SIAM J. IMAGING SCIENCES
Vol. 5, No. 3, pp. 835–856

Multiscale Mining of fMRI Data with Hierarchical Structured Sparsity*

Rodolphe Jenatton†, Alexandre Gramfort‡, Vincent Michel‡, Guillaume Obozinski†, Evelyn Eger§, Francis Bach†, and Bertrand Thirion‡

Abstract. Reverse inference, or *brain reading*, is a recent paradigm for analyzing functional magnetic resonance imaging (fMRI) data based on pattern recognition and statistical learning. By predicting some cognitive variables related to brain activation maps, this approach aims at decoding brain activity. Reverse inference takes into account the multivariate information between voxels and is currently the only way to assess how precisely some cognitive information is encoded by the activity of neural populations within the whole brain. However, it relies on a prediction function that is plagued by the curse of dimensionality, since there are far more features than samples, i.e., more voxels than fMRI volumes. To address this problem, different methods have been proposed, including univariate feature selection, feature agglomeration, and regularization techniques. In this paper, we consider a sparse hierarchical structured regularization. Specifically, the penalization we use is constructed from a tree that is obtained by spatially constrained agglomerative clustering. This approach encodes the spatial structure of the data at different scales into the regularization, which makes the overall prediction procedure more robust to intersubject variability. The regularization used induces the selection of spatially coherent predictive brain regions simultaneously at different scales. We test our algorithm on real data acquired to study the mental representation of objects, and we show that the proposed algorithm not only delineates meaningful brain regions but also yields better prediction accuracy than reference methods.

Key words. brain reading, structured sparsity, convex optimization, sparse hierarchical models, intersubject validation, proximal methods

AMS subject classifications. 90C25, 65F22, 62P10, 62H99, 62J07

DOI. 10.1137/110832380

1. Introduction.

Functional magnetic resonance imaging (fMRI) is a widely used functional neuroimaging modality. Modeling and statistical analysis of fMRI data are commonly done through a linear model, called a general linear model (GLM) in the community, that incorporates information about the different experimental conditions and the dynamics of the hemodynamic response in the design matrix. The experimental paradigm consists of a sequence of stimuli, e.g., visual and auditory stimuli, which are included as regressors in the

*Received by the editors April 28, 2011; accepted for publication (in revised form) March 6, 2012; published electronically July 10, 2012. This research was supported by ANR grants ViMAGINE ANR-08-BLAN-0250-02 and ANR 2010-Blan-0126-01 "IRMGroup." The project was also partially supported by a grant from the European Research Council (SIERRA Project). A preliminary version of this work appeared in [27].
http://www.siam.org/journals/siims/5-3/83238.html
†INRIA Rocquencourt, Sierra Project-Team, Laboratoire d'Informatique de l'Ecole Normale Supérieure (INRIA/ENS/CNRS UMR 8548), 23, Avenue d'Italie CS 81321 75214 Paris, France (rodolphe.jenatton@inria.fr, guillaume.obozinski@inria.fr, francis.bach@inria.fr).
‡INRIA Saclay, Parietal Project-Team, CEA, Neurospin, Bâtiment 145, point courier 156, 91191 Gif sur Yvette, France (alexandre.gramfort@inria.fr, vincent.michel@inria.fr, bertrand.thirion@inria.fr).
§INSERM U562, CEA/ DSV/ I2BM/ Neurospin/ Unicog, Bâtiment 145, point courier 156, 91191 Gif sur Yvette, France (evelyn.eger@cea.fr).

design matrix after convolution with a suitable hemodynamic filter. The resulting model parameters—one coefficient per voxel and regressor—are known as *activation maps*. They represent the local influence of the different experimental conditions on fMRI signals at the level of individual voxels. The most commonly used approach to analyzing these activation maps is called classical inference. It relies on mass-univariate statistical tests (one for each voxel) and yields so-called statistical parametric maps (SPMs) [19]. Such maps are useful for functional brain mapping, but classical inference has some limitations: it suffers from multiple comparison issues, and it is oblivious to the multivariate structure of fMRI data. Such data exhibit natural correlations between neighboring voxels forming clusters with different sizes and shapes, and also between distant but functionally connected brain regions.

To address these limitations, an approach called reverse inference (or "brain reading") [14, 13] was recently proposed. Reverse inference relies on pattern recognition tools and statistical learning methods to explore fMRI data. Based on a set of activation maps, reverse inference estimates a function that can then be used to predict a target (typically, a variable representing a perceptual, cognitive, or behavioral parameter) for a new set of images. The challenge is to capture the correlation structure present in the data in order to improve the accuracy of the fit, which is measured through the resulting prediction accuracy. Many standard statistical learning approaches have been used to construct prediction functions, among them kernel machines (SVM, RVM) [54] or discriminant analysis (LDA, QDA) [22]. For the application considered in this paper, earlier performance results [13, 32] indicate that we can restrict ourselves to mappings that are linear functions of the data.

Throughout the paper, we shall consider a training set composed of n pairs $(\mathbf{x}, y) \in \mathbb{R}^p \times \mathcal{Y}$, where $\mathbf{x}$ denotes a p-dimensional fMRI signal (p voxels) and y stands for the target that we try to predict. Each fMRI data point $\mathbf{x}$ will correspond to an activation map after GLM fitting. In the experiments we carry out in section 5, we will encounter both the regression and the multiclass classification settings, where $\mathcal{Y}$ denotes, respectively, the set of real numbers and a finite set of integers. An example of a regression setting is the prediction of a pain level from fMRI data [37] or, in the context of classification, the prediction of object categories [13]. Typical datasets consist of a few hundred measurements, each defined on a $2 \times 2 \times 2$-mm voxel grid forming $p \approx 10^5$ voxels when working with full brain data. Such numbers, given as illustration, are not an intrinsic limitation of MRI technology and are still regularly improved by experts in the field.

In this paper, we aim at learning a weight vector $\mathbf{w} \in \mathbb{R}^p$ and an intercept $b \in \mathbb{R}$ such that the prediction of y can be based on the value of $\mathbf{w}^\top \mathbf{x} + b$. This is the case for the linear regression and logistic regression models that we use in section 5. The scalar b is not particularly informative; however, the vector $\mathbf{w}$ corresponds to a volume that can be represented in brain space as a volume for visualization of the predictive pattern of voxels. It is useful to rewrite these quantities in matrix form; more precisely, we denote by $\mathbf{X} \in \mathbb{R}^{n \times p}$ the design matrix assembled from n fMRI volumes and by $\mathbf{y} \in \mathbb{R}^n$ the corresponding n targets. In other words, each row of $\mathbf{X}$ is a p-dimensional sample, i.e., an activation map of p voxels related to one stimulus presentation.

Learning the parameters $(\mathbf{w}, b)$ remains challenging since the number of features (10^4 to 10^5 voxels) far exceeds the number of samples (a few hundred volumes). The prediction function is therefore prone to overfitting, that is, the learning set is predicted precisely, while

the algorithm provides inaccurate predictions on new samples (the test set). To address this issue, *dimensionality reduction* attempts to find a low dimensional subspace that concentrates as much of the predictive power of the original set as possible on the problem at hand.

Feature selection is a natural approach to performing dimensionality reduction in fMRI, since reducing the number of voxels makes it easier to identify a predictive region of the brain. This corresponds to discarding some columns of $\mathbf{X}$. This feature selection can be univariate, e.g., analysis of variance (ANOVA) [33], or multivariate. While univariate methods ignore joint information between features, multivariate approaches are more adapted to reverse inference since they extract predictive patterns from the data as a whole. However, due to the huge number of possible patterns, these approaches suffer from combinatorial explosion, and some costly suboptimal heuristics (e.g., recursive feature elimination [21, 39]) can be used. This is why ANOVA is usually preferred in fMRI. Alternatively, two more adapted solutions have been proposed—*regularization* and *feature agglomeration*.

Regularization is a way to encode a priori knowledge about the weight vector $\mathbf{w}$. Possible regularizers can promote, for example, spatial smoothness or sparsity, which is a natural assumption for fMRI data. Indeed, only a few brain regions are assumed to be significantly activated during a cognitive task. Previous contributions on fMRI-based reverse inference include [6, 51, 52, 64]. They can be presented through the following minimization problem:

$$\min_{(\mathbf{w},b)\in\mathbb{R}^{p+1}} \mathcal{L}(\mathbf{y},\mathbf{X},\mathbf{w},b) + \lambda\Omega(\mathbf{w}) \quad \text{with} \quad \lambda \geq 0, \tag{1.1}$$

where $\lambda\Omega(\mathbf{w})$ is the regularization term, typically a non-Euclidean norm, and the fit to the data is measured through a convex loss function $(\mathbf{w},b) \mapsto \mathcal{L}(\mathbf{y},\mathbf{X},\mathbf{w},b) \in \mathbb{R}_+$. The choice of the loss function will be made more specific and formal in the following sections. The coefficient of regularization λ balances the loss and the penalization term. In this notation, a common regularization term in reverse inference is the so-called *Elastic net* [67, 20], which is a combined ℓ_1 and ℓ_2 penalization:

$$\lambda\Omega(\mathbf{w}) = \lambda_1\|\mathbf{w}\|_1 + \lambda_2\|\mathbf{w}\|_2^2 = \sum_{j=1}^{p} \left\{\lambda_1|\mathbf{w}_j| + \lambda_2\mathbf{w}_j^2\right\}. \tag{1.2}$$

For the squared loss, when setting λ_1 to 0, the model is called ridge regression, while when $\lambda_2 = 0$, it is known as Lasso [58] or basis pursuit denoising (BPDN) [9]. The essential shortcoming of the Elastic net is that it does not take into account the spatial structure of the data, which is crucial in this context [43]. Indeed, due to the intrinsic smoothing of the complex metabolic pathway underlying the difference of blood oxygenation measured with fMRI [61], statistical learning approaches should be informed by the three-dimensional (3D) grid structure of the data.

