Comparison of Two Exploratory Data Analysis Methods for fMRI: Unsupervised Clustering Versus Independent Component Analysis

A. Meyer-Baese, Axel Wismueller, and Oliver Lange

***Abstract*—Exploratory data-driven methods such as unsupervised clustering and independent component analysis (ICA) are considered to be hypothesis-generating procedures, and are complementary to the hypothesis-led statistical inferential methods in functional magnetic resonance imaging (fMRI). In this paper, we present a comparison between unsupervised clustering and ICA in a systematic fMRI study. The comparative results were evaluated by 1) task-related activation maps, 2) associated time-courses, and 3) receiver operating characteristic analysis. For the fMRI data, a comparative quantitative evaluation between the three clustering techniques, self-organizing map, "neural gas" network, and fuzzy clustering based on deterministic annealing, and the three ICA methods, FastICA, Infomax and topographic ICA was performed. The ICA methods proved to extract features relatively well for a small number of independent components but are limited to the linear mixture assumption. The unsupervised clustering outperforms ICA in terms of classification results but requires a longer processing time than the ICA methods.**

***Index Terms*—FastICA, functional magnetic resonance imaging (fMRI), Infomax, minimal free energy vector quantization (VQ), "neural gas" network, principal component analysis (PCA), self-organizing map, topographic independent component analysis (ICA).**

I. Introduction

FUNCTIONAL magnetic resonance imaging (fMRI) with high temporal and spatial resolution represents a powerful technique for visualizing rapid and fine activation patterns of the human brain [1]. As is known from both theoretical estimations and experimental results [2], an activated signal variation appears very low on a clinical scanner. This motivates the application of analysis methods to determine the response waveforms and associated activated regions. Generally, these techniques can be divided into two groups: Model-based techniques require prior knowledge about activation patterns, whereas model-free techniques do not. However, model-based analysis methods impose some limitations on data analysis under complicated experimental conditions. Therefore, analysis methods that do not rely on any assumed model of functional response are considered more powerful and relevant. We distinguish two groups of model-free methods: transformation-based and clustering-based.

The first method, principal component analysis (PCA) [3], [4] or independent component analysis (ICA) [5]–[8], transforms original data into high-dimensional vector space to separate functional response and various noise sources from each other.

Among the data-driven techniques, ICA has been shown to provide a powerful method for the exploratory analysis of fMRI data [6], [8]. ICA is an information theoretic approach which enables recovery of underlying signals, or independent components (ICs) from linear data mixtures. Therefore, it is an excellent method to be applied for the spatial localization and temporal characterization of sources of BOLD activation. ICA can be applied to fMRI both temporal [9] or spatial [6]. Spatial ICA has dominated so far in fMRI applications because the spatial dimension is much larger than the temporal dimension in fMRI. However, recent literature results have suggested that temporal and spatial ICA yield similar results for experiments where two predictable task-related components are present.

The second method, fuzzy clustering analysis [10]–[13] or self-organizing map [13], [21], [23], attempts to classify time signals of the brain into several patterns according to temporal similarity among these signals.

In this paper, we perform a detailed comparative study among unsupervised clustering methods ["neural gas" network [22], fuzzy clustering based on deterministic annealing [13], and Kohonen's self-organizing map (SOM)] and spatial ICA techniques (FastICA [14], topographic ICA [16], Infomax [17], PCA) for fMRI. In a systematic manner, we will compare and evaluate the results obtained based on each technique and present the benefits associated with each paradigm.

II. Exploratory Data Analysis Methods

Functional organization of the brain is based on two complementary principles, localization and connectionism. Localization means that each visual function is performed mainly by a small set of the cortex. Connectionism, on the other hand, expresses that the brain regions involved in a certain visual cortex function are widely distributed, and thus, the brain activity necessary to perform a given task may be the functional integration of activity in distinct brain systems. It is important to stress that in neurobiology the term "connectionism" is used in a different sense that that used in the neural network terminology.

Manuscript received June 28, 2003; revised February 12, 2004.

A. Meyer-Baese and O. Lange are with the Department of Electrical and Computer Engineering, Florida State University, Tallahassee, FL 32310-6046 USA (e-mail: amb@eng.fsu.edu).

A. Wismueller is with the Department of Clinical Radiology, Ludwig-Maximilians University, Munich 80336, Germany.

Digital Object Identifier 10.1109/TITB.2004.834406

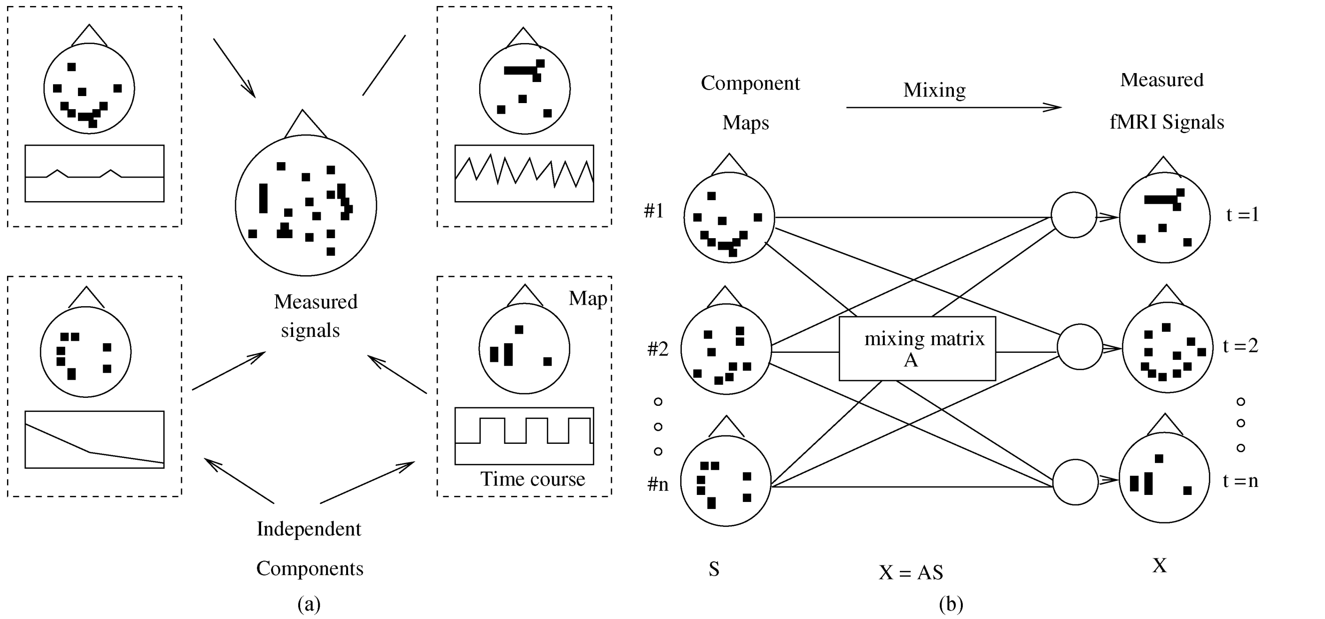


Fig. 1. Visualization of ICA applied to fMRI data. (a) Scheme of fMRI data decomposed into ICs, and (b) fMRI data as a mixture of ICs where the mixing matrix $\mathbf{M}$ specifies the relative contribution of each component at each time point [6].

The following sections are dedicated to presenting the algorithms and evaluate the discriminatory power of the two main groups of exploratory data analysis methods.

A. ICA Algorithms

According to the principle of functional organization of the brain, it was suggested for the first time in [6] that the multifocal brain areas activated by performance of a visual task should be unrelated to the brain areas whose signals are affected by artifacts of physiological nature, head movements, or scanner noise related to fMRI experiments. Every single above mentioned process can be described by one or more spatially ICs, each associated with a single time course of a voxel and a component map. It is assumed that the component maps, each described by a spatial distribution of fixed values, represent overlapping, multifocal brain area of statistically dependent fMRI signals. This aspect is visualized in Fig. 1. In addition, it is considered that the distributions of the component maps are spatially independent, and in this sense uniquely specified. Mathematically, this means that if $p_k(C_k)$ specifies the probability distribution of the voxel values C_k in the kth component map, then the joint probability distribution of all n components yields

$$p(C_1, \ldots, C_n) = \prod_{k=1}^{n} p_k(C_k) \tag{1}$$

where each of the component maps C_k is a vector ($C_{ki}, i = 1, 2, \ldots, M$), where M gives the number of voxels. Independency is a stronger condition than uncorrelatedness. It was shown in [6] that these maps are independent if the active voxels in the maps are sparse and mostly nonoverlapping. Additionally, it is assumed that the observed fMRI signals are the superposition of the individual component processes at each voxel. Based on these assumptions, ICA can be applied to fMRI time series to spatially localize and temporally characterize the sources of BOLD activation.

Different methods for performing ICA decompositions have been proposed which employ different objective functions together with different criteria of optimization of these functions, and it is assumed that they can produce different results.

