Frontiers in Neurosciences 7June 2013

This is Title

Alexandre Abraham [1,2,*], **Fabian Pedregosa** [1,2], **Andreas Muller, Jean Kossaifi, Philippe Gervais** [1,2], **Alexandre Gramfort, Bertrand Thirion** [1,2] **and Gal Varoquaux** [1,2]

[1] *Parietal Team, INRIA Saclay-Île-de-France, Saclay, France*
[2] *Neurospin, I²BM, DSV, CEA, 91191 Gif-Sur-Yvette, France*

Correspondence*:
Alexandre Abraham
Parietal Team, INRIA Saclay-Île-de-France, Saclay, France,
alexandre.abraham@inria.fr

Research Topic

ABSTRACT

Statistical learning methods are increasingly used to perform neuroimaging analysis. Their main virtue for this type of application is their ability to model high-dimensional datasets, e.g. multivariate analysis of activation images, or capturing intersubject variability. Supervised learning is typically used in decoding setting to relate brain images to behavioral or clinical observations, while unsupervised learning is typically used to uncover hidden structure in sets of images (e.g. resting state functional MRI) or to find subpopulations in large cohorts of subjects. By considering functional neuroimaging use cases, we illustrate how the Scikitlearn, a Python machine learning library, can be used to perform some key analysis steps. Scikitlearn contains a large set of statistical learning algorithms, both supervised and unsupervised, that can be applied to neuroimaging data after a proper preprocessing. Combined with other Python libraries, neuroimaging data can be loaded, processed and the results can be visualised easily.

Keywords: Text Text Text Text Text Text Text Text

1 INTRODUCTION

1.1 PYTHON SCIENTIFIC AND NEUROIMAGING ECOSYSTEM

1.1.1 Scipy and Numpy

1.1.2 nibabel

1.1.3 nipy

1.1.4 scikit-learn

2 SCIKIT-LEARN CONCEPTS

2.1 ESTIMATOR

2.2 DATA REPRESENTATION

Explain that the scikit process 2D data. This is an introduction to masking.

2.3 TRANSFORMER

2.4 PIPELINE

Make a reference to SVM example

2.5 CROSS VALIDATION

It seems more right to me to put it in this part

3 DATASETS

I think this should be good to talk about datasets from the beginning to introduce what we will want (introducing examples) and to be able to cite them from the beginning (we may want to precise what kind of preprocessing is requirend on which dataset and why).

3.1 HAXBY

3.2 KAMITANI

BOLD response is usually 4/6 seconds after the stimulus. For Kamitani, we shift the stimuli by 2 TR.

3.3 NYU TEST-RETEST

4 PREPARING THE DATA

4.1 DATA PREPARATION

4.2 RESAMPLING

- Necessary for multi subjects (do we speak of this problem before ?)
- This is one way of decreasing data size
- Removing confounds is necessary for some treatments

4.3 SIGNAL CLEANING

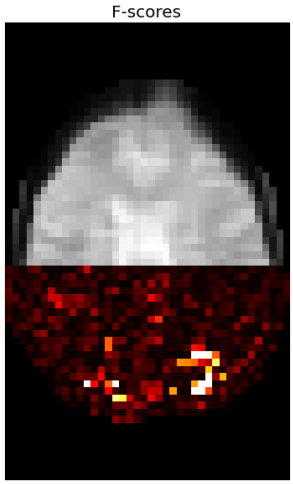
- Remove high frequency (scanner artefacts)

4.4 DIMENSION REDUCTION

- Data is often too big for computation, we need to reduce its dimensionality

4.4.1 Resampling Resampling is a way to reduce dimensionality.

4.4.2 Feature selection Speak of Anova here. This is one of the simplest way to reduce efficiently dimensionality.



4.4.3 Clustering / ROI We can select regions to reduce dimensionality. For example, V1 for a visual task. We can also segment automatically the brain thanks to a Ward, or use a reference atlas.

4.4.4 PCA The PCA is good to reduce dimensionality in the time series dimension (other methods are for spatial reduction).

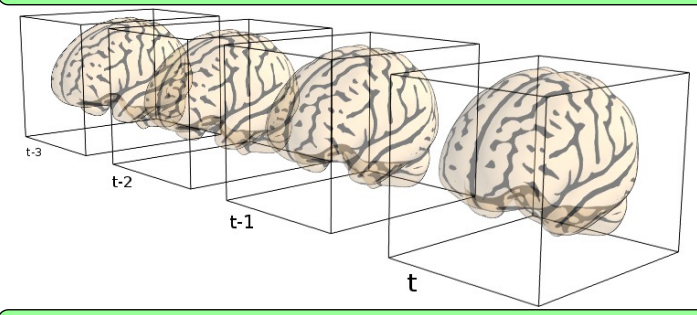
4.5 MASKING

4.5.1 From 4-dimensional image to 2-dimensional array Neuroimaging data are represented in 4 dimensions: 3 dimensions for the scans, which are positioned in a coordinate space, and one dimension for the time. Scikit-learn algorithms, on the other hand, only accept 2-dimensional data: one dimension for the features and one for the samples.

Consequently, in order to use neuroimaging data in the scikit-learn, a conversion is needed. The most simple way to achieve that would be to *flatten* the 3D scans into a 1D array. However, we know that not every voxels in a neuroimaging scan is useful. In particular, outter-brain voxels are of no use and, worse, they can bring spurious noise and scanner artefacts (such as ghosts).

To sort out voxels of interest, we will have to apply a mask on the data. Most of public datasets provide a mask, come of them even provide several, isolating different functional or anatomical brain regions.

ref to Haxby



Should tell here that some algorithms, like logistic regression, do not like colinear features.