In order to achieve dimensionality reduction, while taking into account the spatial structure of the data, one can resort to *feature agglomeration*. Precisely, new features called *parcels* are naturally generated via the averaging of groups of neighboring voxels exhibiting similar activations. The advantage of agglomeration is that no information is discarded a priori and that it is reasonable to hope that averaging might reduce noise. Although this approach has been successfully used in previous work for brain mapping [18, 57], existing work does not

typically consider the supervised information (i.e., the target y) while exploring the parcels. A recent approach has been proposed to address this issue, based on a supervised greedy top-down exploration of a tree obtained by hierarchical clustering [42]. This greedy approach has proven to be effective especially for intersubject analyses, i.e., when training and evaluation sets are related to different subjects. In this context, methods need to be robust to intrinsic spatial variations that exist across subjects: although a preliminary coregistration to a common space has been performed, some variability remains between subjects, which implies that there is no perfect voxel-to-voxel correspondence between volumes. As a result, the performances of traditional voxel-based methods are strongly affected. Therefore, averaging in the form of parcels is a good way to cope with intersubject variability. This greedy approach is nonetheless suboptimal, as it explores only a subpart of the whole tree.

Based on these considerations, we propose integrating the multiscale spatial structure of the data *within* the regularization term Ω while preserving convexity in the optimization. This notably guarantees global optimality and stability of the obtained solutions. To this end, we design a sparsity-inducing penalty that is directly built from the hierarchical structure of the spatial model obtained by Ward's algorithm [62] using a contiguity constraint [45]. This kind of penalty has already been successfully applied in several contexts, e.g., in bioinformatics to exploit the tree structure of gene networks for multitask regression [30], in log-linear models for the selection of potential orders [53], in image processing for wavelet-based denoising [3, 28, 49], and also in topic models [28]. Other applications have emerged in natural language [40] and audio processing [55].

We summarize here the contributions of our paper:

- We explain how the multiscale spatial structure of fMRI data can be taken into account in the context of reverse inference through the combination of a spatially constrained hierarchical clustering procedure and a sparse hierarchical regularization.
- We provide a convex formulation of the problem and propose an efficient optimization procedure.
- We conduct an experimental comparison of several algorithms and formulations on fMRI data and illustrate the ability of the proposed method to localize in space and in scale some brain regions involved in the processing of visual stimuli.

The rest of the paper is organized as follows: we first present the concept of structured sparsity-inducing regularization and then describe the different regression/classification formulations we are interested in. After exposing how we handle the resulting large-scale convex optimization problems thanks to a particular instance of proximal methods—the forward-backward splitting algorithm—we validate our approach both in a synthetic setting and on a real dataset.

2. Combining agglomerative clustering with sparsity-inducing regularizers. As suggested in the introduction, it is possible to construct a tree-structured hierarchy of new features on top of the original voxels using hierarchical clustering. Moreover, spatial constraints can be enforced in the clustering algorithm so that the underlying voxels corresponding to each of these features form localized spatial patterns on the brain similar to those we hope to retrieve [10]. Once these features are constructed, instead of selecting features in the tree greedily, we propose casting the feature selection problem as a supervised learning problem of

the form (1.1). One of the qualities of the greedy approach, however, is that it is allowed to select potentially more noisy features, corresponding to smaller clusters, only after the a priori more stable features associated with ancestral clusters in the hierarchy have been selected. As we will show, it is possible to construct a convex regularizer Ω that has the same property, i.e., that respects the hierarchy and prioritizes the selection of features in the same way. Naturally, the regularizer has to be constructed directly from the hierarchical clustering of the voxels.

2.1. Spatially constrained hierarchical clustering. Starting from n fMRI volumes $\mathbf{X} = [\mathbf{x}^1, \ldots, \mathbf{x}^p] \in \mathbb{R}^{n \times p}$ described by p voxels, we seek to cluster these voxels so as to produce a hierarchical representation of $\mathbf{X}$.

To this end, we consider *hierarchical agglomerative clustering* procedures [29]. These begin with every voxel $\mathbf{x}^j$ representing a singleton cluster $\{j\}$, and, at each iteration, a selected pair of clusters—according to a criterion discussed below—is merged into a single cluster. This procedure yields a hierarchy of clusters represented as a binary tree $\mathcal{T}$ (also often called a dendrogram) [29], where each nonterminal node is associated with the cluster obtained by merging its two children clusters. Moreover, the root of the tree $\mathcal{T}$ is the unique cluster that gathers all the voxels, while the leaves are the clusters consisting of a single voxel. From now on, we refer to each nonterminal node of $\mathcal{T}$ as a *parcel*, which is the union of its children's voxels (see Figure 1).

Among different hierarchical agglomerative clustering procedures, we use the variance-minimizing approach of Ward's algorithm [62]. In short, two clusters are merged if the resulting cluster minimizes the sum of squared differences of the fMRI signal within all clusters (also known as the *inertia criterion*). More formally, at each step of the procedure, we merge the clusters c_1 and c_2 that minimize the criterion

$$\begin{aligned}\Delta(c_1, c_2) &= \sum_{j \in c_1 \cup c_2} \|\mathbf{x}^j - \langle \mathbf{X} \rangle_{c_1 \cup c_2}\|_2^2 - \Big(\sum_{j \in c_1} \|\mathbf{x}^j - \langle \mathbf{X} \rangle_{c_1}\|_2^2 + \sum_{k \in c_2} \|\mathbf{x}^k - \langle \mathbf{X} \rangle_{c_2}\|_2^2 \Big) \\ &= \frac{|c_1||c_2|}{|c_1| + |c_2|} \|\langle \mathbf{X} \rangle_{c_1} - \langle \mathbf{X} \rangle_{c_2}\|_2^2, \end{aligned} \tag{2.1}$$

where we have introduced the average vector $\langle \mathbf{X} \rangle_c \triangleq \frac{1}{|c|} \sum_{j \in c} \mathbf{x}^j$. In order to take into account the spatial information, we also add connectivity constraints in the hierarchical clustering algorithm, so that only neighboring clusters can be merged together. In other words, we try to minimize the criterion $\Delta(c_1, c_2)$ only for pairs of clusters which share neighboring voxels (see Algorithm 1). This connectedness constraint is important since the resulting clustering is likely to differ from standard Ward hierarchical clustering.

2.1.1. Augmented space of features. Based on the output of the hierarchical clustering presented previously, we define the following augmented space of variables (or *features*): instead of representing the n fMRI volumes only by its individual voxel intensities, we add to it a vector with levels of activation of each of the parcels at the interior nodes of the tree $\mathcal{T}$, which we obtained from the agglomerative clustering algorithm. Since $\mathcal{T}$ has $q \triangleq |\mathcal{T}| = 2p - 1$ nodes,[1] the data obtained in the augmented space can be gathered in a matrix $\tilde{\mathbf{X}} \in \mathbb{R}^{n \times q}$.

[1] We can then identify nodes (and parcels) of $\mathcal{T}$ with indices in $\{1, \ldots, q\}$.

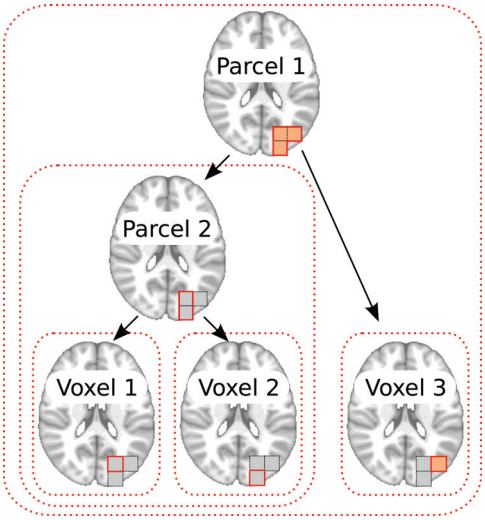

Figure 1. *Example of a tree $\mathcal{T}$ when $p = 5$ with three voxels and two parcels. Parcel 2 is defined as the averaged intensity of the voxels $\{1, 2\}$, while Parcel 1 is obtained by averaging the values of the voxels $\{1, 2, 3\}$. Red dashed lines represent the five groups of variables that compose $\mathcal{G}$. For instance, if the group containing Parcel 2 is set to zero, the voxels $\{1, 2\}$ are also (and necessarily) zeroed out.*

Algorithm 1. Spatially constrained agglomerative clustering and augmented feature space.

Input: n fMRI volumes $\mathbf{X} = [\mathbf{x}^1, \ldots, \mathbf{x}^p] \in \mathbb{R}^{n \times p}$ described by p voxels.
Output: n fMRI volumes $\tilde{\mathbf{X}} \in \mathbb{R}^{n \times q}$ in the augmented feature space.
Initialization: $\mathcal{C} = \big\{\{j\};\ j \in \{1, \ldots, p\}\big\}$, $\tilde{\mathbf{X}} = \mathbf{X}$.
while $|\mathcal{C}| > 1$ **do**
  Find a pair of clusters $c_1, c_2 \in \mathcal{C}$ which have neighboring voxels and minimize (2.1).
  $\mathcal{C} \leftarrow \mathcal{C} \setminus \{c_1, c_2\}$.
  $\mathcal{C} \leftarrow \mathcal{C} \cup (c_1 \cup c_2)$.
  $\tilde{\mathbf{X}} \leftarrow [\tilde{\mathbf{X}}, \langle \mathbf{X} \rangle_{c_1 \cup c_2}]$.
end while
Return: $\tilde{\mathbf{X}}$.

In the following, the level of activation of each parcel is simply the averaged intensity of the voxels it is composed of (i.e., local averages) [18, 57]. This produces a multiscale representation of the fMRI data that has the advantage of becoming increasingly invariant to spatial shifts of the encoding regions within the brain volume. We summarize the procedure to build the enlarged matrix $\tilde{\mathbf{X}} \in \mathbb{R}^{n \times q}$ in Algorithm 1. Let us now illustrate through the example of linear models (such as those we use in section 3) the implications of considering the augmented space of features. For a node j of $\mathcal{T}$, we denote by $P_j \subseteq \{1, \ldots, p\}$ the set of voxels of the corresponding parcel (or, equivalently, the set of leaves of the subtree rooted at node j). In this notation, and for any fMRI volume $\tilde{\mathbf{x}} \in \mathbb{R}^q$ in the augmented feature space,

linear functions indexed by $\mathbf{w} \in \mathbb{R}^q$ take the form

$$f_{\mathbf{w}}(\tilde{\mathbf{x}}) = \mathbf{w}^\top \tilde{\mathbf{x}} = \sum_{j=1}^{q} \mathbf{w}_j \Big[\frac{1}{|P_j|} \sum_{k \in P_j} \mathbf{x}_k \Big] = \sum_{k=1}^{p} \Big[\sum_{j \in A_k} \frac{\mathbf{w}_j}{|P_j|} \Big] \mathbf{x}_k,$$

where A_k stands for the set of ancestors of a node k in $\mathcal{T}$ (including itself).

