

Total variation regularization for fMRI-based prediction of behaviour.

Vincent Michel, Alexandre Gramfort, Gaël Varoquaux, Evelyn Eger, Bertrand Thirion

▶ **To cite this version:**

Vincent Michel, Alexandre Gramfort, Gaël Varoquaux, Evelyn Eger, Bertrand Thirion. Total variation regularization for fMRI-based prediction of behaviour.. IEEE Transactions on Medical Imaging, 2011, 30 (7), pp.1328 - 1340. 10.1109/TMI.2011.2113378 . inria-00563468

HAL Id: inria-00563468
https://inria.hal.science/inria-00563468

Submitted on 5 Feb 2011

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Total variation regularization for fMRI-based prediction of behaviour

Vincent Michel*‡[1], Alexandre Gramfort*[1], Gaël Varoquaux*, Evelyn Eger† and Bertrand Thirion*
*INRIA, Saclay-Île-de-France, Parietal team, France - CEA/ DSV/ I2BM/ Neurospin/ LNAO
† INSERM U562, France - CEA/ DSV/ I2BM/ Neurospin/ Unicog
‡Corresponding author - vincent.michel@inria.fr

***Abstract*—While medical imaging typically provides massive amounts of data, the extraction of relevant information for predictive diagnosis remains a difficult challenge. Functional *MRI* (*fMRI*) data, that provide an indirect measure of task-related or spontaneous neuronal activity, are classically analyzed in a mass-univariate procedure yielding statistical parametric maps. This analysis framework disregards some important principles of brain organization: population coding, distributed and overlapping representations. Multivariate pattern analysis, *i.e.*, the prediction of behavioural variables from brain activation patterns better captures this structure. To cope with the high dimensionality of the data, the learning method has to be regularized. However, the spatial structure of the image is not taken into account in standard regularization methods, so that the extracted features are often hard to interpret. More informative and interpretable results can be obtained with the ℓ_1 norm of the image gradient, *a.k.a.* its Total Variation (TV), as regularization. We apply for the first time this method to fMRI data, and show that TV regularization is well suited to the purpose of brain mapping while being a powerful tool for brain decoding. Moreover, this article presents the first use of TV regularization for classification.**

***Index Terms*—fMRI; regression; classification; regularization; Total Variation; spatial structure**

I. Introduction

Functional Magnetic Resonance Imaging (or fMRI) has been widely used for more than fifteen years for neuroscientific and cognitive studies. The analysis of these data largely relies on the general linear model (GLM), introduced for functional imaging by Friston et al. [1]. The GLM is a simple yet powerful framework for deciding which brain regions exhibit a significantly positive task-related effect. This inference, also called *classical inference*, is based on statistical tests applied to each voxel separately, yielding significance maps (*a.k.a.* Statistical Parametric Maps - SPMs) for the effects under consideration. However, despite its simplicity and the accuracy of the SPMs, classical inference suffers from a major drawback: it analyzes each voxel separately and consequently cannot fully exploit the correlations existing between different brain regions to improve the inference. Spatial information is only taken into account in testing procedures, *e.g.* by using the cluster size tests in Random Field Theory.

Correlations between brain activations are likely to arise as a consequence of processing in distributed populations of neurons [2], [3], [4]. This is particularly the case in *population coding* models [5], [6]. For the purpose of statistical inference, these models suggest that effects that differ between experimental conditions are not optimally characterized by the effect significance at individual voxels [7], and that one should rather consider the combined information from different voxels/regions of the brain [8]. Moreover, statistical power in the case of classical inference is limited by the multiple comparison problem (one statistical test is performed for each voxel and the number of comparisons has to be corrected for).

Recently, the inference of behavioral information or cognitive states from brain activation images such as those obtained with fMRI has emerged as an alternative neuroimaging data analysis paradigm [9], [10], [11]. It can be used to assess the specificity of several brain regions for certain cognitive or perceptual functions, by evaluating the accuracy of the prediction of a behavioral variable of interest – the *target* – based on the activations measured in these regions. This inference relies on a prediction function, the accuracy of which depends on whether it uses the relevant variables, *i.e.*, the correct brain regions. This approach, called *inverse inference*, has some major advantages:

- As multivariate approach, it is consistent with population coding models. Indeed, the neural information, which can be encoded by different populations of neurons, can be decoded using a *pattern* of voxels [9], [12].
- It avoids the multiple comparison issue, as it performs only one statistical test (on the predicted behavioral variable). In that sense, it can detect significant links between image data and target that would not have been detected by standard statistical parametric mapping procedures [13]; note however that the statistical interpretation of these two tests are clearly different.
- It addresses new challenges, in particular by allowing to identify a new stimulus in a large dataset, based on already seen stimuli (as visual stimuli [14], or nouns associated with new images [15]). Moreover, it can be used for the more challenging generalization of the prediction to unknown high level stimuli [16], which opens a deeper understanding on brain functional organization.

Many machine learning methods have been applied to fMRI activation images. Among them are linear discriminant analysis [9], support/relevance vector machines [10], neural networks [17], *Lasso* [18], elastic net regression [19], kernel

[1] Contributed equally.

ridge regression [20], boosting [21], sparse logistic regression [22], [23] or Bayesian regularization [24], [25], [26]. Moreover, fMRI data are intrinsically smooth, so that their spatial structure has to be taken into account. Spatial information has thus been considered within the inverse inference framework, by using specific priors in a Bayesian framework [27] or by creating spatially informed features [28]. In the inverse inference problem the main objective remains the extraction of informative regions within the brain volume (see [12] for a review). Besides prediction accuracy, an even greater challenge in brain functional imaging, is the ability of the method to provide an interpretable model (see e.g. [19]). Ultimately, the predictive function learned from the data should be as explicit as standard statistical mapping results. This double objective is addressed by the present contribution.

In practice, selecting the relevant voxels – called features in machine learning – is fundamental in order to achieve accurate prediction. However, when the number of *features* (voxels) is much larger than the numbers of *samples* (images), the prediction method may overfit the training set. In other words, it fits seemingly predictive information from noise in the training set, and thus does not generalize well to new data. To address this issue, one can reduce the number of features. A classical strategy consists of preceding the learning algorithm with a feature selection procedure that drastically reduces the spatial support of predictive regions. To date, the most widely used method for feature selection is voxel-based *Anova* (Analysis of Variance), that evaluates each voxel independently. This is often combined with the use of Support Vector Machine as prediction function (see [29], [10], [30], [31], [32]). An alternative approach consists in performing the model estimation by taking the high dimensional data as input while using relevant regularization methods. These regularizations are performed with two possible goals: stabilizing the estimation of the weights of the features, and/or forcing a majority of features to have close to zero weights (*i.e.* promoting sparsity).

Let us introduce the following predictive linear model:

$$\mathbf{y} = f(\mathbf{X}, \mathbf{w}, b) = F(\mathbf{X}\mathbf{w} + b) \ , \tag{1}$$

where $\mathbf{y}$ represents the behavioral variable and $(\mathbf{w}, b)$ are the parameters to be estimated on a training set. A vector $\mathbf{w} \in \mathbb{R}^p$ can be seen as an image; p is the number of features (or voxels) and $b \in \mathbb{R}$ is called the *intercept*. The matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$ is the design matrix. Each row is a p-dimensional sample, *i.e.*, an activation map related to the observation. It has been shown [9], [10] that using a non-linear classifier does not improve the prediction accuracy, and yields interpretation issues. Thus, we only focus on linear classifiers in this paper. Depending on whether the variable to be predicted takes scalar or discrete values, the learning problem is either a regression or a classification problem. In a linear regression setting, f reads:

$$f(\mathbf{X}, \mathbf{w}, b) = \mathbf{X}\,\mathbf{w} + b \ , \tag{2}$$

with $\mathbf{y} \in \mathbb{R}^n$. In the case of classification with a linear model, f is defined by:

$$f(\mathbf{X}, \mathbf{w}, b) = \text{sign}(\mathbf{X}\,\mathbf{w} + b) \ , \tag{3}$$

where "sign" denotes the sign function and $\mathbf{y} \in \{-1, 1\}^n$.

The crucial issue here is that $n \ll p$, so that estimating $\mathbf{w}$ is an ill-posed problem. The estimation requires therefore adapted regularization. A standard approach to perform the estimation of $\mathbf{w}$ with regularization uses penalization of a maximum likelihood estimator. It leads to the following minimization problem:

$$\hat{\mathbf{w}} = \underset{\mathbf{w},b}{\text{argmin}} \ \mathcal{L}(\mathbf{y}, F(\mathbf{X}\mathbf{w} + b)) + \lambda J(\mathbf{w}) \ \ , \ \ \lambda \geq 0 \tag{4}$$

where $\lambda J(\mathbf{w})$ is the regularization term and $\mathcal{L}(\mathbf{y}, F(\mathbf{X}\mathbf{w}+b))$ is the loss function. The parameter λ balances the loss function and the penalty $J(\mathbf{w})$. Note that the intercept b is not included in the regularization term.

The use of the intercept is fundamental in practice as it allows the separating hyperplane to be offset from 0. However for the sake of simplicity in the presentation of the method, we will from now on consider b as an added coefficient in the vector $\mathbf{w}$. This is classically done by concatenation of a column filled with 1 in the matrix $\mathbf{X}$. The loss function will also be abbreviated $\mathcal{L}(\mathbf{w})$.

In the formalism of (4), the reference method is *elastic net* [33], which is a combined ℓ_1 and ℓ_2 penalization:

$$\lambda J(\mathbf{w}) = \lambda_1 \|\mathbf{w}\|_1 + \lambda_2 \|\mathbf{w}\|_2^2 = \sum_{i=1}^{p} \lambda_1 |w_i| + \lambda_2 w_i^2 \tag{5}$$

Elastic net has two limit cases: $\lambda_2 = 0$ is the *Lasso* [34] which yields an extreme sparsity in the selected features, and $\lambda_1 = 0$ corresponds to *Ridge regression* [35].

A major limitation of the methods cited above, including the latter penalization, is that they do not take into account the underlying structure of $\mathbf{w}$. In the case of brain images, $\mathbf{w}$ is defined on a spatial 3-dimensional grid. The main motivation for using this spatial structure is that the predictive information is most likely organized in regions, and not randomly spread across voxels [36], [28]. As it is demonstrated in this contribution, one can both decrease the complexity of the results (*i.e* increase the *interpretability* of the results by extracting a small set of spatially coherent regions of interest) as well as increase the *accuracy* of the prediction by taking into account the spatial relations between voxels.

