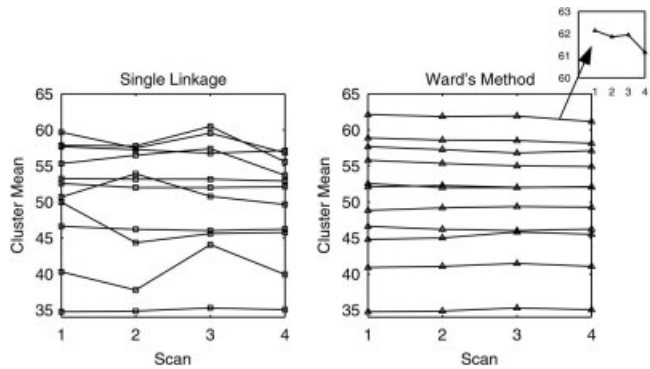


Figure 4. Mean profiles for the 11 clusters produced by the single linkage algorithm and Ward’s method. The single-linkage clusters exhibit poor separation between groups, in contrast to Ward’s procedure.

Classifying Noisy Data

To examine the sensitivity of the results to increased noise, we repeat our simulation experiment with a 125% increase in the error variances. Specifically, we draw data from multivariate normal distributions with variance parameters of the covariance matrices from the defined clusters increased from σ_g^2 to $(1.5\sigma_g)^2$. The focus of this sensitivity analysis is to examine the dependability of the clustering methods and stopping rules for artificially noisy data, but we expect the original simulation study to better reflect characteristics of experimental PET data. In practice, cluster analyses often classify summary statistics, which protects against noisy data by reducing variability present in the original images.

For the noisy data, the CCC stopping rule fails to identify the correct number of clusters. CCC consistently selects the optimal number of clusters as $G \leq 7$, for all algorithms. Pseudo- T^2 and pseudo-F criteria yield mixed results, with Ward's and beta-flexible methods continuing to perform very well, but the increased noise deteriorates the performances of several other methods. For both Ward's and beta-flexible methods, pseudo-F correctly detects 11 as the number of clusters in the data for all 100 simulation runs, and pseudo- T^2 identifies the correct number in 95 and 72 of the runs, respectively. Pseudo-F also works reasonably well for K -means, selecting 11 as the number of clusters in 57 of the simulation runs. The noise increase causes K -means to produce 11 clusters with more separation between the mean profiles than for the unscaled data, resulting in larger increases in between-cluster variation relative to the increases in within-cluster variation. Consequently, K -means does not consistently underestimate the number of true clusters as it does with the unscaled data, and it produces high-quality classifications with 11 clusters (see F-scores below). Average linkage has a mode at 11 clusters for both pseudo-F and pseudo- T^2 (34%), and average linkage results in 11 or 12 clusters in 62% of the simulations for pseudo-F and 66% for pseudo- T^2 .

Several of the clustering algorithms accurately identify classifications of brain activity from the noisy data. The highest quality partitions overall are from Ward's and beta-flexible methods with 11 clusters and have corresponding F-scores 0.967 and 0.964. Also several of the classifications with 11 or more clusters from average linkage, centroid linkage, variable linkage, Ward's, and beta-flexible methods yield competitive F-scores. The K -means and fuzzy K -means classifications with 11 clusters tied as the 15th best overall, both with F-scores of 0.923. Figure 2 shows a comparison of all 10 algorithms with 11 clusters for the noisy data and again reveals strong performances by Ward's and the beta-flexible methods, followed by K -means, fuzzy K -means, and average linkage. Outlying values have less of an impact on K -means and fuzzy K -means than for some hierarchical methods because these methods maintain a fixed number of clusters and can reallocate extreme values throughout the algorithm. The best classifications for average linkage, centroid linkage, and variable linkage algorithms contain more than 11 clusters. In addition to the 11 prominent clusters,

these partitions include a few small classes with extreme data that do not fit well within any of the other groups. Despite the additional clusters, these algorithms still capture the essence of the functional classification structure, as reflected by F-score values exceeding 0.941.

The final classifications of one noisy simulated data set appear in Figure 3 (six images in the bottom panel) for six of the algorithms with 11 clusters. With the exception of single linkage, the algorithms still recover much of the classification structure in the data, but the partitions also depict the increased noise for these data. For these particular data, Ward's, beta-flexible, K -means, and fuzzy K -means methods all appear to perform well, and variable linkage gives an informative classification, but the classification appears more crude. Single linkage, once again, exhibits very poor performance.

