Region-network hierarchical sparsity priors for high-dimensional inference in brain imaging

Danilo Bzdok, Michael Eickenberg, Gaël Varoquaux, Bertrand Thirion
Department of Psychiatry, Psychotherapy and Psychosomatics, RWTH Aachen, Germany
INRIA, Parietal team, Saclay, France
CEA, Neurospin, Gif-sur-Yvette, France
firstname.lastname@inria.fr

Abstract

keywords: Sparsity-inducing norms, hierarchical structured sparsity, numerical optimization, systems neuroscience, brain imaging, functional specialization, functional integration

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") [32]. 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 high-dimensional 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 [11, 24]. 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 [21]. Generally, *structured sparsity* can impose domain knowledge on the statistical estimation, thus shrinking and selecting variables guided by expected data distributions [3]. Such restrictions to complexity are an attractive plan of attacked for the $>$100,000 variables of brain maps. Yet, what generally accepted neurobiological structure lends itself to harness the *curse of dimensionality* by 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 solve distinct classes of computational processes [34]. Functional integration, in turn, emphasizes that brain function is enabled by complex connections between these distinct brain regions [48]. These notions were predominantly derived from invasive examination of anatomy (i.e., histological preparation), connectivity (i.e., axonal tracing), and functional properties (i.e., single-cell recordings) in animals. Regarding functional segregation into specialized regions, early histological investigations into the microscopic heterogeneity of the human cerebral cortex have resulted in several detailed anatomical maps [9, 52]. Regarding axonal connections, each such cortical area has been observed to possess a unique set of incoming and outgoing connections [39, 54, 44]. Both local infrastructure and its unique global connectivity profile together are thought to realize brain function. In sum, cortical brain modules versus connections between them reflect functional specialization versus functional integration [20, 37]. Importantly, probably no existing brain analysis method acknowledges that both functional organizations are inextricably involved in the realization of mental operations [50, 43].

Functional specialization has been explored and interpreted based on many different research methods. Single-cell recordings and microscopic examination revealed, for instance, the specialization in the visual cortex into V1, V2, V3, V3A, and V4 [26, 56]. Tissue lesion of the mid-fusiform gyrus of the visual system, in turn, was frequently reported to impair recognition of others' identity from faces [27]. The whole-brain localization of sensory, motor, and emotional functions to cortical areas was later enabled by non-invasive brain imaging with functional magnetic resonance imaging (fMRI) and positron emission tomography (PET) [19]. Further, radioactive mapping of neurotransmitter receptors rendered accessible yet another local characteristic of neuronal populations [58]. In the computational era, automatic clustering methods are increasingly employed to regionally differentiate the cerebral cortex, which can partly be more fine-grained than classical microscopical borders [7, 17]. High-thoughput approaches today enable ultrahigh-resolution 3D models of brain anatomy at near-cellular scale [2]. 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 [55, 49]. Invasive axonal tracing studies in monkeys were complemented by diffusion MRI tractography in humans as a now frequently employed method to outline fiber bundles between brain regions [28]. Besides analyses of electrophysiological oscillations [13] and graph-theoretical properties [12], studies of functional connectivity [10] and independent component analysis (ICA) [6] became the workhorses of network discovery in neuroimaging. These revealed the important implication of canonical brain networks across psychological tasks, including the so-called "default-mode network" [41], "salience network" [45], and "dorsal attention network" [14]. Characteristic changes in the configuration of these macroscopical networks were repeatedly observed to be induced by the onset of given psychological tasks [18]. As a common point of all these 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 streams in systems neuroscience, the present study proposes to incorporate established neurobiological structure underlying functional segregation and integration into supervised estimators by hierarchical structured sparsity. Learning techniques exploiting structured sparsity have recently made much progress in various application domains from processing of auditory signals [16], natural images [23] and videos [33, 36] to astrophysics [51], genetics [42, 35], and conformational dynamics of protein complexes [31]. This is extended by the present work that introduced neuroscience-specific estimators capitalizing on neurobiologically plausible region and network priors. Based on the largest neuroimaging repository, 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-naïve estimators.

2 Methods

Rationale we need to inject domain knowledge into statistical estimations to harness the curse of dimensionality. two neurobiological design principles

imposing parsimony integrative processes

This L1/L2 norm for group lasso has been extended to a more general setting to

designed groups the child nodes enter the set of relevant inputs only if its parent node does.

should be able to estimate voxel level while taking into account known supravoxel structure. is instrumental in Developmentally, such large-scale networks emerge during late fetal growth (Doria et al., 2010), before cognitive capacities mature in childhood.

