Feature Detection in fMRI Data: The Information Bottleneck Approach

Bertrand Thirion, Olivier Faugeras

Odyssée Laboratory (ENPC-Cermics/ENS-Ulm/INRIA)

Abstract Clustering is a well-known technique for the analysis of fMRI data, whose main advantage is certainly flexibility: given a metric on the dataset, it defines the main features contained in the data. But intrinsic to this approach are also the problem of defining correctly the quantization accuracy, and the number of clusters necessary to describe the data. The Information Bottleneck (IB) approach to vector quantization, proposed by Bialek and Tishby, addresses these difficulties: 1) it deals with an explicit tradeoff between quantization and data fidelity; 2) it does so during the clustering procedure and not post hoc; 3) it takes into account the statistical distribution of the features within the feature space and not only their most likely value; last, it is principled through an information theoretic formulation, which is relevant in many situations. In this paper, we present how to benefit from this method to analyze fMRI data. Our application is the clustering of voxels according to the magnitude of their responses to several experimental conditions. The IB quantization provides a consistent representation of the data, allowing for an easy interpretation.

1 Introduction

Functional Magnetic Resonance Imaging (fMRI) of blood oxygen level-dependent (BOLD) contrast is a common tool for the localization of brain processes associated with any kinds of psychological tasks. It is in fact an indirect measure of the latter, based on brain oxygenation. Moreover, many confounds are known to be present in fMRI data (subject movements, respiratory and heart artifacts, temperature drift, machine noise), making the analysis of such data a challenging task. Commonly used methods belong to one of the following families: i) hypotheses-based techniques (e.g. [6]), which parametrically fit a prior model to the data through analysis of variance or correlation and ii) exploratory techniques, that give an account of the data content with little prior knowledge, like Principal Components Analysis (PCA), Independent Components Analysis (ICA) or clustering.

In this paper, we describe another use of clustering which is based upon the recognition of structures present in the data, after some preprocessing. Actually, clustering analysis (C-means algorithm [1] , fuzzy C-means [2], [4], dynamical cluster analysis [3], deterministic annealing [12]) has been essentially used in fMRI data analysis to give a simplified account of the data by gathering voxels which have similar time courses. This similarity can be measured by the Euclidean distance in the signal space of origin [12] or another distance based on cross correlation [5], or a Mahalanobis metric [7]. These methods are efficient [2] and can isolate interesting patterns in the data, but they suffer from several limitations

– The choice of a correct metric is not obvious; an Euclidean metric can represent a suboptimal choice [8].
– Clustering algorithms can spend a lot of efforts trying to isolate patterns of no interest; this is due to the absence of prior information.
– The quality of clustering results is difficult to assess. To solve these problems, authors have proposed some heuristics [5] [9], but these are not necessarily optimal; moreover they are used after convergence of the algorithms, or sometimes yield complex multistage strategies [4].
– This point is related to the problem of the selection of the number of clusters [5]: It is intuitively clear that the choice of a given number of clusters corresponds to a certain bias/variance tradeoff, but this tradeoff is usually implicit.

These problems motivate the introduction of a new clustering method by Bialek et al. [11], namely the Information Bottleneck (IB) method. This method performs a kind of fuzzy quantization of the data, but by minimizing a function that explicitly balances quantization efficiency and data fidelity. Here we try to preserve the estimated voxel-based response to the experimental stimuli, given the uncertainty of these responses measured by a dispersion matrix.

The paper is organized as follows: in section 2, we present how to build a low dimensional *feature space* from fMRI datasets. Then we show how to quantify it with the IB formulation . We illustrate and validate the method on a synthetic example in section 3, and present results on a real dataset. Last, we discuss the limitations of the method and possible extensions in section 4.

