Random Forests based feature selection for decoding fMRI data

Robin Genuer, Vincent Michel, Evelyn Eger, Bertrand Thirion

Université Paris-Sud 11, INRIA Saclay-Ile-de-France

INSERM, CEA NeuroSpin

August 26th

COMPSTAT'2010, Paris

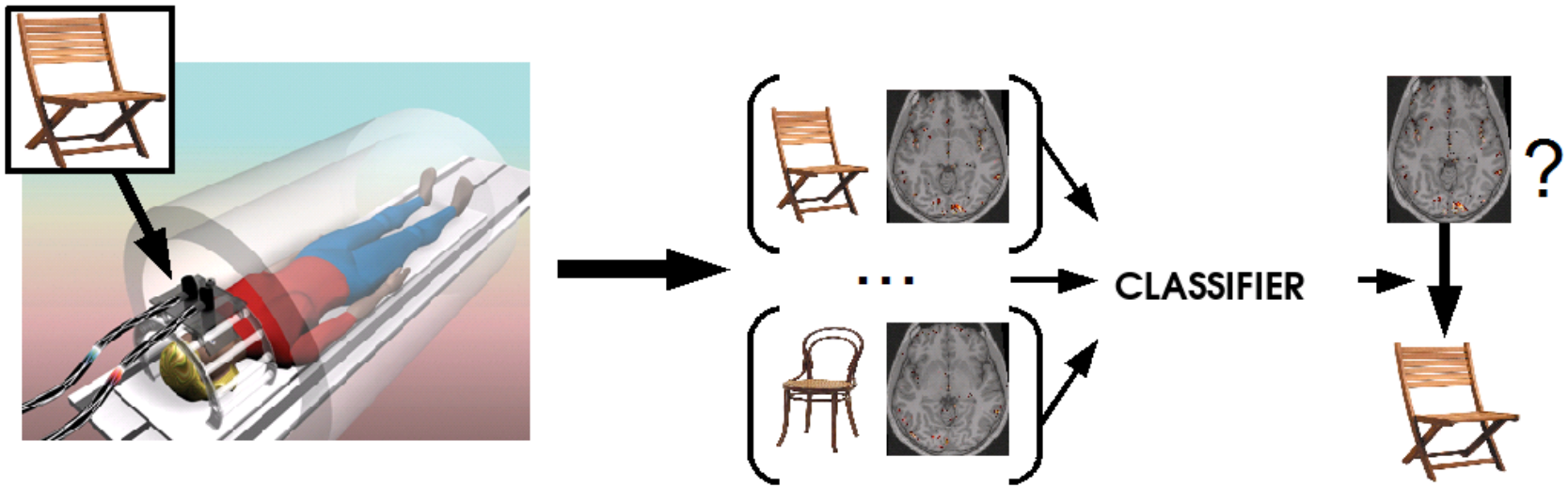


Figure: Experimental framework

4 kinds of chair (shapes)
100 000 voxels (variables)
72 observations

Random Forests

- introduced by L. Breiman in 2001
- ensemble methods, Dietterich (1999) and (2000)
- popular and very efficient algorithm of statistical learning, based on model aggregation ideas, for both classification and regression problems.

We consider a learning set $\mathcal{L}_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ made of n i.i.d. observations of a random vector (X, Y).

