Predictive support recovery with TV-Elastic Net penalty and logistic regression: an application to structural MRI

Mathieu Dubois*, Fouad Hadj-Selem*, Tommy Löfstedt*, Matthieu Perrot†, Clara Fischer†,
Vincent Frouin* and Édouard Duchesnay*
* Neurospin, I2BM, CEA, Gif-sur-Yvette - France
† Centre d'Acquisition et de Traitement des Images (CATI), Gif-sur-Yvette - France
Corresponding author: edouard.duchesnay@cea.fr

***Abstract*—The use of machine-learning in neuroimaging offers new perspectives in early diagnosis and prognosis of brain diseases. Although such multivariate methods can capture complex relationships in the data, traditional approaches provide irregular (ℓ_2 penalty) or scattered (ℓ_1 penalty) predictive pattern with a very limited relevance. A penalty like Total Variation (TV) that exploits the natural 3D structure of the images can increase the spatial coherence of the weight map. However, TV penalization leads to non-smooth optimization problems that are hard to minimize. We propose an optimization framework that minimizes any combination of ℓ_1, ℓ_2, and TV penalties while preserving the exact ℓ_1 penalty. This algorithm uses Nesterov's smoothing technique to approximate the TV penalty with a smooth function such that the loss and the penalties are minimized with an exact accelerated proximal gradient algorithm. We propose an original continuation algorithm that uses successively smaller values of the smoothing parameter to reach a prescribed precision while achieving the best possible convergence rate. This algorithm can be used with other losses or penalties. The algorithm is applied on a classification problem on the ADNI dataset. We observe that the TV penalty does not necessarily improve the prediction but provides a major breakthrough in terms of support recovery of the predictive brain regions.**

I. Introduction

Multivariate machine-learning applied in neuroimaging offers new perspectives in early diagnosis and prognosis of brain diseases. However, it is essential that the method provides meaningful predictive patterns in order to reveal the neuroimaging biomarkers of the pathologies. Penalized linear models (such as linear SVM, penalized logistic regression) are often used in neuroimaging since the weight map might provide clues about biomarkers.

In particular, we are interested in penalized logistic regression in order to predict the clinical status of patients from neuroimaging data and link this prediction to known neuroanatomical structures. When using the ℓ_2 penalty with such data, the weight maps are dense and potentially irregular (i.e. with abrupt, high-frequency changes). With the ℓ_1 penalty, they are scattered and sparse with only a few voxels with non-zero weight. In both cases, the weight maps are hard to interpret in terms of neuroanatomy. The combination of both penalties in Elastic Net (see [1]), promotes sparse models while still maintaining the regularization properties of the ℓ_2 penalty.

978-1-4799-4149-0/14/$31.00 ©2014 IEEE

A major limitation of the Elastic Net penalty is that it does not take into account the spatial structure of brain images, which leads to scattered patterns.

The Total Variation (TV) penalty is widely used in 2D or 3D image processing to account for this spatial structure. In this paper, we propose to add TV to the Elastic Net penalty to improve the interpretability and the accuracy of logistic regression. We hypothesize that the predictive information is most likely organized in regions rather than scattered across the brain.

The difficulty is that ℓ_1 and TV are convex but not smooth functions (see section II for the precise definition of smoothness used in this paper). Therefore, we cannot use classic gradient descent algorithms. In [2], the authors use a primal-dual approach for ℓ_1 and TV penalties (which can be extended to include ℓ_2) but their method is not applicable to logistic regression because the proximal operator of the logistic loss is not known. Another strategy for non-smooth problems is to use methods based on the proximal operator of the penalties. For the ℓ_1 penalty alone, the proximal operator is analytically known and efficient iterative algorithms such as ISTA and FISTA are available (see [3]). However, as the proximal operator of the TV penalty is not analytically defined, those algorithms won't work in our case.

There are two general strategies to address this problem. The first one involves using an iterative algorithm to numerically approximate the proximal operator of each convex non-smooth penalty (see [4]). This algorithm is then run for each iteration of ISTA or FISTA (leading to nested optimization loops). This was done for TV alone in [5] where the authors use FISTA to approximate the proximal operator of TV. The problem with such methods is that by approximating the proximal operator we may loose the sparsity induced by the ℓ_1 penalty. The second strategy is to approximate the non-smooth penalties for which the proximal operator is not known (e.g. TV) with a smooth function (of which the gradient is known). Non-smooth penalties with a known proximal operator (e.g. ℓ_1) are not changed. Therefore it is possible to use an exact accelerated proximal gradient algorithm. Such a smoothing technique has been proposed by Nesterov in [6].

