# Total variation meets sparsity: region-inducing penalties segment functional brain modules

**Anonymous Author(s)**
Affiliation
Address
email

## Abstract

Functional MRI has been intensively used to chart the functional organization of the brain, highlighting a myriad of specialized brain regions More recently, it has also been used in a predictive modeling setting to *decode* brain function from images. However, the notion of brain regions is lost by this type of analysis. We address this challenge by developing efficient convex region-selecting penalties, combining sparsity and total variation, that can be used to regularize linear models. Given the size and the 3D nature of brain images, computational efficiency is key. Keeping this in mind, we contribute an efficient optimization scheme that performs all dense operations in the dual, as well as a relaxation of total variation that opens the door to using screening rules. On an fMRI study of the visual system, the penalties not only improve prediction of cognitive state from the brain images, but also segment well the known functional brain modules.

## 1 Introduction

Functional MRI (fMRI) provides very noisy, but spatially-resolved, images of brain function. While stringent statistics are required to tease out signal from noise, analyzing these images for many cognitive tasks has led to the notion of spatially-localized brain modules that are recruited in specific situations. For instance, the way the brain performs visual object recognition is understood as involving a chain of object-sensitive areas such as the famed Fusiform Face Area (FFA) [**?**].

From a formal standpoint, statistical analysis performed on these images implies detecting continuous regions, here brain modules, on a noisy background. In image analysis, this operation can be seen as a foreground-segmentation problem. In fMRI analysis, Markov random fields have been used as post-processing on statistical maps [**?**]. However, the images are themselves often the result of a preceding estimation or reconstruction procedure. In functional imaging, they can be the output of statistical parametric mapping, in which a series of images are analyzed via a linear model to test for differences. In computed tomography (CT), raw measurements are noisy projections of 3D objects onto planes and recovering the image requires inverting a Radon transform. If the statistical power of the first estimation procedure is low, for instance because of a high level of noise, or a low number of samples, there is great benefit in combining it with the segmentation process in a single estimation step. Indeed, injecting in CT reconstruction the information that the image to recover consists of flat regions can have a dramatic effect on recovery quality, related to compressed sensing [**?**]. The empirical risk minimization formulation combines the measurement model with the prior information expressed as a penalty. We are thus interested in developing convex penalties for the segmentation of smooth, but well delineated foreground structure, from constant zero background.

Our goal is to detect spatially-continuous patches in statistically estimated images, and to inform the estimation of the image with these detections. For example, in CT reconstruction of a small set of objects, knowing the contours of the objects makes the estimation of gray-level internal structure

easier. In fMRI analysis, the challenge is to find extended patches of differences between noisy images of brain function that carry information to predict cognition. Our methodological contributions are the following: *i)* we introduce a new penalty selecting spatially-continuous weights but without enforcing any structure inside the domain; *ii)* we derive an efficient optimization procedure for squared-loss regression with this penalty; *iii)* we introduce a relaxation of this penalty to a simple group-lasso problem that leads to efficient path algorithms leveraging screening rules. We apply this penalty to segmenting functional modules of the visual cognitive system from fMRI data.

## 2 Prior work on sparse penalties and segmentation

Our work draws from two bodies of literature: the fields of sparsity and segmentation. Sparse penalties have the remarkable property to minimize the risk in linear models, whether it be prediction error in machine-learning settings, or estimation error in signal-processing settings. In particular, under incoherence conditions that bound the correlations of the design matrix (or sensing operator in signal processing), the $\ell_1$ penalty can recover the weights and their support in noisy conditions and low numbers of samples, provided that a sparsity assumption can be made [**?**, **?**, **?**]. In fMRI, this assumption of sparsity is very relevant: specialized brain modules under study occupy only a small fraction of the images. Specifically, we are interested a foreground-segmentation-like problem: recovering non-zero functional regions on a background, *i.e.* sparse images. However, in many real-world applications, such as CT or medical imaging, the underlying measurement process leads to strong correlations in columns of the design matrix corresponding to neighboring pixels. The standard way to overcome this challenge is to impose sparsity on groups of correlated variables, *e.g.* via mixed $\ell_{2,1}$ penalties that extend the group-lasso [**?**]. This strategy boils down to engineering penalties with specific structure, often overlapping groups, that encode domain knowledge [**?**]. On sparse images, local structure can be imposed by grouping neighboring pixels [**?**, **?**], while long-distance overlapping groups can enforce a more global structure [**?**]. The limitation of these approaches is that the group structure is not data-dependent. Thus if the sparsity pattern is not known to some degree, the number of groups must be large enough to span the various configurations of interest, and much more so when the number of spatial dimensions of the images increases –as with medical images that are often 3D. From a statistical standpoint, increasing the number of groups is detrimental to the recovery properties of the problem: it inflates the number of samples necessary for recovery and can bias the shape of the regions recovered to that of the groups. In addition, the computational cost increases at least linearly with the number of group overlaps [**?**], which becomes prohibitive in 3D settings, particularly so to enforce long-distance structure.