2.2. Hierarchical sparsity-inducing norms. From the perspective of intersubject validation, the augmented space of variables can be exploited in the following way: Since the information of single voxels may be unreliable, *the deeper the node in $\mathcal{T}$, the more variable the corresponding parcel's intensity is likely to be across subjects.* This property suggests that, while looking for sparse solutions of (1.1), we should preferentially select the variables near the root of $\mathcal{T}$ before trying to access smaller parcels located further down in $\mathcal{T}$.

Traditional sparsity-inducing penalties, e.g., the ℓ_1-norm $\Omega(\mathbf{w}) = \sum_{j=1}^{p} |\mathbf{w}_j|$, yield sparsity at the level of single variables $\mathbf{w}_j$, disregarding potential structures—for instance, spatial—existing between larger subsets of variables. We leverage here the concept of *structured sparsity* [3, 7, 66, 24, 26, 25, 41, 16], where Ω penalizes some predefined subsets, or *groups*, of variables that reflect prior information about the problem at hand.

When these groups form a *partition* of the space of variables, the resulting penalty has been shown to improve the prediction performance and/or interpretability of the learned models, provided that the block structure is relevant (see, e.g., [60, 65, 35, 31, 56, 23] and references therein).

If the groups overlap [3, 66, 24, 26, 25, 36], richer structures can then be encoded. In particular, we follow [66], which first introduced hierarchical sparsity-inducing penalties. Given a node j of $\mathcal{T}$, we denote by $g_j \subseteq \{1, \ldots, q\}$ the set of indices that record all the descendants of j in $\mathcal{T}$, including itself. In other words, g_j contains the indices of the subtree rooted at j; see Figure 1. If we now denote by $\mathcal{G}$ the set of all g_j, $j \in \{1, \ldots, q\}$, that is, $\mathcal{G} \triangleq \{g_1, \ldots, g_q\}$, we can define our hierarchical penalty as

$$\Omega(\mathbf{w}) \triangleq \sum_{g \in \mathcal{G}} \|\mathbf{w}_g\|_2 \triangleq \sum_{g \in \mathcal{G}} \Big[\sum_{j \in g} \mathbf{w}_j^2 \Big]^{1/2}. \tag{2.2}$$

As formally shown in [26], Ω is a norm on $\mathbb{R}^q$, and it promotes sparsity at the level of groups $g \in \mathcal{G}$ in the sense that it acts as an ℓ_1-norm on the vector $(\|\mathbf{w}_g\|_2)_{g \in \mathcal{G}}$. Regularizing by Ω therefore causes some $\|\mathbf{w}_g\|_2$ (and equivalently $\mathbf{w}_g$) to be zeroed out for some $g \in \mathcal{G}$. Moreover, since the groups $g \in \mathcal{G}$ represent rooted subtrees of $\mathcal{T}$, this implies that if one node/parcel $j \in g$ is set to zero by Ω, the same occurs for all its descendants [66]. To put it differently, *if one parcel is selected, then all the ancestral parcels in $\mathcal{T}$ will also be selected.* This property is likely to increase the robustness of the methods to voxel misalignments between subjects, since large parcels will be considered for inclusion in the model before smaller ones.

The family of norms with the previous property is actually slightly larger, and we consider throughout the paper norms Ω of the following form [66]:

$$\Omega(\mathbf{w}) \triangleq \sum_{g \in \mathcal{G}} \eta_g \|\mathbf{w}_g\|, \tag{2.3}$$

where $\|\mathbf{w}_g\|$ denotes either the ℓ_2-norm $\|\mathbf{w}_g\|_2$ or the ℓ_∞-norm $\|\mathbf{w}_g\|_\infty \triangleq \max_{j\in g} |\mathbf{w}_j|$ and $(\eta_g)_{g\in\mathcal{G}}$ are (strictly) positive weights that can compensate for the fact that some features are overpenalized as a result of being included in a larger number of groups than others. In light of the results from [28], we will see in section 4 that a large class of optimization problems regularized by Ω—as defined in (2.3)—can be solved efficiently.

3. Supervised learning framework. In this section, we introduce the formulations which we consider in our experiments. As further discussed in section 5, the target y that we try to predict corresponds to (discrete) sizes of objects, i.e., a one-dimensional *ordered* variable. It is therefore sensible to address this prediction task from both a regression and a classification viewpoint. In the remainder of this section, we shall denote by $\{\mathbf{w}^*, b^*\}$ (or $\{\mathbf{W}^*, \mathbf{b}^*\}$) *a* solution of the optimization problems which we present below.[2] For simplicity, the formulations we review next are all expressed in terms of a matrix $\mathbf{X} \in \mathbb{R}^{n\times p}$ with p-dimensional parameters, but they are of course immediately applicable to the augmented data $\tilde{\mathbf{X}} \in \mathbb{R}^{n\times q}$ and q-dimensional parameters.

3.1. Regression. In this first setting, we naturally consider the squared loss function, so that problem (1.1) can be reduced to

$$\min_{\mathbf{w}\in\mathbb{R}^p} \frac{1}{2n}\|\mathbf{y} - \mathbf{X}\mathbf{w}\|_2^2 + \lambda\Omega(\mathbf{w}) \quad \text{with } \lambda \geq 0.$$

Note that in this case, we have omitted the intercept b since we can center the vector $\mathbf{y}$ and the columns of $\mathbf{X}$ instead. Prediction for a new fMRI volume $\mathbf{x}$ is then simply performed by computing the dot product $\mathbf{x}^\top\mathbf{w}^*$.

3.2. Classification. We can look at our prediction task from a multiclass classification viewpoint. Specifically, we assume that $\mathcal{Y}$ is a finite set of integers $\{1,\dots,c\}$, $c > 2$, and consider both multiclass and "one-versus-all" strategies [50]. We need to slightly extend the formulation (1.1): To this end, we introduce the weight matrix $\mathbf{W} \triangleq [\mathbf{w}^1,\dots,\mathbf{w}^c] \in \mathbb{R}^{p\times c}$, composed of c weight vectors, along with a vector of intercepts $\mathbf{b} \in \mathbb{R}^c$.

A standard way of addressing multiclass classification problems consists in using a multilogit model, also known as multinomial logistic regression (see, e.g., [22] and references therein). In this case, class-conditional probabilities are modeled for each class by a softmax function; namely, given a fMRI volume $\mathbf{x}$, the probability of having the kth class label reads

$$\text{Prob}(y = k|\mathbf{x};\mathbf{W},\mathbf{b}) = \frac{\exp\{\mathbf{x}^\top\mathbf{w}^k + \mathbf{b}_k\}}{\sum_{r=1}^c \exp\{\mathbf{x}^\top\mathbf{w}^r + \mathbf{b}_r\}} \quad \text{for } k \in \{1,\dots,c\}. \tag{3.1}$$

The parameters $\{\mathbf{W},\mathbf{b}\}$ are then learned by maximizing the resulting (conditional) log-likelihood, which leads to the following optimization problem:

$$\min_{\substack{\mathbf{W}\in\mathbb{R}^{p\times c}\\ \mathbf{b}\in\mathbb{R}^c}} \frac{1}{n}\sum_{i=1}^n \log\Big[\sum_{k=1}^c e^{\mathbf{x}_i^\top(\mathbf{w}^k - \mathbf{w}^{\mathbf{y}_i}) + \mathbf{b}_k - \mathbf{b}_{\mathbf{y}_i}}\Big] + \lambda\sum_{k=1}^c \Omega(\mathbf{w}^k).$$

[2] In the absence of strict convexity, we cannot in general guarantee the uniqueness of $\mathbf{w}^*$ (or $\mathbf{W}^*$).

Whereas the regularization term is separable with respect to the different weight vectors $\mathbf{w}^k$, the loss function induces a coupling in the columns of $\mathbf{W}$. As a result, the optimization has to be carried out over the entire matrix $\mathbf{W}$. In this setting, and given a new fMRI volume $\mathbf{x}$, we make predictions by choosing the label that maximizes the class-conditional probabilities (3.1), that is, $\text{argmax}_{k\in\{1,\ldots,c\}}\text{Prob}(y = k|\mathbf{x}; \mathbf{W}^*, \mathbf{b}^*)$.

In section 5, we consider another multiclass classification scheme. The "one-versus-all" strategy (OVA) consists in training c different (real-valued) binary classifiers; each one is trained to distinguish the examples in a single class from the observations in all remaining classes. In order to classify a new example, among the c classifiers, the one which outputs the largest (most positive) value is chosen. In this framework, we consider binary classifiers built from both the squared and the logistic loss functions. If we denote by $\bar{\mathbf{Y}} \in \{-1, 1\}^{n\times c}$ the indicator response matrix defined as $\bar{\mathbf{Y}}_i^k \triangleq 1$ if $\mathbf{y}_i = k$ and -1 otherwise, we obtain

$$\min_{\mathbf{W}\in\mathbb{R}^{p\times c}} \frac{1}{2n}\sum_{k=1}^{c} \|\bar{\mathbf{Y}}^k - \mathbf{X}\mathbf{w}^k\|_2^2 + \lambda\sum_{k=1}^{c}\Omega(\mathbf{w}^k)$$

and

$$\min_{\substack{\mathbf{W}\in\mathbb{R}^{p\times c}\\ \mathbf{b}\in\mathbb{R}^c}} \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{c}\log\left[1 + e^{-\bar{\mathbf{Y}}_i^k(\mathbf{x}_i^\top\mathbf{w}^k+\mathbf{b}_k)}\right] + \lambda\sum_{k=1}^{c}\Omega(\mathbf{w}^k).$$

By invoking the same arguments as in section 3.1, the vector of intercepts $\mathbf{b}$ is again omitted in the above problem with the squared loss. Moreover, given a new fMRI volume $\mathbf{x}$, we predict the label k that maximizes the response $\mathbf{x}^\top[\mathbf{w}^*]^k$ among the c different classifiers. The case of the logistic loss function parallels the setting of the multinomial logistic regression, where each of the c OVA classifiers leads to a class-conditional probability; the predicted label is the one corresponding to the highest probability.