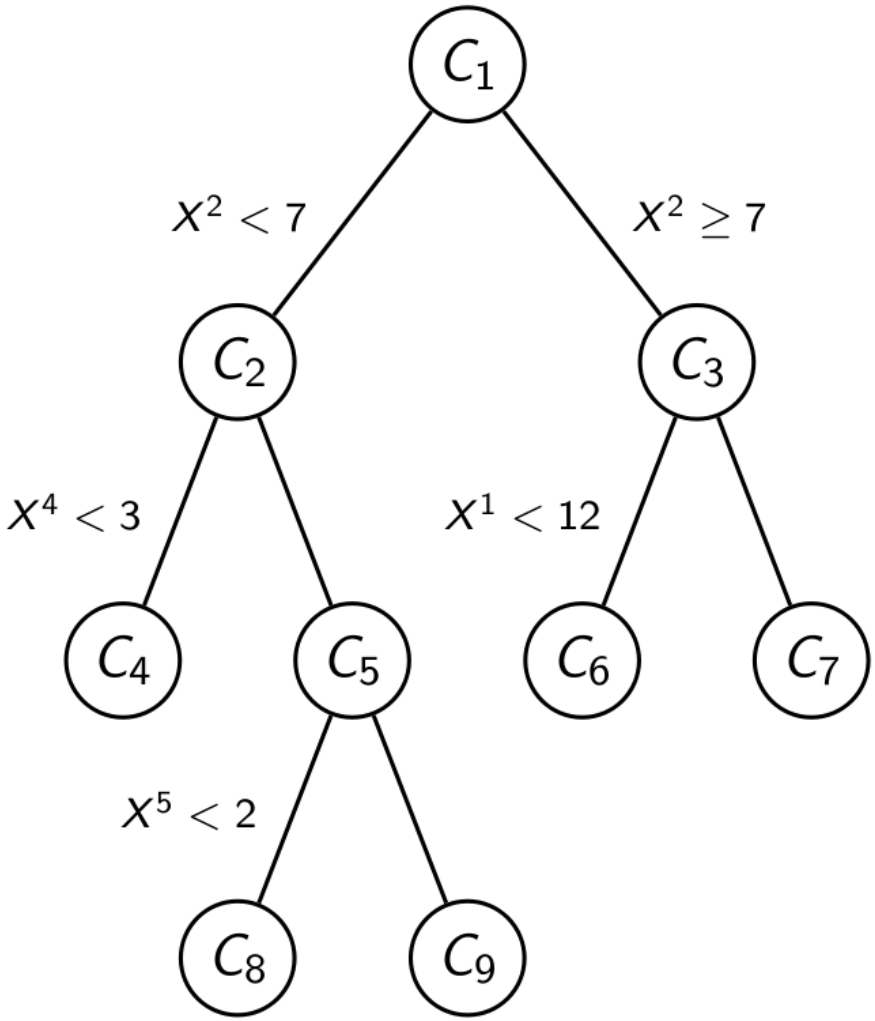
Vector $X = (X^1, ..., X^p)$ contains explanatory variables, say $X \in \mathbb{R}^p$, and $Y \in \mathcal{Y}$ is a class label.

A classifier h is a mapping $h : \mathbb{R}^p \to \mathcal{Y}$.

CART (Classification And Regression Trees, (Breiman et al. (1984)) can be viewed as the base rule of a random forest. Recall that CART design has two main stages:

- maximal tree construction to build the family of models
- pruning for model selection

With CART, we get a classifier, which is a piecewise constant function obtained by partitioning the predictor's space.



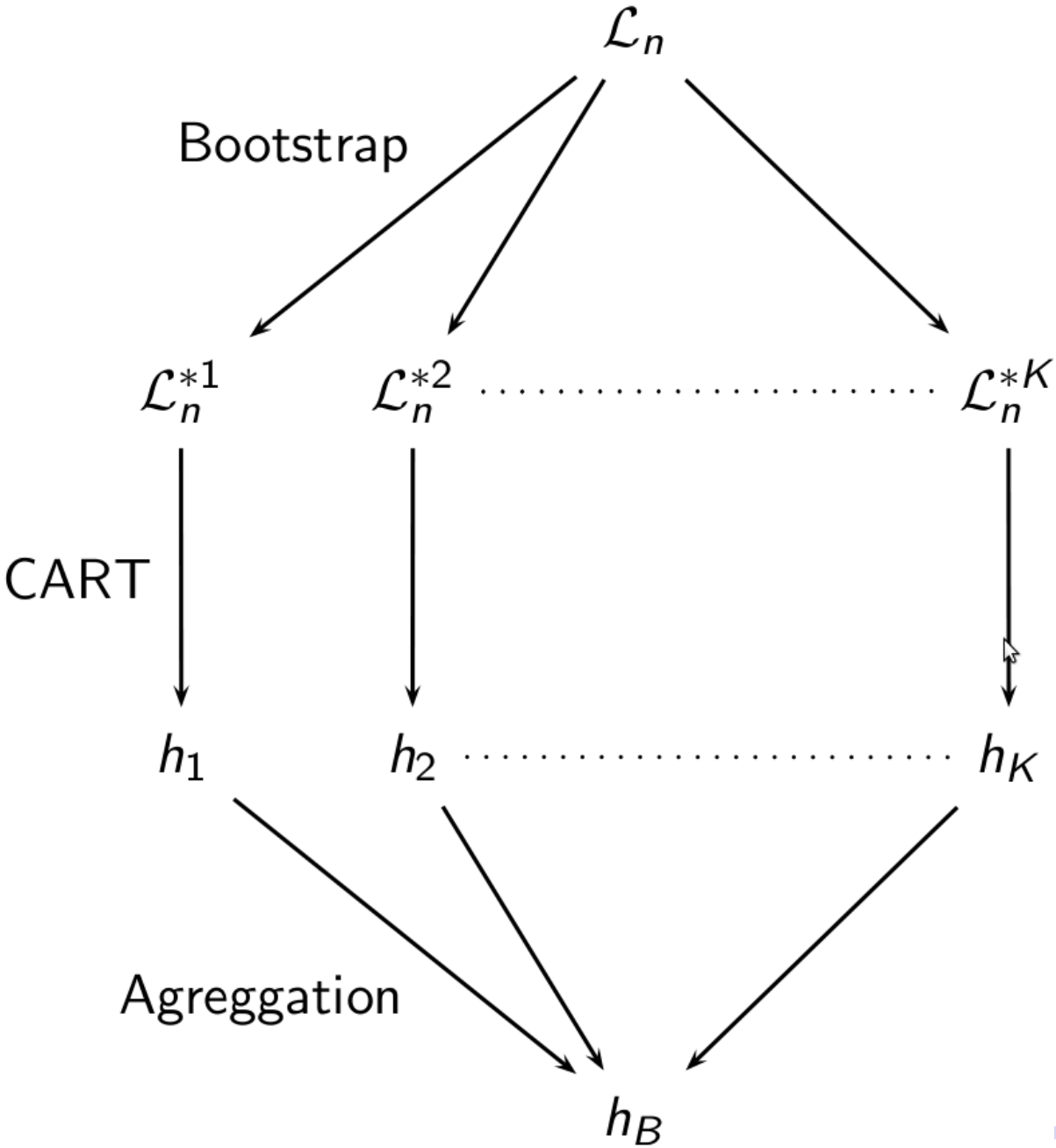
$$
\begin{aligned}
h(x) \quad = \quad & C_4 \mathbb{1}_{x^2<7, x^4<3} \\
& + C_8 \mathbb{1}_{x^2<7, x^4\geq 3, x^5<2} \\
& + C_9 \mathbb{1}_{x^2<7, x^4\geq 3, x^5\geq 2} \\
& + C_6 \mathbb{1}_{x^2\geq 7, x^1<12} \\
& + C_7 \mathbb{1}_{x^2\geq 7, x^1\geq 12}
\end{aligned}
$$