We choose to apply the second strategy. We will present an algorithm able to solve TV-Elastic Net penalized logistic

regression with exact ℓ_1 penalty and evaluate it on the prediction of the clinical status of patients from structural magnetic resonance imaging (MRI) scans. The paper is organized as follows: we present the minimization problem and our algorithm in section II, the experimental dataset is described in section III, and section IV presents the classification rates and weight maps. Finally, we conclude in section V.

II. Method

We first detail the notations of the problem. Then we develop the TV regularization framework. Finally, we detail the algorithm used to solve the minimization problem.

A. Problem statement

We place ourselves in the context of logistic regression models. Let $X \in \mathbb{R}^{n \times p}$ be a matrix of n samples, where each sample lies in a p-dimensional space and let $y \in \{0,1\}^n$ denote the n-dimensional response vector. In the logistic regression model the conditional probability of y_i given the data X_i is defined through a linear function of the unknown predictors $\beta \in \mathbb{R}^p$ by

$$p_i := p(y_i = 1|X_i) = \frac{1}{1+\exp(-X_i^T\beta)},$$

and $p(y_i = 0|X_i) = 1 - p_i$. Therefore, looking for the maximum of the log-likelihood with structured and sparse penalties, we consider the following minimization problem of a logistic regression objective function with Elastic Net and TV penalties:

$$\beta^* := \arg\min_{\beta \in \mathbb{R}^p} f(\beta), \tag{1}$$

where $f(\beta)$ is the sum of a smooth part, $g(\beta)$, and of a non-smooth part, $h(\beta)$, such that

$$\begin{aligned} f(\beta) \quad &:= \quad \underbrace{\frac{1}{n}\sum_{i=1}^{n}\left\{y_i X_i\beta - \log\left[1+\exp(X_i\beta)\right]\right\} + \lambda_{\ell_2}\left\|\beta\right\|_2^2}_{g(\beta)} \\ &+ \quad \underbrace{\lambda_{\ell_1}\left\|\beta\right\|_1 + \lambda_{TV} TV(\beta)}_{h(\beta)}, \end{aligned} \tag{2}$$

where λ_{ℓ_2}, λ_{TV} and λ_{ℓ_1} are constants that control the relative strength of each penalty. In this context, a function is said to be smooth if it is differentiable everywhere and its gradient is Lipschitz-continuous.

Given a 3D image I of size (p_x, p_y, p_z), TV is defined as

$$TV(I) = \sum_{(i,j,k)} \left\|\mathrm{grad}_{i,j,k}(I)\right\|_2 \tag{3}$$

where $\mathrm{grad}_{i,j,k}(I) \in \mathbb{R}^3$ is the numerical gradient of I at coordinates (i,j,k) and the sum runs over all voxels of I.

In our case, rows of X are composed of masked and flattened 3D images arranged into vectors of size $p < p_x \times p_y \times p_z$. Similarly, the vector β belongs to $\mathbb{R}^p$. For now on each voxel is identified by its linear index in X, noted i ($1 \leq i \leq p$). Special care must be taken for the computation of the gradient on the flattened vector β, because, due to the existence of a mask and border conditions, not all the neighbors of a voxel i exist in the data. Given this precaution, we can compute the gradient for each β_i and then compute $TV(\beta)$. More details regarding the TV penalty in the context of 3D image analysis can be found in [5].

B. Regularization framework

A sufficient condition for the application of Nesterov's smoothing technique to a given convex function s is that it can be written on the form

$$s(\beta) = \max_{\boldsymbol{\alpha} \in K_s} \langle \boldsymbol{\alpha} \,|\, A_s\beta \rangle, \tag{4}$$

for all $\beta \in \mathbb{R}^p$, with K a compact convex set in a finite-dimensional vector space and A_s a linear operator between two finite-dimensional vector spaces.