B. Models of Spatial ICA in fMRI

In the following, we will assume that $\mathbf{X}$ is a $T \times M$ observed fMRI signal data matrix, $\mathbf{C}$ is the $N \times M$ random matrix of component map values, and A is a $T \times N$ mixing matrix containing in its columns the associated time-courses of the N components. Furthermore, T corresponds to the number of scans, and M is the number of voxels included in the analysis. Matrix $\mathbf{X}$ is the so-called matrix of observed voxel time courses (VTCS).

The spatial ICA (sICA) problem is given by the following linear combination model for the data:

$$\mathbf{X} = \mathbf{AC} \tag{2}$$

where no assumptions are made about the mixing matrix $\mathbf{A}$ and the rows $\mathbf{C_i}$ being mutually statistically independent.

Then the ICA decomposition of $\mathbf{X}$ can be defined as an invertible transformation

$$\mathbf{C} = \mathbf{WX} \tag{3}$$

where $\mathbf{W}$ is an unmixing matrix providing a linear decomposition of data. $\mathbf{A}$ is the pseudoinverse of $\mathbf{W}$.

C. Infomax Approach

The Infomax was the first ICA application to fMRI time series [6], and is based on minimization of mutual information [17]. The algorithmic description is given in the following. A self-organizing learning algorithm is described that maximizes the information transferred in a network of linear units. It was shown that the neural network is able to perform ICA estimation and that the nonlinearities in the transfer function introduce higher order statistics.

The network has N input and output neurons, and an $N \times N$ weight matrix $\mathbf{W}$ connecting the input layer neurons with the output layer neurons. Assuming sigmoidal units, the neurons outputs are given by

$$\mathbf{y} = g(\mathbf{C}), \qquad \text{with} \quad \mathbf{C} = \mathbf{W}\mathbf{X} \tag{4}$$

where g is a logistic function $g(u_i) = 1/(1 + \exp -u_i)$.

The idea of this algorithm is to find an optimal weight matrix $\mathbf{W}$ such that the output entropy $H(\mathbf{y})$ is maximized. The algorithm initializes $\mathbf{W}$ to the identity matrix $\mathbf{I}$. The elements of $\mathbf{W}$ are updated based on the following rule:

$$\mathbf{W} \leftarrow -\eta \left(\frac{\partial H(\mathbf{y})}{\partial \mathbf{W}} \right) \mathbf{W}^T\mathbf{W} = -\eta(\mathbf{I} + f(\mathbf{C})\mathbf{C}^T)\mathbf{W} \tag{5}$$

where η is the learning rate. The term $\mathbf{W}^T\mathbf{W}$ in (5) was first proposed in [18], and it avoids matrix inversions and speeds convergence. The vector-function f has the elements

$$f_i(u_i) = \frac{\partial}{\partial u_i} \ln g_i'(u_i) = (1 - 2y_i). \tag{6}$$

During training, the learning rate is reduced gradually until the weight matrix $\mathbf{W}$ stops changing appreciably.

Equation (5) represents the so-called “Infomax” algorithm.

The choice of a nonlinearity is determined by the application type. In the context of fMRI, where small activity foci in a large volume are usually expected, the distribution of the target components is assumed to be super-Gaussian or sparse. Therefore, a sigmoidal function $g(\cdot)$ is relevant for such an application.

D. *FastICA Approach*

The FastICA approach is based on minimization of mutual information but using the negentropy as a measure of non-Gaussianity. This approach is both symmetric and hierarchical, and is based on fixed-point iterations. To apply this ICA approach, the data must be preprocessed by centering and whitening. A single artificial neuron has a weight vector $\mathbf{w}$ that is updated based on a learning algorithm. It finds a vector $\mathbf{w}$ such that the projection $\mathbf{w}^T\mathbf{x}$ maximizes non-Gaussianity.

The FastICA algorithm is a fixed-point iteration scheme for finding a maximum of the non-Gaussianity of $\mathbf{w}^T\mathbf{x}$. To estimate several ICs, the one-unit FastICA is employed using several units (neurons) with weight vectors $\mathbf{w}, \ldots, \mathbf{w}_n$. To prevent different vectors from converging to the same maxima the outputs $\mathbf{w}_1^T\mathbf{x}, \ldots, \mathbf{w}_n^T\mathbf{x}$ have to be decorrelated after every iteration. For a whitened $\mathbf{x}$, this is equivalent to orthogonalization. There are several known methods to achieve this [15]. Here, only the symmetric decorrelation is considered. It has several advantages over other methods: 1) the weight vectors $\mathbf{w}_i$ are estimated in parallel and not one by one and 2) it does not perpetuate the errors from one weight vectors to the next.

The symmetric orthogonalization of $\mathbf{W}$ can be accomplished by involving matrix square roots

$$\mathbf{W} = (\mathbf{W}\mathbf{W}^T)^{-1/2}\mathbf{W}. \tag{7}$$

Numerical simplifications of the above equation are given in [15].

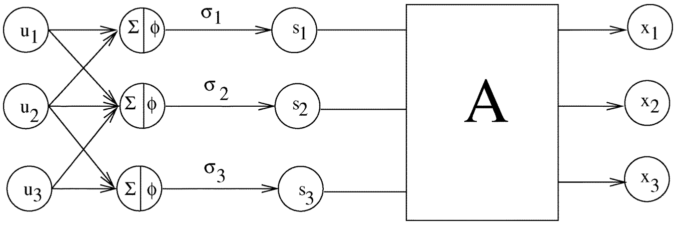


Fig. 2. Topographic ICA model [16]. The variance generated variables u_i are randomly generated, and mixed linearly inside their topographic neighborhoods. This forms the input to nonlinearity ϕ, thus giving the local variance σ_i. Components s_i are generated with variances σ_i. The observed variables are x_i are obtained as with standard ICA from the linear mixture of the components s_i.

An algorithmic description of the FastICA algorithm for estimating several ICs is given in [15]. The main difference between Infomax and FastICA lies in the updating rule: It is adaptive for the Infomax depending on a learning rate, and it is nonadaptive for FastICA.

E. *Topographic ICA Approach*

Topographic ICA represents a generative model which combines topographic mapping with ICA. As in all topographic mappings, the distance in the representation space given by the topographic grid is related to the distance of the represented components. This distance is defined for topographic ICA by the mutual information implied by higher order correlations [16]. Thus, a natural distance measure is given in the context of ICA. Traditional topographic mapping methods define distance either based on the Euclidean distance or correlation. The ICA distance measure enables the definition of a topography even if the Euclidean distances are all equal as it is the case with an orthogonal vector space.

In the generative model described in [16] and shown in Fig. 2, s_i represent the unknown sources and are independent given their variances σ_i^2. Dependence among the s_i is enforced by the variance dependence. Obeying the principle of topography, the variances of only neighboring components are positively correlated, while the others are independent. By using a neighborhood function $h(i, j)$, the variance σ_i is given by

$$\sigma_i = \phi \left(\sum_{k=1}^{n} h(i, k) u_k \right) \tag{8}$$

where u_i are the higher order ICs used to generate the variances, while ϕ describes some nonlinearity. The neighborhood function $h(i, j)$ can be either a two-dimensional grid or have a ring-like structure. The components s_i are given by the following relationship:

$$s_i = z_i \sigma_i \tag{9}$$

where z_i is a random variable having the same distribution as s_i while σ_i^2 is fixed to unity. u_i and z_i are mutually independent.

The most important properties of the topographic ICA are: 1) all the components are uncorrelated, 2) components far from each other are independent, 3) neighboring components tend to be active (nonzero) at the same time, and thus, have positively correlated energies s_i^2 and s_j^2. The classic ICA results from the topographic ICA by setting $h(i, j) = \delta_{ij}$.

The learning rule is based on the maximization of the likelihood. First, it is assumed that the data is preprocessed by whitening and that the estimates of the components are uncorrelated.

The update rule for the weight vector $\mathbf{w}_i$ is derived from a gradient algorithm

$$\Delta \mathbf{w}_i \propto E\{\mathbf{x}(\mathbf{w}_i^T \mathbf{x}) r_i)\} \tag{10}$$

where

$$r_i = \sum_{k=1}^{n} h(i,k) g \left(\sum_{j=1}^{n} k(k,j)(\mathbf{w}_j^T \mathbf{x})^2 \right). \tag{11}$$

The function g is the derivative of $G = -\alpha_1 \sqrt{u} + \beta_1$. After every iteration, the vectors $\mathbf{w}_i$ in (10) are normalized to unit variance and orthogonalized. This equation represents a modulated learning rule, where the learning term is modulated by the term r_i.

It is useful to point out some differences between topographic ICA and other topographic mappings: 1) topographic ICA finds a decomposition into ICs, while topographic mappings find cluster centers or codevectors, and 2) the similarity of two vectors in topographic ICA is based on higher order correlations and not defined by an Euclidean distance or dot-product. However, if the data is prewhitened, the dot-product in the data space is equivalent to correlation in the original space [16].

