Cluster Analysis of Biomedical Image Time-Series

Axel Wismüller[1], Oliver Lange[1],
Dominik R. Dersch[2], Gerda L. Leinsinger[1],
Klaus Hahn[1], Benno Pütz[3], and Dorothee Auer[3]

[1]Institut für Radiologische Diagnostik,
Ludwig-Maximilians-Universität München,
Klinikum Innenstadt, Ziemssenstr. 1, 80336 München, Germany
e-mail: Axel.Wismueller@physik.uni-muenchen.de

[2]Crux Cybernetics Corp., Sydney, Australia

[3]Max Planck Institute of Psychiatry, Munich, Germany

This is a preprint version.
Please use the following reference for citation:

A. Wismüller, O. Lange, D.R. Dersch,
G.L. Leinsinger, K. Hahn, B. Pütz, and D. Auer:
Cluster Analysis of Biomedical Image Time-Series.
International Journal of Computer Vision, 46(2):103-128, 2002.

Abstract

In this paper, we present neural network clustering by deterministic annealing as a powerful strategy for self-organized segmentation of biomedical image time-series data identifying groups of pixels sharing common properties of local signal dynamics. After introducing the theoretical concept of minimal free energy vector quantization and related clustering techniques, we discuss its potential to serve as a multi-purpose computer vision strategy to image time-series analysis and visualization for many fields of medicine ranging from biomedical basic research to clinical assessment of patient data. In particular, we present applications to (i) functional MRI data analysis for human brain mapping, (ii) dynamic contrast-enhanced perfusion MRI for the diagnosis of cerebrovascular disease, and (iii) magnetic resonance mammography for the analysis of suspicious lesions in patients with breast cancer. This wide scope of completely different medical applications illustrates the flexibility and conceptual power of neural network vector quantization in this context. Although there are obvious methodological similarities, each application requires specific careful consideration w.r.t. data preprocessing, postprocessing and interpretation. This challenge can only be managed by close interdisciplinary cooperation of medical doctors, engineers, and computer scientists. Hence, this field of research can serve as an example for lively cross-fertilization between computer vision and related research.

Keywords: clustering, time-series, neural networks, deterministic annealing, medical imaging, functional MRI, dynamic perfusion MRI, MRI mammography.

This is a preprint version. Please use the following reference for citation:
A. Wismüller, O. Lange, D.R. Dersch, G.L. Leinsinger, K. Hahn, B. Pütz, and D. Auer. Cluster analysis of biomedical image time-series. *International Journal of Computer Vision*, 46(2):103-128, 2002.

1 Introduction

In recent years a wide scope of advanced non-invasive medical imaging techniques such as positron emission tomography (PET), dynamic computer tomography (CT), and magnetic resonance imaging (MRI) has been introduced into biomedical practice.

Beyond the plain imaging of morphological structure, the analysis of biological function increasingly moves into the focus of attention. Thus, the analysis and visualization of medical image time-series data is a challenge with growing importance for both basic research and clinical application.

The basic principle of dynamic biomedical imaging can be summarized as follows: after administration of some contrast agent, one can observe, as a function of time, contrast agent-induced dynamic changes with regard to the imaging properties of the examined tissue. These changes can be recorded by a series of rapid subsequent images. The term 'contrast agent' should be considered very general here, as it comprises both exogenous and endogenous effects: It ranges from exogenous intravenous administration of iodinated or paramagnetic contrast media by a medical doctor to the concentration shifts of endogenous substances that may be induced by functional changes in biological tissue.

For each pixel, we obtain a signal time-series that may be interpreted in the light of the experimental conditions. There are two approaches to the analysis of dynamic biomedical image time-series data: The first one is a model-driven approach that tries to interprete the time-series data as a consequence of dynamically changing experimental input parameters, given a specific model that defines the pattern of interaction between underlying physiological processes and the observed signal dynamics. This "supervised" cause and effect concept is the mainstream approach to the interpretation of biomedical image time-series data. Important examples for the imaging procedures discussed in this paper are:

(i) Correlation-based methods of functional MRI (fMRI) data analysis (e. g. [3]) of the human brain compare gray levels and stimulus protocol at each time frame of image acquisition in the fMRI experiment. Thus, they evaluate the similarity of the signal time-series at each pixel (Pixel Time Course, PTC) and the protocol time-series employing an appropriate distance measure like cross-correlation or Euclidean distance in order to define regions of 'activation'. The underlying conceptual model is based on the so-called BOLD- (blood oxygen level dependent) contrast that was first described by [40]. The physical basis of this BOLD contrast is given by the different magnetic properties of deoxyhemoglobin (DeoxHb, paramagnetic) and oxyhemoglobin (OxHb, diamagnetic) that lead to differences in signal intensity in appropriately susceptibility weighted MRI sequences depending on the concentration of deoxyhemoglobin.

The physiological nature of the complex temporo-spatial pattern of the BOLD response to sensory, motor, or cognitive brain activation, however, is only partially understood and currently subject to thorough investigation. In principle, the BOLD signal shows a regionally varying tri-phasic response: initial signal decrease, delayed signal increase for the duration of the stimulus followed by a post-stimulation decrease. For standard fMRI experiments, the signal increase phase is exploited using long time-series during the resting and activation state to account for the typically low contrast-to-noise ratio. The main objective of fMRI data processing is to segment active from non-active brain regions by analyzing the dynamics of these time-series.

(ii) In dynamic contrast-material enhanced MRI studies, such as dynamic cerebral perfusion MRI or MRI mammography, the contrast-agent uptake can be studied as a function of time. From the resulting tissue-concentration-time curves, perfusion parameters can be computed (e.g. [41]). The main objective of dynamic MRI is to separate abnormal from normal contrast enhancement reflecting disturbed hemodynamics or pathological vascularization.

The model-based, supervised data analysis based on the underlying knowledge of experimental conditions and model assumptions, may be difficult or even impossible. For the examples sketched above, we have to face the following limitations:

(i) Supervised techniques for fMRI data analysis imply presumptive knowledge of expected stimulus-response patterns. Thus, they may sometimes fail in unveiling complex signal changes that exceed the scope of the model, thus discarding valuable information about the fMRI signal. Moreover, in fMRI studies of spontaneously occurring mental processes like hallucinations, sleep, or epileptic fits, the exact time course of the stimulus may not even be known. This also holds for the analysis of spontaneous brain activity in resting state fMRI.

(ii) Exact blood flow quantification based on the indicator dilution theory for nondiffusible intravascular tracers (e.g. [41]) is a difficult issue which cannot be performed on a routine clinical basis. Under pathological conditions, however, the computation of perfusion parameters is even impossible, as the conditions for the applicability of the indicator dilution theory may be violated.

In the context of these limitations, *unsupervised* or solely *data-driven* techniques may offer a powerful alternative strategy for image time-series analysis. A simple approach is to use appropriate *transformations,* i. e. to apply mappings that help to unveil inherent properties of the data. Principal Component Analysis (PCA) for example, performs a basis transformation by diagonalization of the covariance matrix and rotating the coordinate system according to the direction of the resulting

eigenvectors. The corresponding (squared) eigenvalues represent the variances of the data along these principal components. In fMRI data analysis, for example, usually some of the eigenvectors reflect the time course of the stimulus protocol. Subsequently, a rough preliminary screening for pixels representing areas of 'activation' may be performed by computing maps of the coefficients of each PTC with respect to these principal components. Recently, an additional component decomposition method has been introduced into fMRI data processing: Independent Component Analysis (ICA) tries to identify how observed data can be generated by a process of linearly mixing statistically independent basic signals [30], [6].

A different approach to unsupervised image time-series data analysis is *clustering* or Vector Quantization (VQ). In this paper, we use the terms VQ and clustering interchangeably. VQ methods have been proposed for a wide scope of biomedical signal processing problems (see, e. g. [51]). VQ procedures map a data space onto a finite set of prototypical feature vectors, a so-called codebook. Examples of this class of algorithms are elementary VQ methods like LBG [35] and k-means clustering [1], [19], [23] or more refined, 'soft-competing' algorithms like Kohonen's self-organizing maps (SOMs) [33], minimal free energy VQ [45], [15], [14], [12], and the 'neural gas' algorithm [36].

Interestingly, there are different approaches to the latter class of VQ algorithms: a neurobiology-oriented approach based on algorithms which model the development of sensory maps [50], an alternative approach based on information theory [5], and approaches derived from statistics and statistical mechanics [44]. The mathematical properties of these algorithms as well as their drawbacks have been thoroughly investigated in the literature (see, e. g. [32], [42], [21], [20], [44], [15], [16], [12], [45], [37]).

In the field of fMRI data analysis, various clustering methods have been introduced by several authors [18], [47], [48], [17], [26], [25] [22], [27], [10], [39], [4] [53], [54]. For VQ analysis, the PTCs of gray levels at each pixel obtained from biomedical time-series data can be interpreted as feature vectors sampled from a multidimensional probability distribution. VQ methods try to identify clusters of similar PTCs within this feature space. The corresponding Codebook Vectors (CVs) represent prototypical PTCs sharing similar temporal characteristics. As a result, each PTC can be assigned to a specific CV according to a minimal distance criterion, or, in the case of soft-competing algorithms, be attributed to several CVs with respect to a 'fuzzy' cluster membership. Accordingly, the output of VQ procedures for medical image time-series data may be plotted as 'hard' or 'soft' cluster assignment maps.

In this paper, we present neural network clustering techniques as a multi-purpose approach to image time-series analysis that can be applied to many fields of medicine ranging from biomedical basic research to clinical assessment of patient data. In particular, we will present applications to (i) fMRI data analysis for human brain mapping, (ii) dynamic contrast-enhanced perfusion MRI for the diagnosis of cere-

brovascular disease (iii) magnetic resonance mammography for the analysis of suspicious lesions in patients with breast cancer. This wide scope of completely different medical applications illustrates the flexibility and conceptual power of neural network vector quantization. Although there are obvious methodological similarities, each application requires specific careful consideration with regard to data preprocessing, postprocessing and interpretation. This challenge can only be met by close interdisciplinary cooperation of medical doctors, engineers, and computer scientists. Hence, this field of research can serve as an example for lively cross-fertilization between computer vision and related research.

2 Theory

Let n denote the number of subsequent scans in a biomedical image time-series. Let K denote the number of pixels in each scan. The dynamics of each pixel $\nu \in \{1, \ldots, K\}$, i.e. the sequence of gray levels $x^\nu(\tau)$ over all scan acquisition time frames τ can be interpreted as a vector $\mathbf{x}^\nu \in \mathbb{R}^n$ in the n-dimensional feature space of possible PTCs. Clustering identifies N groups j, $j \in \{1, \ldots, N\}$ of pixels with similar PTCs. These groups are represented by prototypical time-series called Codebook Vectors (CVs) $\mathbf{w}_j$ located at the center of the corresponding clusters.

Some VQ procedures determine these cluster centers by an iterative adaptive update according to

$$\mathbf{w}_j(t+1) = \mathbf{w}_j(t) + \epsilon(t)\, a_j(\mathbf{x}(t), W(t), \kappa)\, (\mathbf{x}(t) - \mathbf{w}_j(t)), \tag{1}$$

where $\epsilon(t)$ denotes a learning parameter, a_j a so-called cooperativity function which, in general, depends on the codebook $W(t) = \{\mathbf{w}_j(t) | j \in \{1, \ldots, N\}\}$, a cooperativity parameter κ, and a, in general, randomly chosen feature vector $\mathbf{x} \in \{\mathbf{x}^\nu | \nu \in \{1, \ldots, K\}\}$. In the context of biomedical image data analysis, one should clearly note the difference between the time τ of the time-series representing a specific component of the feature vector $\mathbf{x}^\nu$ and the time t which denotes the iteration step of the VQ procedure.

It should be emphasized that not all clustering methods can be written in a sequential form (1) which has only been chosen here because of its intuitive simplicity. The implementation of all the VQ algorithms used in this paper is indeed non-sequential, i.e. all the feature vectors $\mathbf{x}$ are presented in a single batch cycle. This allows the application of well-defined optimization strategies from statistical learning theory, e.g. the Expectation Maximization (EM) algorithm [11] for parameter estimation.

VQ algorithms can be classified according to the cooperativity function a_j. One can distinguish between so-called **'hard clustering'** methods that assign each fea-

ture vector $\mathbf{x}$ to exactly one cluster and **'soft clustering'** methods implying the concept of fuzzy membership of feature vectors in several clusters.

For example, a simple method for hard clustering is the algorithm proposed by Linde, Buzo and Gray (LBG) [35]. Here, a_j is calculated by the 'winner takes all' learning rule

$$a_j(\mathbf{x}(t), W(t)) := \delta_{i(\mathbf{x}),j}, \tag{2}$$

where $i(\mathbf{x})$ is determined according to the minimal distance criterion

$$\|\mathbf{x} - \mathbf{w}_i\| = \min_j \|\mathbf{x} - \mathbf{w}_j\|. \tag{3}$$

If we change a_j in a way that more than one CV can take part in the update at the same time t, (1) becomes a 'winner takes most' learning rule. The resulting soft clustering methods can be classified according to the metric the cooperativity function a_j is based on.