The other body of literature that we are concerned with is that of segmentation, with a specific interest on convex variational approaches, as they can be expressed as penalties in a risk minimizer. A central aspect is the Mumford-Shah functional that yields piecewise smooth approximations of images [**?**]. Chan and Vese [**?**] introduced a variant for segmentation purposes computing piecewise constant approximations: the minimal partition problem. These variational formulations are not convex, but [**?**] have shown that good solutions to the minimal partition problem can be achieved with a similar but convex functional, based on the total variation norm (TV), *i.e.* the $\ell_1$ norm of the image gradient. For our purposes, this approach is appealing, as TV can be used as a penalty –technically an analysis sparsity penalty [**?**] that imposes sparse gradients– and has good properties for image denoising [**?**] or estimation in a linear model [**?**]. However, all these related segmentation approaches model an object as a homogeneous constant-valued domain, thus washing out internal structure. Here, in the context of foreground-background segmentation, we want to impose a flat structure on the background, but not in the selected image domain. In this setting, imposing sparsity of the background seems a better candidate for segmentation than imposing constant domains.

The challenge that we address here is that penalties creating long-distance sparse domains with structured sparsity require many overlapping groups and thus entail high computational cost and bias. For this we revisit TV, introducing a new sparse analysis penalty that does not impose flat non-zero domains and can be solved efficiently. We also introduce a relaxation of this new problem into a non-overlapping group-lasso on a small set of auxiliary variables. The benefit of this formulation is that it is a synthesis sparsity problem that can be solved very efficiently for very sparse solutions.

## 3 Sparse variation: combining sparsity and total variation

**TV-$\ell_1$** In fMRI, total variation (TV) [**?**] and $\ell_1$ have been used in combination [**?**], giving the following estimation problem, in the case of the square loss:

$$\hat{\boldsymbol{w}} = \underset{\boldsymbol{w}}{\operatorname{argmin}} \left\|\boldsymbol{y} - \boldsymbol{X}\,\boldsymbol{w}\right\|_2^2 + \lambda\big(\mu\|\boldsymbol{w}\|_1 + (1-\mu)\,\|\nabla\boldsymbol{w}\|_{2,1}\big) \tag{1}$$

where $\boldsymbol{X} \in \mathbb{R}^{n\times p}$ is the design matrix, $\boldsymbol{y} \in \mathbb{R}^n$ is the vector of observations, $\boldsymbol{w} \in \mathbb{R}^p$ the weights we are searching for, and the parameters $\lambda$ and $\mu$ control respectively the overall amount of penalization and the trade-off between sparsity and spatial structure. $\nabla$ is the finite-difference operator; on a 3D grid of size $p = p_x p_y p_z$, $\nabla \in \mathbb{R}^{3p\times p}$. This formulation can be seen as an extension of the fused lasso [**?**] to 3-dimensional settings. The TV and TV-$\ell_1$ problems can be understood as an analysis sparse penalty [**?**]: sparsity is formulated on a linear transformation of the weights $\boldsymbol{K}\,\boldsymbol{w}$ (thus on an operator that *analyzes* the signal), as opposed to synthesis sparsity, where the goal is to re*synthesize* the signal $\boldsymbol{w}$ from a small number of atomic components. By setting $\boldsymbol{K} = [\mu\boldsymbol{I}_p, (1-\mu)\nabla]^\mathsf{T} \in \mathbb{R}^{4p\times p}$ where $\boldsymbol{I}_p$ is the identity on $\mathbf{R}^p$, the penalty in (**??**) can be written $\|\boldsymbol{K}_{\text{TV}}\,\boldsymbol{w}\|_{2,1}$ for a certain group structure on the $\ell_{2,1}$ norm.