To gauge performances of the clustering algorithms under larger noise increases, where outliers are more probable, we examine the results from a single simulated data set with 250% increases in the error variances. The increased noise yields reduced F-scores ranging from 0.19 to 0.64 for classifications with 11 clusters. The noisy data produces singleton clusters for the single, complete, centroid, variable, median, and average linkage methods. Consistent with the findings from our full simulation studies, Ward's and the beta-flexible methods have the highest F-scores (0.64) and single and complete linkage exhibit the poorest performances (0.19). The K -means and average linkage methods also perform fairly well relative to the other algorithms with F-scores exceeding 0.59, followed by median and centroid linkage with F-scores above 0.55. To mitigate the impact of noise on classification procedures, it is often helpful to increase the number of clusters slightly from the targeted number, allowing outlying observations to form separate clusters (possibly singletons), while retaining the essential spatial patterns of the underlying functional classifications.

DISCUSSION

In this article, we introduce several clustering algorithms and stopping rules to the neuroimaging literature. We present simulation results that shed light on the relative performances of the clustering algorithms in context of data based on an in vivo PET study. Neuroimaging analyses tend to utilize K -means and fuzzy clustering algorithms, but our simulations direct attention to five hierarchical clustering methods that exhibit excellent performances relative to the other algorithms. We also examine the performances of three stopping rules and identify the stopping rules that work well with specific clustering algorithms.

Ward's and the beta-flexible methods both perform extremely well in our simulation study, along with average linkage, variable linkage, and centroid linkage. The strong performance of Ward's method, particularly for CR, relates to the fact that each level of the hierarchy joins clusters to maximize a quantity that is proportional to the exponential argument in a multivariate normal likelihood function. Clustering least-squares estimates from a GLM

will often lead to good performance by Ward's method, since these estimates asymptotically follow multivariate normal distributions. Despite desirable analytical properties of the single and complete linkage algorithms, such as invariance to any monotone transformation of the quadratic distances, both algorithms perform very poorly in our simulation study. Our study shows that the top-performing hierarchical algorithms also successfully classify noisy data, and an exploration of the algorithms' performances for data with substantial noise increases (250% increases) suggest that Ward's and beta-flexible methods continue to yield the best functional classifications. Our results are based on the use of a standard distance metric for cluster analysis that, in our setting, measures the dissimilarity in brain function between voxels or clusters. An interesting extension that may have appeal in functional neuroimaging applications is to consider a metric that incorporates the spatial proximity of voxels.

Hierarchical clustering circumvents a clear shortcoming of the K -means and fuzzy clustering procedures by not requiring advanced specification of the number of clusters. Hierarchical clustering provides the flexibility of "cutting the tree" of nested merges at varying levels after completion of the algorithm, resulting in different numbers of total groups. For our data, combining the stopping rules with the best hierarchical methods provides accurate functional classifications. Another limitation of K -means is that it may heavily depend on initial cluster assignments. A useful method is to examine the sensitivity of the resulting partitions to several different cluster initializations. The pseudo- T^2 , pseudo- F , and CCC stopping rules all perform well for the five best hierarchical algorithms. When we artificially increase the noise in our data by inflating the error variances by 125%, CCC fails to correctly identify the number of clusters. However, pseudo- T^2 and pseudo- F continue to provide good performances for beta-flexible and Ward's methods. Although we expect other PET neuroimaging applications to produce images much more like the unscaled simulated data than the noisy images, we still favor the pseudo- F and pseudo- T^2 criteria, given the apparent breakdown of CCC for the noisy data. Overall, we recommend the combination of either Ward's or the beta-flexible clustering algorithm with the pseudo- F or pseudo- T^2 stopping criterion for classifying PET neuroimaging data.

CR and F-score give very similar results, given a specified number of clusters. In the overall rankings, CR favors some classifications with larger numbers of clusters. Having fewer clusters produces larger cluster sizes, and the large cluster sizes dominate the model complexity penalty term in the likelihood function. One approach to mitigate the tendency of CR toward large numbers of clusters is to formulate a harsher penalty for increasing the number of clusters. Although this approach may reduce bias by directing the optimal classification toward a smaller number of clusters, it deviates from the original derivation of CR, which is based on the posterior proba-

bility of a partition. CR relates to methodology by Schwarz [1978] since it is expressible in terms of the Bayesian information criterion (BIC).

Given the reliance of clustering algorithms and stopping rules on specific data characteristics, potential limitations always exist when generalizing the findings of empirical performance evaluations. Nonetheless, our study takes several measures to increase the extent to which our results generalize to other PET neuroimaging applications. First, our simulations are based on data from an actual PET experiment, so that the location and scale (variability) of our simulated data are realistic. The amount of data and number of defined clusters, which both impact clustering computations, match a neuroimaging application. In addition, characteristics of the defined clusters resemble functional classifications determined from an *in vivo* study of the effects of ethanol on rCBF. We also evaluate the sensitivity of our results to noise increases and to an alternative selection of the number of true clusters, and our results regarding the best (and worst) performing methods appear to be robust to these various specifications. Therefore, the findings of our simulation study should be informative for PET neuroimaging applications, although they may not generalize to all settings.