2 Methods

fMRI Data Analysis: Let us denote Y a fMRI dataset, considered as an $N \times T$ matrix, where N is the number of voxels in the dataset, and T the length of the time series. $Y_n(t)$ is thus the signal at voxel n and time t. We assume that the subject undergoes different conditions of a given experimental paradigm. We model the effects of interest in the experiment as temporal regressors $G = (g_r(t)), r = 1..R, t = 1..T$. For instance, the temporal regressors may include the time courses of the experimental conditions convolved with a hemodynamic filter (hrf) and potential confounding signals (motion estimates, constant, low frequency signals).

As in [6] we compute the projection of the data in the space spanned by the regressors $(g_r, r = 1..R)$:

$$Y_n(t) = \sum_{r=1}^{R} \gamma_r(n) g_r(t) + \epsilon_n(t), \tag{1}$$

for $t = 1..T$ and $n = 1..N$, where $\gamma(n) = (\gamma_r(n))_{r=1..R}$ is the projection of Y_n onto the rows of G. $\gamma(n)$ is obtained in a least-square sense, together with the dispersion matrix:

$$\hat{\gamma}(n) = (GG^T)^{-1} G Y_n^T \tag{2}$$

$$\Lambda_\gamma(n) = (GG^T)^{-1} \frac{\sum_{t=1}^{T} \epsilon_n(t)^2}{T - rank(G)} \tag{3}$$

Since some of the regressors are potentially of no interest (they are used only for estimation improvement purposes), we only consider $(\gamma_r(n), r = 1..S \leq R)$ and the corresponding reduced dispersion matrix, which we still note $\Lambda_\gamma(n)$.

Next we propose to study the estimates of $\gamma(n)$ as a *feature space* through clustering/vector quantization, taking into account the uncertainty in the estimation of the response, $(\Lambda_\gamma(n))$.

Data Quantization within the Information Bottleneck Framework: The IB method, described in [11], addresses the following problem: Given a discrete dataset X (the set of voxels, isomorphic to $[1, .., N]$), a space of interest Γ (the set of possible values for γ), and the conditional probability densities (pdf), here the normal densities

$$p(\Gamma|X = n) = \mathcal{N}(\hat{\gamma}(n), \Lambda_\gamma(n)) \tag{4}$$

find the *fuzzy clusters* $\tilde{\mathrm{X}}$ that maximize compression while retaining most of the information on $p(\Gamma|X)$. In mathematical terms this leads to the minimization of the quantity

$$I(X, \tilde{X}) - \beta I(\tilde{X}, \Gamma) \tag{5}$$

with respect to $\tilde{X}$, where $I(X, \tilde{X})$ is the mutual information between the dataset and its compressed representation, $I(\tilde{X}, \Gamma)$ is the mutual information between the compressed representation and the variable of interest, and β a positive scalar. The minimization of $I(X, \tilde{X})$ yields compression of the original data X into $\tilde{X}$, while the maximization of $I(\tilde{X}, \Gamma)$ implies that the compressed data must preserve as much information as possible on Γ. The problem, when stated in this manner, has been shown to have a formal solution: Given

$$p(\gamma|\tilde{x}) = \frac{1}{p(\tilde{x})} \sum_x p(\gamma|x) p(\tilde{x}|x) p(x), \tag{6}$$

$$Z(x, \beta) = \sum_{\tilde{x}} p(\tilde{x}) \exp\left(-\beta \sum_\gamma p(\gamma|x) \log \frac{p(\gamma|x)}{p(\gamma|\tilde{x})} \right), \tag{7}$$

in terms of $p(\tilde{x}|x)$, the solution satisfies the equation

$$p(\tilde{x}|x) = \frac{p(\tilde{x})}{Z(x, \beta)} \exp\left(-\beta \sum_\gamma p(\gamma|x) \log \frac{p(\gamma|x)}{p(\gamma|\tilde{x})} \right) \tag{8}$$

$\sum_\gamma p(\gamma|x) \log \frac{p(\gamma|x)}{p(\gamma|\tilde{x})}$ is nothing but the Kullback-Leibler divergence between the two pdfs $p(\gamma|x)$ and $p(\gamma|\tilde{x})$, which we write henceforth as $d(x, \tilde{x})$. Equation (7) rewrites

$$Z(x, \beta) = \sum_{\tilde{x}} p(\tilde{x}) \exp\left(-\beta d(x, \tilde{x})\right) \tag{9}$$

