Region-Network Hierarchical Sparsity Priors for High-Dimensional Inference in Brain Imaging

Danilo Bzdok DANILO.BZDOK@INRIA.FR
Department of Psychiatry, Psychotherapy and Psychosomatics, RWTH Aachen, Germany
Parietal, INRIA, NeuroSpin, bat 145, CEA, 91191 Gif-sur-Yvette France

Michael Eickenberg MICHAEL.EICKENBERG@INRIA.FR
Department d'Informatique, Ecole Normale Superieure, Paris

Gaël Varoquaux GAEL.VAROQUAUX@INRIA.FR
Bertrand Thirion BERTRAND.THIRION@INRIA.FR
Parietal, INRIA, NeuroSpin, bat 145, CEA, 91191 Gif-sur-Yvette France

Abstract

Structured sparsity penalization has recently improved statistical models applied to high-dimensional data in various domains. As an extension to imaging neuroscience, the present work incorporates priors on network hierarchies of brain regions into logistic-regression estimators to distinguish neural activity. This remarries the perspectives of functional segregation and functional integration that are routinely divorced in neuroscientific research. Region-network hierarchical priors for classifying 18 psychological tasks from a reference dataset are shown to outperform naive ℓ_1-norm, elastic-net, and trace-norm penalization, as well as neurobiologically informed ℓ_1/ℓ_2-block-norm group sparsity. Weighing the relative importance of region and network structure within the hierarchical tree penalty recovered complementary aspects of the neural activity patterns. It is thus demonstrated that priors of local and global neurobiological knowledge can enhance generalization performance, sample complexity, and domain interpretability by hierarchical tree sparsity.

1. Introduction

Many quantitative scientific domains underwent a recent passage from the classical regime (i.e., "long data") to the high-dimensional regime (i.e., "wide data"). Also in the brain imaging domain, many contemporary methods for acquiring brain signals yield more variables per observation than total observations per data sample. This $n << p$ scenario challenges various statistical methods from classical statistics. For instance, estimating generalized linear models without additional assumptions yields an underdetermined system of equations. Many such ill-posed estimation problems have benefited from *sparsity* assumptions (Hastie et al., 2015). They act as a regularizer and can be used for model selection. Sparse supervised and unsupervised learning algorithms have proven to yield statistical relationships that can be readily estimated, reproduced, and interpreted (Giraud, 2014). Generally, *structured sparsity* can impose domain knowledge on the statistical estimation, thus shrinking and selecting variables guided by expected data distributions (Bach et al., 2012). Such restrictions to statistical complexity are an attractive plan of attack for the >100,000 variables per brain map. Yet, what generally accepted neurobiological structure lends itself to fruitful exploitation using structured sparsity priors?

Concepts on human brain organization have long been torn between the two extremes *functional specialization* and *functional integration*. Functional specialization emphasizes that microscopically distinguishable brain regions are responsible for distinct classes of computational processes (Kanwisher, 2010). Conversely, functional integration emphasizes that brain function is enabled by a complex interplay between these distinct brain regions (Sporns, 2014). Local infrastructure and unique global connectivity profiles are thought to go hand-in-hand to realize brain function. However, probably no existing brain analysis method acknowledges that both functional design principles are inextricably involved in the realization of mental operations.

Proceedings of the 33^{rd} International Conference on Machine Learning, New York, NY, USA, 2016. JMLR: W&CP volume 48.

Functional specialization has long been explored and interpreted based on different research methods. For instance, single-cell recordings and microscopic examination revealed the segregation of the occipital visual cortex into V1, V2, V3, V3A/B, and V4 regions (Hubel & Wiesel, 1962; Zeki, 1978). Tissue lesion of the mid-fusiform gyrus of the visual system, in turn, was frequently reported to impair recognition of others' identity from faces (Iaria et al., 2008). As a crucial common point, all these methodological approaches yield neuroscientific findings that are naturally interpreted according to non-overlapping, discrete region compartments as the basic architecture of brain organization. It is more recent that the main interpretational focus has shifted from circumscribed regions to network stratifications in systems neuroscience (Yuste, 2015). Besides analyses of electrophysiological oscillations and graph-theoretical properties, studies of functional connectivity correlation (Buckner et al., 2013) and independent component analysis (ICA) (Beckmann et al., 2005) became the workhorses of network discovery in neuroimaging. As a common point of this second set of methods, interpretation of findings naturally embraces cross-regional integration by overlapping network compartments as the basic architecture of brain organization, in stark constrast to methods examining regional specialization.

Building on these two major interpretational traditions in neuroscience, the present study proposes to incorporate established neurobiological structure underlying functional segregation and integration into supervised estimators by hierarchical structured sparsity. Learning algorithms exploiting structured sparsity have recently made much progress in various domains from processing auditory signals (Daudet, 2004), natural images (Harzallah et al., 2009) and videos (Kang, 2015) to astrophysics (Vinci et al., 2014), genetics (Kim et al., 2012), and conformational dynamics of protein complexes (Jenatton et al., 2009). The hierarchical tree penalties recently suggested for imaging neuroscience (Jenatton et al., 2011) will be extended by introducing neurobiologically plausible region and network priors to design neuroscience-specific estimators. Based on the largest neuroimaging repository (Human Connectome Project [HCP]), we demonstrated that domain-informed supervised models gracefully tackle the curse of dimensionality, yield more human-interpretable results, and generalize better to new samples than domain-naive estimators.

2. Methods

The main contribution is the domain adaptation of hierarchical structured tree penalties to jointly incorporate region specialization and network integration priors into supervised estimators. We capitalize on hierarchical group lasso as recently introduced (Jenatton et al., 2011) to create a set of convex penalty terms. These allow acknowledging the interplay between local specialization and global connectivity when discriminating defined psychological tasks from brain activity. Rather than inferring brain activity from psychological tasks in a univariate setting, this study infers psychological tasks from brain activity in a multivariate setting to allow for prediction in unseen neuroimaging data.