In this article, we develop an approach for regularized prediction based on Total Variation (TV), $J(\mathbf{w}) = TV(\mathbf{w})$. TV, mathematically defined as the ℓ_1 norm of the image gradient, has been primarily used for image denoising [37], [38] as it preserves edges. The motivation for using TV for brain imaging is that it promotes estimates $\hat{\mathbf{w}}$ of $\mathbf{w}$ with a block structure, creating regions with piecewise constant weights, and therefore outlining the brain regions correlated to the target behavioral variable. Indeed, we are expecting that the spatial layout of the neural code is sparse and spatially structured in the sense that non-zero weights are grouped into connected clusters. Weighted maps showing such characteristics will be called interpretable, as they fulfill our hypothesis on the spatial layout of neural coding [39]. This approach is closely related to the one developed in [40], that introduce proximity information about the features in the regularization term.

In this contribution, the mathematical and implementation details of *TV regression/classification* are first detailed. As far as we know, the present work is the first to use TV in the context of image classification and also the first one to propose the use of the image structure in the learning framework of (4) in the context of fMRI inverse inference. We apply both TV regression and TV classification to an fMRI paradigm that studied the processing of object shape and size in the human brain. Results show that TV outperforms other state of the art methods, as it yields better prediction performance while providing weights $\hat{\mathbf{w}}$ with an interpretable spatial structure.

II. Total Variation and prediction

We first detail the notations of the problem. We then develop the TV regularization framework. Finally, we detail the algorithm used for regression and classification.

A. Notations

Let us define $\Omega \subset \mathbb{R}^3$ the 3D image domain, discretized on a finite grid. The coefficients $\mathbf{w}$ define a function from Ω to $\mathbb{R}$, *i.e.*, $\mathbf{w} : \Omega \to \mathbb{R}$. Its TV reads:

$$\begin{aligned} TV(\mathbf{w}) &= \sum_{\omega \in \Omega} \|\nabla \mathbf{w}\|(\omega) \\ &= \sum_{\omega \in \Omega} \sqrt{\nabla_x \mathbf{w}(\omega)^2 + \nabla_y \mathbf{w}(\omega)^2 + \nabla_z \mathbf{w}(\omega)^2} \end{aligned}$$

Let us assume that ω stands for the voxel at position (i, j, k), away from the border of Ω, then $\nabla_x \mathbf{w}(\omega)^2$ corresponds to $(\mathbf{w}_{i+1,j,k} - \mathbf{w}_{i,j,k})^2$ (see appendix A for more details). TV can be used with different discretizations, such as an anisotropic discretization. However, such a discretization is biased in the direction of the axes of the image, which is problematic especially with a strong regularization. Indeed, an isotropic discretization promotes sparse gradient along the image axes. We use therefore the standard isotropic discretization of TV [38], [41].

We denote $\mathbf{y} \in \mathbb{R}^n$ the targets to be predicted, and $\mathbf{X} \in \mathbb{R}^{n \times p}$ the set of activation images related to the presentation of different stimuli. The integer p is the number of voxels and n the number of samples (images). Typically, $p \sim 10^3$ to 10^5 (for a whole volume), while $n \sim 10$ to 10^2. We denote M the mask of the brain that comes from standard fMRI analysis, and that is used to avoid computation outside of the brain volume. M is a $p_i \times p_j \times p_k$ three dimensional grid, with:

$$\begin{cases} M_{i,j,k} = 1 & \text{if the voxel is in the mask} \\ M_{i,j,k} = 0 & \text{if the voxel is not in the mask} \end{cases}$$

with $\sum_{i,j,k} M_{i,j,k} = p$. Additionally, we define grad : $\mathbb{R}(\Omega) \to \mathbb{R}^3(\Omega)$ a gradient operator and div : $\mathbb{R}^3(\Omega) \to \mathbb{R}(\Omega)$ the associated adjoint divergence operator (the adjoint operator is used in the convex optimization algorithm, see appendix A for more details, in particular Eq. 13).

Let K the convex set defined by:

$$K = \{g : \Omega \to \mathbb{R}^3 \mid \forall \omega \in \Omega, \|g(\omega)\| \leq 1\}$$

and Π_K the projection operator onto the set K:

$$\begin{cases} \Pi_K(g)(\omega) = g(\omega) & \text{if } \|g(\omega)\| \leq 1 \\ \Pi_K(g)(\omega) = g(\omega)/\|g(\omega)\| & \text{otherwise.} \end{cases}$$

This projection operator will be used in the optimization loop solving Eq. 2, to apply the constraint. It can be viewed as the projection on the ℓ_∞ norm (dual of the ℓ_1 norm) ball.

B. Convex optimization

We consider the minimization problem (4). When $J(\mathbf{w})$ is non-smooth (*i.e.* not differentiable), an analytical solution does not exist and the optimization can unfortunately not be performed with simple algorithms such as Gradient descent and Newton method. This is for example the case with $J(\mathbf{w}) = \|\mathbf{w}\|_1$ (ℓ_1 norm *a.k.a. Lasso* penalty) and with $J(\mathbf{w}) = TV(\mathbf{w})$, both of which require advanced optimization strategies.

A recently studied strategy ([42], [43], [44], [45]) is based on iterative procedures involving the computation of *proximity operators* (see def. 1) [46]. Such approaches are adapted to composite problems with both a smooth term and a non-smooth term as it is the case here (see [47] for a recent review). In the context of neuroimaging, such optimization schemes have been proposed recently in order to solve the inverse problem of magneto- and electro-encephalography (collectively M/EEG) when considering non ℓ_2 priors [48], [49].

Definition 1 (Proximity operator)**.** *Let $J : \mathbb{R}^p \to \mathbb{R}$ be a proper convex function. The proximity operator associated with J and $\lambda \in \mathbb{R}_+$ denoted by $prox_{\lambda J} : \mathbb{R}^p \to \mathbb{R}^p$ is given by:*

$$prox_{\lambda J}(\mathbf{w}) = \arg\min_{\mathbf{v} \in \mathbb{R}^p} \left(\frac{1}{2}\|\mathbf{v} - \mathbf{w}\|_2^2 + \lambda J(\mathbf{v}) \right)$$

The iterative procedure known as *ISTA* (*Iterative Shrinkage-Thresholding Algorithm, a.k.a Forward-Backward iterations*) [42], [43], is based on the alternate minimization of the *loss* term $\mathcal{L}(\mathbf{w})$, by gradient descent, and the penalty $J(\mathbf{w})$, by computing a proximity operator. One can show (see appendix B for a sketch of the proof), that this can be done in one single step by iterating:

$$\mathbf{w}^{(k+1)} = \text{prox}_{\lambda J/L} \left(\mathbf{w}^{(k)} - \frac{1}{L} \nabla \mathcal{L}(\mathbf{w}^{(k)}) \right) \quad , \qquad (6)$$

where $\frac{1}{L}\nabla\mathcal{L}(\mathbf{w}^{(k)})$ is the gradient descent term with a stepsize $\frac{1}{L}$, $\text{prox}_{\lambda J/L}$ is the proximity operator of the penalty and the scalar L is an upper bound on the *Lipschitz constant* L_0 of the gradient of the loss function. The pseudo code of the *ISTA* procedure is defined in Algo. 1.

Inspired by previous findings [44], the *FISTA* (*Fast Iterative Shrinkage-Thresholding Algorithm*) procedure [45], [50] has been developed to speed up the convergence of *ISTA*. While *ISTA* converges in $\mathcal{O}(1/K)$, *FISTA* is proven to converge in $\mathcal{O}(1/K^2)$, where K is the number of iterations. The pseudo code of the *FISTA* procedure is given in Algo. 2. The main

improvement in *FISTA* is to compute the next descent direction using the previous one. Such an idea is also present in the well known conjugate gradient algorithm that uses all previous iterates to compute the next descent direction.

Algorithm 1: *ISTA* procedure

Compute the Lipschitz constant L_0 of the operator $\nabla\mathcal{L}$.
Initialize $\mathbf{w}^{(0)} \in \mathbb{R}^p$
repeat
  $\mathbf{w}^{(k+1)} = \text{prox}_{\lambda J/L}\left(\mathbf{w}^{(k)} - \frac{1}{L}\nabla\mathcal{L}(\mathbf{w}^{(k)})\right)$
  where $L > L_0$.
until *convergence*;
return $\mathbf{w}$

Algorithm 2: *FISTA* procedure

Compute the Lipschitz constant L_0 of the operator $\nabla\mathcal{L}$.
Initialize $\mathbf{w}^{(0)} \in \mathbb{R}^p$, $\mathbf{v}^{(1)} = \mathbf{w}^{(0)}$ and $t_1 = 1$.
repeat

$$\begin{aligned}\mathbf{w}^{(k)} &= \text{prox}_{\lambda J/L}\left(\mathbf{v}^{(k)} - \frac{1}{L}\nabla\mathcal{L}(\mathbf{v}^{(k)})\right)\\ t_{k+1} &= \frac{1+\sqrt{1+4t_k^2}}{2}\\ \mathbf{v}^{(k+1)} &= \mathbf{w}^{(k)} + \left(\frac{t_k - 1}{t_{k+1}}\right)(\mathbf{w}^{(k)} - \mathbf{w}^{(k-1)})\end{aligned}$$

until *convergence*;
return $\mathbf{w}$

Let us introduce now the notion of *Duality gap*. The duality gap is a natural stopping condition for approaches as *ISTA* and *FISTA*. In practice, if the duality gap is below a value $\epsilon > 0$, it guarantees that the solution obtained is ϵ-optimal, *i.e.*, that the value of the cost-function reached by the algorithm is not greater than ϵ more the globally optimal value. A comprehensive presentation of this notion [51] is beyond the scope of this paper, and we now give some details in the particular case of the proximity operator $\text{prox}_{\lambda TV}$ known as the *ROF* problem [37] (named after the authors L. Rudin, S. Osher and E. Fatemi) in the image processing literature.

The computation of $\text{prox}_{\lambda TV}$ and the associated duality gap requires the derivation of a Lagrange dual problem [51].

