Beyond brain reading: randomized sparsity and clustering to simultaneously predict and identify

Alexandre Gramfort, Gaël Varoquaux, Bertrand Thirion

Beyond brain reading: randomized sparsity and clustering to simultaneously predict and identify

A. Gramfort[1,2], G. Varoquaux[1,2], B. Thirion[1,2]

[1] INRIA, Parietal team, Saclay, France
[2] LNAO/NeuroSpin, CEA Saclay, Bat. 145, 91191 Gif-sur-Yvette, cedex France

Abstract. The prediction of behavioral covariates from functional MRI (fMRI) is known as brain reading. From a statistical standpoint, this challenge is a supervised learning task. The ability to predict cognitive states from new data gives a model selection criterion: prediction accuracy. While a good prediction score implies that some of the voxels used by the classifier are relevant, one cannot state that these voxels form the brain regions involved in the cognitive task. The best predictive model may have selected by chance non-informative regions, and neglected relevant regions that provide duplicate information. In this contribution, we address the support *identification* problem. The proposed approach relies on randomization techniques which have been proved to be consistent for support recovery. To account for the spatial correlations between voxels, our approach makes use of a spatially constrained hierarchical clustering algorithm. Results are provided on simulations and a visual experiment.

1 Introduction

Functional MRI (fMRI) is an imaging technique that measures Blood Oxygen-Level Dependent signal changes caused by brain activity. Detecting and localizing these changes can be used to improve our understanding of brain function. Over the last decade, many contributions have proposed to tackle this challenge using statistical learning and more specifically supervised learning methods [10]. The data are fMRI volumes –3D images made of *voxels*– and the target to predict is, for example, the stimulus that was presented to the subject in the scanner. This formulation of the problem is commonly called *brain reading* or *decoding*.

A strong benefit that supervised learning methods bring to brain mapping is the ability of the estimator to account for a distributed pattern of active voxels. While standard statistics for brain mapping model only one voxel at a time, or local clusters, brain reading can be applied to full brain data. The method is said to be *multivariate*: the learned prediction function relies on correlations between distant brain regions.

To actually achieve *brain mapping*, the learning method used for decoding should inform about which voxels are useful for the prediction. This constraint naturally favors linear classifiers for which the prediction function is obtained from a linear combination of the voxel amplitudes. We call the *coefficients* of this linear combination the *weights* of the estimator. They form a spatial map.

Functional MRI data are a spatially smoothed representation of the underlying neural signals. Consequently, the activations are not only distributed over the entire brain but also spatially correlated. For better prediction performance, the estimators should incorporate this prior knowledge. It is natural to promote prediction functions relying on only a few brain regions, for instance using sparsity inducing regularization methods [6, 15, 2]. In addition, to account for the spatial structure in the signal, the estimator can make use of the three-dimensional grid structure over which the signal is defined. This can be achieved with convex regularization promoting piecewise constant weights [8] or by constructing hierarchically organized spatial features using a spatially constrained Ward clustering method [14] and learning a linear decision function defined over this new set of features. [9] perform the learning step with a greedy top-down approach while [5] use a hierarchical convex prior.

A caveat with the decoding approach is that the model is selected to optimize the prediction, while the localization of brain function requires instead to optimize the *identification* of the brain regions involved in the task. A good prediction indicates that the identified regions are sufficient to predict but it means neither that they are the true ones nor that they cannot be better estimated. A simple illustration is that different weights can lead to the same prediction accuracy [13]. In this paper we address the identification problem while taking into account the specificity of fMRI data: distributed patterns and spatial correlations.

Notations We write vectors with bold letters, $\mathbf{a} \in \mathbb{R}^n$, matrices with capital bold letters, $\mathbf{A} \in \mathbb{R}^{n \times n}$. We denote $\|\mathbf{a}\|_1 = \sum_{i=1}^n |\mathbf{a}_i|$ the ℓ_1 norm and $\|\mathbf{a}\|_2 = \sqrt{\sum_{i=1}^n \mathbf{a}_i^2}$ the ℓ_2 norm.