The formulations which we have reviewed in this section can be solved efficiently within the same optimization framework that we now introduce.

4. Optimization. The convex minimization problem (1.1) is challenging, since the penalty Ω as defined in (2.3) is nonsmooth and the number of variables to consider is large (we have $q \approx 10^5$ variables in the following experiments). These difficulties are well addressed by *forward-backward splitting methods*, which belong to the broader class of proximal methods. Forward-backward splitting schemes date back (at least) to [38, 34] and have been further analyzed in various settings (see, e.g., [59, 8, 12]); for a thorough review of proximal splitting techniques, we refer the interested readers to [11].

Our convex minimization problem (1.1) can be handled well by such techniques since it is the sum of two semi–lower-continuous, proper, convex functions with nonempty domain, and where one element—the loss function $\mathcal{L}(\mathbf{y}, \mathbf{X}, \ldots)$—is assumed differentiable with Lipschitz-continuous gradient (which notably covers the cases of the squared and simple/multinomial logistic functions, as introduced in section 3).

To describe the principle of forward-backward splitting methods, we need to introduce the concept of the *proximal operator*. The proximal operator associated with our regularization term $\lambda\Omega$, which we denote by $\text{Prox}_{\lambda\Omega}$, is the function that maps a vector $\mathbf{w} \in \mathbb{R}^p$ to the unique

solution of

$$\min_{\mathbf{v}\in\mathbb{R}^p} \frac{1}{2}\|\mathbf{v}-\mathbf{w}\|_2^2 + \lambda\Omega(\mathbf{v}). \tag{4.1}$$

This operator was initially introduced by Moreau [44] to generalize the projection operator onto a convex set; for a complete study of the properties of $\text{Prox}_{\lambda\Omega}$, see [12]. Based on definition (4.1), and given the current iterate $\mathbf{w}^{(k)}$,[3] the typical update rule of forward-backward splitting methods has the form[4]

$$\mathbf{w}^{(k+1)} \leftarrow \text{Prox}_{\frac{\lambda}{L}\Omega}\Big(\mathbf{w}^{(k)} - \frac{1}{L}\nabla\mathcal{L}_{\mathbf{w}}(\mathbf{y},\mathbf{X},\mathbf{w}^{(k)})\Big), \tag{4.2}$$

where $L>0$ is a parameter which is an upper bound on the Lipschitz constant of the gradient of $\mathcal{L}$. In light of the update rule (4.2), we can see that efficiently solving problem (4.1) is crucial to good performance. In addition, when the nonsmooth term Ω is not present, the previous proximal problem (4.2), also known as the implicit or backward step, exactly leads to the standard gradient update rule.

For many regularizations Ω of interest, the solution of problem (4.1) can actually be computed in closed form in simple settings: in particular, when Ω is the ℓ_1-norm, the proximal operator is the well-known soft-thresholding operator [15]. The work of [28] recently showed that the proximal problem (4.1) could be solved efficiently and exactly with Ω as defined in (2.3). The underlying idea of this computation is to solve a *well-ordered* sequence of simple proximal problems associated with each of the terms $\|\mathbf{w}_g\|$ for $g\in\mathcal{G}$. We refer the interested readers to [28] for further details on this norm and to [2] for a broader view.

In the subsequent experiments, we focus on accelerated multistep versions of forward-backward splitting methods (see, e.g., [46, 4, 63]),[5] where the proximal problem (4.2) is solved not for a current estimate but for an auxiliary sequence of points that are linear combinations of past estimates. These accelerated versions have increasingly drawn the attention of a broad research community since they can deal with large nonsmooth convex problems, and their convergence rates on the objective achieve the complexity bound of $O(1/k^2)$, with k denoting the iteration number. As a side comment, note that as opposed to standard one-step forward-backward splitting methods, nothing can be said about the convergence of the sequence of the iterates themselves. In our case, the cost of each iteration is dominated by the computation of the gradient (e.g., $O(np)$ for the squared loss) and the proximal operator, whose time complexity is linear, or close to linear, in p for the tree-structured regularization [28].

5. Experiments and results.

We now present experimental results on simulated data and real fMRI data.

[3] For clarity of the presentation, we do not consider the optimization of the intercept that we leave unregularized in all our experiments.

[4] For simplicity, we present only a constant-stepsize scheme; adaptive line search can also be used in this context and can lead to larger stepsizes [11].

[5] More precisely, we use the accelerated proximal gradient scheme (FISTA) taken from [4]. The Matlab/C++ implementation we use is available at http://www.di.ens.fr/willow/SPAMS/.

5.1. Simulations. In order to illustrate the proposed method, the hierarchical regularization with the ℓ_2-norm and $\eta_g = 1$ for all $g \in \mathcal{G}$ was applied in a regression setting on a small two-dimensional simulated dataset consisting of 300 square images (40×40 pixels, i.e., $\mathbf{X} \in \mathbb{R}^{300\times 1600}$). The weight vector $\mathbf{w}$ used in the simulation—itself an image of the same dimension—is presented in Figure 2(a). It consists of three localized regions of two different sizes that are predictive of the output. The images $\mathbf{x}^{(i)}$ are sampled so as to obtain a correlation structure which mimics fMRI data. Precisely, each image $\mathbf{x}^{(i)}$ was obtained by smoothing a completely random image, where each pixel was drawn independent and identically distributed (i.i.d.) from a normal distribution, with a Gaussian kernel (standard deviation 2 pixels), which introduces spatial correlations between neighboring pixels. Subsequently, additional correlations between the regions corresponding to the three patterns were introduced in order to simulate coactivations between different brain regions by multiplying the signal by the square-root of an appropriate covariance matrix $\mathbf{\Sigma}$. Specifically, $\mathbf{\Sigma} \in \mathbb{R}^{1600\times 1600}$ is a spatial covariance between voxels, with diagonal set to $\mathbf{\Sigma}_{i,i} = 1$ for all i, and with two off-diagonal blocks. Let us denote $\mathcal{C}_1$ and $\mathcal{C}_2$ the set of voxels forming the two larger patterns, and $\mathcal{C}_3$ the voxels in the small pattern. The covariance coefficients are set to $\mathbf{\Sigma}_{i,j} = 0.3$ for $i \in \mathcal{C}_1$ and $j \in \mathcal{C}_2$, and $\mathbf{\Sigma}_{i,j} = -0.2$ for $i \in \mathcal{C}_2$ and $j \in \mathcal{C}_3$. The covariance is of course symmetric.

The choice of the weights and of the correlation introduced in the images aims at illustrating how the hierarchical regularization estimates weights at different resolutions in the image. The targets were simulated by forming $\mathbf{w}^\top \mathbf{x}^{(i)}$ corrupted with an additive white noise (signal-to-noise ration (SNR) = 10dB). The loss used was the squared loss as detailed in section 3.1. The regularization parameter was estimated with two-fold cross-validation (150 images per fold) on a logarithmic grid of 30 values between 10^3 and 10^{-3}.

The components of the estimated weight vector $\mathbf{w}^*$ at different scales are presented in the images of Figure 2, with each image corresponding to a different depth in the tree. For a given tree depth, an image is formed from the corresponding parcellation. All the voxels within a parcel are colored according to the associated scalar in $\mathbf{w}^*$. It can be observed that all three patterns are present in the weight vector but at different depths in the tree. The small activation in the top right-hand corner shows up mainly at scale 3, while the bigger patterns appear higher in the tree at scales 5 and 6. This simulation clearly illustrates the ability of the method to capture informative spatial patterns at different scales.

5.2. Description of the fMRI data. We apply the different methods to analyze the data of ten subjects from an fMRI study originally designed to investigate object coding in visual cortex (see [17] for details). During the experiment, ten healthy volunteers viewed objects of two categories (each one of the two categories is used in half of the subjects) with four different exemplars in each category. Each exemplar was presented at three different sizes (yielding 12 different experimental conditions per subject). Each stimulus was presented four times in each of the six sessions. We averaged data from the four repetitions, resulting in a total of $n = 72$ volumes by subject (one volume of each stimulus by session). Functional volumes 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 2s (echo time, 30ms; flip angle, 70°; $2 \times 2 \times 2$-mm voxels; 0.5-mm gap). Realignment, spatial normalization to common space (Montreal Neurological

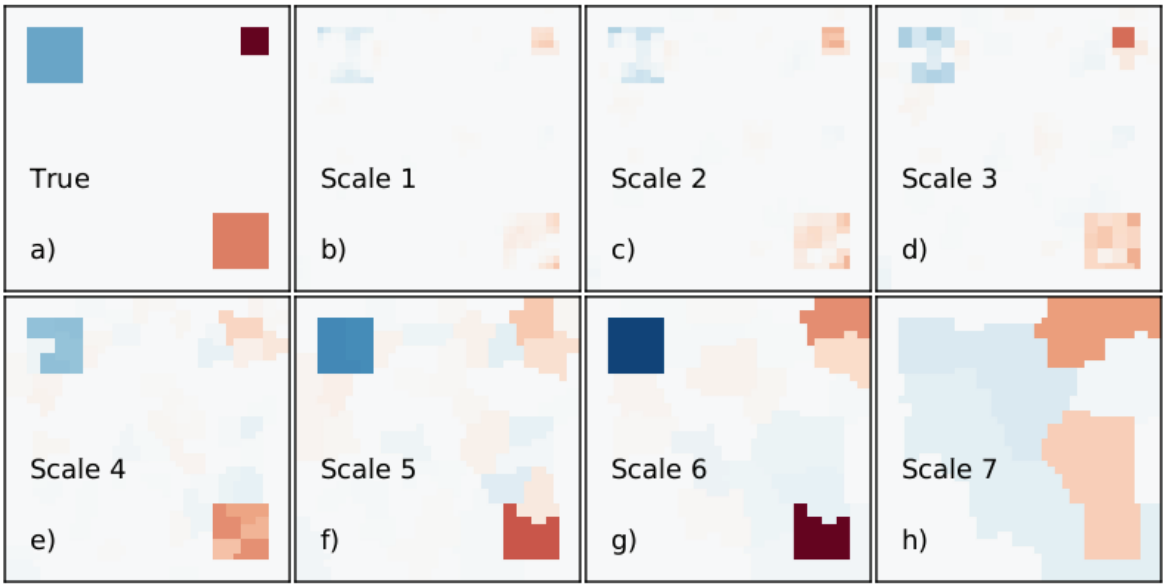

Figure 2. *Weights* $\mathbf{w}^*$ *estimated in the simulation study. The true coefficients are presented in* (a) *and the estimated weights at different scales, i.e., different depths in the tree, are presented in* (b)–(h) *with the same colormap.*

Institute (MNI)) convention, slice timing correction, and GLM fit were performed with the SPM5 software.[6] In the GLM, the time course 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 with a high-pass filter with a cut-off of 128s (7.8×10^{-3}Hz). In the present work we used the resulting sessionwise parameter estimate volumes. Contrary to common practice in the field, the data were not smoothed with an isotropic Gaussian filter. All the analyses are performed on the whole acquired volume.