Proposition 2 ($\text{prox}_{\lambda TV}$ Dual problem)**.** *A dual problem associated with* $prox_{\lambda TV}$ *is given by*

$$\mathbf{z}^* = \underset{\mathbf{z}\in K}{\text{argmax}} -\|div\,\mathbf{z} + \mathbf{w}/\lambda\|_2^2 \ , \tag{7}$$

where z *is the dual variable that satisfies* $\mathbf{v}^* = \mathbf{w} + \lambda\, div\,\mathbf{z}^*$, *with* $\mathbf{v}^* = prox_{\lambda TV}(\mathbf{w})$

This result is adapted from [38] (see appendix C for a sketch of the proof). The problem (7) is a maximization of a smooth concave function over a convex set. As shown in [50], it can be solved with the *FISTA* iterative procedure. The resolution of the *ROF* problem is therefore achieved by solving the dual problem. Once $\mathbf{z}^*$ is obtained, $\mathbf{v}^* = \text{prox}_{\lambda TV}(\mathbf{w})$ is given by $\mathbf{v}^* = \mathbf{w} + \lambda\,\text{div}\,\mathbf{z}^*$.

The latter result also gives an estimate of the duality gap.

Proposition 3 (Duality gap)**.** *The duality gap* δ_{gap} *associated with the* ROF *problem is given by:*

$$\delta_{gap}(\mathbf{v}) = \frac{1}{2}\|\mathbf{w} - \mathbf{v}\|_2^2 + \lambda TV(\mathbf{v}) - \frac{1}{2}(\|\mathbf{w}\|_2^2 + \|\mathbf{v}\|_2^2) \geq 0 \ , \tag{8}$$

where the primal variable $\mathbf{v}$ *is obtained during the iterative procedure from the current estimate of the dual variable* $\mathbf{z}$ *with* $\mathbf{v} = \mathbf{w} + \lambda\, div\,\mathbf{z}$.

See appendix C for more details. This duality gap will be used as a stopping criterion for the *FISTA* procedure solving the *ROF* problem. At each iteration of the *FISTA* procedure, we will stop the iterative loop if the duality gap is below a given threshold ϵ. In practice, ϵ is set to $10^{-4} \times \|\mathbf{w}\|_2^2$ to be invariant to the scaling of the data.

Note that the *ROF* problem can be also solved using very efficient combinatorial optimization methods [41], when using the anisotropic discretization of TV.

C. Prediction framework

We now detail the original contribution of this work, that is the construction of a predictive framework using the TV regularization. For $J(\mathbf{w}) = TV(\mathbf{w})$, the global algorithm for solving the minimization problem defined in (4) consists in a *FISTA* procedure (resolution of the *ROF* problem) nested inside an *ISTA* procedure (resolution of the main minimization problem). The *FISTA* procedure is performed at each step of *ISTA* with a *warm restart* on the dual variable $\mathbf{z}$. We do not use *FISTA* for solving the main minimization problem, as this procedure requires an exact proximity operator. The resolution of the *ROF* problem only leads to an ϵ-optimal solution. The pseudo-code of the global algorithm for the TV regularization is provided in Table 3.

A difficulty specific to fMRI data is the computation of the gradient and divergence over a mask of the brain with correct border conditions (see appendix A for details). Moreover, with such an irregular domain, the upper bound $\tilde{L}$ for the Lipschitz constant of the *FISTA* procedure also needs to be estimated on each input data. To do this we use a power method that is classically used to estimate the spectral norm of a linear operator, here equal to the Laplacian $\Delta : \Omega \to \Omega$ defined by $\Delta(\omega) = \text{div}(\text{grad}(\omega))$.

TV regression: The regression version of the TV is called *TV regression*. In this case, we use the least-squares *loss*:

$$\begin{cases}\mathcal{L}(\mathbf{w}) = \frac{1}{2n}\|\mathbf{y} - \mathbf{X}\mathbf{w}\|^2\\ \nabla\mathcal{L}(\mathbf{w}) = -\frac{1}{n}\mathbf{X}^T(\mathbf{y} - \mathbf{X}\mathbf{w})\end{cases}$$

The Lipschitz constant L_0 of the operator $\nabla\mathcal{L}$ is $L_0 = |||\mathbf{X}^T\mathbf{X}|||/n$, where $|||.|||$ stands for the spectral norm equal to largest singular value. The constant L is set in practice to $L = 1.1L_0$.

TV classification: The classification version of the TV is called *TV classification*. This algorithm is based on a logistic loss [52]. We now give the mathematical formulation for the binary case with $\mathbf{y} \in \{-1, 1\}^n$. The logistic regression model

Algorithm 3: TV regularization solver

Set maximum number of iterations K (*ISTA*).
Set the threshold ϵ on the dual gap (*FISTA*).
Set $L = 1.1L_0$ where L_0 is the Lipschitz constant of $\nabla\mathcal{L}$.
Set $\tilde{L} = 1.1\tilde{L_0}$ where $\tilde{L_0}$ is the Lipschitz constant of the Laplacian operator $\Delta : w \in \mathbb{R}(\Omega) \rightarrow \text{div}(\text{grad}(w))$.
Initialize $\mathbf{z} \in \mathbb{R}(\Omega^3)$ with zeros.
ISTA loop
for $k = 1 \dots K$ **do**
  $\mathbf{u} = \mathbf{w} - \frac{1}{L}\nabla\mathcal{L}(\mathbf{w})$
  ### FISTA loop ###
  Initialize $\mathbf{z}_{aux} = \mathbf{z}$, $t = 1$
  repeat
    $\mathbf{z}_{old} = \mathbf{z}$
    $\mathbf{z} = \Pi_K\left(\mathbf{z}_{aux} - \frac{1}{\lambda L}\text{grad}(L\mathbf{u} + \lambda\text{div}(\mathbf{z}_{aux}))\right)$
    $t_{old} = t$
    $t = (t + \sqrt{1 + 4t^2})/2$
    $\mathbf{z}_{aux} = \mathbf{z} + \frac{t_{old}-1}{t}(\mathbf{z} - \mathbf{z}_{old})$
  until $\delta_{gap}(\mathbf{u} + \lambda div(\mathbf{z})) \leq \epsilon$;
  $\mathbf{w} = \mathbf{u} + \lambda\text{div}(\mathbf{z})$
return $\mathbf{w}$

defines the conditional probability of y_i given the data $\mathbf{x}_i$ as:

$$p(y_i|\mathbf{x}_i, \mathbf{w}) = \frac{1}{1 + \exp^{-y_i(\mathbf{x}_i^T\mathbf{w})}} \tag{9}$$

The corresponding loss and the loss gradient read:

$$\begin{cases} \mathcal{L}(\mathbf{w}) = \frac{1}{n}\sum_{i=1}^{n}\log\left(1 + \exp^{-y_i(\mathbf{x_i}^T\mathbf{w})}\right) \\ \nabla\mathcal{L}(\mathbf{w}) = -\frac{1}{n}\sum_{i=1}^{n}\frac{y_i\mathbf{x_i}}{1+\exp^{y_i(\mathbf{x_i}^T\mathbf{w})}} \end{cases}$$

The Lipschitz constant L_0 of the operator $\nabla\mathcal{L}$ is $L_0 = \|\mathbf{X}\|^2/(4n)$. The classification framework developed in this paper treats the binary case with a logistic model, *a.k.a.*, binomial model. In our analysis, we expand this framework to multiclass classification using a one-versus-one voting heuristic. The number of classifiers used is $(k) \times (k-1)/2$, where k is the number of classes. The predicted class is then selected as the class which yields the highest probability across the predictions of all of the classifiers, as defined in (9). Note that a multinomial approach could also be used [53]. However the resulting weights $\mathbf{w}$ become impossible to interpret, so that the multinomial model may not adapted to the applicative context. Indeed, with three classes, one gets two hyperplanes from which it is hard to draw any neuroscientific conclusions. The weights of each binary classifier have a simpler meaning. This one-versus-one voting heuristic is the one used in LibSVM [54].

D. Performance evaluation

Our method is evaluated with a cross-validation procedure that splits the available data into training and validation sets. In the following, $(\mathbf{X}^l, \mathbf{y}^l)$ are a learning set, $(\mathbf{X}^t, \mathbf{y}^t)$ a test set and $\hat{\mathbf{y}}^t = F(\mathbf{X}^t\hat{\mathbf{w}})$ refers to the predicted target, where $\hat{\mathbf{w}}$ is estimated from the training set.

For regression analysis, the performance of the different models is evaluated using ζ, the ratio of explained variance:

$$\zeta(\mathbf{y}^t, \hat{\mathbf{y}}^t) = \frac{\text{var}(\mathbf{y}^t) - \text{var}\,(\mathbf{y}^t - \hat{\mathbf{y}}^t)}{\text{var}(\mathbf{y}^t)}$$

This is the amount of variability in the response that can be explained by the model (perfect prediction yields $\zeta = 1$, while $\zeta < 0$ if prediction is worse than chance).

For classification analysis, the performance of the different models is evaluated using the classification score denoted κ , classically defined as:

$$\kappa(\mathbf{y}^t, \hat{\mathbf{y}}^t) = \frac{\sum_{i=1}^{n^t}\delta(y_i^t, \hat{y}_i^t)}{n^t}$$

where n^t is the number of samples in the test set, and δ is Kronecker's delta.

The p-values are computed using a Wilcoxon signed-rank test on the prediction score.

E. Competing methods

In our experiments, TV regression is compared to different state of the art regularization methods:

- *Elastic net* regression [33], that requires setting two parameters λ_1 and λ_2 (5). In our analyzes, a cross-validation procedure within the training set is used to optimize these parameters. Here, we use $\lambda_1 \in \{0.2\tilde{\lambda}, 0.1\tilde{\lambda}, 0.05\tilde{\lambda}, 0.01\tilde{\lambda}\}$, where $\tilde{\lambda} = \|\mathbf{X}^T\mathbf{y}\|_\infty$, and $\lambda_2 \in \{0.1, 0.5, 1., 10., 100.\}$ (λ_1 and λ_2 parametrize two different types of norm).
- *Support Vector Regression* (*SVR*) with a linear kernel [55], which is the reference method in neuroimaging. The C parameter is optimized by cross-validation in the range 10^{-3} to 10^1 in multiplicative steps of 10.