Growing step, stopping rule:

- do not split a pure node
- do not split a node containing less than `nodesize` data

Pruning step:

- the maximal tree overfits the data
- an optimal tree is pruned subtree of the maximal tree which realizes a good trade-off between the variance and the bias of the associated model


$\mathcal{L}_n$
Bootstrap
$\mathcal{L}_n^{*1}$
$\mathcal{L}_n^{*2}$
$\mathcal{L}_n^{*K}$
CART
h_1
h_2
h_K
Agreggation
h_B

CART-RF

We define CART-RF as the variant of CART consisting to select at random, at each node, `mtry` variables, and split using only the selected variables. The maximal tree obtained is not pruned.

`mtry` is the same for all nodes of all trees in the forest.

Random forest (Breiman 2001)

To obtain a random forest we proceed as in bagging. The difference is that we now use the CART-RF procedure on each bootstrap sample.

OOB = Out Of Bag.

OOB error

Consider a forest. For one data (X_i, Y_i), we only keep the classifiers h_k built on a bootstrap sample which does not contain (X_i, Y_i), and we aggregate these classifiers. We then compare the predicted label we get to the real one Y_i.

After doing that for each data (X_i, Y_i) of the learning set, the OOB error is the proportion of misclassified data .

R package:

- seminal contribution of Breiman and Cutler (early update in 2005)
- described in Liaw, Wiener (2002)

Focus on the `randomForest` procedure whose main parameters are:

- `ntree`, the number of trees in the forest (default value : 500)
- `mtry`, the number of variables randomly selected at each node (default value : $\sqrt{p}$)

Breiman (2001), Strobl *et al.* (2007) and (2008), Ishwaran (2007), Archer *et al.* (2008).

Variable importance

Let $j \in \{1, \ldots, p\}$. For each OOB sample we permute at random the j-th variable values of the data. The variable importance of the j-th variable is the mean increase of the error of a tree.

The more the increase is, the more important is the variable.

We distinguish two different objectives:

1. to magnify all the important variables, even with high redundancy, for interpretation purpose
2. to find a sufficient parsimonious set of important variables for prediction

Two earlier works must be cited:

- Díaz-Uriarte, Alvarez de Andrés (2006)
- Ben Ishak, Ghattas (2008)

Our aim is to build an automatic procedure, which fulfills these two objectives.