In adults, nodes of a same cohesive network have more similar functional profiles than nodes from different networks (Anderson et al., 2013).

data exhibit natural correlations between neighboring voxels forming clusters

representing some phenomenon with as few variables as possible

neurobiologically motivated restrictions to complexity circumvented the curse of dimensionality three-dimensional spatial arrangement that respects the functional anatomy of the brain not ignore the spatial configuration

incorporate rich prior knowledge

If meaningful structures exist, we show that one can take advantage of such structures

Statistically, l1 and l2 are local sparsity priors -¿ resulting sparsity does yield structure we want to priviledge representations with structure

=¿ a biologically and statistically desirable bias

Problem formulation Sparse linear models encode geometric prior information topology local sets of voxels

Group-sparsity is a first step towards the more general idea that a regularization function can encourage sparse solutions with a particular structure.

it is not realistic to assume that all of the tasks share the same set of rele- vant inputs as in the L1/L2-regularized regression. A subset of highly related outputs may share a common set of relevant inputs, whereas weakly related outputs are less likely to be affected by the same inputs.

structured regularization We might therefore gain in the quality of the factors induced by enforcing directly this a priori

groups at multiple granularity

tree-guided group lasso

encourage structured shrinkage effect

l1 = unstructured sparsity-inducing penalty

Our method extends the L1/L2 penalty to the tree-lasso penalty by letting the hierarchically-defined groups overlap. the tree lasso is a special case of overlapping group lasso

for every column u of U, it compute a column v of V solving

we aim at learning a weight vector w Rp and an intercept b R such that the prediction of y can be based on the value of wŁx + b.

We omit a bias term, since the data were mean-centered and unit-variance scaled. The scalar b is not particularly informative

however the vector w corresponds to a volume that can be represented in brain space as a volume

hierarchial tree = more generally into a directed acyclic graph

more precisely, we denote by X Rnp the design matrix assembled from n fMRI volumes and by y Rn the corresponding n targets. In other words, each row of X is a p-dimensional sample, i.e., an activation map of p voxels related to one stimulus presentation. for visualization of the predictive pattern of voxels.

Learning the parameters (w, b) remains challenging since the number of features (104 to 105 voxels) exceeds by far the number of samples (a few hundreds of volumes).

The scalar b is not particularly informative, however the vector w corresponds to a volume that can be represented in brain space as a volume for visualization of the predictive pattern of voxels.

each row of X is a p-dimensional sample, i.e., an activation map of p voxels related to one stimulus presentation.

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 for the problem at hand. -¿ we do not want to do preliminary feature selection or dimensionality reduction or feature agglomeration because we want to fit one model parameter to each brain voxel for maximal interpretability This corresponds to discarding some columns of X.

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

Craddock clusters are often used for feature agglomeration into parcels -¿ exploits only a part of the data

dual-level spatial structure sparse hierarchical regularization structured sparsity-inducing regularization the root of the tree T is the unique cluster that gathers all the voxels,

It is a generalization of the traditional ℓ_1-norm $\Omega(\mathbf{w}) = \sum_{j=1}^{p} |\mathbf{w}_j|$ ignores structure

[30]

structured sparsity

[25, 38, 29]

a node j of $\mathcal{T}$, we denote by $g_j \subseteq \{1, .., q\}$ the set of indices that record all the descendants of j in $\mathcal{T}$

the family of sparsity-inducing norms has recently been extended by hierarchical sparsity penalty terms [57].

$$\Omega(\mathbf{w}) = \sum_{g \in G} ||\mathbf{w}_g||_2 = \sum_{g \in G} \sqrt{\sum_{j \in g} \mathbf{w}_j^2} \tag{1}$$

For example, when G is the set of all singletons, is the usual l1 norm (assuming that all the weights are equal to 1).

l1/l2 mixed norm is convex

Discarding coefficients belonging to a network group will naturally enforce discarding the coefficients belonging to each of its descendent region groups. Conversely, variable selection of a network group will also enforce selection of all voxel of its descendent group regions. Single region groups can however be set to zero (unselected) or non-zero (selected) without analogous effect on the parent network group.

At the between-group level,... Ω exerts ℓ_1-like variable selection on the $(||\mathbf{w}_g||_2)_{g \in G}$ groups, yielding a maximum of $g \in G$ to be zeroed out [29]. The important consequence is that also all descentes of such a zeroed group $g \in G$ will be descarded. Conversely, if one group g is selected, then all the ancestral groups will also be selected. Thus, statistical estimation will be improved by enticing entire voxel sets to be selected or discarded as predictive, although one individual coefficient is computed for each voxel.