If we do **not** use the metric of the feature space itself, but define a mapping

$$\mathbf{r} : \mathbb{N} \to Y \subset \mathbb{R}^k, \quad j \mapsto \mathbf{r}(j) \tag{4}$$

of the CVs onto a so-called 'cortex model', we obtain a scenario that may be interpreted with respect to a neural network model of sensory maps (e. g. [50]). Usually k is chosen $k \ll N$ and $k \ll n$. The $\mathbf{r}(j)$'s can be thought of as representing the physical location of neurons in the cortex model. To avoid misunderstandings, the term "cortex model" does not refer to the fMRI data space discussed later in this paper. A common choice would be the ordering of the $\mathbf{r}_j$'s on a regular low-dimensional grid ($k \in \{2, 3\}$). However, this is not mandatory (see e. g. [56]). By choosing the metric of the cooperativity function a_j according to the neuron positions $\mathbf{r}_j$ in the target space, we obtain Kohonen's SOM algorithm [32], [33]. a_j is then chosen as a monotonically declining function of the distance to the 'winner neuron' $\mathbf{r}'(\mathbf{x}(t))$ on the cortex model. The winner neuron is determined according to (3). Thus, the cooperativity function $a_j(\mathbf{x}(t), W(t), \kappa)$ can be written as $a(\mathbf{r}(j), \mathbf{r}'(\mathbf{x}(t)), \sigma(t))$, where $\mathbf{r}(j)$ denotes the position of neuron j on the cortex model and $\sigma(t)$ a distance in the metric of the cortex model. Usual choices are Gaussian cooperativity functions

$$a(\mathbf{r}(j), \mathbf{r}'(\mathbf{x}(t)), \sigma(t)) := \exp\left(-\frac{(\mathbf{r} - \mathbf{r}'(\mathbf{x}(t)))^2}{2\sigma(t)^2}\right) \tag{5}$$

or characteristic functions on a k-dimensional hypersphere around $\mathbf{r}'(\mathbf{x}(t))$

$$\begin{aligned} a_j(\mathbf{r}(j), \mathbf{r}'(\mathbf{x}(t)), \sigma(t)) &:= \chi_{\|\mathbf{r}-\mathbf{r}'(\mathbf{x}(t))\| \le \sigma(t)} \\ &:= \begin{cases} 1 & : \quad \|\mathbf{r} - \mathbf{r}'(\mathbf{x}(t))\| \le \sigma(t) \\ 0 & : \quad \|\mathbf{r} - \mathbf{r}'(\mathbf{x}(t))\| > \sigma(t). \end{cases} \end{aligned} \tag{6}$$

The update rule (1) then becomes

$$\mathbf{w}_j(t+1) = \mathbf{w}_j(t) + \epsilon(t)\, a_j(\mathbf{r}(j), \mathbf{r}'(\mathbf{x}(t)), \sigma(t))\, (\mathbf{x}(t) - \mathbf{w}_j(t)). \tag{7}$$

$\sigma(t)$ denotes a concrete choice of the cooperativity parameter κ in (1). It is a measure for the 'range' of the cooperativity function $a_j(\mathbf{r}(j), \mathbf{r}'(\mathbf{x}(t)), \sigma(t))$ on the cortex model. It is usually updated according to some heuristic annealing scheme, as is the learning parameter $\epsilon(t)$. A common strategy is an exponential decay, i. e.

$$\sigma(t) = \sigma(0) \left(\frac{\sigma(t_{\max})}{\sigma(0)} \right)^{\frac{t}{t_{\max}}}, \quad t \in [0, t_{\max}]. \tag{8}$$

and

$$\epsilon(t) = \epsilon(0) \left(\frac{\epsilon(t_{\max})}{\epsilon(0)} \right)^{\frac{t}{t_{\max}}}, \quad t \in [0, t_{\max}]. \tag{9}$$

In contrast to algorithms in which the cooperativity function is based on the metric of the feature space like k-means type clustering or minimal free energy VQ, the SOM algorithm has the interesting property of *topology preservation*, i. e. neighboring points of the feature space are mapped onto neighboring neurons of the cortex model. In many fields of data analysis, this allows for the construction of graphically appealing 2D maps that represent the CVs in some topological order [32]. However, this order may be misleading due to possible folding of the feature map (see e. g. [43]) in cases where the data are located on manifolds with a dimension greater than k, a common situation with biomedical time-series data ($n \gg 2$).

Recently, a generative topographic mapping (GTM) technique has been proposed [9], [8] as a principled alternative to SOMs. Although conceptually motivated by the SOM algorithm, this technique avoids several of its theoretical disadvantages, e.g. the non-existence of an objective function [20] or the lack of a convergence proof.

If we do not use a mapping like (4) of the CVs onto a cortex model but apply a metric of the feature space a_j directly, e.g.

$$a_j(\mathbf{x}(t), W(t), \kappa \equiv \rho(t)) = \frac{\exp(-E_j(\mathbf{x}(t))/2\rho^2)}{\mathcal{Z}}, \tag{10}$$

we obtain the cooperativity function for the soft clustering scheme proposed by Rose, Gurewitz, and Fox [45].

Here, the 'error' $E_j(\mathbf{x}(t)) = \|\mathbf{x}(t) - \mathbf{w}_j(t)\|^2$ measures the distance between the codebook vector $\mathbf{w}_j$ and the data vector $\mathbf{x}$. $\mathcal{Z}$ denotes a partition function given by

$$\mathcal{Z} = \sum_j \exp(-E_j(\mathbf{x})/2\rho^2), \tag{11}$$

and ρ is the cooperativity parameter of the model. This so-called 'fuzzy range' $\rho(t)$ defines a length scale in the feature space, not in a cortex model like $\sigma(t)$ in

(5) or (6). Noting the analogy to statistical mechanics, ρ can be interpreted as the temperature T in a multiparticle system employing $T = 2\rho^2$. It is annealed to repeatedly smaller values in the VQ procedure, e. g. according to an annealing scheme (8).

To avoid misunderstandings, equation (11) does not make the assumption that all the feature vectors are represented by a single Gaussian probability distribution, but that they can be estimated in the sense of a Gaussian mixture model. It can be shown that the partitions obtained by this algorithm fulfill the conditions of a maximum likelihood estimation based on a set of univariate Gaussian distributions that are constrained to identical width for each CV in each annealing step. Of course, by skipping this restriction one may think of further increasing the approximation quality achieved by Gaussian mixture models by multiplying the number of model parameters, e.g. by introducing mixtures of multivariate Gaussian distributions. A classical approach for this is the straight-forward maximum likelihood estimation of Gaussian mixture models using arbitrary multivariate Gaussian distributions (e.g. based on the well-known EM algorithm [11] mentioned above). This concept suffers from stability problems with regard to singular solutions, as has already been pointed out by [19]. It has been shown [31] that the clustering algorithm based on the cooperativity function (10), in contrast, does not suffer from such stability problems and is more robust against being captured in local minima of the likelihood function. For a discussion of this tradeoff between model complexity, estimation precision, and stability issues with regard to Gaussian mixture models, we refer to the literature, e.g. [19], [11], [31].

The learning rule (1) with a_j given by (10) describes a stochastic gradient descent on the error function

$$F_\rho(W) = \langle \sum_j a_j(\mathbf{x})(\mathbf{x} - \mathbf{w}_j)^2 \rangle + 2\rho^2 \langle \sum_j a_j(\mathbf{x}) \ln(a_j(\mathbf{x})) \rangle, \tag{12}$$

which is a free energy in a mean-field approximation [13], [45]. Here, $\langle . \rangle$ denotes the expectation value. The values of the cooperativity function a_j are normalized

$$\sum_{j=1}^{N} a_j(\mathbf{x}, W, \rho) = 1 \tag{13}$$

and can be interpreted as the assignment probabilities of a given feature vector $\mathbf{x}$ to each of the CVs $\mathbf{w}_j$, $j \in \{1, \ldots, N\}$ [45].

The CVs $\mathbf{w}_j$ mark local centers of the multidimensional probability distribution $f(\mathbf{x})$. Thus, for the application to biomedical signal analysis, the CV $\mathbf{w}_j$ is the weighted average time-series of all the PTCs $\mathbf{x}^\nu$ belonging to cluster j with respect to a fuzzy tessellation of the feature space according to (10).

In contrast to SOMs, minimal free energy VQ

(i) can be described as a stochastic gradient descent on an explicitly given energy function (see (12)) [45],

(ii) preserves the probability density without distortion as discretization density, i. e. the number of CVs, grows towards infinity [15], [16], and, most important for practical applications,

(iii) allows for hierarchical data analysis on different scales of resolution with regard to the metric of the feature space [14].

Furthermore, the procedure can be monitored by various control parameters like the free energy, entropy, reconstruction error, etc. which enable easy detection of cluster splitting. Properties of resulting optimal codebooks have been thoroughly investigated [15] and allow for a self-control process of the VQ procedure with respect to theoretically proven conservation laws [14].

Fig. 1 illustrates the application of hierarchical minimal free energy VQ clustering to a simple two-dimensional toy example.

As the 'fuzzy range' ρ declines in the course of the annealing scenario, the VQ procedure passes a sequence of cluster splitting phase transitions, until finally a fine-grained tessellation of the feature space is obtained.

Tracing the repetitive cluster splitting through the whole VQ procedure leads to a 'genealogy' of cluster centers, i. e. a resemblance tree of codebook vectors. At the same time, the scope of clustering resolution can be adapted according to the observer's needs.

3 Methods

3.1 Imaging Protocol

Functional MRI

Functional imaging was performed in 6 healthy subjects (3 male, 3 female, age range 20 – 37years) on a 1.5 T system (Signa™, General Electrics, Milwaukee) using a gradient echo echoplanar imaging (EPI) sequence. In five subjects, five slices with 100 images (TR/TE = 3000/60msec) were acquired with five periods of rest and five photic stimulation periods starting with rest. Stimulation and rest periods comprised 10 repetitions each, i.e. 30 s. In one subject eight slices with 64 images (TR/TE= 4000/66 msec) were obtained starting with a period of rest lasting 10 repetitions (i.e. 40 s) followed by three periods of stimulation alternating with three periods of rest comprising 9 repetitions each, i.e. 36 s. Resolution was $3 \times 3 \times 4$ mm. The slices were oriented parallel to the calcarine fissure. A respective pilot scan including the acquired slice positions is shown in fig. 2. The data analysis results in

this paper are plotted for the central slice in the 100 scan experiments and for slice #4 counted from top to bottom in the 64 scan experiment. 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. The first scan was discarded for remaining saturation effects. Motion artifacts were compensated by automatic image alignment (AIR 3.08, [57]).

Dynamic Contrast-Enhanced Cerebral Perfusion MRI

Clustering results are presented for dynamic cerebral perfusion MRI data sets from two patients: (i) One man aged 37 years without focal neurological deficit. He was referred to MRI to rule out perfusion abnormality. (ii) One female patient aged 66 years with subacute stroke (3 days elapsed since onset of neurological symptoms) who underwent MRI examination as a routine clinical diagnostic procedure. Dynamic susceptibility contrast-enhanced perfusion weighted MRI was performed on a 1.5 T system (Magnetom Vision, Siemens, Erlangen, Germany) using a standard circularly polarized head coil for radio frequency transmission and detection. First, fluid-attenuated inversion recovery, T2-weighted spin echo, and diffusion weighted MRI sequences were obtained in transversal slice orientation enabling initial localization and evaluation of the cerebrovascular insult in the patients with stroke. Then dynamic contrast-enhanced perfusion weighted MRI was performed using a 2D gradient echo echoplanar imaging (EPI) sequence employing 10 transversal slices with a matrix size of 128×128 pixels, pixel size 1.88×1.88 mm, and a slice thickness of 3.0 mm (TR = 0.8 ms, TE = 0.54 ms, FA = 90°). The dynamic study consisted of 38 scans with an interval of 1.5 s. The perfusion sequence and an antecubital vein bolus injection (injection flow 3 ml/s) of gadopentetate dimeglumine (0.15 mmol/kg body weight, Magnevist™, Schering, Berlin, Germany) were started simultaneously in order to obtain several (more than six) scans before cerebral first pass of the contrast agent.

Dynamic Contrast-Enhanced MRI Mammography

Clustering results are presented for dynamic contrast-enhanced MRI mammography in a 48 year-old woman with histologically confirmed breast cancer in the upper medial quadrant of the left breast. These data were acquired according to an MRI protocol to investigate focal breast lesions classified as indeterminate from clinical findings and conventional X-ray mammography. MRI was performed with a 1.5 T system (Magnetom Vision, Siemens, Erlangen, Germany) equipped with a dedicated surface coil to enable simultaneous imaging of both breasts. The patients were placed in a prone position. First, transversal images were acquired with a STIR (short TI inversion recovery) sequence (TR = 5600 ms, TE = 60 ms, FA = 90°, TI = 150 ms,

matrix size 256 × 256 pixels, slice thickness 4 mm). Then a dynamic T1 weighted gradient echo sequence (3D fast low angle shot sequence) was performed (TR = 12 ms, TE = 5 ms, FA = 25°) in transversal slice orientation with a matrix size of 256 × 256 pixels and an effective slice thickness of 4 mm. The dynamic study consisted of 6 measurements with an interval of 110 s. The first frame was acquired before injection of paramagnetic contrast agent (gadopentetate dimeglumine (0.1 mmol/kg body weight, Magnevist™, Schering, Berlin, Germany), immediately followed by the other 5 measurements. The initial localization of suspicious breast lesions was performed by computing difference images, i.e. subtracting the image data of the first from the fourth acquisition.

3.2 Data Analysis

In an initial presegmentation step, the surrounding parts of the data set were excluded by a mask. Although semiautomatic or fully automatic presegmentation procedures could be employed, we performed interactive manual contour tracing by a human observer for all three applications.

Functional MRI

For the analysis of fMRI data, the raw gray level time-series were average-corrected, i.e. the individual mean gray-level for all scans was subtracted for each pixel. Subsequently, the resulting time-series were normalized, i.e. they were scaled for a standard deviation of one. The resulting average-corrected, normalized PTCs $\mathbf{x}$ served as input for the VQ procedure. Average-corrected normalized PTCs were clustered employing 36 CVs, i.e. a maximal number of 36 separate clusters at the end of the hierarchical VQ procedure. In addition, exploratory analyses were performed using different numbers of CVs (16 – 144).

Dynamic Contrast-Enhanced Cerebral Perfusion MRI

The raw gray level time-series $S(\tau)$, $\tau \in \{1, \ldots, 38\}$ was transformed into a PTC of relative signal reduction $x(\tau)$ for each voxel by normalizing with the pre-contrast gray level. The first two scans were discarded from analysis for remaining saturation effects.

Clustering of dynamic perfusion MRI time-series $x(\tau)$ was performed employing 16 CVs. Cluster-specific regional cerebral blood flow (rCBF), regional cerebral blood volume (rCBV), and mean transit time (MTT) were calculated according to the literature [58], [2]. For this purpose, the CVs were transformed into estimates for cluster-specific concentration-time curves (CTCs) according to [46].

Dynamic Contrast-Enhanced MRI Mammography

The raw gray level time-series $S(\tau)$, $\tau \in \{1, \ldots, 6\}$ was transformed into a PTC of relative signal change $x(\tau)$ for each voxel by normalizing with the gray-level of the pre-contrast scan at $\tau = 1$. The PTCs $\mathbf{x}$ were clustered using 9 CVs.