This problem does not have a closed form solution. Nevertheless, the following result holds: *Equation (8) is satisfied at the minima of the functional*

$$\mathcal{F}(p(\tilde{x}|x), p(\tilde{x}), p(\gamma|\tilde{x})) = -\left\langle \log Z(x, \beta) \right\rangle_{p(x)} = I(X, \tilde{X}) + \beta \left\langle d(x, \tilde{x}) \right\rangle_{p(x, \tilde{x})} \tag{10}$$

where $\langle S(a) \rangle_{p(a)}$ stands for the expectation of the quantity S for the probability law p. The minimization can be done independently over the sets of the normalized

distributions $p(\tilde{x})$, $p(\tilde{x}|x)$ and $p(\gamma|\tilde{x})$ by the converging alternating iterations (t being here the iteration step):

$$p_t(\tilde{x}|x) = \frac{p_t(\tilde{x})}{Z_t(x,\beta)} exp(-\beta.d(x,\tilde{x})) \tag{11}$$

$$p_{t+1}(\tilde{x}) = \sum_x p(x) p_t(\tilde{x}|x) \tag{12}$$

$$p_{t+1}(\gamma|\tilde{x}) = \sum_x p(\gamma|x) p_t(x|\tilde{x}) \tag{13}$$

The above algorithm provides a solution which may be suboptimal (in fact a local minimum of $\mathcal{F}$, exactly as any EM algorithm). An excessive number of clusters are generated randomly at the beginning; the IB algorithm (eq. (11), (12), (13)) is applied to the data until convergence (typically a few hundred iterations); we then use the final probability laws $p(\tilde{x}|n)$ for a hard clustering of the data ($cl(n) = argmax_{\tilde{x}}\, p(\tilde{x}|n)$); the final number of clusters is given by the ones whose probability has not canceled during the iterations (i.e. $\{\tilde{x}/\exists n/\tilde{x} = cl(n)\}$). The number of remaining clusters is thus provided by the algorithm and depends highly on the choice of β, whose interpretation as a scale parameter is obvious.

In practice, the use of a finite grid for the sampling of the pdfs is important. From our experiments, it seems that the grid precision does not have much importance on the final result, as long as it is not coarser than the intrinsic data dispersion.

3 Experimental Results

Synthetic data: We have created a synthetic dataset by simulating one slice of fMRI data containing $N = 1963$ voxels. 3 small foci of 21 voxels are created and an independent gaussian noise is added to all voxels, so that the SNR is 0.5 in the *activated areas*. The length of the series is $T = 200$; the simulated paradigm comprises two conditions (see figure 1(a)); the simulated time courses, and spatial maps are presented in figure 1(b), (c). The data has been smoothed spatially as commonly done for fMRI. Through equations (2) (3), we obtain a $S = 2$ dimensional feature-space. We have displayed the estimated feature at each voxel in figure 1(d). Then, we have discretized the feature space on a (20×20) grid and analyzed it with the IB method. To study the dependence of the number k of final clusters on β, we present the cluster hierarchy, indexed by β, in figure 1 (e).

Figure 1 (e) shows that the 4 clusters configuration is the main non-trivial one. The associated p.d.f. $p(\gamma|\tilde{x})$ (figure 1(f)) confirms the pertinence of this model. For comparison, we have applied a fuzzy-C-Means algorithm with 4 clusters on the same feature space, with 10^4 independent random initializations. In no case did we obtain the results described in figure 1. This may be attributable to the small number of activated voxels, and to the inadequate choice of the Euclidean metric.