Topographic ICA represents a new paradigm for fMRI signal analysis since the strict independence condition imposed by standard ICA techniques is relaxed among neighboring components, such that neighboring components (voxels) are positively correlated.

F. PCA Approach

Principal component analysis is a basic technique used for data reduction in bioimaging. The idea is that similar input patterns belong to the same class. Thus, the input data can be normalized within the unit interval and then chosen based on their variances. In this sense, the larger the variances, the better discriminatory properties the input features have.

PCA involves a mathematical procedure that transforms a number of (possibly) correlated variables into a (smaller) number of uncorrelated variables called principal components. This is a highly desirable property, since besides being optimally uncorrelated, the redundancy in data information is removed. By selecting the eigenvectors having the largest eigenvalues, we lose as little information as possible in the mean-square sense. A fixed number of eigenvectors and their respective eigenvalues can be chosen to obtain a consistent representation of the data.

Let $\mathbf{X}$ be a $T \times M$ matrix of observed cortical time courses. The correlation matrix $\mathbf{R}$ of the cortical time courses is defined as

$$\mathbf{R} = E[\mathbf{X}\mathbf{X}^T]. \tag{12}$$

This correlation matrix $\mathbf{R}$ can be rewritten in terms of the eigenvalues as

$$\mathbf{R} = \sum_{i=1}^{m} \lambda_i \mathbf{q}_i \mathbf{q}_i^T \tag{13}$$

where $\mathbf{q}_i$ is the ith eigenvector and λ_i the corresponding ith eigenvalue of the matrix $\mathbf{R}$.

G. Clustering Algorithms

The previous sections showed that ICA techniques can be applied to fMRI by considering brain function as consisting of sets of nonsystematically overlapping networks. In other words, ICA works by assuming that during a given fMRI experiment there are a number of brain regions (networks) that are spatially independent from one another (sources) and are mixed together via a network specific hemodynamic time course.

In this section, we will review cluster analysis as an alternative technique which is based on grouping image voxels together based on the similarity of their intensity profile in time (i.e., their time courses).

Let T denote the number of subsequent scans in an fMRI study, and let M be the number of voxels. The dynamics of each voxel $\mu \in \{1, \ldots, M\}$, i.e., the sequence of signal values $\{\mathbf{x}^\mu(1), \ldots, \mathbf{x}^\mu(T)\}$, can be interpreted as a vector $\mathbf{x}^\mu(i) \in \mathbf{R}^T$ in the T-dimensional feature space of possible signal time series at each voxel.

Cluster analysis groups image voxels together based on the similarity of their intensity profile in time. In the clustering process, a time course with T points is represented by one point in an T-dimensional Euclidean space which is subsequently partitioned into clusters based on the proximity of the input data.

Here, we employ several vector quantization (VQ) approaches as a method for unsupervised image time-series analysis. VQ clustering identifies several groups of voxels with similar VTC, while these groups or clusters are represented by prototypical time series called codebook vectors (CVs) located at the center of the corresponding clusters. The CVs represent prototypical VTCs sharing similar temporal characteristics. Thus, each VTC can be assigned in the crisp clustering scheme to a specific CV according to a minimal distance criterion, while in the fuzzy scheme according to a membership to several CVs. Accordingly, the outcomes of VQ approaches for fMRI data analysis can be plotted as “crips” or “fuzzy” cluster assignment maps.

VQ approaches determine the cluster centers $\mathbf{w}_i$ by an iterative adaptive update based on the following equation:

$$\mathbf{w}_i(t+1) = \mathbf{w}_i(t) + \epsilon(t) a_i(\mathbf{x}(t), C(t), \kappa)(\mathbf{x}(t) - \mathbf{w}_i(t)) \tag{14}$$

where $\epsilon(t)$ represents the learning parameter, a_i a codebook $C(t)$ dependent cooperativity function, κ a cooperativity parameter, and $\mathbf{x}$ a randomly chosen feature vector. For fMRI, the feature vector represents the VTC.

H. Kohonen’s Self-Organizing Map

Kohonen’s self-organizing map generates nodes on a two-dimensional lattice in which the distribution of these nodes corresponds to the proximity of their associated node patterns in the signal intensity space. The benefits of this clustering technique are: 1) if started with an adequate number of neurons, it can find distinctive features in the data even if they are less predominant, and 2) the emerging node patterns are ordered according to their proximity properties in the data space. This topology-preserving technique enables the forming of superclusters by fusing nodes, and thus, provides a way to visualize high-dimensional data sets. Its advantages in analyzing fMRI data were demonstrated in [23].

The update equation for the CVs based on this VQ approach can be derived from (14). The cooperativity function a_i is given by

$$a_i(\mathbf{x}(t), C(t), \kappa = \rho(t)) = \exp\left(-\frac{d_{ij}}{\sigma^2}\right) \tag{15}$$

where d_{ij} is a distance between neurons i and j determined by a neighborhood relation and σ^2 is an adjusting parameter. a_i takes the maximum value of one when $i = j$, namely for the firing neuron, and decreases when the distance becomes large.

The resulting learning rule for the Kohonen self-organizing map is given below

$$\mathbf{w}_i(t+1) = \mathbf{w}_i(t) + \epsilon(t)\exp\left(-\frac{d_{ij}}{\sigma^2}\right)(\mathbf{x}(t) - \mathbf{w}_i(t)). \tag{16}$$

I. Fuzzy Clustering Based on Deterministic Annealing

Another proven tool for the analysis of fMRI time series is given by a clustering fuzzy clustering technique based on deterministic annealing [13], [19].

The update equation for the CVs based on this VQ approach can be derived from (14). The cooperativity function a_i is given by

$$a_i(\mathbf{x}(t), C(t), \kappa = \rho(t)) = \frac{\exp -\dfrac{||\mathbf{x}(t) - \mathbf{w}_i(t)||^2}{2\rho^2}}{\sum\limits_i \exp -\dfrac{||\mathbf{x}(t) - \mathbf{w}_i(t)||^2}{2\rho^2}} \tag{17}$$

where ρ is the "fuzzy range" of the model, and defines a length scale in data space and is annealed to repeatedly smaller values in the VQ approach. In parlance of statistical mechanics, ρ represents the temperature T of a multiparticle system by $T = 2\rho^2$.

The cooperativity function a_j is the so-called *softmax* activation function, and accordingly, the outputs lie in the interval [0,1] and they sum up to one. The resulting learning rule for fuzzy clustering based on deterministic annealing is given below

$$\mathbf{w}_i(t{+}1) {=} \mathbf{w}_i(t) {+} \epsilon(t) \frac{\exp -\dfrac{||\mathbf{x}(t) - \mathbf{w}_i(t)||^2}{2\rho^2}}{\sum\limits_i \exp -\dfrac{||\mathbf{x}(t) - \mathbf{w}_i(t)||^2}{2\rho^2}} (\mathbf{x}(t) {-} \mathbf{w}_i(t)). \tag{18}$$

This learning rule describes a stochastic gradient descent on an error function which is a free energy in a mean-field approximation. The algorithm starts with one cluster representing the center of the whole data set. Gradually, the large clusters split up into smaller ones representing smaller regions in the feature space. This represents a major advantage over standard fuzzy c-means clustering since this algorithm does not employ prespecified cluster centers.

This clustering procedure identifies groups of voxels sharing similar properties of signal dynamics, and thus, enables the interpretation of the physiological part of the experiment. The main differences between SOM and fuzzy clustering based on deterministic annealing were pointed out in [13]: 1) the hierarchical and multiresolution aspect of data analysis; 2) monitoring based on different control parameters (free energy, entropy) facilitates straightforward cluster splitting; and 3) the learning rule based on a stochastic gradient descent on an explicitly given error function.

J. "Neural Gas" Network

The "neural-gas" algorithm [22] is an efficient approach which, applied to the task of VQ, 1) converges quickly to low distortion errors, 2) reaches a distortion error E lower than that from Kohonen's feature map, and 3) at the same time obeys a gradient descent on an energy surface.

Instead of using the distance $||\mathbf{x} - \mathbf{w_i}||$ or of using the arrangement of the $||\mathbf{w_i}||$ within an external lattice, it utilizes a neighborhood-ranking of the reference vectors $\mathbf{w_i}$ for the given data vector $\mathbf{x}$.

The update equation for the CVs based on this VQ approach can be derived from (14). The cooperativity function a_i is given by

$$a_i(\mathbf{x}(t), C(t), \kappa = \rho(t)) = \exp -k_i\left(\frac{\mathbf{x}, \mathbf{w_i}}{\lambda}\right) \tag{19}$$

where $k_i = 0, \ldots, N-1$ represents the rank index describing the neighborhood-ranking of the neural units, N is the number of units in the network, and λ determines the number of neural units changing their synapses with every iteration. The resulting learning rule for the "neural gas" network is given below

$$\mathbf{w}_i(t+1) = \mathbf{w}_i(t) + \epsilon(t)\exp -k_i\left(\frac{\mathbf{x}, \mathbf{w_i}}{\lambda}\right)(\mathbf{x}(t) - \mathbf{w}_i(t)). \tag{20}$$

The step size $\epsilon \in [0, 1]$ describes the overall extent of the modification.

