ELSEVIER

NeuroImage

www.elsevier.com/locate/ynimg
NeuroImage 23 (2004) 260–268

Identifying spatial relationships in neural processing using a multiple classification approach

F. DuBois Bowman* and Rajan Patel

Department of Biostatistics, Rollins School of Public Health, Emory University, Atlanta, GA 30322, USA

Received 11 February 2004; revised 4 April 2004; accepted 21 April 2004

The application of statistical classification methods to in vivo functional neuroimaging data makes it possible to explore spatial patterns in task-related changes in neural processing. Cluster analysis is one group of descriptive statistical procedures that can assist in identifying classes of brain regions that exhibit similar task-related functionality. In practice, a limitation of cluster analysis is that the performances of clustering algorithms rely on unknown characteristics of the data, making it difficult to determine which procedure best suits a particular analysis. We present a multiple classification approach that incorporates numerous algorithms, evaluates the associated classifications, and either selects a plausible partition relative to the others considered or pools the results from the numerous methods. The multiple classification approach utilizes a new performance criterion, called the relative information (RI) measure, to evaluate the quality of the candidate partitions and as the basis for producing a composite classification image. Employing multiple classifications, rather than a single algorithm, our methodology increases the chance of detecting the functional relationships within the data and, therefore, produces more reliable results. We apply our methodology to a PET study to explore spatial relationships in measured brain function associated with increasing blood alcohol concentration levels, and we perform a simulation study to evaluate the performance of RI.

Keywords: Evaluative classification; Aggregate classification; Relative information measure; Clustering; PET; Ethanol

Introduction

The neural processing associated with specific behaviors, cognitions, and emotions often relies on the interplay between spatially distinct brain regions. Vast systems of anatomical connections allow communication between areas of the cerebral cortex (Hendelman, 2000), with white matter association bundles joining cortical areas within the same hemisphere and commissural bundles linking regions in separate hemispheres. The enormous number of axonal pathways in the human brain and the complexity of these systems make it difficult to succinctly describe spatial associations in brain function. Additionally, the anatomical circuitry does not directly uncover functional relationships, which may be modified or initiated by specific behaviors. Applying statistical classification methods to in vivo positron emission tomography (PET) or functional magnetic resonance imaging (fMRI) data makes it possible to examine spatial relationships between correlates of blood flow or metabolic activity in the brain.

Cluster analysis is a class of data-driven statistical procedures that can assist in identifying functionally related brain regions. Other methods such as partial least squares (McIntosh et al., 1996) as well as exploratory statistical procedures such as principal component analysis and independent component analysis are also useful for identifying functional classifications (Baumgartner et al., 2000; Friston et al., 1993; Frutiger et al., 2000; McKeown et al., 1998; Thireou et al., 2001). Clustering has a long history in the statistical literature and has recent applications in functional neuroimaging. Neuroimaging applications typically utilize the *K* means (Balslev et al., 2002; Goutte et al., 1999, 2001; MacQueen, 1967) or fuzzy clustering (Baumgartner et al., 2000; Fadili et al., 2000, 2001; Somorjai and Jarmasz, 2003) approaches. Select hierarchical clustering methods also appear in the neuroimaging literature including single, complete, and average linkage methods and a hybrid hierarchical *K* means approach (Cordes et al., 2002; Filzmoser et al., 1999; Goutte et al., 1999; Stanberry et al., 2003).

To validate the performance of a cluster analysis, it is important to evaluate the quality of the resulting classification. Fadili et al. (2001) discuss several methods of examining cluster validity for fuzzy clustering applications. The *F* score statistic is a more general measure of classification quality that compares an algorithm's computed clusters to the true group memberships (Larsen and Aone, 1999). The true clusters are not available in neuroimaging studies, so the *F* score has limited utility in practice.

An algorithm's performance partially depends on whether the data contain compact clusters, elongated clusters, balanced clusters, clusters with equal variances, or clusters with other specific characteristics. Because these data characteristics are generally unknown in neuroimaging analyses, it is challenging to determine the best clustering method to employ for a particular application.

* Corresponding author. Department of Biostatistics, The Rollins School of Public Health, Emory University, 1518 Clifton Road, N.E., Atlanta, GA 30322. Fax: +1-404-727-1370.

E-mail address: dbowma3@sph.emory.edu (F. DuBois Bowman).

Available online on ScienceDirect (www.sciencedirect.com.)

1053-8119/$ - see front matter © 2004 Elsevier Inc. All rights reserved.
doi:10.1016/j.neuroimage.2004.04.022

Bowman et al. (in press) conduct an empirical comparison of numerous hierarchical and partition-based clustering algorithms. Their findings provide some guidance for advanced selection of a particular clustering method, but the performance of clustering algorithms may vary for different data applications. There are potential advantages of executing multiple methods, increasing the chance that at least one resulting classification effectively detects specific aspects of the underlying clusters of unknown size, shape, and dispersion. When considering multiple clustering algorithms in practice, how does one either select a credible clustering solution, given the data, or combine results to produce a composite classification?

We take a multiple classification approach that begins with a collection of clustering methods and either performs a selection process, yielding a plausible partition, or aggregates results from the various procedures. The multiple classification approach entails an evaluation of the quality of each resulting partition. We present a new performance evaluation criterion to identify classifications that provide a good fit to the data. The performance criterion is applicable to a range of clustering algorithms and is useful in practice because, unlike benchmark measures, it does not require knowledge of the true cluster memberships. We illustrate our methods using a PET neuroimaging study that examines neural correlates of drinking alcohol.

Methods

Notation

We apply the multiple classification approach to spatially organize changes in neural processing associated with the use of ethanol. There are 10 subjects, indexed by $k = 1, \ldots, K$, and we consider four scans for each individual, indexed by $s = 1, \ldots, S$. Injections of a low dose of ethanol and a high dose of ethanol preceded the second and third scans, respectively, resulting in increasing blood alcohol concentration levels across the scans. Subjects completed continuous performance tasks after each scan, giving behavioral assessments of attention. PET images for each subject were aligned and resliced (Woods et al., 1998a) and spatially normalized to a population-representative PET atlas (Woods et al., 1998b) centered in Talairach stereotaxic coordinates (Talairach and Tournoux, 1988).

Let $\boldsymbol{Y}_k(v) = (Y_{k1}(v), \ldots, Y_{kS}(v))'$ represent a vector of regional cerebral blood flow (rCBF) measurements for the kth subject at voxel v, where v=1, …, V. Also, let $\boldsymbol{X}_k$ $(S \times q)$ denote a matrix of design variables with columns corresponding to the scan numbers or to polynomial trends over time (scans) and a final column containing a covariate adjustment for global cerebral blood flow (gCBF). A clustering procedure, indexed by i, produces a classification $(\Re_i, G_i)$ (or simply $\Re_i$) that maps the V voxels from an image into G_i clusters. The set Θ_{g_i} represents the collection of V_{g_i} voxels comprising a cluster g_i, where $V = \sum_{g_i=1}^{G_i} V_{g_i}$.

