Machine Learning for Neuroimaging with Scikit-Learn

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

[1] *Parietal Team, INRIA Saclay-Île-de-France, Saclay, France*
[2] *Neurospin, I^2BM, 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 inter-subject 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 sub-populations in large cohorts of subjects. By considering functional neuroimaging use cases, we illustrate how the Scikit-learn, a Python machine learning library, can be used to perform some key analysis steps. Scikit-learn 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: Machine learning, Statistical Learning, Neuroimaging, Scikit-learn, Python

1 INTRODUCTION

1.1 SCIENTIFIC PYTHON 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 CROSS VALIDATION

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

3 FROM MR VOLUMES TO A DATA MATRIX

As any domain specific data, MR volumes holds particular properties. Understanding them is crucial to be sure to make proper use of the data.

$$\begin{bmatrix} r_x & 0 & 0 & o_x \\ 0 & r_y & 0 & o_y \\ 0 & 0 & r_z & o_z \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \\ 1 \end{bmatrix}$$

3.1 DATA PREPARATION

At this point, we suppose that standard preprocessings have been applied to the data. They should be registrated on a common template (MNI for example). However, data is not yet ready to be processed by the scikit-learn. In fact, preprocessed data may have different shapes. Moreover, it is essential to get rid of some remaining scanner artefacts and individual trends.

3.1.1 Detrending Detrending is an essential step when dealing with fMRI data. It removes a best-fit linear trend (in the least square sense) over the time series of each voxel. It is obviously needed when you want to study the correlation between features.

This step is essential because voxel intensity may differ from one subject to another because of the scanner. It may even differ in a single subject scans because of a birdcage coil artefact (some brain regions appears brighter than others, see Fig.).

3.2 RESAMPLING

Resampling is necessary when your dataset is composed of several scan sessions taht may not have the same shape. It may also be used to downsample the data : this is the easiest way to reduce data-size and speed up computation.

- Removing confounds is necessary for some treatments

3.3 SIGNAL CLEANING

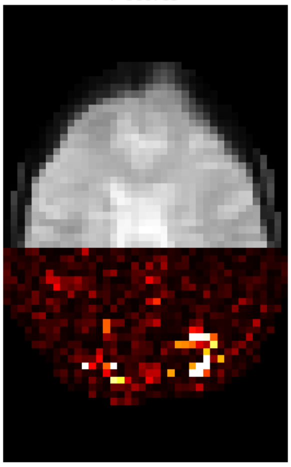
- Remove high frequency (scanner artefacts)

3.4 DIMENSION REDUCTION

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

3.4.1 Resampling Resampling is a way to reduce dimensionality.

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



3.4.3 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.

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

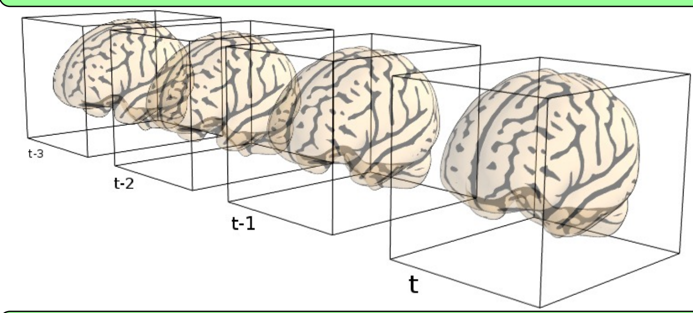
3.5 MASKING

3.5.1 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.

3.5.2 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.

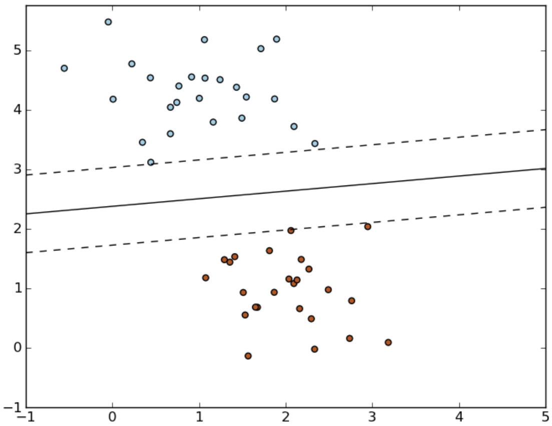


Figure 1. Example of SVC on toy problem

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.