Rationale. 3D brain maps obtained by neuroimaging techniques are high-dimensional but, luckily, their signal is highly structured. Its explicit dimensionality, the number of brain voxels, typically exceeds 100,000 variables, while the number of samples rarely exceeds few hundreds. This $n << p$ scenario directly implies underdetermination of any linear model based on dot products with the voxel values. However, the effective dimensionality of functional neuroimaging data has been shown to be much lower (Bzdok et al., 2015). Two types of low-dimensional neighborhoods will be exploited by injecting established knowledge of regional specialization (i.e., region priors) and spatiotemporal interactions (i.e., network priors) into the statistical estimation process.

Large-scale brain networks emerge in human individuals around birth (Doria et al., 2010), before mental capacities mature in childhood. In adults, the sets of regions that compose a given spatiotemporally cohesive network are known to have more similar functional profiles than nodes from different networks (Anderson et al., 2013). As a well-established method for network estimation (Beckmann et al., 2005), ICA yields continuous brain maps with weights for each voxel. The region nodes of ICA networks are spatially disjoint sets of voxel groups that agree with boundaries of brain atlases. Such brain atlases can be obtained by segregating the brain's grey matter into spatially compact voxel groups using clustering algorithmisms (Eickhoff et al., 2015). Consequently, each region from a brain atlas can be uniquely associate to one the extracted ICA networks. In the present work previously published network definitions obtained using ICA (Smith et al., 2009) and previously published region definitions obtained from spatially constrained ward clustering (Craddock et al., 2012) allow constructing a hierarchy of global ICA networks with their assigned local cluster regions. The ensuing network-region tree is used as a structural prior of expected weight distributions during supervised model fitting.

This tree structure can be seamlessly plugged into a recently introduced hierarchical sparsity estimator (Jenatton et al., 2011). It extends the group lasso (Yuan & Lin, 2006) by permitting variable groups that contain each other in a nested tree structure. The first hierarchical level of that tree

are the network groups that contain all the voxels of the brain regions associated to them. Each network node, in turn, descends into a second hierarchical level of brain regions of spatially neighboring voxel groups. As a consequence of such a region-network sparsity tree, a child node enters the set of relevant voxel variables only if its parent node has been selected (Bach et al., 2012). Conversely, if a parent node is deselected, also the voxel variables of all child nodes are deselected. Moreover, the coefficient of all region groups and all network groups can be weighed individually. Trading off the voxel penalties of the network level against the voxel penalties of the region level, we can design different estimation regimes. Setting a low penalty on the network groups makes it probable that all of them are active in the estimated weight map. If we then select higher penalties on region groups, selection of relevant region groups is forced without negligable bias of the network definitions. Conversely, setting low penalties on the region definitions makes it possible for all voxels to be active. Selecting higher penalties on the networks then leads to a selection of networks with all regions associated to it. This region-network tradeoff allows inspecting the relative contributions of the region and network levels of the hierarchical tree prior.

Problem formulation. We formulate our estimation problem in the framework of regularized risk estimation applied to linear models: We would like to estimate a good predictor of cognitive task given a brain image. Let the set $\mathcal{X} \subset \mathbb{R}^p$ represent brain images of $p > 0$ voxels.

Then we would like to minimize the risk $\mathcal{L}(\hat{y}, y)$, where $\hat{y} = X\hat{w} + \hat{b}$, while regularizing to incorporate a useful prior. Taken together, this can be framed as an optimization problem

$$\arg\min_{w,b} \mathcal{L}(Xw + b, y) + \lambda\Omega(w),$$

where $\lambda > 0$ and Ω is the regularizer.

Brain regions are defined as disjoint groups of voxels. Let $\mathcal{G}$ be a partition of $\{1, \ldots, p\}$, i.e.

$$\bigcup_i g_i = \{1, \ldots, p\} \text{ and } g_i \cap g_j = \emptyset \quad \forall i \neq j$$

Brain networks consist of regions and are thus super-regions or groups of regions. The set of brain networks $\mathcal{H}$ is also a partition of $\{1, \ldots, p\}$ and in addition it is consistent with $\mathcal{G}$ in the sense that

$$\text{for all } g \in \mathcal{G}, h \in \mathcal{H}, \quad \text{either } g \subset h \text{ or } g \cap h = \emptyset.$$

This allows a clear association of each region $g \in \mathcal{G}$ to a network $h \in \mathcal{H}$ and thus establishes a tree structure (up to adding a root node containing all voxels).

For a brain image $w \in \mathbb{R}^p$ and a group g, the vector $w_g \in \mathbb{R}^{|g|}$ is defined as the restriction of w to the coordinates in g. The structured penalty incorporating network and region information can then be written as

$$\Omega(w) = \alpha \sum_{h \in \mathcal{H}} \eta_h \|w_h\|_2 + \beta \sum_{g \in \mathcal{G}} \eta_g \|w_g\|_2.$$

According to (Yuan & Lin, 2006) we set $\eta_g = 1/\sqrt{|g|}$ to account for varying group size. The hierarchy-level-specific factors $\alpha > 0$ and $\beta > 0$ are used to trade-off region-weighted and network-weighted models against each other.

One can read off the formula that decreasing α leads to less penalization of brain networks and thus to tendentially fully active groups and dense brain maps. If at the same time β is increased to a point of inducing group sparsity, then only the structure of brain regions encoded by $\mathcal{G}$ is present. Conversely, if β is decreased to become non-selective, and α is increased to group selectivity, then the structure imposed will come from $\mathcal{H}$, leading to the selection of brain networks rather than regions.