2 Method for sparse recovery with spatial structure

Let us consider the linear classification model:

$$y = \text{sign}(\mathbf{x}\,\mathbf{w} + b) = \text{sign}\big(\sum_{v=1}^{p} \mathbf{x}_v \mathbf{w}_v + b\big) \ , \tag{1}$$

where $y \in \{-1, 1\}$ represents the target to predict, *sign* stands for the sign function, p is the number of voxels in the grid, and $(\mathbf{w}, b)$ are the model parameters to be estimated: the *weight* vector and the intercept, also called bias term. The vector $\mathbf{x} \in \mathbb{R}^p$ is an fMRI activation volume. Using a logistic regression model, the estimation problem over a training set formed by n volumes reads:

$$(\widehat{\mathbf{w}}, \widehat{b}) = \underset{(\mathbf{w},b)}{\text{argmin}} \frac{1}{n} \sum_{i=1}^{n} \log\left(1 + e^{-y^i(\mathbf{x}^i\mathbf{w}+b)}\right) + \lambda\Omega(\mathbf{w}), \quad \lambda > 0, \tag{2}$$

where λ controls the level of regularization and Ω is the regularizing function that typically promotes sparse and potentially spatially structured weights. Sparse

logistic regression (SLR) refers to the case where $\Omega(\mathbf{w}) = \|\mathbf{w}\|_1$. We denote by $\mathbf{X} \in \mathbb{R}^{n \times p}$ the design matrix formed by the concatenation of all $\mathbf{x}^i$.

One of the issues with sparse methods is the instability of the estimated support of $\mathbf{w}$, particularly when the columns of $\mathbf{X}$ are very correlated as it is the case with fMRI data. To stabilize the estimated support, it has been proposed to randomly perturb the design matrix [7] by taking only a fraction of the training samples and randomly scale each column, in our case each voxel. By repeating the later procedure and then counting how often each voxel is selected across the repetitions, each voxel can be assigned a score. The higher is the score, the more likely is the voxel likely to be truly informative. In a regression setting, this procedure is called Randomized Lasso [7]. We apply it here to a binary classification problem (see *e.g.* [11]).

Let $k \in [1 \dots K]$ denote the repetition and $\mathbf{w}^k$ be the corresponding estimated weight vector. The design matrix $\mathbf{X}^k$ is formed by a random fraction π of the training data. Each column of $\mathbf{X}^k$ is then randomly scaled to 1 or to $1 - a$ with equal probability. The procedure is a subsampling of the data and a random perturbation of each voxel. The stability score of each voxel v is then the percentage of the repetitions for which the voxel has a non-zero weight, *i.e.*, is used for the prediction. A voxel v is used if the corresponding entry in the weight vector $\mathbf{w}^k$ estimated at repetition k is non-zero. We denote it by $v \in \text{supp}(\mathbf{w}^k)$. The stability score can then defined as $\mathbf{s}_v = \#\{k \text{ s.t. } v \in \text{supp}(\mathbf{w}^k)\}/K \in [0, 1]$. The estimated support is then defined as $\{v \text{ s.t. } \mathbf{s}_v \geq \tau\}$. In the following experiments a is set to 0.5 and π to 75%, while τ is estimated by cross-validation in a discrete set of values $\mathcal{T}$. Following experiments use a fixed value $\mathcal{T} = \{0.25\}$ or a grid $\mathcal{T} = \{0.1, 0.2, 0.3, 0.4, 0.5\}$. For every threshold $\tau \in \mathcal{T}$, a cross-validation score is obtained using a ℓ_2-logistic regression model ($\Omega(\mathbf{w}) = \|\mathbf{w}\|_2^2$). The estimated τ is the one yielding the highest score.