Neural Network Clustering Procedure

Hierarchical neural network clustering of the preprocessed PTCs $\mathbf{x}$ was performed by minimal free energy VQ employing a batch Expectation Maximization (EM) version of minimal free energy VQ [12], [55] in which, in contrast to (1), the computation of CVs $\mathbf{w}_j$ (M-step) and assignment probabilities a_j (E-step) is decoupled and iterated until convergence at each annealing step characterized by a given 'temperature' $T = 2\rho^2$. The batch EM version of our clustering procedure and its relation to the sequential VQ learning rule (1) is explained in the following: Let $\langle f \rangle_X$ denote the expectation value of a random variable f, i.e.

$$\langle f \rangle_X := \int_X f(\mathbf{x})\, p(\mathbf{x}) d\mathbf{x} \tag{14}$$

with probability density $p(\mathbf{x})$ and let

$$\langle f \rangle_j := \int_X f(\mathbf{x})\, p(\mathbf{x}|j) d\mathbf{x} \tag{15}$$

denote the class-specific expectation values, then $\langle f \rangle_j$ can be computed according to Bayes rule

$$p(\mathbf{x}|j) = \frac{p(\mathbf{x})\, p(j|\mathbf{x})}{p(j)} \tag{16}$$

as

$$\langle f \rangle_j = \int_X f(\mathbf{x}) \frac{p(\mathbf{x})\, p(j|\mathbf{x})}{p(j)} d\mathbf{x} = \frac{\langle a_j(\mathbf{x})\, f \rangle_X}{\langle a_j(\mathbf{x}) \rangle_X}, \tag{17}$$

if we interpret the activations $a_j(\mathbf{x})$ as the a posteriori probabilities $p(j|\mathbf{x})$ (see (13)) for the assignment of feature vector $\mathbf{x}$ to the hidden layer neuron j with

$$p(j) = \int_X p(j \cap \mathbf{x}) d\mathbf{x} = \int_X p(\mathbf{x})\, p(j|\mathbf{x}) d\mathbf{x} = \langle a_j(\mathbf{x}) \rangle_X \tag{18}$$

as the average activation (so-called "load") of neuron j. The stationarity condition

$$\langle a_j(\mathbf{x})\, (\mathbf{x} - \mathbf{w}_j) \rangle_X \stackrel{!}{=} 0 \tag{19}$$

of the learning rule (1) yields

$$\mathbf{w}_j = \frac{\langle a_j(\mathbf{x})\,\mathbf{x}\rangle_X}{\langle a_j(\mathbf{x})\rangle_X}, \tag{20}$$

where the CVs $\mathbf{w}_j$ represent the class-specific averages $\langle\mathbf{x}\rangle_j$ of the data distribution according to (17).

Equation (20) represents a batch version of minimal free energy VQ. The right side of equation (20) is influenced via the activations $a_j(\mathbf{x})$ by the CV positions $\mathbf{w}_j$. The procedure is iterated until some convergence threshold is reached for the $\mathbf{w}_j$ or a maximum number of iterations (50 in our implementation) has been performed. A learning parameter ϵ according to the sequential update rule (1) is not needed in the batch version of the algorithm. For a detailed explanation with regard to the theoretical relation between clustering according to batch algorithms such as equation (20) and maximum likelihood estimation by EM, we refer to [28].

Clustering was performed employing 128 annealing steps corresponding to approx. 8×10^3 EM iterations within an exponential annealing schedule for ρ according to (8). The initial value for ρ was chosen $1.2\,\lambda_{\max}^{1/2}$, where $\lambda_{\max}$ denotes the largest eigenvalue of a preceding PCA. The final value for ρ and the annealing strategy were determined in an exploratory manner guided by minimization of the hard assignment criterion function

$$E = \sum_{j=1}^{N}\sum_{\mathbf{x}\in X_j}\|\mathbf{x}-\mathbf{w}_j\|^2, \tag{21}$$

where $X_j \subset X$ is the subset of data vectors attributed to cluster j according to (2). However, we did not find relevant differences of clustering results (as quantified by (21)) when choosing among different annealing strategies. Codebook control parameters like load balance deviation, center of mass shift, entropy, and reconstruction error were computed for each annealing step [12].

Besides minimal free energy VQ, we performed alternative neural network clustering procedures on our data: SOM clustering [34] and k-means type (fuzzy c-means) VQ [7]. All the learning parameters and the respective parameter update strategies were chosen in accordance with the suggestions of the quoted literature, the minimization of the criterion function (21) serving as guideline for parameter selection.

Image processing routines were realized using IDL software, whereas computer simulations of neural networks were implemented as C programs.

3.3 Evaluation

The clustering results were evaluated by (i) assessment of cluster assignment maps, i. e. cluster membership maps according to a minimal distance criterion in the metric

of the PTC feature space, and (ii) the analysis of corresponding CV time courses. For fMRI data, we performed comparative quantitative evaluation of the results obtained by different clustering algorithms using the 'hard assignment' criterion function (21). In addition, we compared the clustering results to PCA performed on the same data.

4 Results

4.1 Functional MRI

Clustering results for a 100 scan visual stimulation experiment are presented in figs. 3, 4, and 5. For illustration, figs. 6a and b show a respective cross-correlation map along with a single raw EPI scan from the fMRI image time-series. Fig. 3 shows the 'hard' cluster assignment maps. In these maps, all the pixels are highlighted that belong to a specific cluster. The decision on assigning a pixel ν characterized by the PTC $\mathbf{x}_\nu = (x_\nu(\tau)), \tau \in \{1, \ldots, n\}$ to a specific cluster j is based on a minimal distance criterion in the n-dimensional time-series feature space, i.e. ν is assigned to cluster j, if the distance $\|\mathbf{x}_\nu - \mathbf{w}_j\|$ is minimal, where $\mathbf{w}_j$ denotes the CV belonging to cluster j. Each CV shown in fig. 5 can be interpreted as the cluster-specific weighted centroid of all PTCs. Figs. 3 and 5 represent cluster assignment maps and codebook vectors for the VQ step with minimal codebook reconstruction error corresponding to a cooperativity parameter $\rho = 1.24$.

Fig. 4 provides the corresponding 'soft' cluster assignment maps which visualize the clustering results of the same data set as fig. 3, but here the soft assignments $a_j(\mathbf{x})$ according to equation (10) are gray-scaled on a continuous scale. This figure has been included to provide a visual comparison to the results obtained by plain PCA in fig. 7.

Note the apparent similarity of the CVs #1, #2, and #3 in fig. 5 compared with the stimulus function shown at the bottom of each CV plot. For these clusters, the corresponding cluster assignment maps in fig. 3 can be attributed to activation of the visual cortex.

Fig. 9 shows a bifurcation diagram of repetitive cluster splitting within the deterministic annealing procedure. Cluster assignment maps in figs. 9b, e, f, i, and j are ordered according to the branch structure of figs. 9a, c, d, g, and h.

Fig. 9 covers three VQ steps of the deterministic annealing procedure. Fig. 9a presents one of 15 cluster centers present at the observed stage of VQ with a cooperativity parameter $\rho = 1.81$. Note the apparent similarity compared with the stimulus (fig. 5). Fig. 9b shows the corresponding hard cluster assignment map, i. e. all the pixels belonging to this cluster center according to a minimal distance criterion in the metric of the PTC feature space. The highlighted regions can be

attributed to the response of the visual cortex. They clearly correspond to the activated regions in the cross-correlation image fig. 6a.

In the subsequent VQ step a phase transition occurs at $\rho = 1.79$, and the cluster of fig. 9a splits up into two descendant clusters representing smaller regions of the visual cortex with different pixel dynamics. They are presented in figs. 9c and d, the corresponding hard cluster assignment maps in figs. 9e and f. Note that the sum of the activated areas in figs. 9e and f is greater than the area of the cluster in fig. 9b. This is based on a reduction in local reconstruction error due to the fact that the new codebook structure fits the underlying local probability density more closely. Thus, the descendant clusters can take over pixels which were formerly attributed to adjacent CVs. The two descendant clusters differ with respect to signal baseline drift and some subtle details of signal dynamics. At a later stage of the VQ procedure, another phase transition occurs at $\rho = 1.64$, and the cluster of fig. 9d again splits up into two descendant clusters represented by figs. 9g and h. The corresponding hard cluster assignment maps are presented in figs. 9i and j.

These results show that minimal free energy VQ can explore fMRI time-series at different scales of resolution with regard to the metric of the feature space. This is not the result of some heuristic cluster splitting or merging procedure that requires meta-knowledge of a human observer about 'relevant' cluster similarity, but evolves naturally in a self-organized manner from a solely data-driven approach (see discussion section). Thus, the cluster structure at smaller values of the cooperativity parameter ρ provides insight into the fine-grained structure of the fMRI activation patterns, representing a higher clustering resolution with respect to subtle differences of signal dynamics.

In the light of these observations, we want to emphasize that the self-organized hierarchical phase transitions represented by the bifurcation diagram and resulting cluster distances are determined by the metric of the feature space, which does not necessarily imply neurobiological relatedness. For the cluster splitting observation of fig. 9, clusters representing activation are derived from a common branch of the bifurcation diagram. For the cluster assignment maps in fig. 3, cluster 21 may be assigned to a co-activation of the frontal eye fields induced by the stimulus onset. However, it cannot be derived from a common branch with regard to clusters 1, 2 and 3. It is negatively correlated to the stimulus function. Although there may be some kind of physiological relatedness between cluster 21 and the stimulus (and, thus, clusters 1, 2, and 3), this relatedness is not expressed by the metric of the feature space, and, thus, cannot be displayed by clustering procedures based on this metric. This observation holds for all clustering procedures currently used for fMRI analysis, such as SOMs and fuzzy c-means VQ, and may be considered as a general weakness of unsupervised fMRI data processing. However, a subset of physiological interdependencies that can be expressed by the metric of the feature space can be successfully detected by clustering approaches, such as demonstrated for the clusters

representing activation in fig. 3.

As an additional observation, clustering is able to identify data sets that are highly indicative of artifacts, e.g. ventricular pulsation or through plane motion. As an example, figs. 10a and 10b present a respective single CV and the corresponding assignment map gained from a 64 scan visual stimulation experiment. The CV reflects a considerable drift over time, and the respective pixels are distributed at the edge of the data set. Thus, clustering results can be used for the detection of artifacts, which sometimes may focus the observer's attention to potential improvements of experimental conditions, e.g. in order to avoid motion artifacts. These artifacts could hardly be detected by classical supervised methods of fMRI time-series analysis.

We analyzed the anatomical features of the 36 clusters of pixels in all six subjects separately. In all subjects, our method resulted in identification of at least two (up to four) pixel clusters confined to the expected activation areas in the visual cortex. The PTCs of the corresponding codebook vectors correlated well to the stimulation protocol (cross-correlation between 0.65 and 0.93). In addition, in all subjects anatomically coherent clusters of pixels were regularly seen in the frontal eye fields, in three subjects in the precuneus, and in three subjects also in the temporal areas. These clusters frequently showed negative cross-correlation for the PTCs (up to -0.43) such as cluster #21 in figs. 3 and 5. The ventricles were also represented separately in all subjects by up to three clusters with complex-shaped CVs such as clusters #6 and #8 in figs. 3 and 5.

For comparison, the results of a principal component analysis on these data are presented in figs. 7 and 8.

Computation times for clustering of these fMRI data by minimal free energy VQ, fuzzy c-means VQ, and SOMs using different numbers of codebook vectors on a Sun Ultra5 workstation with an UltraSPARC-IIi 333 MHz CPU can be found in table 6. Computation time for plain PCA was 3 s. Our computation times for fuzzy c-means VQ are much lower than those observed on a comparable data set in a recently published paper on fMRI clustering [10], where fuzzy c-means VQ using 25 CVs required 61191 s.

A comparative quantitative evaluation of different unsupervised learning techniques using the 'hard assignment' criterion function (21) for this visual stimulation fMRI experiment is presented in fig. 11. The criterion function has been computed for quality assessment of minimal free energy VQ, SOMs, and fuzzy c-means VQ using different numbers of codebook vectors for each method along with principal component analysis using different numbers of components in the order of descending eigenvalues.

4.2 Dynamic Contrast-Enhanced Cerebral Perfusion MRI

Clustering results for a 38 scan dynamic contrast-enhanced MRI perfusion study in a patient with a large stroke in the supply area of the left middle cerebral artery are presented in figs. 12 and 13. After discarding the first two scans, a relative signal reduction time-series $x(\tau), \tau \in \{1, \ldots, n\}, n = 36$ can be computed for each voxel. Clustering these PTCs identifies groups of pixels with similar signal dynamics. Fig. 12 shows the 'cluster assignment maps' overlaid onto an EPI scan of the perfusion sequence. As explained for the fMRI data in fig. 3, all the pixels are highlighted in these maps that belong to a specific cluster according to a minimal distance criterion in the n-dimensional time-series feature space.

Fig. 13 shows the prototypical cluster-specific concentration-time curves (CTCs) computed from the CVs belonging to the clusters of fig. 12. The region of the cerebrovascular insult in the supply region of the left middle cerebral artery is clearly depicted by cluster #15 and #16. Note the small corresponding CTC amplitude, i.e. the small cluster-specific rCBV and rCBF. Cluster #14 contains peripheral and adjacent regions.

The clustering results for the control subject without evidence of cerebrovascular disease were similar to those of the non-affected hemisphere in fig. 12. Clustering could differentiate between larger vessels located in the cerebral sulci and parenchymal vascularization. In contrast to the results of the stroke patient data in figs. 12 and 13, there was no evidence of side-asymmetry with regard to both the temporal pattern and the amplitude of brain perfusion. This was expressed by the observation that, in contrast to fig. 12, each cluster contained pixels in roughly symmetrical regions of both hemispheres. Furthermore, there was no cluster indicating a localized perfusion deficit.

4.3 Dynamic Contrast-Enhanced MRI Mammography

Clustering results for a 6 scan dynamic contrast-enhanced MRI mammography study in a patient with breast cancer in the upper medial quadrant of the left breast are presented in figs. 14 and 15 for VQ employing 9 CVs. The figures are analogous to figs. 3 and 5 for functional MRI and figs. 12 and 13 for dynamic contrast-enhanced perfusion MRI. They show clusters of pixels and CVs for minimal free energy VQ of MRI mammography data covering a supramamillar transversal slice of the left breast containing a suspicious lesion that has been proven to be malignant by subsequent histological examination.

The procedure is able to segment the lesion from the surrounding breast tissue as can be seen from clusters #0 and #1 in fig. 14. Note the rapid and strong contrast agent uptake followed by subsequent plateau and washout phases in the round central region of the lesion as indicated by the corresponding CV of cluster

#0 in fig. 15. Furthermore, clustering results enable a subclassification within this lesion with regard to regions characterized by different MRI signal time-courses: The central cluster #0 is surrounded by the peripheral circular cluster #1 which primarily can be separated from both the central region and the surrounding tissue by the amplitude of its contrast agent uptake ranging between CV #0 and all the other CVs. Furthermore, an additional, even more peripheral ringlike structure can be detected by careful observation of cluster #2.

5 Discussion

In this paper, we have presented the concept of self-organized neural network clustering as a flexible and powerful computer vision strategy for data analysis and visualization in the field of medical image time-series analysis. Deterministic annealing by the minimal free energy VQ algorithm can serve as a useful method to unveil the properties of medical image sequences by gradually resolving the fine-grained structure of the data set in an unsupervised repetitive cluster splitting process. It has been shown to be useful for the hierarchical subclassification of local tissue properties with respect to regions of different signal dynamics and/or amplitude.

Although we found that non-hierarchical neural network clustering methods like SOMs or k-means type clustering can be useful for medical image time-series data analysis as well, we want to point out some potential advantages of the deterministic annealing procedure based on minimal free energy VQ:

(i) Deterministic annealing, in contrast to SOMs and k-means type clustering, is a hierarchical method that allows data exploration on different scales of clustering resolution by gradually increasing the number of cluster centers in a self-organized, solely data-driven procedure. This may sometimes unveil some insight for exploring the coarse-grained structure of the data set.