"Toys data", Weston *et al.* (2003)

an interesting equiprobable two-class problem, $Y \in \{-1, 1\}$, with 6 true variables, the others being noise:

- two near independent groups of 3 significant variables (highly, moderately and weakly correlated with response Y)
- an additional group of noise variables, uncorrelated with Y

Model defined through the conditional distributions of the X^i for $Y = y$:

- for 70% of data, $X^i \sim y\mathcal{N}(i, 1)$ for $i = 1, 2, 3$ and $X^i \sim y\mathcal{N}(0, 1)$ for $i = 4, 5, 6$
- for the 30% left, $X^i \sim y\mathcal{N}(0, 1)$ for $i = 1, 2, 3$ and $X^i \sim y\mathcal{N}(i - 3, 1)$ for $i = 4, 5, 6$
- the other variables are noise, $X^i \sim \mathcal{N}(0, 1)$ for $i = 7, \ldots, p$

Genuer, Poggi, Tuleau (2010)

1. Preliminary ranking and elimination:
   - Sort the variables in decreasing order of RF scores of importance
   - Cancel the variables of small importance. Let m be the number of remaining variables
2. Variable selection:
   - For *interpretation*:
     - Construct the nested collection of RF models involving the k first variables, for $k = 1$ to m
     - Select the variables involved in the model leading to the smallest OOB error
   - For *prediction* (conservative version):
     - Starting from the ordered variables retained for interpretation, construct an ascending sequence of RF models, by invoking and testing the variables stepwise
     - The variables of the last model are selected

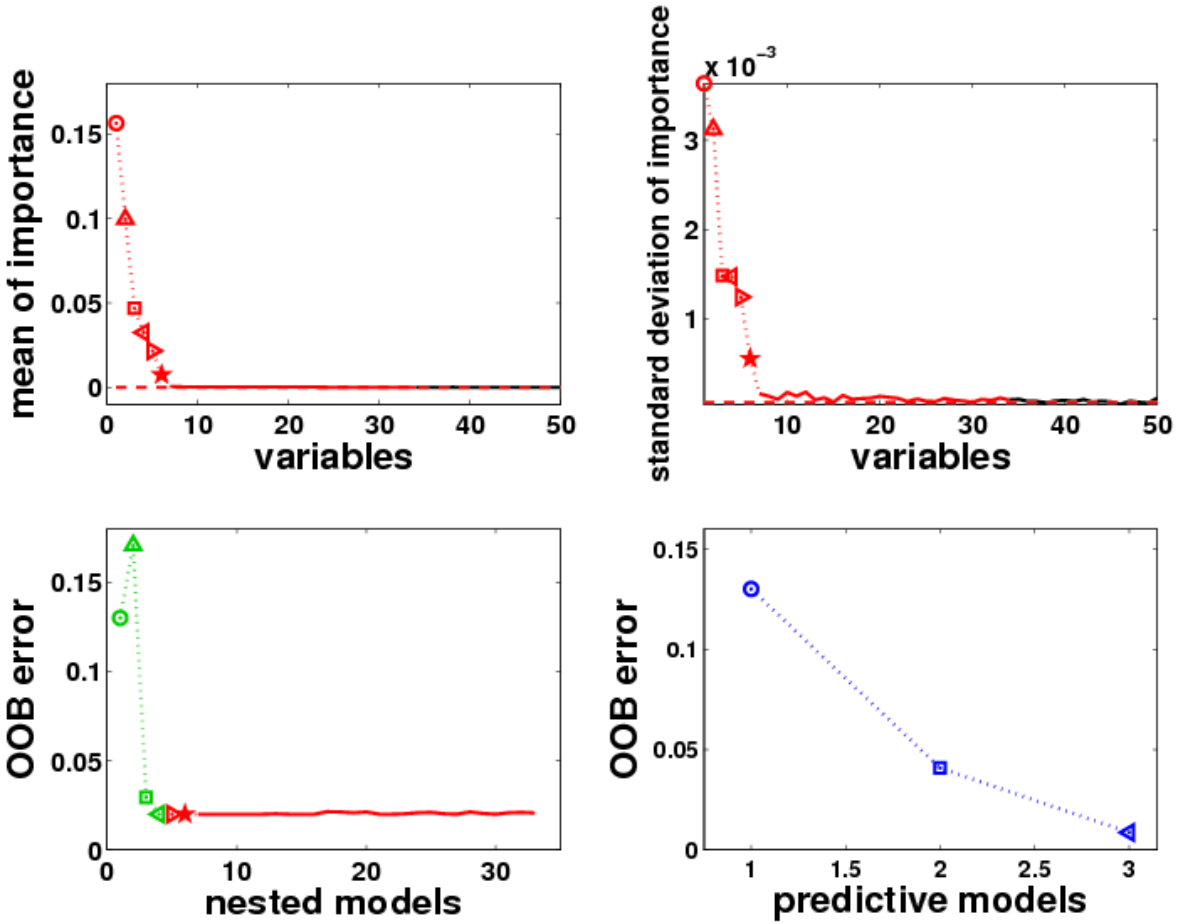


Figure: Variable selection procedure for interpretation and prediction: toys data $n = 100$, $p = 200$

