A supervised clustering approach for extracting predictive information from brain activation images

Vincent Michel, Evelyn Eger, Christine Keribin, Jean-Baptiste Poline, Bertrand Thirion

A supervised clustering approach for extracting predictive information from brain activation images

Vincent Michel
Parietal team, INRIA Saclay-Ile-de-France, France
Université Paris-Sud 11, Orsay, France
CEA, DSV, I2BM, NeuroSpin, Gif/Yvette, France
vincent.michel@inria.fr

Evelyn Eger
INSERM U562, Gif/Yvette, France
CEA, DSV, I2BM, NeuroSpin, Gif/Yvette, France

Christine Keribin
Université Paris-Sud 11, Orsay, France
Select team, INRIA Saclay-Ile de France, France

Jean-Baptiste Poline
CEA, DSV, I2BM, NeuroSpin, Gif/Yvette, France

Bertrand Thirion
Parietal team, INRIA Saclay-Ile-de-France, France
CEA, DSV, I2BM, NeuroSpin, Gif/Yvette, France

Abstract

It is a standard approach to consider that images encode some information such as face expression or biomarkers in medical images; decoding this information is particularly challenging in the case of medical imaging, because the whole image domain has to be considered a priori to avoid biasing image-based prediction and image interpretation. Feature selection is thus needed, but is often performed using mass-univariate procedures, that handle neither the spatial structure of the images, nor the multivariate nature of the signal. Here we propose a solution that computes a reduced set of high-level features which compress the image information while retaining its informative parts: first, we introduce a hierarchical clustering of the research domain that incorporates spatial connectivity constraints and reduces the complexity of the possible spatial configurations to a single tree of nested regions. Then we prune the tree in order to produce a parcellation (division of the image domain) such that parcel-based signal averages optimally predict the target information. We show the power of this approach with respect to reference techniques on simulated data and apply it to enhance the prediction of the subject's behaviour during functional Magnetic Resonance Imaging (fMRI) scanning sessions. Besides its superior performance, the method provides an interpretable weighting of the regions involved in the regression or classification task.

1. Introduction

Inferring behavioral information or cognitive states from activation brain images such as those obtained with functional magnetic resonance imaging (fMRI) is a recent approach in neuroimaging [1] that can provide more sensitive analyses than standard statistical parametric mapping procedures [7]. Specifically, it can be used to assess the involvement of some brain regions in certain cognitive or perceptual functions, by evaluating the accuracy of the prediction of a behavioral variable of interest (the *target*) when the classifier is instantiated on these brain regions.
This inference generally uses a prediction function such as a classifier that relates the image data to relevant variables. Many methods have been tested for classification or regression of activation images (Linear Discriminant Analysis,

Support/Relevance Vector Machines, Lasso, elastic net regression and many others), but in this problem the major bottleneck remains the extraction of predictive information within the brain volume (see [6] for a review). Feature selection is important both to achieve accurate prediction (by alleviating the curse of dimensionality) and to understand the spatial distribution of the informative features. In particular, when the number of *features* (voxels, regions) is much larger than the numbers of samples (images), the prediction method overfits the training set, and thus does not generalize well.

Multivariate feature reduction is an NP-hard problem, that can only be solved approximately. To date, the most widely used method for feature selection is voxel-based Anova (Analysis of Variance), that evaluates each brain voxel independently. In that case, spatial information is not used, and selected features can be redundant. By contrast, an algorithm for extracting information from image-based datasets can be specified as follows:

(i) A multivariate model: The information of interest can be distributed over distant brain regions. Feature selection should be able to account for combinations of signals over these different brain sites, hence it should be a multivariate approach. For instance, [10] shows how crucial multivariate pattern analysis is to make accurate predictions.

(ii) Taking into account the spatial structure of the data: Due to the spatial structure of fMRI data, there is a local redundancy of the predictive information, which should be considered in the feature building procedure, e.g. by replacing voxel-based signals by local averages. For instance, the searchlight approach [9], takes into account the local information in the image, but it cannot handle long-range interactions in the information coding.

(iii) A multi-scale approach: Given that the investigated regions are wide if there is little prior information, while the truly informative regions can be relatively tiny, we need an approach that focuses on compact subregions of the search volume: a multi-scale approach might thus be useful to optimize the definition of predictive regions. Unlike purely geometrical clustering approaches such as [8], procedures that use the signal for clustering might better respect the underlying data structure.