Real data: The data is taken from an experiment published in [10]. The present analysis is reduced to one subject performing the following experiment (called fMRI2 in [10]): A visual stimulation is performed, with 4 conditions: Heading, dimming static, dimming flow, and baseline. The *Heading* condition means that the subject views a ground plane optic flow pattern that simulates self-motion;

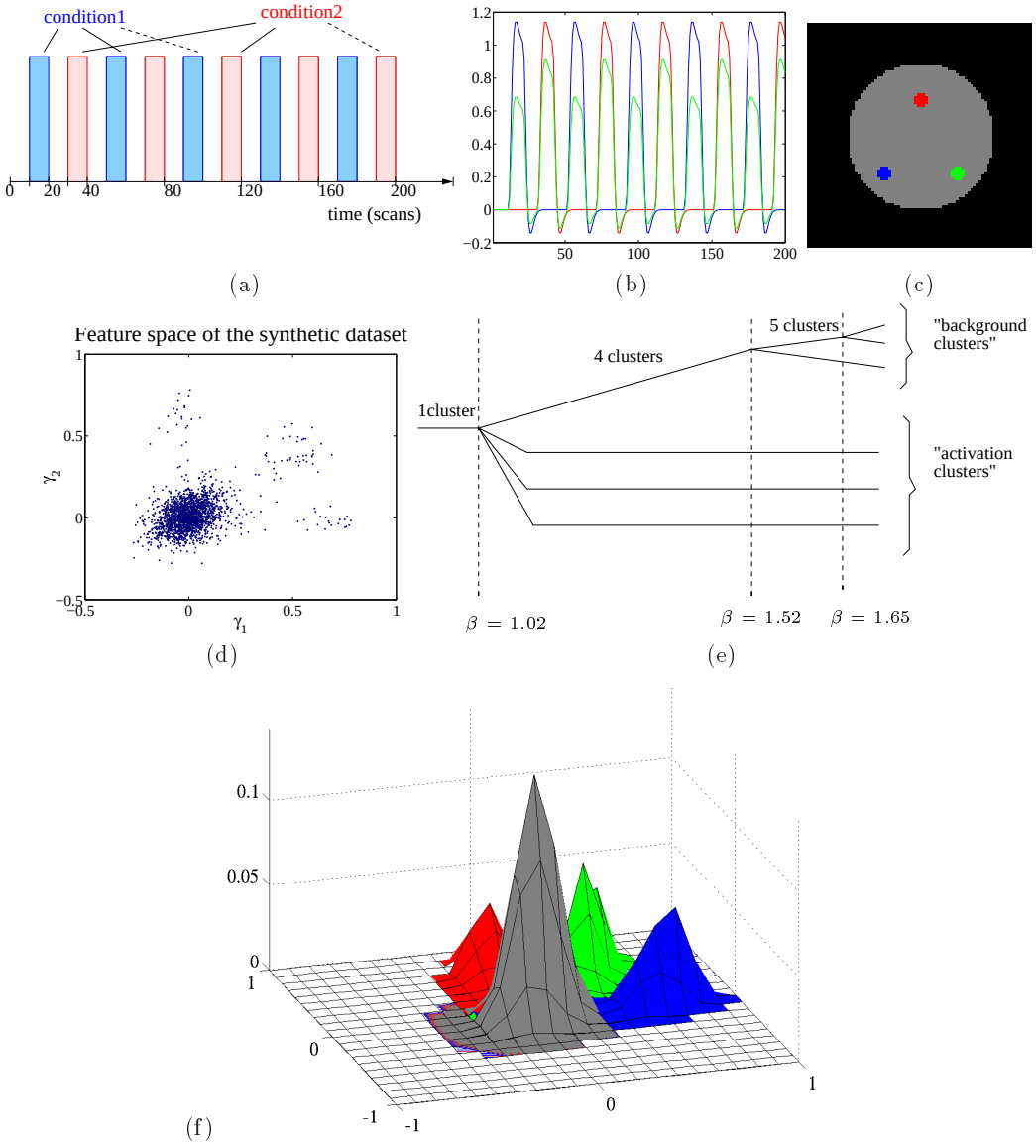

Figure 1. (a) Simulated experimental paradigm (two conditions, alternating block design with resting periods). (b) Synthetic activations time courses. The three patterns are obtained by convolving the canonical hrf with three different linear combination of the stimuli time courses. (c) Spatial layout of the activations simulated in the experiment. The colors are those of the time course. (d) Estimated features at each voxel $\hat{\gamma}(n) = (\hat{\gamma}_1(n), \hat{\gamma}_2(n))$ (the dispersion is not represented). (e) Cluster hierarchy obtained by letting the scale parameter β vary. Clusters appear by successive bifurcations or splittings. The terms *activation clusters* and *background clusters* refer to post hoc interpretation. The configuration with 4 clusters is stable over a large scale interval; we refer to this configuration in the remainder of the section. The associated spatial map $cl(n)$ (not shown) is identical to the original activation map displayed in (c). (f) Probability density functions associated with the four clusters $p(\gamma|\tilde{x})$. Note that they correspond readily to the main mode and the three "arms" of the feature distribution clearly visible in figure (d). Their color match with the color of figures (b) and (c).

dimming static is a control task where no self-motion is simulated, but a part of the stimulus display is slightly dimmed, and *dimming flow* is another control condition specially designed to disentangle spatial and featural attention through the dimming of the ground plane flow.

Details about the data are available in [10]. Let us only notice that the number of scans is $T = 720$ scans, and that the number of voxels considered here is $N = 30094$. We derive a 3 dimensional feature-space $(\gamma_1, \gamma_2, \gamma_3)$, which is discretized on a finite $(13 \times 13 \times 13)$ grid. The hierarchy of clusters obtained with the IB algorithm is described in figure 2 (d). We concentrate only on the three clusters that are significantly far from 0: one of them shows negative patterns, and the other two present positive responses, one cluster having higher scores. The cluster maps are given in figure 2 (a), together with spatial maps of the contrasts of interest *heading-dimming static* (b) and *heading-dimming flow* (c) obtained from the standard SPM procedure. The average feature per cluster is displayed in figure 2 (e).