To improve the stability of the estimation and inform the estimator about the grid structure of the data, we propose using Ward hierarchical clustering as in [9] to train the classifiers on data-driven spatial clusters. At each repetition the brain volume is first segmented in q spatially connected regions in which the fMRI signal is averaged. As the hierarchical tree is estimated each time on a random fraction of the data, the tree is different for every randomization. Note that a similar procedure is performed in the Random Forests algorithm [1]. One obvious benefit of this procedure is that it tends to produce an "average" tree which balances with the greedy hierarchical construction of a single tree. The SLR is then fitted on a q-dimensional dataset. A voxel is marked as active in repetition k if it belongs to a region with a non-zero weight. Although the estimated $\text{supp}(\mathbf{w}^k)$ is in $\mathbb{R}^q$, we will still write $v \in \text{supp}(\mathbf{w}^k)$.

The main benefit of the additional clustering step is to reduce the correlations in the design matrix, therefore improving the behavior of sparse methods. Our method can thus select more voxels than the number of observations, which would be impossible with standard SLR and difficult with only randomization.

The procedure is summarized in Algorithm 1.

Algorithm 1 Randomized Sparse Logistic Regression with hierarchical features

Input: Set $0 < a < 1$, $\mathcal{T}$ (e.g. $\{0.1, 0.2, 0.3, 0.4, 0.5\}$), K (e.g. 200), π (e.g. 0.75).

1: Estimate q and λ with cross-validation
2: **for** $k = 1, ..., K$ **do**
3: Randomize design $\mathbf{X}^k$ with data subsampling and random feature scaling
4: Hierarchical clustering (segment brain in q regions)
5: Estimate $\mathbf{w}^k \in \mathbb{R}^q$ with SLR (2)
6: **end for**
7: Set scores $\mathbf{s}_v = \#\{k \text{ s.t. } v \in \text{supp}(\mathbf{w}^k)\}/K \in [0, 1]$
8: Set estimated support $\{v \text{ s.t. } \mathbf{s}_v \geq \tau\}$ ($\tau \in \mathcal{T}$ estimated by cross-validation with ℓ_2-logistic regression)

Table 1. Area under Precision-Recall curve as a function of the active region size.

Methods	Ward Rand. LR	F-test	ℓ_1-LR	ℓ_2-LR	SVM
1×1×1	0.84	0.589	**1.0**	0.773	0.773
2×2×2	**0.98**	0.829	0.523	0.775	0.775
3×3×3	**0.786**	0.749	0.456	0.535	0.631

3 Simulation study and fMRI results

We now present a simulation study followed by results on fMRI data recorded during an object recognition experiment. Experiments were performed with the scikit-learn [12] using LibLinear [3].

Simulation study – The simulation data consist of training and test sets each of 160 volumes. Each volume is a cube of (9×9×9 voxels). The active regions are 2 cubes of size 2×2×2 located at two opposite sides of the cube (see Fig. 1-a). Data are corrupted by a Gaussian additive noise and smoothed. The parameters λ and q are estimated by 5-fold cross-validation on the training set. Then stability scores are estimated with $K = 200$ repetitions.

Figure 1 presents the F-values for each voxel, as in conventional brain mapping, and the selection scores $\mathbf{s}$. Accuracy is quantified, for the identification, with a Precision-Recall (PR) curve on the recovered support and, for the prediction, with a Receiver-Operating-characteristic (ROC) curve on the predicted labels. Prediction performances using the known true weights are also given as baseline. A first interesting observation is that although SLR outperforms a linear SVM and a ℓ_2-logistic regression for prediction (Figure 1-e), it is clearly worse for the identification (Figure 1-d). This illustrates that the model that predicts the best may not be the model built from the true active voxels. What is also interesting it that the proposed method clearly outperforms all alternative methods for support recovery, while also giving almost optimal prediction accuracy. Identification results with different active regions sizes are present in Tab. 1. Our approach consistently provides the best estimation, except when the solution is very sparse (1 voxel), in which case it is outperformed by sparse estimators.