-¿ it is a (ℓ_1, ℓ_2)-mixed norm -¿ between-group sparsity effect by l1 -¿ within-group shrinkage effect by l2

$$\Omega(\mathbf{w}) = \sum_{g \in G} \eta_g ||\mathbf{w}_g||_2 \tag{2}$$

$(\eta_g)_{g \in G}$ are positive weights for the groups

fit to the data is measured through a convex loss function (w, b) L(y, X, w, b) R+.

Classification logistic loss function

$$P(y = k|x, W, b) = \frac{exp\{x^T w^k + b_k\}}{\sum_{m=1}^{c} exp\{x^T w^m + b_m\}} \arg\min \frac{1}{2}||u - v||_2^2 + \lambda\Omega(\mathbf{w}) \tag{3}$$

$\lambda > 0$

bias is omitted because $\mathbf{X} and \mathbf{y}$ are mean-centered and unit-variance scaled.

In this setting, and given a new fMRI volume x, we make predictions by choosing the label that maximizes the class-conditional probabil- ities (3.1), that is, argmaxk1,...,cProb(y = k—x; W, b)

One-versus-rest scheme

Regression squared error as loss

$$\arg\min \frac{1}{2}||\mathbf{y} - \mathbf{X}\mathbf{w}||_2^2 + \lambda\Omega(\mathbf{w}) \tag{4}$$

$\lambda > 0$

Prediction for a new fMRI volume x is then simply performed by computing the dot product xŁw

Numerical optimization Difficult because high-dimensional setting

empirical risk minimization was performed by

The intercept b is left unregularized

Implementation. The analyses were performed in Python. We used *nilearn* to handle the large quantities of neuroimaging data [1] and *Theano* for automatic, numerically stable differentiation of symbolic computation graphs [5, 8]. All Python scripts that generated the results are accessible online for reproducibility and reuse (`http://github.com/banilo/nips2015`).

all algorithm from a same software library -¿ SPAMs

Data. As the currently biggest openly-accessible reference dataset, we chose resources from the Human Connectome Project (HCP) [4]. 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 (Table 1). 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.

Cognitive Task	**Stimuli**	**Instruction for participants**
1 Reward	Card game	Guess the number of a mystery card for gain/loss of money
2 Punish		
3 Shapes	Shape pictures	Decide which of two shapes matches another shape geometrically
4 Faces	Face pictures	Decide which of two faces matches another face emotionally
5 Random	Videos with objects	Decide whether the objects act randomly or intentionally
6 Theory of mind		
7 Mathematics	Spoken numbers	Complete addition and subtraction problems
8 Language	Auditory stories	Choose answer about the topic of the story
9 Tongue movement	Visual cues	Move tongue
10 Food movement		Squeezing of the left or right toe
11 Hand movement		Tapping of the left or right finger
12 Matching	Shapes with textures	Decide whether two objects match in shape or texture
13 Relations		Decide whether object pairs differ both along either shape or texture
14 View Bodies	Pictures	Passive watching
15 View Faces	Pictures	Passive watching
16 View Places	Pictures	Passive watching
17 View Tools	Pictures	Passive watching
18 Two-Back	Various pictures	Indicate whether current stimulus is the same as two items earlier

Table 1: **Description of psychological tasks to predict.**

These labeled data were complemented by unlabeled activity maps from HCP acquisitions of unconstrained resting-state activity [46]. These reflect brain activity in the absence of controlled thought. In sum, the HCP rest data concatenated 8000 unlabeled, noise-cleaned rest maps with 40 brain maps from each of 200 randomly selected participants.