**Sparse variation** We propose a new penalty, called *sparse variation*, that adapts the TV-$\ell_1$ penalty to the segmentation of smooth regions on a zero background. Indeed, TV-$\ell_1$ enforces flat regions in the image space as well as sparsity. In the specific context of foreground segmentation of regions, we are interested in the long-distance sparsity it brings, but not the staircasing effect that it can impose on continuously varying images. As TV-$\ell_1$, sparse variation is an $\ell_{2,1}$ penalty on finite differences:

$$\text{in 1D,} \qquad \text{SV}(\boldsymbol{w}) \overset{\text{def}}{=} \sum_{i=1}^{p} \sqrt{\mu^2\,\boldsymbol{w}_i^2 + \frac{1}{2}(1-\mu)^2\Big((\boldsymbol{w}_i - \boldsymbol{w}_{i+1})^2 + (\boldsymbol{w}_i - \boldsymbol{w}_{i-1})^2\Big)} \tag{2}$$

here we use circular boundary conditions: $\boldsymbol{w}_{p+1} = \boldsymbol{w}_1$, $\boldsymbol{w}_0 = \boldsymbol{w}_p$. To better highlight the link with the TV-$\ell_1$ problem, this penalty can be written in terms of a new analysis operator $\boldsymbol{K}_{\text{SV}} = [\mu\,\boldsymbol{I}_p, \frac{1}{\sqrt{2}}(1-\mu)\nabla_{\text{Left}}, \frac{1}{\sqrt{2}}(1-\mu)\nabla_{\text{Right}}]^\mathsf{T} \in \mathbb{R}^{7p\times p}$, where $\nabla_{\text{Right}}$ and $\nabla_{\text{Left}}$ are the finite-difference operators respectively left-shifted, and right-shifted. Penalty (**??**) is then written $\|\boldsymbol{K}_{\text{SV}}\,\boldsymbol{w}\|_{2,1}$ with the $\ell_{2,1}$ norm defined using a simple group structure on the output space of $\boldsymbol{K}_{\text{SV}}$, that groups together at one image location all the corresponding terms of $\boldsymbol{K}_{\text{SV}}$: the image value itself, and the finite differences in all directions, left and right shifted. Note that for $\mu \in (0, 1]$, $\text{Ker}(\boldsymbol{K}_{\text{SV}}) = \{0\}$ and thus $\text{SV}(\boldsymbol{w})$ is a proper norm. This penalty uses group sparsity to enforce jointly zero values for the spatial differences and the values in the image; in other terms where the image values are not zero, the left and right differences are not forced to zero, thus alleviating the staircasing effect of TV.

**Optimization strategy** From a computation standpoint, solvers for very high-dimensional problems with dense unstructured design matrix such as fMRI are very costly, unless they can optimize in the dual (the famous kernel trick) or on an active set. With elaborate penalties such as ours, neither of these tricks are straightforwardly available. Here, we introduce an efficient optimization strategy for the squared loss relying on an ADMM scheme [**?**] as in [**?**]. Unlike previous work solving TV-related problems, we devise our strategy so that the updates related to the data-fit term are optimized in the dual, relying on the fact that with circular boundary conditions, we readily know an eigenvalue decomposition for the operators $\boldsymbol{K}_{\text{TV}}$ and $\boldsymbol{K}_{\text{SV}}$. In general, circular boundary conditions when dealing with images are a useful numerical trick, but they introduce meaningless effects at the boundaries. Here we avoid these by spatially padding the design matrix $\mathbf{X}$ with a layer of zeros at the boundary, augmenting by 2 in each direction the size of the estimated images. In the ADMM framework, we introduce a split variable $\boldsymbol{z} = \boldsymbol{K}\,\boldsymbol{w}$ where $\boldsymbol{K}$ is $\boldsymbol{K}_{\text{TV}}$ or $\boldsymbol{K}_{\text{SV}}$ to solve the TV-$\ell_1$ regression or the sparse-variation regression. The ADMM update equations [**?**] are then written ($\rho$ is the ADMM control parameter, and $\boldsymbol{a}$ the lagrangian variable):