- True variables (1 to 6) represented by $(\triangleright, \triangle, \circ, \star, \triangleleft, \square)$
- VI based on 50 forests with $ntree = 2000$, $mtry = 100$

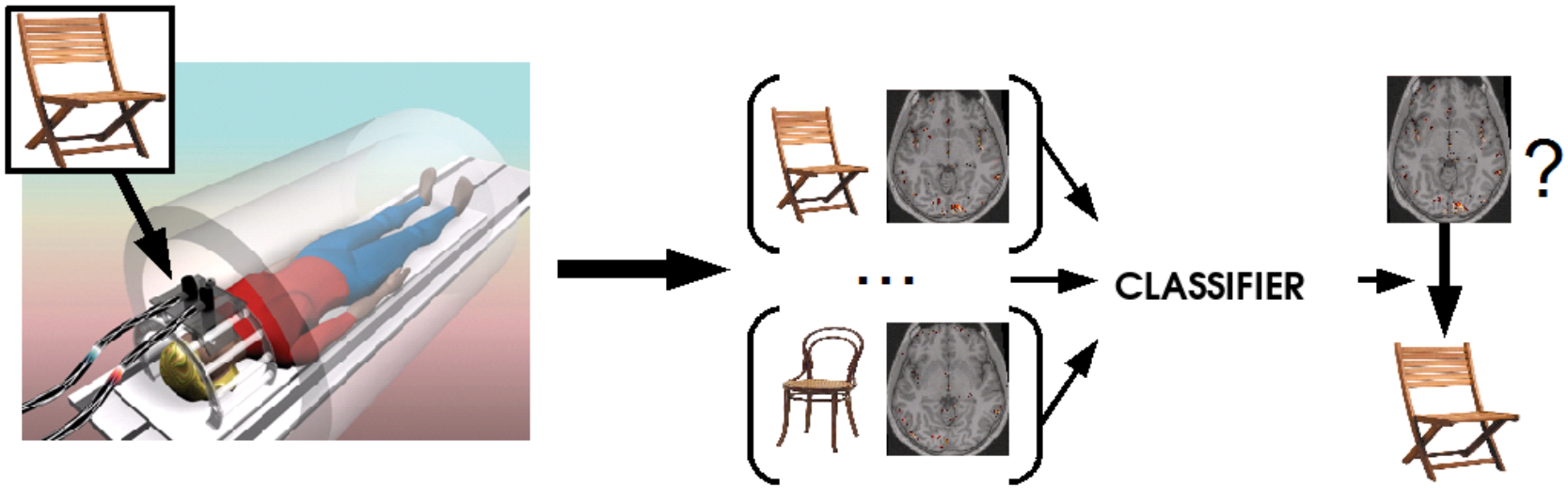


Figure: Experimental framework

4 kinds of chair $\Rightarrow$ 4 classes.

Whole brain: raw data are made of 100 000 voxels (variables) and 72 observations. A parcellation obtained by Ward algorithm reduces to 1000 parcels.

