Identifying predictive regions from fMRI with TV-ℓ_1 prior

Alexandre Gramfort[1,2]
[1]*Institut Mines-Télécom, Télécom ParisTech, CNRS LTCI*
[2]*Neurospin bât 145, CEA, DSV, I²BM*
Gif-Sur-Yvette, France
alexandre.gramfort@telecom-paristech.fr

Bertrand Thirion[2,3]
Gaël Varoquaux[2,3]
[3]*Parietal Team*
INRIA Saclay-Île-de-France, France
firstname.lastname@inria.fr

***Abstract*—"Decoding", *i.e.* predicting stimulus related quantities from functional brain images, is a powerful tool to demonstrate differences between brain activity across conditions. However, unlike standard brain mapping, it offers no guaranties on the localization of this information. Here, we consider decoding as a statistical estimation problem and show that injecting a spatial segmentation prior leads to unmatched performance in recovering predictive regions. Specifically, we use ℓ_1 penalization to set voxels to zero and Total-Variation (TV) penalization to segment regions. Our contribution is two-fold. On the one hand, we show via extensive experiments that, amongst a large selection of decoding and brain-mapping strategies, TV+ℓ_1 leads to best region recovery. On the other hand, we consider implementation issues related to this estimator. To tackle efficiently this joint prediction-segmentation problem we introduce a fast optimization algorithm based on a primal-dual approach. We also tackle automatic setting of hyper-parameters and fast computation of image operation on the irregular masks that arise in brain imaging.**

***Keywords*-fMRI; supervised learning; total-variation; sparse; decoding; primal-dual optimization; support recovery;**

I. Introduction

Functional MRI (fMRI) gives images of brain activity via Blood Oxygen-Level Dependent (BOLD) signal changes. Though noisy and indirect, it is the workhorse of functional brain mapping, relating cognition or functional pathologies to their neural basis. Brain "decoding" can extract maps that predict behavior from fMRI. While it is an impressive evidence that the observed brain activity expresses differences between the behavioral conditions, it offers no guaranties on the localization of this information. In this regard, brain mapping with univariate statistics remains the reference tool. Likewise, thresholding predictive maps to retain predictive regions defeats the multivariate model and does not bring guarantees for recovery.

The primary use of decoding in neuroscience research is to provide a gage on the presence in the brain images of discriminant information with regards to the experimental conditions. In this sense, predictive power on left-out data is the figure of merit of decoding: a prediction above chance establishes the presence of a significant effect in the data. A critical point for neuroscientists is then to know what are the variables driving this effect, *i.e.* what brain regions. This goal naturally favors linear classifiers for which the prediction is obtained from a linear combination of the voxel amplitudes. The *weights* of the estimator then form a spatial map that can be defined over the entire brain, hence exploiting correlations between distant brain regions: the decoder is said to be *multivariate*. As we can see, another purpose of decoders is to perform estimation of this weight map so that it highlights the predictive regions [1]. For this, certain decoding procedures [2], [3] output a statistical test per voxel, a multivariate extension of standard analysis. However, as highlighted by [4], these procedures can outline different brain regions.

The challenge that we address here is to reconcile the two goals of prediction and region recovery in one decoding method: providing a unique map of weights that gives good out-of-sample prediction and segments clearly predictive regions. For this purpose, we study empirically a large variety of decoding approaches on simulations where the ground truth is known. In addition, we contribute an efficient method that uses a prior specifically-crafted for our purpose, building upon previous work [5].

Indeed, from a statistical standpoint, estimation of this linear model is ill-posed, as the number of unknowns is commonly 50 000 voxels, while the number of observations never exceeds a few thousands. It thus requires regularization, preferably compatible with prior knowledge on the data. Functional MRI data are a spatially-smoothed representation of the underlying neural signals. Consequently, the activations are spatially correlated. For better prediction performance, a decoder should account for this structure by using a spatial model. This can be achieved with convex penalization promoting isotropic smooth weights via a graph [6] or piecewise constant weights with Total-Variation [7]. The second important insight for fMRI decoding is that the extent of the regions involved in the task is limited. It is therefore natural to promote weight maps with only a small fraction of non-zero voxels, *e.g.* using sparsity inducing norms such as the ℓ_1 norm [1]. Combining both of these insights leads to consider TV-ℓ_1 penalization [5] that achieves the segmentation of a limited number of predictive brain regions when decoding from full brain data.