The results obtained here are consistent with those obtained with standard Statistical Maps: the *green* cluster corresponds broadly to the negative part of either SPM map (in blue), while the *red* and *blue* clusters correspond more to the positive patterns, with the *red* cluster corresponding to the maxima of the activation maps; nevertheless, the interpretation of the two latter clusters in terms of contrasts is more subtle, as can be seen in figure 2(e), the contrast *heading-dimming flow* being positive for only the *blue* cluster. The relative decrease of the contrast amplitude between the two SPM maps has been described in [10]; our clustering method makes it clearer by distinguishing between two patterns. Last, the existence of negative signals had not been reported in [10]; we do not have other interpretation for them than the competition between different cognitive tasks.

4 Discussion

Our method is based on the specification of the feature space made prior to data analysis. This is a difference with respect to current exploratory methods used for fMRI datasets (independent components analysis, clustering). The rationale for that choice is that the whole signal is not interesting to the experimenter, but only parts of it that contain relevant information, i.e. essentially consistently task-related responses. It is thus dependent on the correct specification of the space of interest; but in the bloc experiments considered here, at least a very good approximation to the actual response can be computed a priori using a standard hemodynamic response function (hrf).

In spite of the simplicity of the simulated dataset, it appeared that a fuzzy C-means method does not yield the intuitively correct *4 clusters* solution that corresponds actually to the generative process of the dataset. Here the IB performs clearly better; moreover, it gives a truly hierarchical representation of the data indexed by β and takes into account the dispersion of the estimators. On the other hand, the implementation of the method implies the use of a discretized pdf of the feature vector for each voxel; this can be done only within low dimensional feature spaces. We face here the well-known curse of dimensionality which standard clustering techniques (C-means, fuzzy C-means) do not suffer from, or at least less critically. The computational cost of the method is reasonable in our implementation, in spite of the number of voxels considered, and the number of iterations that