In [7] the authors show that $TV(\beta)$ can be written as

$$TV(\beta) = \sum_{i=1}^{p} \max_{\alpha_i \in K_i} \langle \alpha_i | A_i\beta \rangle$$

where $K_i = \left\{\alpha \in \mathbb{R}^3, \|\alpha\|_2^2 \leq 1\right\}$ and A_i is a sparse matrix that allows to compute the gradient at position i (A_i depends on the mask M). This can be further written as

$$TV(\beta) = \max_{\alpha \in K} \langle \alpha | A\beta \rangle$$

where α is the concatenation of all the α_i, A is the vertical concatenation of all the A_i matrices and K is the product of all the compact convex spaces K_i (as such, K is itself a compact convex space). Note that K and A are specific to TV.

Given this expression for TV, we can apply Nesterov's smoothing. For a given smoothing parameter $\mu > 0$, TV is approximated by the smooth function

$$TV_\mu(\beta) = \max_{\alpha \in K} \left\{ \langle \alpha | A\beta \rangle - \frac{\mu}{2}\|\alpha\|_2^2 \right\}. \tag{5}$$

The value that maximizes Equation 5 is

$$\boldsymbol{\alpha}_\mu^*(\beta) = \mathrm{proj}_K\left(\frac{A\beta}{\mu}\right)$$

The function TV_μ is convex and differentiable. Its gradient can be written (see [6]) as

$$\nabla TV_\mu(\beta) = A^\top \boldsymbol{\alpha}_\mu^*(\beta).$$

The gradient is Lipschitz continuous with Lipschitz constant

$$\frac{\|A\|_2^2}{\mu},$$

where $\|A\|_2$ is the matrix spectral norm of A.

C. Algorithm

A new optimization problem, closely related to problem 1, arises from this regularization:

$$\beta_\mu^* := \arg\min_{\beta \in \mathbb{R}^p} f_\mu(\beta) \tag{6}$$

where

$$f_\mu(\beta) := \underbrace{g(\beta) + \lambda_{TV} TV_\mu(\beta)}_{\text{smooth}} + \underbrace{\lambda_{\ell_1}\left\|\beta\right\|_1}_{\text{non-smooth}}. \tag{7}$$

β^*_μ approximates β^*, the solution to the original problem 1, since $\|f_\mu - f\| \leq \frac{\mu p}{2}$.

Since we are now able to explicitly calculate the gradient of the smooth part, its Lipschitz constant and the proximal operator of the non-smooth part, this new problem can be solved by FISTA [3]. The convergence rate of FISTA is governed by

$$f_\mu(\beta^{(k)}) - f_\mu(\beta^*_\mu) \leq \frac{2}{t_\mu(k+1)^2}\|\beta^{(0)} - \beta^*_\mu\|_2^2, \qquad (8)$$

where $k \geq 1$ is the iteration number and t_μ is the step size that must be chosen smaller than or equal to the inverse of the known Lipschitz constant of the gradient of the smooth part. Note that the convergence depends on the initial value $\beta^{(0)}$.

If μ is small the algorithm will converge with a high precision (i.e. β^*_μ will be close to β^*) but in this case it will converge slowly (because small μ leads to small t_μ). Thus, there is a trade-off between speed and accuracy. We therefore propose to perform successive runs of FISTA with decreasing values of the smoothing parameter (to increase precision) but using the regression vector obtained at the previous run as a starting point for FISTA to increase convergence speed. We denote $\beta^{(i)}$ the regression vector after the ith run of FISTA.

The key point is how to derive the sequence of smoothing parameter $\mu^{(i)}$. Our approach involves two steps. First, we describe how to obtain a value of the smoothing parameter $\mu_{opt}(\varepsilon)$ that minimizes the number of iterations needed to achieve a prescribed precision $\varepsilon > 0$ when minimising 1 via 6 (i.e. such that $f(\beta^{(k)}) - f(\beta^*) < \varepsilon$). Next, given a predefined sequence $\varepsilon^{(i)}$ of decreasing precision values, we can define a continuation sequence of smoothing parameters such that $\mu^{(i)} = \mu_{opt}(\varepsilon^{(i)})$. Concerning the first point we can prove that for any given $\varepsilon > 0$, selecting the smoothing parameter as