We now present the model and the convex optimization procedure we employed. We then discuss key practical

details that significantly improve performance and usability of the method, before showing some results on simulations and publicly available fMRI data.

Notation: We write vectors with bold letters, $\mathbf{a} \in \mathbb{R}^N$ and matrices with capital bold letters, $\mathbf{A} \in \mathbb{R}^{N\times N}$. $\mathbf{a}[i]$ stands for the ith entry in $\mathbf{a}$ and $\mathbf{A}[i,\cdot]$ the ith row of $\mathbf{A}$. $\|\mathbf{a}\|_2 = \sqrt{\sum_i^N \mathbf{a}[i]^2}$ is the ℓ_2 norm. $\mathbf{A}^T$ stands for the matrix transpose. $\mathrm{mod}(\cdot, p)$ stands for the integer p-modulo.

II. Solving the TV-ℓ_1 regression

A. An efficient algorithm

Let us consider the standard linear supervised model $y = f(\mathbf{x}\,\mathbf{w} + b)$ where $y \in \mathcal{Y}$ represents the target to predict, $\mathbf{x} \in \mathbb{R}^P$ is an fMRI volume made of P voxels, $\mathbf{w} \in \mathbb{R}^P$ is a *weight* vector and b is a scalar called *intercept*, or *bias* term. For regression $\mathcal{Y} = \mathbb{R}$ and f is the identity. Let N denote the number of fMRI volumes. The matrix $\mathbf{X} \in \mathbb{R}^{N\times P}$ is formed by the concatenation of the data from all subjects.

The estimation of the model parameters $(\mathbf{w}, b)$ can then be done by minimization of the errors over the training data. In a regression setup, mean squared error (MSE) is a natural way to quantify training errors. The estimation, formalized as a variational problem, reads:

$$\hat{\mathbf{w}},\, \hat{b} = \underset{\mathbf{w},b}{\operatorname{argmin}} \frac{1}{N}\sum_{i=1}^{N}\big(\mathbf{y}[i] - \mathbf{X}[i,\cdot]\mathbf{w} + b\big)^2 + \lambda\,\Omega(\mathbf{w}) \ . \quad (1)$$

with $\lambda \geq 0$, where Ω is the penalization term. Here we propose to use a combined TV and ℓ_1 regularization $\Omega(\mathbf{w}) = (1-\rho)\mathrm{TV}(\mathbf{w}) + \rho\|\mathbf{w}\|_1$, $0 \leq \rho \leq 1$. When $\rho = 0$ it is equivalent to TV [7] while for $\rho = 1$ the spatial model is ignored in favor of an ℓ_1 penalty, a.k.a. Lasso. Let $\nabla_x \in \mathbb{R}^{P\times P}$ (resp. ∇_y and ∇_z) denote the spatial gradient obtained by finite differences along the x direction (resp. y and z directions). Let $\nabla \in \mathbb{R}^{3P\times P}$ be the concatenation of the 3 gradients. The sparse TV regularization can be written as: $\Omega(\mathbf{w}) = \|\mathbf{K}\,\mathbf{w}\|_{21+1}$ where $\mathbf{K} \in \mathbb{R}^{4P\times P}$ is obtained by concatenating $(1-\rho)\nabla$ and $\rho\mathbf{I}$ matrices, and the structured norm reads $\|\mathbf{z}\|_{21+1} = \sum_{p=1}^{P}\sqrt{\mathbf{z}[p]^2 + \mathbf{z}[p+P]^2 + \mathbf{z}[p+2P]^2} + |\mathbf{z}[p+3P]|$. After discarding the bias term b from the estimation by centering the data and the target, (1) can be written as:

$$\underset{\mathbf{w}}{\operatorname{argmin}}\ G(\mathbf{w}) + F(\mathbf{K}\,\mathbf{w}) \ , \quad (2)$$

where G is quadratic, $G(\mathbf{w}) = \frac{1}{N}\|\mathbf{y} - \mathbf{X}\mathbf{w}\|_2^2$, and F is the convex structured norm ℓ_{21+1}. We now introduce the tools necessary to minimize such a function.