$$\boldsymbol{w}_{t+1} = \big(\boldsymbol{X}^\mathsf{T}\boldsymbol{X} + \rho\,\boldsymbol{K}^\mathsf{T}\boldsymbol{K}\big)^{-1}\big(\boldsymbol{X}^\mathsf{T}\boldsymbol{y} + \boldsymbol{K}^\mathsf{T}(\rho\,\boldsymbol{z}_t - \boldsymbol{a})\big) \tag{3}$$

$$\boldsymbol{z}_{t+1} = \text{prox}_{\frac{\lambda}{\rho}\ell_{2,1}}\big(\boldsymbol{K}\,\boldsymbol{w}_{t+1} + \frac{1}{\rho}\boldsymbol{a}_t\big) \tag{4}$$

$$\boldsymbol{a}_{t+1} = \boldsymbol{a}_t + \rho\,(\boldsymbol{K}\,\boldsymbol{w}_{t+1} - \boldsymbol{z}_{t+1}). \tag{5}$$

$\text{prox}_{\frac{\lambda}{\rho}\ell_{2,1}}$ is the proximal operator [**?**] for the $\ell_{2,1}$, which can be computed in closed form with computationally cheap operations. For both total variation and sparse variation, $\boldsymbol{K}^\mathsf{T}\boldsymbol{K} = \mu^2\boldsymbol{I}_n - (1-\mu)^2\boldsymbol{\Delta}$ with $\boldsymbol{\Delta}$ the image Laplacian, which is diagonal in a Fourier basis, with known eigenvalues, thanks to the circular boundary conditions. Using the Woodbury matrix identity and a bit of algebra, we can avoid inverting a $p \times p$ matrix in Eq. (**??**):

$$\left(\boldsymbol{X}^\mathsf{T}\boldsymbol{X} + \rho\,\boldsymbol{K}^\mathsf{T}\boldsymbol{K}\right)^{-1} = \frac{1}{\rho}\boldsymbol{F}\,\boldsymbol{\Lambda}^{-1}\left(\boldsymbol{I}_p - \tilde{\boldsymbol{X}}^\mathsf{T}\left(\rho\,\boldsymbol{I}_n + \tilde{\boldsymbol{X}}\boldsymbol{\Lambda}^{-1}\tilde{\boldsymbol{X}}^\mathsf{T}\right)^{-1}\tilde{\boldsymbol{X}}\boldsymbol{\Lambda}^{-1}\right)\boldsymbol{F}^\mathsf{T} \tag{6}$$

where $\boldsymbol{F}$ is the Fourier transform on the image domain –implemented using an FFT–, $\boldsymbol{\Lambda}$ is the diagonal matrix of the eigenvalues of $\mu^2\boldsymbol{I}_n - (1-\mu)^2\boldsymbol{\Delta}$, and $\tilde{\boldsymbol{X}} = \boldsymbol{X}\,\boldsymbol{F}^\mathsf{T}$. This formula involves only rank-$n$ matrix multiplications (as $\boldsymbol{F}$ is implemented via an FFT), which greatly speeds up the operation for $p \gg n$. Indeed, in our fMRI application, we find that 80% of the time is spent in the multiplications by the $\boldsymbol{X}$ and $\boldsymbol{X}^\mathsf{T}$ matrix.

## 4 Variation lasso: relaxing to a synthesis sparsity

Next, we introduce a synthesis sparsity formulation related to sparse variation. Indeed, for very high dimensional problems with solutions sparse in synthesis, very fast solvers [**?**] can rely on screening rules to optimize the problem on a small set of variables, with a cardinality close to that of the active set of the final solution. For this reason, very sparse synthesis problems can be solved much faster than analysis problems: recall that the dimensionality of the $\boldsymbol{X}$ matrix is the limiting factor for computation speed in our algorithm, as in any gradient based one.