The four different exemplars in each of the two categories were pooled, leading to volumes labeled according to the three possible sizes of the object. We are thus interested in finding discriminative information to predict the size of the presented object.

This can be reduced to either a regression problem in which our goal is to predict the class label of the size of the presented object (i.e., $y \in \{0, 1, 2\}$)[7] or a three-category classification problem, each size corresponding to a category. We perform an intersubject analysis on the sizes both in regression and classification settings. This analysis relies on subject-specific fixed-effects activations; i.e., for each condition, the six activation maps corresponding to the six sessions are averaged together. This yields a total of 12 volumes per subject, one for each experimental condition. The dimensions of the real dataset are $p \approx 7 \times 10^4$ and $n = 120$ (divided into three different sizes). We evaluate the performance of the method by

[6] http://www.fil.ion.ucl.ac.uk/spm/software/spm5.

[7] An interesting alternative would be to consider some real-valued dimension such as the field of view of the object.

cross-validation with a natural data splitting—*leave-one-subject-out*. Each fold consists of 12 volumes. The parameter λ of all methods is optimized over a grid of 30 values of the form 2^k, with a nested leave-one-subject-out cross-validation on the training set. The exact scaling of the grid varies for each model to account for different Ω.

5.3. Methods involved in the comparisons. In addition to considering standard ℓ_1- and squared ℓ_2-regularizations in both our regression and multiclass classification tasks, we compare various methods that we now review.

First, when the regularization Ω as defined in (2.3) is employed, we consider three settings of values for $(\eta_g)_{g\in\mathcal{G}}$ which leverage the tree structure $\mathcal{T}$. More precisely, we set $\eta_g = \rho^{\text{depth}(g)}$ for g in $\mathcal{G}$, with $\rho \in \{0.5, 1, 1.5\}$ and where depth(g) denotes the depth of the root of the group g in $\mathcal{T}$. In other words, the larger ρ is, the more averse we are to selecting small (and variable) parcels located near the leaves of $\mathcal{T}$. As the results illustrate, the choice of ρ can have a significant impact on the performance. More generally, the problem of selecting ρ properly is a difficult question which is still under investigation, both theoretically and practically; see, e.g., [1].

The greedy approach from [42] is included in the comparisons for both the regression and classification tasks. It relies on a top-down exploration of the tree $\mathcal{T}$. In short, starting from the root parcel that contains all the voxels, we choose at each step the split of the parcel that yields the highest prediction score. The exploration step is performed until a given number of parcels is reached, and it yields a set of nested parcellations with increasing complexity. Similarly to a model selection step, we choose the best parcellation among those found in the exploration step. The selected parcellation is thus used on the test set. In the regression setting, this approach is combined with Bayesian ridge regression, while it is associated with a linear support vector machine for the classification task (whose regularization parameter, often referred to as C in the literature [54], is found by nested cross-validation in $\{0.01, 0.1, 1\}$).

5.3.1. Regression setting. In order to evaluate whether the level of sparsity is critical in our analysis, we implemented a reweighted ℓ_1-scheme [5]. In this case, sparsity is encouraged more aggressively as a multistage convex relaxation of a concave penalty. Specifically, it consists in iteratively using a weighted ℓ_1-norm, whose weights are determined by the solution of previous iteration. Moreover, we additionally compare this to Elastic net [67], whose second regularization parameter is set by cross-validation as a fraction of λ, that is, $\alpha\lambda$ with $\alpha \in \{0.5, 0.05, 0.005, 0.0005\}$.

To better understand the added value of the hierarchical norm (2.3) over unstructured penalties, we consider not only the plain ℓ_1-norm in the augmented feature space but also another variant of weighted ℓ_1-norm. The weights are manually set and reflect the underlying tree structure $\mathcal{T}$. By analogy with the choice of $(\eta_g)_{g\in\mathcal{G}}$ made for the tree-structured regularization, we take exponential weights depending on the depth of the variable j, where $\eta_j = \rho^{\text{depth}(j)}$ with $\rho \in \{0.5, 1.5\}$.[8] We also tried weights $(\eta_j)_{j\in\{1,\dots,p\}}$ that are linear with respect to the depths, i.e., $\eta_j = \frac{\text{depth}(j)}{\max_{k\in\{1,\dots,p\}} \text{depth}(k)}$, but those led to worse results. In Table 1, we present only the best result of this weighted ℓ_1-norm, obtained with the exponential weight

[8] Formally, the depth of the feature j is equal to depth(g_j), where g_j is the smallest group in $\mathcal{G}$ that contains j (*smallest* is understood here in the sense of the inclusion).

Table 1

*Prediction results obtained on fMRI data (see text) for the regression setting. From the left, the first column contains the mean and standard deviation of the test error (mean squared error), computed over leave-one-subject-out folds. The best performance is obtained with the greedy technique and the hierarchical ℓ_2 penalization ($\rho = 1$) constructed from the Ward tree. Methods with performance significantly worse than the latter are assessed by Wilcoxon two-sample paired signed rank tests. (The superscript * indicates a rejection at 5%.) Levels of sparsity reported are in the augmented space whenever it is used.*

Loss function	Squared loss		
	Error (mean,std)	P-value w.r.t. Tree ℓ_2 ($\rho = 1$)	Median fraction of nonzeros (%)
Regularization			
ℓ_2 (ridge)	(13.8, 7.6)	0.096	100.00
ℓ_1	(20.2, 10.9)	0.013*	0.11
$\ell_1 + \ell_2$ (elastic net)	(14.4, 8.8)	0.065	0.14
Reweighted ℓ_1	(18.8, 14.6)	0.052	0.10
ℓ_1 (augmented space)	(14.2, 7.9)	0.096	0.02
ℓ_1 (tree weights)	(13.9, 7.9)	0.032*	0.02
Tree ℓ_2 ($\rho = 0.5$)	(13.0, 7.4)	0.137	99.99
Tree ℓ_2 ($\rho = 1$)	(**11.8**, **6.7**)	-	9.36
Tree ℓ_2 ($\rho = 1.5$)	(13.5, 7.0)	0.080	0.04
Tree ℓ_∞ ($\rho = 0.5$)	(13.6, 7.8)	0.080	99.99
Tree ℓ_∞ ($\rho = 1$)	(12.8, 6.7)	0.137	1.22
Tree ℓ_∞ ($\rho = 1.5$)	(13.0, 6.8)	0.096	0.04
	Error (mean,std)	P-value w.r.t. Tree ℓ_2 ($\rho = 1$)	Median fraction of nonzeros (%)
Greedy	(12.0, 5.5)	0.5	0.01

and $\rho = 1.5$. We now turn to the models taking part in the classification task.

5.3.2. Classification setting. As discussed in section 3.2, the optimization in the classification setting is carried out over a matrix of weights $\mathbf{W} \in \mathbb{R}^{p \times c}$. This makes it possible to consider regularization schemes that couple the selection of variables across rows of that matrix.

In particular, we apply ideas from *multitask* learning [47] by viewing each class as a task. More precisely, we use a regularization norm defined by $\Omega_{\text{multitask}}(\mathbf{W}) \triangleq \sum_{j=1}^{p} \|\mathbf{W}_j\|$, where $\|\mathbf{W}_j\|$ denotes either the ℓ_2- or ℓ_∞-norm of the jth row of $\mathbf{W}$. The rationale for the definition of $\Omega_{\text{multitask}}$ is to assume that the set of relevant voxels is the same across the c different classes, so that sparsity is induced simultaneously over the columns of $\mathbf{W}$. It should be noted that, in the OVA setting, although the loss functions for the c classes are decoupled, the use of $\Omega_{\text{multitask}}$ induces a relationship that ties the optimization problems together.

Note that the tree-structured regularization Ω we consider does not impose a joint pattern-selection across the c different classes. It would, however, be possible to use Ω over the matrix $\mathbf{W}$ in a multitask setting. More precisely, we would define $\Omega(\mathbf{W}) = \sum_{g \in \mathcal{G}} \|\mathbf{W}_g\|$, where $\|\mathbf{W}_g\|$ denotes either the ℓ_2- or ℓ_∞-norm of the vectorized submatrix $\mathbf{W}_g$ defined by $\mathbf{W}_g \triangleq [\mathbf{W}_{jk}]_{j \in g, k \in \{1,\ldots,c\}}$. This definition constitutes a direct extension of the standard nonoverlapping ℓ_1/ℓ_2- and ℓ_1/ℓ_∞-norms used for multitask. Furthermore, it is worth noting that the optimization tools from [28] would still apply for this tree-structured matrix regularization.

5.4. Results. We present results comparing our approach based on the hierarchical sparsity-inducing norm (2.3) with the regularization listed in the previous section. For each method, we computed the cross-validated prediction accuracy and the percentage of nonzero coefficients, i.e., the level of sparsity of the model.

5.4.1. Regression results. The results for the intersubject regression analysis are reported in Table 1. The lowest error in prediction accuracy is obtained by both the greedy strategy and the proposed hierarchical structured sparsity approach (Tree ℓ_2 with $\rho = 1$) whose performances are essentially indistinguishable. Both also yield one of the lowest standard deviations indicating that the results are most stable. This can be explained by the fact that the use of local signal averages in the proposed algorithm is a good way to increase the robustness against intersubject variability. We also notice that the sparsity-inducing approaches (Lasso and reweighted ℓ_1) have the highest error in prediction accuracy, probably because the obtained solutions are too sparse, and suffer from the absence of perfect voxel-to-voxel correspondence between subjects.