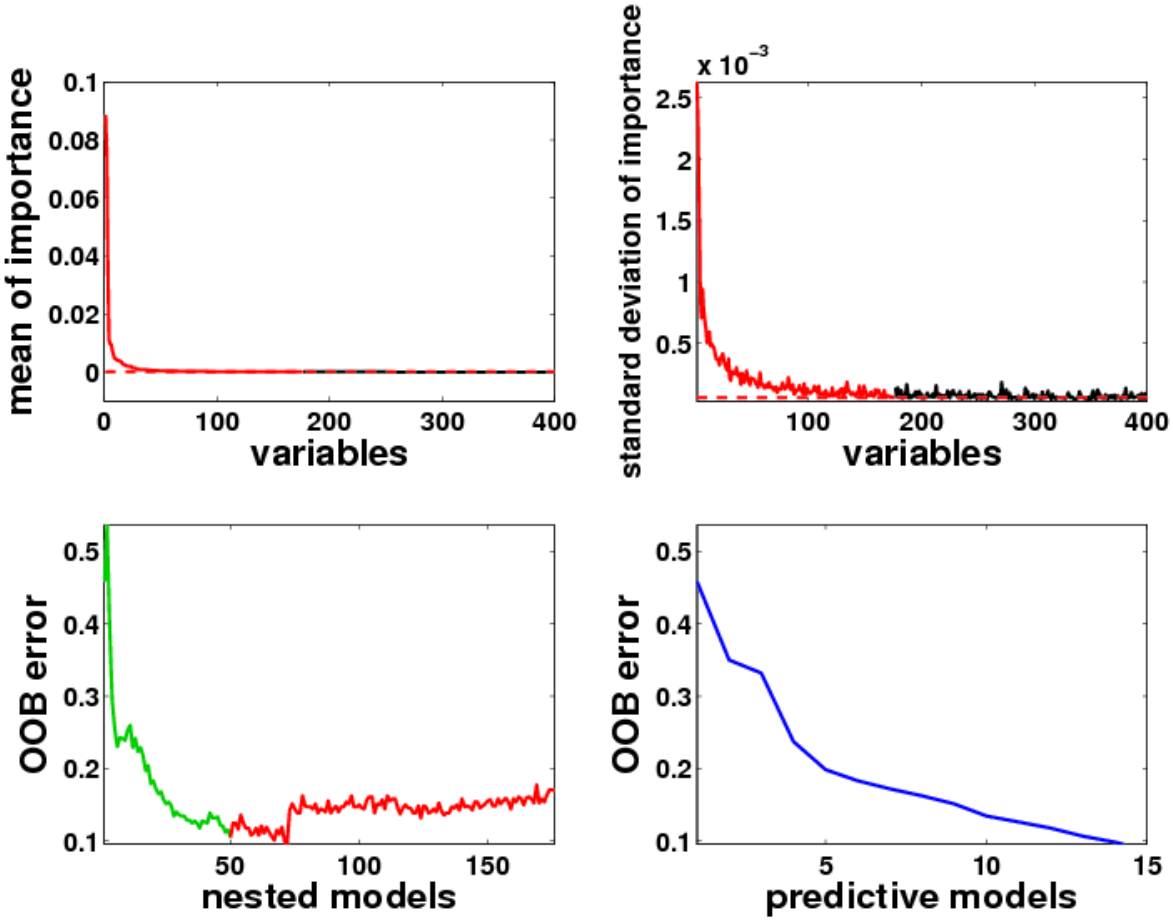


Figure: Variable selection procedures for a real subject, $ntree = 2000$, $mtry = p/3$

- Key point: it selects 176 variables after the threshold step, 50 variables for interpretation, and 15 variables for prediction (very much smaller than $p = 1000$)