In this article, we develop spatial models that rely on hierarchical clustering to improve fMRI-based decoding. It has already been shown [4, 5] that a hierarchical multi-scale parcellation is pertinent for understanding brain network structure; here we develop this idea in the case of supervised classification. We call *parcellation* a division of the image domain into spatially connected units. Using parcel-based averages of fMRI signals to fit the *target* naturally reduces the number of features, hence allows tractable computations and accurate modeling. This raises the new challenge of optimizing the parcellation of the brain volume for the particular prediction task.
To address it, we first construct a hierarchical subdivision of the search domain. As the resulting nested parcel sets is isomorphic to a tree, we identify any tree cut with a given parcellation of the domain, and thus to a reduction of the available signal into parcel-based averages. We optimize the cut in order to maximize prediction accuracy by using a greedy approach and internal cross-validation. This is presented in Section 2. Importantly, this approach can focus on strongly informative, though spatially tiny regions, while leaving large uninformative regions of the search volume unsegmented. It is important to note that the cut definition takes into account the joint distribution of the data (across clusters), so that the final predictive model deals effectively with the feature covariance structure and is thus expected to be accurate – though global optimality cannot be guaranteed. We show in Section 3 that our method can recover the true spatial support of a discriminative pattern embedded in an image: as a consequence, it achieves higher prediction performances. Finally, we apply our approach to a real fMRI experiment, where we analyze brain activations associated with the mental processing of quantities. With the proposed approach, we achieve very significant fit of processing differences associated with the quantities involved, both within and across subjects. Moreover, our results on high-dimensional, but structured data such as brain activation images suggest that our approach can be applied to any type of data, where spatial structure is important, such as medical images.

2. Methods

After introducing the notations, we present the regression method that is used in this work, then we turn to our new feature selection method.

2.1. Introduction and notations

It is assumed that a set of activation images related to the presentation of different stimuli has been pre-computed, so that the image data can be viewed as a $N_p \times N_v$ matrix X, where N_v the number of voxels and N_p the number of samples (images). Typically, $N_v \sim 10^3$ to 10^5 (for a whole volume), while $N_p \sim 10$ to 10^2. In the sequel, we reduce the number of features to a certain number $N_f \ll N_v$. Note that the target Y is real-valued, and is thus fit through regression techniques.

Inverse inference is based on a framework that includes a

feature selection step to extract the informative features, a prediction method to infer the relationship between the fMRI signal X and the target Y to be predicted, and a cross-validation scheme that splits the available data into training and validation sets (here we use a leave-one-out procedure). See the flowchart in Fig. 2 for an overview of the whole procedure. Let X^l, Y^l be a learning set, X^t, Y^t a test set and X_i refer to the i^{th} feature.

We use elastic net regression (see [13]) to predict the target Y from a subset of features that we still denote X. These features can be the signal in a voxel or the mean signal within a parcel. We thus have:

$$Y = \sum_{i=1}^{N_f} X_i \beta_i + \epsilon \tag{1}$$

Estimation of the parameters $(\beta_i)_{i=1..N_f}$ requires a regularization since $N_p \ll N_f$. Elastic net criterion is defined as $L(\lambda_1, \lambda_2, \beta)_{X,Y} = \|Y - X\beta\|^2 + \lambda_2\|\beta\|_2^2 + \lambda_1\|\beta\|_1$, and predictions $\hat{Y}$ on the test set are computed as:

$$\hat{Y}^t(X^l, Y^l, X^t) = \sum_{i=1}^{N_f} X_i^t \hat{\beta}_i(X^l, Y^l), \tag{2}$$

where $\hat{\beta}(X^l, Y^l) = (1 + \lambda_2)\text{argmin}_\beta \left(L(\lambda_1, \lambda_2, \beta)_{X^l, Y^l}\right)$. The conventional parameterization for elastic net is specified by λ_2 (that we denote λ) and $s = \frac{\lambda_1}{\lambda_1 + \lambda_2}$, the fraction of the L_1 norm.