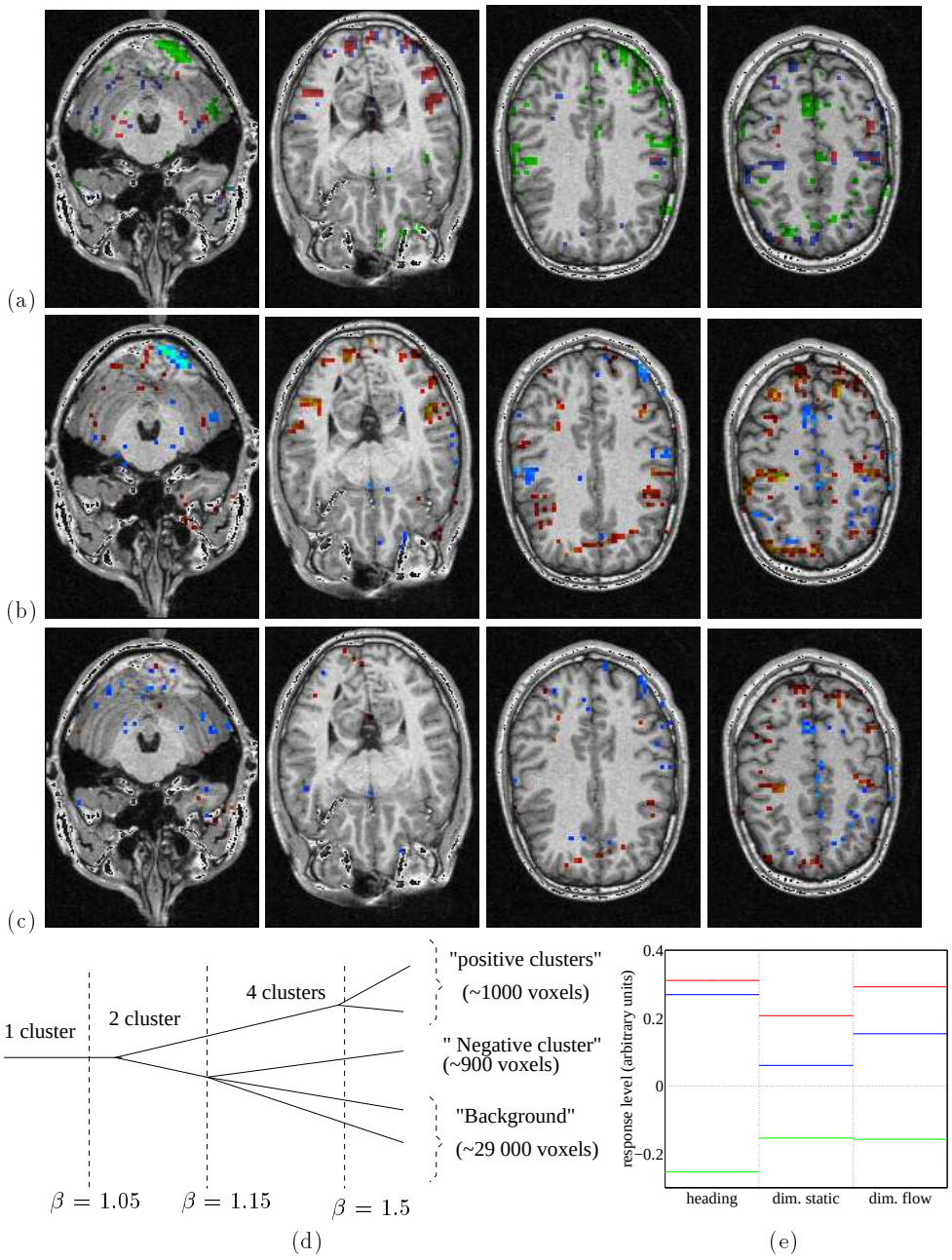

Figure 2. (a) 4 axial slices extracted from the spatial maps obtained with our clustering method, (b) a SPM map of the test of the contrast *heading-dimming static*, and (c) a SPM map of the test of the contrast *heading-dimming flow*, both thresholded at the level $|t| = 2.5$. On the cluster maps, the colors correspond to those employed in figure (e); on the SPM maps, positive activations are represented in red-yellow, while negative responses are represented in blue. (d) Cluster hierarchy obtained by letting the scale parameter β vary. The clusters appear by successive bifurcations or splittings. The terms *positive clusters* and *negative clusters* refer to the post hoc interpretation. We do not further study the "background" clusters. (e) Centers of the three clusters in the feature space. Two of the clusters (red and blue) represent a positive signal, while the green one represents a negative signal. Interestingly, the contrast "heading-dimming flow" is null for the red pattern while it is positive for the blue pattern.