Summary statistics for multisubject studies

Many neuroimaging clustering applications focus on single-subject analyses, although the concepts extend to studies with multiple subjects. To apply cluster analyses to multisubject studies, we formulate a statistic $\boldsymbol{T}(v) = (T_1(v), \ldots, T_P(v))' = f(\boldsymbol{Y}_1(v), \ldots, \boldsymbol{Y}_K(v))$, $P \leq S$, which summarizes the raw data vectors across individuals. Perhaps the simplest procedure is to compute voxel-specific mean functions

$$T_s(v) = K^{-1}\sum_{k=1}^{K} Y_{ks}(v), \qquad (1)$$

where $P = S$. Balslev et al. (2002) applied this method in a cluster analysis of 18 subjects. Extending the raw means approach, we cluster voxel-specific linear combinations of estimated regression coefficients from our data. We obtain ordinary least squares (OLS) estimates of the regression parameters from the general linear model (GLM)

$$\boldsymbol{Y}_k(v) = \boldsymbol{X}_k\beta(v) + \varepsilon_k(v), \qquad (2)$$

where $\beta(v)(q \times 1)$ is the unknown parameter vector at voxel v relating design variables and covariates to rCBF, and $\varepsilon_k(v)$ contains mean-zero random errors. We utilize quantities $\boldsymbol{T}(v) = \boldsymbol{C}\hat{\beta}(v)$ to classify measured brain activity, where $\boldsymbol{C}(P \times q)$ is a matrix of specified constants.

We focus on PET, in light of our ethanol data example, but similar extensions for model-based summary statistics apply to fMRI studies. Viewing a localized fMRI time series from one individual as a functional datum and the activation stimuli in an analogous functional representation, a useful approach for summarizing data across subjects is the functional linear model of Ramsay and Silverman (1997). Extended linear models are also available for PET and fMRI that incorporate temporal correlations into the estimation of summary statistics (Bowman and Kilts, 2003; Friston et al., 1995; Worsley and Friston, 1995), but simplified methods such as OLS typically suffice for our subsequent descriptive statistical procedures.

Clustering methods

Using distances between rCBF measurements from pairs of voxels (or clusters), clustering algorithms attempt to spatially organize voxels into G well-separated classes that exhibit similar patterns of brain activity within groups. The typical measure of distance between a pair of voxels (v_i, v_j) is

$$d(\boldsymbol{T}(v_i),\boldsymbol{T}(v_j)) = \left[(\boldsymbol{T}(v_i)-\boldsymbol{T}(v_j))'\,\boldsymbol{B}_{v_iv_j}(\boldsymbol{T}(v_i)-\boldsymbol{T}(v_j))\right]^{1/2}, \qquad (3)$$

where $\boldsymbol{B}_{v_ivj}$ is a $P \times P$ positive definite matrix, often selected as an inverse covariance matrix or the identity matrix. We propose an analysis that incorporates numerous clustering algorithms, including hierarchical procedures and the popular partition-based K means and fuzzy clustering methods. Specifically, the hierarchical procedures that we employ are single linkage, complete linkage, average linkage, centroid, median linkage, Ward's minimum variance, beta-flexible, and variable linkage algorithms, several of which do not commonly appear in neuroimaging applications. K means and fuzzy clustering procedures require advanced specification of G, while hierarchical methods allow selection of G following completion of the algorithm.

Hierarchical clustering algorithms begin with each voxel representing a separate cluster, and these procedures sequentially merge the most similar clusters until forming a single group. The distinctions between the algorithms stem from the particular method that each uses to determine the clusters with the smallest corresponding distance, i.e., the clusters that exhibit the most similar patterns of neural processing. For example, the centroid method measures the distance between two clusters g and g* using

$$d(g,g^{*}) = d\left(V_{g}^{-1}\sum_{v\in\Theta_{g}}\boldsymbol{T}(v),\, V_{g^{*}}^{-1}\sum_{v^{*}\in\Theta_{g^{*}}}\boldsymbol{T}(v^{*})\right), \tag{4}$$

and the beta-flexible method updates the matrix of pairwise distances at each iteration using

$$d(g,g^{*}) = (1-\beta)\left[\frac{1}{2}(d(g,g_i)+d(g,g_j))\right]+\beta d(g_i,g_j), \tag{5}$$

after merging clusters g_i and g_j to form g*, where $\beta \in [-1,1]$. Variable linkage is a kth nearest neighbor clustering procedure (Wong and Lane, 1983). We modify the general kth nearest neighbor approach by specifying $k = \alpha(V_gV_g*)$, where α is a constant in the interval $[(V_gV_g*)^{-1}, 1]$, so k increases proportionally with the product of the sizes of the two joining clusters (Bowman et al., in press). For the analyses presented in this paper, we use $\beta = -0.5$ for the beta-flexible method and $\alpha = 0.15$ for variable linkage. Additional details about the clustering methods used in our analysis are available in numerous references including Rencher (2002), Hartigan (1975), and Milligan and Cooper (1985).

Classification quality

Many cluster characteristics generally improve with the addition of new clusters. For example, designating each voxel as its own cluster minimizes the within-group inertia, which provides a measure of intracluster variability (Goutte et al., 1999). We present a relative measure of cluster quality that evaluates the goodness-of-fit of each classification, while penalizing for the addition of new clusters, and it combines with statistics for selecting the optimal number of clusters such as pseudo-T^2 (Duda and Hart, 1973), pseudo-F (Calinski and Harabasz, 1974), and the cubic clustering criterion (CCC) (Sarle, 1983).