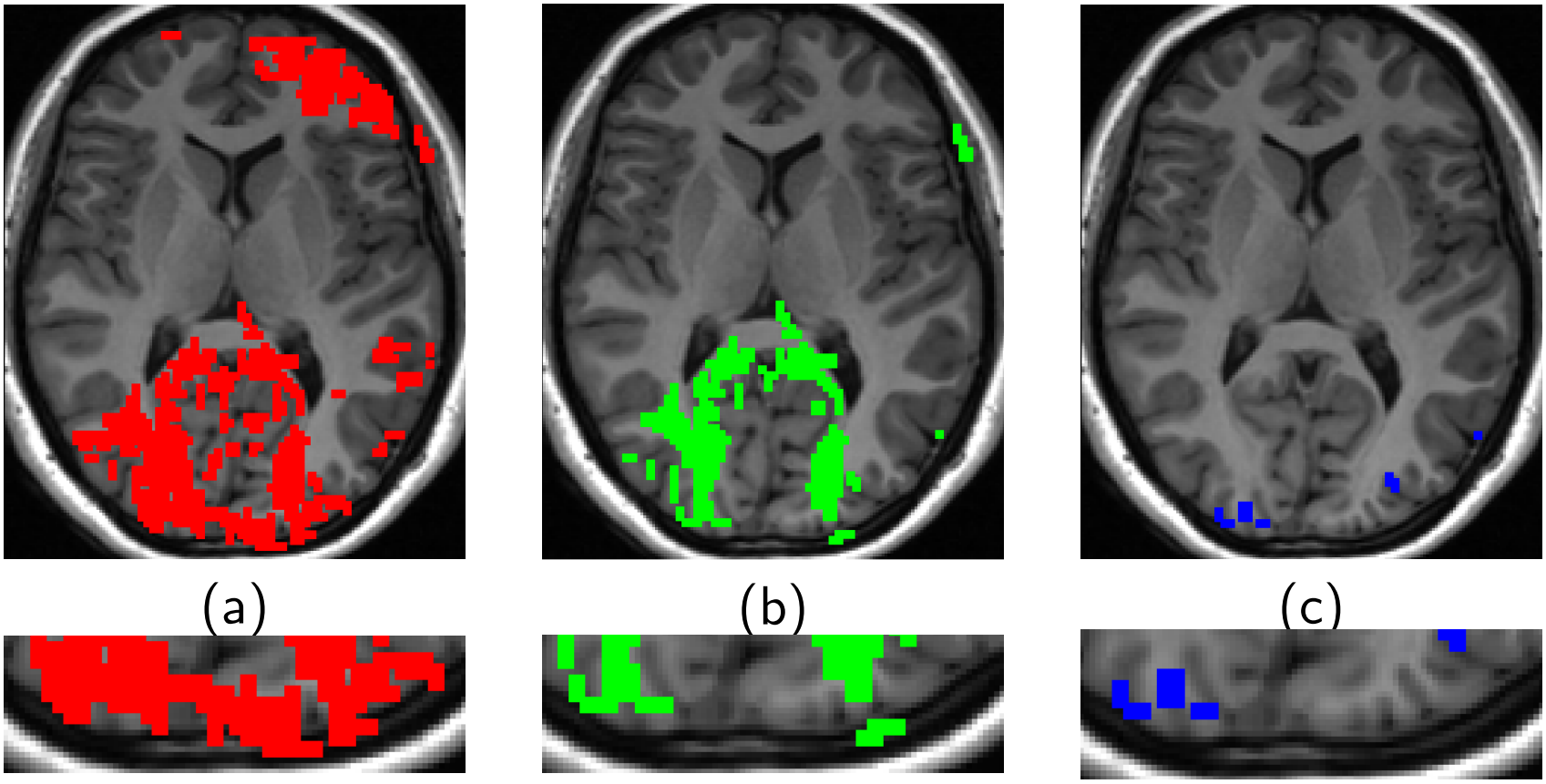
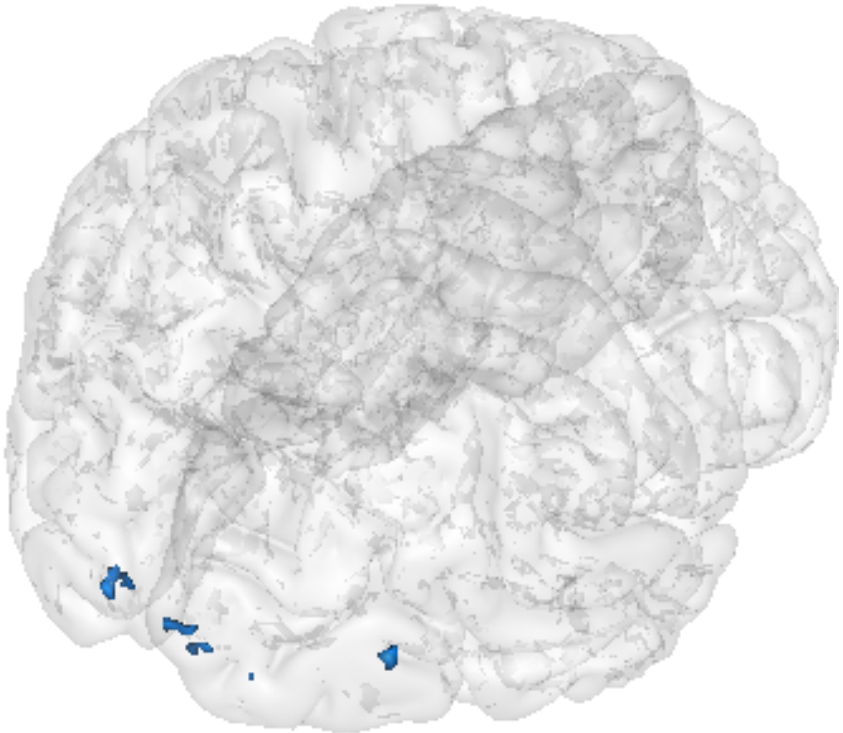


Figure: Example of the different steps of the framework on a real subject.

(a) Elimination Step (b) Interpretation Step (c) Prediction Step

Figure: Regions selected in at least 3 subjects among 12 by the last step of the RF-based selection.