In terms of sparsity, we can see, as expected, that ridge regression does not yield any sparsity and that the Lasso solution is very sparse (in the feature space, with approximately 7×10^4 voxels). Our method yields a median value of 9.36% of nonzero coefficients (in the augmented space of features, with about 1.4×10^5 nodes in the tree). The maps of weights obtained with Lasso and the hierarchical regularization for one fold, are reported in Figure 3. The Lasso yields a scattered and overly sparse pattern of voxels, which is not easily readable, while our approach extracts a pattern of voxels with a compact structure that clearly outlines brain regions expected to activate differentially for stimuli with different low-level visual properties, e.g., sizes; it corresponds to the early visual cortex in the occipital lobe at the back of the brain. Interestingly, the patterns of voxels show some symmetry between left and right hemispheres, especially in the primary visual cortex, which is located at the back and center of the brain. This observation is consistent with the current understanding in neuroscience that the symmetric parts of this brain region process, respectively, the visual contents of each of the visual hemifields. The weights obtained at different depth levels in the tree, corresponding to different scales, show that the largest coefficients are concentrated at the higher scales (scale 6 in Figure 3), which suggests that the object sizes cannot be well decoded at the voxel level but require features corresponding to larger clusters of voxels.

5.4.2. Classification results. The results for the intersubject classification analysis are reported in Table 2. The best performance is obtained with a multinomial logistic loss function, also using the hierarchical ℓ_2 penalization ($\rho = 1$).

It should be noted that the sparsity level of the different models varies widely depending on the loss and regularization used. With the squared loss, ℓ_1-type regularization, including the multitask regularizations based on the ℓ_1/ℓ_2- and ℓ_1/ℓ_∞-norms, tends to select quite sparse models, which keep around 0.1% of the voxels. When using logistic-type losses, these regularizations tend to select a significantly large number of voxels, which suggests that the selection problem is really difficult and that these methods are unstable. For the methods with hierarchical regularization, on the contrary, the sparsity tends to improve with the choice of loss functions that are better suited to the classification problem. Tuning ρ trades sparsity of the model for performance, resulting in models that are not sparse when ρ is small

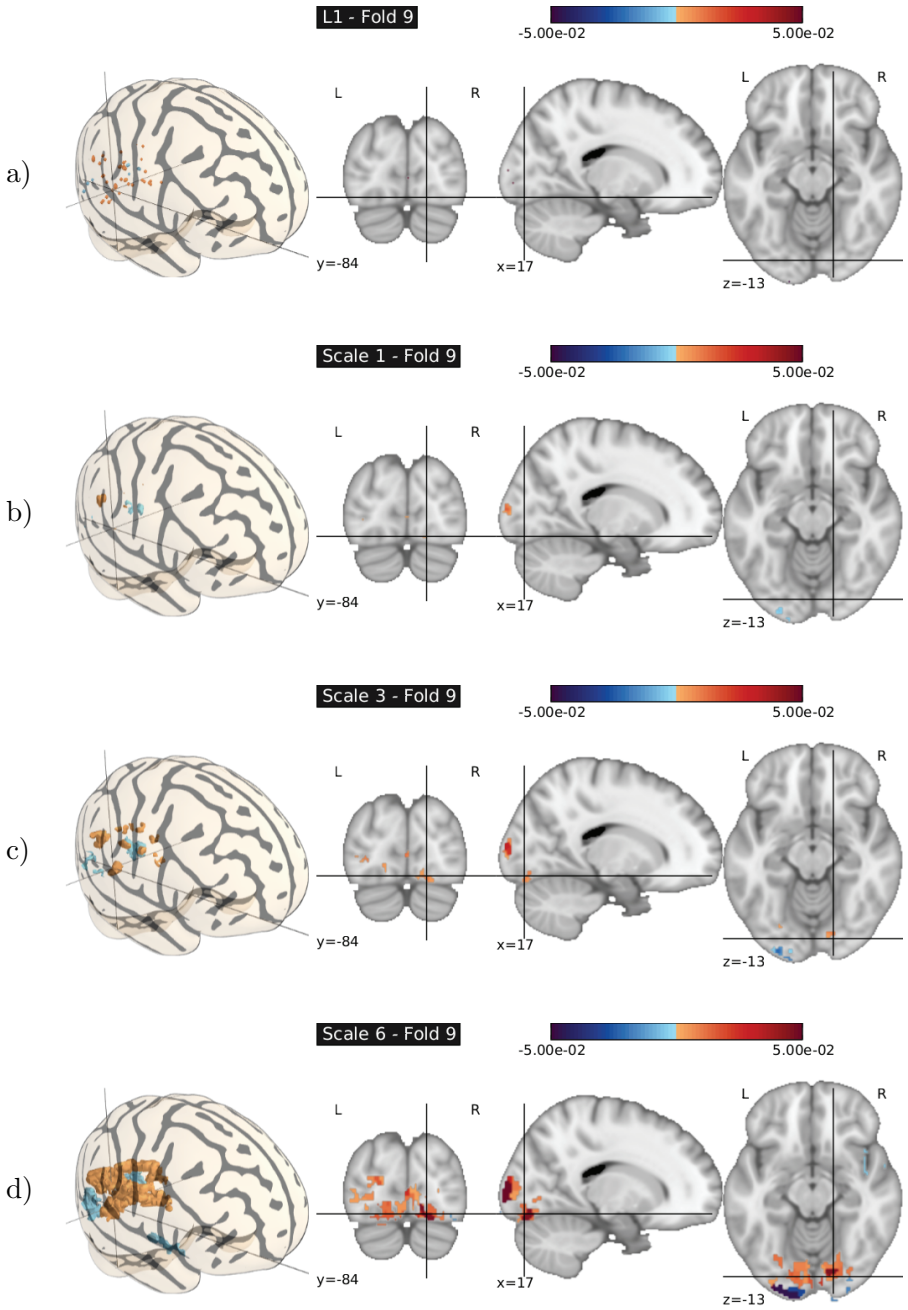

Figure 3. *Maps of weights obtained using different regularizations in the regression setting.* (a) ℓ_1 *regularization. We can notice that the predictive pattern obtained is excessively sparse and is not easily readable despite being mainly located in the occipital cortex.* (b)–(d) *Tree* ℓ_2 *regularization* ($\rho = 1$) *at different scales. In this case, the regularization algorithm extracts a pattern of voxels with a compact structure that clearly outlines early visual cortex which is expected to discriminate between stimuli of different sizes. 3D images were generated with Mayavi* [48].

Table 2

Prediction results obtained on fMRI data (see text) for the multiclass classification setting. From the left, the first column contains the mean and standard deviation of the test error (percentage of misclassification), computed over leave-one-subject-out folds. The best performance is obtained with the hierarchical ℓ_2 penalization ($\rho = 1$) constructed from the Ward tree, coupled with the multinomial logistic loss function. Methods with performance significantly worse than this combination are assessed by Wilcoxon two-sample paired signed rank tests. (The superscript * indicates a rejection at 5%.) Levels of sparsity reported are in the augmented space whenever it is used.

Loss function	Squared loss ("one-versus-all")		
	Error (mean,std)	P-value w.r.t. Tree ℓ_2 ($\rho = 1$)-ML	Median fraction of nonzeros (%)
Regularization			
ℓ_2 (ridge)	(29.2, 5.9)	0.004*	100.00
ℓ_1	(33.3, 6.8)	0.004*	0.10
ℓ_1/ℓ_2 (multitask)	(31.7, 9.5)	0.004*	0.12
ℓ_1/ℓ_∞ (multitask)	(33.3,13.6)	0.009*	0.22
Tree ℓ_2 ($\rho = 0.5$)	(25.8, 9.2)	0.004*	99.93
Tree ℓ_2 ($\rho = 1$)	(25.0, 5.5)	0.027*	10.08
Tree ℓ_2 ($\rho = 1.5$)	(24.2, 9.9)	0.130	0.05
Tree ℓ_∞ ($\rho = 0.5$)	(30.8, 8.8)	0.004*	59.49
Tree ℓ_∞ ($\rho = 1$)	(24.2, 7.3)	0.058	1.21
Tree ℓ_∞ ($\rho = 1.5$)	(25.8, 10.7)	0.070	0.04
Loss function	Logistic loss ("one-versus-all")		
	Error (mean,std)	P-value w.r.t. Tree ℓ_2 ($\rho = 1$)-ML	Median fraction of nonzeros (%)
Regularization			
ℓ_2 (ridge)	(25.0, 9.6)	0.008*	100.00
ℓ_1	(34.2, 15.9)	0.004*	0.55
ℓ_1/ℓ_2 (multitask)	(31.7, 8.6)	0.002*	47.35
ℓ_1/ℓ_∞ (multitask)	(33.3, 10.4)	0.002*	99.95
Tree ℓ_2 ($\rho = 0.5$)	(25.0, 9.6)	0.007*	99.93
Tree ℓ_2 ($\rho = 1$)	(20.0, 11.2)	0.250	7.88
Tree ℓ_2 ($\rho = 1.5$)	(18.3, 6.6)	0.500	0.06
Tree ℓ_∞ ($\rho = 0.5$)	(30.8, 10.4)	0.004*	59.42
Tree ℓ_∞ ($\rho = 1$)	(24.2, 6.1)	0.035*	0.60
Tree ℓ_∞ ($\rho = 1.5$)	(21.7, 8.9)	0.125	0.03
Loss function	Multinomial logistic loss (ML)		
	Error (mean,std)	P-value w.r.t. Tree ℓ_2 ($\rho = 1$)-ML	Median fraction of nonzeros (%)
Regularization			
ℓ_2 (ridge)	(24.2, 9.2)	0.035*	100.00
ℓ_1	(25.8, 12.0)	0.004*	97.95
ℓ_1/ℓ_2 (multitask)	(26.7, 7.6)	0.007*	30.24
ℓ_1/ℓ_∞ (multitask)	(26.7, 11.6)	0.002*	99.98
Tree ℓ_2 ($\rho = 0.5$)	(22.5, 8.8)	0.070	83.06
Tree ℓ_2 ($\rho = 1$)	(**16.7**, **10.4**)	-	4.87
Tree ℓ_2 ($\rho = 1.5$)	(18.3, 10.9)	0.445	0.02
Tree ℓ_∞ ($\rho = 0.5$)	(26.7, 11.6)	0.015*	48.82
Tree ℓ_∞ ($\rho = 1$)	(22.5, 13.0)	0.156	0.34
Tree ℓ_∞ ($\rho = 1.5$)	(21.7, 8.9)	0.460	0.05
	Error (mean,std)	P-value w.r.t. Tree ℓ_2 ($\rho = 1$)-ML	Median fraction of nonzeros (%)
Greedy	(21.6, 14.5)	0.001*	0.01

to very sparse models when ρ is large. In particular, a better compromise between sparsity and prediction performance can probably be obtained by tuning $\rho \in [1, 1.5]$.