TV classification is compared to different state of the art classification methods:

- *Sparse multinomial logistic regression* (*SMLR*) classification [53], that requires a double optimization, for the two parameters λ_1 and λ_2. A cross-validation procedure within the training set is used to optimize these parameters. Here, we use $\lambda_1 \in \{0.2\tilde{\lambda}, 0.1\tilde{\lambda}, 0.05\tilde{\lambda}, 0.01\tilde{\lambda}\}$, where $\tilde{\lambda} = \|\mathbf{X}^T\mathbf{y}\|_\infty$, and $\lambda_2 \in \{0.1, 0.5, 1., 10., 100.\}$.
- *Support Vector Classification* (*SVC*) with a linear kernel [55], which is the reference method in neuroimaging. The C parameter is optimized by cross-validation in the range 10^{-3} to 10^1 in multiplicative steps of 10.

All these methods are used after an *Anova*-based feature selection as this maximizes their performance. Indeed, irrelevant features and redundant information can decrease the accuracy of a predictor [56]. The optimal number of voxels is selected within the range $\{50, 100, 250, 500\}$, through a nested cross-validation within the training set. We do not select directly a threshold on p-value or cluster size, but rather a number of features. Additionally, we check that increasing the range of voxels (*i.e.* adding 2000 in the range of number of selected voxels) does not increase the prediction accuracy on our datasets. The parameter estimation of the learning function is also performed using a nested cross-validation within the

training set, and thus, the cross-validation framework is used rigorously in all the experiments of this paper. All methods are developed in *C* and used in *Python*. The implementation of *Elastic net* is based on *coordinate descent* [57], while *SVR* and *SVC* are based on LibSVM [54]. Methods are used from *Python* via the *Scikit-learn* open source package [58].

III. Experiments

A. Details on simulated data

The simulated data set $\mathbf{X}$ consists of $n = 100$ images (size $12 \times 12 \times 12$ voxels) with a set of four square Regions of Interest (ROIs) (size $2 \times 2 \times 2$). We call $\mathcal{R}$ the support of the ROIs (*i.e.* the 32 resulting voxels of interest). Each of the four ROIs has a fixed weight in $\{-0.5, 0.5, -0.5, 0.5\}$. We call $w_{i,j,k}$ the weights of the (i, j, k) voxel. The resulting images are smoothed with a Gaussian kernel with a standard deviation of 2 voxels, to mimic the correlation structure observed in real fMRI data. To simulate the spatial variability between images (inter-subject variability, movement artifacts in intra-subject variability), we define a new support of the ROIs, called $\tilde{\mathcal{R}}$ such as, for each image l^{th}, 50% (randomly chosen) of the weights $\mathbf{w}$ are set to zero. Thus, we have $\tilde{\mathcal{R}} \subset \mathcal{R}$. We simulate the target $\mathbf{y}$ for the l^{th} image as:

$$y_l = \sum_{(i,j,k)\in\tilde{\mathcal{R}}} w_{i,j,k} X_{i,j,k,l} + \epsilon_l \tag{10}$$

with the signal in the (i, j, k) voxel of the l^{th} image simulated as:

$$X_{i,j,k,l} \sim \mathcal{N}(0, 1) \tag{11}$$

and $\epsilon_l \sim \mathcal{N}(0, \gamma)$ is a Gaussian noise with standard deviation $\gamma > 0$. We choose γ in order to have a signal-to-noise ratio of 5 dB. We compare TV regression cross-validated with different values of λ in the range $\{0.01, 0.05, 0.1, \}$, with the two reference algorithms, *elastic net* and *SVR*. All three methods are optimized by 4-folds cross-validation in the range described below.

B. Details on real data

We apply the different methods on a real fMRI dataset related to an experiment studying the representation of objects, on ten subjects, as detailed in [59]. During this experiment, ten healthy volunteers viewed objects of two categories (each one of the two categories used in equal halves of subjects) with 4 different exemplars each shown in 3 different sizes (yielding 12 different experimental conditions), with 4 repetitions of each stimulus in each of the 6 sessions. We pooled data from the 4 repetitions, resulting in a total of $n = 72$ images by subject (one image of each stimulus by session). Functional images were acquired on a 3-T MR system with eight-channel head coil (Siemens Trio, Erlangen, Germany) as T2*-weighted echo-planar image (EPI) volumes. Twenty transverse slices were obtained with a repetition time of 2 s (echo time, 30 ms; flip angle, 70°; $2 \times 2 \times 2$-mm voxels; 0.5-mm gap). Realignment, normalization to MNI space, and General Linear Model (GLM) fit were performed with the SPM5 software (http://www.fil.ion.ucl.ac.uk/spm/software/spm5). The normalization is the conventional one of SPM (implying affine and non-linear transformations) and not the one using unified segmentation. The normalization parameters are estimated on the basis of a whole-head EPI acquired in addition, and are then applied to the partial EPI volumes. The data are not smoothed. In the GLM, the effect of each of the 12 stimuli convolved with a standard hemodynamic response function was modeled separately, while accounting for serial autocorrelation with an AR(1) model and removing low-frequency drift terms by a high-pass filter with a cut-off of 128 s. The GLM is fitted separately in each session for each subject, and we used in the following analyzes the resulting session-wise parameter estimate images the β-maps are used as rows of $\mathbf{X}$). All the analyzes are performed without any prior selection of regions of interest, and use the whole acquired volume.

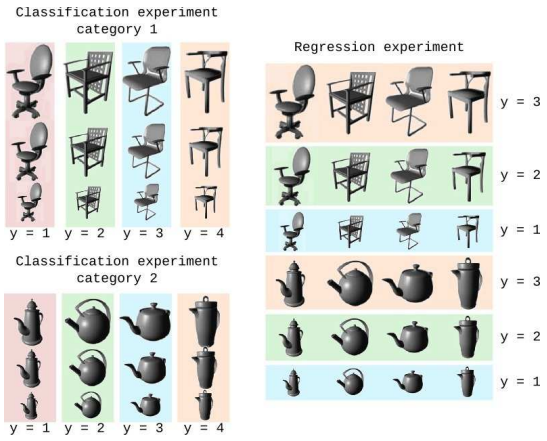


Fig. 1. Experiment paradigm for the classification of object in each of the category (left) and regression (right) experiments. Each color represents the stimuli which are pooled together in one of the three experiments (classification category 1, classification category 2 and regression).

Regression experiments: First, we perform an intra-subject regression analysis. The four different shapes of objects (for the two categories) were pooled across for each one of the three sizes, and we are interested in finding discriminative information between sizes. This reduces to a regression problem, in which our goal is to predict a simple scalar factor (size of an object) (see Fig. 1). Each subject is evaluated independently, in a 12-fold cross-validation. The dimensions of the real data set for one subject are $p \sim 7 \times 10^4$ and $n = 72$ (divided in 3 different sizes, 24 images per size). We evaluate the performance of the method by a leave-one-condition-out cross-validation (*i.e.*, leave-6-images-out), and doing so the GLM is performed separately for the training and test sets. The parameters of the reference methods are optimized with a nested leave-one-condition-out cross-validation within the training set, in the ranges given before.

Additionally, we perform an inter-subject regression analysis on the sizes. The inter-subject analysis relies on subject-specific fixed-effects activations, *i.e.* for each condition, the 6 activation maps corresponding to the 6 sessions are averaged together. This yields a total of 12 images per subject, one for each experimental condition. The dimensions of the real data set are $p \sim 7 \times 10^4$ and $n = 120$ (divided in 3

different sizes). We evaluate the performance of the method by cross-validation (leave-one-subject-out). The parameters of the reference methods are optimized with a nested leave-one-subject-out cross-validation within the training set, in the ranges given before.

Classification experiments: We evaluate the performance on a second type of discrimination which is object classification (see Fig. 1). In that case, we averaged the images for the three sizes and we are interested in discriminating between individual object exemplars/shapes. For each of the two categories, this can be handled as a classification problem, where we aim at predicting the shape of an object corresponding to a new fMRI scan. In order to investigate the performance of TV classification, which is an original contribution, we perform an inter-subject analysis in the same way as described for the regression study, except that now, we perform two analyzes corresponding to the two categories used, each one including 5 subjects.

Statistical Parametric Maps: For comparison purposes, the corresponding maps of *Anova* (*F-score*), or *SPMs*, for the inter-subject analysis are given Fig. 2, for the representation of sizes (top) and representation of objects for the two categories (middle and bottom). As expected, the sizes are mostly processed in primary visual cortex, while for objects, discrimination is additionally observed in lateral occipital regions [59].

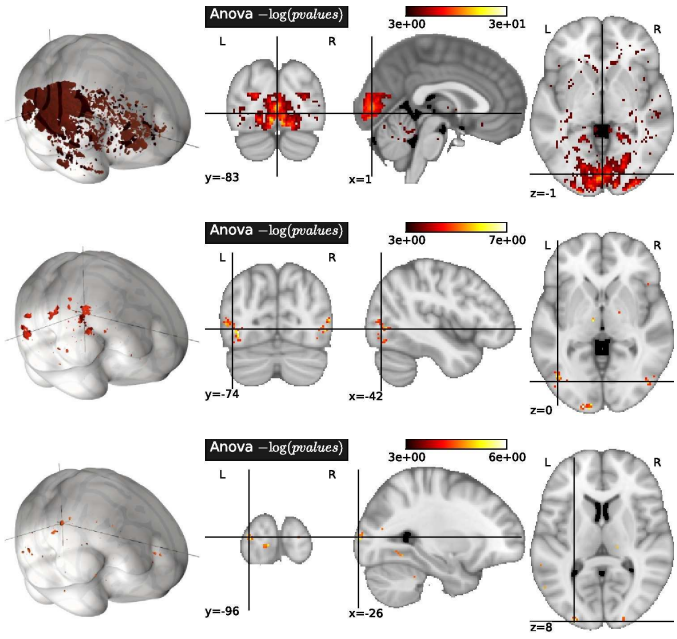


Fig. 2. Inter-subject analysis - Maps of *Anova* ($-\log$(p-values)) for the sizes prediction experiment (top) and the objects identifications for category 1 (middle) and category 2 (bottom). We threshold the p-values higher than 10^{-3} (i.e. $-\log$(p-values) > 3).

IV. Results

A. Results on simulated data

We compare the different methods on the simulated data, see Fig. 3. The true weights (a) and resulting *Anova* F-scores (b) are shown. Only TV regression (e) is able to extract the simulated discriminative regions. *Elastic net* (d) only retrieves part of the support of the weights, and yields an overly sparse solution. This sparsity pattern obtained with *elastic net* is the one that yields the highest prediction accuracy: one could seek a less sparse solution, but this would decrease the prediction accuracy. We note that the weights in the primal space estimated by *SVR* (c) are everywhere non-zero and do not retrieve the support of the weights.