It is important to note that the above trade-off shows that predominance can be either attributed to regions or networks, but that nevertheless the penalty structure is hierarchical: If the network penalty layer sets a network group to zero, then all the contained region groups are forced to have activity zero. Conversely, if a brain region has non-zero coefficients, then necessarily the network containing it must be active. This relation is asymmetric, since when swapping the roles of $\mathcal{G}$ and $\mathcal{H}$, these statements do not hold true anymore. A brain region can set all its coefficients to zero without forcing the corresponding network to zero. A brain network can be active without its subregions necessarily being active. When evaluating the trade-off in (α, β), this needs to be taken into account.

The prediction problem at hand is a multiclass classification. We choose to attack this using one-vs-rest scheme on a binary logistic regression, whose loss can be written as

$$\sum_{i=1}^{n} \log(1 + \exp(-y_i \langle x_i, w \rangle)) + \lambda\Omega(w),$$

if $y \in \{-1, 1\}$ and with $x_i \in \mathbb{R}^p$ the training sample brain images.

Optimization. We optimize the data loss function using an iterative forward-backward scheme analogous to (F)ISTA for the lasso (**?**). Michael: Algorithm specification to be removed for ICML - supposed known, except possibly for hierarchical prox. We divide the loss $\mathcal{L}(w)$ into a smooth, convex data fit term $F(w)$ (here: logistic regression loss) and a potentially non-smooth, convex regularization term $G(w)$. Define $\text{prox}_{\gamma G}(v) = (\text{Id} + \gamma\partial G)^{-1}(v)$,

where ∂G is the subdifferential of G and let $\nabla F(v)$ be the gradient of F in v. Further let $L > 0$ be a Lipschitz constant of the gradient of F and choose a step size $\varepsilon \in (0, \frac{2}{L})$. Starting with an arbitrary initialization of w_0, putting $z_1 = w_0$ and $t_1 = 1$, then for $k \geq 1$ until convergence do:

$$\begin{aligned} x_k &= \mathrm{prox}_{\varepsilon G}(y_k - \varepsilon \nabla F(y_k)) \\ t_{k+1} &= \frac{1 + \sqrt{1 + 4t_k^2}}{2} \\ y_{k+1} &= x_k + \left(\frac{t_k - 1}{t_{k+1}}\right)(x_k - x_{k-1}) \end{aligned}$$

In our case $G(v) = \lambda\Omega(v)$. Splitting $\Omega(v) = \Omega_{\mathcal{H}}(v) + \Omega_{\mathcal{G}}(v)$ with $\Omega_{\mathcal{H}}(v) = \alpha \sum_{h \in \mathcal{H}} \eta_h \|v_h\|_2$ and $\Omega_{\mathcal{G}}(v) = \beta \sum_{g \in \mathcal{G}} \eta_g \|v_g\|_2$, we observe that

$$\mathrm{prox}_{\varepsilon\lambda\Omega_{\mathcal{H}}}(v) = \sum_{h \in \mathcal{H}} \iota_h \left([\|v_h\|_2 > \eta_h \varepsilon \lambda \alpha] \left(v_h - \eta_h \varepsilon \lambda \alpha \frac{v_h}{\|v_h\|_2} \right) \right)$$

where ι_h is the canonical injection of the group h into $\mathbb{R}^p$ (if $\pi_g : v \mapsto v_g$, then $\iota_g = \pi_g^*$ the adjoint/transposed operator). This holds analogously for $\Omega_{\mathcal{G}}$. According to (**?**), the proximal operator of the hierarchical penalty $\varepsilon\lambda\Omega$ is

$$\mathrm{prox}_{\varepsilon\lambda\Omega}(v) = \mathrm{prox}_{\varepsilon\lambda\Omega_{\mathcal{H}}} \circ \mathrm{prox}_{\varepsilon\lambda\Omega_{\mathcal{G}}}(v).$$

Hyperparameter optimization. The stratified and shuffled training data were submitted to a nested cross-validation scheme for model selection and model assessment. In the inner CV layer, the logistic regression estimators have been trained in a one-versus-rest design that distinguishes each class from the respective 17 other classes (number of maximal iterations=100, tolerance=0.001). In the outer CV layer, grid search selected among candidates for the respective λ parameter by searching between 10^{-2} and 10^{1} in 9 steps on a logarithmic scale. Importantly, the thus selected sparse logistic regression classifier was evaluated on an identical test set in all settings.

Implementation. All analyses were performed in Python. We used *nilearn* to process and resphape the extensive neuroimaging data (Abraham et al., 2014), *scikit learn* to design machine-learning data processing pipelines (Pedregosa et al., 2011), and *SPAMs* for numerically optimized implementations of the sparse estimators (`http://spams-devel.gforge.inria.fr/`). All Python scripts that generated the results are accessible online for reproducibility and reuse (`http://github.com/anonymous/anonymous`).

Data. As the currently biggest openly-accessible reference dataset, we chose resources from the Human Connectome Project (HCP) (Barch et al., 2013). Neuroimaging task data with labels of ongoing cognitive processes were drawn from 500 healthy HCP participants (cf. Appendix for details on datasets). 18 HCP tasks were selected that are known to elicit reliable neural activity across participants. In sum, the HCP task data incorporated 8650 first-level activity maps from 18 diverse paradigms administered to 498 participants (2 removed due to incomplete data). All maps were resampled to a common $60 \times 72 \times 60$ space of 3mm isotropic voxels and gray-matter masked (at least 10% tissue probability). The supervised analyses were thus based on labeled HCP task maps with 79,941 voxels of interest representing z-values in gray matter.