As a general probability model for the data, we assume that given $\Re_i$, the summary statistic $T(v)$ arises from a distribution with density function

$$f(\boldsymbol{T}(v)\,|\,\Re_i,\tau^{(i)}) = \sum_{g_i=1}^{G_i} f_{g_i}(\boldsymbol{T}(v)\,|\,\tau_{g_i})\mathrm{I}(v\in\Theta_{g_i}), \tag{6}$$

where $\tau^{(\mathrm{i})} = (\tau_1, \ldots, \tau_{G_i})$ represents a vector of parameters corresponding to $\Re_i$, the indicator function $\mathrm{I}(v \in \Theta_{g_i})$ equals 1 if $v \in \Theta_{g_i}$ and zero otherwise, and f_{g_i} is the component density for cluster g_i. The density suggests that measurements from voxels in different clusters have distinct underlying probability models. The relative information (RI) associated with a partition $\Re_i$, among J competing classifications, is given by

$$\eta_i = 1-\frac{\omega_i+\psi_i}{\min\limits_{j=1,\ldots,J}(\omega_j+\psi_j)}, \tag{7}$$

where

$$\omega_i = G_i\log\theta-\log(\exp(\theta)-1)-\log(G_i!)-\log J_{G_i}, \tag{8}$$

with J_{G_i} representing the number of competing classifications with G_i groups and θ denoting the anticipated number of clusters, and

$$\psi_i = \sum_{v=1}^{V}\log\left[f\left(\boldsymbol{T}(v)\,|\,\Re_i,\hat{\tau}^{(i)}\right)\right]-\frac{1}{2}[G_i(q+1)\Big]\log(qV). \tag{9}$$

RI measures the proportion increase in the log of the classification probability, relative to the clustering procedure yielding the least probable partition, and assumes that the classification probability is nonzero. The quantity ψ_i can be regarded as a function of the likelihood of the estimated parameter values, given the data, penalized for the number of clusters G_i. ω_i involves the chance of obtaining a particular classification and is based on a discrete uniform probability model and a left-truncated Poisson model with parameter θ. Additional details are available in Appendix A. Methods for determining the optimal number of clusters, as well as biological considerations, should drive the selection of θ. Values of RI are in the range $0 \le \eta_i \le 1$, and larger values indicate higher classification quality, relative to a reference clustering solution.

Multiple classification approach

The multiple classification approach produces a set of J classifications, assesses the relative performances based on η_i, and either conducts an evaluative procedure that selects a plausible partition or performs an aggregation process that yields a quality-weighted composite classification map. We include an assessment of the optimal number of clusters in the evaluative clustering method. The final classification given by the evaluative method is deemed probable among the competing clustering solutions. The aggregate clustering procedure pools the J classifications using RI to apply relative weights. The composite classification contains an average vector $\gamma(v)$ at each voxel, defined as

$$\gamma(v) = \sum_{j=1}^{J}\frac{\eta_j\bar{\boldsymbol{T}}_{g_j}(v)}{\bar{\eta}J}, \tag{10}$$

where $v \in \Theta_{g_j}$ and $\bar{\boldsymbol{T}}_{g_j}(v) = V_{g_j}^{-1}\sum_{v*\in\Theta_{g_j}}\boldsymbol{T}(v^{*})$ is the average summary vector for cluster g from the jth clustering procedure. Voxels in the composite classification map contain weighted averages of cluster memberships (means), rather than discrete classifications identifying clusters.

Data example and results

Multiple classification approach

We apply the multiple classification approach to the PET ethanol study to identify classifications of neural processing associated with changes in blood alcohol concentrations. First, we model rCBF from the ethanol data using a GLM with separate means for each of the four scans, and we adjust for gCBF by including a mean-centered global flow value in the final column of the design matrix. We obtain least squares estimates of the fixed-effects parameters $\hat{\beta}(v)$ and associated variance estimates

$\hat{\sigma}_v^2$ $(\boldsymbol{X}'\,\boldsymbol{X})^{-1}$. We define a contrast vector $\boldsymbol{C}$ to obtain summary statistics $T(v) = \hat{\beta}_1(v) - \hat{\beta}_4\,(v)$ at each voxel, and we incorporate $T(v)$ $T(v)$ into the multiple classification approach.

In the clustering procedures, we compute distances using Eq. (3), measuring the dissimilarity in ethanol-induced changes in brain activity between two locations (voxels). We normalize the distance between two voxels using $B_{v_1 v_2} = (\sigma_{v_1}^2 + \sigma_{v_2}^2)^{-1} H$, the distance between two cluster centroids (for clusters g and g*) using

$$B_{gg^*} = \left(V_g^{-2} \sum_{i \in \Theta_g} \sigma_i^2 + V_{g^*}^{-2} \sum_{j \in \Theta_{g^*}} \sigma_j^2 \right)^{-1} H, \tag{11}$$

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

$$B_{vg} = \left(a\sigma_v^2 + \sum_{i \in \Theta_g} \frac{\sigma_i^2}{V_g^2} \right)^{-1} H, \text{ where} \begin{cases} \text{a} = 1, & \text{if } v \notin \Theta_g \\ \text{a} = (1 - 2\text{V}_\text{g}^{-1}), & \text{if } \nu \in \Theta_g \end{cases} \tag{12}$$

and $H = [\boldsymbol{C}(\boldsymbol{X}'\,\boldsymbol{X})^{-1}\boldsymbol{C}'\,]^{-1}$. In each case, the normalizing constant is interpretable as the inverse variance of the associated difference $T(v_1) - T(v_2)$.

Before clustering, we explore the number of groups in the data using the CCC, pseudo-T^2, and pseudo-F statistics. Fig. 1 displays CCC for Ward's minimum variance clustering procedure. The figure strongly suggests that G = 15 clusters should adequately describe the data, indicated by the large peak relative to the other values, and similar CCC plots for the other algorithms (not shown) corroborate our selection of 15 clusters. The multiple classification approach permits a range of cluster numbers in the analysis. Our analysis includes 10 clustering procedures, and we let the number of clusters range from 12 to 18, giving a total of 70 classifications. We consider a subset of axial slices ranging from −20 to +28 mm. To facilitate computations, one can implement data reductions by applying thresholding procedures such as those proposed by Goutte et al. (1999), Balslev et al. (2002), or Fadili et al. (2000).

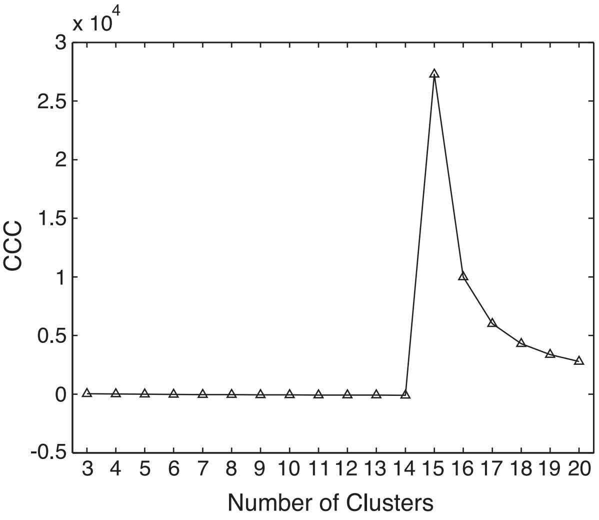

Fig. 1. The line profile shows the cubic clustering criterion (CCC) for Ward's minimum variance method, with the number of clusters ranging from 3 to 20. CCC suggests the use of 15 clusters.