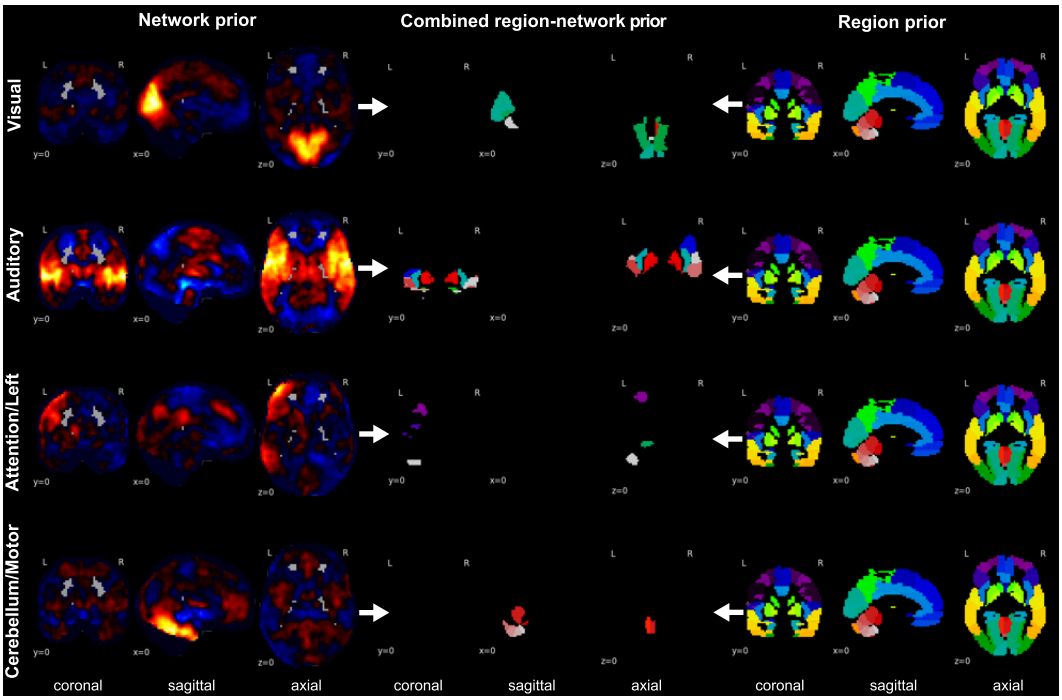

Figure 1: **Building blocks of the region-network tree.** Depicts neurobiological priors introduced into the classification problem by hierarchical structured sparsity. *Left:* Continuous, partially overlapping brain network priors (*hot-colored*, taken from [47]) accommodate the functional integration perspective of brain organization. *Right:* Discrete, non-overlapping brain region priors (*single-colored*, taken from [15]) 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.

We were further interested in the utility of the optimized low-rank projection in one task dataset for dimensionality reduction in another task dataset. To this end, the HCP-derived network decompositions were used as preliminary step in the classification problem of another large sample. The ARCHI dataset [40] provides activity maps from diverse experimental tasks, including auditory and visual perception, motor action, reading, language comprehension and mental calculation. Analogous to HCP data, the second task dataset thus incorporated 1404 labeled, grey-matter masked, and z-scored activity maps from 18 diverse tasks acquired in 78 participants.

sparse statistical models have only few nonzero parameters

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 expectes 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