3. Experimental Results

Benchmarking hierarchical tree sparsity against common sparsity penalties. Hierarchical region-network priors have been systematically evaluated against other popular choices of sparse classification algorithms in an 18-class scenario (Figure 2). Logistic regression with ℓ_1/ℓ_2 block norm penalization incorporated a hierarchy of previously known region and network neighborhoods for a neurobiological bias of the statistical estimation. Vanilla logistic regression with ℓ_1-penalization does not assume any previously known special structure. This classification estimator embraces a vision of neural activity structure that expects a minimum of topographically and functionally independent brain voxel to be relevant. Logistic regression with (sparse) group sparsity imposes a structured ℓ_1/ℓ_2 block norm (with additional ℓ_1 term) with a known atlas of region voxel groups onto the statistical estimation process. This supervised estimator shrinks and selects the coefficients of topographically compact voxel groups expected to be relevant together. Logistic regression with trace-norm penalization imposed low-rank structure. This supervised classification algorithm expected a minimum of unknown network patterns to be relevant.

Across analyses, hierarchial tree sparsity was most successful in distinguishing unseen neural activity maps from 18 psychological tasks (89.7% accuracy, mean precision 86.7, mean recall 91.5). It was closely followed by logistic regression structured by trace-norm regularization (89.4%, precision 86.4, recall 91.3). Lasso featured an average performance comparing to the other sparse estimators (88.60%, precision 85.7, recall 90.3). Elastic-Net, in turn, featured an average performance comparing to the other sparse estimators (88.1%, precision 84.8, recall 84.1). Introducing a priori knowledge of brain region compartments by sparse group sparsity (87.9%, precision 84.8, recall 89.5) and by group sparsity (87.9%, precision 85.3, recall 90.3) performed worst. In an important subanalysis, the gain of the combined region-network prior was also confirmed by selectively zeroing the η_g coefficients

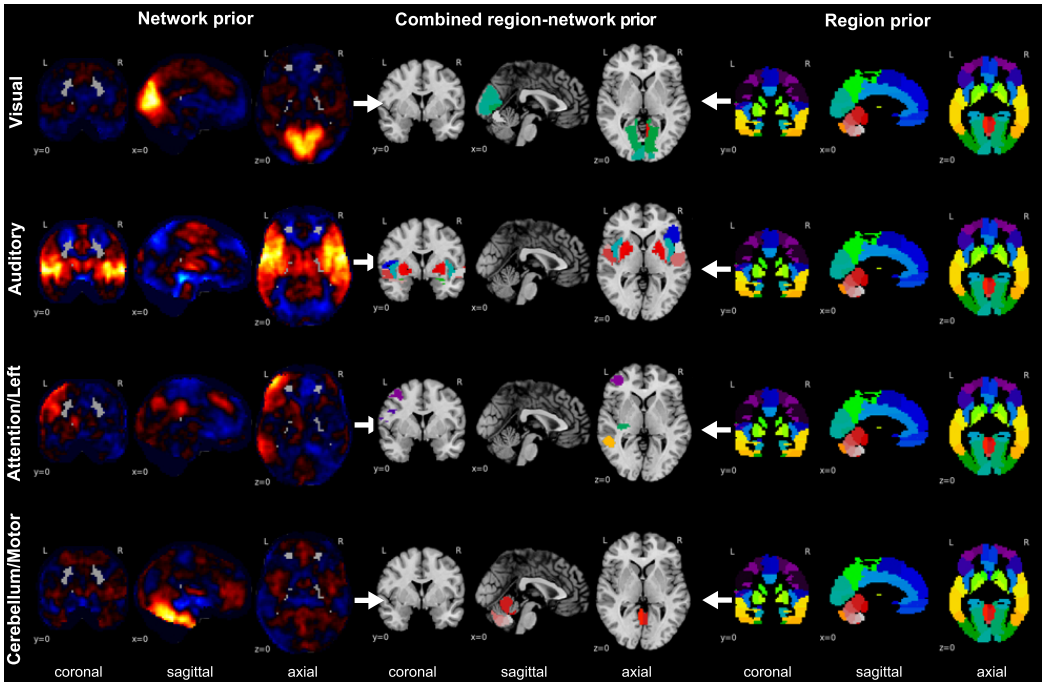

Figure 1. **Building blocks of the region-network tree.** Depicts neurobiological priors introduced into the classification problem by hierarchical structured sparsity. *Left:* Continuous, partially overlapping brain network priors (*hot-colored*, taken from (Smith et al., 2009)) accommodate the functional integration perspective of brain organization. *Right:* Discrete, non-overlapping brain region priors (*single-colored*, taken from (Craddock et al., 2012)) accommodate the functional segregation perspective. *Middle:* These two types of predefined voxel groups are incorporated into hierarchical priors of parent networks with their descending region nodes. *Top to bottom:* Four examplary region-network priors are shown, including the early cortex that processes visual and sound information from the environment, a well-known attentional circuit in the left brain hemisphere, and the cerebellum that is involved in motor behavior.