**Dual formulation of sparse-variation regression** Due to the injectivity of the extended finite difference operator $\boldsymbol{K}$, the sparse-variation optimization problem can be rewritten as follows:

$$\min_{\boldsymbol{v}} \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{X}\,\boldsymbol{K}^{+}\boldsymbol{v}\|_2^2 + \lambda\|\boldsymbol{v}\|_{2,1} \quad \text{subject to} \quad \boldsymbol{v} \in \operatorname{im}\boldsymbol{K}, \tag{7}$$

where $\boldsymbol{K}^+$ is the Moore-Penrose pseudoinverse of $\boldsymbol{K}$. Indeed, using $\boldsymbol{v} = \boldsymbol{K}\boldsymbol{w}$ and the injectivity of $\boldsymbol{K}$, one can revert back to the original problem. The "unbridgeable gap" [**?**] between synthesis and analysis sparsity is now expressed by a linear constraint on a group lasso optimization problem on an augmented variable $\boldsymbol{v}$. The corresponding dual problem to this formulation is

$$\max_{\boldsymbol{\mu},\boldsymbol{\eta}} \frac{1}{2}\|\boldsymbol{y}\|_2^2 - \frac{1}{2}\|\boldsymbol{\nu} - \boldsymbol{y}\|_2^2 \quad \text{subject to } \|\boldsymbol{K}^{\mathsf{T},+}\boldsymbol{X}^\mathsf{T}\boldsymbol{\nu} + \boldsymbol{\eta}\|_{2,\infty} \leq \lambda \text{ and } \boldsymbol{\eta} \in \ker\boldsymbol{K}^\mathsf{T} \tag{8}$$

It is worth noting that the maximal penalty yielding a non-zero solution follows from this equation and KKT conditions (non-saturation of the dual constraint in a group leads to inactivity of the corresponding primal group) and can be written as

$$\lambda_{\max}^{\text{sv}} = \min_{\eta \in \ker\boldsymbol{K}^\mathsf{T}} \|\boldsymbol{K}^{\mathsf{T},+}\boldsymbol{X}^\mathsf{T}\boldsymbol{y} + \boldsymbol{\eta}\|_{2,\infty} \tag{9}$$

This quantity is an optimization problem in itself, but setting $\eta = 0$ gives a simple upper bound:

$$\lambda_{\max}^{\text{sv}\prime} = \|\boldsymbol{K}^{\mathsf{T},+}\boldsymbol{X}^\mathsf{T}\boldsymbol{y}\|_{2,\infty} \geq \lambda_{\max}^{\text{sv}} \tag{10}$$

**Variation lasso: a new optimization problem** Solving analysis sparsity problems is difficult compared to synthesis problems. The difficulty is very visible in the situations of TV, TV-$\ell_1$ and sparse variation, since reaching the optimum requires passing information spatially across variables in the image; unless the loss term has long distance coupling, this information propagation is slow as it is mediated by the neighbor coupling of TV. In dropping the linear constraint $\boldsymbol{v} \in \operatorname{im}\boldsymbol{K}$ from the reformulated primal problem (**??**) we create a synthesis problem in an augmented variable which eliminates the spatial cross-talk between variables:

$$\min_{\boldsymbol{v}} \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{X}\,\boldsymbol{K}^{+}\boldsymbol{v}\|_2^2 + \lambda\|\boldsymbol{v}\|_{2,1}, \tag{11}$$

We dub this optimization problem *variation lasso*. After optimization, a uniform cross-talk between variables is re-introduced by projecting the augmented solution back using

$$\hat{\boldsymbol{w}}_{\boldsymbol{vl}} = \boldsymbol{K}^{+}\boldsymbol{v}. \tag{12}$$

Note that the corresponding dual problem is stricter than (**??**), as it enforces $\boldsymbol{\eta} = 0$ instead of $\boldsymbol{\eta} \in \ker \boldsymbol{K}^\mathsf{T}$. In fact the maximum penalization value $\lambda^{\text{vl}}_{\max}$ for this problem is exactly $\lambda^{\text{vl}}_{\max} = \lambda^{\text{sv}\prime}_{\max}$.

Intuition on the link between variation lasso and the reformulation (**??**) of sparse variation can be gained by implementing the purely primal *Generalized Forward Backward Splitting* algorithm [**?**] to solve it: At each iteration, after a gradient descent step of the smooth loss, the proximal operators of $\lambda\|\cdot\|_{2,1}$ and that of the indicator $\chi_{\operatorname{im}\boldsymbol{K}}$ are applied in parallel on split variables and the results averaged. In contrast, variation lasso corresponds to applying $\operatorname{prox}\chi_{\operatorname{im}\boldsymbol{K}}$ only once at the end.