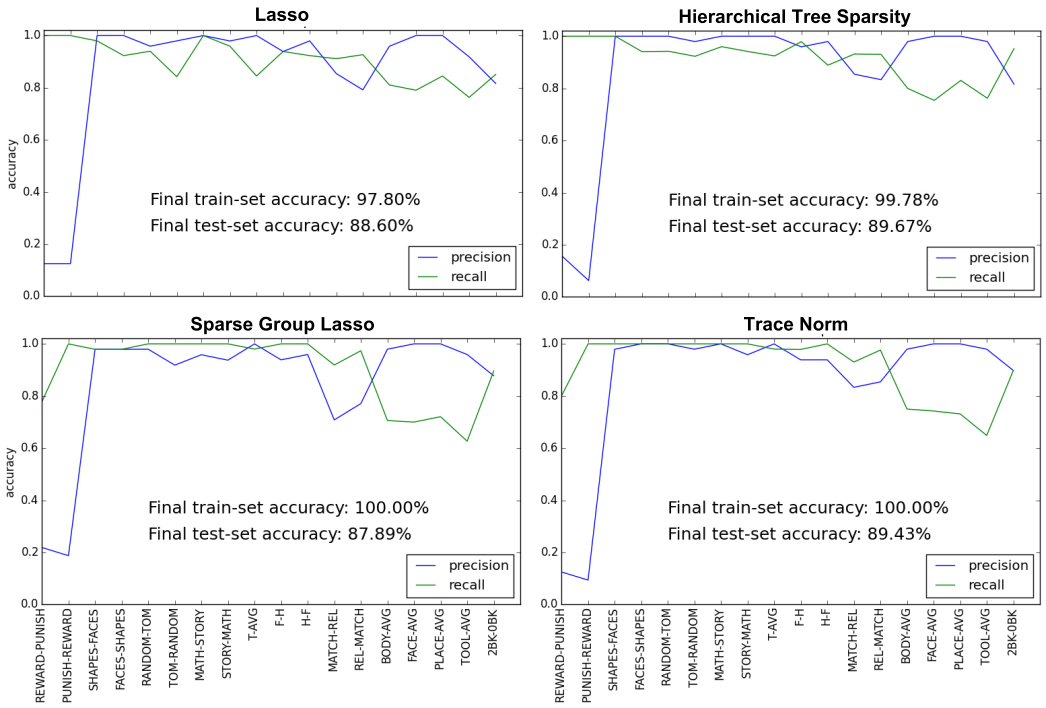

Figure 2: **Model performance across sparsity priors.** Compares the performance of logistic regression estimators with 4 different structured and unstructered sparsity terms in classifying neural activity from 18 psychological tasks. The class-wise precision and recall metrics were obtained on the same test set. Unstructured ℓ_1-penalized logistic regression (*upper left*) imposed a minimum of relevant brain voxels without assuming special structure. Structured ℓ_1/ℓ_2 block norm with additional ℓ_1 term (*lower left*) imposed region compartments, but naïve to network structure. Structured trace-norm penalization (*lower right*) imposed low-rank structure with sparsity of network patterns, but naïve to region structure. Structured ℓ_1/ℓ_2 block norm with a hierarchy of both region and network priors (*upper right*) exhibited the best out-of-sample performance. A priori knowledge of both region and network neighborhoods was hence most beneficial for predicting psychological tasks from brain maps.

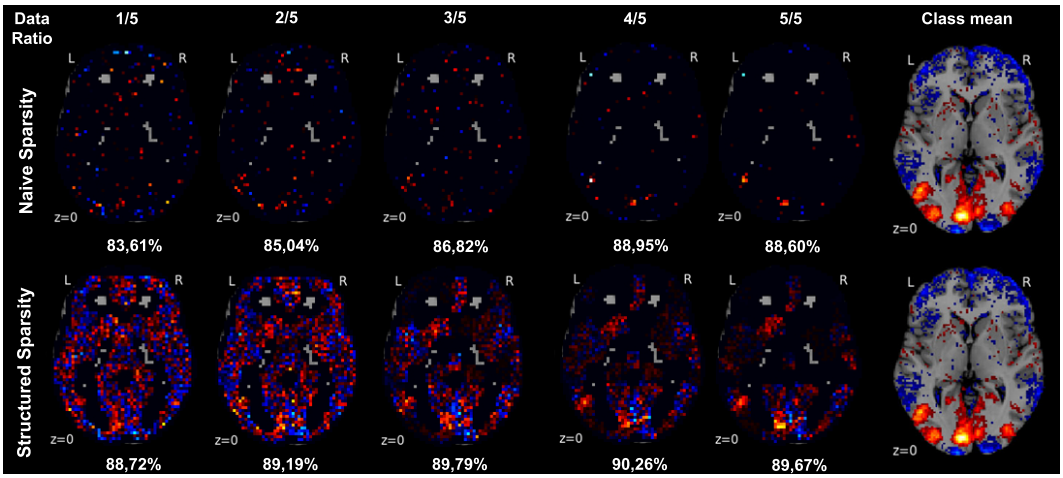

Figure 3: **Naïve versus informed sparse model selection across training set sizes.** Ordinary ℓ_1-penalized logistic regression (*upper row*) is compared to hierarchical-tree-penalized logistic regression (*lower row*) with increasing fraction of the available training data (*left to right columns*). 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 classes 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.

regression with trace-norm penalization imposed low-rank structure. This supervised classification algorithm expected a minimum of unknown network patterns to be relevant. The stratified and shuffled training data were submitted to 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. Hierarchial tree sparsity demonstrated the best generalization in distinguihsing unseen neural activity maps from 18 tasks (89.67%, mean recall XX.YY, mean precision XX.YY). It was closely followed by logistic regression with trace-norm regularization that is structured but not neurobiologically informed (89.43%, mean recall XX.YY, mean precision XX.YY). Lasso featured an average performance comparing to the other sparse estimators (88.60%, mean recall XX.YY, mean precision XX.YY) Introduced a priori knowledge of brain region compartments by sparse group sparsity performed worst (87.89%, mean recall XX.YY, mean precision XX.YY). In sum, biasing sparse model selection by domain knowledge of region-network hierarchies outperformed other types of frequently used sparse penalization techniques.