Definition 1 (Proximity operator): Let $\varphi : \mathbb{R}^M \to \mathbb{R}$ be a proper convex function. The proximity operator associated with φ, denoted by $\mathrm{prox}_\varphi : \mathbb{R}^M \to \mathbb{R}^M$ reads:

$$\mathrm{prox}_\varphi(\mathbf{y}) = \underset{\mathbf{x}\in\mathbb{R}^M}{\operatorname{argmin}} \frac{1}{2}\big\|\mathbf{y} - \mathbf{x}\big\|_2^2 + \varphi(\mathbf{x}) \ .$$

Definition 2 (Fenchel conjugate): The Fenchel conjugate $\varphi^* : \mathbb{R}^M \to \mathbb{R}$ associated to $\varphi : \mathbb{R}^M \to \mathbb{R}$ reads:

$$\varphi^*(\mathbf{y}) = \max_{\mathbf{x}\in\mathbb{R}^M} \mathbf{x}^T\mathbf{y} - \varphi(\mathbf{x}) \ .$$

Lemma 1 (Proximity operator for F): Let $\mathbf{z} \in \mathbb{R}^{4P}$, the proximity operator reads $\mathbf{z} = \mathrm{prox}_{\lambda F}(\mathbf{y})$ where

$$\mathbf{z}[i] = \begin{cases} \mathbf{y}[i]\left(1 - \dfrac{s[\mathrm{mod}(i-1,P)+1]}{\lambda}\right)_+ & \text{if } 1 \leq i \leq 3P \\ \mathbf{y}[i]\left(1 - \dfrac{|\mathbf{y}[i]|}{\lambda}\right)_+ & \text{if } i > 3P \end{cases} ,$$

with $(a)_+ = \max(a, 0)$, $a \in \mathbb{R}$, $s \in \mathbb{R}^P$ such that $s[i] = \sqrt{\mathbf{y}[i]^2 + \mathbf{y}[i+P]^2 + \mathbf{y}[i+2P]^2}$, and setting $\frac{0}{0} = 0$.

Proof: See e.g. [8]

The proximity operator for F^* can then be obtained from the identity: $\mathbf{x} = \mathrm{prox}_{\tau F^*}(\mathbf{x}) + \tau\mathrm{prox}_{F/\tau}(\mathbf{x}/\tau)$ with $\mathbf{x} \in \mathbb{R}^{4P}$ and $\tau \in \mathbb{R}$. The proximity operator associated to G yields a quadratic problem whose solution is obtained by solving a linear system. We can now solve problem in (2) using the primal dual iterative algorithm proposed in [9].

Algorithm 1: Primal-dual iterative solver [9]

Compute the spectral norm $\mathcal{L}$ of the operator $\mathbf{K}$.
Set $0 \leq \theta \leq 1$, τ and σ such that $\sigma\tau\mathcal{L}^2 < 1$.
Initialize $\mathbf{w}^{(0)} \in \mathbb{R}^P$, $\mathbf{u}^{(0)} \in \mathbb{R}^P$ and $\mathbf{v}^{(0)} \in \mathbb{R}^{4P}$
repeat
- $\mathbf{v}_{k+1} = \mathrm{prox}_{\sigma F^*}(\mathbf{v}_k + \sigma\mathbf{K}\mathbf{u}_k)$
- $\mathbf{w}_{k+1} = \mathrm{prox}_{\tau G}(\mathbf{w}_k + \tau\mathbf{K}^T\mathbf{v}_k)$
- $\mathbf{u}_{k+1} = \mathbf{w}_{k+1} + \theta(\mathbf{w}_{k+1} - \mathbf{w}_k)$

until *convergence*;
return $\mathbf{w}^{k+1}$

Contrary to [7] and [5] that proposed to use two nested loops of proximal solvers (ISTA/FISTA), we have here a single loop. As the proximal operator for G leads to a linear system with the same operator to invert, the SVD factorization can be precomputed to speed up the computation.

B. Practical considerations

Two major challenges arise in the use of decoding approaches in practical setting: computation time and setting the regularization hyper-parameters.