B. Sensitivity study on real data

Before any further analysis on real data, we have performed a sensitivity analysis of our model with regards to the parameter λ. In the inter-subject analysis for the size regression, we compute the cross-validated prediction accuracy for twelve different values of λ between 10^{-4} and 0.95. The aim of the sensitivity study is to assess the stability of the prediction with respect to the regularization parameter. The results, detailed in Fig. 4, are extremely stable with respect to λ in the range $[5.10^{-4}, 5.10^{-1}]$. For this reason, we can fix $\lambda = 0.05$ in the following analyzes. The value of λ is the same for all the experiments, in both classification and regression settings. The correct way of choosing the regularization parameter is to embed the TV regularization within an internal cross-validation on the training set. However, such approach can be computationally costly.

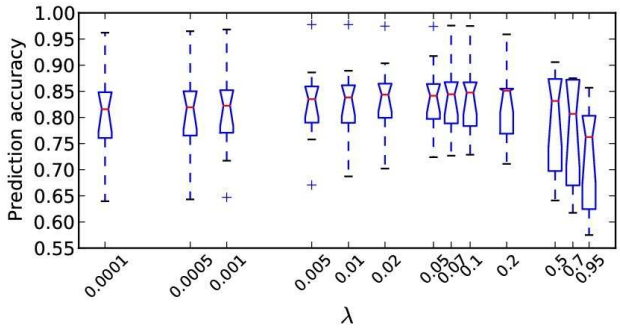


Fig. 4. Explained variance ζ for different values of λ, in the inter-subjects regression analysis. The accuracy is very stable regarding to λ in the range $[5.10^{-4}, 5.10^{-1}]$.

C. Results for regression analysis

In a first set of analyzes, we assess the performance of TV regression in both intra-subject and inter-subject cases, where the aim is to predict the size of an object seen by the subject during the experiment.

Intra-subject analysis: The results obtained by the three methods are given in Tab. I. TV regression outperforms the two alternative methods, yielding an average explained variance of 0.92 across the subjects. The difference with *SVR* is significant, but not with *elastic net*. Moreover, the results of the regularized methods (*TV, elastic net*) are more stable (standard deviation three times smaller) across subjects, than the results of the *SVR*.

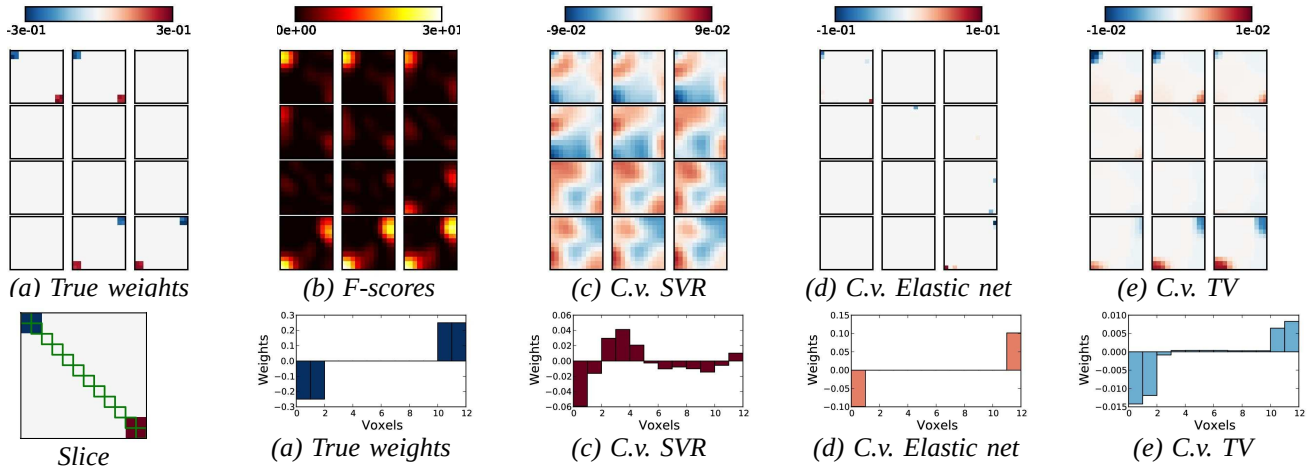


Fig. 3. Two-dimensional slices of the three-dimensional volume of simulated data (top), and weights found on the diagonal (green squares) of the first two-dimensional slice (bottom). Comparisons of the weights found by different methods, with the true target (a), and the *F-score* found by *Anova* (b). The TV method (e) retrieves the true weights. The reference methods ((c), (d)) yield less accurate maps. Indeed, the support of the weights found by *elastic net* is too sparse and does not yield convex regions. *SVR* yields smooth maps that do not look like the ground truth.

Methods	mean ζ	std ζ	max ζ	min ζ	p-value to TV
SVR	0.82	0.07	0.9	0.67	0.0051
Elastic net	0.9	0.02	0.93	0.85	0.0745
TV $\alpha = 0.05$	0.92	0.02	0.95	0.88	-

TABLE I
Regression - Sizes prediction experiment - Intra-subject analysis. EXPLAINED VARIANCE ζ FOR THE THREE DIFFERENT METHODS. TV REGRESSION YIELDS THE BEST PREDICTION ACCURACY, WHILE BEING MORE STABLE THAN THE TWO REFERENCE METHODS (STANDARD DEVIATION OF ζ THREE TIMES SMALLER THAN *SVR*). THE P-VALUES ARE COMPUTED ON THE EXPLAINED VARIANCE OF THE DIFFERENT SUBJECTS.

Inter-subject analysis: The results obtained with the three methods are given in Tab. II. As in the intra-subject analysis, *TV regression* outperforms the two alternative methods, yielding an average explained variance of 84%, and also more stable predictions. Such stability can be illustrated on the subject 3, where both reference methods yield poor results, while TV regression yields an explained variance 0.2 higher. Moreover, we have tested that feature selection minimizes overfitting. Indeed, without such feature selection, we obtain a smaller explained variance of 76% for *SVR* and 64% for *elastic net*.

Methods	mean ζ	std ζ	max ζ	min ζ	p-value to TV
SVR	0.77	0.11	0.97	0.58	0.0284
Elastic net	0.78	0.1	0.97	0.65	0.0469
TV $\lambda = 0.05$	0.84	0.07	0.97	0.72	-

TABLE II
Regression - Sizes prediction experiment - Inter-subject analysis. EXPLAINED VARIANCE ζ FOR THE THREE DIFFERENT METHODS. TV REGRESSION STILL YIELDS THE BEST PREDICTION ACCURACY, WITH AN EXPLAINED VARIANCE 0.06 HIGHER THAN THE BEST REFERENCE METHOD (*elastic net*). THE P-VALUES ARE COMPUTED ON THE EXPLAINED VARIANCE OF THE DIFFERENT SUBJECTS.

The weighted maps found by the different methods are given in Fig. 5. One can notice that, as λ increases, the spatial support of these maps tends to be aggregated in few clusters within the occipital cortex, and that the maps have a nearly constant value on these clusters. By contrast, both reference methods yield uninterpretable (*i.e.* more complex) maps, with a few informative voxels scattered in the whole occipital cortex. The average positions and the sizes of the three main clusters found by the TV algorithm, using all the subjects, are given Tab. III. TV regression is able to adapt the regularization to tiny regions, yielding ROIs from 25 to 193 voxels. The clusters are found within the occipital cortex. The majority of informative voxels are located in the posterior part of the occipital cortex ($y \leq -90$ mm), most likely corresponding to primary visual cortex, with one additional slightly more anterior cluster in posterior lateral occipital cortex. This is consistent with the previous findings [59] where a gradient of sensitivity to size was observed across object selective lateral occipital ROIs, and the most accurate discrimination of sizes in primary visual cortex.

x (mm)	y (mm)	z (mm)	Sizes (voxels)
24	-92	-16	25
-26	-96	-10	103
16	-96	12	193

TABLE III
INTER-SUBJECT REGRESSION ANALYSIS: POSITIONS AND SIZES OF THE THREE MAIN CLUSTERS FOR THE TV REGRESSION METHOD.

D. Results of classification experiments

In a second analysis, we assess the performance of TV classification in an inter-subject classification analyzes, in which the aim is to predict which of 4 object exemplars is seen by the different subjects.

The results (average across the two categories) found by the three methods are given in Tab. IV. As in the inter-subject regression analysis, the TV-based method outperforms the *SMLR* method. Moreover, it yields an average classification score similar to the *SVC* while being more stable. Seeking clusters of activation thus seems a reasonable way to cope with inter-subject variability. The average number of selections of each voxels within one of the three larger clusters for each

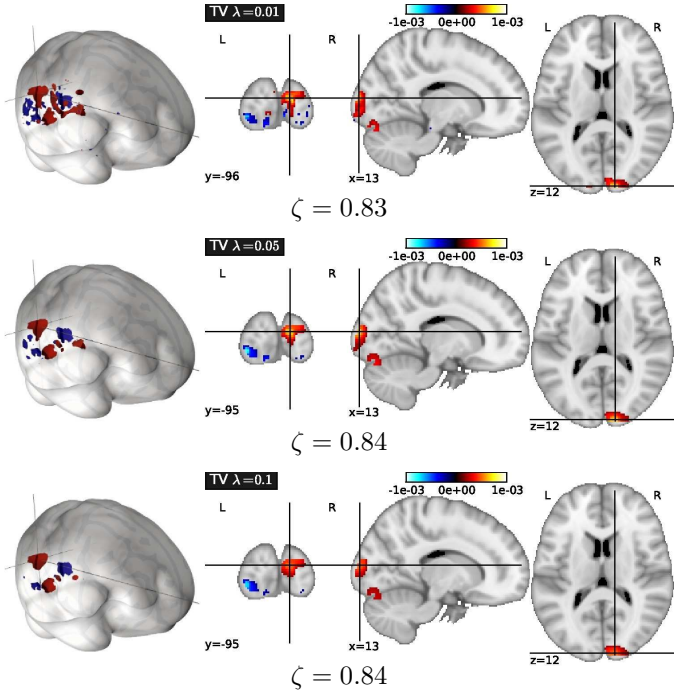


Fig. 5. *Regression - Sizes prediction experiment - Inter-subject analysis*. Maps of weights found by TV regression for various values of the regularization parameter λ. When λ decreases, the TV regression algorithm creates different clusters of weights with constant values. These clusters are easily interpretable, compared to voxel-based map (see below). The TV regression algorithm is very stable for different values of λ, has shown by the explained variance ζ.