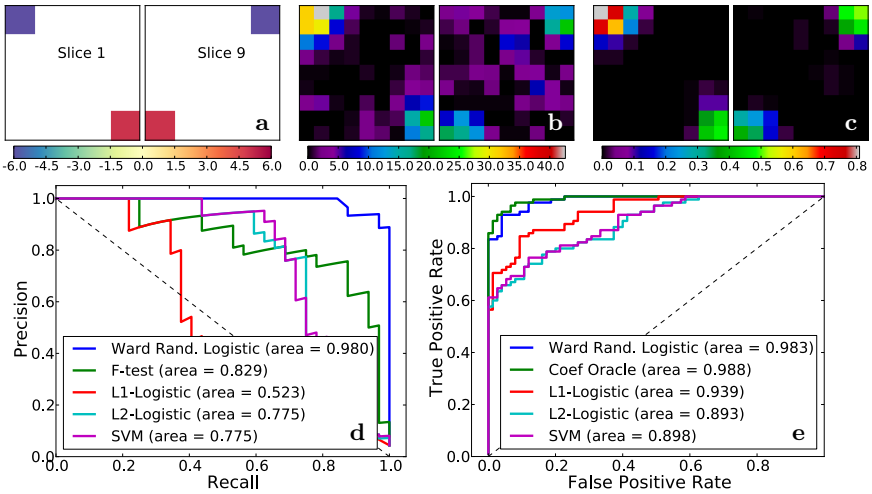

Fig. 1. Simulation results: **a.** ground truth, **b.** F-values, **c.** scores, **d.** identification precision-recall, **e.** prediction ROC

fMRI data analysis – The investigated fMRI data consist of five subjects recorded during a visual object recognition protocol [4]. In this experiment subjects were asked to recognize 8 different types of objects. We focus here on the binary classification task that consist in predicting whether the subject was viewing a *house* or a *face*. The data consist of 12 sessions, that were split in a training set and an independent test set. Each session contained 18 volumes (9 in each category). Preprocessing of the data consisted in motion correction using FSL MCFLIRT and a removal of linear trends in each session.

We first present results using respectively 4 and 8 sessions for the train and test sets. The first subject is presented in details in Fig. 2. The ROC curves show that the best prediction accuracy is obtained with the proposed method, followed by ℓ_1-logistic regression and then the ℓ_2 penalized methods (Logistic and linear SVM). As shown on Fig. 2-a, voxels with strong selection scores (above 0.25) are located within Fusiform gyrus in a region known as the Fusiform Face Area (FFA). The ROC curves for the other subjects are presented in Fig. 3. The mean ROC area across subjects is 0.989 while it is only 0.869 for the SLR, 0.807 for the ℓ_2-LR and 0.808 for the linear SVM. These results show that the proposed method consistently outperforms alternative approaches in terms of prediction accuracy. This method also yields a spatially structured and a meaningful estimated support in the FFA which suggests that the randomization procedure employed improves the support recovery as shown in [7].

In order to further investigate the performance of the method as a function of the number of training data, we have conducted the same experiments when varying the number of sessions used from estimating the support and fitting the predictive model. Results are presented in Figure 4. A first interesting observation is that all methods tend to predict almost perfectly when using a

