Improved brain pattern recovery through ranking approaches

Fabian Pedregosa, Elodie Cauvet, Gaël Varoquaux, Christophe Pallier,

Bertrand Thirion, Alexandre Gramfort

Improved brain pattern recovery through ranking approaches

Fabian Pedregosa *†§, Elodie Cauvet ‡†, Gaël Varoquaux*†, Christophe Pallier ‡*†, Bertrand Thirion*†
and Alexandre Gramfort*†,

*Parietal Team, INRIA Saclay-Île-de-France, Saclay, France
† CEA, DSV, I²BM, Neurospin bât 145, 91191 Gif-Sur-Yvette, France
‡ Inserm, U992, Neurospin bât 145, 91191 Gif-Sur-Yvette, France
§ SIERRA Team, INRIA Paris - Rocquencourt, Paris, France

Abstract—**Inferring the functional specificity of brain regions from functional Magnetic Resonance Images (fMRI) data is a challenging statistical problem. While the General Linear Model (GLM) remains the standard approach for brain mapping, supervised learning techniques (*a.k.a.* decoding) have proven to be useful to capture multivariate statistical effects distributed across voxels and brain regions. Up to now, much effort has been made to improve decoding by incorporating prior knowledge in the form of a particular regularization term. In this paper we demonstrate that further improvement can be made by accounting for non-linearities using a ranking approach rather than the commonly used least-square regression. Through simulation, we compare the recovery properties of our approach to linear models commonly used in fMRI based decoding. We demonstrate the superiority of ranking with a real fMRI dataset.**

Index Terms—**fMRI, supervised learning, decoding, ranking**

I. INTRODUCTION

The prediction of behavioral information or cognitive states from brain activation images such as those obtained with fMRI can be used to assess the specificity of several brain regions for certain cognitive or perceptual functions. This kind of analysis is implemented by learning a classifier or regression function that fits a given *target* variable given fMRI activations. The accuracy of this prediction depends on whether it uses the relevant variables *i.e.* the correct brain regions. *Recovering* the truly predictive pattern has proven to be challenging from a statistical point of view: the high dimensionality of the data together with the limited number of images makes the problem of brain pattern recovery an ill-posed problem.

So far, the approaches proposed to address this issue have relied on linear models, with univariate, *i.e.* voxel-based, Anova (analysis of variance) for hypothesis testing, or, for predictive modeling, with the choice of a regularizer using a priori domain-specific knowledge, such as the ℓ_1-norm to promote sparsity [1], [2], total variation to promote spatial smoothness [3]. Various data fit terms have been used, Logistic Regression (LR) [2], Linear SVM [4], Lasso [1]. While Linear SVM and LR cannot address the recovery problem for multiclass problems, linear regression models assume a linear relationship between the quantity to predict and the amplitude of the fMRI signals. If the linear relationship does not hold in practice, then the estimation of the predictive patterns may suffer from a loss of statistical power. This can be particularly relevant with Blood Oxygen-Level Dependent (BOLD) signals observed in fMRI, where a saturation effect is expected as the level of signal increases.

When targets to predict consist of ordered values, as in a parametric design, such as clinical scores, pain levels or the complexity of a cognitive task, the response to these different conditions can reflect the non-linearities in the data. In such situation, we propose to use a data fit term, known as loss function, not relying on an assumption of linearity but only of increasing response. We show on simulations that this new formulation opens the door to capturing the non-linearity and leads to better recovery of the predictive brain patterns. On an fMRI dataset we show that the new formulation leads to models with better recovery properties.

a) Notations: We write vectors in bold, $\boldsymbol{a} \in \mathbb{R}^n$, matrices with capital bold letters, $\boldsymbol{A} \in \mathbb{R}^{n \times p}$. The dot product between two vectors is denoted $\langle \boldsymbol{a}, \boldsymbol{b} \rangle$. We denote by $\|\mathbf{a}\| = \sqrt{\langle \boldsymbol{a}, \boldsymbol{a} \rangle}$ the ℓ_2 norm of a vector.

