Speeding-up model-selection in GraphNet via early-stopping and univariate feature-screening

Elvis Dohmatob, Michael Eickenberg, Bertrand Thirion, Gaël Varoquaux

Speeding-up model-selection in GraphNet via early-stopping and univariate feature-screening

Elvis DOHMATOB *†, Michael EICKENBERG *, Bertrand THIRION *, Gaël VAROQUAUX *
* INRIA Parietal, Neurospin, Bât 145, CEA Saclay, 91191 Gif sur Yvette, France
firstname.lastname@inria.fr
†Corresponding author

***Abstract*—The GraphNet (aka S-Lasso), as well as other "sparsity + structure" priors like TV (Total-Variation), TV-L1, etc., are not easily applicable to brain data because of technical problems relating to the selection of the regularization parameters. Also, in their own right, such models lead to challenging high-dimensional optimization problems. In this manuscript, we present some heuristics for speeding up the overall optimization process: (a) Early-stopping, whereby one halts the optimization process when the test score (performance on leftout data) for the internel cross-validation for model-selection stops improving, and (b) univariate feature-screening, whereby irrelevant (non-predictive) voxels are detected and eliminated before the optimization problem is entered, thus reducing the size of the problem. Empirical results with GraphNet on real MRI (Magnetic Resonance Imaging) datasets indicate that these heuristics are a win-win strategy, as they add speed without sacrificing the quality of the predictions. We expect the proposed heuristics to work on other models like TV-L1, etc.**

***Index Terms*—MRI; supervised learning; pattern-recognition; sparsity; GraphNet; S-Lasso; TV-L1; spatial priors; model-selection; cross-validation; univariate feature-screening**

I. Introduction

Sparsity- and *structure*-inducing priors are used to perform jointly the prediction of a target variable and region segmentation in multivariate analysis settings. Specifically, it has been shown that one can employ priors like Total Variation (TV) [1], TV-L1 [2], [3], TV-ElasticNet [4], and GraphNet [5] (aka S-Lasso [6] outside the neuroimaging community) to regularize regression and classification problems in brain imaging. The results are brain maps which are both sparse (i.e regression coefficients are zero everywhere, except at predictive voxels) and structured (blobby). The superiority of such methods over methods without structured priors like the Lasso, ANOVA, Ridge, SVM, etc. for yielding more interpretable maps and improved prediction scores is now well established (see for example [2], [3]). These priors are fast becoming popular for brain decoding and segmentation. Indeed, they leverage a feature-selection function (since they limit the number of active voxels), and also a structuring function (since they penalize local differences in the values of the brain map). Also, such priors produce state-of-the-art methods for automatic extraction of functional brain atlases [7].

However, these rich multivariate models lead to difficult optimization and model-selection problems which render them impractical on brain data. In this paper, we provide heuristic techniques for speeding-up the application of GraphNet[6], [5]