Non regular grids: The fMRI data are not defined over the entire $P_x \times P_y \times P_z$ grid but over a mask so that $P < P_xP_yP_z$. Implementing the gradient computation on the mask leads to tedious expressions [7] that are inefficient in terms of memory access patterns. We denote by $\mathbf{\Pi} \in \mathbb{R}^{P\times P_xP_yP_z}$ the masking operator that ignores values outside of the mask. If we replace the matrix $\mathbf{X}$ by $\mathbf{X\Pi}$ so that $\mathbf{w} \in \mathbb{R}^{P_xP_yP_z}$ is defined over a full regular grid, the gradients of $\mathbf{w}$ can be obtained much more efficiently.

Parameter scanning: To speed up scanning hyper-parameter space, we leverage the convexity of the optimization problem, and use warm restarts to update a solution after changing the value of λ (both primal and dual variables need to be updated). We start with a high λ for which convergence is faster and then progressively reduce it. During K-Fold cross-validation (CV), a path is computed for each fold and for each value of ρ on grid from 0 to 1, with a step of 0.1. The same grid of ρ was used for the ElasticNet.

Setting parameters for recovery: The common practice for hyper-parameter tuning is CV. A caveat is however that CV optimizes the *prediction*, while the segmentation of predictive regions requires instead to optimize the *recovery* of the predictive variables. In order to achieve both good prediction and recovery, when using convex sparsity promoting priors, one need to compensate for the amplitude bias due to the shrinkage of the weights. Indeed, maximizing prediction score leads to choosing a small penalization to minimize the bias, which in turn leads to noisy weight maps. To alleviate this limitation we correct for the amplitude bias by rescaling the weights by a scalar value [6]: in the prediction function, we use $\hat{\mathbf{w}}_{scaled} = \kappa\hat{\mathbf{w}}$ where $\kappa = \mathbf{y}^T\mathbf{X}\hat{\mathbf{w}}/\|\mathbf{X}\hat{\mathbf{w}}\|^2$.

III. Empirical results

To investigate the performance of the TV-ℓ_1 estimator, we simulated active regions in a cube of $12 \times 12 \times 12$ voxels as in [7]. Four regions of interest of size $4 \times 4 \times 4$ voxels were positioned on corners of the cube (2 positive activations and 2 negative). We simulated 400 volumes corrupted by Gaussian smoothing ($\sigma = 2$ voxels) and added noise to the targets to be predicted with different signal-to-noise ratios (SNR). The TV-ℓ_1 estimator was compared with a precision-recall metric to a standard univariate F-test, ElasticNet, Ridge regression, regression with linear Support Vector Machines (SVR) [10], without and with z-scores [3], as well as a searchlight [2] using a linear SVR (C=1) and balls of radius 2 voxels. All estimators were tuned by 3-Fold CV over a grid of hyperparameters.

On the results in Fig. 1, one can observe that the TV-ℓ_1 estimator yields the best recovery performance, followed by the F-test and the ElasticNet. Scaling coefficients improves the recovery. The two estimators using ℓ_2 regularization, Ridge and SVR, yield overly smooth maps and fail to isolate active regions. Computing z-score for SVR maps [3] improves recovery but cannot compete with TV-ℓ_1. The searchlight leads also to a very smooth map of CV scores with overestimated predictive regions. Interestingly, the recovery performance of TV-ℓ_1 varies from 0.89 to 0.95 as SNR varies from 2.5 to 10.0, while for ElasticNet it varies from 0.71 to 0.83. The ElasticNet, that is commonly advertised for support recovery, suffers much more from poor SNR conditions than the TV-ℓ_1 model. We also quantified the prediction accuracy of the predictive models compared, and obtain on average the best performance with TV-ℓ_1, with here little impact of the scaling of $\hat{\mathbf{w}}$.

The TV-ℓ_1 estimator was also tested on the fMRI data from [4]. This data is of specific interest as decoding and univariate analysis have shown different results in terms of regions highlighted. It is a gambling task where the subject is asked to accept or reject gambles that offered a 50/50 chance of gaining or losing money. Each gamble has an amount that can be used as target in a regression setting. We refer to [4] for a detailed description of the experimental protocol. Data are publicly available on http://openfmri.org. After standard preprocessing (slice timing, motion correction, first level analysis with a general linear model, inter-subject spatial normalization), the dataset consists of 16 subjects with 48 fMRI observations per subject. For the prediction task, only the gain condition was used (see [4]): 8 levels of gain (targets y coded between 1 and 8.). FMRI volumes were downsampled to $4\times4\times4$ mm voxels. The full dataset of 16 subjects consist of 768 samples with approximately 33 000 voxels. The prediction here is inter-subject: the estimator learns on some subjects and predicts on left out subjects. Parameter estimation was performed with 5-Folds CV.