Table 1
Classification performances and composite weights

Algorithm	RI	AC weights
K means	0.2977	0.019693
Fuzzy clustering	0.2877	0.019031
Beta-flexible	0.2834	0.018749
Ward's method	0.2764	0.018281
Average linkage	0.2616	0.017304
Centroid linkage	0.2562	0.016946
Median linkage	0.2554	0.016892
Variable linkage	0.2419	0.016004
Single linkage	0.0015	0.000102
Complete linkage	0.0012	0.000079

Relative information (RI) for classifications with G = 15 clusters and the aggregate classification (AC) weights (among 70 total) for pooling across the classifications.

We assume Gaussian probability distributions for each f_{g_i} in Eq. (4) to calculate RI, but our methodology permits the use of alternative probability models. Table 1 displays RI values for all of the classifications with G = 15 clusters. The RI index gives a measure of relative fit among all classifications considered, and we use complete linkage with G = 12 clusters as the reference classification because it performs worst overall. Therefore, RI indicates the extent to which a classification improves the fit for the data, relative to the reference. Table 1 shows that several procedures with G = 15 clusters provide substantially better classifications than the reference, including K means, fuzzy clustering, beta-flexible, and Ward's methods. Generally, the classifications with G = 15 or more clusters provide the best relative fits. The evaluative classification procedure selects among the best-performing partitions. With such consistent performances by several of the algorithms, any of the top four methods should reveal reliable classifications of changes in blood flow associated with drinking alcohol, relative to the reference.

Fig. 2(a) displays the four clusters from Ward's method that exhibit the most notable ethanol-related declines in rCBF as well as a cluster that shows an increase in brain activity. The colors in the figure correspond to cluster-specific mean changes in rCBF from scan 1 to scan 4, so positive values indicate that brain activity decreases with rising blood alcohol concentration levels, and negative values reveal increases in brain activity. The mean changes (standard errors) in the four clusters with decreased brain activity are 1.55 (0.18), 1.26 (0.06), 1.07 (0.07), and 0.79 (0.06), and the cluster with increased activity has mean −1.0350 (0.07). Fig. 3 presents mean changes in measured brain activity for all 15 clusters produced by Ward's method.

As illustrated in Fig. 2, functional similarities exist between the anterior and posterior cingulate, which both belong to the cluster that reveals the largest decrease in brain activity (dark red). The mean rCBF change in this cluster suggests that ethanol diminishes activity within a functional structure that plays an important role in shifting attention and transitioning from one idea to another. The cluster that exhibits the second largest decrease in brain activity (red-orange) includes regions of the cerebellum, the right middle temporal gyrus [Brodmann area (BA) 21], and left inferior frontal gyrus (BA 46 and 47). The maps also portray a cluster with declining brain activity (orange) that reveals functional connections between a region of the right postcentral gyrus (BA 40) and portions of the cuneous, the lingual gyrus (BA 18), and the cerebellum. This cluster also extends to the parahippocampal gyrus, which is thought to play an important role in limbic function and memory and is known to have widespread white matter connections with many areas of the cerebral cortex (Hendelman, 2000). The cluster showing the fourth largest ethanol-related decrease in rCBF (yellow) primarily consists of voxels in the cerebellum, the right thalamus, and the right inferior frontal gyrus (BA 47), with sparse inclusion of voxels in the right middle frontal gyrus (BA 10), and superior frontal gyrus (BA 9) (from axial views higher than +20 mm—not shown). The functional connection between the cerebellum and the thalamus uncovered in the ethanol data may result from the fact that the cerebellum projects fibers to the cerebral cortex via the thalamus.

The blue cluster represents a functional network that exhibits an increase in brain activity from scan 1 to scan 4. This cluster primarily consists of voxels within the middle and superior frontal gyri, an area that borders a lateral section of the left cerebellum and extends into the temporal lobe, and a posterior region of the right thalamus. Intriguingly, clusters that show decreased activity and the cluster that shows increased activity include common anatomical structures. One possible explanation is that an anatomical structure such as the thalamus contains a multitude of white matter projections to various cortical areas, which may lead to distinct functional connections with other areas. Potentially, an alternative explanation is that localized changes may induce compensatory alterations in neighboring voxels, given the spatial proximity of the regions and the overall constraints on global blood flow.

The cerebellum primarily belongs to clusters that show ethanol-related decreases in brain activity. The cerebellum generally helps to control and coordinate familiar movements. Cerebellar dysfunction leads to difficulty performing basic motor tasks such as walking in a straight line, finger-to-nose tests, eye movements, and complex movements of lips, cheeks, and tongue involved in speech. Although drinking alcohol is known to result in difficulty performing such behavioral tasks, our findings suggest that these

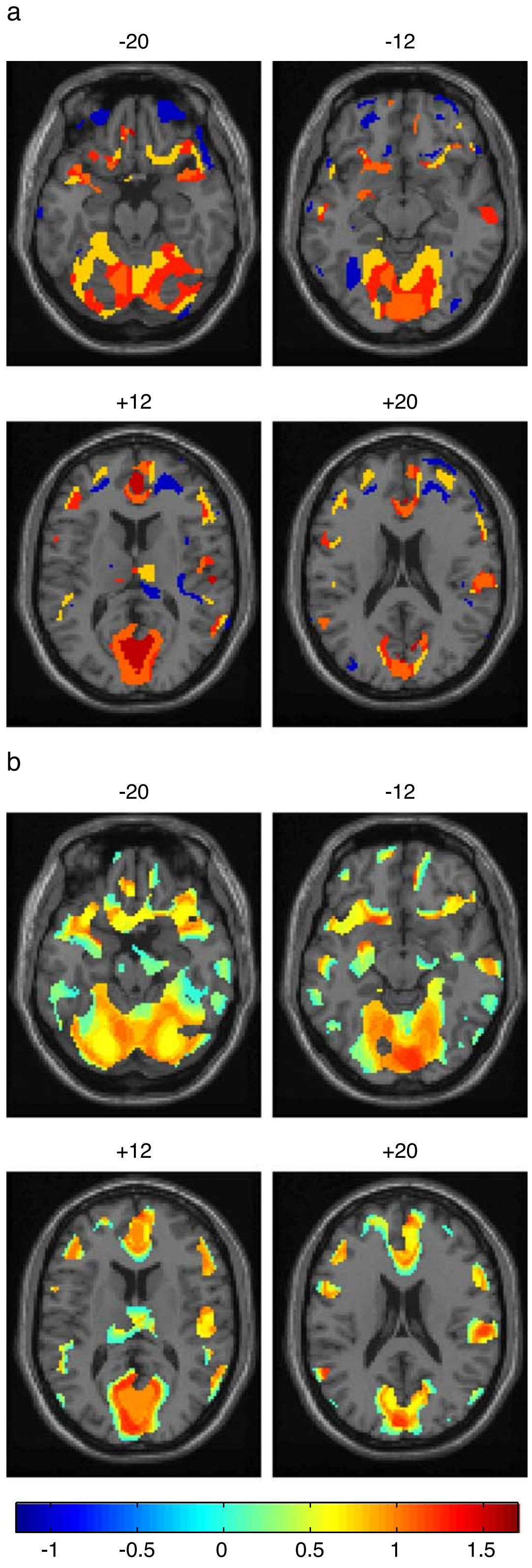

Fig. 2. Cluster maps (a) from Ward's method, illustrating results from the evaluative classification procedure and (b) from a composite classification pooling across 70 clustering solutions. The maps display mean differences from scan 1 to scan 4 (scan 1 – scan 4) for four axial slices corresponding to − 20, − 12, + 12, and + 20 mm. The maps in (a) depict the four clusters with the largest ethanol-related decreases in measured brain activity and a cluster that exhibits increased activity. The maps in (b) show areas with associated declines in measured activity.