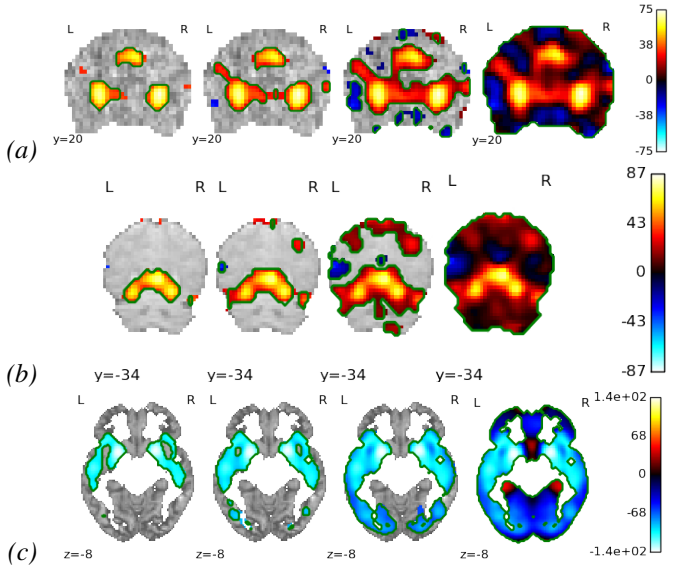

Fig. 1. Univariate feature-screening for the GraphNet problem (2) on different datasets. This figure shows spatial maps of $X_j^T y$, thresholded so that only voxels j with (from left to rightmost column) $|X_j^T y| \geq p_{10\%}(|X^T y|)$, $|X_j^T y| \geq p_{20\%}(|X^T y|)$, $|X_j^T y| \geq p_{50\%}(|X^T y|)$, and $|X_j^T y| \geq p_{100\%}(|X^T y|)$ (full-brain) respectively, survive. The green contours enclose the elite voxels which are selected by the screening procedure at the respective threshold levels. *(a)*: Mixed Gambles dataset [8]. Remarkably, the geometry of the regions obtained here for the 10th and 20th screening-percentiles match pretty well the results obtained in [3] with their TV-L1 penalty. *(b)*: Face vs House contrast of the visual recognition dataset [9]. Weights maps obtained for the GraphNet model (2) with these different screening-percentiles are shown in Figure 3. *(c)*: OASIS dataset [10] with VBM. See Figure 2 for weights maps and age predictions obtained using these different screening-percentiles.

on neuro-imaging data. The first heuristic termed *univariate feature-screening*, provides a principled way to a priori detect and eliminate voxels which are the most irrelevant to the learning task, thus reducing the size of the underlying optimization problem (2). The second heuristic, *early-stopping*, detects when the model has "statistically" converged so that pushing further the numerical optimization leads to no gain in prediction / classification performance, so that the process can be halted safely, without sacrificing predictive performance.

The GraphNet [5] (aka S-Lasso [6]): We denote by $y \in \mathbb{R}^n$ the targets to be predicted (age, sex, IQ, etc.); the *design matrix* $X \in \mathbb{R}^{n \times p}$ are the brain images related to the presentation of different stimuli, or other brain acquisition (e.g

gray-matter concentration maps from anatomy, etc.). p is the number of voxels and n the number of samples (images). In brain imaging, $p \gg n$; typically, $p \sim 10^3 - 10^6$ (in full-brain analysis), while $n \sim 10 - 10^3$ (n being limited by the cost of acquisition, etc.). Let $\Omega \subset \mathbb{R}^3$ be the 3D image domain representing the region occupied by the brain –or ROI (region of interest) thereof– under study, discretized regularly on a finite grid. The coefficients w define a spatial map in $\mathbb{R}^p$. The spatial gradient of w at a voxel $j \in \Omega$ reads:

$$\nabla w(j) := [\nabla_x w(j), \nabla_y w(j), \nabla_z w(j)] \in \mathbb{R}^3, \quad (1)$$

where ∇_u is the finite-difference operator along the u-axis. Thus ∇ defines a $3p$-by-p linear operator, with adjoint $= -div$.

GraphNet then corresponds to the following problem:

$$\text{Find } (w, b) \in \mathbb{R}^{p+1} \text{ minimizing } \mathcal{L}(y, Xw, b) + \alpha J(w) \quad (2)$$

where:

- w is the *weights map* of *regressor coefficients*, and b is the *intercept*; $(\hat{w}, \hat{b})$ denotes a solution to problem (2).
- $\mathcal{L}(y, Xw, b)$ is the *loss* term, and measures how well the coefficients (w, b) explain the data (X, y). Typically, $\mathcal{L}(y, Xw, b)$ is *Mean Square Error (MSE)* in regression problems, and *logistic loss* in classification problems. For details, refer to subsection II.C of [1], for example.
- $J(w) := \rho\|w\|_{\ell_1} + \frac{1-\rho}{2}\|\nabla w\|_2^2$ is the *regularization*. $\alpha \geq 0$ controls the amount of regularization, and the parameter $\rho \in [0, 1]$, also known as the ℓ_1*-ratio*, is the trade-off between the *sparsity-inducing* penalty ℓ_1 (Lasso) and *spatial-structure-promoting* ℓ_2 term $\|\nabla w\|_2^2$.