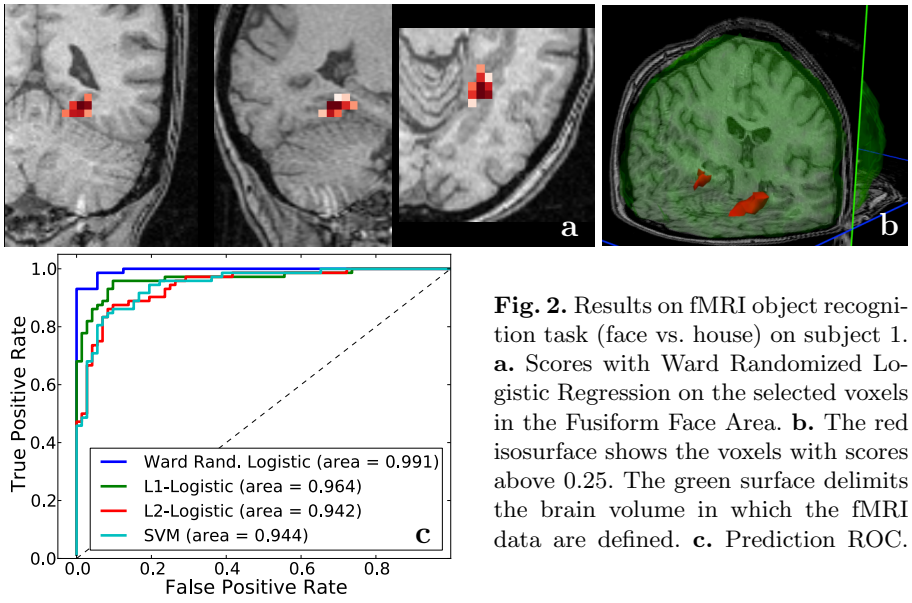

Fig. 2. Results on fMRI object recognition task (face vs. house) on subject 1. **a.** Scores with Ward Randomized Logistic Regression on the selected voxels in the Fusiform Face Area. **b.** The red isosurface shows the voxels with scores above 0.25. The green surface delimits the brain volume in which the fMRI data are defined. **c.** Prediction ROC.

large training set. Another observation is that here again, a linear SVM and an ℓ_2-logistic regression yield very similar results. The SLR outperforms the later methods when using more than 3 sessions for fitting the model. Finally, the proposed method is the only one yielding almost perfect predictions as soon as the number of sessions exceeds three.

4 Conclusion

In this work, we have shown that a randomization technique coupled with a spatial clustering algorithm could significantly improve the identification of predictive brain regions while yielding better prediction scores. The sparse randomized logistic used for that purpose allowed to stabilize the support estimation while the clustering pre-processing addresses successfully the problem of strong spatial correlations.

This contribution illustrates a somehow unintuitive fact that among the set of models, like the one obtained with a sparse method when varying the regularization parameter, the model that predicts the best is not always the model that identifies best the good voxels. The optimization of the prediction score on unseen data or the identification of the good voxels can lead to different models. The nice observations presented in this work, is that the proposed procedure improves both aspects, the support identification and the prediction scores.

References

1. Breiman, L.: Random forests. Machine Learning 45, 5–32 (2001)

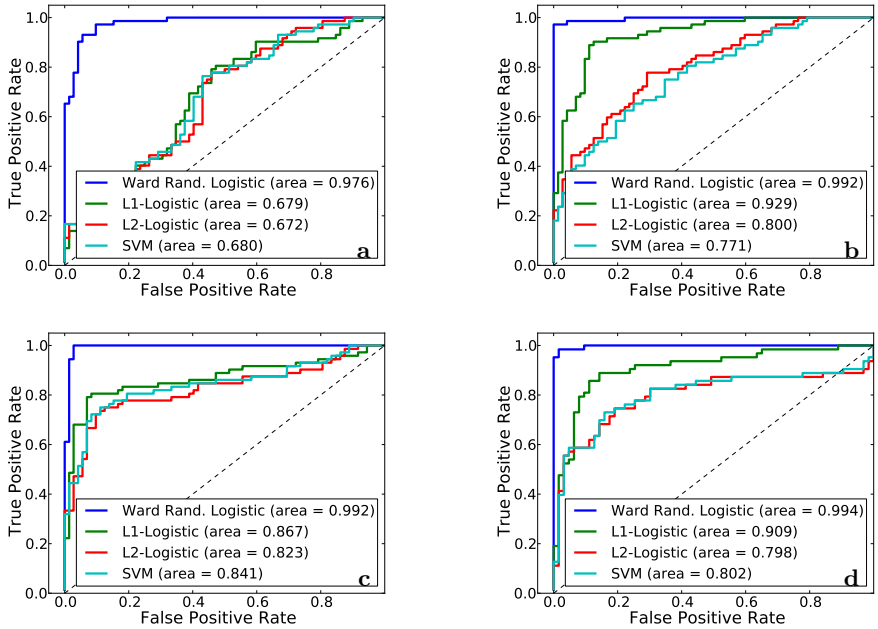

Fig. 3. Prediction ROC on fMRI object recognition task (face vs. house) for the four other subjects. **a.** Subject 2 **b.** Subject 3 **c.** Subject 4 **d.** Subject 5

