Structured Sparsity Models for Brain Decoding from fMRI data

Luca Baldassarre, Janaina Mourão-Miranda and Massimiliano Pontil
Department of Computer Science
University College London
London, UK
{*l.baldassarre, j.mourao-miranda, m.pontil*}*@cs.ucl.ac.uk*

***Abstract*—Structured sparsity methods have been recently proposed that allow to incorporate additional spatial and temporal information for estimating models for decoding mental states from fMRI data. These methods carry the promise of being more interpretable than simpler Lasso or Elastic Net methods. However, despite sparsity has often been advocated as leading to more interpretable models, we show that by itself sparsity and also structured sparsity could lead to unstable models.**

We present an extension of the Total Variation method and assess several other structured sparsity models on accuracy, sparsity and stability. Our results indicate that structured sparsity via the Sparse Total Variation can mitigate some of the instability inherent in simpler sparse methods, but more research is required to build methods that can reliably infer relevant activation patterns from fMRI data.

***Keywords*-brain decoding; structured sparsity; stability; fMRI**

I. Introduction

Supervised machine learning techniques are being increasingly used in the analysis of brain imaging data for their inherent ability to deal with multi-variate data, higher sensibility and possibility of incorporating specific prior-information.

Given the high-dimensionality of neuroimaging, and especially fMRI, data and the few number of samples, linear models have been proven to be sufficient in order produce effective classifiers [1], [2], [3], [4].

However, ordinary linear models, for example Least Squares or Ridge Regression [5], are incapable of discriminating which areas of the brain mostly contribute to the model's predictions, in the sense that all voxels contribute to generate a predictive function.

Sparse methods, like the Lasso [6] or the Elastic Net [7], are able to estimate solutions for which only few voxels are deemed relevant, therefore aiding interpretation. However, often these models provide overly sparse solutions, or activation patterns, where the non-zero coefficients are assigned to disparate regions across the brain, without exploiting any spatial or temporal prior information [3], [4], [8].

Recently, structured sparsity models [9], [10] have been proposed to extend the well-known methods of Lasso and Elastic Net by enforcing more structured constraints on the solution. These include constancy or closeness of the regression coefficients over neighbouring or connected regions or graph structures.

Despite sparsity has traditionally been connected with interpretability, in the sense that sparser models are easier to interpret, these new structured sparsity models promise an even greater ease of interpretation of the activation patterns, because the active voxels are grouped together in possibly few clusters, which fits well with our knowledge about the brain's specialized regions and networks. However, sparsity alone is not sufficient for making reasonable inferences from these models, because a sparse model could be unstable under resampling or slight changes of the experimental conditions. Therefore, we advocate stability as the natural counterpart of sparsity in order to obtain interpretable inferences from sparse supervised learning methods.

In this paper, we assess several structured sparsity methods that have been recently used for decoding fMRI data and assess their performance with respect to accuracy, sparsity and stability. The methods we consider include Lasso [6] and Elastic Net [7], Total Variation [3], Graph Laplacian Elastic Net (GraphNET) [4] and an extension of the Total Variation method which, up our knowledge, is applied to fMRI data for the first time.

For our comparison we use a dataset of fMRI scans collected from 16 healthy volunteers while watching pleasant or unpleasant images in a block experimental design [1], [2], [11]. We discuss the relevance of our findings with respect to using classification accuracy as a proxy for statistical significance of a given model.

II. Supervised Learning for Classification

Given a training set of input-output pairs $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^m$, with $x_i \in \mathbb{R}^p$ and $y_i \in \mathbb{R}$, a supervised learning method infers the relationship between x and y by estimating a prediction function $f : \mathbb{R}^p \to \mathbb{R}$ such that, for every $x \in \mathbb{R}^p$, $f(x)$ provides the prediction of y given x.