The performance of a regression model is evaluated using ζ, the ratio of explained variance (or R^2 coefficient):

$$\zeta(X^l, Y^l, X^t, Y^t) = \frac{\text{var}(Y^t) - \text{var}\left(Y^t - \hat{Y}^t(X^l, Y^l, X^t)\right)}{\text{var}(Y^t)} \tag{3}$$

This is the amount of variability in the response that can be explained by the model (perfect prediction yields $\zeta = 1$, while we might have $\zeta < 0$ if the prediction error is high). In the following, this value will be referred to as the fit criterion.

Next, we introduce spatial models in order to build a reduced number of fMRI features to fit the target data.

2.2. Features definition and selection

Construction of the hierarchical parcellation

To break the complexity of the problem, we first perform a hierarchical clustering of the voxel-based signals, under connectivity constraints, so that only spatially connected clusters are created. At that stage, we ignore the target information, but use the variance-minimizing approach of Ward's algorithm [12] in order to ensure that cluster-based averages provide a fair representation of the signal within each cluster. Only adjacent clusters can be merged together. The purpose of this procedure is to use the hierarchical parcellation to guide the search of informative regions within the volume of interest. Thus, at a given level in the hierarchy, the data is reduced to N_C cluster-based averages, which significantly decreases the computational complexity compared to a voxel-based approach with $N_v \gg N_C$ voxels.

Pruning of the tree

The hierarchical subdivision of the brain volume (by successive inclusions) is naturally identified as a tree; choosing a parcellation adapted to the regression problem means optimizing a cut of the tree, where the sub-trees created by the cut represent a region whose average signal is used for regression. As no optimal solution is currently available to solve that problem, we consider two approaches to perform such a cut (see Fig.1) :

- The first one consists in using the inertia criterion from Ward's algorithm: the cut consists in a subdivision of Ward's tree into its N_f main branches. As this does not take into account target information Y, we call it *unsupervised cut*.

- The second solution consists in initializing the cut to the highest level of the hierarchy and then successively finding the new subtree cut that maximizes the fit criterion. As in a greedy approach, successive cuts iteratively create a finer parcellation of the search volume. More specifically, one parcel is split at each step, where the choice of the split is driven by the prediction problem. After Δ such steps of exploration, the brain is divided into $\Delta + 1$ parcels. This procedure, that we call *supervised cut* is detailed in the algorithm 1.

Selection of the optimal subtree

In both cases, a set of nested parcellations is produced, and the optimal model among the available cuts still has to be chosen. This is done by computing a cross-validated generalization score within the training set, i.e. by averaging the values of ζ within a k-fold cross-validation on the training set. We select the subtree that yields the highest score.

Validation of the method

These procedures are performed on a learning dataset, which is split into train and test sets to optimize N_f. After learning, the validation dataset is subject to the same parcellation, and results are given in terms of cross-validated explained variance.

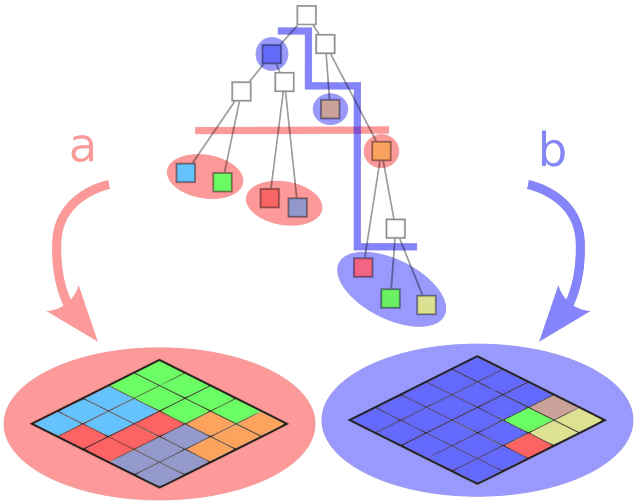


Figure 1. Illustration of two possible approaches to perform the pruning of the tree in order to obtain a given number of parcels (here 5). In the *unsupervised cut* approach (a), Ward's tree is divided into 5 parcels through a horizontal cut (red). The size of the parcels are similar. In the *supervised cut* approach (b), by choosing the cut (blue) in the tree that optimizes the prediction score, we are able to let large regions unsegmented, and to focus on some specific regions of the tree that are more informative.