are necessary for convergence. For example, we need about one minute to process the real dataset.

Future work involves the test on more realistic simulations, the use of analytical approximations of Kullback-Leibler divergence between pdfs, e.g. with gaussian mixture models, the statistical inference at the cluster level, and multiple initializations of the algorithm on pre-clustered data in order to approach optimal solutions.

Conclusion: Clustering can be used for analyzing fMRI data beyond purely exploratory analysis: it is also a tool to study the data structure within a specified feature space. Among known clustering techniques, the Information Bottleneck gives a principled way to handle the robustness/accuracy tradeoff -but does not solve it- and gives a practical solution to the selection of the cluster numbers. Additionally, it takes into account the dispersion in the estimation of the feature data, which is more realistic than what usual clustering procedures allow for.

Acknowledgments: We wish to thank Professor G. Orban, S.Sunaert and H. Peuskens, who provided us with the functional MR images we used. The work was developed in collaboration with the laboratory of Neurophysiology, K.U.Leuven, Medical School, Leuven, Belgium (LEUNEURO), directed by G. Orban.

References

1. Daniela Balslev, Finn A. Nielsen, et al. Cluster analysis of activity-time series in motor learning. *Human Brain Mapping*, 15:135–145, 2002.
2. 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.
3. A. Baune, F.T. Sommer, M. Erb, D. Wildgruber, B. Kardatzaki, and W. Palm, G.and Grodd. Dynamical cluster analysis of cortical fMRI activation. *NeuroImage*, 9:477–489, 1999.
4. M.J. Fadili, S. Ruan, D. Bloyet, and B. Mazoyer. A multistep unsupervised fuzzy clustering analysis of fMRI time series. *Human Brain Mapping*, 10:160–178, 2000.
5. M.J. Fadili, S. Ruan, D. Bloyet, and B. Mazoyer. On the number of clusters and the fuzziness index for unsupervised fca applications to bold fMRI time series. *Medical Image Analysis*, 5:55–67, 2001.
6. K.J. Friston, J. Ashburner, et al. *spm 97 course notes*. Wellcome Department of Cognitive Neurology, University College london, 1997.
7. C. Goutte, P. Troft, E. Rostrup, A. Nielsen, and L.K. Hansen. On clustering fMRI time series. *NeuroImage*, 9(3):298–310, 1998.
8. Cyril Goutte, Lars Kai Hansen, Mattew G. Liptrot, and Egill Rostrup. Feature space clustering for fMRI meta-analysis. *Human Brain Mapping*, 13(3):165-183.
9. U. Möller, M. Ligges, P. Georgiewa, C. Grünling, W. A. Kaiser, H. Witte, and B. Blanz. How to avoid spurious cluster validation ? A methodological investigation on simulated and fMRI data. *NeuroImage*, 17:431–446, 2002.
10. H. Peuskens, S. Sunaert, P. Dupont, P. Van Hecke, and G.A. Orban. Human brain regions involved in heading estimation. *Journal of Neuroscience*, 21(7):2451–61, 2001.
11. Naftali Tishby, Fernando C. Pereira, and William Bialek. The information bottleneck method. In *Proc. of the 37-th Annual Allerton Conference on Communication, Control and Computing*, pages 368–377, 1999.
12. Axel Wismüller and Olivier Lange. Cluster analysis of biomedical image time-series. *International Journal of Computer Vision*, 46(2):103–128, 2002.