In [22], it was shown that the average change of the reference vectors corresponds to an overdamped motion of particles in a potential that is given by the negative data point density. Superimposed on the gradient of this potential is a "force," which points toward the direction of the space where the particle density is low. This "force" is the result of a repulsive coupling between the particles (reference vectors). In its form, it resembles an entropic force and tends to homogeneously distribute the particles (reference vectors) over the input space, like in case of a diffusing gas. This suggests the name for the "neural-gas" algorithm. It is interesting also to mention that the reference vectors change their locations slowly but permanently and, therefore, pointers that are neighboring at an early stage of the adaptation procedure might not be neighboring anymore at a more advanced stage. Connections that have not been refreshed for a while die out and are removed.

Another important feature of the presented algorithm compared to Kohonen algorithm is that it does not require a prespecified graph (network). In addition, it can produce topologically preserving maps, which is only possible if the topological structure of the graph matches the topological structure of the data manifold. However, in cases where it is not possible to *a priori* determine an appropriate graph, for example, in cases where the topological structure of the data manifold is not known *a priori* or is too complicated to be specified, Kohonen's algorithm necessarily fails in providing perfectly topology preserving maps.

III. Results and Discussion

FMRI data were recorded from six subjects (three female, three male, age 20–37) performing a visual task. In five subjects, five slices with 100 images ($\mathrm{TR/TE} = 3000/60$ ms) were acquired with five periods of rest and five photic simulation periods with rest. Simulation and rest periods comprised ten repetitions each, i.e., 30 s. Resolution was $3 \times 3 \times 4$ mm. The slices were oriented parallel to the calcarine fissure. Photic stimulation was performed using an 8-Hz alternating checkerboard stimulus with a central fixation point and a dark background with a central fixation point during the control periods [13]. The first scans were discarded for remaining saturation effects. Motion artifacts were compensated by automatic image alignment [20].

The clustering results were evaluated by 1) task-related activation maps, 2) associated time-courses, and 3) receiver operating characteristic (ROC) analysis.

In the following, we will give the set of parameters chosen for the comparative evaluation for both exploratory data analysis techniques. For PCA, no parameters had to be set. For Infomax we choose the following: 1) the learning rate $\eta = 10^{-6}$; and 2) 10^5 as the maximal number of iterations. For FastICA we choose the following: 1) 10^5 as the maximal number of iterations; and 2) the nonlinearity $g(u) = \tanh u$. And last, for topographic ICA we set the following: 1) stop criterion is fulfilled if the synaptic weights difference between two consecutive iterations is less than $10^{-5}\times$ number of IC; 2) the function $g(u) = u$; and 3) 10^4 is the maximal number of iterations.

For SOM, we employed the SOMPAK (http://www.cis.hut.fi/research/sompak) and we choose the following: A rectangular grid, the neighborhood function is a bubble (step function), and for the initial ordering training phase we set $\epsilon_i = 0.05$, $\sigma = 3.0$, and the maximum iteration number equals the number of data points. For the fine-tuning training phase we set $\epsilon_i = 0.01$, $\sigma = 1.5$, and the maximum iteration number equals $1.5 \times$ the number of data points. For "neural gas" network we choose the following: 1) the learning parameters $\epsilon_i = 0.5$ and $\epsilon_f = 0.005$; 2) the lattice parameters λ_i equals half the number of classes and $\lambda_f = 0.01$; and 3) the maximal number of iterations equals 1000. And last, for fuzzy clustering based on deterministic annealing we set the following: 1) neurons' initialization with principal components; 2) learning parameter $\rho_{\text{final}} = 0.01$ and updating based on a linear annealing scheme; and 3) the maximal number of iterations equals 100.

It is important to determine the differences and implications on the analysis of fMRI time series. The two main groups which are subject to our comparative study are 1) the transformation-based methods such as PCA and ICA, and 2) the clustering methods such as SOM, "neural gas" network, and the fuzzy clustering based on deterministic annealing.

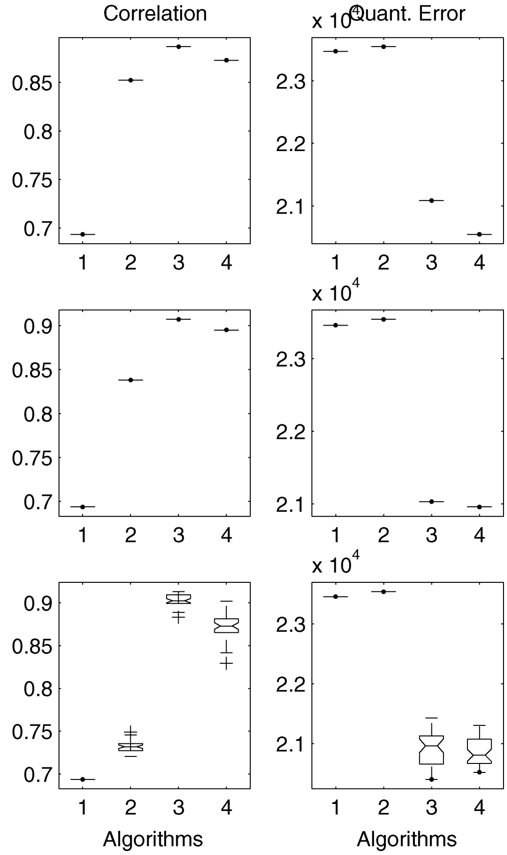


Fig. 3. Results from simulation optimization for transformation-based techniques for 8, 16, and 36 ICs from top to bottom. The algorithms 1–4 are: PCA, Infomax, FastICA, and topographical ICA. The plots show the correlation of the CTR component with the reference function and the corresponding quantization error per voxel.

The main difference between these two model-free fMRI analysis techniques lies in their categorization properties. The transformation-based techniques allow us to determine the contribution of the consistently task-related (CTR) component and the artifacts components (head movement) to a single voxel. While based on ICA, it is possible to separate the artifacts from the CTR; clustering techniques try to identify clusters of similar VTCs within the data space. In other words, they try to determine clusters describing similar VTCs based on a minimal distance criterion, such that the resulting CVs are then prototypes of VTCs of similar temporal characteristics.

A. Estimation of Optimal Number of ICs and CVs

The estimation of the optimal number of target components, IC number in case of transformation-based techniques and CV for clustering techniques, is of critical importance for exploratory data analysis in fMRI.

To compare uniformly the performances of the seven algorithms with a varying number of target components, we use the maximum achieved correlation of the CTR component with the reference function and the quantization error per voxel.

The achieved results for both ICA and clustering techniques for 8, 16, and 36 components are plotted in Figs. 3 and 4.

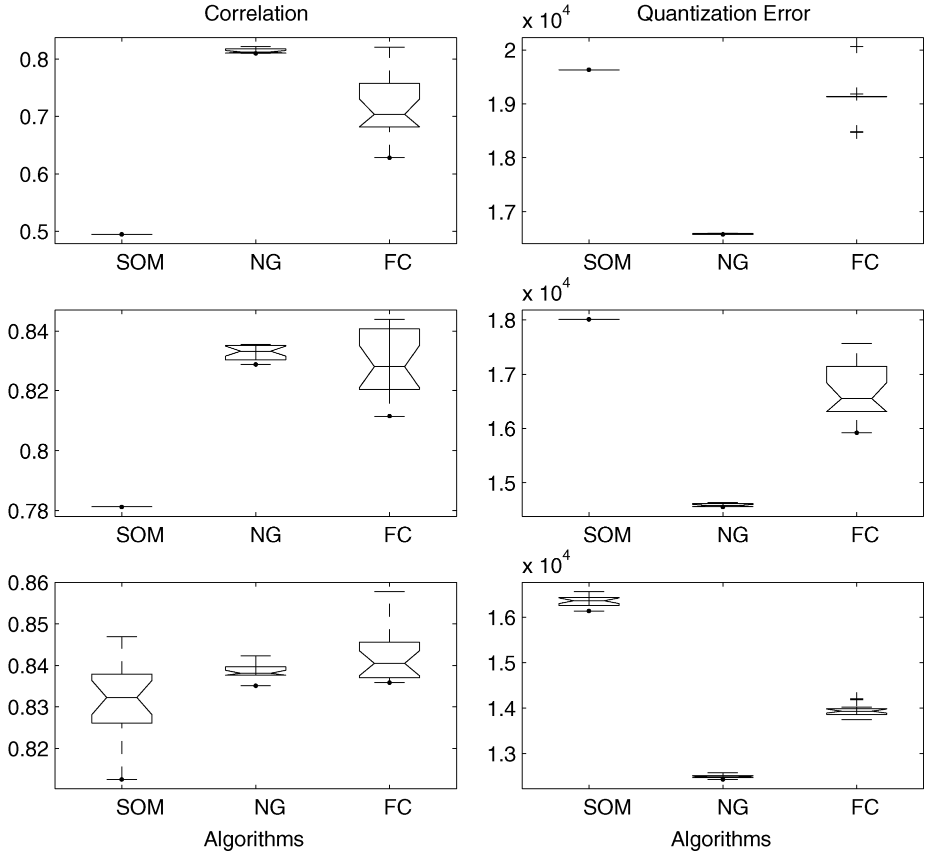


Fig. 4. Results from simulation optimization for clustering techniques for 8, 16, and 36 ICs from top to bottom. The plots show the correlation of the CTR component with the reference function and the corresponding quantization error per voxel.