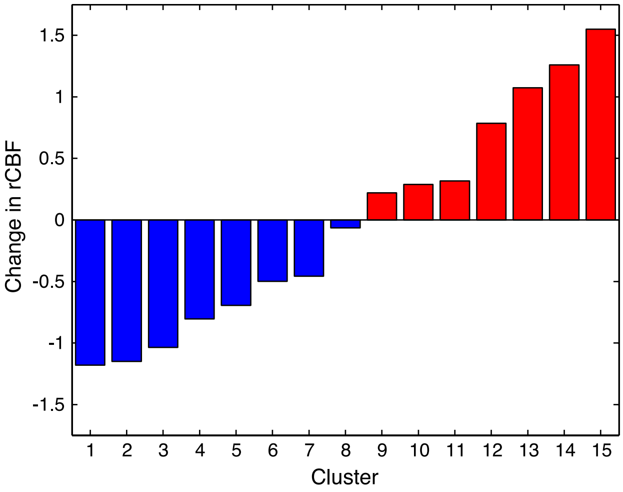

Fig. 3. Display of the mean change in rCBF for each of the 15 clusters produced by Ward's method. The mean change corresponds to scan 1 – scan 4, so that positive values (red) indicate declines in brain activity and negative values (blue) represent increases in brain activity associated with increased blood alcohol concentration levels.

deficiencies may directly relate to compromised cerebellar activity associated with increased blood alcohol concentration levels. Also, the decreased activity observed in the posterior cingulate and lateral portions of the middle and inferior frontal gyri may reflect the negative effects of alcohol on self-monitoring and self-awareness as well as executive cognitive functioning and decision making.

Fig. 2(b) displays results from the aggregate classification procedure. Applying the multiple classification approach in practice typically involves the use of either the evaluative clustering procedure, producing maps similar to Fig. 2(a), or the composite classification procedure, yielding maps similar to Fig. 2(b). We illustrate both procedures here, but generally the objectives of the analysis will help guide the choice of methods. For example, the evaluative approach may be preferable for studies that aim to identify a single cluster for each voxel, whereas the composite approach is better suited for settings where discrete classifications are not required.

The composite classification maps in Fig. 2(b) pool across all 70 clustering solutions obtained by our analyses, with weights ranging from 0 to 0.0207 applied to the individual classifications. The weights for classifications with $G = 15$ clusters appear in Table 1. The weights in Table 1 are fairly even across all algorithms, except for single and complete linkage, because most of the classifications with 15 clusters provide similar improvements in fit relative to the reference. The composite classification image shows voxels with positive weighted means across all clustering solutions. Fig. 2(b) reflects the same basic structure as that depicted in Fig. 2(a), but with coverage that is slightly more spatially expansive. The weighted clusters provide views of functional classifications with slightly smoothed edges. Smoothing arises because the cluster mean for a particular voxel is a linear combination of the means from all of the computed classifications, rather than a single number obtained from one classification.

The multiple classification approach provides increased assurance of obtaining a plausible classification from the data. In contrast to the functional organization identified by the multiple classification approach, an analysis solely based on the single linkage or complete linkage algorithm essentially fails to extract any functional patterns from the data. The $G = 15$ clusters from single linkage appear spatially as three separate groups (not shown) and essentially contain information that reflects three clusters (see Fig. 4). One cluster spans nearly the entire volume and contains 99.4% of the brain voxels within the slice range (−20 to +28 mm). There are three other relatively small clusters and all other clusters are singletons. Fig. 4(a) displays distances between the single linkage cluster centroids, illustrating the poor separation between some of the clusters. Cluster 10 in the diagram is the largest cluster and appears distinct from the other groups. This cluster essentially exhibits no change (−0.01) in rCBF from scan 1 to scan 4. Clusters 1 through 9 have similar characteristics, with respect to the ethanol-related change in blood flow, and show very little separation between the cluster means. Similarly, clusters 11 through 15 (especially 12 through 15) show similar changes in activity. Fig. 4(b) reveals much more balanced separation between the clusters produced by Ward's method.

Performance of the relative information measure

To assess the performance of RI, we use a simulation study to compare it to two benchmark methods, *F* score and clustering reliability (CR) (Bowman et al., in press), both of which rely on knowledge of the true classifications. *F* score attempts to maximize the overlap between the true and computed classifications, and RI and CR both target the most plausible classification, given the data. In the simulation study, we define clusters by employing *K* means to organize the voxel-specific OLS vectors into 11 groups. We simulate 100 data sets, with values in each voxel drawn from a