Sample complexity of naïve 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 ℓ_1-constrained logistic regression naïve to neurobiology in the data-scarce and data-rich scenarios. The HCP task data with examples from 18 classes were first divided into 90% of training set (i.e., 7584 neural activity maps) and 10% of test set (i.e., 842 maps). Both learning algorithms were fitted based on the training set at different subsampling fractions: 20% (1516 maps), 40% (3033 maps), 60% (4550 maps), 80% (6067 maps), and 100% (7584 maps). The stratified and shuffled training data were submitted to 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. Three observation have been made. In the data-scarce scenario (i.e., 1/5 of actual 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 training data average. In the case of scarce training data, which is typical for the brain imaging domain, regularization by region-network priors indeed allowed for more effective extraction of classification-relevant structure from the neural activity scans. Across scenarios, the weight maps from ordinary logistic regression exhibit higher variance and many 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 profits 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 and network emphasis. Finally, the relative importance of the region and network priors within the hierarchical tree prior was quantified (Figure 4). The η_g group of region priors was multiplied with a region-network ratio, while the η_g group of network priors was biased by the corresponding network-region ratio. A region-network ratio of 3, for instance, increased the relative importance of known region structure by multiplying $\frac{3}{1}$ to the η_g factor of all region groups and multiplying $\frac{1}{3}$ to all network groups. The data splitting cross-validation scheme was identical to the above modelling experiments. As the most important observation, a range between region-dominant and network-dominant structured penalties yields 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 extrapolation to unseen neural activity maps. Second, these perform similar 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 maps of the most necessary region or 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 Discussion

Relevant structure in neuroimaging data has long been investigated according to two separate organizational principles: functional segregation into discrete brain regions and functional integration by interregional brain networks. This paper 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. Domain-informed hierarchical structured sparsity 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 neuroimaging studies that capitalized on dimensionality reduction to harness the curse of dimensionality. They often used preliminary region-wise pooling functions or regression against network templates for subsequent supervised learning on the aggregated feature space. Such lossy two-step approaches of feature engineering and inference *i)* can only account for either functional specialization or functional integration of brain organization, *ii)* depend on the ground truth being a region or network effect, and *iii)* cannot issue