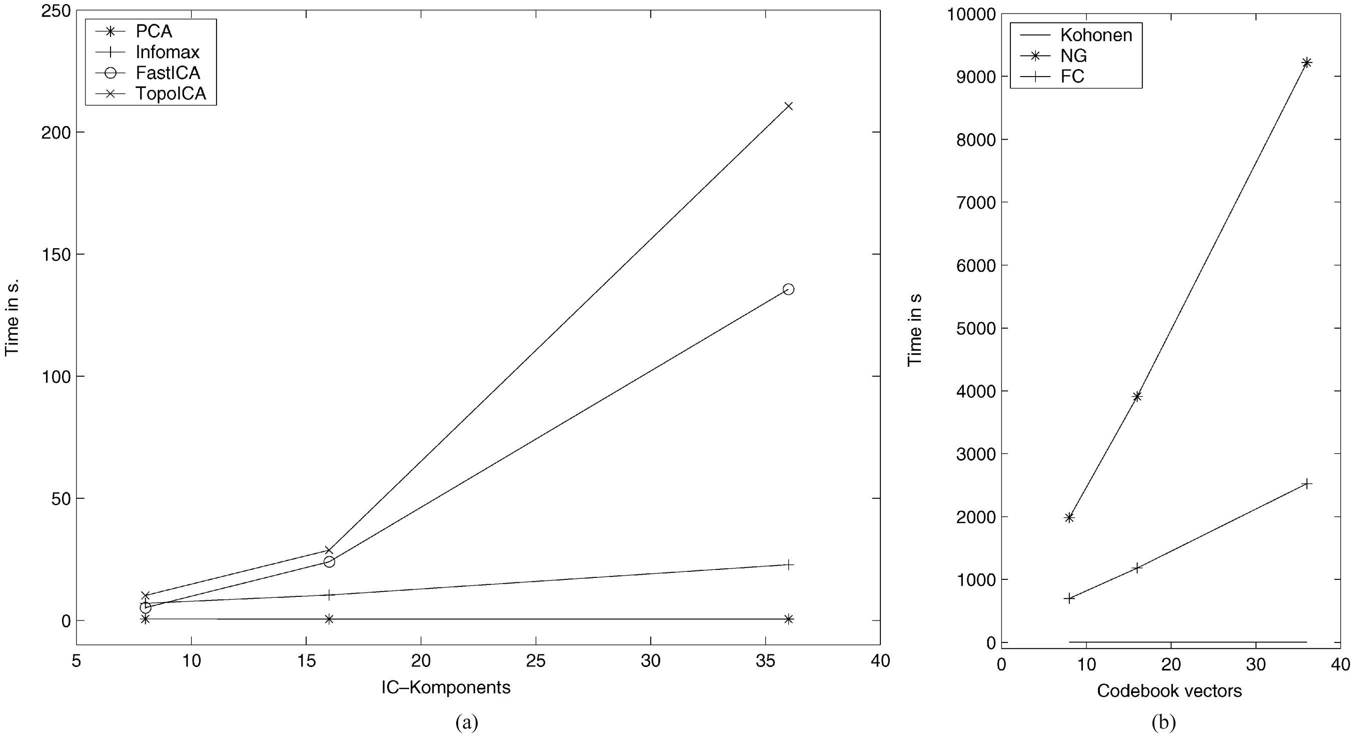


Fig. 5. Simulation time aspects for transformation-based and clustering techniques in function of the number of ICs or CVs. (a) Transformation-based techniques. (b) Clustering techniques.

The observed results suggest that the optimal number of target components regarding the maximum correlation is equal to 36 for clustering techniques, while for transformation-based techniques is equal to eight.

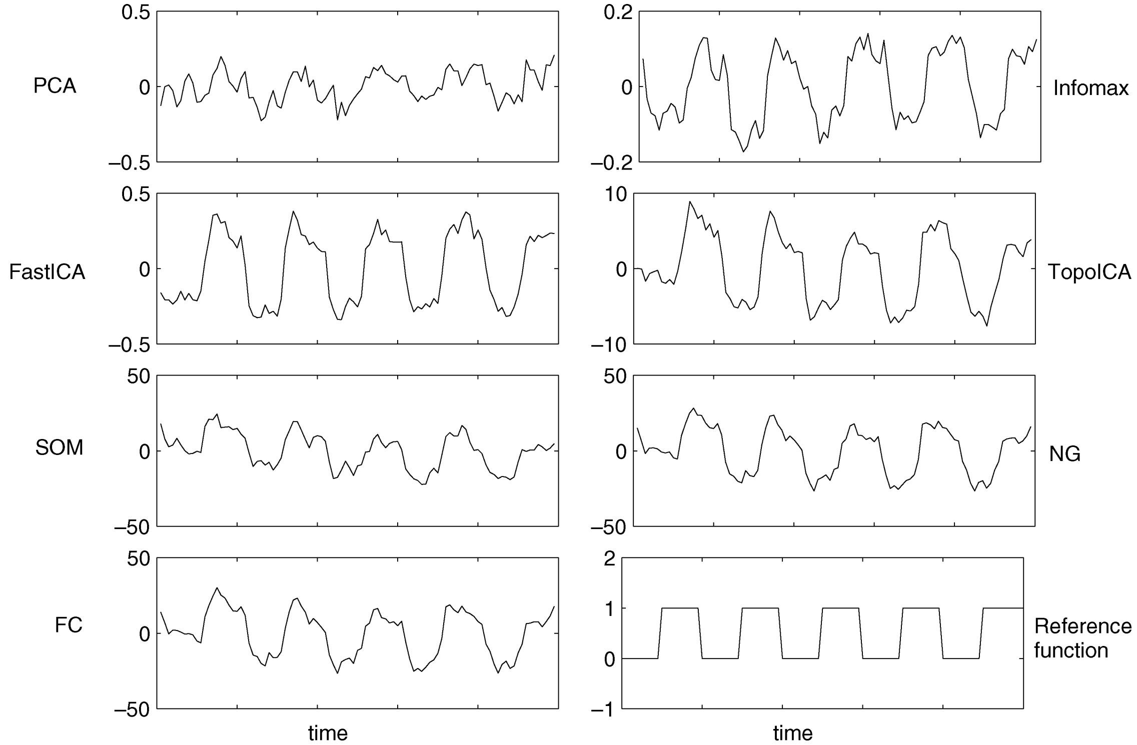


Fig. 6. Computed reference functions for the three clustering techniques (SOM, "neural gas" network, and fuzzy clustering based on deterministic annealing) for eight code vectors, and the transformation-based methods (PCA and ICA techniques) for eight different components. The determined correlation coefficients are: 0.69 (PCA), 0.85 (Infomax), 0.93 (FastICA), 0.86 (TopoICA), 0.72 (SOM), 0.83 ("neural gas"), 0.82 (fuzzy clustering based on deterministic annealing).