Computational considerations

Our algorithm can be used to search informative regions in very high-dimensional data, where other algorithms such as elastic net do not scale well. At the current iteration $\delta \in [1, \Delta]$, $\delta + 1$ possible features are considered in the regression model, and the regression function is fit $N_p(\delta + 1)$ times, each call having a complexity $\mathcal{O}(\delta^3)$ when no particular optimization of the fit is performed. The overall cost complexity of all the procedure is thus $\mathcal{O}(N_p \Delta^5)$. In general $\Delta \ll N_f$, and the cost remains affordable as long as $\Delta < 10^3$, which was the case in all our experiments. Higher values for Δ might also be used, but in that case, some optimizations of Elastic Net should be used (early stopping, see [13]). Lasso could also be a cheaper alternative in such cases.

3. Experiments and results

We compare the results of the *supervised clustering* on different experiments with the results of the *unsupervised cut* algorithm and a univariate feature selection based on an F-test. The reference algorithms are optimized within a wide range of values for their respective parameters.

- This univariate selection is used with an elastic net regression (called *Enet*), with an optimized number of voxels found by cross-validation within the training set, in the range 50 to 250 in steps of 50 (only 50 to 150 for the simulated data). In the inter-subject study, the parameter of the elastic net are cross-validated within the training set, in the range 10^{-3} to 10^3 in multiplicative steps of 10 for λ, and in the range 0.1 to 1 in steps of 0.1 for s.

- Moreover, we use linear Support Vector Regression

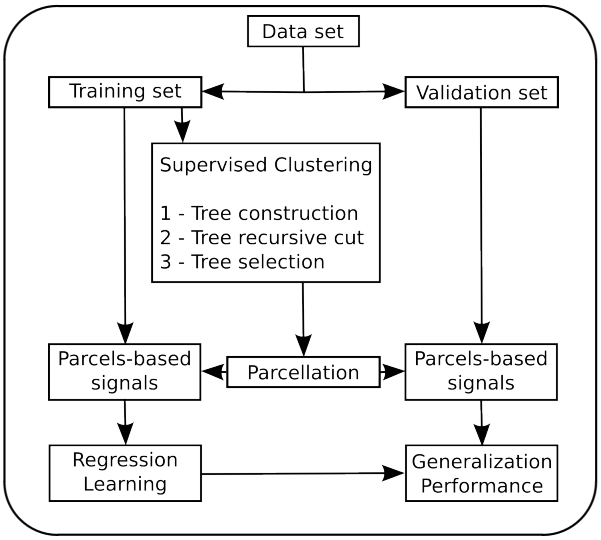


Figure 2. Global Flowchart for the supervised cut procedure.

Algorithm 1 Pseudo-code for the supervised cut algorithm.

Ensure: Let $\mathbb{T}$ be the tree constructed from the fMRI dataset X by Ward's algorithm. Let $\mathbb{P}_\delta$ be the set of parcels defined at the current iteration δ and X_δ the average signal per parcel obtained from $\mathbb{P}_\delta$.

Require: Set a number of exploration steps Δ. Let $\mathbb{P}_0 = \{P_1\}$, the top parcel of the tree.

for $\delta = 1$ to Δ **do**

for all $P_i \in \mathbb{P}_{\delta-1}$ **do**

- split $P_i \rightarrow (P_i^1, P_i^2)$ according to $\mathbb{T}$.
- set $P_{\delta-1,i} = \mathbb{P}_{\delta-1} - \{P_i\} \cup \{P_i^1, P_i^2\}$.
- compute the corresponding parcel-average signals $X_{\delta-1,i}$.
- estimate:

$$SC_i = \zeta(X^l_{\delta-1,i}, Y^l, X^l_{\delta-1,i}, Y^l)$$

end for

- perform the split $i^\star$ with the highest score $SC_{i^\star} \rightarrow$ new set of parcels : $\mathbb{P}_\delta = \mathbb{P}_{\delta-1,i^\star}$.
- compute the cross-validated regression score VC_δ of this new set of parcels using a leave-one-out procedure.

$$VC_\delta = \text{mean}_{j \in l} \left(\zeta(X_\delta^{l-\{j\}}, Y^{l-\{j\}}, X_\delta^{\{j\}}, Y^{\{j\}}) \right)$$

end for

return Retain the parcellation $\mathbb{P}_{\hat{\delta}}$ with the highest score $VC_{\hat{\delta}}$.