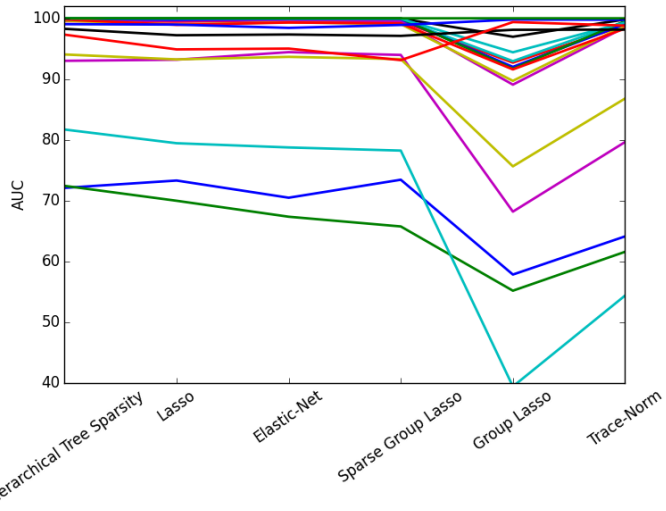

Figure 2. **Model performance across sparsity priors.** Compares the performance of logistic regression estimators with 4 different structured and unstructered sparsity terms in classifying neural activity from 18 psychological tasks. The class-wise area under the curve (AUC) is obtained on the same test set. *Hierarchical Tree Sparsity*: Structured ℓ_1/ℓ_2-block-norm with a hierarchy of both region and network priors exhibited the best out-of-sample performance. *Lasso*: Unstructured ℓ_1-penalized logistic regression imposed a minimum of relevant brain voxels without assuming special structure. *Elastic-Net*: Unstructured logistic regression with interpolation between ℓ_1- and ℓ_2-norm imposed an equilibrium between sparsity and model fit. *Trace-norm*: Structured trace-norm penalization imposed low-rank structure with sparsity of network patterns, but naive to region structure. *(Sparse) Group Sparsity*: Structured ℓ_1/ℓ_2-block norm with additional ℓ_1 term imposed region compartments, but naive to network structure. A priori knowledge of both region and network neighborhoods was hence most beneficial for predicting psychological tasks from brain maps.

of all region groups or all network groups in the hierarchical prior. Removing region structure from the prior achieved 88,84% accuracy, while removing network structure from the prior achieved 87,05% accuracy. These results from partial priors are indeed outperformed the full region-network tree prior at 89,67% accuracy. In sum, biasing sparse model selection by domain knowledge of region-network hierarchies outcompeted other types of frequently used sparse penalization techniques.

Sample complexity of naive versus informed sparse model selection. Subsequently, the sample complexity of ℓ_1-penalized and hierarchical-tree-penalized logistic regression were quantitatively compared (Figure 3). Region-network priors should bias model selection towards more neurobiologically plausible classification estimators. This should yield better out-of-sample generalization and support recovery than neurobiology-naive ℓ_1-constrained logistic regression in the data-scarce and data-rich scenarios. The HCP task data with examples from 18 psychological tasks were first divided into 90% of training set (i.e., 7584 neural activity maps) and 10% of test set (i.e., 842 neural activity maps). Both learning algorithms were fitted based on the training set at different subsampling fractions: 20% (1516 neural activity maps), 40% (3033 maps), 60% (4550 maps), 80% (6067 maps), and 100% (7584 maps).

Regarding classification performance on the always same test set, ℓ_1-penalized versus hierarchical-tree-penalized logistic regression achieved 83.6% versus 88.7% (20% of training data), 85.0% versus 89.2% (40%), 86.8% versus 89.8% (60%), 88.9% versus 90.3% (80%), 88.6% versus 89.7% (100%) accuracy. Regarding model sparsity, the measure $s = \frac{||w||_1}{||w||_F}$ was computed from the model weights w of both penalized estimators for each of the 18 classes. The ℓ_1-penalized logistic regression yielded the mean (+/- standard deviation) sparsities 50.0 (+/- 2.6), 45.4 (+/- 2.3), 40.0 (+/- 2.4), 30.9 (+/- 2.0), and 24.0 (+/- 2.3) after model fitting with 20% to 100% training data. The hierarchical-tree-penalized logistic regression yield the sparsities 163.2 (+/- 0.7), 160.2 (+/- 1.8), 132.1 (+/- 3.3), 116.2 (+/- 4.4), and 88.4 (+/- 8.4) after fitting 20% to 100% of the training data. Finally, the support recovery was quantified by Pearson correlation ρ between vectors of the zscored model coefficients and the zscored sample maps for each class. ℓ_1-penalized versus hierarchical-tree-penalized logistic regression achieved a mean ρ (+/- standard deviation) recovery of 0.10 (+/- 0.03) versus 0.13 (+/- 0.04), 0.11 (+/- 0.03) versus 0.13 (+/- 0.05), 0.13 (+/- 0.04) versus 0.17 (+/- 0.05), 0.16 (+/- 0.04) versus 0.22 (+/- 0.05), and 0.19 (+/- 0.05) versus 0.29 (+/- 0.06) based on 20% to 100% training data.

Three observations have been made. In the data-scarce scenario (i.e., 1/5 of available training data), hierarchical tree sparsity achieved the biggest advantage in out-of-sample performance by 5,11% as well as better support recovery with weight maps already much closer to the class averages (Varoquaux et al., 2012). In the case of scarce training data, which is typical for the brain imaging domain, regularization by region-network priors thus allowed for more effective extraction of classification-relevant structure from the neural activity maps. Across training data fractions, the weight maps from ordinary logistic regression exhibited higher variance and more zero coefficients than hierarchical tree logistic regression. Given the usually high multicollinearity in neuroimaging data, this observation is likely to reflect instable selection of representatives among class-responsive predictor groups due to the ℓ_1-norm penalization. In the data-rich scenario (i.e., entire training data used for model fitting), neurobiologically informed logistic regression profited more from the increased information quantities than neurobiologically naive logistic regression. That is, the region-network priors actually further enhance the similarity to the weight maps even in abundant input data. This was the case although the maximal classification performance of ≈90% has already been reached with small training data fractions by the structured estimator. In contrast, the unstructured estimator reached this generalization performance only with bigger input data quantities.

Support recovery as a function of region-network emphasis. Finally, the relative importance of the region and network priors within the hierarchical tree prior was quantified (Figure 4). The group weight η_g of region priors was multiplied with a region-network ratio, while the group weight η_g of network priors was divided by that region-network ratio. For instance, a region-network ratio of 3 increased the relative importance of known region structure by multiplying $\frac{3}{1}$ to η_g of all region group penalties and multiplying $\frac{1}{3}$ to η_g of all network group penalties (Table 1).