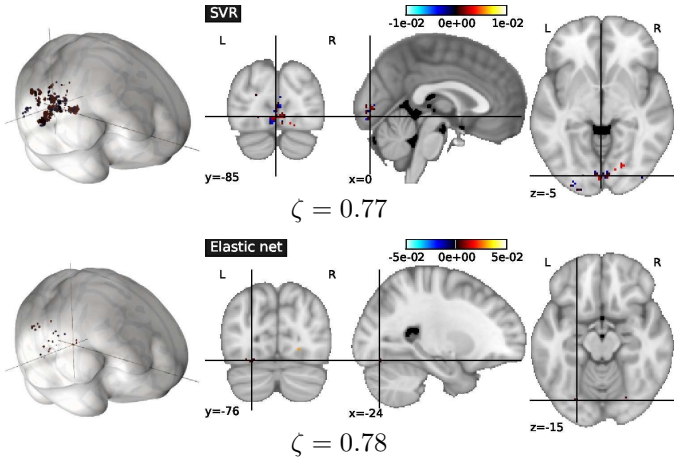


Fig. 6. *Regression - Sizes prediction experiment - Inter-subject analysis*. Maps of weights found by the *SVR* (top) and *elastic net* (bottom) methods. The optimal number of voxels selected by *Anova* is 500, but *elastic net* further reduces this set to 21 voxels. These voxel-based methods do not yield interpretable features (especially when compared to TV regression), which is due to the fact that they do not consider the spatial structure of the image.

one-versus-one map are given Fig. 7. The informative clusters are more anterior and more ventral than the ones found within the sizes prediction paradigm. We thus confirm the results found by classical brain mapping approach, such as *Anova* (see Fig. 2), while providing a classification score based on cross-validation on independent data which allows to check the actual implication of these regions in the cognitive process.

Methods	mean κ	std κ	max κ	min κ	p-value to SVC
SVC	48.33	15.72	75.0	25.0	-
SMLR	42.5	9.46	58.33	33.33	0.0 **
TV $\lambda = 0.05$	46.67	11.3	66.67	25.0	1.0

TABLE IV
Classification - Objects prediction experiment . AVERAGED CLASSIFICATION SCORE κ FOR THE THREE DIFFERENT METHODS, ACROSS THE TWO CATEGORIES. TV CLASSIFICATION YIELDS SIMILAR PREDICTION ACCURACY THAN THE REFERENCE METHOD *SVC*. THE P-VALUES ARE COMPUTED ON THE CLASSIFICATION SCORE OF THE DIFFERENT SUBJECTS.

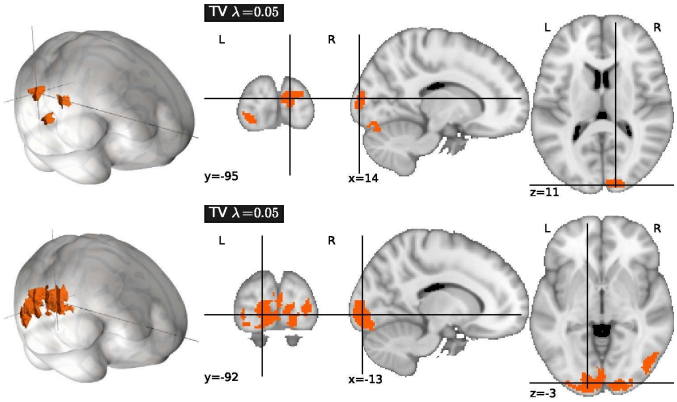


Fig. 7. Inter-subject analysis. Top - voxels selected within one of the three main clusters by TV regression, for the *Sizes prediction experiment*. Bottom - voxels selected at least one time within one of the three main clusters for each of the one-vs-one TV classification, for the *Objects prediction experiment*. Some clusters found in the *Objects prediction experiment* ($y = -40$ mm, $y = -74$ mm) are more anterior than the ones found for the *Sizes prediction experiment* ($y = -92$ mm, $y = -96$ mm). This is coherent with the hypothesis that the processing of shapes is done at a higher level in the processing of visual information, and thus the implied regions are found further in the ventral pathway.

V. DISCUSSION

In this article, we present the first use of TV regularization for brain decoding. This method outperforms the reference methods on prediction accuracy, and yields sparse brain maps with clear informative foci.

Moreover, with regard to the classification paradigm, we integrate the TV in a logistic regression framework. This approach, which to our knowledge, has not been used before, yields high prediction accuracy, and seems to be a promising method for more machine learning problems beyond the scope of neuroimaging.

One major advantage of the proposed method is that, in the case of a multi-subject studies, considering extended regions is expected to compensate for spatial misalignment, hence it can better generalize than voxel-based methods. As proven on both inter-subject analyzes, the proposed TV approach yields significantly higher prediction accuracy than reference voxel-based methods. In addition, the proposed approach yields weight maps very similar to the maps obtained by a classical brain mapping approach (such as *Anova*). We note that the solution found by our method has a sparse block structure and is sufficient for good prediction accuracy, which explains the fact that the regions observed may be more compactly localized than the ones from *Anova*. Thus, the TV approach has the assets of

a predictive framework, while leading to accurate brain maps. It is important to notice that, even if TV does not promote a strict sparsity of the weights, most voxels are associated with very small weights, and only a few clusters get high weightings. Moreover, TV regression allows to consider the whole brain in the analysis, without requiring any prior feature selection. As many accurate dimension reduction approaches such as *Recursive Feature Elimination* [60] can be extremely costly in computational time, avoiding this step is a major asset. An important feature of our implementation is thus that it reduces computation time to a reasonable amount, so that it is not significantly more costly than SVR or elastic net in practical settings (*i.e.*, including the cross-validation loops). In the inter-subject regression analysis, the average computational time is 185 seconds for TV regression, 131 seconds for *Anova + SVR* and 121 seconds for *Anova + elastic net*, on a *Intel(R) Xeon(R) CPU* at 2.83GHz. Regularization of the voxel weights significantly increases the generalization ability in regression problems, by performing feature selection and training of the prediction function jointly. However, to date, regularization has most often been performed without using the spatial structure of the images. By applying a penalization on the gradient of the weight and thus taking into account the spatial structure of the image-based information, our approach performs an adaptive and efficient regularization, while creating sparse weight maps with regions of quasi constant weights. TV regularization method fulfills thus the two requirements that make it suitable for neuroimaging brain mapping: a good prediction accuracy (better than the reference methods for regression experiments, and equal for classification), and a set of interpretable features, made of clusters of similarly-tuned voxels. In that sense, it can be seen as the first method for performing a large scale multivariate brain mapping (the searchlight [61] only consider the multivariate information in a small neighborhood).

From a neuroscientific point of view, the regions extracted from the whole analysis volume in the size discrimination task are concentrated in the early visual cortex. This is consistent with the fact that early visual cortex yields highly reliable signals that are discriminative of feature/shape differences between object exemplars, which holds as long as no high-level generalization across images is required (see *e.g.* [9] and [59]). This is expected given the small receptive fields of neurons in these regions that will reliably detect differences in the spatial envelop or other low-level structure of the images. Most importantly, the predictive spatial pattern is stable enough across individuals to make reliable predictions in new subjects. In fact our method compares best with regards to the state of the art in the inter-subjects setting, because it selects predictive regions that are not very sensitive to anatomo-functional variability. In the object discrimination task, the clusters found by our approach are also in the visual cortex, but including more anterior ones (probably corresponding to posterior *lateral occipital* region) compared to size discrimination, which is consistent with the fact that shape discrimination requires intermediate/higher level visual areas. The finding that also large parts of early visual cortex were discriminative here is explained by the fact that generalization across viewing conditions was not a part of the analysis and classification can therefore be driven by lower-level features. However, even if similar maps as the ones found by our method can be obtained using *Anova*, they do not provide a prediction score for generalization to independent data (*i.e.* a global measure of the involvement of the regions in the cognitive process).

VI. Conclusion

In this paper we introduce TV regularization for extracting information from brain images, both in regression or classification settings. Feature selection and model estimation are performed jointly and capture the predictive information present in the data better than alternative methods. A particularly important property of this approach is its ability to create spatially coherent regions with similar weights, yielding simplified and still informative sets of features. Experimental results show that this algorithm performs well on real data, and is far more accurate than voxel-based reference methods for multi-subject analysis. In particular, the segmented regions are robust to inter-subject variability. These observations demonstrate that TV regularization is a powerful tool for understanding brain activity and spatial mapping of cognitive process, and is the first method that is able to derive statistical weight maps, as in the standard *SPM* approach, within the inverse inference framework.

Acknowledgments: The authors acknowledge support from the ANR grant ViMAGINE ANR-08-BLAN-0250-02.

References

[1] K. J. Friston, A. P. Holmes, K. J. Worsley, J.-B. Poline, C. Frith, and R. S. J. Frackowiak, "Statistical parametric maps in functional imaging: A general linear approach," *Human Brain Mapping*, vol. 2, pp. 189–210, 1995. 1

[2] A. Georgopoulos, A. Schwartz, and R. Kettner, "Neuronal population coding of movement direction," *Science*, vol. 233, no. 4771, pp. 1416–1419, 1986. 1

[3] D. Y. Tsao, W. A. Freiwald, R. B. H. Tootell, and M. S. Livingstone, "A Cortical Region Consisting Entirely of Face-Selective Cells," *Science*, vol. 311, no. 5761, pp. 670–674, 2006. 1

[4] A. Nevado, M. P. Young, and S. Panzeri, "Functional imaging and neural information coding," *NeuroImage*, vol. 21, no. 3, pp. 1083 – 1095, 2004. 1

[5] A. Pouget, T. Dyan, and R. Zemel, "Information processing with population codes," *Nature Reviews Neuroscience*, vol. 1, pp. 125–132, 2000. 1

[6] P. Dayan and L. Abbott, *Theoretical Neuroscience: Computational and Mathematical Modeling of Neural Systems*. The MIT Press, 2001. 1

[7] K. M. O'Craven and N. Kanwisher, "Mental imagery of faces and places activates corresponding stimulus-specific brain regions," *J. Cogn. Neurosci.*, vol. 12, no. 6, pp. 1013–1023, 2000. 1

[8] J. J. Sidtis, S. C. Strother, and D. A. Rottenberg, "Predicting performance from functional imaging data: methods matter." *Neuroimage*, vol. 20, no. 2, pp. 615–24, 2003. 1