4.5.2 Automatically computing a mask The simplest strategy to compute a mask is a binarization by a selected threshold. Due to the nature of the neuroimaging data, there exists some strategy to choose this threshold in order to obtain a decent segmentation.

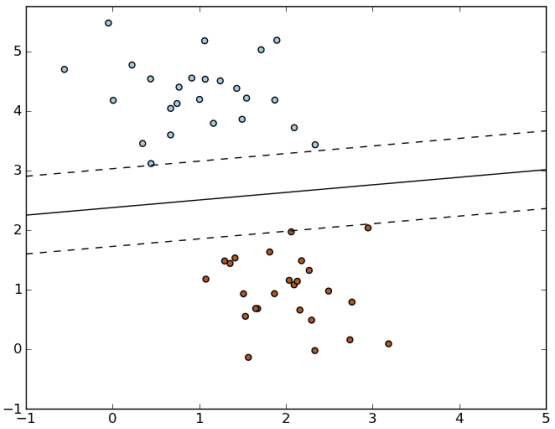


Figure 1. Example of SVC on toy problem

There is a reference for the method used in Nisl. We should put it there and in the code. Add a figure with an histogram to illustrate.

Multi subject computation is simply done by intersecting subjects maps relatively to a chose threshold.

4.5.3 Conserving geometrical structure Applying a mask on the data obviously remove the 3-dimensional structure of the data. However, some algorithms, like the Ward, need this structural information to run.

- Speak about connectivity graphs / adjacency matrices

5 DECODING

5.1 SVM

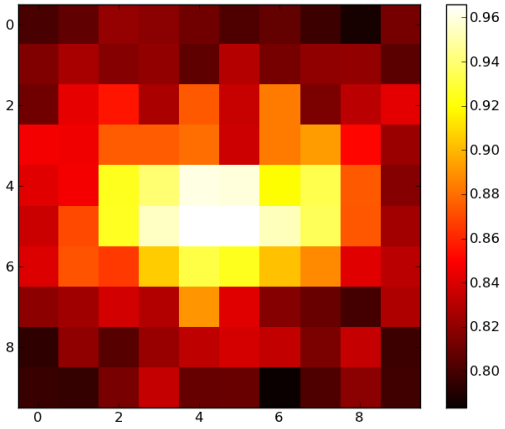
- Precise that we use ANOVA and give an example of pipeline

5.2 SEARCHLIGHT

- Present the Searchlight problem
- Say it is less a pain to implement thanks to scikit-learn bricks (estimator and cross_val). Plus it is easily customizable.

5.3 CLASSIFICATION OF M/EEG SENSOR SPACE DATA

5.4 ORTHOGONAL MATCHING PURSUIT



6 ENCODING

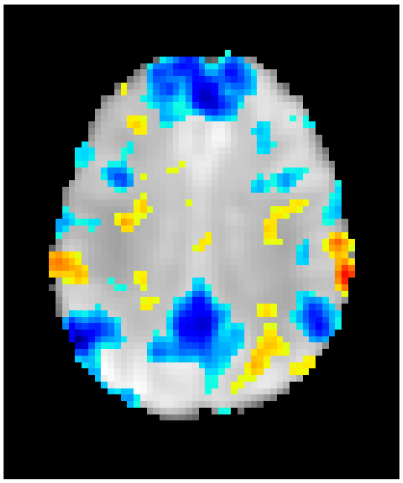
After talking with Michael, he told me that he could make a fairly simple example for encoding, which I think is a plus for the paper. The example will be integrated in Nisl.

7 FUNCTIONAL CONNECTIVITY

Should we speak of correlation matrices to represent interaction between regions?

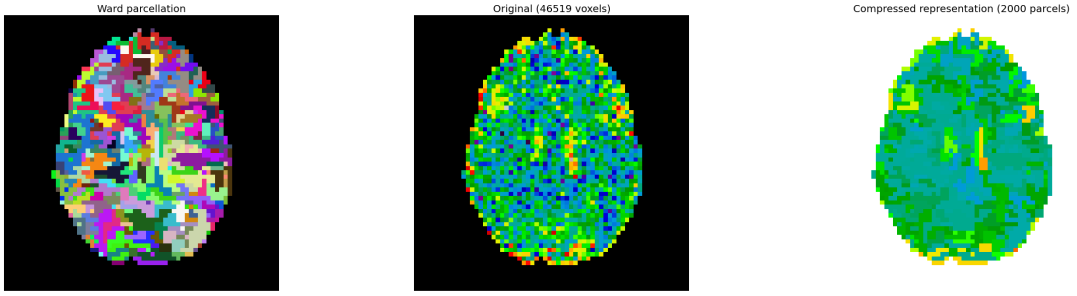
7.1 ICA

- Explain principle
- Put several refs
- Put maps
- Should we talk about CANICA ?



7.2 CLUSTERING

Make an example with Ward Clustering. Indicate then that other algorithms can be used such as KMeans and Spectral clustering and only give results.



8 DATA SHARING

Frontiers supports the policy of data sharing, and authors are advised to make freely available any materials and information described in their article, and any data relevant to the article (while not compromising confidentiality in the context of human-subject research) that may be reasonably requested by others for the purpose of academic and non-commercial research. In regards to deposition of data and data sharing through databases, Frontiers urges authors to comply with the current best practices within their discipline.

DISCLOSURE/CONFLICT-OF-INTEREST STATEMENT

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

ACKNOWLEDGEMENT

Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text.

Funding: Text Text Text Text Text Text Text Text.

SUPPLEMENTAL DATA

Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text.

REFERENCES