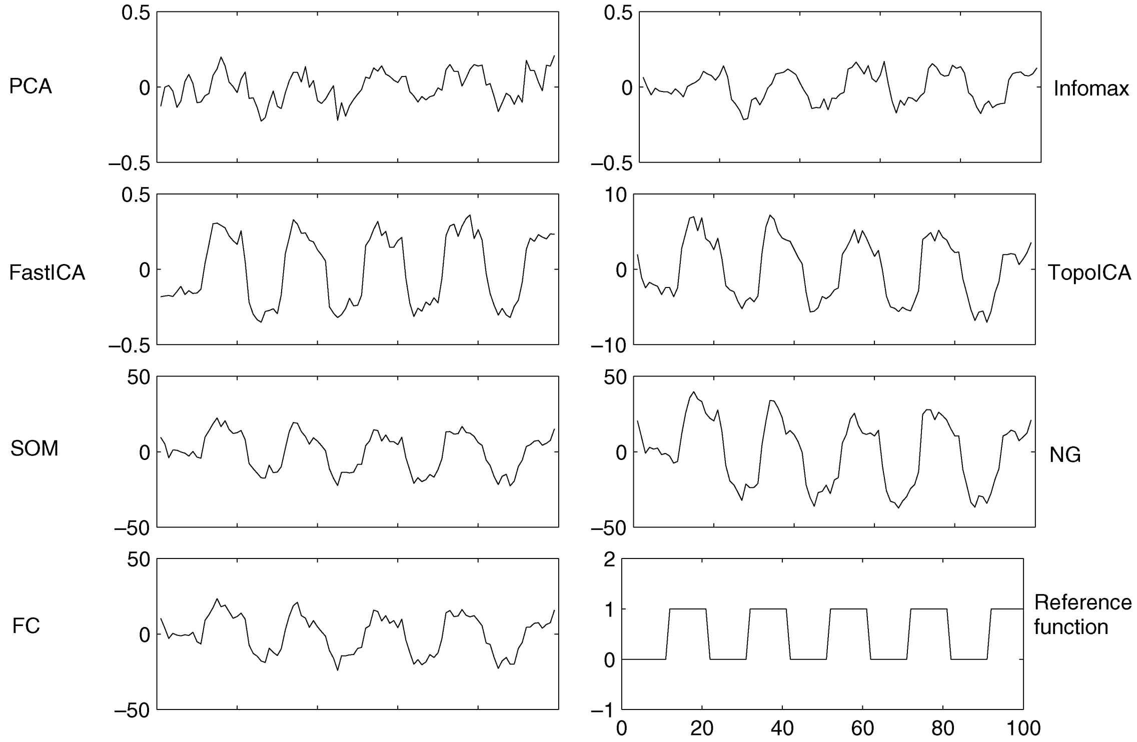


Fig. 7. Computed reference functions for the three clustering techniques (SOM, "neural gas" network, and fuzzy clustering based on deterministic annealing) for 36 code vectors, and the transformation-based methods (PCA, and ICA techniques) for 16 different components. The determined correlation coefficients are: 0.69 (PCA), 0.85 (Infomax), 0.93 (FastICA), 0.90 (TopoICA), 0.84 (SOM), 0.84 ("neural gas"), 0.84 (fuzzy clustering based on deterministic annealing).

An important aspect in real-time analysis is the required processing time associated with each class of techniques. A comparison between these techniques is given in Fig. 5(a) and (b). For the same number of target components, ICA techniques are faster than clustering techniques.

B. Correlation With the Reference Function

An interesting aspect can be observed if we compare the computed reference functions at the maximum correlation for both the clustering and the transformation-based techniques.

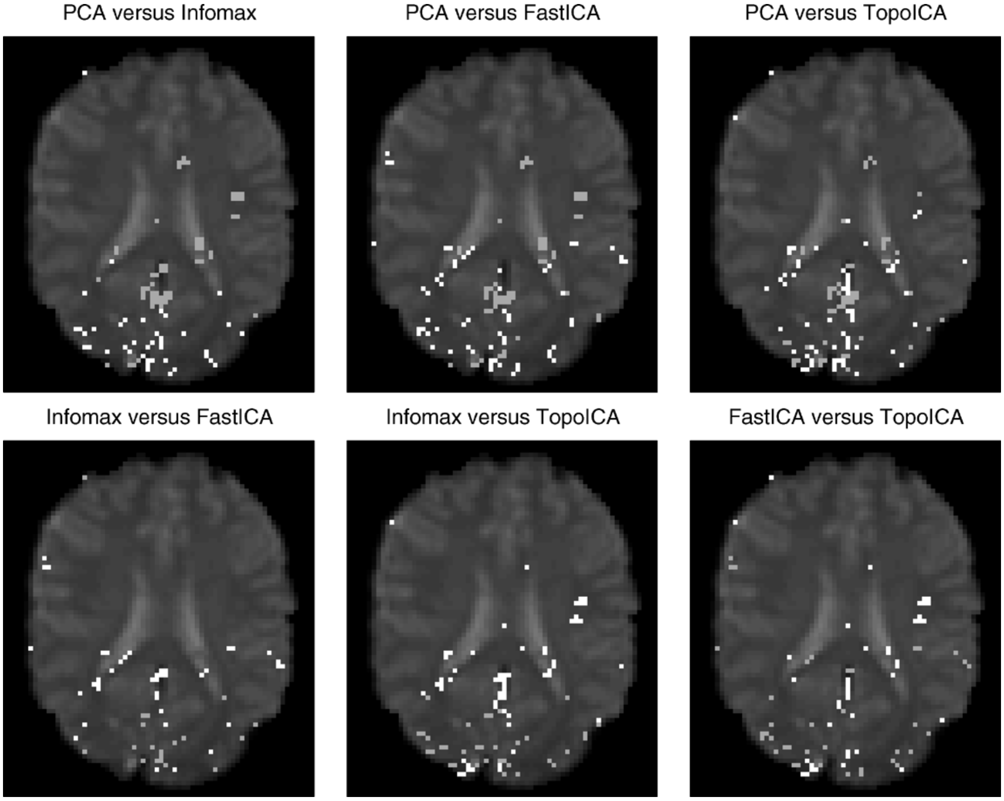


Fig. 8. Comparison of the results for the four techniques, PCA, Infomax, FastICA, and topographic ICA. The shown activation maps show only the pairwise differences. The darker pixels show active voxels for the first technique, while the lighter pixels show the active voxels for the second technique.

Figs. 6 and 7 visualizes the computed reference functions for all model-free methods. We can see that increasing the number of IC components leads to a deterioration of the computed reference function, while for the unsupervised cluster techniques the opposite is true. A larger number of code vectors improves the classifier's efficiency. The only exception pertains to the "neural gas" network: The correlation coefficient for 8 codevectors has almost the same value as that for 36 code vectors, thus suggesting that this clustering technique works very efficiently even at a lower resolution.

All CTR components obtained based on the clustering techniques show, in general, an important aspect: The curve drops during the first time step due to the presence of an artifact. Some of the ICA techniques are able to annihilate this artifact, providing, thus, the proof for the IC separation's feasibility. In general, the obtained correlations for the clustering methods are higher for a larger number of CV, suggesting that unsupervised clustering works more robustly.

Fig. 7 shows for 36 CVs the component time course most closely associated with the visual task for all two main techniques. The best results are achieved by the "neural gas" network and the fuzzy clustering based on deterministic annealing, yielding almost identical reference functions, and a correlation coefficient of $r = 0.9984$ between those two.

Both Figs. 6 and 7 show clearly an increase in sensitivity of the CTR components based on the unsupervised clustering methods. A sensitivity range of $[-50, 50]$ is common for all clustering techniques. Among the ICA techniques, topographic ICA shows the largest sensitivity range of $[-10, 10]$ compared to the remaining ICA techniques.

C. Activation Maps

To further differentiate in terms of detecting activation clusters, we take a look at the activation z maps of the CTR component for all model-free analysis methods.

Fig. 8 shows the activation maps as a comparison of results obtained by the four ICA techniques.

The striking similarity between the "neural gas" network and the fuzzy clustering based on deterministic annealing is also visualized in Fig. 9 which shows the activation maps as a comparison of results obtained by the three clustering techniques.

To further differentiate in terms of detecting activation clusters, we take a look at the differences in the activation z maps for topographic ICA versus "neural gas," and fuzzy clustering based on deterministic annealing, as illustrated by Fig. 10. Here again, it is evident that the topographic ICA unifies the strengths of the two other techniques, and that the resulting pairwise pixel differences are minimal and scattered.

This finding suggests that the relaxation of the independence condition for neighboring components in ICA applied to fMRI proves to be a better feature extraction method than traditional ICA relying solely on a strict independence.

D. ROC Analysis

It is important to perform a quantitative analysis of the relative performance of the introduced exploratory data analysis techniques. To do so, we compared the proposed algorithms for 8 and 36 components in terms of ROC analysis using correlation map with a chosen threshold of 0.4. We report the ROC performances for the five subjects in Fig. 11. The Figure illustrates the average area under the curve and its deviations for 20 different

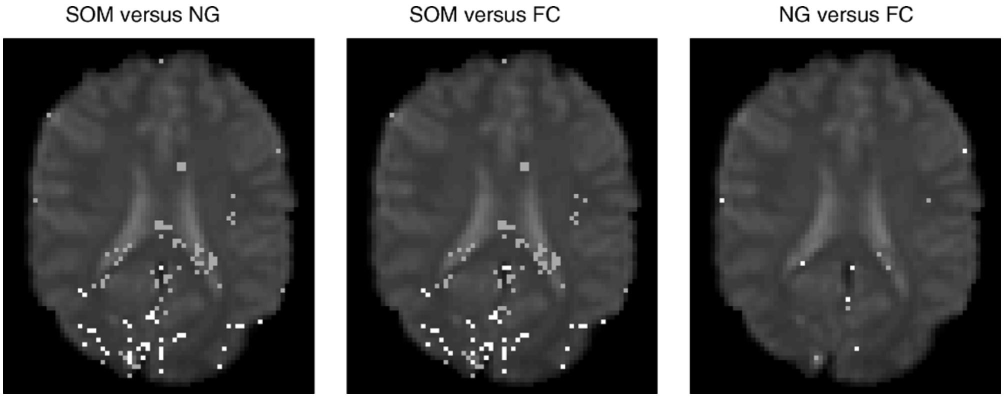


Fig. 9. Comparison of the results for the three techniques, SOM, "neural gas" network, and FC for 16 code vectors. The shown activation maps show only the pairwise differences for 16 code vectors. The darker pixels show active voxels for the first technique, while the lighter pixels show the active voxels for the second technique.