II. LEARNING A LINEAR MODEL FROM FMRI DATA

Following standard statistical learning notations we denote by $\boldsymbol{x}_i \in \mathbb{R}^p$, $1 \le i \le n$, the data and $y_i \in \mathcal{Y}$ the target variables. In this paper, we aim at learning a weight vector $\mathbf{w} \in \mathbb{R}^p$ such that the prediction of $\boldsymbol{y}$ can be non-linearly related to the value of $\mathbf{w}^T \mathbf{x}$. The vector $\mathbf{w}$ corresponds here to a brain map that can be represented in brain space as a volume for visualization of the predictive pattern of voxels. It is useful to rewrite these quantities in matrix form; more precisely, we denote by $\mathbf{X} \in \mathbb{R}^{n \times p}$ the design matrix assembled from n fMRI volumes and by $\boldsymbol{y} \in \mathbb{R}^n$ the corresponding n targets. In other words, each row of $\mathbf{X}$ is a p-dimensional sample, i.e., an activation map of p voxels related to one stimulus presentation.

A standard approach to perform the estimation of $\mathbf{w}$ leads to the following minimization problem

$$\hat{\mathbf{w}} = \underset{\mathbf{w}}{\operatorname{argmin}} \ \mathcal{L}(\boldsymbol{y}, \boldsymbol{X}, \boldsymbol{w}) + \lambda \, \Omega(\mathbf{w}) \ , \ \ \lambda \ge 0 \qquad (1)$$

where $\lambda \, \Omega(\mathbf{w})$ is the regularization term and $\mathcal{L}(\mathbf{y}, \mathbf{X}, \mathbf{w})$ is the loss function. The parameter λ balances the loss function and the penalty $\Omega(\mathbf{w})$.

If the explained variable is a linear combination of the images, $\mathbf{y} = \mathbf{Xw} + \boldsymbol{\epsilon}$, we can estimate $\hat{\mathbf{w}}$ using the mean squared error loss function $\mathcal{L}(\mathbf{y}, \mathbf{X}, \mathbf{w}) = \|\boldsymbol{y} - \mathbf{Xw}\|^2$.

However, with fMRI the linearity assumption may not be valid. Instead, we model our explained variable as $\mathbf{y} = F(\mathbf{Xw}) + \epsilon$, where F is a non-decreasing function.

We introduce the use of pairwise loss functions. These loss functions only assume the target values to be a non-decreasing function of the data. They have been widely used in ranking, a type of supervised machine learning problem whose goal is to automatically construct an order from the training data. A pairwise loss function operates on pairs of images: given a pair of images $(\mathbf{x}_i, \mathbf{y}_i)$ and $(\mathbf{x}_j, \mathbf{y}_j)$, $\mathbf{y}_i \neq \mathbf{y}_j$ we build a model that predicts the sign of $\mathbf{y}_i - \mathbf{y}_j$.

Let $\mathcal{I}$ denote the index set of all considered pairs. Note that in some settings it might be important to restrict ourselves to a selected subgroup of all pairs, *e.g.* to the pairs of images of a single subject or to the pairs of images corresponding to a single session. For this purpose we define $a_{ij} \in \mathbb{R}, (i,j) \in \mathcal{I}$ to be a weight associated to each pair. We will now present the pairwise loss functions used in this article:

- *Pairwise hinge loss* [5]. This is a natural extension of the loss function used by Support Vector Machines and has been successfully used in information retrieval [6].

$$\sum_{(i,j)\in\mathcal{I}} a_{ij}[1 - \mathbf{w}^T(\mathbf{x}_i - \mathbf{x}_j)]_+ \quad (2)$$

where $[z]_+ = \max\{z, 0\}$.
- *Pairwise logistic loss* [7]. This is the pairwise extension of the logistic regression loss function.

$$\sum_{(i,j)\in\mathcal{I}} a_{ij} \log(1 + \exp(\mathbf{w}^T(\mathbf{x}_i - \mathbf{x}_j))) \quad (3)$$

When noise is present in the model, the order of two samples might get inverted, a phenomenon known as *label switching*. Because this only affects labels that lie close, it is natural to penalize more the misclassification of distant labels. By setting the sample weights to a_{ij} to a value that increases as labels become more separated, we become more robust to label switching. In the case of hinge loss functions, several strategies for choosing the appropriate weights are discussed in [6].

On the implementation side, both pairwise hinge loss and pairwise logistic loss can be implemented on top of existing SVM and Logistic Regression solvers, respectively, by taking the difference of pairs as input values. In practice, we used the liblinear [8] library via the scikit-learn [9] library.