For neuroimaging studies, the x_i represent the brain scans in vector format and the number of variables p corresponds to the number of recorded voxels. In the present paper we consider a binary classification task, so that $y \in \{-1, 1\}$, but our results can easily be

extended to the regression or the multi-class setting. Furthermore, we limit our analysis to linear models, so that the decision function can be written as $f(x) = \text{sign}(x^T\beta)$, where $\beta \in \mathbb{R}^p$ is a vector of coefficients to be estimated, one associated to each voxel.

The aim of a machine learning algorithm is to find a coefficient vector β able to correctly classify new examples and with specific properties such as sparsity (i.e. few non-zero coefficients) or smoothness. Regularized methods find β minimizing an objective function consisting of a data fit term $I(\beta)$ and a penalty term $\Omega(\beta)$ that favours certain properties and improves the generalization over unseen examples (outside the training set $\mathcal{D}$).

As data fit term we consider the square loss that can be concisely written as

$$I(\beta) = \frac{1}{m}\|X\beta - Y\|_2^2$$

where $X \in \mathbb{R}^{m\times p}$ is the matrix that contains the training examples as rows and $Y = (y_1, \ldots, y_m)^T$.

A. Structured Sparsity Models

Note that, since for a linear model each regression coefficient is associated to a voxel, the vector β can also be interpreted as 3D matrix of the same size as the brain scans and we use this 3D structure to define particular penalty terms $\Omega(\beta)$. We define the ℓ_1 norm of β as $\|\beta\|_1 = \sum_{i=1}^n |\beta_i|$; the discrete gradient of β in 3 dimensions as $\nabla\beta$, with

$$\begin{aligned}(\nabla\beta)^1_{i,j,k} &= \beta(i,j,k) - \beta(i-1,j,k)\\ (\nabla\beta)^2_{i,j,k} &= \beta(i,j,k) - \beta(i,j-1,k)\\ (\nabla\beta)^3_{i,j,k} &= \beta(i,j,k) - \beta(i,j,k-1)\end{aligned}$$

and $(\nabla\beta)^\ell_{i,j,k} = 0$ if (i,j,k) is on the boundary w.r.t. the direction ℓ. Finally, $\sum_{i\sim j}(\beta_i - \beta_j)^2$ means that the sum is only for neighbouring voxels i and j.

For each method, the estimated model $\hat{\beta}$ is the minimizer of the functional $I(\beta)+\Omega(\beta)$, where $\Omega(\beta)$ is defined as follows.

Elastic Net and Lasso:

$$\Omega(\beta) := \lambda_1||\beta||_1 + \lambda_2\|\beta\|_2^2 \ .$$

This method favours coefficients vectors that have few non-zero components whose location is not constrained in any manner. For $\lambda_2 = 0$, we obtain the Lasso, while $\lambda_2 \neq 0$ allows for correlated features to be selected together.

Total Variation:

$$\Omega(\beta) := \lambda||\nabla\beta||_1 \ .$$

This method favours solutions that have constant value in contiguous regions and has its origins in image de-noising applications [12], however it does not enforce any coefficient to be exactly zero.

Sparse Total Variation:

$$\Omega(\beta) := \lambda\left(\|\nabla\beta\|_1 + \|\beta\|_1\right) \ .$$

By adding a ℓ_1-penalty term to the Total Variation functional, this method favours solutions whose coefficients are constant within contiguous regions, but also promotes sparsity. This hybrid method has been proposed in other domains, such as image de-noising using Fourier or wavelet representations (e.g. [13]), and, up to our knowledge, this is its first application to brain decoding.

Sparse Laplacian (SLAP):

$$\Omega(\beta) := \lambda(1-\alpha)\sum_{i\sim j}(\beta_i - \beta_j)^2 + \lambda\alpha||\beta||_1 \ .$$

This method relaxes the constancy requirement of the Total Variation method, allowing for smooth variations within regions, but it still enforces sparsity. It is equivalent to the GraphNET model [4], with $\lambda_1 = \lambda\alpha$ and $\lambda_G = \lambda(1-\alpha)$.