$$\mu_{opt}(\varepsilon) = \frac{-\lambda_{TV}\|A\|_2^2}{L_0} + \frac{\sqrt{(\lambda_{TV}M\|A\|_2^2)^2 + \varepsilon M L_0 \|A\|_2^2}}{ML_0}$$

where $M = p/2$ and L_0 is the Lipschitz constant of $\nabla(g)$ (following [5], we have $L_0 = 2\lambda_{\ell_2} + \|A\|_2/(4n)$) minimizes the worst case bound on the number of iterations needed to achieve the precision ε when minimizing 1 via 6. The proof is inspired by the proof of Lemma 3 in [8]. In this article, we use a fixed sequence of precision $\varepsilon^{(i)} = (1/2)^{i-1}$. The only parameter of the algorithm is then the initial point β^0. In these experiments, we used a random vector with a unit norm.

We call this algorithm CONESTA (for **CO**ntinuation with **NE**sterov smoothing in a **S**hrinkage-**T**hresholding **A**lgorithm). The algorithm is presented in Algorithm 1. The convergence proof will be presented in an upcoming paper. We denote the total number of FISTA loops used in CONESTA by K. We have experimentally verified that the convergence rate to the solution of problem 1 is $O(1/K^2)$ (which is the optimal convergence rate). Also, the algorithm works even if some of the weights λ_{ℓ_1}, λ_{ℓ_2} or λ_{TV} are zero, which thus allows us to solve e.g. the Elastic Net or pure lasso using CONESTA.

Algorithm 1 CONESTA

Require: β^0, the initial regression vector.
1: $i = 1$
2: **repeat**
3: $\quad\epsilon^{(i)} \leftarrow (1/2)^{i-1}$
4: $\quad\mu^{(i)} \leftarrow \mu_{opt}(\epsilon^{(i)})$
5: $\quad\beta^{(i)} \leftarrow FISTA(\beta^{(i-1)}, \mu_i)$
6: $\quad i = i + 1$
7: **until** Convergence

III. Dataset

The data used in the preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (http://adni.loni.usc.edu/). The MR scans are T1-weighted MR image acquired at 1.5 T according to the ADNI acquisition protocol (see [9]). The image dimensions were $p_x = 121$, $p_y = 145$, $p_z = 121$. The 510 T1-weighted MR images were segmented into GM (Gray Matter), WM (White Matter) and CSF (Cerebrospinal Fluid) using the SPM8 unified segmentation routine [10]. 456 images were retained after quality control on GM probability. These images were spatially normalized using DARTEL [11] without any spatial smoothing. From the 456 registered images we use only the 148 control (CTL) subjects and the 122 Alzheimer's Disease (AD) subjects. Thus, the total number of images was $n = 270$. A brain mask was obtained by thresholding the modulated gray matter map, leading to the selection of $p = 311{,}341$ voxels. According to the assignments found in [12], those 270 images were split into 132 training images, used in the learning phase, and 138 images used to test the algorithms.

IV. Experimental results

Table I presents the prediction results obtained on the test samples. It shows that using the ℓ_1 penalty alone decreases the predictive performance. We suspect that the ℓ_1 penalty is inefficient in recovering the predictive support on non-smoothed images. The TV penalty does not significantly increase nor decrease the performances except when it is combined with the ℓ_1 penalty.

Figure 1 demonstrates that the TV penalty provides a major breakthrough in terms of support recovery of the predictive brain regions. Conversely to the ℓ_2 penalty that highlights an irregular and meaningless pattern, $\ell_2 + TV$ provides a smooth map that match the well-known brain regions involved in AD [13]. A large region of negative weights was found in the temporal lobe. This region includes the superior and middle temporal gyri, the parahippocampal gyrus and the entorhinal cortex, the fusiform gyrus, the amygdala, the insula and the hippocampus. As expected, this pattern was predominantly found on the left hemisphere. The bi-lateral ventricular enlargement is sharply identified, the surprising positive sign of the weights is explained in Figure 1. Atrophy in the frontal lobe (inferior frontal gyrus) was found. Positive weights within the whole cingulum region reflect tissue shift due to periventricular atrophy. In the occipital lobe, positive weights were observed within the calcarine fissure and the cuneus.

As hypothesized, the combination of the $\ell_1 + \ell_2$ penalties provides scattered patterns with a very limited relevance.