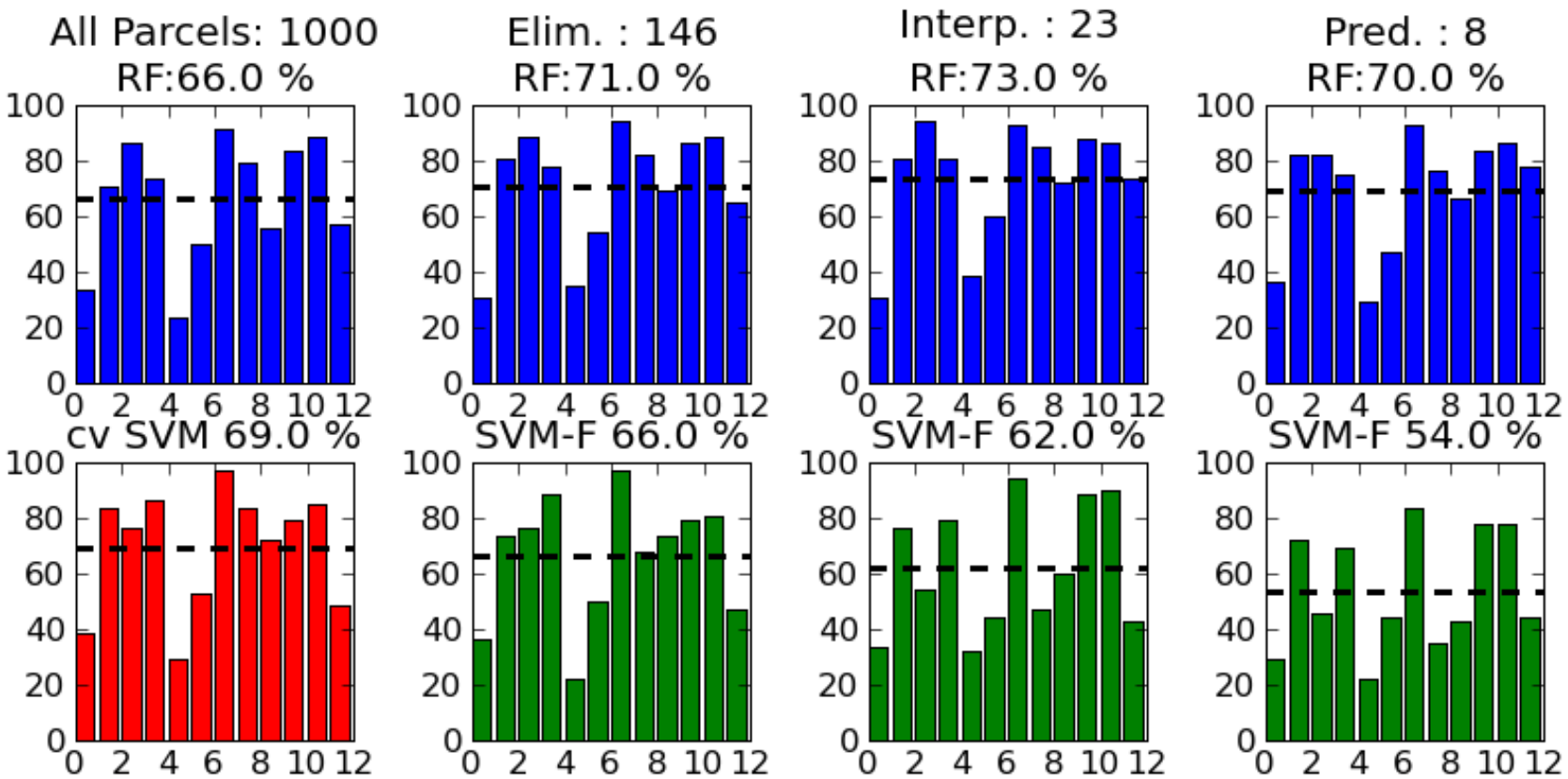


Figure: Classification rates (whole brain)

Short bibliography


Breiman, L. Random Forests. *Machine Learning* (2001)


Cox, D.D. and Savoy, R.L. Functional magnetic resonance imaging (fMRI) "brain reading": detecting and classifying distributed patterns of fMRI activity in human visual cortex. *NeuroImage* (2003)


Dayan, P. and Abbott, L.F. *Theoretical Neuroscience: Computational and Mathematical Modeling of Neural Systems*. The MIT Press (2001)


Dietterich, T.. Ensemble Methods in Machine Learning. *Lecture Notes in Computer Science* (2000)


Eger, E., Kell, C. and Kleinschmidt, A. Graded size sensitivity of object exemplar evoked activity patterns in human LOC subregions. *Journal of Neurophysiology* (2008)


Genuer R., Poggi J.-M. and Tuleau C. Variable selection using random forests. To appear in *Pattern Recognition Letters* (2010)