In all cases, λ and α are regularization parameters that control the trade-off between fitting the training data and minimizing the penalties. These parameters must be chosen in an unbiased way during learning.

We compare these methods to two least squares models: one trained on all variables and one trained only on 10% of the variables selected via thresholding the p-values obtained by performing a two-sample t-test for each variable with respect to the two classes.

B. Optimization

In order to estimate the solutions for each method, given the high-dimensionality of the problem ($p \approx 10^5$ in the following experiments), we implemented in MATLAB accelerated proximal methods [14], which are a form of gradient-descent that can deal with non-smooth functions and scale nicely to large problem sizes. They rely on the computation of the gradient of the smooth data-fit term, $I(\beta)$, and on the computation of the proximity operator [15] associated to $\Omega(\beta)$. For the Lasso and the SLAP methods, this operator can be computed analytically, while for the methods that employ the Total Variation penalty, we use a recently proposed efficient algorithm [16]. The efficacy of accelerated proximal methods when the proximity operator is computed numerically has been studied in [17], [18]. We use the weaker requirements in [17] on the decay of the errors.

C. Experimental Protocol and Assessment

We performed two nested loops of Leave-One-Subject-Out Cross-Validation (LOSO-CV): the external loop is used for assessing the classification accuracy, the sparsity and the stability of the methods; the internal loop is used for selecting the model hyper-parameters (λ and α). Hence, for each method, we

train N different models, where N is the number of subjects in the dataset.

A recent work [8] studies the impact of model selection on the reproducibility and stability of the estimated models for simpler learning methods. In the present work, we select the hyper-parameters in the usual way of maximizing the classification accuracy over the internal LOSO-CV and we assess the stability of the resulting models in the external LOSO-CV. Our comparison is focused on the effects of structured sparsity on the stability and not on the particular method of model selection, which might still have an impact and will be subject to future research.

Due to the large number of voxels, the optimization algorithms were slow to reach convergence. We stopped the iterations when the relative decrease in the objective function was smaller than 10^{-3}. This choice had the impact that some of the estimated coefficients have not been set exactly to zero. Therefore we adopted the heuristic of setting to zero the smallest components of the regression vector which contribute only 1% to the $\|\beta\|_1$. We also applied this thresholding to the non-sparse methods in order to assess their stability.

Let $\beta(s)$ be the coefficient vector estimated when the data for subject s is left out for testing. We define the model support $I_s := \{i|\beta(s)_i \neq 0\}$ as the index set of the location of the non-zero coefficients, the model sparsity $S(s) := \frac{|I_s|}{p}$ as the relative number of non-zero coefficients and the corrected pairwise relative overlap as

$$O_{s,s'} := \frac{|I_s \cap I_{s'}| - E}{\max(|I_s|, |I'_s|)} ,$$

where E is the expected overlap between the support of two random vectors with sparsity $S(s)$ and $S(s')$ respectively, given by the formula

$$E = \frac{S(s)S(s')}{p} .$$

We use the average corrected pairwise overlap $\overline{O} := \frac{1}{N(N-1)} \sum_{s \neq s'=1}^{N} O_{s,s'}$ as a measure of stability.

The accuracy is the average percentage of correctly classified examples over all the LOSO folds, namely

$$\text{Accuracy} = \frac{1}{N} \sum_{s=1}^{S} \frac{1}{m_s} \sum_{i=1}^{m_s} \delta(f_s(x_i) = y_i)$$

where $f_s(x_i) = \text{sign}(\beta(s)^T x_i)$ and m_s is the number of examples for subject s.

III. Experiments

A. Dataset

We used fMRI data from 16 male healthy US college students (age $20 - 25$) [1], [2], [11]. Participants did not have any history of neurological or psychiatric illness and had normal vision.