For both ℓ_1 and hierarchical regularizations, one of the three vectors of coefficients obtained for one fold and for the loss leading to sparser models is presented in Figure 4. For ℓ_1, the active voxels are scattered all over the brain, and for losses other than the squared-loss the models selected tend not to be sparse. By contrast, the tree ℓ_2-regularization yields clearly delineated sparsity patterns located in the visual areas of the brain. Like for the regression results, the highest coefficients are obtained at scale 6 showing how spatially extended the brain region involved in the cognitive task is. The symmetry of the pattern at this scale is also particularly striking in the primary visual areas. It also extends more anteriorly into the inferior temporal cortex, known for high-level visual processing.

6. Conclusion. In this article, we introduced a hierarchically structured regularization which takes into account the spatial and multiscale structure of fMRI data. This approach copes with intersubject variability in a similar way to feature agglomeration by averaging neighboring voxels. Although alternative agglomeration strategies do exist, we simply used the criterion which appears as the most natural, Ward's clustering criterion, and which builds parcels with little variance.

Results on a real dataset show that the proposed algorithm is a promising tool for mining fMRI data. It yields similar or higher prediction accuracy than reference methods, and the map of weights it obtains exhibits a cluster-like structure. The map is easily readable compared to the overly sparse patterns found by classical sparsity-promoting approaches.

For the regression problem, both the greedy method from [42] and the proposed algorithm yield better results than unstructured and nonhierarchical regularizations, whereas in the classification setting, our approach leads to the best performance. Moreover, our proposed methods enjoy the different benefits of convex optimization. In particular, while the greedy algorithm relies on a two-step approach that may be far from optimal, the hierarchical regularization induces simultaneously the selection of the optimal parcellation and the construction of the optimal predictive model given the initial hierarchical clustering of the voxels. Moreover, convex methods yield predictors that are essentially stable with respect to perturbations of the design or the initial clustering, which is typically not the case of greedy methods. It is important to distinguish here the stability of the predictors from that of the only learned map $\mathbf{w}$, which could be enforced via a squared ℓ_2-norm regularization.

Finally, it should be mentioned that the performance achieved by this approach in intersubject problems suggests that it could potentially be used successfully for medical diagnosis, in a context where brain images—not necessarily functional images—are used to classify individuals into a diseased or a control population. Indeed, for difficult problems of that sort, where the reliability of the diagnostic is essential, the stability of models obtained from convex formulations and the interpretability of sparse and localized solutions are quite relevant to providing a credible diagnostic.

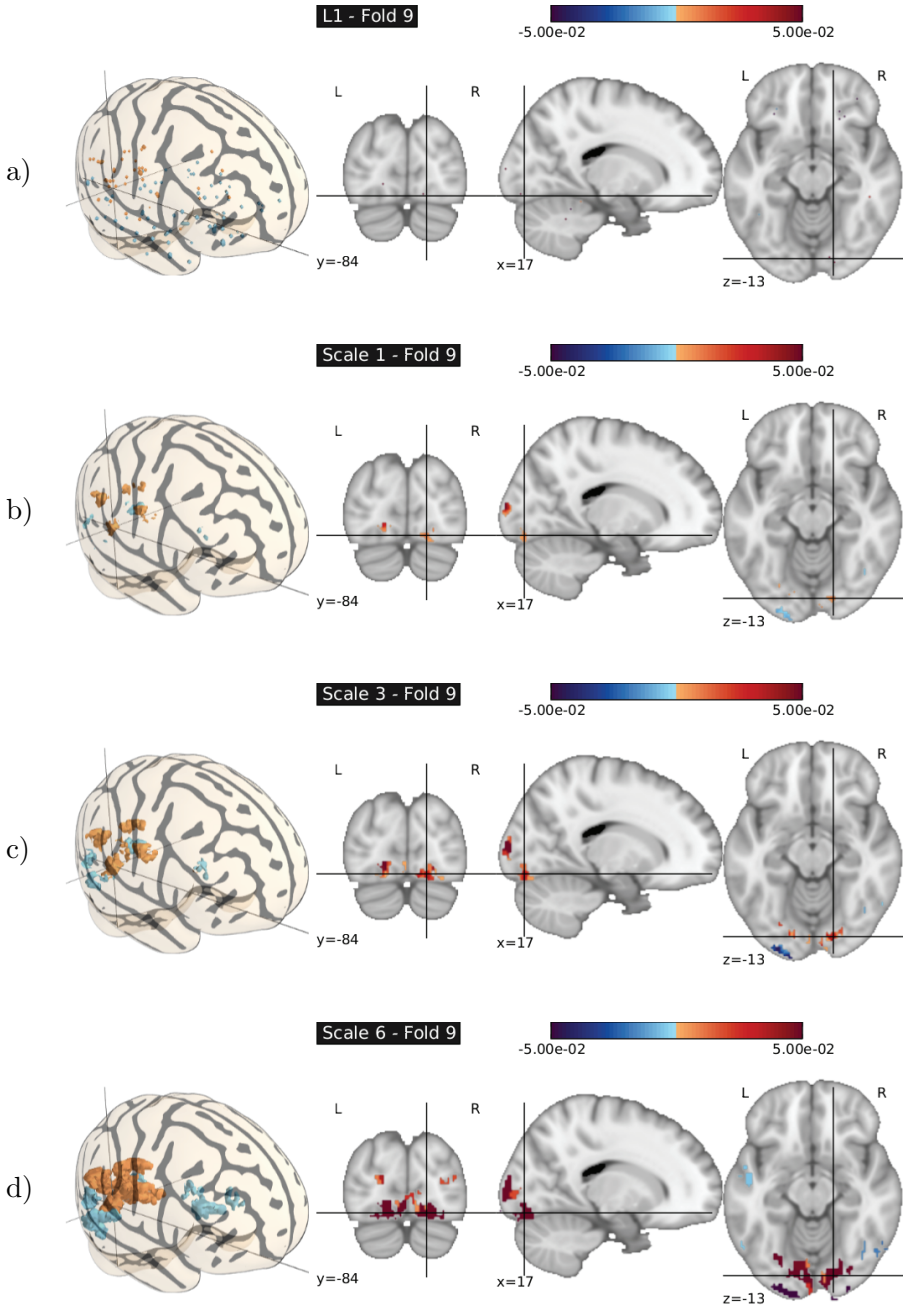

Figure 4. *Maps of weights obtained using different regularizations in the classification setting.* (a) ℓ_1 *regularization (with squared loss). We can notice that the predictive pattern obtained is excessively sparse and is not easily readable with voxels scattered all over the brain.* (b)–(d) *Tree regularization (with multinomial logistic loss) at different scales. In this case, the regularization algorithm extracts a pattern of voxels with a compact structure that clearly outlines early visual cortex which is expected to discriminate between stimuli of different sizes.*

REFERENCES

[1] F. Bach, *High-Dimensional Non-Linear Variable Selection Through Hierarchical Kernel Learning*, preprint, http://arxiv.org/abs/0909.0844, 2009.

[2] F. Bach, R. Jenatton, J. Mairal, and G. Obozinski, *Convex optimization with sparsity-inducing norms*, in Optimization for Machine Learning, S. Sra, S. Nowozin, and J. S. Wright, eds., MIT Press, Cambridge, MA, 2011.

[3] R. G. Baraniuk, V. Cevher, M. F. Duarte, and C. Hegde, *Model-based compressive sensing*, IEEE Trans. Inform. Theory, 56 (2010), pp. 1982–2001.

[4] A. Beck and M. Teboulle, *A fast iterative shrinkage-thresholding algorithm for linear inverse problems*, SIAM J. Imaging Sci., 2 (2009), pp. 183–202.

[5] E. J. Candes, M. B. Wakin, and S. P. Boyd, *Enhancing sparsity by reweighted L1 minimization*, J. Fourier Anal. Appl., 14 (2008), pp. 877–905.

[6] 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, 44 (2009), pp. 112 – 122.

[7] V. Cevher, M. F. Duarte, C. Hegde, and R. G. Baraniuk, *Sparse signal recovery using markov random fields*, in Proceedings of the 22nd Annual Conference on Neural Information Processing Systems, Advances in Neural Information Processing Systems 21, MIT Press, Cambridge, MA, 2008, pp. 257–264.

[8] G. H.-G. Chen and R. T. Rockafellar, *Convergence rates in forward–backward splitting*, SIAM J. Optim., 7 (1997), pp. 421–444.

[9] S. S. Chen, D. L. Donoho, and M. A. Saunders, *Atomic decomposition by basis pursuit*, SIAM J. Sci. Comput., 20 (1998), pp. 33–61.

[10] D. B. Chklovskii and A. A. Koulakov, *Maps in the brain: What can we learn from them?*, Ann. Rev. Neurosci., 27 (2004), pp. 369–392.

[11] P. L. Combettes and J.-C. Pesquet, *Proximal splitting methods in signal processing*, in Fixed-Point Algorithms for Inverse Problems in Science and Engineering, Springer, New York, 2011, pp. 185–212.

[12] P. L. Combettes and V. R. Wajs, *Signal recovery by proximal forward-backward splitting*, Multiscale Model. Simul., 4 (2005), pp. 1168–1200.

[13] 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, 19 (2003), pp. 261–270.

[14] S. Dehaene, G. Le Clec'H, L. Cohen, J.-B. Poline, P.-F. van de Moortele, and D. Le Bihan, *Inferring behavior from functional brain images*, Nature Neurosci., 1 (1998), pp. 549–550.