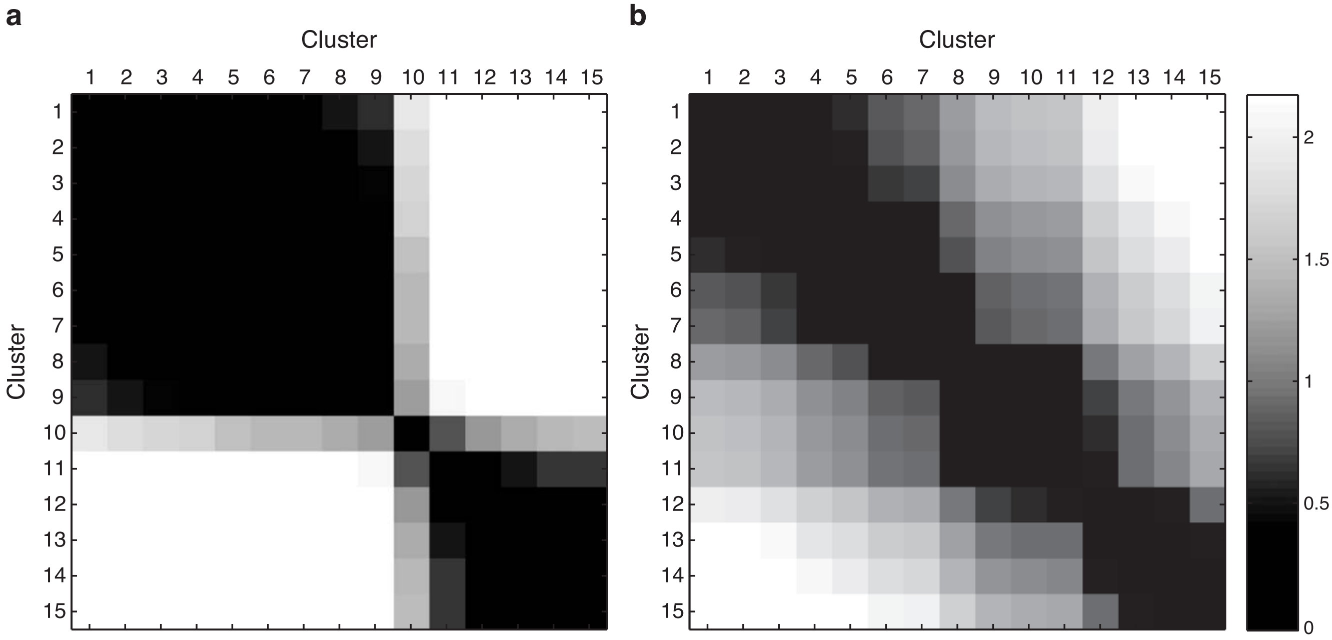

Fig. 4. Distance matrices between cluster means for (a) single linkage and (b) Ward's minimum variance methods. Row and column labels represent cluster numbers, arranged in increasing order of the cluster centroids. Each intensity element indicates the separation between the clusters in the corresponding row and column.

multivariate Gaussian distribution with the mean vector and covariance matrix matching the corresponding cluster-specific values in the experimental data.

The simulation results reveal close correspondence between the RI and both *F* score and CR clustering performance measures (see Table 2 and Fig. 5). To match the range of the other performance measures, we scale CR using [CR − min(CR)] / [max(CR) − min(CR)], preserving the ranks of the classifications on the original scale. For all three measures, larger values indicate higher quality classifications. Table 2 reveals the small discrepancies between RI and the two benchmark measures. The only notable difference between the performance measures arises for the rankings of median linkage and fuzzy clustering methods. As revealed by Fig. 5, the differences are slight, and RI aligns with the CR ranking, which assigns a higher value to the fuzzy clustering classification. All other discrepancies between the classification quality measures are negligible. The performances of the top five algorithms tend to vary much less between simulation runs than the performances of the other procedures. While *F* score and CR depend on knowledge of the true clusters, the close agreement with RI suggests that RI may serve as a useful criterion for evaluating the quality of classifications in practice.

In some instances, RI gravitates toward many clusters for some methods, but this tendency is not too severe for our simulation results. This inclination does not affect the evaluative classification procedure, because we employ stopping rules to target a specific number of clusters. However, the bias toward more refined classifications may prompt the use of alternative weights in the aggregate classification procedure. To examine the sensitivity of our composite classification to the influence of large *G* in the analysis of the ethanol data, we also pool classification results using weights adjusted for the number of clusters. Specifically, we calculate weights as the product of RI values η_{iG}, calculated among the algorithms with a given number of clusters, and a proportional weight across the total number of clusters. The η_{iG} sum to one within each group, the proportional weights sum to one across groups, and the total sum of the products of these two terms equals 1. In our sensitivity analysis, we apply weights 0.0045, 0.0540, 0.2420, 0.3989, 0.2420, 0.0540, and 0.0045 corresponding to the number of clusters ranging from 12 through 18, respectively. The group weights are symmetric, with the largest weight applying to the anticipated number of clusters—G = 15. The alternative weighting reduces the influence of large *G* and essentially eliminates the contribution of the worst performing method (complete linkage) within each *G*. Overall, the composite classifications given by the alternative weighting scheme yield results that are quite consistent with those obtained from our initial analysis.

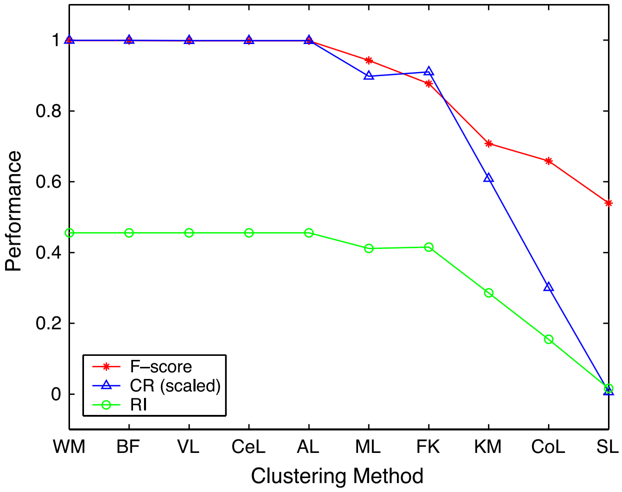

Fig. 5. Simulation results of *F* score, clustering reliability (CR), scaled to the interval [0,1], and relative information (RI) measures for 10 algorithms producing 11 clusters. There is high concordance among the performance criteria.

Table 2
Comparison of performance measures

Algorithm	*F* score	CR	RI
1. Ward's method	0.9988 (0.0008)	0.9996 (0.0005)	0.4559 (0.0290)
2. Beta-flexible	0.9986 (0.0008)	0.9994 (0.0007)	0.4557 (0.0290)
3. Variable linkage	0.9985 (0.0008)	0.9991 (0.0009)	0.4556 (0.0290)
4. Centroid linkage	0.9984 (0.0008)	0.9991 (0.0008)	0.4556 (0.0290)
5. Average linkage	0.9983 (0.0009)	0.9990 (0.0009)	0.4556 (0.0290)
6. Median linkage	0.9429 (0.0506)	0.8983 (0.0943)	0.4115 (0.0513)
7. Fuzzy clustering	0.8772 (0.0419)	0.9105 (0.0773)	0.4156 (0.0448)
8. *K* means	0.7081 (0.0320)	0.6095 (0.0601)	0.2862 (0.0435)
9. Complete linkage	0.6586 (0.0976)	0.3007 (0.2302)	0.1544 (0.1065)
10. Single linkage	0.5396 (0.0432)	0.0062 (0.0252)	0.0164 (0.0295)

Mean classification performance measures and standard deviations (in parentheses) from 100 simulations. The relative information (RI) measure corresponds closely to the benchmark *F* score and clustering reliability (CR), scaled to [0,1]. Larger values of the performance criteria indicate higher-quality partitions. The classifications consist of 11 clusters of voxel-specific brain activity profiles.