The fMRI data were acquired on a 3T Allegra Head-only MRI system, using a T2* sequence with 43 axial slices (slice thickness, 3mm; gap between slices, 0mm; TR=3sec; TE=30ms; FA=$80°$; FOV=192×192mm; matrix, 64×64; voxel dimensions, $3 \times 3 \times 3$ mm). Stimuli were presented in a blocked fashion. There were two different active conditions: viewing unpleasant (dermatological diseases) and pleasant images (pretty women in swimsuits), and a control condition (fixation). Each run comprised six blocks of the active condition (each consisting of 7 images volumes) alternating with fixation control blocks (of 7 images volumes). Blocks of each of the two stimuli classes were presented in random order.

The data were pre-processed using SPM2[1]. All the scans were realigned to remove residual motion effects and transformed into standard space [19]. The data were de-trended and smoothed in space using an 8mm Gaussian filter. Finally, a mask was applied to select voxels which contain brain tissue according to the SPM template, excluding the eyeballs. The dataset consists of 1344 scans of size 219727 voxels, with 42 scans per subject per active condition.

B. Results

We applied the protocol described in Sect.II-C to all the considered methods. Table I reports the average and standard deviation of the accuracy, sparsity and stability measures computed for all the methods on the external LOSO-CV. Using the simple two-sample t-test for selecting a subset of features leads to poorer classification accuracy, especially compared to sparse models with the same level of sparsity. From the table it is evident that all other methods achieve almost the same accuracy, albeit with different sparsity and stability performances. We note that the non-sparse methods (Least Squares, Total Variation and Laplacian without ℓ_1 term) are also the less stable ones, once correcting for the expected overlap. After thresholding, their sparsity is around 44%, meaning that most of the magnitude of the coefficients is concentrated in less than half of the voxels. We also observe that sparser models do not necessarily have a higher stability, in fact it is difficult to see any direct correlation between sparsity and stability. The Sparse Total Variation is, among the sparse models, the one that achieves the highest stability with smaller standard deviation and therefore could lead to more interpretable solutions.

IV. Conclusion

Sparsity has often been advocated as a proxy for interpretability, however we show that sparsity in itself could produce highly unstable models. We investigated the effect of using structured sparsity methods

[1]Wellcome Department of Imaging Neuroscience, http://www.fil.ion.ucl.ac.uk/spm/

Table I
COMPARISON OF THE DIFFERENT METHODS.

Method	Accuracy	Sparsity	Stability
Least Squares	$83.0 \pm 5.9\%$	$44 \pm 1\%$	$41 \pm 1\%$
T-Test (10%) + LS	$78.6 \pm 5.8\%$	$4.43 \pm 0.02\%$	62 ± 3
Lasso	$85.8 \pm 6.6\%$	$6.4 \pm 1.2\%$	$64 \pm 15\%$
Elastic Net	$85.9 \pm 6.8\%$	$44.4 \pm 0.2\%$	$39 \pm 9\%$
TV	$85.0 \pm 6.4\%$	$42 \pm 2\%$	$34 \pm 19\%$
Sparse TV	$87.4 \pm 6.2\%$	$9.4 \pm 0.4\%$	$71 \pm 3\%$
Laplacian ($\alpha = 0$)	$83.2 \pm 5.7\%$	$44.2 \pm 0.1\%$	$40 \pm 1\%$
SLAP	$85.5 \pm 6.2\%$	$7 \pm 10\%$	$52 \pm 22\%$

on stability on a dataset of fMRI scans of very high dimensionality. We found that the methods perform similarly with respect to classification performance, but the resulting models differ in term of sparsity and stability. The proposed Sparse Total Variation seems to produce the most stable sparse model, an indication that structured sparsity could alleviate some of the instability inherent in the non-structured methods such as Lasso or Elastic Net.