(ii) As deterministic annealing attempts to track the global minimum of the free energy landscape through different stages of cost function complexity, the procedure seems to be more resistant against being trapped in local minima in contrast to other clustering techniques, see below. The resulting increase of result quality can be objectively determined by computing respective quantitative measures such as the clustering criterion function (21).

SOMs provide the advantage of 'topology preservation' that allows the construction of graphically appealing 2D maps. As pointed out in the theory section, this property may be misleading due to folding of the feature map if the data are located on a manifold that exceeds the dimensionality of the mapping grid (e. g. [43]).

At what stage should the clustering procedure be stopped? Deterministic annealing can be continued until cluster degeneracy has been removed completely, i.e. each CV represents a different cluster. In this stage, the codebook reconstruction error approaches a minimum at $\rho := \rho_0$ which can serve as a stop criterion. However, clustering can be stopped at any earlier stage, i.e. $\rho > \rho_0$, if one is interested in large-scale differences of signal dynamics. The 'high-level' clusters in figs. 9a – f present examples for this.

How many CVs should be used for clustering? If the number of CVs is chosen too small, interesting phenomena involving only a small subset of pixels cannot be detected in a suitable clustering resolution. If it is chosen too large, however, large-scale phenomena are represented by a great number of CVs which have to be recollected by heuristic merging procedures to form 'superclusters' of several CVs coding for similar PTCs (e. g. [22]), or, for deterministic annealing, one has to trace higher levels of the splitting hierarchy. As can be seen from fig. 11, a simple clustering criterion function cannot be used for a priori defining a reasonable number of CVs without knowledge of the data or the respective goals of data analysis. There have been various attempts in the literature to find and evaluate solutions to this problem, however, it has also been shown that their performance is data-dependent and frequently poor. For a detailed discussion of this issue, we refer to [38].

Although we conjecture that a data-driven hierarchical procedure can be useful to data analysis and visualization on different scales of clustering resolution in our applications, we do not claim that it is always possible to attribute a specific neurophysiological interpretation to each hierarchy level. However, there may be situations where hierarchical clustering can provide a useful subclassification of brain regions that may be relevant in terms of neurophysiological or neuroanatomical interpretation. As an example for the potential of our method, we refer to previously published data where the visual cortex identified by a higher level in hierarchical clustering of non-normalized photic stimulation data could be subdivided into two clusters representing central and peripheral extrastriatal areas of the visual cortex [52].

The question arises, how the data drives the splitting process, i.e. what in the data provokes the phase transition leading to a split. For this purpose, deterministic annealing can be defined as an optimization on a family of smoothed cost functions, i.e. the free energy (12) parameterized by the cooperativity parameter ρ. In the analogy to statistical mechanics, the first term in (12) can be interpreted as the so-called "internal energy" in a canonic ensemble, and the second term defines the entropic contribution. Note that the first term resembles a conventional clustering criterion function ("within-class inertia") such as (21), although, in contrast to (21), soft assignments are used according to the "fuzzy" character of the procedure.

The cooperativity parameter determines the amount of smoothing. For high values of ρ, the clustering cost function is smoothed to such a degree that only

one global minimum remains, which is easily determined by gradient descent or an EM approach. For lower values of ρ, the structure of the original cost function is increasingly reflected by the free energy. For $\rho \to 0$, the entropic contribution finally vanishes, and the free energy approaches the exact form of the original clustering cost function. Deterministic annealing starts by determining the minimum of the free energy at high values of ρ and attempts to track the minimum for lower values of ρ, until the minimum of the free energy at $\rho \to 0$ coincides with the minimum of the internal energy, i.e. the original clustering cost function. We do not claim that the algorithm always reaches a global minimum. However, according to the literature, e.g. [28], [29], deterministic annealing frequently finds better local minima.

Cluster splitting occurs, when a minimum of the free energy is transformed into a saddle point within the annealing process of the cooperativity parameter ρ, which gradually unveils the original complexity of the clustering cost function. For a toy example with explicit computation of the free energy landscape and its respective contributions, we refer to [12].

However, even if one may deny any advantage of self-organized hierarchical clustering, e.g. by criticizing the relevance of the resulting multi-resolution hierarchy with regard to its physiological meaning in specific examples, the performance of minimal free energy VQ based on the comparative evaluation of the clustering criterion function (21) in fig. 11 nevertheless justifies our approach. We conjecture that the result quality is based on the property of deterministic annealing to be more resistant against being trapped in local minima in comparison to direct minimization of a clustering cost function. Toy examples supporting this conjecture can be found in [12] as well.

Although the computation of the criterion function (21) in fig. 11 may enable a quantitative comparative evaluation of conceptually different data decomposition techniques, it may be criticized that it does not take into account the specific characteristics of each algorithm with regard to 'soft' data partitioning which is an inherent property of both PCA and the clustering techniques presented in our paper. In this context, one should not overemphasize the poor result for PCA in the comparison of fig. 11. It corresponds to our experience that PCA can nevertheless serve as a valuable tool for unsupervised fMRI analysis, picking up stimulus-related activation or motion artifacts as well.

Careful consideration should be made concerning the metric implied by appropriate preprocessing of the data before clustering [49]. The concrete choice depends on both the observer's focus of interest and the specific clinical application.

Should one perform normalization of PTCs before clustering? The discussion of this issue resembles the problem of choosing between the computation of cross-correlation or cross-covariance images in classical supervised fMRI analysis (e. g. [3]). The concrete choice depends on the observer's focus of interest. If cluster analysis is to be performed with respect to signal dynamics rather than amplitude, clustering

should be preceded by time-series normalization. Normalization, however, leads to amplification of noise in low-amplitude PTCs. Furthermore, PTC amplitude can provide useful information that may be valuable for the physiological or clinical interpretation of image time-series data depending on the focus of data analysis. The decision should be reflected in the light of the specific biological and clinical application.

For the analysis of contrast-enhanced studies, PTC normalization usually does not make sense, as the amplitude information is crucial for the clinical interpretation of the observed time-series data which would be discarded by normalization. In the analysis of dynamic contrast-enhanced cerebral perfusion MRI, for example, locally reduced cerebral perfusion of the infarcted region can only be documented by avoiding PTC normalization (see fig. 13), although there may be interesting side asymmetries with regard to temporal signal dynamics as well. The same holds for dynamic MRI mammography: Here, the initial identification of suspicious tissue lesions is primarily based on the high contrast-agent uptake amplitude, whereas the temporal signal dynamics may provide additional information for subclassification of different regions within this lesion, see fig. 15.

In this context one may speculate about the validity of spatial coherence measures among voxels assigned to the same cluster as a criterion for selecting "interesting" clusters, where "spatial" can refer to both the geometrical 3D space of pixel coordinates and the metric of the high-dimensional feature space of fMRI time-series. In so far as the geometrical feature space of brain coordinates is concerned, considerable effort has been made in the past in order to correctly assess significant activation in fMRI by use of cluster-size thresholds. For a detailed discussion of this issue, we refer to the relevant literature, e.g. [24].

An additional approach to define a cluster validity measure by exploiting cluster coherence based on the metric of the high-dimensional feature space of fMRI time-series can be obtained by observing the norm of the resulting CVs, when normalized PTCs are used as input for the clustering procedure. As can be shown, this measure reflects changes in cluster-specific covariance in the feature space. This effect may be used for the selection of clusters representing 'interesting' features with regard to the interpretation of physiological phenomena or experimental conditions, such as motion artifacts. However, we want to emphasize that, like in conventional supervised MRI time-series data analysis, the physiological or clinical interpretation of clustering results is still in the responsibility of the neuroradiologist or the biomedical researcher. Neural network clustering can only provide computer-aided support to appropriate data exploration in order to assist, not to replace this interpretation. In particular, we want to emphasize that we would consider it 'dangerous' to solely rely on cluster validity indices in order to decide about the neurobiological relevance of clustering results.

In summary, we obtain a complex picture of various pros and cons for choosing

among specific clustering algorithms and preprocessing strategies. Careful consideration should be made with respect to all of these issues in order to extract maximal information from the data according to the specific clinical focus of the examination. This can only be achieved by close cooperation between computer vision scientists and biomedical researchers. The clinical applications presented in this paper may serve as an example for this endeavour aiming at interdisciplinary communication.

6 Conclusion

Our study shows that deterministic annealing by the minimal free energy VQ is a useful strategy for unsupervised analysis of medical image time-series data. It realizes a hierarchical unsupervised learning procedure unveiling the structure of the data set with gradually increasing clustering resolution. Repetitive cluster splitting can be analyzed by back-tracking a bifurcation diagram which allows for convenient ordering of clusters according to their underlying multidimensional probability distribution. Thus, the properties of the data set can be explored in a natural manner.

In particular, the presented neural network clustering methods could identify activated brain regions in functional MRI studies of visual stimulation experiments. They unveiled regional abnormalities of brain perfusion characterized by differences of signal amplitude and dynamics in contrast-enhanced cerebral perfusion MRI. Here, they could provide a rough segmentation with regard to vessel size and identify regions of perfusion deficit in patients with stroke. Furthermore, dynamic MRI mammography data could be analyzed concerning the identification and regional subclassification of pathological breast tissue lesions.

Although we conjecture from our results that neural network clustering can provide a useful extension of conventional methods for medical image time-series analysis, large controlled clinical trials are needed to establish the validity of our method and to finally evaluate its contribution to clinical decision making.

References

[1] M.R. Anderberg, editor. *Cluster analysis for applications.* Academic Press, New York, 1973.

[2] L. Axel. Cerebral blood flow determination by rapid-sequence computed tomography. *Radiology*, 137:679–686, 1980.

[3] P.A. Bandettini, A. Jesmanowicz, E.C. Wong, and J.S. Hyde. Processing strategies for time-course data sets in functional MRI of the human brain. *Magn. Reson. Med.*, 30:161–173, 1993.

[4] R. Baumgartner, C. Windischberger, and E. Moser. Quantification in functional magnetic resonance imaging: fuzzy clustering vs. correlation analysis. *Magnetic Resonance Imaging*, 16(2):115–125, 1998.

[5] S. Becker. Mutal information maximization: models of cortical self-organization. *Network*, 7:7–31, 1996.

[6] A.J. Bell and T.J. Sejnowski. An information-maximization approach to blind separation and blind deconvolution. *Neural Computation*, 7:1129–1159, 1995.

[7] J.C. Bezdek. *Pattern recognition with fuzzy objective function algorithms.* Plenum Press, 1981.

[8] C. M. Bishop, M. Svensen, and C. K. I. Williams. Developments of the generative topographic mapping. Technical Report NCRG-98-012, Neural Computing Research Group, Aston University, Birmingham, 1998.

[9] C. M. Bishop, M. Svensen, and C. K. I. Williams. GTM: the generative topographic mapping. *Neural Computation*, 10(1):215–234, 1998.

[10] K.H. Chuang, M.J. Chiu, C.C. Lin, and J.H. Chen. Model-free functional MRI analysis using Kohonen clustering neural network and fuzzy c-menas. *IEEE Transactions on Medical Imaging*, 18(12):1117–1128, 1999.

[11] A.P. Dempster, N.M. Laird, and D.B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *J. Royal. Statist. Soc. Ser. B (methodological)*, 39:1–38, 1977.

[12] D.R. Dersch. *Eigenschaften neuronaler Vektorquantisierer und ihre Anwendung in der Sprachverarbeitung.* Verlag Harri Deutsch, Reihe Physik, Bd. 54, Thun, Frankfurt am Main, 1996. ISBN 3-8171-1492-3.

[13] D.R. Dersch, S. Albrecht, and P. Tavan. Hierarchical fuzzy clustering. In A. Wismüller and D.R. Dersch, editors, *Symposion über biologische Informationsverarbeitung und Neuronale Netze – SINN '95, Konferenzband.* Hanns-Seidel-Stiftung, München, 1996.

[14] D.R. Dersch and P. Tavan. Control of annealing in minimal free energy vector quantization. In *Proceedings of the IEEE International Conference on Neural Networks ICNN'94*, pages 698–703, Orlando, Florida, 1994.

[15] D.R. Dersch and P. Tavan. Load balanced vector quantization. In *Proceedings of the International Conference on Artificial Neural Networks ICANN*, pages 1067–1070. Springer, 1994.

[16] D.R. Dersch and P. Tavan. Asymptotic level density in topological feature maps. *IEEE Transactions on Neural Networks*, 6(1):230–236, 1995.

[17] X. Ding, T. Masaryk, P. Ruggieri, and J. Tkach. Detection of activation patterns in dynamic functional MRI with a clustering technique. In *Proc., SMR, 4th Annual Meeting*, page 1798, New York, 1996.

[18] X. Ding, J. Tkach, P. Ruggieri, and T. Masaryk. Analysis of time-course functional mri data with clustering method without use of reference signal. In *Proc., SMR, 2nd Annual Meeting*, page 630, San Francisco, 1994.

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

[20] E. Erwin, K. Obermayer, and K. Schulten. Self-organizing maps: ordering, convergence properties, and energy functions. *Biological Cybernetics*, 61:47–55, 1992.

[21] E. Erwin, K. Obermayer, and K. Schulten. Self-organizing maps: stationary states, metastability, and convergence rate. *Biological Cybernetics*, 61:35–45, 1992.

[22] H. Fischer and J. Hennig. Neural-network based analysis of MR time series. *Magn. Reson. Med.*, 41(1):124–131, 1999.

[23] E.W. Forgy. Cluster analysis of multivariate data: Efficiency vs. interpretability of classifications. *Biometrics*, 21:768, 1965.

[24] S.D Forman, J.D. Cohen, M. Fitzgerald, W.F. Eddy, M.A. Mintun, and D.C. Noll. Improved assessment of significant activation in functional magnetic resonance imaging (fMRI): use of a cluster-size threshold. *Magn. Reson. Med.*, 33:636–647, 1995.

[25] M. Galicki, U. Möller, and H. Witte. Neural clustering networks based on global optimisation of prototypes in metric spaces. *Neural Computing and Applications*, 5:2–13, 1997.

[26] X. Golay, S. Kollias, D. Meier, and P. Boesinger. Optimization of a fuzzy clustering technique and comparison with conventional post processing methods in fMRI. In *Proc., SMR, 4th Annual Meeting*, page 1787, New York, 1996.

[27] C. Goutte, P. Toft, E. Rostrup, F.A. Nielsen, and L.K. Hans en. On clustering fMRI time series. *Neuroimage*, 9:298–310, 1999.

[28] T. Graepel, M. Burger, and K. Obermayer. Phase transitions in stachastic self-organizing maps. *Physical Review E*, 56(4):3876–3890, 1997.

[29] T. Graepel, M. Burger, and K. Obermayer. Self-organizing maps: Generalizations and new optimization techniques. *Neurocomputing*, 20:173–190, 1998.