As the most important observation, a range between region-dominant and network-dominant structured penalties yielded quantitatively almost identical generalization to new data but qualitatively different decision functions manifested in the weight maps (Figure 4, second and forth column). Classification models with many zero coefficients but high absolute coefficients in either region compartments or network compartments can similarly extrapolate to unseen neural activity maps. Second, these achieve classification performance comparable to equilibrated region-network priors that set less voxel coefficients to zero and spread the probability mass with lower absolute coefficients across the whole brain (Figure 4, third column in the middle). Third, overly strong emphasis on either level of the hierarchical prior can yield the neurobiologically informative results with maps of the most necessary region or

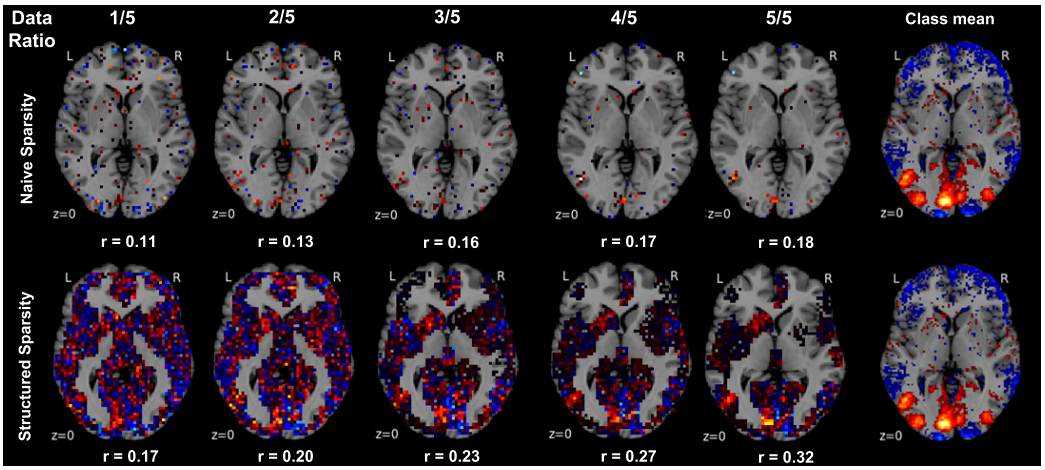

Figure 3. **Naive versus informed sparse model selection across training set sizes.** Ordinary ℓ_1-penalized logistic regression (*upper row*) is compared to hierarchical-tree-penalized logistic regression (*lower row*) with increasing fraction of the available training data (*left to right columns*). For one example (i.e., "View tools") from 18 psychological tasks, unthresholded axial maps of model weights are shown for comparison against the sample average of that class (*rightmost column*, thresholded at the 75^{th} percentile). The out-of-sample accuracies for predicting all 18 psychological tasks is given in percent. In the data-scarce scenario, typical for brain imaging, hierarchical tree sparsity achieves much better support recovery with the biggest difference in model performance. In the data-rich scenario, neurobiologically informed logistic regression profits more from the available information quantities than neurobiologically naive logistic regression.

Table 1. Out-of-sample performance by region-network emphasis

Reg-Net Ratio	500	100	50	10	5	2	1	$\frac{1}{2}$	$\frac{1}{5}$	$\frac{1}{10}$	$\frac{1}{50}$	$\frac{1}{100}$	$\frac{1}{250}$	$\frac{1}{500}$	$\frac{1}{1000}$
Accuracy [%]	3.8	41.7	63.7	89.9	89.9	88.5	87.9	87.9	87.9	87.8	88.6	87.8	85.6	67.2	39,3