II. Methods

(a) A note on implementation of the solver: Problem (2) is a nonsmooth convex-optimization problem. One notes that in the penalty term $J(w)$, the $\|\nabla w\|_2^2$ sub-term is smooth (i.e differentiable) with *Lipschitz* gradient, whilst the ℓ_1 –though nonsmooth– is *proximable*[1] by means of the *soft-thesholding* operator [11]. Thus problem (2) is amenable to the FISTA (Fast Iterative Shrinkage-Thresholding Algorithm) [12], with a provable $\mathcal{O}(1/\sqrt{\epsilon})$ convergence rate. Our implementation of FISTA uses technical recommendations (line-searching, parametrization, etc.) which were provided in [13], in the context of TV-L1 [2], [3]. The model parameters α and ρ in (2) are set by *internal* cross-validation.

(b) Univariate feature-screening: In machine-learning, feature-screening aims at detecting and eliminating irrelevant (non-predictive) features thus reducing the size of the underlying optimization problem (here problem (2)). The general idea is to compute for each value of the regularization parameter, a *relevance measure* for each feature, which is then compared with a threshold (produced by the screening procedure itself). Features which fall short of this threshold are detected as irrelevant and eliminated. For the Lasso and similary models (including Group Lasso), *exact*[2] screening techniques include those developed in [14], [15], [16], [17]. Inexact screening techniques (e.g [18]) have also been proposed in the literature.

Our proposed heuristic screening technique is inspired by the *Marginal screening* technique developed in Algorithm 1 of [15], and operates as follows. The data (X, y) are standardized so that y has unit variance and zero mean, likewise each row of the design matrix X. To ensure obtention of a smooth mask, a Gaussian-smoothed version of X is used in the screening procedure (but not in the actual model fit). For each voxel j (voxels are the features here) the absolute dot-product $|X_j^T y|$ of y with the jth column of X is computed. For a given screening-percentile $sp \in [0, 100]$, the spth percentile value of the vector $|X^T y| := (|X_1^T y|, ..., |X_p^T y|)$, denoted $p_{sp}(|X^T y|)$, is computed. The case $sp = 100$ corresponds to full-brain analysis. 25 means we keep the quarter of the brain made of voxels with the highest $|X_j^T y|$ values. And so on. A brain-mask is then formed, keeping only those voxels j for which $|X_j^T y| \geq p_{sp}(|X^T y|)$. Next, this brain-mask is morphologically eroded and then dilated, to obtain a more structured mask. Figure 1 shows results of applying this screening heuristic to various datasets, prior to model fitting.

(c) Early-stopping: In each train sub-sample (for example a fold, in the case of K-fold cross-validation) of the internal cross-validation loop for setting the parameters of the GraphNet model (2), a pass is done on the 2-dimensional parameter grid, and each parameter pair (α, ρ) is scored according to its prediction / classification performance. For a fixed parameter pair (α, ρ), an instance of problem (2) is solved iteratively using FISTA [12]. At each iteration, the prediction / classification performance of the current (not yet optimal) solution $\hat{w}_k$ in (2) is computed. If in a time-window of 5 iterations this score has not increased above an a priori fixed threshold, called the *early-stopping tolerance (es tol)*, then the optimization process is halted for the currrent model parameter pair (α, ρ) under inspection. This heuristic is motivated by the intuition that, for a particular problem, sub-optimal solutions $\hat{w}_k$ can give the same score as an optimal solution $\hat{w}$ (i.e statistical convergence may happen before numerical convergence). By default we set this early-stopping tolerance to -10^{-4} for classification and -10^{-2} for regression problems. A value of $+\infty$ (in fact, any value above 10, say) corresponds to no early-stopping at all (i.e, solve problem (2) until numerical convergence).