3.5.3 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

4 DECODING

The process of predicting behavioral or comportamental data from fMRI scan is called decoding.

4.1 SVM

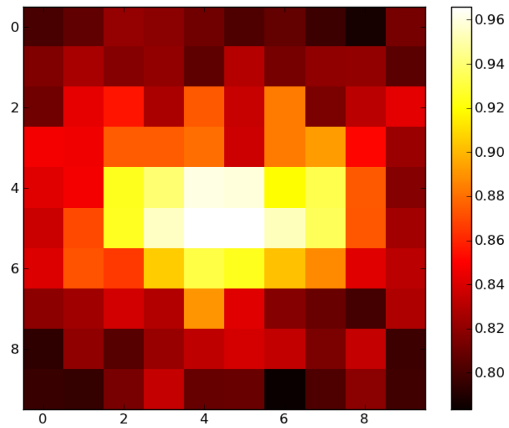
- Precise that we use ANOVA
- Introduce the pipeline
- Introduce Haxby dataset ?

4.2 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.

4.3 CLASSIFICATION OF M/EEG SENSOR SPACE DATA

4.4 ORTHOGONAL MATCHING PURSUIT



5 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.

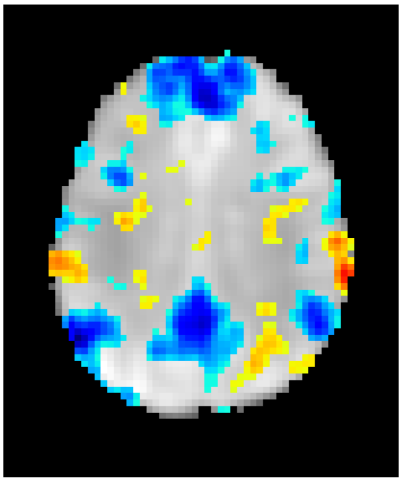
6 FUNCTIONAL CONNECTIVITY

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

6.1 INDEPENDENT COMPONENT ANALYSIS (ICA)

6.1.1 Intuition ICA is a blind source separation method. Its principle is to separate a multivariate signal into several components by maximizing their non-gaussianity. A typical example is the *cocktail party problem* where ICA separates the voices of people using signal from several mikes.

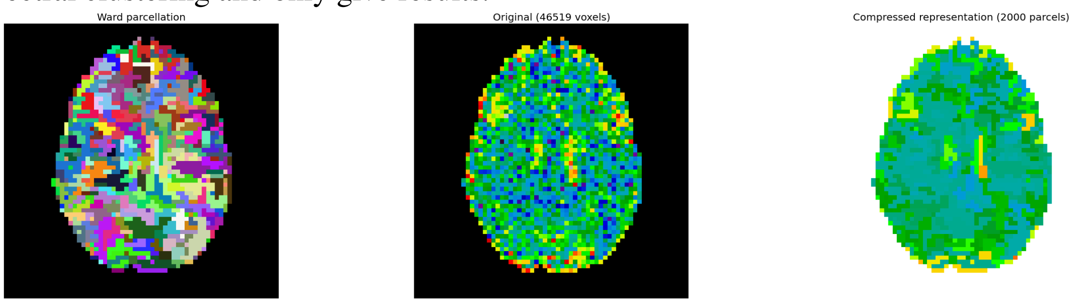
It is historically the reference method to extract networks from resting state fMRI Biswal and Ulmer (1999).



6.1.2 Application

6.2 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.



7 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.

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 Text.

REFERENCES

Biswal, B. and Ulmer, J. (1999) Blind source separation of multiple signal sources of fMRI data sets using independent component analysis. *Journal of computer assisted tomography* 23 265.