[15] D. L. Donoho and I. M. Johnstone, *Adapting to unknown smoothness via wavelet shrinkage*, J. Amer. Statist. Assoc., 90 (1995), pp. 1200–1224.

[16] M. F. Duarte and Y. C. Eldar, *Structured Compressed Sensing: From Theory to Applications*, preprint, http://arxiv.org/abs/1106.6224, 2011.

[17] E. Eger, C. Kell, and A. Kleinschmidt, *Graded size sensitivity of object exemplar evoked activity patterns in human loc subregions*, J. Neurophysiol., 100 (2008), pp. 2038-2047.

[18] G. Flandin, F. Kherif, X. Pennec, G. Malandain, N. Ayache, and J.-B. Poline, *Improved detection sensitivity in functional MRI data using a brain parcelling technique*, in Medical Image Computing and Computer-Assisted Intervention (MICCAI'02), Springer, New York, 2002, pp. 467–474.

[19] 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, 2 (1995), pp. 189–210.

[20] L. Grosenick, S. Greer, and B. Knutson, *Interpretable classifiers for FMRI improve prediction of purchases*, IEEE Trans. Neural Systems and Rehabilitation Engineering, 16 (2009), pp. 539–548.

[21] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, *Gene selection for cancer classification using support vector machines*, Machine Learning, 46 (2002), pp. 389–422.

[22] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed., Springer, New York, 2009.

[23] J. Huang and T. Zhang, *The benefit of group sparsity*, Ann. Statist., 38 (2010), pp. 1978–2004.

[24] J. Huang, T. Zhang, and D. Metaxas, *Learning with structured sparsity*, in Proceedings of the International Conference on Machine Learning (ICML), Montreal, Canada, 2009, pp. 417–424.
[25] L. Jacob, G. Obozinski, and J.-P. Vert, *Group Lasso with overlaps and graph Lasso*, in Proceedings of the International Conference on Machine Learning (ICML), Montreal, Canada, 2009, pp. 433–440.
[26] R. Jenatton, J.-Y. Audibert, and F. Bach, *Structured variable selection with sparsity-inducing norms*, J. Mach. Learn. Res., 12 (2011), pp. 2777–2824.
[27] R. Jenatton, A. Gramfort, V. Michel, G. Obozinski, F. Bach, and B. Thirion, *Multi-scale mining of fMRI data with hierarchical structured sparsity*, in International Workshop on Pattern Recognition in Neuroimaging (PRNI), Seoul, Korea, 2011, pp. 69–72.
[28] R. Jenatton, J. Mairal, G. Obozinski, and F. Bach, *Proximal methods for hierarchical sparse coding*, J. Mach. Learn. Res., 12 (2011), pp. 2297–2334.
[29] S. C. Johnson, *Hierarchical clustering schemes*, Psychometrika, 32 (1967), pp. 241–254.
[30] S. Kim and E. P. Xing, *Tree-guided group Lasso for multi-task regression with structured sparsity*, in Proceedings of the International Conference on Machine Learning (ICML), Haifa, Israel, 2010, pp. 543–550.
[31] M. Kowalski, *Sparse regression using mixed norms*, Appl. Comput. Harmon. Anal., 27 (2009), pp. 303–324.
[32] S. LaConte, S. Strother, V. Cherkassky, J. Anderson, and X. Hu, *Support vector machines for temporal classification of block design fMRI data*, NeuroImage, 26 (2005), pp. 317–329.
[33] E. L. Lehmann and J. P. Romano, *Testing statistical hypotheses*, Springer-Verlag, New York, 2005.
[34] P. L. Lions and B. Mercier, *Splitting algorithms for the sum of two nonlinear operators*, SIAM J. Numer. Anal., 16 (1979), pp. 964–979.
[35] J. Liu, S. Ji, and J. Ye, *Multi-task feature learning via efficient $\ell_{2,1}$-norm minimization*, in Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence (UAI), Montreal, Canada, 2009, pp. 339–348.
[36] J. Liu and J. Ye, *Fast Overlapping Group Lasso*, preprint, http://arxiv.org/abs/1009.0306, 2010.
[37] A. Marquand, M. Howard, M. Brammer, C. Chu, S. Coen, and J. Mourao-Miranda, *Quantitative prediction of subjective pain intensity from whole-brain fMRI data using gaussian processes*, NeuroImage, 49 (2010), pp. 2178–2189.
[38] B. Martinet, *Régularisation d'inéquations variationnelles par approximations successives.*, Revue Française Inform. Recherche Opérationnelle Ser. R-3, 4 (1970), pp. 154–158.
[39] F. D. Martino, G. Valente, N. Staeren, J. Ashburner, R. Goebel, and E. Formisano, *Combining multivariate voxel selection and support vector machines for mapping and classification of fMRI spatial patterns*, NeuroImage, 43 (2008), pp. 44–58.
[40] A. F. T. Martins, N. A. Smith, P. M. Q. Aguiar, and M. A. T. Figueiredo, *Structured sparsity in structured prediction*, in Proceedings of the Conference on Empirical Methods in Natural Language Processing, Edinburgh, Scotland, 2011, pp. 1500–1511.
[41] C. A. Micchelli, J. M. Morales, and M. Pontil, *A family of penalty functions for structured sparsity*, in Advances in Neural Information Processing Systems, MIT Press, Cambridge, MA, 2010, pp. 1612–1623.
[42] V. Michel, E. Eger, C. Keribin, J.-B. Poline, and B. Thirion, *A supervised clustering approach for extracting predictive information from brain activation images*, in MMBIA'10, San Francisco, CA, 2010.
[43] V. Michel, A. Gramfort, G. Varoquaux, E. Eger, and B. Thirion, *Total variation regularization for fMRI-based prediction of behaviour*, IEEE Trans. Medical Imaging, PP (2011), p. 1.
[44] J.-J. Moreau, *Fonctions convexes duales et points proximaux dans un espace hilbertien*, C. R. Acad. Sci. Paris, 255 (1962), pp. 2897–2899.
[45] F. Murtagh, *A survey of algorithms for contiguity-constrained clustering and related problems*, The Computer Journal, 28 (1985), pp. 82–88.
[46] Y. Nesterov, *Gradient Methods for Minimizing Composite Objective Function*, Tech. report, Center for Operations Research and Econometrics (CORE), Catholic University of Louvain, Louvain,-la-Neuve, Belgium, 2007.
[47] G. Obozinski, B. Taskar, and M. I. Jordan, *Joint covariate selection and joint subspace selection for multiple classification problems*, Statist. Comput., 20 (2010), pp. 231–252.

[48] P. Ramachandran and G. Varoquaux, *Mayavi: 3D visualization of scientific data*, Computing in Science Engineering, 13 (2011), pp. 40 –51.
[49] N. S. Rao, R. D. Nowak, S. J. Wright, and N. G. Kingsbury, *Convex approaches to model wavelet sparsity patterns*, in International Conference on Image Processing (ICIP), Brussels, Belgium, 2011, pp. 1917–1920.
[50] R. Rifkin and A. Klautau, *In defense of one-vs-all classification*, J. Mach. Learn. Res., 5 (2004), pp. 101–141.
[51] J. Rissman, H. T. Greely, and A. D. Wagner, *Detecting individual memories through the neural decoding of memory states and past experience*, Proc. Nat. Acad. Sci. USA, 107 (2010), pp. 9849–9854.
[52] S. Ryali, K. Supekar, D. A. Abrams, and V. Menon, *Sparse logistic regression for whole-brain classification of fMRI data*, NeuroImage, 51 (2010), pp. 752–764.
[53] M. Schmidt and K. Murphy, *Convex structure learning in log-linear models: Beyond pairwise potentials*, in Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS), Sardinia, Italy, 2010, pp. 709–716.
[54] B. Schölkopf and A. J. Smola, *Learning with kernels: Support vector machines, regularization, optimization, and beyond*, MIT Press, Cambridge, MA, 2002.
[55] P. Sprechmann, I. Ramirez, P. Cancela, and G. Sapiro, *Collaborative Sources Identification in Mixed Signals via Hierarchical Sparse Modeling*, preprint, http://arxiv.org/abs/1010.4893, 2010.
[56] M. Stojnic, F. Parvaresh, and B. Hassibi, *On the reconstruction of block-sparse signals with an optimal number of measurements*, IEEE Trans. Signal Process., 57 (2009), pp. 3075–3085.
[57] 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., 27 (2006), pp. 678–693.
[58] R. Tibshirani, *Regression shrinkage and selection via the Lasso*, J. Roy. Statist. Soc. Ser. B, 58 (1996), pp. 267–288.
[59] P. Tseng, *Applications of a splitting algorithm to decomposition in convex programming and variational inequalities*, SIAM J. Control Optim., 29 (1991), pp. 119–138.
[60] B. A. Turlach, W. N. Venables, and S. J. Wright, *Simultaneous variable selection*, Technometrics, 47 (2005), pp. 349–363.
[61] K. Ugurbil, L. Toth, and D. Kim, *How accurate is magnetic resonance imaging of brain function?*, Trends Neurosci., 26 (2003), pp. 108–114.
[62] J. H. Ward, *Hierarchical grouping to optimize an objective function*, J. Amer. Statist. Assoc., 58 (1963), pp. 236–244.
[63] S. J. Wright, R. D. Nowak, and M. A. T. Figueiredo, *Sparse reconstruction by separable approximation*, IEEE Trans. Signal Process., 57 (2009), pp. 2479–2493.
[64] O. Yamashita, M. Sato, T. Yoshioka, F. Tong, and Y. Kamitani, *Sparse estimation automatically selects voxels relevant for the decoding of fMRI activity patterns*, NeuroImage, 42 (2008), pp. 1414–1429.
[65] M. Yuan and Y. Lin, *Model selection and estimation in regression with grouped variables*, J. Roy. Statist. Soc. Ser. B, 68 (2006), pp. 49–67.
[66] P. Zhao, G. Rocha, and B. Yu, *The composite absolute penalties family for grouped and hierarchical variable selection*, Ann. Statist., 37 (2009), pp. 3468–3497.
[67] H. Zou and T. Hastie, *Regularization and variable selection via the elastic net*, J. Roy. Statist. Soc. Ser. B, 67 (2005), pp. 301–320.