[9] D. D. Cox and R. L. Savoy, "Functional magnetic resonance imaging (fmri) "brain reading": detecting and classifying distributed patterns of fmri activity in human visual cortex." *Neuroimage*, vol. 19, pp. 261–270, 2003. 1, 2, 10

[10] S. LaConte, S. Strother, V. Cherkassky, J. Anderson, and X. Hu, "Support vector machines for temporal classification of block design fmri data," *NeuroImage*, vol. 26, no. 2, pp. 317 – 329, 2005. 1, 2

[11] A. J. O'Toole, F. Jiang, H. Abdi, N. Pnard, J. P. Dunlop, and M. A. Parent, "Theoretical, statistical, and practical perspectives on pattern-based classification approaches to the analysis of functional neuroimaging data," *Journal of Cognitive Neuroscience*, vol. 19, no. 11, pp. 1735–1752, 2007. 1

[12] J.-D. Haynes and G. Rees, "Decoding mental states from brain activity in humans," *Nat. Rev. Neurosci.*, vol. 7, p. 523, 2006. 1, 2

[13] Y. Kamitani and F. Tong, "Decoding the visual and subjective contents of the human brain," *Nat. Neurosci.*, vol. 8, p. 679, 2005. 1

[14] K. N. Kay, T. Naselaris, R. J. Prenger, and J. L. Gallant, "Identifying natural images from human brain activity," *Nature*, vol. 452, pp. 352–355, 2008. 1

[15] T. M. Mitchell, S. V. Shinkareva, A. Carlson, K.-M. Chang, V. L. Malave, R. A. Mason, and M. A. Just, "Predicting Human Brain Activity Associated with the Meanings of Nouns," *Science*, vol. 320, no. 5880, pp. 1191–1195, 2008. 1

[16] A. Knops, B. Thirion, E. Hubbard, V. Michel, and S. Dehaene, "Mathematics as cortical recycling : Recruitment of an area involved in eye movements during mental arithmetic," *Science*, vol. 324, pp. 1583 – 1585, June 2009. 1

[17] J. V. Haxby, M. I. 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, no. 5539, pp. 2425–2430, 2001. 1

[18] H. Liu, M. Palatucci, and J. Zhang, "Blockwise coordinate descent procedures for the multi-task lasso, with applications to neural semantic basis discovery," in *ICML '09: Proceedings of the 26th Annual International Conference on Machine Learning*, 2009, pp. 649–656. 1

[19] M. K. Carroll, G. A. Cecchi, I. Rish, R. Garg, and A. R. Rao, "Prediction and interpretation of distributed neural activity with sparse models," *NeuroImage*, vol. 44, no. 1, pp. 112 – 122, 2009. 1, 2

[20] C. Chu, Y. Ni, G. Tan, C. J. Saunders, and J. Ashburner, "Kernel regression for fMRI pattern prediction," *NeuroImage*, 2010. 2

[21] M. Martnez-Ramn, V. Koltchinskii, G. L. Heileman, and S. Posse, "fmri pattern classification using neuroanatomically constrained boosting," *NeuroImage*, vol. 31, no. 3, pp. 1129 – 1141, 2006. 2

[22] S. Ryali, K. Supekar, D. A. Abrams, and V. Menon, "Sparse logistic regression for whole-brain classification of fmri data," *NeuroImage*, vol. 51, no. 2, pp. 752 – 764, 2010. 2

[23] R. D. Raizada, F.-M. Tsao, H.-M. Liu, I. D. Holloway, D. Ansari, and P. K. Kuhl, "Linking brain-wide multivoxel activation patterns to behaviour: Examples from language and math," *NeuroImage*, vol. 51, no. 1, pp. 462 – 471, 2010. 2

[24] O. Yamashita, M. aki Sato, T. Yoshioka, F. Tong, and Y. Kamitani, "Sparse estimation automatically selects voxels relevant for the decoding of fMRI activity patterns," *NeuroImage*, vol. 42, no. 4, pp. 1414 – 1429, 2008. 2

[25] K. Friston, C. Chu, J. Mourao-Miranda, O. Hulme, G. Rees, W. Penny, and J. Ashburner, "Bayesian decoding of brain images," *NeuroImage*, vol. 39, pp. 181–205, January 2008. 2

[26] V. Michel, E. Eger, C. Keribin, and B. Thirion, "Adaptive multi-class bayesian sparse regression - an application to brain activity classification," in *MICCAI'09 Workshop on Analysis of Functional Medical Images.*, Sept 2009. 2

[27] M. Palatucci and T. Mitchell, "Classification in very high dimensional problems with handfuls of examples," in *Principles and Practice of Knowledge Discovery in Databases (ECML/PKDD)*. Springer-Verlag, September 2007. 2

[28] V. Michel, E. Eger, C. Keribin, J.-B. Poline, and B. Thirion, "A supervised clustering approach for extracting predictive information from brain activation images." *MMBIA'10*, 2010. 2

[29] T. M. Mitchell, R. Hutchinson, R. S. Niculescu, F. Pereira, X. Wang, M. Just, and S. Newman, "Learning to decode cognitive states from brain images," *Machine Learning*, vol. 57, no. 1, pp. 145–175, 2004. 2

[30] J. Mouro-Miranda, A. L. Bokde, C. Born, H. Hampel, and M. Stetter, "Classifying brain states and determining the discriminating activation patterns: Support vector machine on functional mri data," *NeuroImage*, vol. 28, no. 4, pp. 980 – 995, 2005. 2

[31] Z. Wang, A. R. Childress, J. Wang, and J. A. Detre, "Support vector machine learning-based fmri data group analysis," *NeuroImage*, vol. 36, no. 4, pp. 1139 – 1151, 2007. 2

[32] S. J. Hanson and Y. O. Halchenko, "Brain reading using full brain support vector machines for object recognition: There is no face identification area," *Neural Computation*, vol. 20, no. 2, pp. 486–503, 2008. 2

[33] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *J. Roy. Stat. Soc. B*, vol. 67, p. 301, 2005. 2, 5

[34] R. Tibshirani, "Regression shrinkage and selection via the lasso," *J. Roy. Stat. Soc. B*, vol. 58, p. 267, 1996. 2

[35] A. E. Hoerl and R. W. Kennard, "Ridge regression: Biased estimation for nonorthogonal problems," *Technometrics*, vol. 12, pp. 55–67, 1970. 2

[36] B. Thirion, G. Flandin, P. Pinel, A. Roche, P. Ciuciu, and J.-B. Poline, "Dealing with the shortcomings of spatial normalization: Multi-subject parcellation of fMRI datasets," *Hum. Brain Mapp.*, vol. 27, no. 8, pp. 678–693, 2006. 2

[37] L. Rudin, S. Osher, and E. Fatemi, "Nonlinear total variation based noise removal algorithms," *Physica D*, Jan 1992. 2, 4

[38] A. Chambolle, "An algorithm for total variation minimization and applications," *J. Math. Imaging Vis.*, vol. 20, no. 1-2, pp. 89–97, 2004. 2, 3, 4

[39] D. B. Chklovskii and A. A. Koulakov, "Maps in the brain: What can we learn from them?" *Annual Review of Neuroscience*, vol. 27, no. 1, pp. 369–392, 2004. 2

[40] J. Stoeckel and G. Fung, "SVM feature selection for classification of spect images of alzheimer?s disease using spatial information," *Data Mining, IEEE International Conference on*, vol. 0, pp. 410–417, 2005. 2

[41] D. Goldfarb and W. Yin, "Parametric maximum flow algorithms for fast total variation minimization," *SIAM J. Sci. Comput.*, vol. 31, pp. 3712–3743, 2009. 3, 4

[42] I. Daubechies, M. Defrise, and C. De Mol, "An iterative thresholding algorithm for linear inverse problems with a sparsity constraint," *Commun. Pure Appl. Math.*, vol. 57, no. 11, pp. 1413 – 1457, 2004. 3, 12

[43] P. L. Combettes and V. R. Wajs, "Signal recovery by proximal forward-backward splitting," *Multiscale Modeling and Simulation*, vol. 4, no. 4, pp. 1168–1200, 2005. 3

[44] Y. Nesterov, "Gradient methods for minimizing composite objective function," Université catholique de Louvain, Center for Operations Research and Econometrics (CORE), CORE Discussion Papers, Sep 2007. [Online]. Available: http://ideas.repec.org/p/cor/louvco/2007076.html 3

[45] 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. 3

[46] J. Moreau, "Proximit et dualit dans un espace hilbertien," *Bull. Soc. Math. France.*, vol. 93, pp. 273–299, 1965. 3

[47] P. Tseng, "Approximation accuracy, gradient methods, and error bound for structured convex optimization," *submitted to Math Program. B*, 2009. 3

[48] A. Gramfort, "Mapping, timing and tracking cortical activations with MEG and EEG: Methods and application to human vision," Ph.D. dissertation, Telecom ParisTech, 2009. 3

[49] A. Gramfort and M. Kowalski, "Improving m/eeg source localization with an inter-condition sparse prior," in *IEEE International Symposium on Biomedical Imaging*, 2009. 3

[50] A. Beck and M. Teboulle, "Fast gradient-based algorithms for constrained total variation image denoising and deblurring problems," *Trans. Img. Proc.*, vol. 18, no. 11, pp. 2419–2434, 2009. 3, 4

[51] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, 2004. 4, 12

[52] T. Hastie, R. Tibshirani, and J. H. Friedman, *The Elements of Statistical Learning*. Springer, 2003. 4

[53] B. Krishnapuram, L. Carin, M. A. Figueiredo, and A. J. Hartemink, "Sparse multinomial logistic regression: Fast algorithms and generalization bounds," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, pp. 957–968, 2005. 5

[54] C.-C. Chang and C.-J. Lin, *LIBSVM: a library for support vector machines*, 2001, software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm. 5, 6

[55] C. Cortes and V. Vapnik, "Support vector networks," in *Machine Learning*, vol. 20, 1995, p. 273. 5

[56] G. Hughes, "On the mean accuracy of statistical pattern recognizers," *Information Theory, IEEE Transactions on*, vol. 14, no. 1, pp. 55–63, 1968. 5

[57] J. Friedman, T. Hastie, and R. Tibshirani, "Regularization paths for generalized linear models via coordinate descent," *Journal of Statistical Software*, vol. 33, no. 1, 2010. 6