III. Experiments on MRI data

We experimented our early-stopping and (separately) feature-screening heuristics on different MRI datasets.

N.B.: *All experiments were run using a single core of a laptop.*

(a) Regression setting: OASIS dataset [10]: The Open Access Series of Imaging Studies (OASIS) dataset consists of a cross-sectional collection of 416 subjects aged 18 to 96. For each subject, 3 or 4 individual T1-weighted MRI scans obtained in single scan sessions are included. A natural regression problem for this dataset is to predict the age of subject from their anatomical data. To this end, we segmented the gray-matter from the anatomy of each

[1] That is, there is a closed-form analytic expression for its proximal operator.

[2] i.e, techniques which don't mistakenly discard active predictive features.

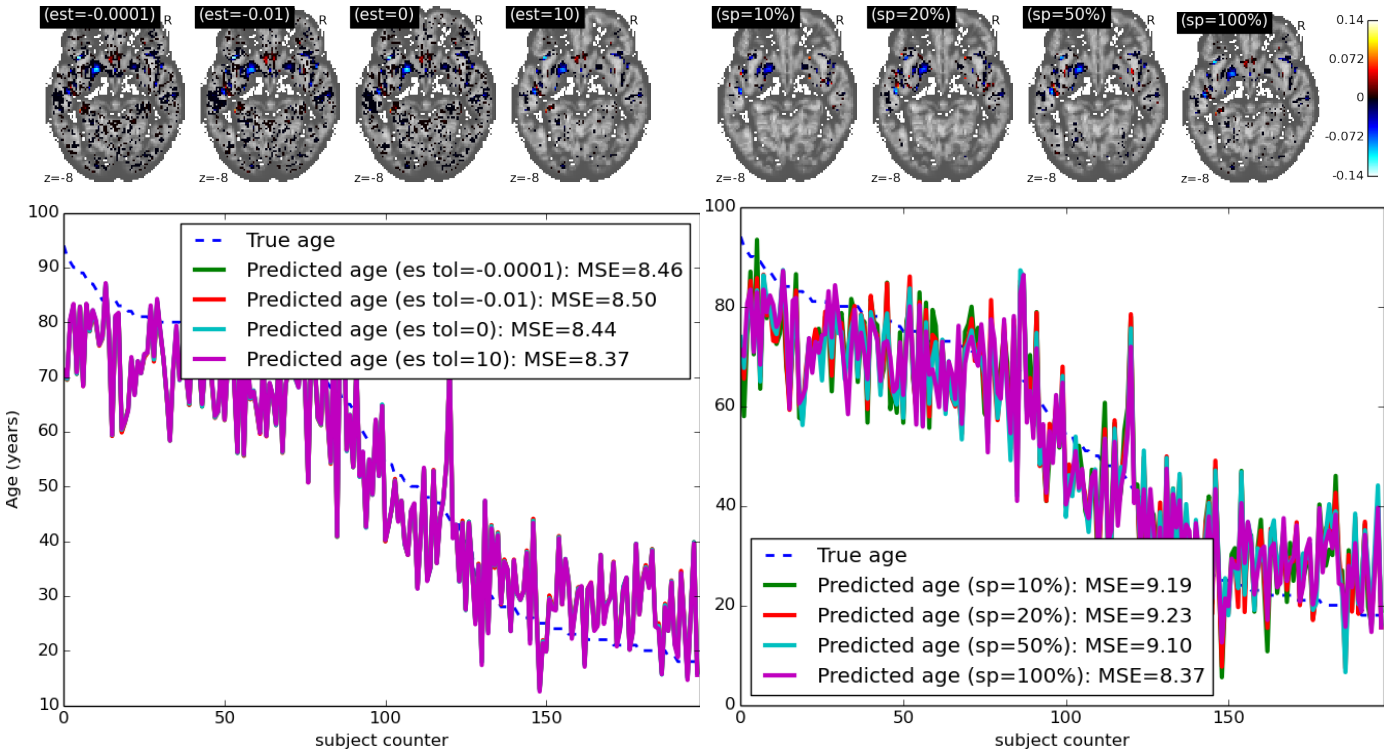

Fig. 2. Predicting age from gray-matter concentration maps from the OASIS dataset [10]. **Top**: Weights maps (solutions to problem (2)). **Bottom-left**: Mean Square Error (MSE) in age prediction, for different subjects of the validation set, for varying levels of the early-stopping toleranace ("es tol" for short), with the screening-percentile (sp) held constant at 100 (full-brain). **Bottom-right**: MSE in age prediction, for varying levels of the screening-percentile (sp). **Running times**: Increasing *est tol* (from -10^{-4} to 10): **100.2m, 171.4m, 188.8m, 289.6m**. For increasing *sp* (10 to 100): **44.2m, 81.3m, 186.5m, 341.3m**

subject (obtained from the T1 images), and used the gray-matter maps as features for predicting age. We split the 416 subjects into two equallly sized and age-balanced groups: a train set and a validation set. The GraphNet model [6], [5] was fitted on the train set, with parameters (α and ρ in (2)) set internally via 8-fold cross-validation. The results for this experiment are shown in Figure 2.