[30] A. Hyvärinen. Survey on independent component analysis. *Neural Computing Surveys*, 2:94–128, 1999.

[31] M. Kloppenburg and P. Tavan. Deterministic annealing for density estimation by multivariate normal mixtures. *Phys. Rev. E*, 55:2089–2092, 1997.

[32] T. Kohonen. *Self-Organization and Associative Memory.* Springer, Berlin, 1989.

[33] T. Kohonen. The self-organizing map. *Proceedings of the IEEE*, 78(9):1464–1480, 1990.

[34] T. Kohonen, J. Hynninen, J. Kangas, and J. Laaksonen. *SOM_PAK: the self-organizing map program package.* Helsinki University of Technology, Laboratory of Computer and Information Science, Rakentajanaukio 2 C, SF-02150 Espoo, Finland, 1996.

[35] Y. Linde, A. Buzo, and R.M. Gray. An algorithm for vector quantizer design. *IEEE Transactions on Communications*, 28:84–95, 1980.

[36] T.M. Martinetz and K. Schulten. A 'neural gas' network learns topologies. In *Proceedings of the International Conference on Artificial Neural Networks ICANN*, pages 397–402, Amsterdam, 1991. Elsevier Science Publishers.

[37] T.M. Martinetz and K. Schulten. Topology representing networks. *Neural Networks*, 7:507–522, 1994.

[38] G.W. Milligan and M.C. Cooper. An examination of procedures for determining the number of clusters in a data set. *Psychometrika*, 50:159–179, 1985.

[39] E. Moser, C. Teichtmeister, and M. Diemling. Reproducibility and postprocessing of gradient echo functional MRI to improve localization of brain activity in the human visual cortex. *Magnetic Resonance Imaging*, 14:567–579, 1996.

[40] S. Ogawa, T. Lee, A.R. Kay, and D.W. Tank. Brain magnetic-resonance-imaging with contrast dependent on blood oxygenation. *Proc Natl Acad Sci USA*, 87(24):9868–9872, Dec 1990.

[41] L. Østergaard, R.M. Weisskopf, D.A. Chesler, C. Gyldensted, and B.R. Rosen. High resolution measurement of cerebral blood flow using intravascular tracer bolus passages. Part I: mathematical approach and statistical analysis. *Magn. Res. Med.*, 36:715–725, 1996.

[42] H. Ritter. Asymptotic level density for a class of vector quantization processes. *IEEE Transactions on Neural Networks*, 1(2):173–175, 1991.

[43] H. Ritter, T. Martinetz, and K. Schulten. *Neuronale Netze.* Addison Wesley, New York, 1991.

[44] K. Rose, E. Gurewitz, and G. Fox. A deterministic annealing approach to clustering. *Pattern Recognition Letters*, 11(11):589–594, 1990.

[45] K. Rose, E. Gurewitz, and G.C. Fox. Vector quantization by deterministic annealing. *IEEE Transactions on Information Theory*, 38(4):1249–1257, 1992.

[46] B.R. Rosen, J.W. Belliveau, J.M. Vevea, and T.J. Brady. Perfusion imaging with NMR contrast agents. *Magn. Reson. Med.*, 14:249–265, 1990.

[47] G. Scarth, M. McIntyre, B. Wowk, and R.L. Somorjai. Detection of novelty in functional images using fuzzy clustering. In *Proc., SMR, 3rd Annual Meeting*, page 238, Nice, 1995.

[48] G. Scarth, E. Moser, R. Baumgartner, M. Alexander, and R.L. Somorjai. Paradigm-free fuzzy clustering-detected activations in fMRI: A case study. In *Proc., SMR, 4th Annual Meeting*, page 1784, New York, 1996.

[49] R. Weisskoff. Personal communication, 1998.

[50] D.J. Willshaw and C. von der Malsburg. How patterned neural connections can be set up by self-organization. *Proceedings of the Royal Society, London*, B 194:431–445, 1976.

[51] A. Wismüller and D.R. Dersch. Neural network computation in biomedical research: chances for conceptual cross-fertilization. *Theory in Biosciences*, 116(3), 1997.

[52] A. Wismüller, D.R. Dersch, B. Lipinski, K. Hahn, and D. Auer. Hierarchical clustering of fMRI time-series by deterministic annealing. *Neuroimage*, 7(4):S593, 1998.

[53] A. Wismüller, D.R. Dersch, B. Lipinski, K. Hahn, and D. Auer. Hierarchical unsupervised clustering of fMRI data by deterministic annealing. In *Proceedings of the Sixth Scientific Meeting of the International Society of Magnetic Resnonace in Medicine 1998*, page 412, Sydney, 1998.

[54] A. Wismüller, D.R. Dersch, B. Lipinski, K. Hahn, and D. Auer. A neural network approach to functional MRI pattern analysis – clustering of time-series by hierarchical vector quantization. In L. Niklasson, M. Bodèn, and T. Ziemke, editors, *ICANN'98 — Proceedings of the 8th International Conference on Artificial Neural Networks, Skövde, Sweden. Perspectives in Neural Computing*, volume 2, pages 123–128, London, Berlin, New York, 1998. Springer-Verlag.

[55] A. Wismüller, F. Vietze, and D.R. Dersch. Segmentation with neural networks. In Isaac Bankman, Raj Rangayyan, Alan Evans, Roger Woods, Elliot Fishman, and Henry Huang, editors, *Handbook of Medical Imaging*, Johns Hopkins University, Baltimore, 2000. Academic Press. ISBN 0120777908.

[56] A. Wismüller, F. Vietze, D.R. Dersch, K. Hahn, and H. Ritter. The deformable feature map — adaptive plasticity in function approximation. In L. Niklasson, M. Bodèn, and T. Ziemke, editors, *ICANN'98 — Proceedings of the 8th International Conference on Artificial Neural Networks, Skövde, Sweden. Perspectives in Neural Computing*, volume 1, pages 222–227, London, Berlin, New York, 1998. Springer-Verlag.

[57] R.P. Woods, S.R. Cherry, and J.C. Mazziotta. Rapid automated algorithm for aligning and reslicing PET images. *Journal of Computer Assisted Tomography*, 16:620–633, 1992.

[58] K.L. Zierler. Theoretical basis of indicator-dilution methods for measuring flow and volume. *Circ. Res.*, 16:393–407, 1965.

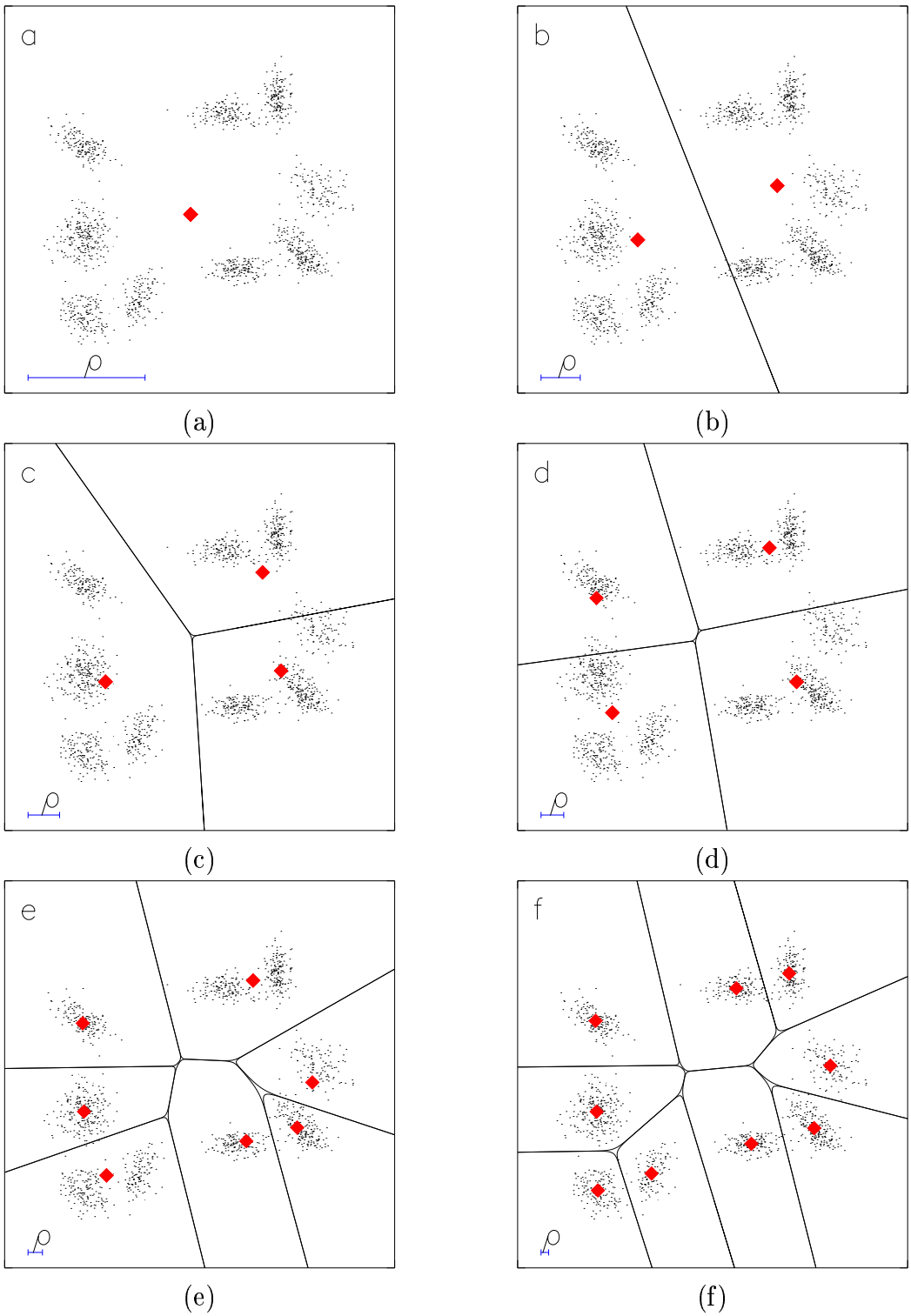


Figure 1: Toy example for hierarchical vector quantization by deterministic annealing. The feature space is represented by a plane with several Gaussian distributions. By gradually decreasing the cooperativity parameter ρ in (a) – (f), the initially degenerate clusters increasingly split up into multiple descendant clusters, thus approaching a fine-grained tessellation of the feature space.

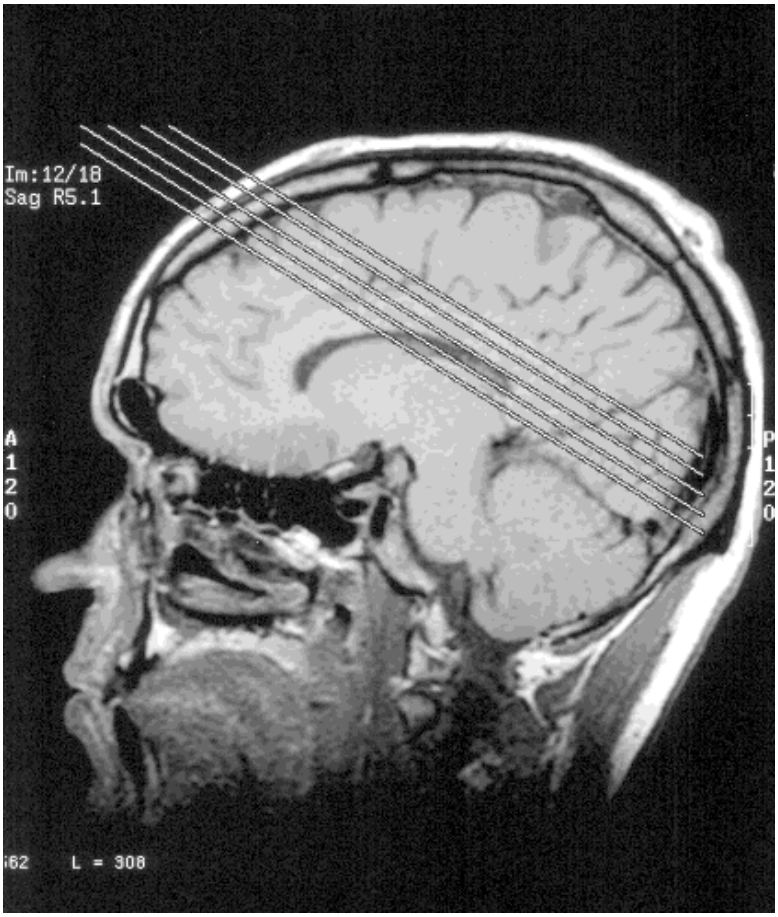


Figure 2: Pilot scan for the visual stimulation fMRI experiment presented in figs. 3 and 5. The data analysis results are plotted for the central slice.

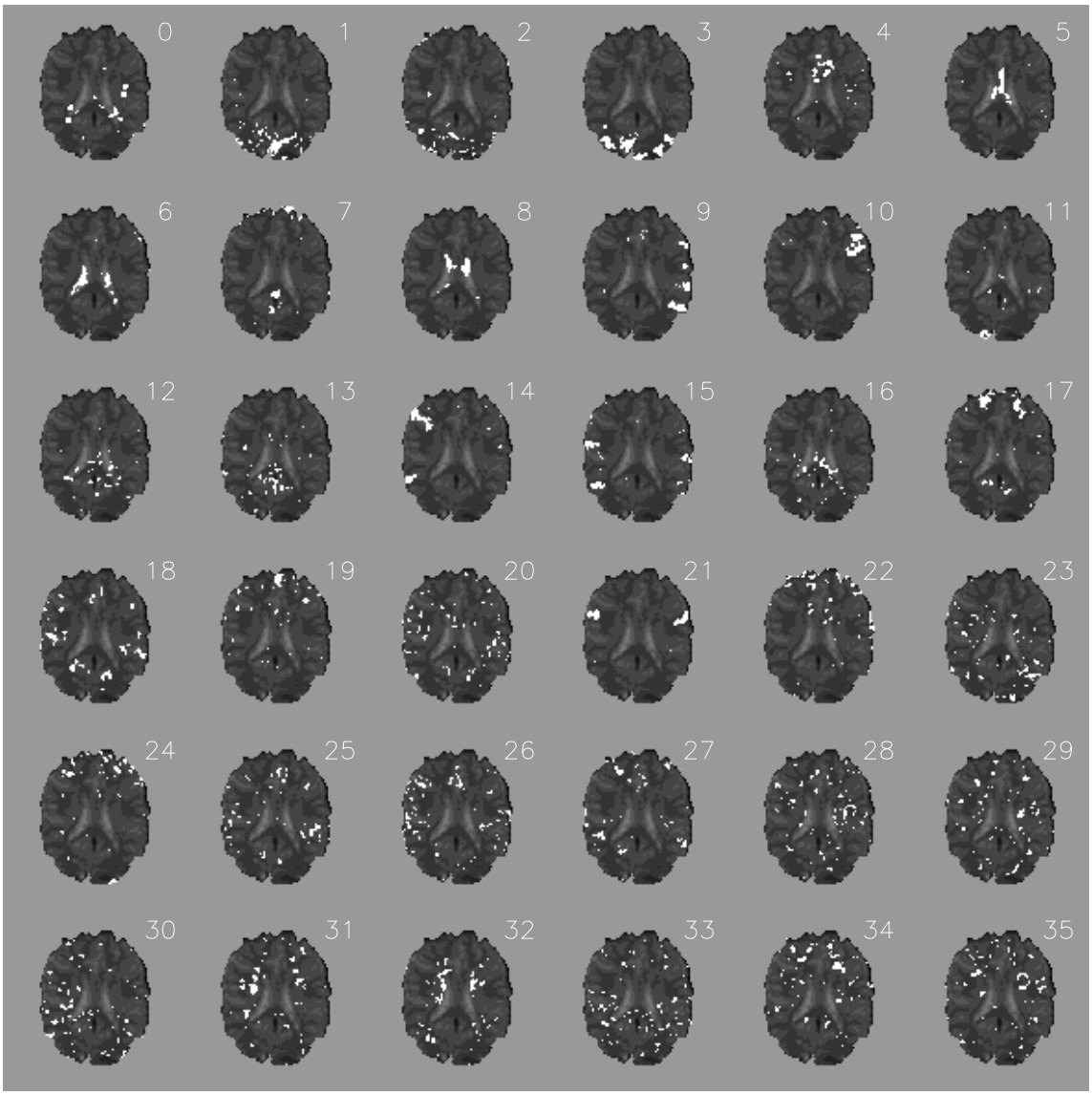


Figure 3: Hard cluster assignment maps for cluster analysis of a visual stimulation fMRI experiment. The figure shows the vector quantization step with minimal reconstruction error. Results are obtained by minimal free energy vector quantization using 36 codebook vectors.