III. SIMULATION

b) Data generation: The simulated data $\mathbf{X}$ contains volumes of size $5 \times 5 \times 5$ and $7 \times 7 \times 7$, each one consisting of Gaussian white noise smoothed by a Gaussian kernel with standard deviation of 2 voxels. This mimics the spatial correlation structure observed in real fMRI data. The simulated vector of coefficients $\mathbf{w}$ has a support restricted to four cubic Regions of Interest (ROIs) of size $(2 \times 2 \times 2)$. The values of $\mathbf{w}$ restricted to these ROIs are $\{1, 1, -1, -1\}$.
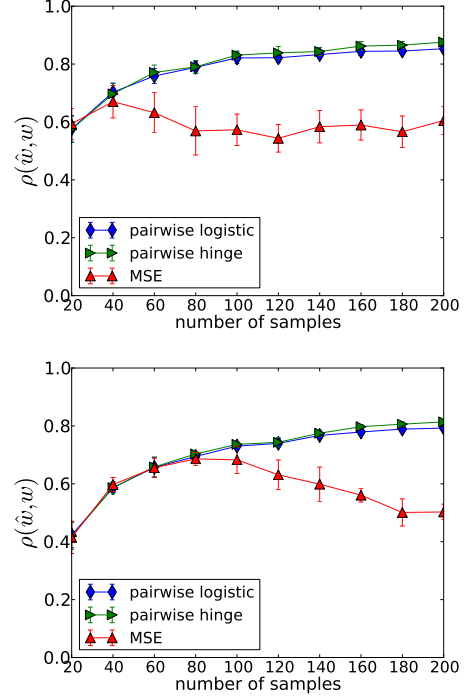

Fig. 1. Correlation between the estimated vector $\hat{\boldsymbol{w}}$ and the ground truth $\boldsymbol{w}$ for different loss functions as the number of considered samples increases (higher is better) for dimensions $5 \times 5 \times 5$ and $7 \times 7 \times 7$ respectively. Pairwise loss functions outperform linear regression as the number of samples increases and tend to a perfect recovery.
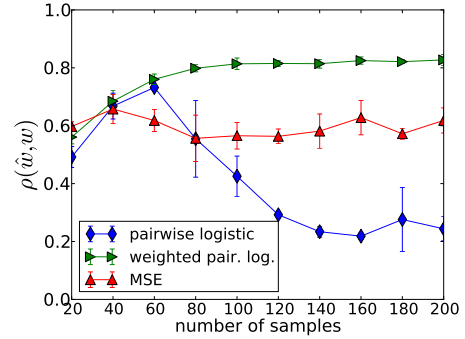

Fig. 2. Correlation between the estimated vector $\hat{\boldsymbol{w}}$ and the ground truth $\boldsymbol{w}$ for different loss functions with a noise level of 5%. Appropriately setting the weights plays a major role in robustness. Without the correct weighting and under noisy conditions, pairwise logistic loss function fails to recover the correct model.

The target variable $\mathbf{y}_l \in \mathbb{R}^n$ is simulated as a linear model:

$$\mathbf{y}_l = \mathbf{Xw} + \epsilon \quad (4)$$

where the noise $\epsilon_i \in [-\frac{\sigma}{2}, \frac{\sigma}{2}]$ follows a uniform distribution. σ is chosen such that the signal-to-noise ratio verifies $\|\epsilon\|/\|\boldsymbol{Xw}\| = 5\%$. We then define another target variable $\mathbf{y}_{nl}$ to be a sigmoid function of $\mathbf{y}_l$, that is,

$$\mathbf{y}_{nl} = \frac{1}{1 + \exp(-\mathbf{y}_l)} \quad (5)$$

For each of the loss functions introduced earlier and mean squared error, we compute the correlation coefficient $\rho(\boldsymbol{w}, \hat{\boldsymbol{w}}) = \langle \boldsymbol{w}, \hat{\boldsymbol{w}} \rangle / (\|\boldsymbol{w}\| \|\hat{\boldsymbol{w}}\|)$. This gives us the goodness of fit for the estimated $\hat{\boldsymbol{w}}$. A method with correlation coefficient of 1 is able to recover perfectly the ground truth. Since we are interested in the estimation of $\boldsymbol{w}$, we restricted ourselves to linear models, leaving out models such as kernel ridge or support vector regression with Gaussian kernel.

For all models we added ℓ_2 regularization in the form of $\lambda \|\boldsymbol{w}\|^2$ and we cross-validated λ separately for each loss. In this setting, the model with mean squared error loss is equivalent to *ridge regression*. In order to focus on the effect of non-linearity, we first considered a noiseless simulation using trivial weights $a_{ij} = 1$.