[58] "scikit-learn," http://scikit-learn.sourceforge.net/, downloaded in Apr. 2010, version 0.2. 6

[59] E. Eger, C. Kell, and A. Kleinschmidt, "Graded size sensitivity of object exemplar evoked activity patterns in human loc subregions," *J. Neurophysiol.*, vol. 100(4):2038-47, 2008. 6, 7, 8, 10

[60] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene selection for cancer classification using support vector machines," *Machine Learning*, vol. 46, no. 1-3, pp. 389–422, 2002. 10

[61] N. Kriegeskorte, R. Goebel, and P. Bandettini, "Information-based functional brain mapping." *Proceedings of the National Academy of*

Sciences of the United States of America, vol. 103, no. 10, pp. 3863–3868, March 2006. 10

[62] J. M. Ortega and W. C. Rheinboldt, *Iterative solution of nonlinear equations in several variables*. Society for Industrial and Applied Mathematics, 2000. 12

APPENDIX A
GRADIENT AND DIVERGENCE

The gradient operator which has to be computed on a mask in our case (mask of the brain). With $I \in \mathbb{R}^{p_i \times p_j \times p_k}$ an image, it is defined by:

$$(\text{grad } I)_x^{i,j,k} = \begin{cases} I_{i+1,j,k} - I_{i,j,k} & \text{if } M_{i,j,k} = M_{i+1,j,k} = 1 \\ 0 & \text{otherwise} \end{cases}$$

$$(\text{grad } I)_y^{i,j,k} = \begin{cases} I_{i,j+1,k} - I_{i,j,k} & \text{if } M_{i,j,k} = M_{i,j+1,k} = 1 \\ 0 & \text{otherwise} \end{cases}$$

$$(\text{grad } I)_z^{i,j,k} = \begin{cases} I_{i,j,k+1} - I_{i,j,k} & \text{if } M_{i,j,k} = M_{i,j,k+1} = 1 \\ 0 & \text{otherwise} \end{cases}$$

The divergence operator for a gradient p is defined by:

$$\begin{aligned}(\text{div } p)^{i,j,k} = &\begin{cases} p^x_{i,j,k} - p^x_{i-1,j,k} & \text{if } M_{i,j,k} = M_{i-1,j,k} = 1 \\ p^x_{i,j,k} & \text{if } M_{i,j,k} \neq M_{i-1,j,k} = 0 \\ -p^x_{i-1,j,k} & \text{if } M_{i,j,k} \neq M_{i-1,j,k} = 1 \end{cases} \\ + &\begin{cases} p^y_{i,j,k} - p^y_{i,j-1,k} & \text{if } M_{i,j,k} = M_{i,j-1,k} = 1 \\ p^y_{i,j,k} & \text{if } M_{i,j,k} \neq M_{i,j-1,k} = 0 \\ -p^y_{i,j-1,k} & \text{if } M_{i,j,k} \neq M_{i,j-1,k} = 1 \end{cases} \\ + &\begin{cases} p^z_{i,j,k} - p^z_{i,j,k-1} & \text{if } M_{i,j,k} = M_{i,j,k-1} = 1 \\ p^z_{i,j,k} & \text{if } M_{i,j,k} \neq M_{i,j,k-1} = 0 \\ -p^z_{i,j,k-1} & \text{if } M_{i,j,k} \neq M_{i,j,k-1} = 1 \end{cases}\end{aligned}$$

APPENDIX B
ISTA PROCEDURE

We give the sketch of proof of (6). The loss $\mathcal{L}(\mathbf{w})$ being differentiable, the second-order linearization of $\mathcal{L}(\mathbf{w})$ reads:

$$\begin{aligned}\mathcal{L}(\mathbf{w}) &\approx \mathcal{L}(\mathbf{w}^{(k)}) + (\mathbf{w} - \mathbf{w}^{(k)})^T \nabla \mathcal{L}(\mathbf{w}^{(k)}) \\ &+ \frac{1}{2}(\mathbf{w} - \mathbf{w}^{(k)})^T \nabla^2 \mathcal{L}(\mathbf{w}^{(k)})(\mathbf{w} - \mathbf{w}^{(k)})\end{aligned}$$

With L_0 the Lipschitz constant of $\nabla\mathcal{L}$, we have:

$$\|\nabla \mathcal{L}(\mathbf{w}) - \nabla \mathcal{L}(\mathbf{w}^{(k)})\|^2 \leq L_0 \|\mathbf{w} - \mathbf{w}^{(k)}\|^2$$

Using [62], we obtain:

$$\begin{aligned}\mathbf{w}^{(k+1)} = \underset{\mathbf{w}}{\text{argmin}}\ &\mathcal{L}(\mathbf{w}^{(k)}) + \frac{L}{2}\|\mathbf{w} - \mathbf{w}^{(k)}\|^2 + \lambda J(\mathbf{w}) \\ &+ (\mathbf{w} - \mathbf{w}^{(k)})^T \nabla \mathcal{L}(\mathbf{w}^{(k)})\end{aligned}$$

Ignoring constant terms, this can be rewritten as:

$$\mathbf{w}^{(k+1)} = \underset{\mathbf{w}}{\text{argmin}}\ \frac{1}{2}\|\mathbf{w} - (\mathbf{w}^{(k)} - \frac{1}{L}\nabla \mathcal{L}(\mathbf{w}^{(k)})\|^2 + \frac{1}{L}\lambda J(\mathbf{w}),$$

where $L \geq L_0$ [42]. Finally, using definition (1) of the proximity operator for $J(\mathbf{w})$, this is equivalent to (6):

$$\mathbf{w}^{(k+1)} = \text{prox}_{\lambda J/L}\left(\mathbf{w}^{(k)} - \frac{1}{L}\nabla \mathcal{L}(\mathbf{w}^{(k)})\right)$$

APPENDIX C
DUAL PROBLEM AND DUALITY GAP COMPUTATION

We give the sketch of proofs of propositions 2 and 3. We recall [51] that the duality between the ℓ_1 norm and the ℓ_∞ norm yields:

$$TV(\mathbf{v}) = \|\nabla \mathbf{v}\|_1 = \max_{\|\mathbf{z}\|_\infty \leq 1} \langle \nabla \mathbf{v}, \mathbf{z} \rangle \tag{12}$$

and that the adjoint relation between the gradient and the divergence operator reads:

$$\langle \nabla \mathbf{v}, \mathbf{z} \rangle = -\langle \mathbf{v}, \text{div}\, \mathbf{z} \rangle \tag{13}$$

Using (12) and (13), we minimize:

$$\begin{aligned}&\min_{\mathbf{v}} \left(\frac{1}{2}\|\mathbf{w} - \mathbf{v}\|_2^2 + \lambda TV(\mathbf{v})\right) \\ = \ &\lambda \min_{\mathbf{v}} \left(\frac{1}{2\lambda}\|\mathbf{w} - \mathbf{v}\|_2^2 + \max_{\|\mathbf{z}\|_\infty \leq 1} \langle \nabla \mathbf{v}, \mathbf{z} \rangle\right) \\ = \ &\lambda \max_{\|\mathbf{z}\|_\infty \leq 1} \left(\min_{\mathbf{v}} \left(\frac{1}{2\lambda}\|\mathbf{w} - \mathbf{v}\|_2^2 + \langle \nabla \mathbf{v}, \mathbf{z} \rangle\right)\right) \\ = \ &\lambda \max_{\|\mathbf{z}\|_\infty \leq 1} \left(\min_{\mathbf{v}} \left(\frac{1}{2\lambda}\|\mathbf{w} - \mathbf{v}\|_2^2 - \langle \mathbf{v}, \text{div}\, \mathbf{z} \rangle\right)\right)\end{aligned}$$

The computation of the minimum and the maximum above can be exchanged because the optimization over $\mathbf{v}$ is convex and the optimization over $\mathbf{z}$ is concave [51].
By setting the derivative with respect to $\mathbf{v}$ to 0 one gets the resulting solution of the minimization problem over $\mathbf{v}$:

$$\min_{\mathbf{v}} \left(\frac{1}{2\lambda}\|\mathbf{w} - \mathbf{v}\|_2^2 - \langle \mathbf{v}, \text{div}\, \mathbf{z} \rangle\right) \ \Rightarrow \ \mathbf{v}^* = \mathbf{w} + \lambda \text{div}\, \mathbf{z}$$

Replacing $\mathbf{v}$ by $\mathbf{v}^*$ in the previous expression leads to:

$$\begin{aligned}&\min_{\mathbf{v}} \left(\frac{1}{2}\|\mathbf{w} - \mathbf{v}\|_2^2 + \lambda TV(\mathbf{v})\right) \\ = \ &\lambda \max_{\|\mathbf{z}\|_\infty \leq 1} \left(\frac{\lambda}{2}\|\text{div}\, \mathbf{z}\|_2^2 - \langle \mathbf{w}, \text{div}\, \mathbf{z} \rangle - \lambda \|\text{div}\, \mathbf{z}\|_2^2\right) \\ = \ &\lambda \max_{\|\mathbf{z}\|_\infty \leq 1} \left(-\frac{\lambda}{2}\|\text{div}\, \mathbf{z}\|_2^2 - \langle \mathbf{w}, \text{div}\, \mathbf{z} \rangle\right) \\ = \ &\frac{1}{2} \max_{\|\mathbf{z}\|_\infty \leq 1} \left(-\lambda^2 \|\text{div}\, \mathbf{z}\|_2^2 - 2\lambda \langle \mathbf{w}, \text{div}\, \mathbf{z} \rangle\right) \\ = \ &\frac{1}{2} \max_{\|\mathbf{z}\|_\infty \leq 1} \left(\|\mathbf{w}\|_2^2 - \|\lambda \text{div}\, \mathbf{z} + \mathbf{w}\|_2^2\right)\end{aligned}$$

This gives the proof of Prop. 2. Also, given a variable $\mathbf{z}$ satisfying $\|\mathbf{z}\|_\infty \leq 1$ and an associated $\mathbf{w}$ such that $\mathbf{v} = \mathbf{w} + \lambda \text{div}\, \mathbf{z}$, one can guarantee that

$$\frac{1}{2}\|\mathbf{w} - \mathbf{v}\|_2^2 + \lambda TV(\mathbf{v}) \geq \frac{1}{2}(\|\mathbf{w}\|_2^2 - \|\mathbf{v}\|_2^2)$$

The strict convexity of the problem guarantees that, at the optimum, the equality holds. This last derivation proves the proposition 3.