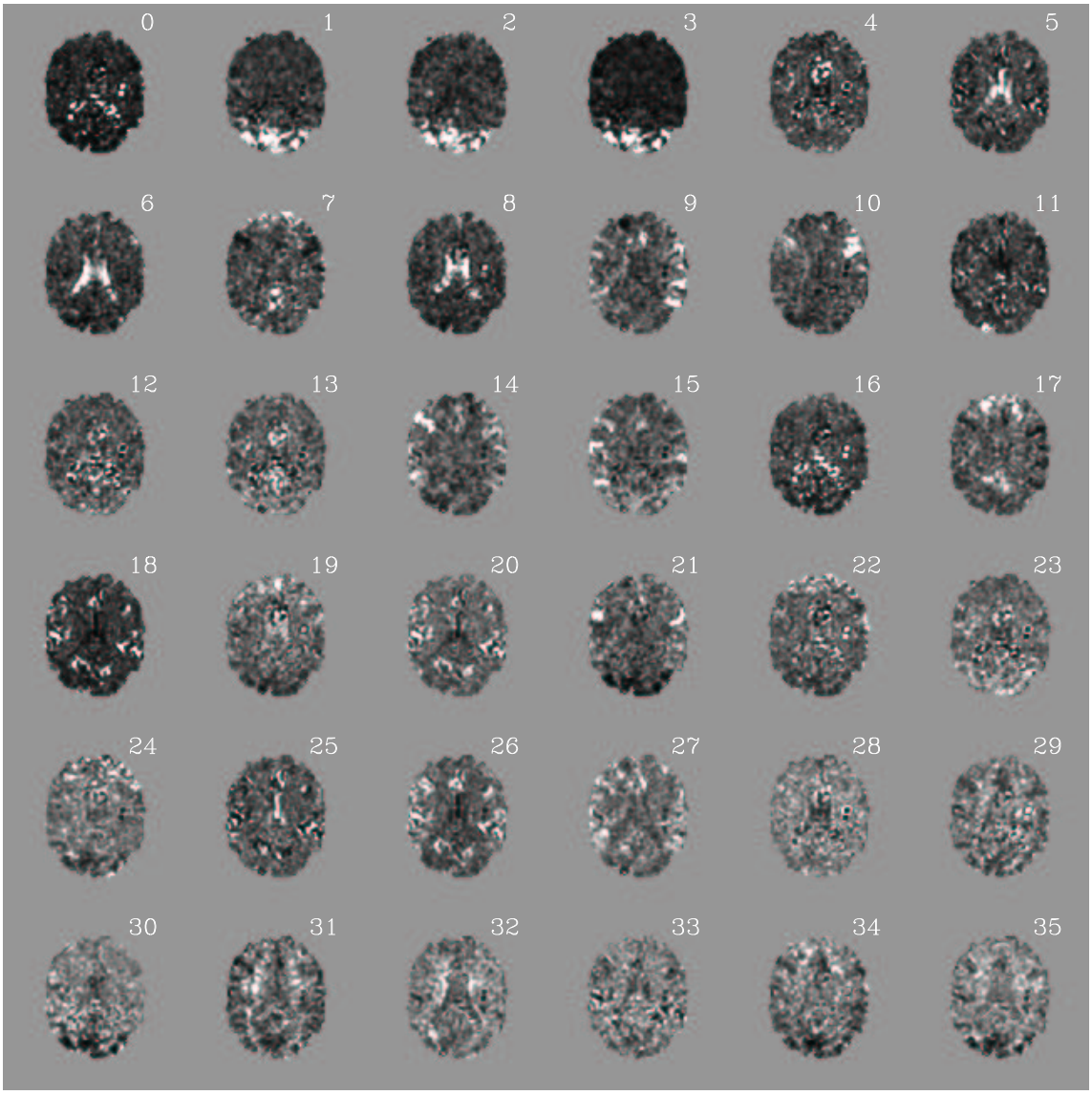


Figure 4: Soft cluster assignment map corresponding to fig. 3.

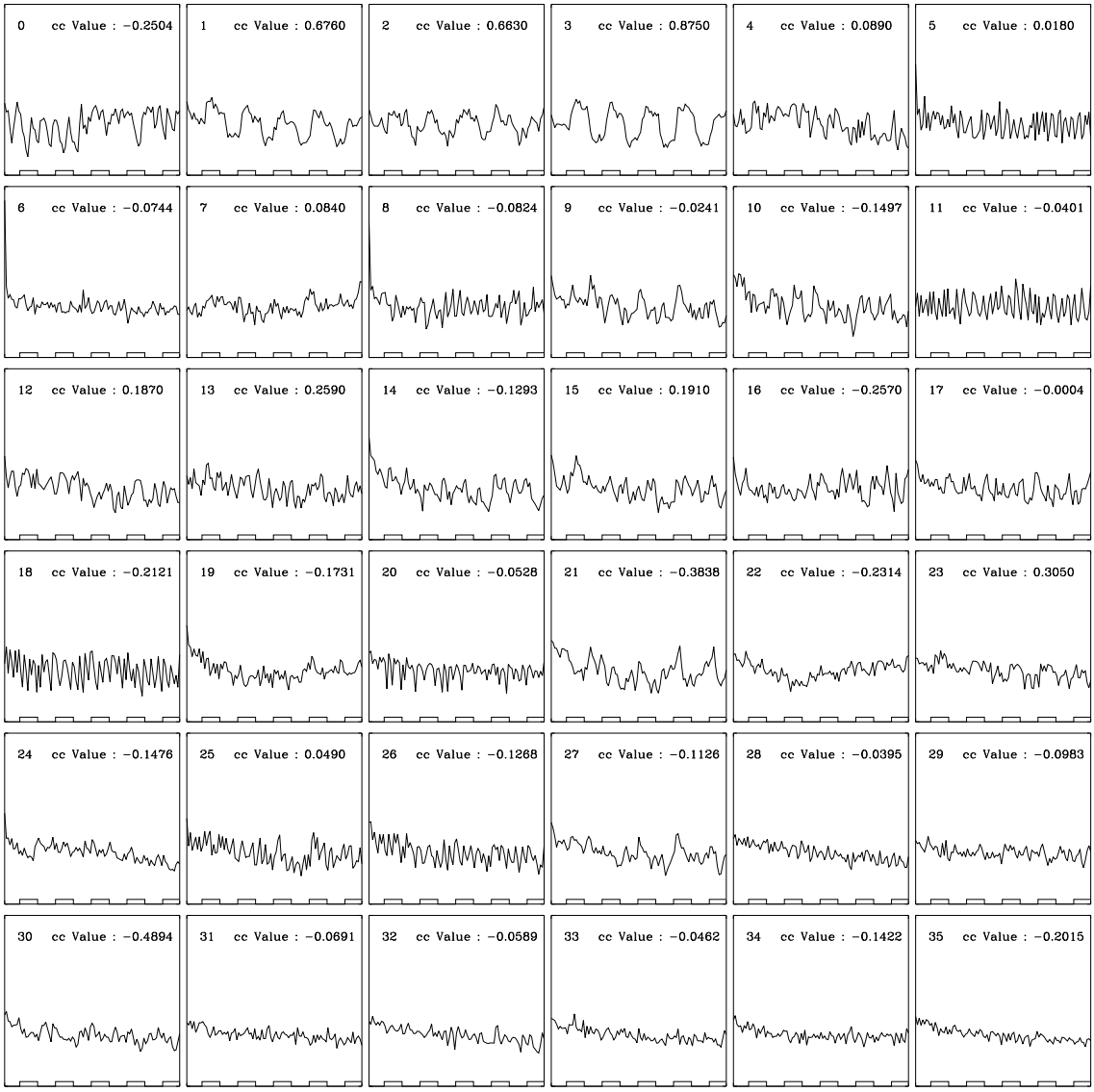
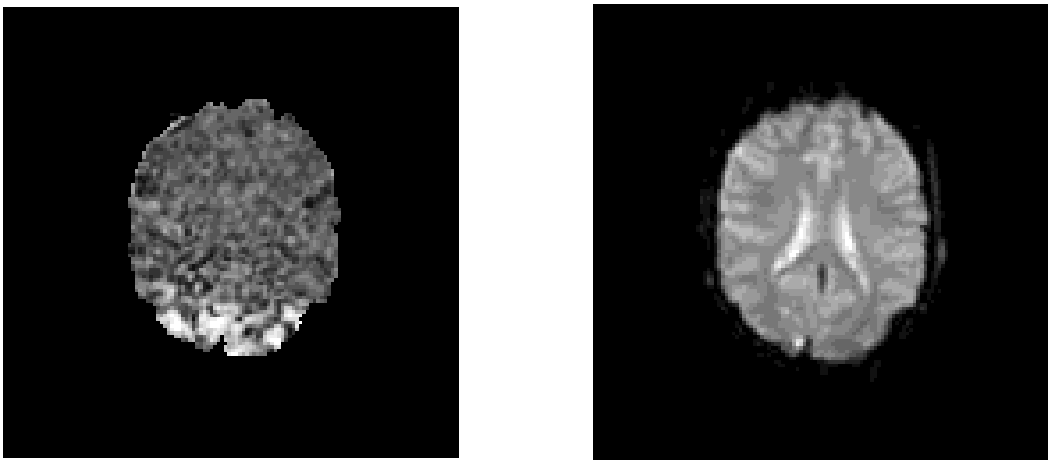


Figure 5: Codebook vectors for vector quantization of the fMRI data set presented in fig. 3. *ccValue* denotes the cross-correlation of each codebook vector with the stimulus function. The time-course of the stimulus is shown at the bottom of each plot.



(a) (b)

Figure 6: (a) Cross-correlation image. (b) EPI image.

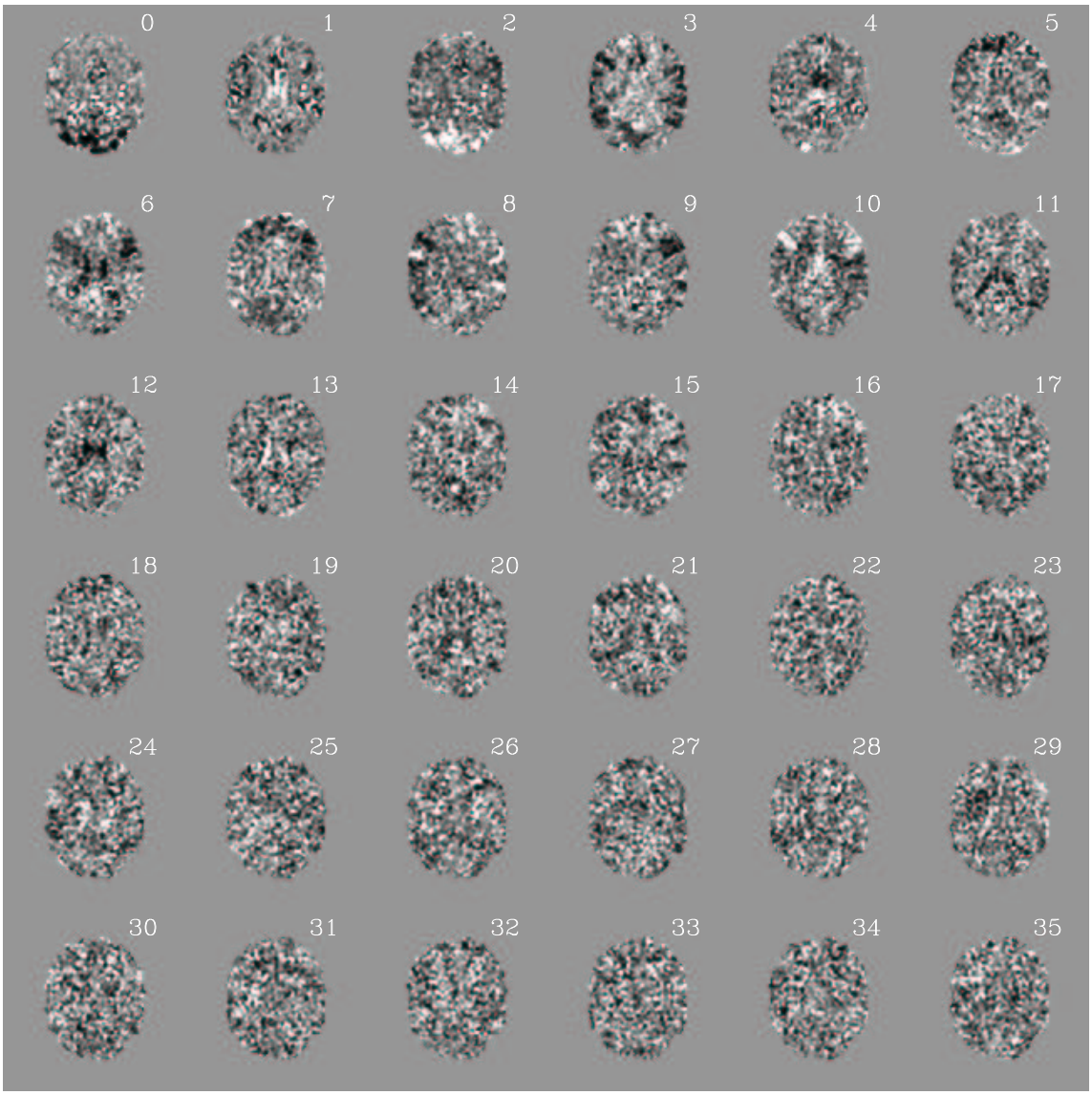


Figure 7: Coefficient maps for principal component analysis of the visual stimulation fMRI experiment presented in figs. 3 and 5. The figure shows the coefficient maps for the components with the 36 largest eigenvalues in descending order.