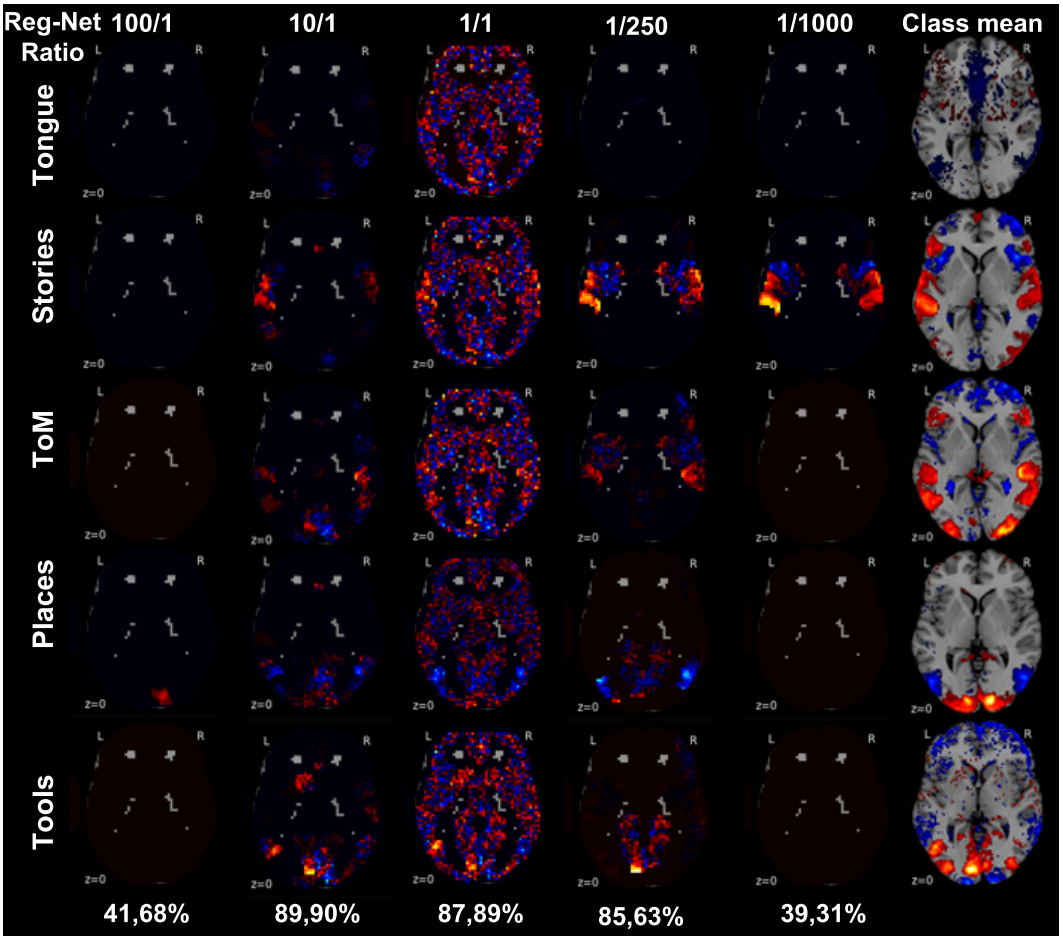

Figure 4: **Support recovery as a function of region and network emphasis.** The relative impact of the region and network priors on model selection is systematically varied against each other. This region-network ratio (*upper fractions*) weighted voxel groups to priviledge sparse models in function 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.

individual coefficients for every brain voxels. Hierarchical region-network sparsity overcomes 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 [31] and k-means clustering [53]. Additionally, model regularization by hierarchical structured sparsity could be extended from the spatial domain of neural activity to priors of coherent spatiotemporal activity structure [22]. Ultimately, successful high-dimensional inference is an important prerequisite for predicting diagnosis, disease trajectories, and treatment responses in personalized psychiatry and neurology.

Acknowledgment. The research leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement no. 604102 (Human Brain Project). Data were provided by the Human Connectome Project. Further support was received from the German National Academic Foundation (D.B.), the German Research Foundation (BZ2/2-1 and BZ2/3-1 to D.B.), and the MetaMRI associated team (B.T., G.V.).

References

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

[2] Amunts, K., Lepage, C., Borgeat, L., Mohlberg, H., Dickscheid, T., Rousseau, M.E., Bludau, S., Bazin, P.L., Lewis, L.B., Oros-Peusquens, A.M., et al.: Bigbrain: an ultrahigh-resolution 3d human brain model. Science 340(6139), 1472–1475 (2013)

[3] Bach, F., Jenatton, R., Mairal, J., Obozinski, G.: Optimization with sparsity-inducing penalties. Foundations and Trends® in Machine Learning 4(1), 1–106 (2012)

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

[5] Bastien, F., Lamblin, P., Pascanu, R., Bergstra, J., Goodfellow, I., Bergeron, A., Bouchard, N., Warde-Farley, D., Bengio, Y.: Theano: new features and speed improvements. arXiv preprint arXiv:1211.5590 (2012)

[6] Beckmann, C.F., DeLuca, M., Devlin, J.T., 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)

[7] Behrens, T.E., Johansen-Berg, H., Woolrich, M.W., Smith, S.M., Wheeler-Kingshott, C.A., Boulby, P.A., Barker, G.J., Sillery, E.L., Sheehan, K., Ciccarelli, O., Thompson, A.J., Brady, J.M., Matthews, P.M.: Non-invasive mapping of connections between human thalamus and cortex using diffusion imaging. Nat Neurosci 6(7), 750–7 (2003)

[8] Bergstra, J., Breuleux, O., Bastien, F., Lamblin, P., Pascanu, R., Desjardins, G., Turian, J., Warde-Farley, D., Bengio, Y.: Theano: a cpu and gpu math expression compiler. Proceedings of the Python for scientific computing conference (SciPy) 4, 3 (2010)

[9] Brodmann, K.: Vergleichende Lokalisationslehre der Groshirnrinde (1909)

[10] Buckner, R.L., Krienen, F.M., Yeo, B.T.: Opportunities and limitations of intrinsic functional connectivity mri. Nature neuroscience 16(7), 832–837 (2013)

[11] Bühlmann, P., Van De Geer, S.: Statistics for high-dimensional data: methods, theory and applications. Springer Science & Business Media (2011)

[12] Bullmore, E., Sporns, O.: Complex brain networks: graph theoretical analysis of structural and functional systems. Nature Reviews Neuroscience 10(3), 186–198 (2009)

[13] Buzsáki, G., Draguhn, A.: Neuronal oscillations in cortical networks. science 304(5679), 1926–1929 (2004)

[14] Corbetta, M., Patel, G., Shulman, G.L.: The reorienting system of the human brain: from environment to theory of mind. Neuron 58(3), 306–24 (2008)

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

[16] Daudet, L.: Sparse and structured decompositions of audio signals in overcomplete spaces. In: In International Conference on Digital Audio Eects (2004)

[17] Eickhoff, S.B., Thirion, B., Varoquaux, G., Bzdok, D.: Connectivity-based parcellation: Critique and implications. Hum Brain Mapp (2015)

[18] Fransson, P.: How default is the default mode of brain function? further evidence from intrinsic bold signal fluctuations. Neuropsychologia 44, 28362845 (2006)