Once we learned a vector $\hat{\boldsymbol{w}}$ for each method we compute the correlation coefficient with the ground truth for different sizes of the training data. The experiment is repeated 10 times with different initialization of the pseudorandom number generator. We compute errorbars and show the results in figure 1. As the number of samples increases, the linear model stalls and pairwise loss functions outperform MSE on both $5 \times 5 \times 5$ and $7 \times 7 \times 7$ dimension. As expected, the higher dimensionality of the second simulation makes the correlation coefficient decrease. However, unlike MSE, ranking tends to a perfect recovery as the number of samples increases. Both pairwise loss functions perform equivalently and have a significatively higher correlation coefficient than MSE. In the rest of the paper we will use pairwise logistic as loss function. As a result, pairwise loss functions should be preferred over MSE in situations where underlying model is non-linear. Notice that we fixed the non-linearity to be a sigmoid function, but the pairwise loss functions only assume that this function is non-decreasing. Unlike linear regression models, pairwise loss functions are indeed able to learn the structure on the non-linear transform F.

We now consider the model with noise as in (4) and use non-trivial weights a_{ij}. To account for label switching, we set a_{ij} to zero for pairs with too similar labels:

$$a_{ij} = \begin{cases} 0 & \text{if } |\boldsymbol{y}_i - \boldsymbol{y}_j| < \sigma \\ 1 & \text{otherwise .} \end{cases} \quad (6)$$

In the case of discrete values, this would be equivalent to zeroing weights for which the labels are adjacent. We now compute the correlation coefficient for weighted and unweighted pairwise logistic models and linear ridge regression model. The result can be seen in figure 2 for dimension $5 \times 5 \times 5$. The unweighted logistic model breaks down in presence of noise and performs worse than linear ridge regression. On the other hand, appropriately setting the weights a_{ij} has a major effect on robustness, where this model outperforms MSE in a noisy setting. Note also that weighted pairwise logistic has smaller variance than MSE.

IV. RESULTS ON fMRI DATA

This dataset, described in [10], consists of 34 healthy volunteers scanned while listening to 16 words sentences with five
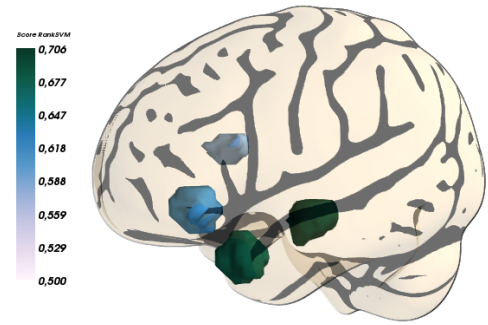

Fig. 3. Scores obtained with the pairwise logistic on the 4 different ROIs. The regions with the best predictive power are the temporal pole the anterior superior temporal sulcus.

different levels of complexity. These were 1 word constituent phrases (the simplest), 2 words, 4 words, 8 words and 16 words respectively, corresponding to 5 levels of complexity which was used as class label in our experiments. To clarify, a sentence with 16 words using 2 words constituents is formed by a series of 8 pairs of words. Words in each pair have a common meaning but there is no meaning between each pair. A sentence has therefore the highest complexity when all the 16 words form a meaningful sentence.

The dataset consists of 8 manually labeled ROIs, some informative and some not. For each ROI separately, we split the data into 60% training samples, 20% for parameter selection and 20% for validation. We trained a pairwise logistic model and set the ℓ_2 regularization by cross validation. We choose a_{ij} to be zero if classes are adjacent, i.e. if $|\boldsymbol{y}_i - \boldsymbol{y}_j| \leq 1$ and if $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ do not belong to the same subject, in order to consider exclusively non-adjacent pairs of images from the same subject. In all other cases, a_{ij} was set to one. We computed the generalization score on the validation set as the mean number of inversions with respect to the order in labels, i.e. the sign flips $\text{sgn}((\boldsymbol{X}_i - \boldsymbol{X}_j)\hat{\boldsymbol{w}}) \neq \text{sgn}(\boldsymbol{y}_i - \boldsymbol{y}_j)$ for all pairs of images in the validation set.

We kept the four ROIs with highest scores (see figure 3). These are: Anterior Superior Temporal Sulcus (aSTS), Temporal Pole (TP), Inferior Frontal Gyrus Orbitalis (IFGorb) and Inferior Frontal Gyrus triangularis (IFG tri).

In order to inspect the properties of the estimated functions F for each ROI, we estimated $\hat{\boldsymbol{w}}$ using a pairwise logistic model. We then projected our data $\boldsymbol{X}$ along this vector $\hat{\boldsymbol{w}}$ and regularized the result using non parametric locally weighted scatterplot smoothing (LOWESS). Results in figure 4 show that the linearity varies in shape across ROIs which suggests that different regions exhibit different sensitivities to the complexity parameter under investigation. We see however a trend in the figures towards non-linear and non-decreasing functions with some saturation effect of the BOLD signal as in the temporal pole (TP).