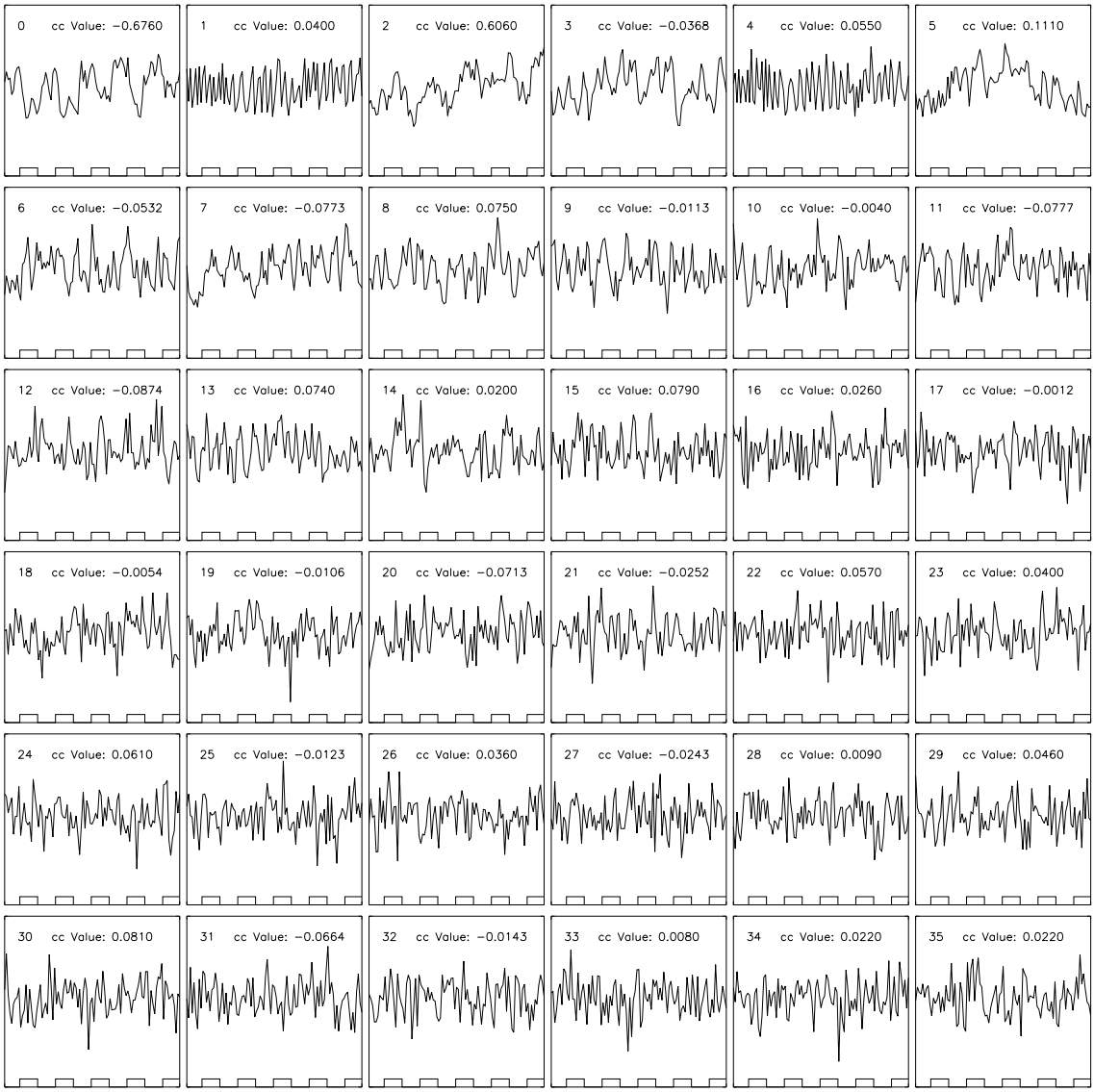


Figure 8: Eigenvectors for principal component analysis of the visual stimulation fMRI experiment presented in figs. 3 and 5. The figure shows the eigenvectors for the components with the 36 largest eigenvalues in descending order. *ccValue* denotes the cross-correlation of each eigenvector with the stimulus function. The time-course of the stimulus is shown at the bottom of each plot.

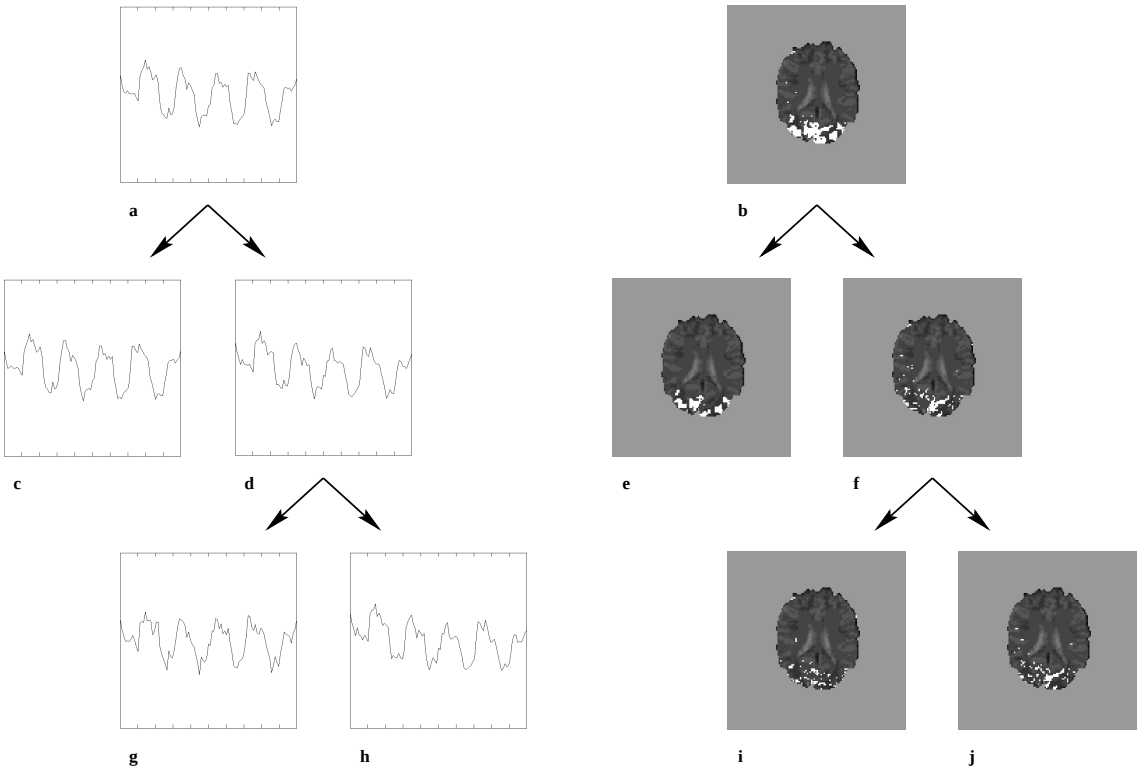
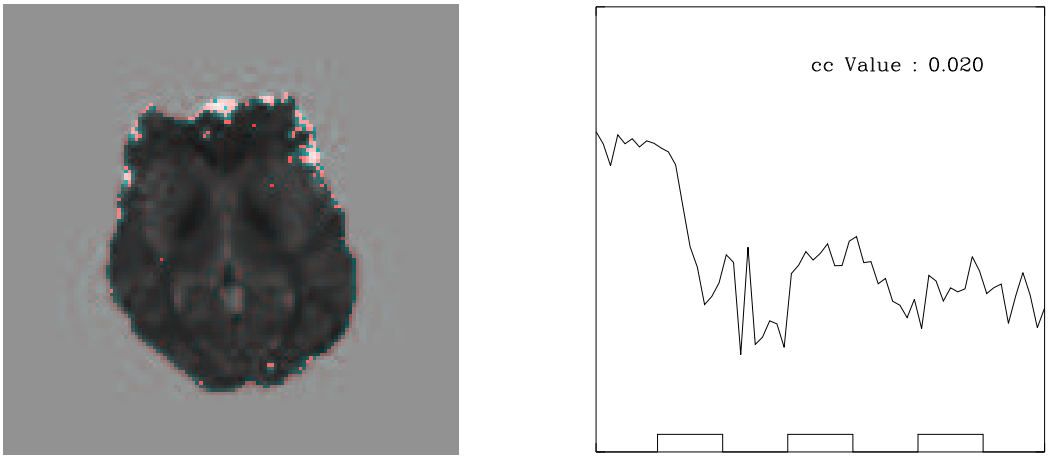


Figure 9: Hierarchical clustering tree demonstrating cluster separation during deterministic annealing by minimal free energy VQ. (a) Cluster center before phase transition, i. e. cluster separation. (b) Corresponding cluster of pixels before phase transition. (c,d) Cluster centers after phase transition. (e,f) Corresponding clusters of pixels after phase transition. The cluster center (d) is split again within a second phase transition into the clusters (g,h). (i,j) Corresponding pixel clusters after the second phase transition. The three lines of the figure represent the status of the VQ procedure at gradually declining cooperativity parameter ρ with $\rho = 1.81$ for (a,b) $\rho = 1.79$ for (c)–(f), and $\rho = 1.64$ for (g)–(j).



(a) (b)

Figure 10: Cluster assignment map and respective codebook vector representing a cluster that is highly indicative of a through-plane motion artifact gained from a 64 scan visual stimulation experiment. (a) Cluster assignment map, (b) codebook vector.

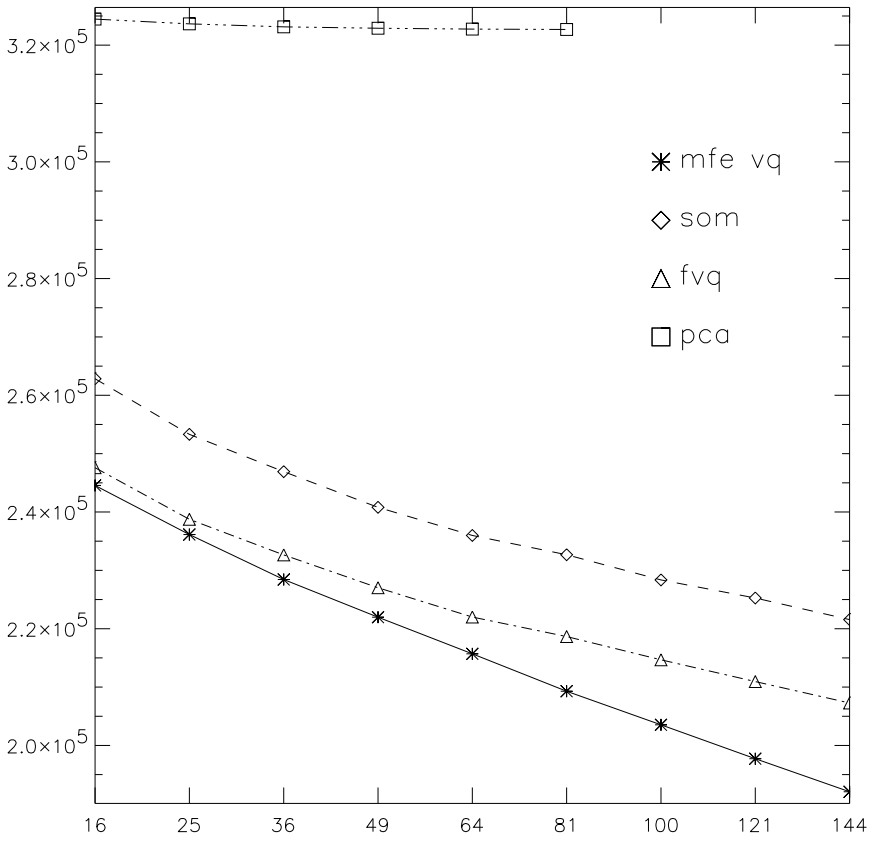


Figure 11: Comparative quantitative evaluation of the different unsupervised learning techniques using the 'hard assignment' criterion function (21) for the visual stimulation fMRI data presented in figs. 3 and 5. The criterion function has been computed for quality assessment of minimal free energy VQ, self-organizing maps, and fuzzy c-means VQ using different numbers of codebook vectors for each method along with principal component analysis using different numbers of components in the order of descending eigenvalues. Note the quadratic scaling of the abscissae.

Algorithm	16 CVs	36 CVs	100 CVs
Minimal free energy VQ	3502	8222	18051
Fuzzy c-means VQ	1654	4392	11260
Self-organizing maps	65	115	175

Table 1: Computation times in seconds for clustering of fMRI data from a 100 scan visual stimulation experiment by minimal free energy VQ, fuzzy c-means VQ, and self-organizing maps using different numbers of codebook vectors on a Sun Ultra5 workstation with an UltraSPARC-IIi 333 MHz CPU. Computation time for principal component analysis is 3 s.