**Screening rules** In variation lasso, the augmented variable $\boldsymbol{v}$ lends itself (among others) to the exact screening rules *enhanced dual polytope projection* (EDPP) [**?**]. Here, the formulas are applicable using the design matrix $\boldsymbol{X}\boldsymbol{K}^+$. For a known solution $\boldsymbol{v_0}$ for a given penalty $\lambda_0 < \lambda_{\max}$, we can screen variables for a $\lambda < \lambda_0$. We introduce the auxiliary variables

$$\boldsymbol{\vartheta_0} = \frac{1}{\lambda_0}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{K}^+\boldsymbol{v_0}) \qquad \boldsymbol{u_1} = \frac{\boldsymbol{y}}{\lambda_0} - \boldsymbol{\vartheta_0} \qquad \boldsymbol{u_2} = \frac{\boldsymbol{y}}{\lambda} - \boldsymbol{\vartheta_0} \qquad \boldsymbol{u_2^\perp} = \boldsymbol{u_2} - \frac{\langle \boldsymbol{u_1}, \boldsymbol{u_2}\rangle}{\langle \boldsymbol{u_1}, \boldsymbol{u_1}\rangle}\boldsymbol{u_1}$$

Indexing the groups of the group lasso problem using $g \in \mathcal{G}$, then for each $g \in \mathcal{G}$, the condition

$$\left\| (\boldsymbol{X}\boldsymbol{K}^+)_g^\mathsf{T} \left( \boldsymbol{\vartheta_0} + \frac{1}{2}\boldsymbol{u_2} \right) \right\|_2 < 1 - \frac{1}{2}\|\boldsymbol{u_2}\|_2 \, \|(\boldsymbol{X}\boldsymbol{K}^+)_g\|_{\mathcal{F}},$$

if verified, indicates that this variable group will be inactive in the solution.

## 5 Empirical results

The structuring sparse penalties that we introduced are well suited to segmentation problems on images obtained via a statistical estimation from a small number of noisy samples. To evaluate their recovery behavior in applications, we start with a simple 1D corrupted measurements problem to develop intuition. We then move on to a 3D problem, namely the segmentation of activation patterns recovered from an fMRI experiment on human object representation. The aforementioned structure imposing penalties are evaluated, along with standard $\ell_1$ regularization (the lasso).

**A simple 1D signal-recovery** Here we study the recovery properties of the structuring penalties on a 1D corrupted measurement problem. We mimic a spectroscopy situation, to use a non-trivial simulation with structure in the weights and design matrix –here a discrete cosine transform (DCT). The signal measurements read $\boldsymbol{y} = \boldsymbol{X}_{\text{DCT}}^{-1}\boldsymbol{w} + \boldsymbol{\varepsilon}$ with $\boldsymbol{X}_{\text{DCT}}$ the DCT operator, $\boldsymbol{w}$ the spectrum to recover and $\varepsilon$ a noise vector. For our experiments we use a ground-truth spectrum of size 200, with around 80% zeros and an activated region resembling that of a chemical compound signature: two overlapping smooth peaks, that we create here using a lower-thresholded, downward-pointing parabola. We use a Gaussian noise of an amplitude of 20% that of the signal.

Fig.**??** shows the ground truth, along with the best $\ell_2$ recovery results for sparse variation, TV + $\ell_1$ and variation lasso: Each method was evaluated on a grid of penalties $\lambda$ and sparsity ratios $\mu$. Observe that the TV + $\ell_1$ method recovers a staircased signal as per its construction, ignoring the smooth nature of the ground truth. Sparse variation and variation lasso are more able to follow the shape of the ground truth. Crucially however, the relaxation of sparse variation to the synthesis type variation lasso incurs a loss in spatial communication of variables and leads to smooth bleeding around the support edges, and a stronger susceptibility to noise (see the kink on the right).

Figure 1: **Signal-recovery in spectroscopy simulation** Best $\ell_2$ recovery of the ground truth measured by mean squared error over a fine grid of penalty and sparsity ratio parameters. Note especially the staircasing effect introduced by the TV + $\ell_1$ penalty. Optimal parameter selection is the determination of the optimal bias variance trade-off to minimize the $\ell_2$ recovery error. The number in parenthesis is the $\ell_2$ error (RMSE).