2. Carroll, M.K., Cecchi, G.A., Rish, I., Garg, R., Rao, A.R.: Prediction and interpretation of distributed neural activity with sparse models. NeuroImage 44(1), 112 – 122 (2009)
3. Fan, R., Chang, K., Hsieh, C., Wang, X., Lin, C.: LIBLINEAR: A library for large linear classification. The Journal of Machine Learning Research 9, 1871–1874 (2008)
4. Haxby, J.V., Gobbini, I.M., Furey, M.L., Ishai, A., Schouten, J.L., Pietrini, P.: Distributed and overlapping representations of faces and objects in ventral temporal cortex. Science 293(5539), 2425–2430 (2001)
5. Jenatton, R., Gramfort, A., Michel, V., Obozinski, G., Eger, E., Bach, F., Thirion, B.: Multi-scale Mining of fMRI data with Hierarchical Structured Sparsity. ArXiv e-prints (May 2011)
6. Krishnapuram, B., Carin, L., Figueiredo, M.A., Hartemink, A.J.: Sparse multinomial logistic regression: Fast algorithms and generalization bounds. IEEE Transactions on Pattern Analysis and Machine Intelligence 27, 957–968 (2005)
7. Meinshausen, N., Bühlmann, P.: Stability selection. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 72(4), 417–473 (2010)
8. Michel, V., Gramfort, A., Varoquaux, G., Eger, E., Thirion, B.: Total variation regularization for fMRI-based prediction of behavior. Medical Imaging, IEEE Transactions on 30(7), 1328–1340 (july 2011)
9. Michel, V., Gramfort, A., Varoquaux, G., Eger, E., Keribin, C., Thirion, B.: A supervised clustering approach for fMRI-based inference of brain states. Pattern Recognition p. epub ahead of print (Apr 2011)

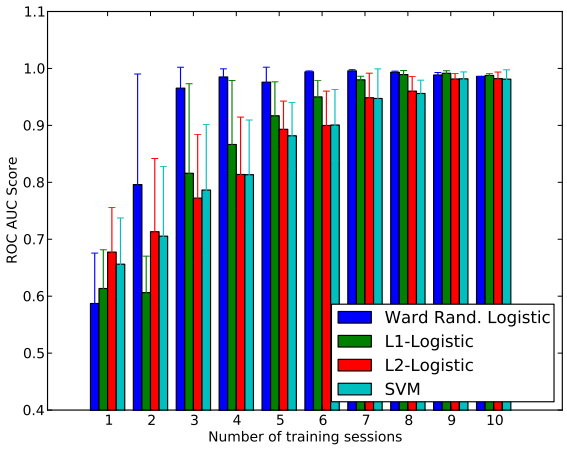

Fig. 4. ROC AUC scores as a function of the number of training sesssions.

10. Mitchell, T.M., Hutchinson, R., Niculescu, R.S., Pereira, F., Wang, X., Just, M., Newman, S.: Learning to decode cognitive states from brain images. Machine Learning 57(1), 145–175 (2004)
11. Ng, A.: Feature selection, l1 vs. l2 regularization, and rotational invariance. In: Proceedings of the twenty-first international conference on Machine learning. p. 78. ACM (2004)
12. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: Machine learning in Python. Journal of Machine Learning Research 12, 2825–2830 (2011)
13. Rish, I., Cecchi, G., Heuton, K., Baliki, M., Apkarian, A.: Sparse regression analysis of task-relevant information distribution in the brain. In: SPIE Medical Imaging (2012)
14. Ward, J.H.: Hierarchical grouping to optimize an objective function. Journal of the American Statistical Association 58, 236 (1963)
15. Yamashita, O., aki Sato, M., Yoshioka, T., Tong, F., Kamitani, Y.: Sparse estimation automatically selects voxels relevant for the decoding of fMRI activity patterns. NeuroImage 42(4), 1414 – 1429 (2008)