Discussion

In this paper, we present methodology for identifying spatial relationships in the neural processing linked to an experimental stimulus. The approach incorporates a collection of clustering solutions, produced by numerous algorithms with varying numbers of classes. The multiple classification approach entails an evaluative process that yields a plausible classification, relative to the others considered. Alternatively, the procedure combines results by assigning weights based on the relative plausibility of each classification. The performance of clustering algorithms depends on unknown characteristics of the data. Therefore, an important advantage of our method is that it provides additional assurance of capturing the underlying structure in the data, which a single algorithm may fail to uncover.

We introduce a measure, RI, for evaluating the quality of a classification produced by a clustering algorithm. RI is useful in practice because it does not involve any information about the true cluster memberships. We use RI as the basis for evaluating the numerous classifications given by our analysis. The RI value assigned to each classification closely relates to the posterior probability of the partition, given the data (see Appendix A). The use of RI in our methodology, therefore, selects or pools classifications based on the relative plausibility of each partition. In the simple case, where all classifications have the same number of clusters, RI provides rankings among the algorithms that are consistent with other likelihood-based methods such as -2 times the log-likelihood function, Akaike's information criterion (AIC) (Akaike, 1974), and the Bayesian information criterion (BIC) (Schwarz, 1978). RI is quite adaptable and readily extends to probability models other than the multivariate Gaussian, left-truncated Poisson, and the discrete uniform distributions that we employ in our data application.

The multiple classification approach provides more reliable results at the expense of increased computations. Our evaluation of the ethanol data, for example, executes eight hierarchical algorithms and nine different classifications (g = 12, . . ., 18) for each of two partitioning methods, for a total of 22 analyses. The total time for conducting the multiple clustering computations was roughly 45 min on a personal computer with a 3.06 GHz Pentium 4 processor and 1 gigabyte of memory. Although the multiple classification approach takes longer to perform than a single analysis, the overall time required is not exorbitant, and we feel that the benefits of our methods far outweigh the costs of additional computing.

We employ simple stopping rules in our analyses as guides toward the optimal number of clusters. Milligan and Cooper (1985) conduct an empirical study, outside of a neuroimaging context, which provides support for the use of the CCC, pseudo-F, and pseudo-T^2 stopping rules. Furthermore, a simulation study based on PET neuroimaging data suggests that all three stopping rules work well with several hierarchical clustering procedures, including Ward's and the beta-flexible methods (Bowman et al., in press). Nonetheless, there are settings in which alternative pruning methods that perform a search over the entire tree, e.g., runt size pruning (Stuetzle, 2003), yield vast improvements over the simple tree-cutting stopping rules that terminate a tree at a given level of the clustering hierarchy. For instance, generating two elongated overlapping clusters from a bivariate Gaussian distribution, some clustering procedures such as complete and average linkage would likely have difficulty discerning the two clusters and consequently create a spherical cover for the data. Single linkage with effective pruning over the entire tree, on the other hand, is capable of detecting the two clusters with a high degree of accuracy, but the algorithm would probably fail to identify the clusters using the simple stopping rules. Tree-search pruning methods can combine with our multiple classification approach by computing RI for the final pruned partitions of each clustering algorithm. By yielding higher-quality clustering solutions, tree-search pruning methods overcome potential limitations of the simple stopping rules and can substantially enhance the performance of our multiple classification procedure.

In addition to considering the RI criterion for evaluating classification performance, one should always incorporate scientific expertise and judgment. RI is intended to provide an objective rule-of-thumb measure for evaluating classifications. When implementing the evaluative clustering procedure, we recommend selecting from the top-performing algorithms and withholding strict interpretations regarding the exact rankings. Furthermore, the analyst should examine the results of our multiple classification analysis for scientific plausibility.

The multiple classification approach is a descriptive procedure that provides insights about the spatial organization of measured brain function, possibly associated with changes in specific behaviors, tasks, or conditions. The method identifies spatial locations that exhibit similar functionality, which may suggest, but not imply, details about functional connections and pathways of neural processing. Because we do not employ formal inference methods, such as Neyman–Pearson testing, our approach can precede an inference-based analysis without requiring any adjustments to the type I error rate in the ensuing analysis. The multiple classification approach is, therefore, an extremely valuable exploratory tool that can reveal important functional associations in neuroimaging data, generate hypotheses for subsequent analyses, and perhaps direct attention to specific anatomical regions of interest.

Acknowledgments

The authors thank Dr. Clinton Kilts, Mr. Timothy Ely, and Mrs. Carmen Santana from the Department of Psychiatry and Behavioral Sciences at Emory University, for discussions about the PET ethanol data employed in this research. This study was funded by NIH, grant number K25-MH65473.

Appendix

RI relates to the posterior probability of obtaining the jth partition, conditional on the data, i.e., $P(\Re=\Re_j,\ G=G_j \mid \boldsymbol{T})$, $j=1, \ldots, J$, which we simply represent as $P(\Re_j, G_j \mid \boldsymbol{T})$. Using a Laplace approximation (Ripley, 1996), we can express the logarithm of this probability as

$$\begin{aligned}&\log P(\Re_j, G_j \mid \boldsymbol{T}) \propto \log P(\Re_j \mid G_j) + \log P(G_j) + \log P(\boldsymbol{T} \mid \Re_j, G_j)\\ &\approx \log P(\Re_j \mid G_j) + \log P(G_j) + \log P(\boldsymbol{T} \mid \Re_j,\ \hat{\boldsymbol{\tau}}^{(j)})\\ &-\frac{G_j(q+1)}{2}\log(qV), \qquad (13)\end{aligned}$$

assuming $P(\Re_j,\ G_j \mid \boldsymbol{T}) > 0$. The first two terms in Eq. (11) represent ω_j in Eq. (6), the last two terms represent ψ_j in Eq. (7), and RI follows directly. We apply a discrete uniform probability model for $\Re$ among the J_{G_j} procedures that produce G_j clusters. We can extend this model to consider all possible partitions with G_j clusters, but the more general formulation is typically difficult to calculate for neuroimaging data due to the large number of voxels. We model the number of clusters using a left-truncated (positive) Poisson distribution with probability mass function

$$P(G=G_j) = \frac{\theta^{G_j}(\exp(\theta)-1)^{-1}}{G_j!}, \quad G_j = 1,2,3,\ldots. \qquad (14)$$

Although G is bounded above by V, we do not use a doubly truncated Poisson distribution because V is generally very large for neuroimaging applications, leading to zero or negligible probability in the range (V, ∞), i.e., $P(G > V) \approx 0$. We compute $P(\boldsymbol{T} \mid \Re_j, \hat{\boldsymbol{\tau}}^{(j)})$ using voxel-specific Gaussian densities, where $\hat{\boldsymbol{\tau}}^{(j)}=(\hat{\boldsymbol{\tau}}_1, \ldots, \hat{\boldsymbol{\tau}}_{G_j})$ and the components $\hat{\boldsymbol{\tau}}_{g_j} = (\hat{\mu}_{g_j}, \hat{\sigma}^2_{g_j})$ are the within-cluster sample mean and variance estimates. Following from Eqs. (13) and (14), the penalty term for the number of clusters is

$$\delta_j^* = G_j \log\theta - \log(G_j!) - \frac{G_j(q+1)}{2}\log(qV). \tag{15}$$

References

Akaike, H., 1974. A new look at the statistical model identification. IEEE Trans. Autom. Control AC-19, 716–723.