(b) Classification setting: Visual recognition dataset [9]: Our second dataset [9], is a popular block-design fMRI dataset from a study on face and object representation in human ventral temporal cortex. It consists of 6 subjects with 12 runs per subject. In each run, the subjects passively viewed images of eight object categories, grouped in 24-second blocks separated by intermittent rest periods. This experiment is a classification task: predicting the object category y. We use a *One-versus-Rest (OvR)* strategy. The design matrix X is made of time-series from the full-brain mask of $p = 23\,707$ voxels over $n = 216$ TRs, of a single subject (subj1). We divided the 12 runs into 6 runs for training and 6 other runs for validation. *Leave-one-label-out* cross-validation was used for selecting the model parameters (α, ρ). The results are depicted in Figure 3.

IV. Results

We now summarize and comment the results of the experiments (refer to section III). Figure 2 shows the effects of early-stopping heuristic and feature-screening heuristic on age prediction scores on the OASIS dataset [10] (416 subjects). We see that in the internal cross-validation, stopping the optimization procedure for fixed (α, ρ) pair of regularization parameters, when test score increases by about -10^{-2} is a good heuristic, and does just as good as running the optimization until numerical convergence. Also (and independently), one gets similar prediction scores using as little as a fifth of the brain volume ($sp = 20$), compared to using the full-brain ($sp = 100$). Figure 3 reports similar results for classification on the visual recognition dataset [9]. Overall, we see from Figures 3 and 2 that we can achieve upto 10-fold speedup with the proposed heuristics, with very little loss in accuracy.

V. Conclusion and Future Work

In this manuscript, we have presented heuristics that provide speedups for optimizing GraphNet [6], [5] in the difficult context of brain data. These heuristics are a *win-win* strategy, as they add speed without sacrificing the quality of the predictions / classifications. In practice, we do a 20% univariate feature-screening by default, which ensures a 5-fold speedup over full-brain analysis, and independently of an approximately 2-fold speedup obtained by the early-stopping heuristic, leading to an overall 10-fold speedup. Our results have been verified empirically on different MRI datasets, namely [10] and [9]. Our heuristics should be applicable to other hard-to-optimize models like TV-L1 [2], [3], etc.