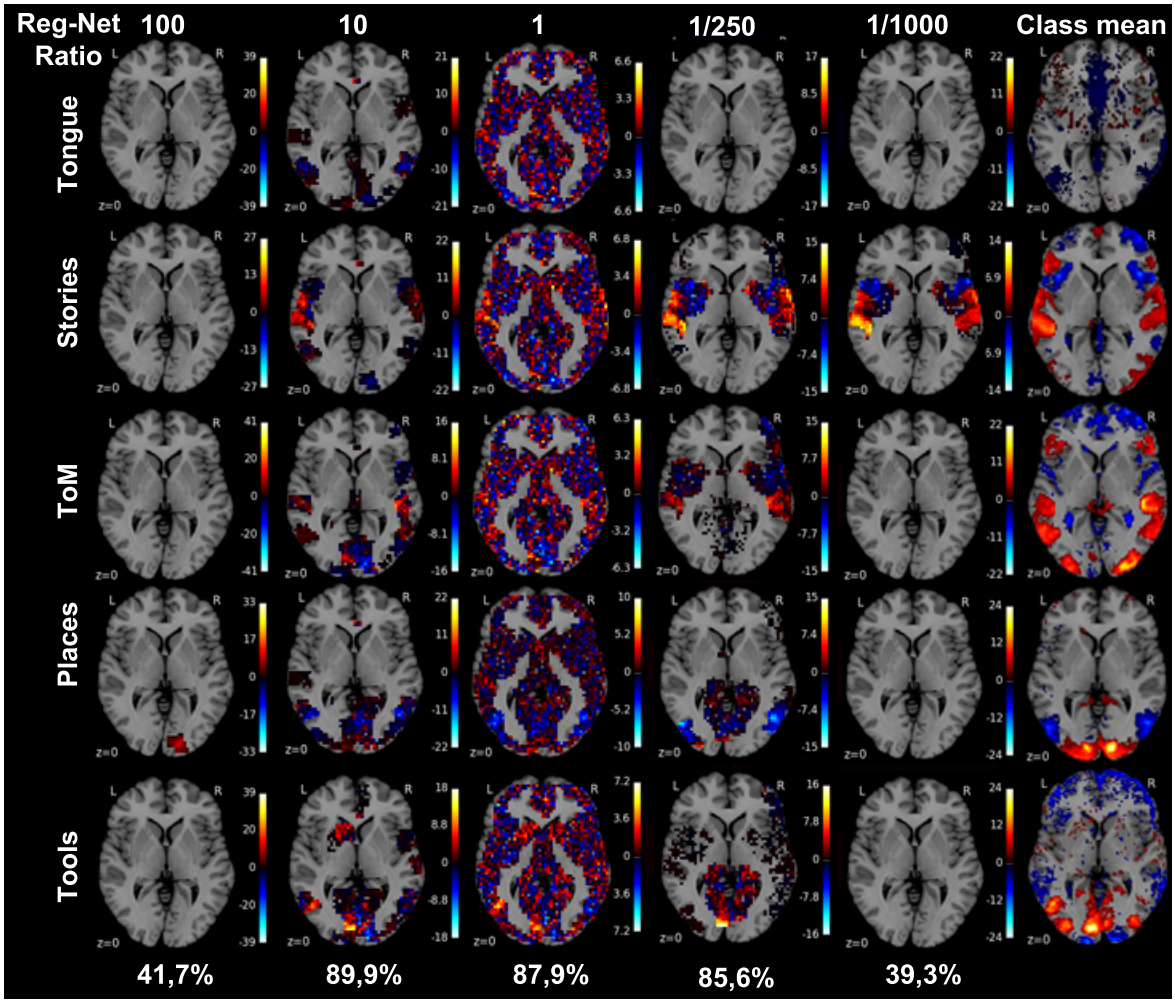

Figure 4. **Support recovery as a function of region and network emphasis.** The relative impact of the region and network priors on model selection is systematically varied against each other. This region-network ratio (*upper fractions*) weighted voxel groups to priviledge sparse models in function space that acknowledge known brain region neighborhoods (*left columns*) or known brain networks neighborhoods (*right columns*). Among the 18 classes, the model weights are shown for the psychological tasks (*from top to bottom*): tongue movement, listening stories, taking somebody else's perspective (ToM, "theory of mind"), as well as viewing locations and tools. The 18-class out-of-sample accuracy is indicated on the *bottom* and the class-wise mean neural activty (*rightmost column*, thresholded at the 75^{th} percentile). Different emphasis on regions versus networks in hierarchical structured sparsity can yield comparable model performance. Favoring region versus network structure during model selection recovers complementary aspects of the neural activity pattern. Equal region and network emphasis yields more dispersed, less interpretable predictive model choices.

network structure for statistically significant out-of-sample performance (Figure 4, leftmost and rightmost columns). In sum, stratifying the hierarchical tree penalty between region and network emphasis suggests that *class-specific region-network weights* might offer more performant and more interpretable classification models in the future.

4. Conclusion

Relevant structure in neuroimaging data has long been investigated according to two separate organizational principles: functional segregation into discrete brain regions (Passingham et al., 2002) and functional integration by interregional brain networks (Sporns, 2014). Both organizational principles are however inextricable because a specialized brain region communicates input and output with other regions and a brain network subserves complex function by orchestrating its region nodes. This suggests hierarchical statistical models as an underexploited opportunity for imaging imaging analysis. This proof-of-concept study demonstrates the simultaneous exploitation of both these neurobiological compartments for sparse variable selection and high-dimensional prediction in a reference dataset. Introducing existing domain knowledge into model selection allowed privileging members of the function space that are most neurobiologically plausible. This statistically and neurobiologically desirable bias is shown to enhance both model interpretability and generalization performance, although these statistical-learning goals are typically in conflict.

The present approach has important advantages over previous analysis strategies that rely on dimensionality reduction of the neuroimaging data to harness the curse of dimensionality. They often use preliminary pooling functions within regions or regression against network templates for subsequent supervised learning on the aggregated feature space. Such lossy approaches divided into feature engineering and inference steps *i)* can only satisfy the specialization or integration account of brain organization, *ii)* depend on the ground truth being a region or network effect, and *iii)* cannot issue individual coefficients for every brain voxels. Hierarchical region-network sparsity addresses these shortcomings by estimating individual voxel contributions while benefitting from their functional segregation and integration to restrict statistical complexity. Viewed from the bias-variance tradeoff, our modification to logistic regression estimators entailed a large decrease in model variance but only a modest increase in model bias. Viewed from the Vapnik-Chervonenkis dimensions, this entailed a healthy decrease in the complexity capacity of the prediction model with a higher chance of generalizing to unobserved data.

In the future, region-network sparsity priors could be incorporated into various pattern-learning methods in systems neuroscience. This includes supervised methods for whole-brain classification and regression with one or several target variables. The principled regularization scheme could even inform unsupervised structure-discovery methods, such as principal component analysis (Jenatton et al., 2009) and k-means clustering (Witten & Tibshirani, 2010). Additionally, model regularization by hierarchical structured sparsity could be extended from the spatial domain of neural activity to priors of coherent spatiotemporal activity structure (Gramfort et al., 2011). Ultimately, successful high-dimensional inference is an important prerequisite for predicting diagnosis, disease trajectories, and treatment response in personalized psychiatry and neurology.

Acknowledgment. Data were provided by the Human Connectome Project.

References

Abraham, A., Pedregosa, F., Eickenberg, M., Gervais, P., Mueller, A., Kossaifi, J., Gramfort, A., Thirion, B., and Varoquaux, G. Machine learning for neuroimaging with scikit-learn. *Front Neuroinform*, 8:14, 2014.

Anderson, M. L., Kinnison, J., and Pessoa, L. Describing functional diversity of brain regions and brain networks. *Neuroimage*, 73:5058, 2013.

Bach, Francis, Jenatton, Rodolphe, Mairal, Julien, and Obozinski, Guillaume. Optimization with sparsity-inducing penalties. *Foundations and Trends® in Machine Learning*, 4(1):1–106, 2012.

Barch, D. M., Burgess, G. C., Harms, M. P., Petersen, S. E., Schlaggar, B. L., Corbetta, M., Glasser, M. F., Curtiss, S., Dixit, S., and Feldt, C. Function in the human connectome: task-fmri and individual differences in behavior. *Neuroimage*, 80:169–189, 2013.