Balslev, D., Nielsen, F.A., Frutiger, S.A., Sidtis, J.J., Christiansen, T.B., Svarer, C., Strother, S.C., Rottenberg, D.A., Hansen, L.K., Paulson, O.B., 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. Imaging 18, 89–94.

Bowman, F.D., Kilts, C., 2003. Modeling intra-subject correlation among repeated scans in Positron Emission Tomography (PET) neuroimaging data. Hum. Brain Mapp. 20, 59–70.

Bowman, F.D., Patel, R., Lu, C., 2004. Methods for detecting functional classifications in neuroimaging data. Hum. Brain Mapp. (in press).

Calinski, R.B., Harabasz, J., 1974. A dendrite method for cluster analysis. Commun. Stat. 3, 1–27.

Cordes, D., Haughton, V., Carew, J., Arfanakis, K., Maravilla, K., 2002. Hierarchical clustering to measure connectivity in fMRI resting-state data. Magn. Reson. Imaging 20, 305–317.

Duda, R.O., Hart, P.E., 1973. Pattern Classification and Scene Analysis Wiley, New York.

Fadili, M.J., Ruan, S., Bloyet, D., Mazoyer, B., 2000. A multistep unsupervised fuzzy clustering analysis of fMRI time series. Hum. Brain Mapp. 10, 160–178.

Fadili, M.J., 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. Med. Image Anal. 5, 55–67.

Filzmoser, P., Baumgartner, R., Moser, E., 1999. A hierarchical clustering method for analyzing functional MR images. Magn. Reson. Imaging 17 (6), 817–826.

Friston, K.J., Frith, C., Liddle, P., Frackowiak, R.S.J., 1993. Functional connectivity: the principal component analysis of large data sets. J. Cereb. Blood Flow Metab. 13, 5–14.

Friston, K.J., Holmes, A.P., Poline, J.B., Grasby, P.J., Williams, S.C.R., Frackowiak, R.S.J., Turner, R., 1995. Analysis of fMRI time-series revisited. NeuroImage 2, 45–53.

Frutiger, S.A., Strother, S.C., Anderson, R., Sidtis, J., Arnold, J.B., Rottenberg, D.A., 2000. Multivariate predictive relationship between kinematic and functional activation patterns in a PET study of visuomotor learning. NeuroImage 12, 515–527.

Goutte, C., Toft, P., Rostrup, E., Nielsen, F.A., Hansen, L.K., 1999. On clustering fMRI time series. NeuroImage 9, 298–310.

Goutte, C., Nielsen, F.Å., Liptrot, M.G., Hansen, L.K., 2001. Feature-space clustering for fMRI meta-analysis. Hum. Brain Mapp. 13 (3), 165–183.

Hartigan, J.A., 1975. Clustering Algorithms. Wiley, New York.

Hendelman, W.J., 2000. Atlas of Functional Neuroanatomy. CRC Press, New York.

Larsen, B., Aone, C., 1999. Fast and effective text mining using linear-time document clustering. Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM Press, San Diego CA, p. 16–22.

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

McIntosh, A.R., Bookstein, F.L., Haxby, J.V., Grady, C.L., 1996. Spatial pattern analysis of functional brain images using partial least squares. NeuroImage 3, 143–157.

McKeown, M.J., Makeig, S., Brown, G.G., Jung, T.-P., Kindermann, S.S., Bell, A.J., Sejnowski, T.J., 1998. Analysis of fMRI data by blind separation into independent spatial components. Hum. Brain Mapp. 6, 160–188.

Milligan, G.W., Cooper, M.C., 1985. An examination of procedures for determining the number of clusters in a data set. Psychometrika 50 (2), 159–179.

Ramsay, J.O., Silverman, B.W., 1997. Functional Data Analysis. Springer, New York.

Rencher, A., 2002. Methods of Multivariate Analysis, second ed. Wiley Inc., New York.

Ripley, B.D., 1996. Pattern Recognition and Neural Networks. Cambridge Univ. Press, Cambridge, MA.

Sarle, W.S., 1983. Cubic Clustering Criterion, SAS Technical Report A-108 SAS Institute Inc., Cary, NC.

Schwarz, G., 1978. Estimating the dimension of a model. Ann. Stat. 6, 461–464.

Somorjai, R.L., Jarmasz, M., 2003. Exploratory analysis of fMRI data by fuzzy clustering—philosophy, strategy, tactics, and implementation. In: Sommer, F.T., and Wichert, A. (Eds.), Exploratory Analysis and Data Modeling in Functional Neuroimaging. The MIT Press, Cambridge, MA, p. 17–48.

Stanberry, L., Nandy, R., Cordes, D., 2003. Cluster analysis of fMRI data using dendrogram sharpening. Hum. Brain Mapp. 20, 201–219.

Stuetzle, W., 2003. Estimating the cluster tree of a density by analyzing the minimal spanning tree of a sample. J. Classif. 20, 25–47.

Talairach, J., Tournoux, P., 1988. Co-Planar Stereotaxic Atlas of the Human Brain. Thieme Medical Publishers Inc, New York.

Thireou, T., Strauss, L., Kontaxakis, G., Pavlopoulos, S., Santos, A., 2001. Principal component analysis in dynamic positron emission tomography. Span. Conf. Biomed. Eng., 241–243.

Wong, M.A., Lane, T., 1983. A kth nearest neighbor clustering procedure. J. Royal Stat. Soc. Ser. B, 45, 362–368.

Woods, R.P., Grafton, S.T., Holmes, C.J., Cherry, S.R., Mazziotta, J.C., 1998a. Automated image registration: I. General methods and intrasubject, intramodality validation. J. Comput. Assist. Tomogr. 22, 139–152.

Woods, R.P., Grafton, S.T., Watson, J.D.G., Sicotte, N.L., Mazziotta, J.C., 1998b. Automated image registration: II. Intersubject validation of linear and nonlinear models. J. Comput. Assist. Tomogr. 22, 153–165.

Worsley, K.J., Friston, K.J., 1995. Analysis of fMRI time-series revisited—again. NeuroImage 2, 173–181.