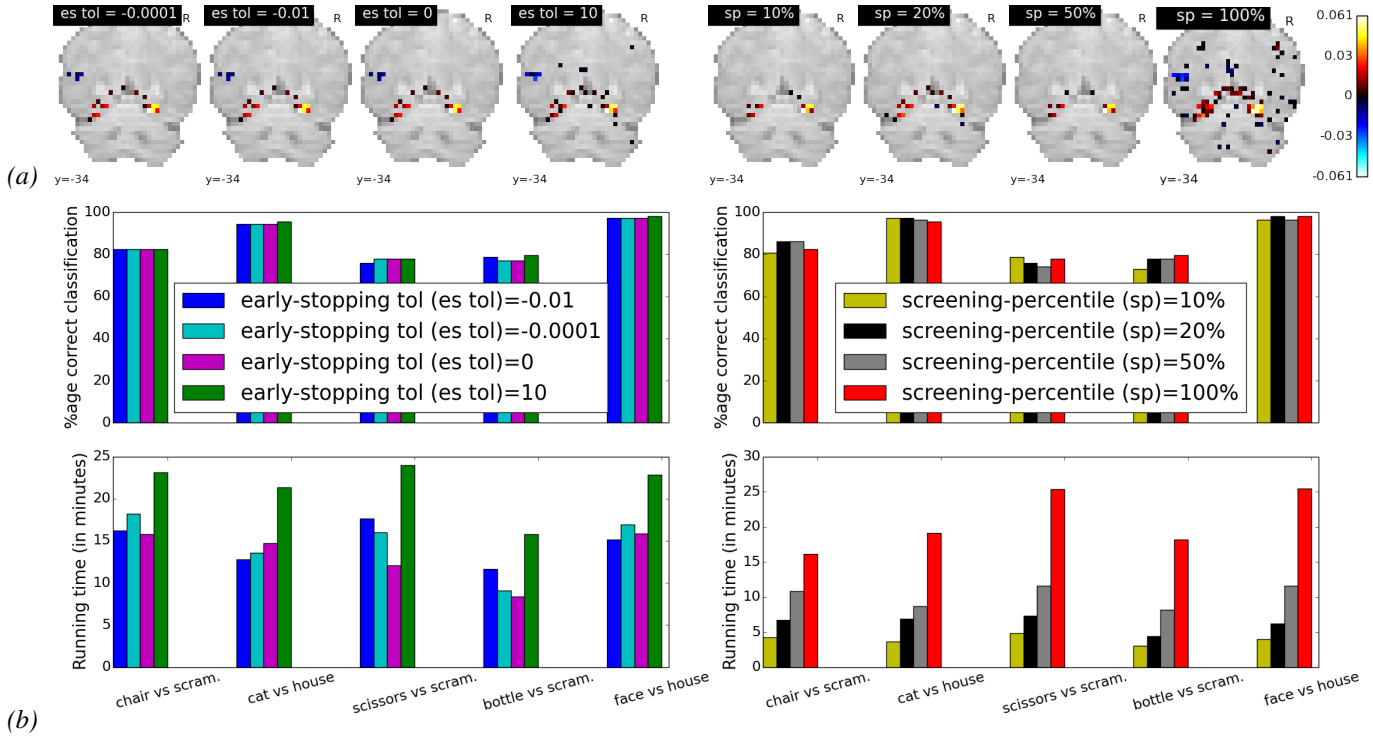

Fig. 3. Visual recognition dataset [9]. ***(a)***: Weights maps for the Face vs House contrast, for different the early-stopping and univariate feature-screening thresholds. One can see that the supports of these maps for different values of the thresholds are quite similar to cases involving no heuristic at all (the case where est $= 10$ and the where case sp $= 100\%$). ***(b)*, top-left**: Prediction scores as a function of the early-stopping tolerance (est), for different task contrasts. It can be seen that contiguous bars are of almost same height, indicating that early stopping does not harm the accuracy of the predictions. ***(b)*, top-right**: Prediction scores as a function of the screening-percentile (sp), for different task contrasts. We can see that contiguous groups of bars are roughly flat at the top, with a sligh increase from lower to high screening-percentile values. The case "chair vs scramped" is an exception, where a slightly reverse tendency if observed. A possible explanation is that 20th percentile feature-screening already selects the right voxels (quasi-exact support recovery), and so including more voxels in the model can only hurt its performance. ***(b)*, bottom-row**: Running times in minutes for the different thresholds of the heuristics. In particular, we see that using only the 20% most relevant voxels achieves a speedup of up 5, while ensuring as much accuracy as in full-brain analysis ($sp = 100$).