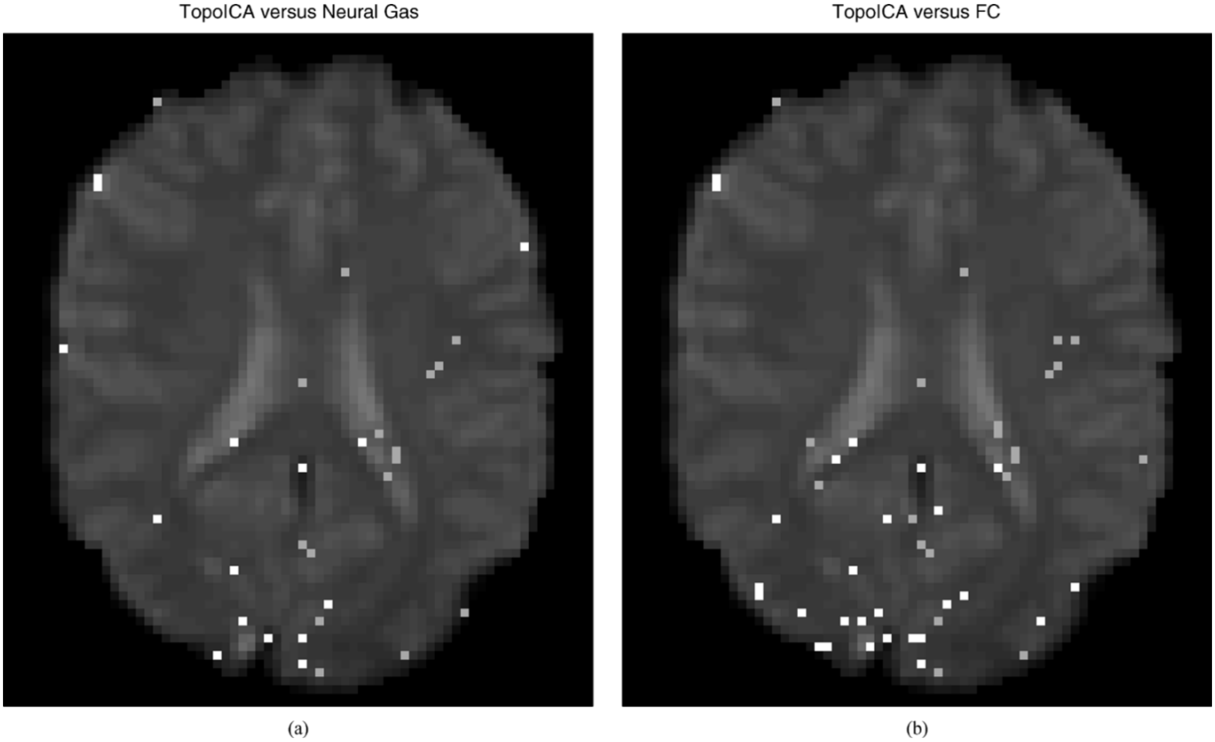


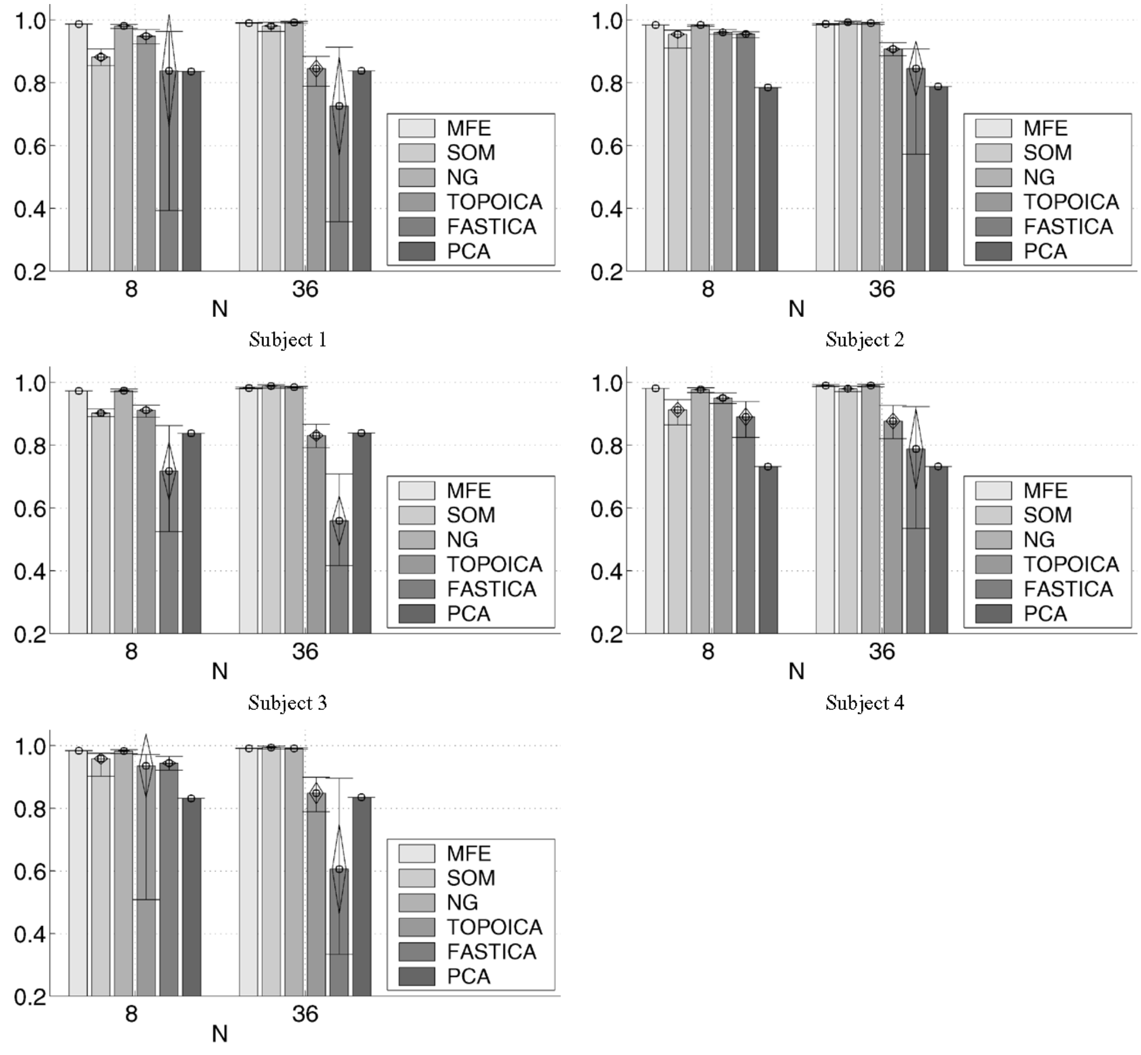
Fig. 10. Differences in the activation maps resulting from the CTR component between topographic ICA, "neural gas," and fuzzy clustering based on deterministic annealing. (a) TopoICA versus "neural gas" network, and (b) TopoICA versus fuzzy clustering based on deterministic annealing. The darker pixels show voxels being active with the first technique, while the lighter ones show only those active with the second technique. The difference activation maps show only the pairwise differences for eight chosen codevectors. The number of code vectors and of components was set equal to eight.

ROC runs using the same parameters but different algorithms' initializations. From Fig. 11, we can see that the clustering methods outperform for 36 components the ICA methods including PCA for all five subjects. For 8 components, we see that for most subjects SOM is outperformed by topographical ICA, while the other two clustering techniques achieve the best results.

IV. Conclusion

In this paper, we have experimentally compared two exploratory data analysis methods for fMRI: the ICA techniques versus unsupervised clustering. The ICA techniques were two standard ICA algorithms, the Infomax and the FastICA, and a new algorithm, the topographic ICA. The unsupervised clustering techniques were two proven clustering algorithms, the SOM and the fuzzy clustering based on deterministic annealing, and a less known algorithm, the "neural gas" network.