However, it is necessary to study the impact of model selection (i.e. choosing the parameters λ and α) on stability and interpretability for these methods. This interesting question will be a subject of further investigations for structured sparsity models. Furthermore, we aim at improving the optimization techniques, allowing us to eschew from thresholding, which might further negatively affect stability.

Other interesting research directions regard the use of even more structured priors, like functional connectivity between regions and stability assessments at different scales.

ACKNOWLEDGMENT

Part of this work was supported by EPSRC Grant EP/H027203/1 and Royal Society International Joint Project Grant 2012/R2. JMM was funded by a Wellcome Trust Career Development Fellowship under grant no. WT086565/Z/08/Z.

REFERENCES

[1] J. Mourão-Miranda, E. Reynaud, F. McGlone, G. Calvert, and M. Brammer, "The impact of temporal compression and space selection on svm analysis of single-subject and multi-subject fmri data," *NeuroImage*, vol. 33, no. 4, pp. 1055–1065, 2006.

[2] J. Mourão-Miranda, K. Friston, and M. Brammer, "Dynamic discrimination analysis: A spatial-temporal svm," *NeuroImage*, vol. 36, no. 1, pp. 88–99, 2007.

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

[4] L. Grosenick, B. Klingenberg, B. Knutson, and J. E. Taylor, "A family of interpretable multivariate models for regression and classification of whole-brain fmri data," *ArXiv e-prints*, Oct. 2011.

[5] A. N. Tikhonov and V. Y. Arsenin, *Solutions of Ill-posed Problems*. John Wiley, 1977.

[6] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society, Series B*, vol. 58, pp. 267–288, 1996.

[7] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *Journal of the Royal Statistical Society: Series B*, vol. 67, no. 2, pp. 301–320, 2005.

[8] P. Rasmussen, L. Hansen, K. Madsen, N. Churchill, and S. Strother, "Model sparsity and brain pattern interpretation of classification models in neuroimaging," *Pattern Recognition*, vol. 45, pp. 2085–2100, 2012.

[9] A. Chambolle, "An algorithm for total variation minimization and applications," *J. Math. Imaging Vis.*, vol. 20, pp. 89–97, 2004.

[10] F. Bach, R. Jenatton, J. Mairal, and G. Obozinski, "Structured sparsity through convex optimization." Arxiv preprint:1109.2397, Tech. Rep., 2011.

[11] D. Hardoon, J. Mourão-Miranda, M. Brammer, and J. Shawe-Taylor, "Unsupervised analysis of fmri data using kernel canonical correlation," *NeuroImage*, vol. 37, no. 4, pp. 1250–1259, 2007.

[12] L. Rudin, S. Osher, and E. Fatemi, "Nonlinear total variation based noise removal algorithms," *Physica D: Nonlinear Phenomena*, vol. 60, no. 1-4, pp. 259–268, 1992.

[13] S. Ma, W. Yin, Y. Zhang, and A. Chakraborty, "An efficient algorithm for compressed mr imaging using total variation and wavelets," in *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*. IEEE, 2008, pp. 1–8.

[14] 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, 2009.

[15] J. Moreau, "Fonctions convexes duales et points proximaux dans un espace hilbertien," *CR Acad. Sci. Paris Sér. A Math*, vol. 255, pp. 2897–2899, 1962.

[16] A. Argyriou, C. Micchelli, M. Pontil, L. Shen, and Y. Xu, "Efficient first order methods for linear composite regularizers," *ArXiv preprint:1104.1436*, 2011.

[17] S. Villa, S. Salzo, L. Baldassarre, and A. Verri, "Accelerated and inexact forward-backward algorithms," Optimization Online, Tech. Rep., 2011.

[18] M. Schmidt, N. Le Roux, and F. Bach, "Convergence rates of inexact proximal-gradient methods for convex optimization," in *Advances of Neural Information Processing Systems (NIPS)*, 2011.

[19] P. Talairach and J. Tournoux, *A stereotactic coplanar atlas of the human brain*. Stuttgart: Thieme, 1988.