In the case of the Temporal Pole (TP), which is the ROI revealing the highest saturation effect, an F-test on the data $(\boldsymbol{X}\hat{\boldsymbol{w}}, y)$ reveals that the quadratic polynomial model fits better the data than a linear model (p-value < 0.03). As
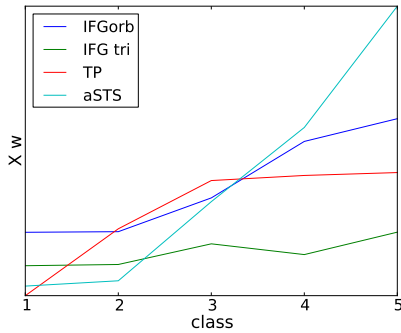
Fig. 4. Data projected along $\hat{w}$ showing the non-linear effect in the 4 regions of interest. This projection gives an insight on the relationship between the BOLD signal and the explained variable. We observe that the the shape of the non-linearity varies across brain regions. Apart from IFG tri, all regions show a saturation effect in the BOLD response.

shown in the simulations, in this particular case pairwise loss functions are likely to improve the recovery of active brain regions.

V. Conclusion

In this paper, we investigated the use of pairwise loss functions to improve the problem of brain pattern recovery with supervised learning applied to fMRI data. Through simulations, we showed the benefit of such loss functions when the target to predict is non-linearly related to the voxel amplitudes. Experimental results on fMRI data confirmed the presence of such non-linear effects in the data which suggest that the pairwise approach should improve the identification of predictive brain patterns on experimental data.

This work shows that improvements in recovery of brain activation patterns should not only rely on the choice of a particular regularizer, but also on an appropriate loss function. Here we have only considered ℓ_2-penalized models, but a natural extension to work with full brain data would be to consider pairwise loss functions combined with sparse structured penalizations which incorporate domain-specific knowledge. This opens the path to further improvements and refinements in the recovery of brain pattern activation via supervised learning.

References

[1] O. Yamashita, M. aki Sato, T. Yoshioka, F. Tong, and Y. Kamitani, "Sparse estimation automatically selects voxels relevant for the decoding of fMRI activity patterns," *NeuroImage*, vol. 42, no. 4, pp. 1414 – 1429, 2008.

[2] S. Ryali, K. Supekar, D. A. Abrams, and V. Menon, "Sparse logistic regression for whole-brain classification of fmri data." *NeuroImage*, vol. 51, no. 2, pp. 752–764, 2010.

[3] V. Michel, A. Gramfort, G. Varoquaux, E. Eger, and B. Thirion, "Total variation regularization for fMRI-based prediction of behaviour." *IEEE Transactions on Medical Imaging*, vol. 30, no. 7, pp. 1328 – 1340, Feb. 2011.

[4] J. Mouro-Miranda, A. L. Bokde, C. Born, H. Hampel, and M. Stetter, "Classifying brain states and determining the discriminating activation patterns: Support vector machine on functional MRI data," *NeuroImage*, vol. 28, p. 980, 2005.

[5] R. Herbrich, T. Graepel, and K. Obermayer, *Large margin rank boundaries for ordinal regression*. MIT Press, Cambridge, MA, 2000, vol. 88, pp. 115–132.

[6] Y. Cao, J. Xu, T.-Y. Liu, H. Li, Y. Huang, and H.-W. Hon, "Adapting ranking svm to document retrieval," in *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, ser. SIGIR '06. New York, NY, USA: ACM, 2006, pp. 186–193.

[7] O. Dekel, C. Manning, and Y. Singer, "Log-linear models for label ranking," *Advances in Neural Information Processing Systems*, vol. 16, no. 2, p. 497504, 2003.

[8] R. E. Fan, K. W. Chang, C. J. Hsieh, X. R. Wang, and C. J. Lin, "Liblinear: A library for large linear classification," *The Journal of Machine Learning Research*, vol. 9, no. 6/1/2008, pp. 1871–1874, 2008.

[9] F. Pedregosa, G. Varoquaux, A. Gramfort *et al.*, "Scikit-learn: Machine Learning in Python ," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[10] E. Cauvet, N. Hara, S. Dehaene, and C. Pallier, "Investigations of musical and linguistic structures using fMRI." *(in prep)*, 2012.