Finally, $\ell_1 + \ell_2 + TV$ provides a summary of the $\ell_2 + TV$ pattern: most of the identified regions are are the same as when using $\ell_2 + TV$ but with limited extent. For example,

Table I. PREDICTION ACCURACIES. SENSITIVITY (SENS.: RECALL RATE OF AD PATIENTS), SPECIFICITY (SPEC.: RECALL RATE OF CTL SUBJECTS), BCR (BALANCED CLASSIFICATION RATE) AND MCNEMAR'S COMPARISON TEST p-VALUE AGAINST ANOTHER METHOD. ALL PREDICTION RATES WERE SIGNIFICANT EXCEPT THOSE OBTAINED WITH THE ℓ_1 METHOD.

Method	$\lambda_{\ell_2}, \lambda_{\ell_1}, \lambda_{TV}$	Sens.	Spec.	BCR	Comp. p-value
ℓ_2	$1.0, 0.0, 0.0$	0.855	0.855	0.855	-
ℓ_1	$0.0, 1.0, 0.0$	0.684	0.484	0.584	-
$\ell_2 + \ell_1$	$0.9, 0.1, 0.0$	0.802	0.742	0.772	-
$\ell_2 + TV$	$0.1, 0.0, 0.9$	0.842	0.726	0.784	0.16 to ℓ_2
$\ell_1 + TV$	$0.0, 0.1, 0.9$	0.829	0.774	0.801	**2e-4** to ℓ_1
$\ell_2 + \ell_1 + TV$	$0.1, 0.1, 0.8$	0.815	0.758	0.787	1 to $\ell_2 + \ell_1$

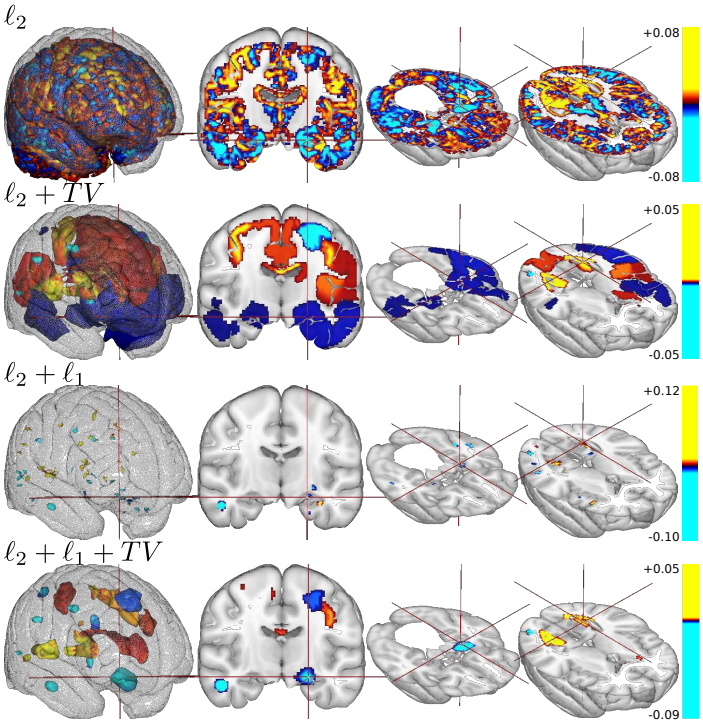


Figure 1. Weight maps: positive/negative values indicate the way regions contribute to predict the AD status. It should generally be interpreted as a increase/decrease of GM in the AD group. Positive weights (increase of GM in AD) may be found where negative weights are expected. For example, positive weights surround the whole bi-lateral ventricles. We hypothesize that we observe the negative pattern of an underlying global atrophy: the GM surrounding the ventricles shift away from them thus we observe GM in AD patients where controls have WM tissue. The map obtained with $\ell_1 + TV$ has been omitted since it provides similar results as those found with $\ell_1 + \ell_2 + TV$. The map obtained with ℓ_1 alone has no relevance.

the whole temporal atrophy found by $\ell_2 + TV$ is now limited to the hippocampus. Noticeably, the right hippocampus is no longer a predictive region due to the property of the ℓ_1 penalty. This suggests that sparse patterns should be considered with caution.