Results obtained with F-scores, ElasticNet, and TV-ℓ_1 are presented in Fig. 2. As opposed to the linear SVM [4], and as confirmed by our simulations, one can observe a good agreement between the F-test and the TV-ℓ_1 predictive model. ElasticNet succeeds in selecting some neuroscientificly meaningful voxels [4] but as expected selects too many of then (false positives) when tuned with CV. The TV-ℓ_1 model, when used with rescaling of the weights, segments neuroscientificly reasonable predictive regions, while yielding similar prediction performance as the ElasticNet. Without weight rescaling, CV underpenalizes and yields very noisy maps (not shown).

IV. Conclusion

Our contributions are the following: First we introduce a principled optimization procedure with convergence guarantees for TV-ℓ_1 regularized predictive models. Second we outline practical details that make the solver more useful for decoding applications, for example adapting cross-validation to recovery purpose with proper rescaling of the coefficients. Finally simulation results as well as experimental data demonstrate the ability of the solver to segment predictive regions in good agreement with simple F-test showing that well-employed decoding models can actually agree with univariate statistics while offering the statistical power of multivariate methods. Further work will investigate a comparison with [6] that also imposes a spatial smoothness on the weights and with [11] that addresses the support recovery problem via randomization and stability scores.

Acknowledgment

This work was supported by the ANR grant BrainPedia, ANR-10-JCJC 1408-01 and by the "FMJH Program Gaspard

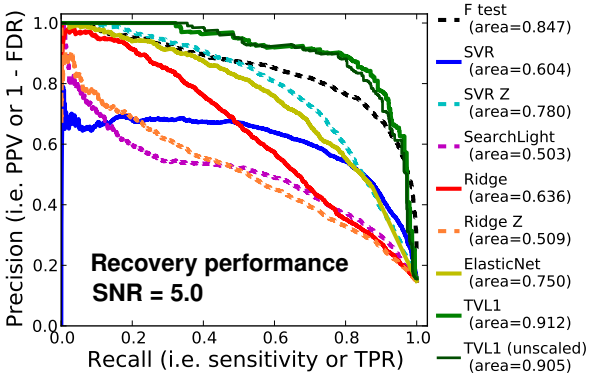

Figure 1. Performance in support recovery quantified by precision-recall (PR) as well as prediction accuracy comparisons evaluated on simulated data (two tables on the right). The TV-ℓ_1 offers both optimal prediction and recovery for all SNR values.

Prediction error (p-value to TV-ℓ_1)

SNR	2.5	5.0	7.5	10.0
SVR	28.4 (.013)	22.6 (.009)	18.1 (.005)	14.5 (.005)
Ridge	27.8 (.007)	21.6 (.005)	17.0 (.005)	13.4 (.005)
ElasticNet	**26.6** (.114)	20.7 (.050)	16.4 (.040)	13.1 (.013)
TV-ℓ_1	**25.7**	**20.1**	**15.8**	**12.5**
TV-ℓ_1 (unscaled)	**25.6** (.878)	**20.3** (.203)	16.2 (.028)	**12.8** (.059)

Recovery performance

SNR	2.5	5.0	7.5	10.0
F-score	.830	.847	.861	.873
SVR	.562	.604	.649	.691
SVR Z	.729	.780	.803	.828
SearchLight	.481	.503	.527	.550
Ridge	.543	.636	.701	.756
Ridge Z	.424	.509	.600	.668
ElasticNet	.705	.750	.791	.827
TV-ℓ_1	**.892**	**.912**	**.93**	**.946**
TV-ℓ_1 (unscaled)	.874	.905	.91	.931