Figure 2: **FFA** (Fusiform Face Area) segmented in a face versus house discrimination on the Haxby 2001 data. **Top**: axial cut at $z = -20$mm, around $x$=14mm, $y$=15mm. **Bottom**: all horizontal lines of the above image. TV-$\ell_1$ shows a blocky behavior –some neighboring voxels share exact same values–, while sparse variation gives smoother maps, and variation lasso has very smooth maps and a background not completely zero.

**Segmenting functional regions from fMRI** FMRI is the tool of choice of cognitive brain mapping. While the signal-to-noise ration is very small –typically only 5% of the variance of brain time series are explained by a known cognitive task– it is widely used to detect regions activated by a task, although the typical data analysis involves spatially smoothing the images, which can be understood as applying a matched filter with the aim of detecting regions. More recently, fMRI images have been also been used to predict the task performed by a subject. Here we propose to do both simultaneously using our region-segmentation penalty in a linear model. An important application of predictive modeling in fMRI has been the study of the visual system: objects presented to a subject recruit different brain regions, possibly with some overlap. Prediction of the object category from brain activity demonstrates the sensitivity of the brain regions to these objects: face-sensitive, place-sensitive, or object-sensitive regions that compose the ventral visual pathway. The challenge is that standard predictors used (kNN, SVM, or $\ell_1$-penalized models) do not segment regions, and neuroscientific conclusions cannot be drawn naturally from the classifier's weight maps.

We revisit the data from a seminal publication in this line of work [**?**]: responses to visual stimuli of different categories - *faces, houses, chairs, scissors, bottles, shoes, cats*, and a control condition named *scrambledpix*, Fourier phase scrambled versions of the other stimuli. We perform on this data a *decoding analysis*: a classification task with brain activity as input and image category as output. The weight maps of the classifier load voxels that distinguish between the tasks, but may not recover the brain modules themselves. We use a one-vs-rest classification scheme, with the squared loss [**?**]. As the problem is very ill-posed ($\sim$ 800 observations total, for 30 000 voxels in the brain), regularization is crucial, and we evaluate the penalties sparse variation, TV-$\ell_1$, variation lasso and the classic lasso. We use cross-validation to choose the best parameters on a grid.

The structure-inducing penalties pull out regions in the weight maps, and, as we will see, these correspond to brain regions involved in object recognition. To better understand the impact of the different penalties, we show on Fig.**??** one of these regions, as segmented by each penalty, for parameters $\lambda$=.01 and $\mu$=.1 falling in the range selected by cross-validation. There is a clear transition from a blocky behavior with TV-$\ell_1$, to a very smooth behavior with variation lasso. Interestingly, for TV-$\ell_1$ on this fMRI data, the staircasing effect is limited to a few neighbooring voxels.

Fig.**??** shows the average weight maps obtained for the best mean score from the folds of cross-validation. This averaging yields a surrogate mean a posteriori, and is useful to reduce the variance of the maps and the blockiness of TV-$\ell_1$. The one-vs-rest approach provides one map for each class, and thus gives a complete brain-mapping view of the visual-recognition task. The voxels marked in green show the weight map for *faces-vs-rest*. To avoid clutter, we show only the contours at 10% of maximum absolute weight for the other classes, using the color of the labels listed on the image.

We observe a flagrant difference between weights obtained with penalties that promote spatial smoothness and contiguity, and those arising from the lasso penalty. Indeed, such settings of strongly spatially correlated design break support-recovery properties of the lasso. As the top row of Fig.**??** shows, it will arbitrarily select a subset of variables amongst the voxels of a region for which have strongly correlated regressors. On the second row, we see that, unlike lasso, the variation lasso establishes spatial contiguity in several regions. Compared to the last two rows of Fig.**??**, which represent true analysis sparse optimizations, the same spatially contiguous structures are recovered. However, the omission of the constraint $\boldsymbol{v} \in \mathrm{im}\, \boldsymbol{K}$ leads to lasso-like susceptibility with spurious activations in the background. The third row of Fig.**??** show the averaged weight maps of TV-$\ell_1$. These are much less noisy than the top two and consist almost solely of spatially contiguous regions selected to perform the classification task. Due to the averaging of weight maps across folds, the staircasing