V. CONCLUSION

We proposed an optimization algorithm that is able to minimize any combination of the ℓ_1, ℓ_2, and TV penalties while preserving the exact ℓ_1 penalty. This algorithm uses Nesterov's technique to smooth the TV penalty such that objective function is minimized with an exact accelerated proximal gradient algorithm. The approximation of TV is controlled by a single smoothing parameter μ . Our contribution was to propose a continuation algorithm with successively smaller values of μ to reach a prescribed precision while achieving the best possible convergence rate. Average execution time is one hour on a standard workstation involving 13,000 FISTA iterations.

We observed that by adding the TV penalty, the prediction does not necessarily improve. However, we demonstrated that it provides a major breakthrough in terms of support recovery of the predictive brain regions.

It should be noted that the algorithm can be extended to minimize any differentiable loss (logistic, least square) with any combination of ℓ_1, ℓ_2 penalties and with any non-smooth penalty that can be written in the form of Equation 4. This includes Group Lasso and Fused Lasso or any penalty that can be expressed as a p-norm of a linear operation on the weight map.

ACKNOWLEDGEMENT

This work was partially funded by grants from the French National Research Agency ANR BRAINOMICS (ANR-10-BINF-04) and from the European Commission MESCOG (FP6 ERA-NET NEURON 01 EW1207).

REFERENCES

[1] J. Friedman, T. Hastie, and R. Tibshirani, "Regularization paths for generalized linear models via coordinate descent," *Journal of statistical software*, vol. 33, no. 1, p. 1, 2010.

[2] A. Gramfort, B. Thirion, and G. Varoquaux, "Identifying predictive regions from fMRI with TV-L1 prior," in *Proceedings of the 2013 International Workshop on Pattern Recognition in Neuroimaging*, ser. PRNI '13. Washington, DC, USA: IEEE Computer Society, 2013, pp. 17–20.

[3] A. Beck and M. Teboulle, "A Fast Iterative Shrinkage-Thresholding Algorithm for Linear Inverse Problems," *SIAM Journal on Imaging Sciences*, vol. 2, no. 1, pp. 183–202, Jan. 2009.

[4] M. Schmidt, N. Le Roux, and F. Bach, "Convergence Rates of Inexact Proximal-Gradient Methods for Convex Optimization," in *NIPS'11*, Dec. 2011.

[5] V. Michel, A. Gramfort, G. Varoquaux *et al.*, "Total Variation Regularization for fMRI-Based Prediction of Behavior," *IEEE Transactions on Medical Imaging*, vol. 30, no. 7, pp. 1328–1340, 2011.

[6] Y. Nesterov, "Smooth minimization of non-smooth functions," *Mathematical Programming*, vol. 103, no. 1, pp. 127–152, 2005.

[7] T. Löfstedt, V. Guillemot, V. Frouin *et al.*, "Simulated Data for Linear Regression with Structured and Sparse Penalties," Jan. 2014.

[8] B. Savchynskyy, S. Schmidt, J. Kappes *et al.*, "A study of Nesterov's scheme for Lagrangian decomposition and MAP labeling," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011, pp. 1817–1823.

[9] C. R. Jack, M. A. Bernstein, N. C. Fox *et al.*, "The Alzheimer's Disease Neuroimaging Initiative (ADNI): MRI methods." *J Magn Reson Imaging*, vol. 27, no. 4, pp. 685–691, Apr 2008.

[10] J. Ashburner and K. J. Friston, "Unified segmentation." *Neuroimage*, vol. 26, no. 3, pp. 839–851, Jul 2005.

[11] J. Ashburner, "A fast diffeomorphic image registration algorithm." *Neuroimage*, vol. 38, no. 1, pp. 95–113, Oct 2007.

[12] R. Cuingnet, E. Gerardin, J. Tessieras *et al.*, "Automatic classification of patients with Alzheimer's disease from structural MRI: a comparison of ten methods using the ADNI database." *Neuroimage*, vol. 56, no. 2, pp. 766–781, May 2011.

[13] G. B. Frisoni, N. C. Fox, C. R. Jack *et al.*, "The clinical use of structural MRI in Alzheimer disease." *Nat Rev Neurol*, vol. 6, no. 2, pp. 67–77, Feb 2010.