The goal of the paper was to determine the robustness and reliability of extracting task-related activation maps and time-courses from fMRI data sets. The success of ICA methods is based on the condition that the spatial distribution of brain areas activated by task performance must be spatially independent of the distributions of areas affected by artifacts. It was also shown that unsupervised clustering techniques represent a successful strategy for the analysis of time-courses from fMRI data sets. The increasing cluster resolution proved to reveal extremely well the structure of the data set. From the ROC



Subject 1

Subject 2

Subject 3

Subject 4

Subject 5

Fig. 11. Results of the comparison between the two different exploratory data analysis methods on fMRI data. Spatial accuracy of the different maps is assessed by ROC analysis using correlation map with a chosen threshold of 0.4. It illustrates the average area under the curve and its deviations for 20 different ROC runs using the same parameters but different algorithms' initializations. The number of chosen ICs or CVs for all techniques is equal to 8 and 36 and results are plotted for all five subjects.

analysis, we observe that for 36 components, the clustering methods outperform the transformation-based methods for all five subjects. For 8 components, we see that for most subjects SOM is outperformed by topographical ICA, while the "neural gas" network and fuzzy clustering based on deterministic annealing achieve the best results.

Both the "neural gas" network and fuzzy clustering based on deterministic annealing outperform ICA in terms of classification results and sensitivity range of the CTR-component but require a longer processing time than the ICA methods. Another important aspect is that topographical ICA represents a unifying paradigm between transformation-based and clustering techniques and, thus, bridges the discriminatory capabilities of FastICA and "neural gas" network. The relaxation of the independence condition for neighboring components leads to an increase in sensitivity range compared to standard ICA, and achieves in most cases, a higher correlation coefficient compared to the other ICA techniques.

The applicability of the new algorithm is demonstrated on experimental data.

Acknowledgment

The authors would like to thank Dr. D. Auer from the Max Planck Institute of Psychiatry in Munich, Germany, for providing the fMRI data. The authors greatly appreciate the programming efforts of T. D. Otto.

References

[1] S. Ogawa, D. Tank, and R. Menon, "Intrinsic signal changes accompanying sensory stimulation: Functional brain mapping with magnetic resonance imaging," in *Proc. Nat. Acad. Sci.*, vol. 89, 1992, pp. 5951–5955.

[2] S. Ogawa, T. Lee, and B. Barrere, "The sensitivity of magnetic resonance image signals of a rat brain to changes in the cerebral venous blood oxygenation activation," *Magn. Reson. Med.*, vol. 29, pp. 205–210, 1993.

[3] J. Sychra, P. Bandettini, N. Bhattacharya, and Q. Lin, "Synthetic images by subspace transforms I. Principal components images and related filters," *Med. Phys.*, vol. 21, pp. 193–201, 1994.

[4] W. Backfrieder, R. Baumgartner, M. Samal, E. Moser, and H. Bergmann, "Quantification of intensity variations in functional MR images using Rotated principal components," *Phys. Med. Biol.*, vol. 41, pp. 1425–1438, 1996.

[5] M. McKeown, T. Jung, S. Makeig, G. Brown, T. Jung, S. Kindermann, A. Bell, and T. Sejnowski, "Spatially independent activity patterns in functional magnetic resonance imaging data during the stroop color-naming task," in *Proc. Nat. Acad. Sci.*, vol. 95, 1998, pp. 803–810.

[6] ——, “Analysis of fMRI data by blind separation into independent spatial components,” *Human Brain Mapping*, vol. 6, pp. 160–188, 1998.

[7] F. Esposito, E. Formisano, E. Seifritz, R. Goebel, R. Morrone, G. Tedeschi, and F. Di Salle, “Spatial independent component analysis of functional MRI time-series: To what extent do results depend on the algorithm used?,” *Human Brain Mapping*, vol. 16, pp. 146–157, 2002.

[8] K. Arfanakis, D. Cordes, V. Haughton, C. Moritz, M. Quigley, and M. Meyerand, “Combining independent component analysis and correlation analysis to probe interregional connectivity in fMRI task activation datasets,” *Magnetic Resonance Imaging*, vol. 18, pp. 921–930, 2000.

[9] B. Biswal and J. Ulmer, “Blind source separation of multiple signal sources of fMRI data sets using independent component analysis,” *J. Comput. Assisted Tomography*, vol. 23, pp. 265–271, 1999.

[10] G. Scarth, M. McIntrye, B. Wowk, and R. Samorjai, “Detection novelty in functional imaging using fuzzy clustering,” in *Proc. SMR 3rd Annu. Meeting*, vol. 95, 1995, pp. 238–242.

[11] K. Chuang, M. Chiu, C. Lin, and J. Chen, “Model-free functional MRI analysis using Kohonen clustering neural network and fuzzy C-means,” *IEEE Trans. Med. Imag.*, vol. 18, pp. 1117–1128, Dec. 1999.

[12] R. Baumgartner, L. Ryder, W. Richter, R. Summers, M. Jarmasz, and R. Somorjai, “Comparison of two exploratory data analysis methods for fMRI: Fuzzy clustering versus principal component analysis,” *Magnetic Resonance Imaging*, vol. 18, pp. 89–94, 2000.

[13] A. Wismüller, O. Lange, D. Dersch, G. Leinsinger, K. Hahn, B. Pütz, and D. Auer, “Cluster analysis of biomedical image time-series,” *Int. J. Comput. Vision*, vol. 18, pp. 102–128, 2002.

[14] A. Hyvarinen, “Fast and robust fixed-point algorithms for independent component analysis,” *IEEE Trans. Neural Networks*, vol. 10, pp. 626–634, May 1999.

[15] A. Hyvarinen and E. Oja, “Independent component analysis: Algorithms and applications,” *Neural Netw.*, vol. 13, pp. 411–430, 2000.

[16] A. Hyvarinen and P. Hoyer, “Topographic independent component analysis,” *Neural Comput.*, vol. 13, pp. 1527–1558, 2001.

[17] A. J. Bell and T. J. Sejnowski, “An information-maximization approach to blind separation and blind deconvoluti on,” *Neural Comput.*, vol. 7, pp. 1129–1159, 1995.

[18] S. Amari, A. Cichocki, and H. H. Yang, “A new learning algorithm for blind signal separation,” in *NIPS’96*, vol. 8, 1996, pp. 757–763.

[19] K. Rose, E. Gurewitz, and G. Fox, “Vector quantization by deterministic annealing,” *IEEE Trans. Inform. Theory*, vol. 38, pp. 1249–1257, July 1992.

[20] R. Woods, S. Cherry, and J. Mazziotta, “Rapid automated algorithm for aligning and reslicing PET images,” *J. Comput. Assisted Tomography*, vol. 16, pp. 620–633, 1992.

[21] S. Ngan and X. Hu, “Analysis of fMRI imaging data using self-organizing mapping with spatial connectivity,” *Magn. Reson. Med.*, vol. 41, pp. 939–946, 1999.

[22] T. Martinetz, S. Berkovich, and K. Schulten, “Neural gas network for vector quantization and its application to time-series prediction,” *IEEE Trans. Neural Networks*, vol. 4, pp. 558–569, July 1993.

[23] H. Fisher and J. Hennig, “Clustering of functional MR data,” in *Proc. ISMRM 4th Annu. Meeting*, vol. 96, 1996, pp. 1179–1183.

A. Meyer-Baese received the Ph.D. degree in electrical and computer engineering from Darmstadt University of Technology, Germany, in 1995.

After four postdoctoral years, one at the Federal Institute of Neurobiology in Magdeburg, Germany, and three at the University of Florida, Gainesville, she joined the faculty at the Department of Electrical and Computer Engineering, Florida State University, Tallahassee. Her research areas include theory and application of neural networks, medical image processing, pattern recognition, and parallel processing. She published over 100 papers in several areas ranging from intelligent systems, medical image processing, signal processing, and neural networks. She is author of two books, one in pattern recognition in medical imaging in Elsevier Science/Academic Press.

Axel Wismueller studied medicine at the Technical University of Munich, Munich, Germany and the University of Regensburg, Germany, and with study exchange programs in Switzerland (Kantonsspital Winterthur) and the U.S. (Yale University). He received the M.D. degree from the Technical University of Munich for a scientific thesis in neurology in 1992. He successfully passed the U.S. medical examinations ECFMG and FLEX. In parallel to his clinical work in internal medicine, he studied physics at the University of Munich where he received a German masters degree (Dipl.-Phys. Univ.) in theoretical physics in 1996 for a scientific thesis on pattern recognition.

Since 1997, he has been working as a Fellow of Radiology in the Department of Clinical Radiology at the University of Munich, where he founded the Digital Image Processing Group. His main research interest is focussed on both the theory of intelligent and self-organizing systems for pattern analysis and its application to real-world biomedical image and signal processing in bioinformatics, radiology, and nuclear medicine, with specific projects on functional MRI for human brain mapping and the diagnosis of breast cancer in MRI mammography. He is the author of more than 70 scientific journal and conference publications related to pattern recognition and biomedicine.

Oliver Lange received the Ph.D. degree in biomedical sciences from the University of Munich, Munich, Germany, in 2004.

In 1998, he joined the Digital Image Processing Group and spent one year as a Research Scientist at the Florida State University where he did research in applying pattern recognition methods to fMRI, breast MRI, and dynamic perfusion MRI. His research interests include biomedical image and signal processing, pattern recognition, and machine learning. He is author of more than 20 articles all in biomedical imaging.