(called *SVR*), the C parameter being optimized by cross-validation in the range 10^{-4} to 10^{4} in multiplicative steps of 10; the number of voxels selected by the univariate feature selection is optimized by cross-validation within the training set in the range 100 to 2000 in steps of 100 (only 50 to 150 for the simulated data). Support Vector methods are reference methods for fMRI data-based prediction, see e.g. [1].

3.1. Simulated data

We test our algorithm on a simulated data set X of N_p images with a set $\mathcal{R}$ of three square Regions of Interest (ROIs). We note b the background (i.e. outside the ROIs). The signal in the (i, j) voxel of the k^{th} image is simulated as:

$$X_{i,j,k} = \sum_{r \in \mathcal{R}} \mathbb{I}_r(i,j)\alpha_{r,k}u_{i,j,k} + \mathbb{I}_b(i,j)u_{i,j,k} + \epsilon_{i,j,k} \quad (4)$$

where $u_{i,j,k}$ is a random value from an uniformed distribution in $[0, 1]$, $\epsilon_{i,j,k}$ a random value from a Gaussian distribution $\mathcal{N}(0, 1)$ smoothed with a parameter of 2 voxels to mimic the correlation structure observed in real fMRI datasets, $\alpha_{r,k} \sim \mathcal{U}[0, 1]$ for ROI r and image k. We have $\mathbb{I}_r(i, j) = 1$ (resp. $\mathbb{I}_b$) if the (i, j) voxel is in r (resp. b), and $\mathbb{I}_r(i, j) = 0$ (resp. $\mathbb{I}_b$) elsewhere. We simulate the target Y as: $Y_k = \sum_{r \in \mathcal{R}} \alpha_{r,k}$

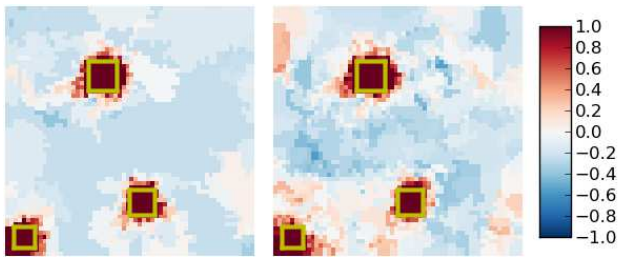
We generate the tree and derive the optimal parcellation using a learning dataset of 40 images, then we validate on 60 other images simulated according to (4). The images have a size of 60×60, with three non-overlapping ROIs of width $5, 6, 7$ pixels. We test the *supervised cut* algorithm with a number of exploration steps set to $\Delta = 60$, and the elastic net parameters $s = 0.2$ and $\lambda = 0.5$.
In the simulation, the leave-one-out cross-validation for the selection step has been replaced by a 5-folds validation.

3.2. Results on simulated data

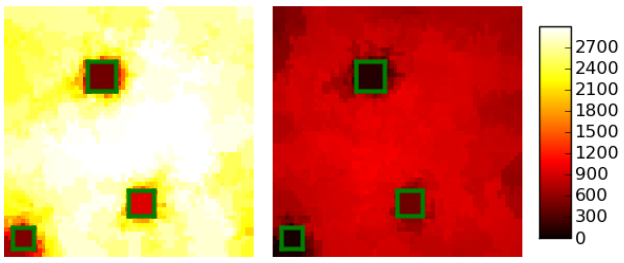
We compare the different methods on twenty sets of simulated data. See the results in Fig.3: both parcel-based approaches are able to extract the simulated discriminative regions (a), but the supervised cut approach has the additional ability to leave very wide regions of the background virtually unsegmented, so that the parcels created by the cut are much larger in the noisy background than when using the unsupervised approach (b). As a consequence, the *supervised cut* approach generalizes better, though both approaches clearly outperform voxel-based elastic net predictions. Only the supervised approach outperforms *SVR*.