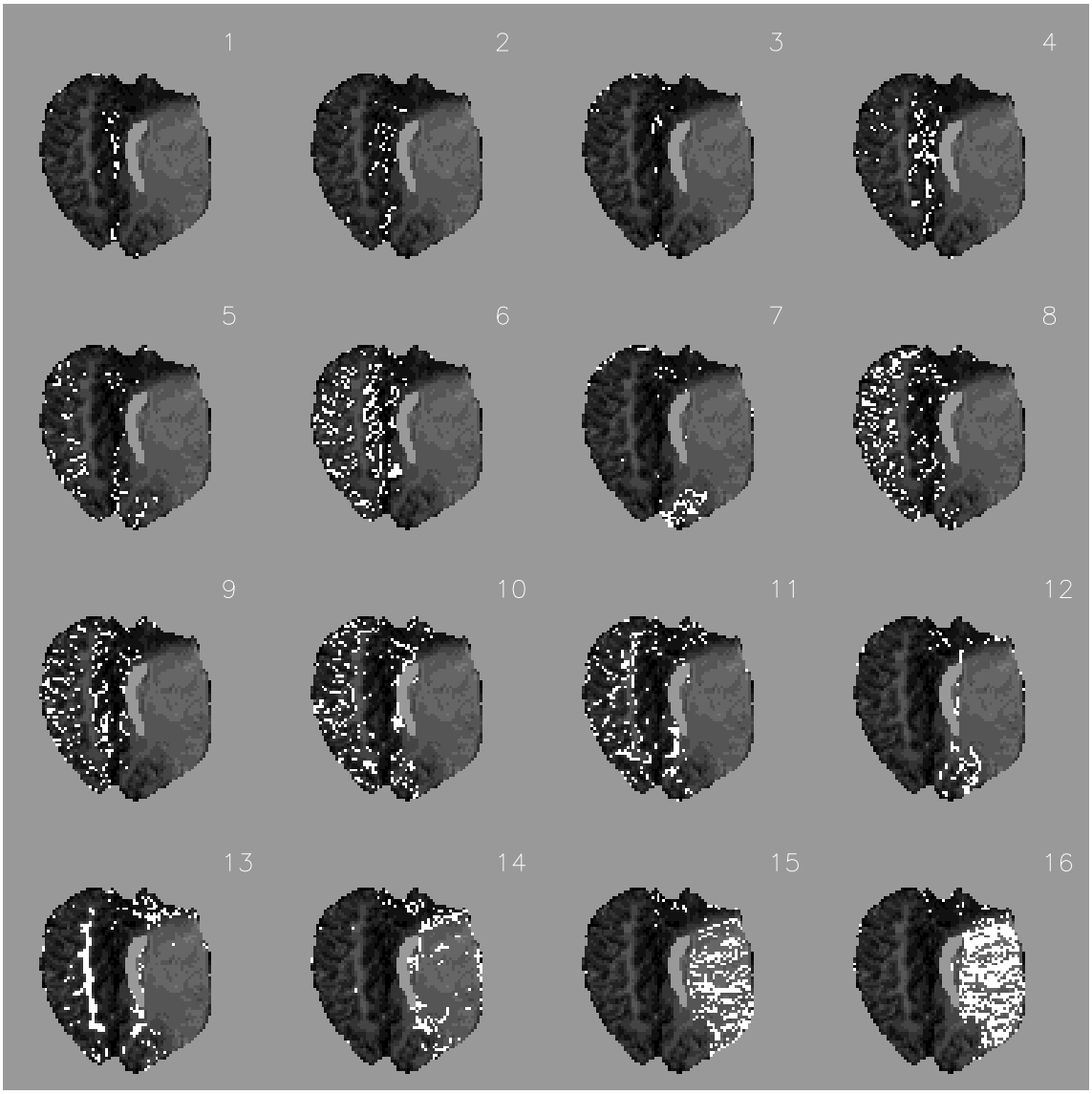


Figure 12: Cluster assignment maps for minimal free energy vector quantization of a dynamic perfusion MRI study in a patient with a large stroke in the supply region of the left middle cerebral artery.

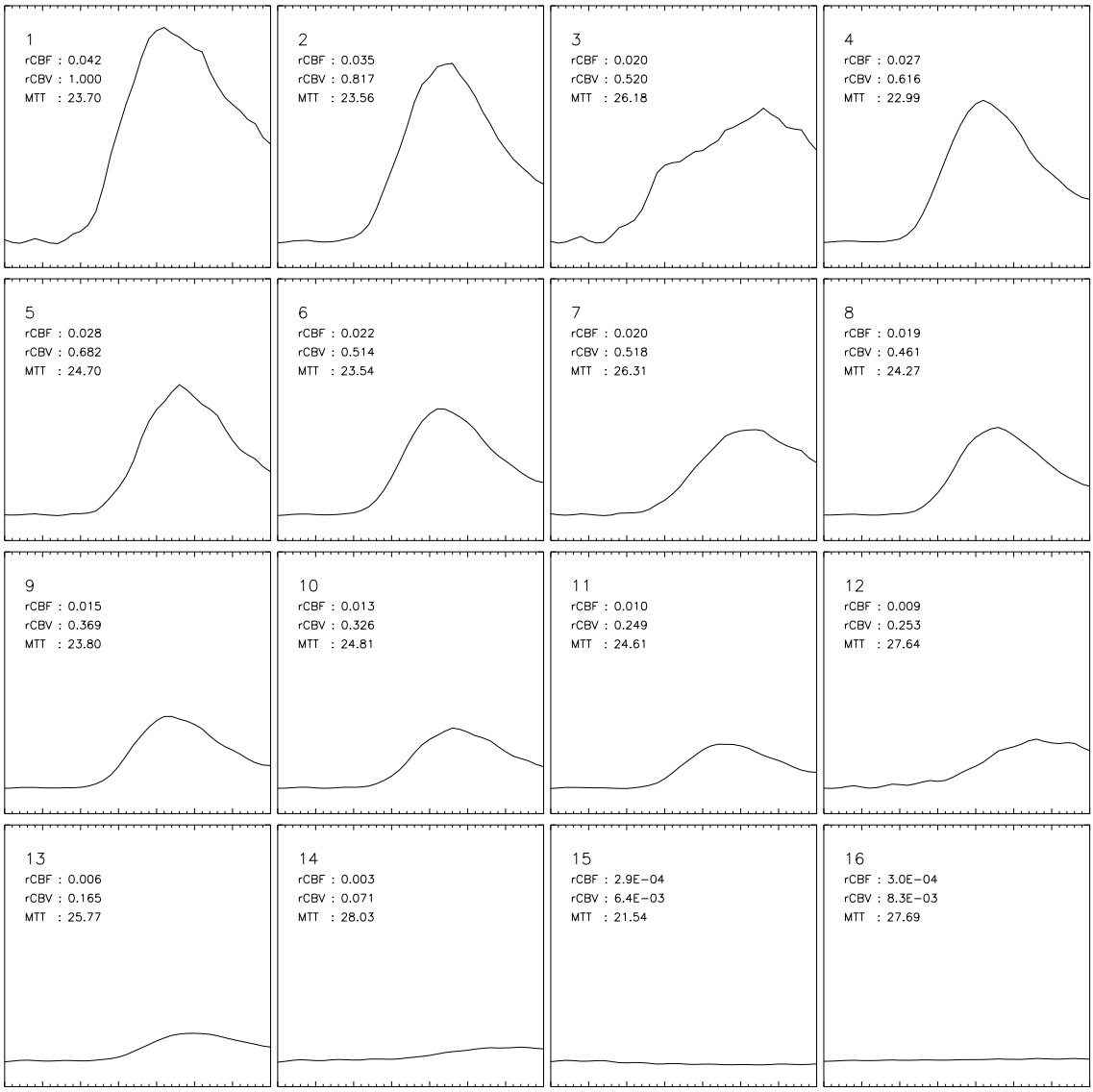


Figure 13: Cluster-specific concentration-time curves for minimal free energy vector quantization of a dynamic perfusion MRI study in a patient with a large stroke in the supply region of the left middle cerebral artery. Cluster numbers correspond to fig. 12. MTT values are indicated as multiples of the scan interval (1.5 s), rCBV values are normalized w.r.t. the maximal value (cluster #1).

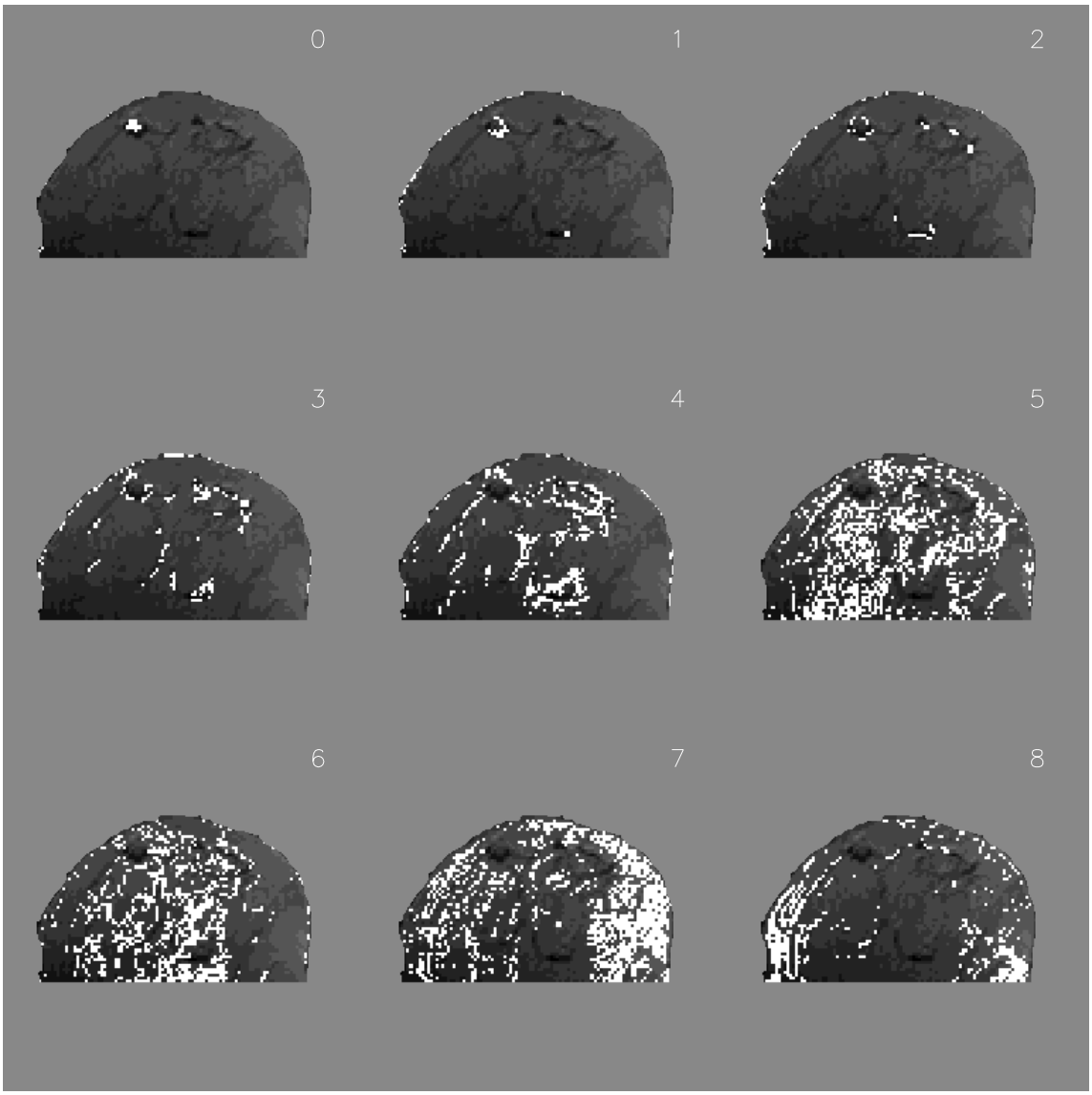


Figure 14: Cluster assignment maps for minimal free energy vector quantization of a dynamic MRI mammography study in a patient with histologically confirmed breast cancer in the upper medial quadrant of the left breast.

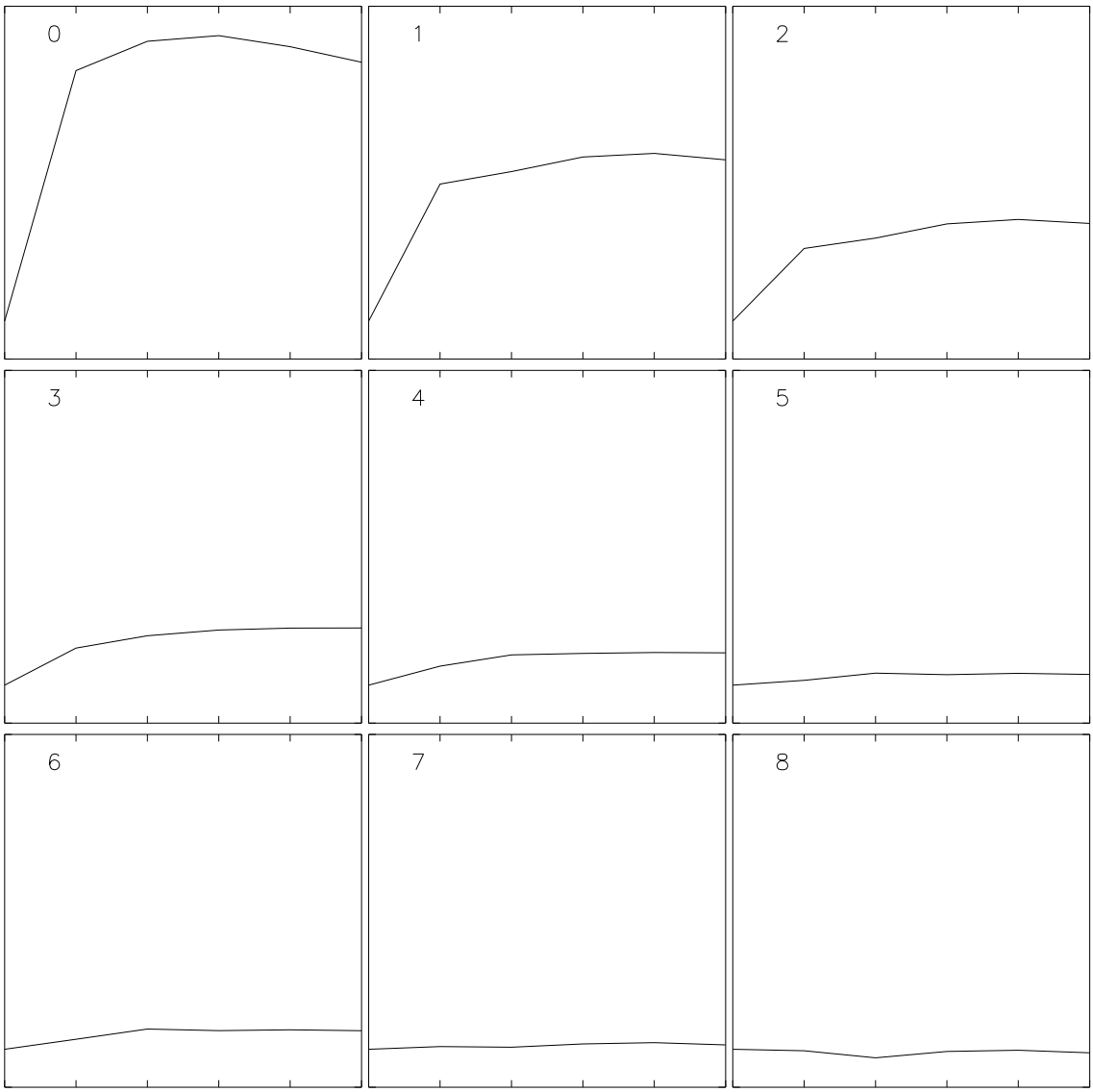


Figure 15: Codebook vectors for minimal free energy vector quantization of a dynamic MRI mammography study in a patient with histologically confirmed breast cancer in the upper medial quadrant of the left breast. Cluster numbers correspond to fig. 14.