Due to time constraints, only 2 datasets [10], [9] were considered in the benchmarks. A natural extention of the empirical results presented here would be to run the experiments on more datasets (for example the OpemfMRI datasets [19]).

REFERENCES

[1] 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, p. 1328, 2011.

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

[3] A. Gramfort, B. Thirion, and G. Varoquaux, "Identifying predictive regions from fMRI with TV-L1 prior," in *PRNI*, 2013, p. 17.

[4] M. Dubois, F. Hadj-Selem, T. Lofstedt, M. Perrot, C. Fischer, V. Frouin, and E. Duchesnay, "Predictive support recovery with tv-elastic net penalty and logistic regression: an application to structural mri," in *PRNI*. IEEE, 2014, p. 1.

[5] 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.

[6] M. Hebiri and S. van de Geer, "The smooth-lasso and other $\ell_1 + \ell_2$-penalized methods," *Electron. J. Stat.*, vol. 5, p. 1184, 2011.

[7] A. Abraham, E. Dohmatob, B. Thirion, D. Samaras, and G. Varoquaux, "Extracting brain regions from rest fMRI with total-variation constrained dictionary learning," in *MICCAI*, 2013, p. 607.

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

[9] J. V. Haxby, I. M. Gobbini, M. L. Furey, A. Ishai, J. L. Schouten, and P. Pietrini, "Distributed and overlapping representations of faces and objects in ventral temporal cortex," *Science*, vol. 293, p. 2425, 2001.

[10] D. S. Marcus, T. H. Wang, J. Parker, J. G. Csernansky, J. C. Morris, and R. L. Buckner, "Open access series of imaging studies (oasis): cross-sectional mri data in young, middle aged, nondemented, and demented older adults," *Journal of cognitive neuroscience*, vol. 19, p. 1498, 2007.

[11] I. Daubechies, M. Defrise, and C. De Mol, "An iterative thresholding algorithm for linear inverse problems with a sparsity constraint," *Comm. Pure Appl. Math.*, vol. 57, p. 1413, 2004.

[12] A. Beck and M. Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," *SIAM Journal on Imaging Sciences*, vol. 2, p. 183, 2009.

[13] E. Dohmatob, A. Gramfort, B. Thirion, and G. Varoquaux, "Benchmarking solvers for tv-l1 least-squares and logistic regression in brain imaging," in *PRNI*. IEEE, 2014.

[14] L. E. Ghaoui, V. Viallon, and T. Rabbani, "Safe feature elimination in sparse supervised learning," *CoRR*, vol. abs/1009.3515, 2010.

[15] J. D. Lee and J. E. Taylor, "Exact post model selection inference for marginal screening," in *NIPS*, 2014, pp. 136–144.

[16] J. Liu, Z. Zhao, J. Wang, and J. Ye, "Safe screening with variational inequalities and its applicaiton to lasso," in *ICML*, 2014.

[17] J. Wang, P. Wonka, and J. Ye, "Lasso screening rules via dual polytope projection," *Journal of Machine Learning Research*, 2015.

[18] R. Tibshirani, J. Bien, J. Friedman, T. Hastie, N. Simon, J. Taylor, and R. J. Tibshirani, "Strong rules for discarding predictors in lasso-type problems," *J. R. Stat. Soc.: Series B*, vol. 74, p. 245, 2010.

[19] R. A. Poldrack, D. M. Barch, J. P. Mitchell, T. D. Wager, A. D. Wagner, J. T. Devlin, C. Cumba, O. Koyejo, and M. P. Milham, "Toward open sharing of task-based fMRI data: the OpenfMRI project," *Front Neuroinform*, vol. 7, p. 12, 2013.