3.3. Real data

We use a part of a real dataset on the mental processing of quantities. During the experiment, ten healthy volunteers (6 males and 4 females, mean age 21.2 +/- 3.0 years)



(a)



(b)

Subjects	SVR	Enet	Super.	Unsuper.
Mean ζ	0.14	0.07	0.17	0.07
Std ζ	0.1	0.19	0.14	0.08

(c)

Figure 3. Results of the simulation experiment. (a) Standardized Elastic net coefficients for the two parcellation techniques: the *supervised cut* (left) and the *unsupervised cut* (right) methods. These results are averaged across 20 simulations and show that both methods recover the simulated active regions (outlined by yellow squares), although the *supervised cut* method yields a spatially more specific pattern. (b) Average size of the parcels that include each voxel, for the *supervised* (left) and *unsupervised cut* (right) methods: the supervised cut creates larger parcels than the unsupervised cut far from the informative spots (outlined by green squares), and smaller parcels in the informative regions. (c) Ratio and Standard deviation of the explained variance for different methods averaged on 20 simulations. Parcel-based techniques outperform voxel-based analyses that use elastic net, and the *supervised cut* algorithm performs best.

viewed dot patterns with different quantities of dots ($\nu = 2$, 4, 6 and 8; we take $Y = \log(\nu)$) with 4 repetitions of each stimulus in each one of 8 sessions : so that we have a total of $N_p = 32$ images per subject. We aimed at predicting the values of Y from the fMRI data through regression.
Functional images were acquired on a 3 Tesla MR system with 12-channel head coil (Siemens Trio TIM) as T2* weighted echo-planar image (EPI) volumes using a high-resolution EPI-sequence. 26 oblique-transverse slices covering parietal and superior parts of frontal lobes were obtained in interleaved acquisition order with a TR of 2.5 s (FOV 192 mm, fat suppression, TE 30 ms, flip angle 78° , $1.5 \times 1.5 \times 1.5$ mm voxels). Standard pre-processings and the fit of the general linear model were performed with the SPM5 software. We used images of parameter estimates, one per condition and repetition.

We use the parameters of elastic net $s = 0.2, \lambda = 0.5$, $\Delta = 100$ exploration steps, for the supervised clustering. We have performed two series of analyses:

- In a first analysis, we launch our algorithm and the reference methods in each subject's dataset in parallel, on the whole parietal lobe, using one-repetition-out cross-validation (8 repetitions by subjects), and compute the average of the method performance (ζ) in this sample of 10 subjects.

- In a second analysis, we run the procedure in a multi-subjects analysis. For each subject, we first compute a fixed-effects activation image that represents the average effect of each stimulus, one for each condition (then, we have 4 images by subjects in 10 subjects). We evaluate the performance of the method by cross-validation (leave-one-subject-out), which yields an average rate of explained variance across subjects. This analysis in launched on the whole brain volume.

3.4. Results on real data

In the intra-subject analysis, we obtain the results given in Tab.1. The parcel-based methods yield the same prediction accuracy as the voxel-based methods, despite the fact that they use fewer features. Thus, the parcels seem to be a good way to compress the information within the whole brain, without loss of performance. Both parcel-based methods yield the same results which may be due to the fact that the information is already well segregated within the tree of parcels, so that a supervised exploration does not improve data representation.
The results of the multi-subjects analysis are given in Tab.2. The fact that a significant proportion of the stimulus variance can be fit using brain activation across subjects means that the spatial layout of the information is relatively stable across subjects. However, this results is probably related to the fact that for small numbers of dots as used here (but not for larger numerosities or symbolic numbers [3]) parametric activity increases can be observed in relatively extended and contiguous parietal regions, see also [11]. Whether these reflect special mechanisms for processing small numbers of objects, or secondary factors not related to numerical representation per se (e.g., increased effort when attempting to count), is currently not clear. The *supervised cut* method outperforms the other approaches. In particular, the explained variance is 19% higher than with the *SVR* method ($p < 0.004$), and 12% higher than with elastic net ($p < 0.04$). Moreover, the parcel-based methods allow us to access interpretable maps, as shown in Fig.4(b), compared to voxel-based methods (Fig.4 (a)).

	SVR	Enet	Super.	Unsuper.
Mean ζ	0.47	0.46	0.46	0.47
Std ζ	0.22	0.25	0.25	0.25
Nb. of features	242.5	158.7	71.0	70.6

Table 1. Results obtained in the intra-subject analysis. Average ratio and corresponding standard deviation of the explained variance, and average number of features (voxels or parcels) across 10 subjects. We can see that all the methods perform equally well, although they use a different number of features. The parcel-based algorithms use far less features than the voxel-based ones, i.e. they create a more compact representation of the data.