Beckmann, C. F., DeLuca, M., Devlin, J. T., and Smith, S. M. Investigations into resting-state connectivity using independent component analysis. *Philos Trans R Soc Lond B Biol Sci*, 360 (1457):1001–13, 2005.

Buckner, Randy L, Krienen, Fenna M, and Yeo, BT Thomas. Opportunities and limitations of intrinsic functional connectivity mri. *Nature neuroscience*, 16(7):832–837, 2013.

Bzdok, Danilo, Eickenberg, Michael, Grisel, Olivier, Thirion, Bertrand, and Varoquaux, Gaël. Semi-supervised factored logistic regression for high-dimensional neuroimaging data. In *Advances in Neural Information Processing Systems*, pp. 3330–3338, 2015.

Craddock, R. C., James, G. A., Holtzheimer, P. E., 3rd, Hu, X. P., and Mayberg, H. S. A whole brain fmri atlas generated via spatially constrained spectral clustering. *Hum Brain Mapp*, 33 (8):1914–28, 2012.

Daudet, Laurent. Sparse and structured decompositions of audio signals in overcomplete spaces. In *In International Conference on Digital Audio Eects*, 2004.

Doria, V., Beckmann, C. F., Arichia, T., Merchanta, N., Groppoa, M., Turkheimerb, F.E., Counsella, S.J., Murgasovad, M., Aljabard, P., Nunesa, R.G., Larkmana, D.J., Reese, G., and Edwards, A. D. Emergence of resting state networks in the preterm human brain. *Proc Natl Acad Sci U S A*, 107(46): 20015–20020, 2010.

Eickhoff, S. B., Thirion, B., Varoquaux, G., and Bzdok, D. Connectivity-based parcellation: Critique and implications. *Hum Brain Mapp*, 2015.

Giraud, C. *Introduction to High-Dimensional Statistics*. CRC Press, 2014.

Gramfort, Alexandre, Papadopoulo, Theodore, Baillet, Sylvain, and Clerc, Maureen. Tracking cortical activity from m/eeg using graph cuts with spatiotemporal constraints. *NeuroImage*, 54(3):1930–1941, 2011.

Harzallah, Hedi, Jurie, Frédéric, and Schmid, Cordelia. Combining efficient object localization and image classification. In *Computer Vision, 2009 IEEE 12th International Conference on*, pp. 237–244. IEEE, 2009.

Hastie, T., Tibshirani, R., and Wainwright, M. *Statistical Learning with Sparsity: The Lasso and Generalizations*. CRC Press, 2015.

Hubel, David H and Wiesel, Torsten N. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The Journal of physiology*, 160(1):106, 1962.

Iaria, G., Fox, C. J., Waite, C. T., Aharon, I., and Barton, J. J. The contribution of the fusiform gyrus and superior temporal sulcus in processing facial attractiveness: neuropsychological and neuroimaging evidence. *Neuroscience*, 155(2):409–22, 2008.

Jenatton, Rodolphe, Obozinski, Guillaume, and Bach, Francis. Structured sparse principal component analysis. *arXiv preprint arXiv:0909.1440*, 2009.

Jenatton, Rodolphe, Gramfort, Alexandre, Michel, Vincent, Obozinski, Guillaume, Bach, Francis, and Thirion, Bertrand. Multi-scale mining of fmri data with hierarchical structured sparsity. In *Pattern Recognition in NeuroImaging (PRNI), 2011 International Workshop on*, pp. 69–72. IEEE, 2011.

Kang, J-W. Structured sparse representation of residue in screen content video coding. *Electronics Letters*, 51(23):1871–1873, 2015.

Kanwisher, Nancy. Functional specificity in the human brain: a window into the functional architecture of the mind. *Proceedings of the National Academy of Sciences*, 107(25):11163–11170, 2010.

Kim, Seyoung, Xing, Eric P, et al. Tree-guided group lasso for multi-response regression with structured sparsity, with an application to eqtl mapping. *The Annals of Applied Statistics*, 6 (3):1095–1117, 2012.

Passingham, R. E., Stephan, K. E., and Kotter, R. The anatomical basis of functional localization in the cortex. *Nat Rev Neurosci*, 3(8):606–16, 2002.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. Scikit-learn: Machine learning in python. *J Mach Learn Res*, 12:2825–2830, 2011.

Smith, S. M., Fox, P. T., Miller, K. L., Glahn, D. C., Fox, P. M., Mackay, C. E., Filippini, N., Watkins, K. E., Toro, R., Laird, A. R., and Beckmann, C. F. Correspondence of the brain's functional architecture during activation and rest. *Proc Natl Acad Sci U S A*, 106(31):13040–5, 2009.

Sporns, O. Contributions and challenges for network models in cognitive neuroscience. *Nat Neurosci*, 17(5):652–60, 2014.

Varoquaux, Gaël, Gramfort, Alexandre, and Thirion, Bertrand. Small-sample brain mapping: sparse recovery on spatially correlated designs with randomization and clustering. *arXiv preprint arXiv:1206.6447*, 2012.

Vinci, Giuseppe, Freeman, Peter, Newman, Jeffrey, Wasserman, Larry, and Genovese, Christopher. Estimating the distribution of galaxy morphologies on a continuous space. *arXiv preprint arXiv:1406.7536*, 2014.

Witten, Daniela M and Tibshirani, Robert. A framework for feature selection in clustering. *Journal of the American Statistical Association*, 105(490), 2010.

Yuan, Ming and Lin, Yi. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67, 2006.

Yuste, Rafael. From the neuron doctrine to neural networks. *Nat Rev Neurosci*, 16(8):487–497, 2015.

Zeki, Semir M. Functional specialisation in the visual cortex of the rhesus monkey. *Nature*, 274(5670):423–428, 1978.