Figure 3: **Regions segmented on the fMRI object recognition data** The different colors correspond to the different weight maps learned in a one-vs-rest approach. The weight maps are cross validation fold averages corresponding to the optimal parameter set for the prediction task. The face-specific map is shown using a green colormap, while only the outline at 10% of the maximum is shown for the other objects. From the top row to the bottom we observe a general increase in smoothness and spatial contiguity of the average weight maps. At the top, the non-spatially-informed lasso penalty creates maps with highly scattered voxels on a sparse background. In the second row, variation lasso yields spatially contiguous regions accompanied by strong scatter. The third and fourth rows display analysis sparsity maps with strong spatial contiguity and little scatter. In terms of functional regions recovered, the FFA face-specific area is well visible, bilateral, on the $z = -24mm$ cut, while the $z = -17mm$ cut shows a bilateral set of regions that select for all the different categories, the LOC (lateral occipital complex). Each object category recruits a number of different brain modules, for instance the face-specific maps also highlight the OFA (Occipital Face Area) on cut $z = -24mm$ and the fSTS in the superior temporal sulcus ($z$=$9mm$) [**?**] which are often left out when describing face regions.

Figure 4: **Convergence plot of optimization path on fMRI data**: log relative primal objective as a function of time for a series of warm-restarted optimizations of variation lasso with decreasing regularization parameter $\lambda$. Use of DPP screening rules yields significant speedup.

effect is also alleviated, leading to relatively smooth maps. Finally, the last row of Fig.**??** show the average maps using the sparse variation penalty. At the optimally predictive parameters, the noise level in the map is at a minimum compared to the other three. Selected regions are predominantly spatially contiguous and vary smoothly within. We can see that while posterior averaging alleviates the blockiness of TV-$\ell_1$, sparse variation segments large regions and has a cleaner background. Sparse variation, TV-$\ell_1$, and to a large extent variation lasso segment well-delineated, albeit small regions, that correspond to the various brain modules recruited in object recognition.

Decoding the 8 classes of objects from the brain images is non trivial, as objects like shoes and scissors probably have very similar neural coding. In terms of prediction accuracy, in a nested cross-validation to select penalty parameters, lasso yields 69.4% (standard error 2.61%) of correct classification in the 8-class prediction problem. TV-$\ell_1$, sparse variation and variation lasso respectively yield 75.3% (1.28%), 76.2% (1.37%), and 77.0% (2.18%). Thus, imposing spatial structure is important for prediction, but the difference between the penalties that we propose is not significant.

**Computation time: speed up in variation lasso** For adoption by the practitioner, computation speed is critical: a 3-fold cross-validation to select parameters on a grid of 6 values for $\mu$ and 10 values for $\lambda$ entails fitting 180 models. In addition, it is most often used in a nested cross-validation. Coordinate-descent algorithms that are fast on dense and very correlated matrices, as they avoid the costly matrix product. Second, as the solution we seek is very sparse, the screening rules vastly reduce the size of the problem. Indeed, in the fMRI segmentation example, the segmented regions are relatively small compared to the number of voxels in the image. Fig. **??** shows convergence of a variation lasso path with and without screening. With variable screening, the path was completed in around 1 hour, without variable screening, it took more than 4 hours on a 2.70 GHz *Xeon 2* CPU.

## 6 Conclusion

To recover brain functional regions from fMRI, we introduced new penalties that segment non-constant regions on a sparse background. These are structure-inducing, in the sense that they enforce joint sparsity of variables, but unlike structured sparse penalties, they do not operate on previously-known structure. They lead to solvers that are computationally efficient on designs made of large dense 3D images, as in fMRI, by leveraging either an optimization of the data-fit in the dual or screening rules. Interestingly, we introduced a synthesis-sparsity variant of TV and found empirically that it performs well. For fMRI, these penalties stabilize the estimation of the highly-ill-posed linear model by introducing in it the final goal of segmenting functional modules, but they do not lead to spatial smoothing, which is the standard practice. In predictive modeling, this approach grounds the prediction on functional brain regions and we have seen that applying it to a simple dataset studying the visual cortex segments many known sub-systems involved in object recognition.