	SVR	Enet	Super.	Unsuper.
Mean ζ	0.42	0.49	0.61	0.52
Std ζ	0.13	0.23	0.2	0.28

Table 2. Results on real data, in a multi subjects analysis: average and the standard deviation of ζ for the different methods. The *supervised cut* algorithm yields the best performance in leave-one-subject-out cross-validation, and is significantly better than the two voxel-based methods (SVR and elastic net).

4. Discussion

Given that an fMRI brain image typically comprises 10^4 to 10^5 voxels, it is perfectly reasonable to use intermediate structures such as parcels, to reduce the information in prediction experiments. Our simulations show that our procedure for parcel definition allows the detection of the most informative regions for the prediction task. Moreover, in the case of a multi-subjects study, parcellations are expected to compensate for spatial misalignment between individual datasets, hence can better generalize than voxel-based methods. The present study confirms that this indeed increases the generalization capability of the trained classifier or regression estimator. Note that it is important to define the parcellation on the training database only to avoid data overfit. This entails the technical difficulty of optimizing the parcellation with respect to the spatial organization of the information within the image. To break the combinatorial complexity of the problem, we have defined a recursive parcellation of the volume using Ward's algorithm, which is furthermore constrained to yield spatially connected clusters. The merit of Ward's clustering is to yield minimal variance parcels at each step, so that it makes sense indeed to use parcel-based signal averages. The sets of possible volume parcellations is then reduced to a tree, so that the problem boils down to finding the optimal cut of the tree.

To define such a cut, we can either use Ward's inertia criteria, which means that the tree is cut *horizontally*, into subtrees with a comparable amount of variance. Model selection then boils down to finding at which level the tree should be cut. The method is relatively powerful, but clearly

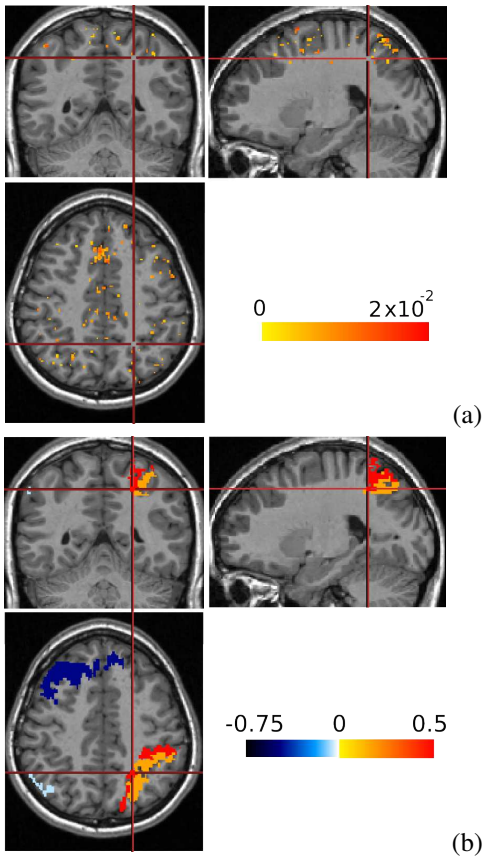


Figure 4. Results obtained with real data in the inter-subject experiment. The functional information is superimposed on the anatomical image of one particular subject: (a) Sum of the absolute values of the weights for the voxels used in the SVR (an optimal number of 2000 voxels have been found by cross-validation). The voxels are spread all over the brain, without any emerging coherence. (b) Coefficients of elastic net for the parcels found when using the *supervised cut* algorithm. We can see that these parcels are embedded along the intra-parietal sulcus, which was expected, see e.g. [2]. Starting from a whole-brain analysis, very few parcels have a non-null weights.

suboptimal with respect to the prediction task. By contrast, the supervised cut approach attempts to optimize the cut with respect to the prediction task. Although finding an optimal solution is infeasible, we adopt a greedy strategy that recursively finds the splits that most improve the prediction score. An important characteristic is that this is a multivariate approach, which always takes into account the joint distribution of the available features. However, there is still no guarantee that the optimal cut might be reached with this strategy. Model selection is then performed a posteriori by considering the best-generalizing parcellations among the available models. We have shown on simulations and real data that this approach has the particular capability to highlight regions of interest, while leaving uninformative regions unsegmented. In that sense it can be viewed as a multi-scale approach. The benefits of parcellation come at a cost regarding CPU time, the parcel definition raising CPU time to 15 minutes on real datasets (with a non optimized python implementation though). Nevertheless, all this remains perfectly affordable for standard neuroimaging data analyses, especially by using fast implementation of elastic net, such as coordinate descent, which yields an average time for the whole analysis (exploration of the tree, selection of the best sub-tree) of 20s on a 1.6 Ghz CPU.