REFERENCES

- Balslev D, Nielsen FA, Frutiger SA, Sidtis JJ, Christiansen TB, Svarer C, Strother SC, Rottenberg DA, Hansen LK, Paulson OB, Law I (2002): Cluster analysis of activity-time series in motor learning. *Hum Brain Mapp* 15:135–145.
- Baumgartner R, Ryner L, Richter W, Summers R, Jarmasz M, Somorjai R (2000): Comparison of two exploratory data analysis methods for fMRI: fuzzy clustering vs. principal component analysis. *Magn Reson Imag* 18:89–94.
- Baune A, Sommer FT, Erb M, Wildgruber D, Kardatzki B, Palm G, Grodd W (1999): Dynamical cluster analysis of cortical fMRI activation. *Neuroimage* 9:477–489.
- Bezdek J, Ehrlich R, Full W (1984): FCM: the fuzzy c-means clustering algorithm. *Comp Geosci* 10:191–203.
- Bowman FD, Patel R (2004): Identifying spatial relationships in neural processing using a multiple classification approach. *Neuroimage* (in press).
- Brodmann K (1909): Vergleichende Lokalisationlehre der Grosshirnrinde in ihren Prinzipien Dargestellt auf Grund des Zellenbaues. [Reprinted as: Garey LJ, translator. Brodmann's "Localisation in the cerebral cortex." London: Smith-Gordon; 1994.]
- Calinski RB, Harabasz J (1974): A dendrite method for cluster analysis. *Comm Stat* 3:1–27.
- Carlin BP, Louis TL. (2000): Bayes and empirical Bayes methods for data analysis, 2nd ed. New York: Chapman and Hall; p.40–42.
- Cordes D, Haughton V, Carew JD, Arfanakis K, Maravilla K (2002): Hierarchical clustering to measure connectivity in fMRI resting-state data. *Magn Reson Imag* 20:305–317.
- Duda RO, Hart PE. 1973. Pattern classification and scene analysis. New York: John Wiley and Sons.
- Fadili MJ, Ruan S, Bloyet D, Mazoyer B (2000): A multistep unsupervised fuzzy clustering analysis of fMRI time series. *Hum Brain Mapp* 10:160–178.

- Fadili MJ, Ruan S, Bloyet D, Mazoyer B (2001): On the number of clusters and the fuzziness index for unsupervised FCA application to BOLD fMRI time series. *Medical Image Analysis* 5:55–67.
- Filzmoser P, Baumgartner R, Moser E (1999): A hierarchical clustering method for analyzing functional MR images. *Magn Reson Imag* 17:817–826.
- Goutte C, Hansen LK, Liptrot MG, Rostrup E (2001): Feature-space clustering for fMRI meta-analysis. *Hum Brain Mapp* 13:165–183.
- Goutte C, Toft P, Rostrup E, Nielsen FA, Hansen LK (1999): On clustering fMRI time series. *Neuroimage* 9:298–310.
- Lance GN, Williams, WT (1967): A general theory of classificatory sorting strategies: I. Hierarchical Systems. *Comput J* 9:373–380.
- Larsen B, Aone C (1999): Fast and effective text mining using linear-time document clustering. In: *Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, p 16–22.
- MacQueen J (1967): Some methods for classification and analysis of multivariate observations. In: Le Cam LM, Neyman J, editors. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*. Berkeley, CA: University of California Press. p 281–297.
- Maitra R (1997): Estimating precision in functional images. *J Comput Graph Stat* 6:132–142.
- Milligan GW, Cooper MC (1985): An examination of procedures for determining the number of clusters in a data set. *Psychometrika* 50:159–179.
- Rencher A. (2002): *Methods of multivariate analysis*, 2nd ed. New York: John Wiley and Sons. p 470.
- Sarle WS (1983): *Cubic clustering criterion*, SAS Technical Report A-108. Cary, NC: SAS Institute Inc.
- Schwarz G (1978): Estimating the dimension of a model. *Ann Stat* 6:461–464.
- Sommer FT, Wichert A (2003): *Exploratory analysis and data modeling in functional neuroimaging*. Cambridge, MA: The MIT Press. p 17–42.
- Stanberry L, Nandy R, Cordes D (2003): Cluster analysis of fMRI data using dendrogram sharpening. *Hum Brain Mapp* 20:201–219.
- Wong MA, Lane T (1983): A kth nearest neighbor clustering procedure. *J R Stat Soc B* 45:362–368.
- Zhu C, Jiang T (2003): Multi-context fuzzy clustering for separation of brain tissues in MRI images. *Neuroimage* 18:685–696.