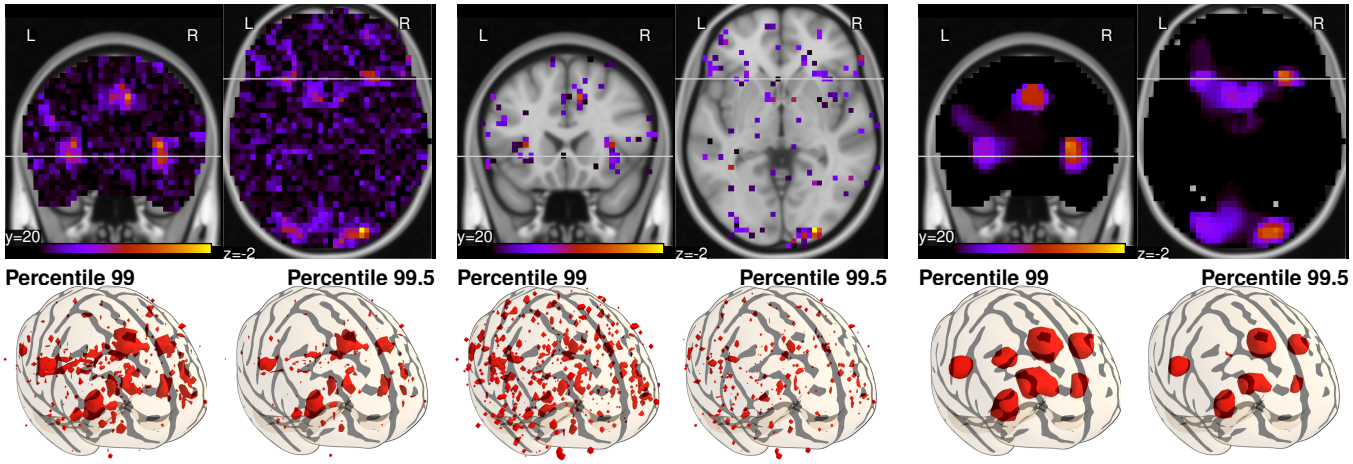

Figure 2. Results on fMRI data from [4] (from left to right F-test, ElasticNet and TV-ℓ_1). The TV-ℓ_1 regularized model segments neuroscientificly meaningful predictive regions in agreement with univariate statistics while the ElasticNet yields sparse although very scattered non-zero weights.

Monge in optimization and operation research", with the support from EDF.

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, p. 1414, 2008.

[2] N. Kriegeskorte, R. Goebel, and P. Bandettini, "Information-based functional brain mapping," *Proc Ntl Acad Sci*, vol. 103, p. 3863, 2006.

[3] B. Gaonkar and C. Davatzikos, "Deriving statistical significance maps for SVM based image classification and group comparisons," in *Proc. MICCAI conf.*, 2012, p. 723.

[4] K. Jimura and R. Poldrack, "Analyses of regional-average activation and multivoxel pattern information tell complementary stories," *Neuropsychologia*, vol. 50, p. 544, 2012.

[5] L. Baldassarre, J. Mourao-Miranda, and M. Pontil, "Structured sparsity models for brain decoding from fMRI data," in *Proc. PRNI Conf.*, 2012, p. 5.

[6] L. Grosenick, B. Klingenberg, K. Katovich, B. Knutson, and J. E. Taylor, "Interpretable whole-brain prediction analysis with graphnet," *NeuroImage*, vol. 72, p. 304, 2013.

[7] V. Michel, A. Gramfort, G. Varoquaux, E. Eger, and B. Thirion, "Total variation regularization for fMRI-based prediction of behaviour." *Trans Med Imag*, vol. 30, p. 1328, 2011.

[8] A. Gramfort, M. Kowalski, and M. Hämäläinen, "Mixed-norm estimates for the M/EEG inverse problem using accelerated gradient methods." *Phys Med Biol*, vol. 57, p. 1937, 2012.

[9] A. Chambolle and T. Pock, "A first-order primal-dual algorithm for convex problems with applications to imaging," *J. Math. Imaging Vis.*, vol. 40, p. 120, 2011.

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

[11] G. Varoquaux, A. Gramfort, and B. Thirion, "Small-sample brain mapping: sparse recovery on spatially correlated designs with randomization and clustering," in *ICML conf.*, 2012.