The proposed methods yield the same results as the reference method *SVR* in the intra-subject study, but they yield better results for the inter-subjects study. Our interpretation is that in the intra-subject case there is a straightforward voxel-to-voxel correspondence across the images, so that *SVR* works optimally. However in the inter-subjects study, voxel-based methods are weakened by the inter-subject spatial variability and their performances are relatively lower; parcel-based models compensate for that effect. Additionally, as our parcellation approach works in the feature space, it can easily incorporate more priors such as anatomical boundaries between brain structures. Our parcellation scheme is further useful to accurately locate contiguous predictive regions, especially in the supervised version, as shown in the comparison with voxel-based methods.

Conclusion In this paper we proposed a new feature building method for extracting information from brain images. This includes the construction of an adapted spatial model that captures the predictive information present in the data better than general feature selection heuristics. A particularly important property of this approach is its ability to focus on relatively small but informative regions while leaving vast but noisy areas unsegmented. This algorithm performs well on real data, and especially in the multi-subjects analysis. Indeed, the spatial averaging of the signal induced by the parcellation seems to be a powerful way to deal with the inter-subject variability. Moreover, this method is not restricted to brain images, and might be used in any dataset where multi-scale structure is considered as important (e.g. medical or satellite images).

References

[1] D. D. Cox and R. L. Savoy. Functional magnetic resonance imaging (fMRI) "brain reading": detecting and classifying distributed patterns of fMRI activity in human visual cortex. *NeuroImage*, 19(2):261–270, 2003. 1, 5

[2] S. Dehaene, M. Piazza, P. Pinel, and L. Cohen. Three parietal circuits for number processing. *Cognitive Neuropsychology*, (20):487–506, 2003. 7

[3] E. Eger, V. Michel, B. Thirion, A. Amadon, S. Dehaene, and A. Kleinschmidt. Deciphering cortical number coding from human brain activity patterns. 19(19):1608–1615, 2009. 6

[4] S. Ghebreab, A. Smeulders, and P. Adriaans. Predicting brain states from fMRI data: Incremental functional prin-

cipal component regression. In *Advances in Neural Information Processing Systems 20*. MIT Press, 2008. 2

[5] P. Golland, Y. Golland, and R. Malach. Detection of spatial activation patterns as unsupervised segmentation of fMRI data. pages 110–118. Med Image Comput Comput Assist Interv. MICCAI 2007, 2007. 2

[6] J.-D. Haynes and G. Rees. Decoding mental states from brain activity in humans. *Nature Reviews Neuroscience*, 7:523–534, 2006. 2

[7] Y. Kamitani and F. Tong. Decoding the visual and subjective contents of the human brain. *Nature Neuroscience*, 8(5):679–685, April 2005. 1

[8] D. Kontos, V. Megalooikonomou, D. Pokrajac, A. Lazarevic, Z. Obradovic, O. B. Boyko, J. Ford, F. Makedon, and A. J. Saykin. Extraction of discriminative functional MRI activation patterns and an application to alzheimer's disease. pages 727–735. Med Image Comput Comput Assist Interv. MICCAI 2004, 2004. 2

[9] N. Kriegeskorte, R. Goebel, and P. Bandettini. Information-based functional brain mapping. *Proceedings of the National Academy of Sciences of the United States of America*, 103:3863 – 3868, 2006. 2

[10] K. A. Norman, S. M. Polyn, G. J. Detre, and J. V. Haxby. Beyond mind-reading: multi-voxel pattern analysis of fMRI data. *Trends Cogn Sci*, 10(9):424–430, 2006. 2

[11] S. Santens, C. Roggeman, W. Fias, and T. Verguts. Number processing pathways in human parietal cortex. *Cereb. Cortex*, 20:77–88, 2010. 6

[12] J. H. Ward. Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58(301):236–244, 1963. 3

[13] H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society B*, 67:301–320, 2005. 3, 4