[19] Fristen, K.J.: Imaging cognitive anatomy. Trends in cognitive sciences 1(1), 21–27 (1997)

[20] Friston, K.: Beyond phrenology: what can neuroimaging tell us about distributed circuitry? Annual review of neuroscience 25(1), 221–250 (2002)

[21] Giraud, C.: Introduction to High-Dimensional Statistics. CRC Press (2014)

[22] Gramfort, A., Papadopoulo, T., Baillet, S., Clerc, M.: Tracking cortical activity from m/eeg using graph cuts with spatiotemporal constraints. NeuroImage 54(3), 1930–1941 (2011)

[23] Harzallah, H., Jurie, F., Schmid, C.: Combining efficient object localization and image classification. In: Computer Vision, 2009 IEEE 12th International Conference on. pp. 237–244. IEEE (2009)

[24] Hastie, T., Tibshirani, R., Wainwright, M.: Statistical Learning with Sparsity: The Lasso and Generalizations. CRC Press (2015)

[25] Huang, J., Zhang, T., Metaxas, D.: Learning with structured sparsity. J Mach Learn Res 12, 3371–3412 (2011)

[26] Hubel, D.H., Wiesel, T.N.: Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. The Journal of physiology 160(1), 106 (1962)

[27] Iaria, G., Fox, C.J., Waite, C.T., Aharon, I., 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)

[28] Jbabdi, S., Behrens, T.E.: Long-range connectomics. Annals of the New York Academy of Sciences 1305(1), 83–93 (2013)

[29] Jenatton, R., Audibert, J.Y., Bach, F.: Structured variable selection with sparsity-inducing norms. The Journal of Machine Learning Research 12, 2777–2824 (2011)

[30] Jenatton, R., Gramfort, A., Michel, V., Obozinski, G., Bach, F., Thirion, B.: 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)

[31] Jenatton, R., Obozinski, G., Bach, F.: Structured sparse principal component analysis. arXiv preprint arXiv:0909.1440 (2009)

[32] Jordan, M.I.: Frontiers in massive data analysis. National Academies Report (2015)

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

[34] Kanwisher, N.: 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)

[35] Kim, S., Xing, E.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)

[36] Kim, T., Shakhnarovich, G., Urtasun, R.: Sparse coding for learning interpretable spatio-temporal primitives. In: Advances in neural information processing systems. pp. 1117–1125 (2010)

[37] Mesulam, M.M.: From sensation to cognition. Brain 121, 1013–52 (1998)

[38] Morales, J., Micchelli, C.A., Pontil, M.: A family of penalty functions for structured sparsity. In: Advances in Neural Information Processing Systems. pp. 1612–1623 (2010)

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

[40] Pinel, P., Thirion, B., Meriaux, S., Jobert, A., Serres, J., Le Bihan, D., Poline, J.B., Dehaene, S.: Fast reproducible identification and large-scale databasing of individual functional cognitive networks. BMC Neurosci 8, 91 (2007)

[41] Raichle, M.E., MacLeod, A.M., Snyder, A.Z., Powers, W.J., Gusnard, D.A., Shulman, G.L.: A default mode of brain function. Proc Natl Acad Sci U S A 98(2), 676–82 (2001)

[42] Rapaport, F., Barillot, E., Vert, J.P.: Classification of arraycgh data using fused svm. Bioinformatics 24(13), i375–i382 (2008)

[43] Saygin, Z.M., Osher, D.E., Koldewyn, K., Reynolds, G., Gabrieli, J.D., Saxe, R.R.: Anatomical connectivity patterns predict face selectivity in the fusiform gyrus. Nat Neurosci 15(2), 321–7 (2012)

[44] Scannell, J.W., Blakemore, C., Young, M.P.: Analysis of connectivity in the cat cerebral cortex. J Neurosci 15(2), 1463–83 (1995)

[45] Seeley, W.W., Menon, V., Schatzberg, A.F., Keller, J., Glover, G.H., Kenna, H., Reiss, A.L., Greicius, M.D.: Dissociable intrinsic connectivity networks for salience processing and executive control. J Neurosci 27(9), 2349–2356 (2007)

[46] Smith, S.M., Beckmann, C.F., Andersson, J., Auerbach, E.J., Bijsterbosch, J., Douaud, G., Duff, E., Feinberg, D.A., Griffanti, L., Harms, M.P., et al.: Resting-state fmri in the human connectome project. Neuroimage 80, 144–168 (2013)

[47] 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., 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)

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

[49] Stephan, K.E., Friston, K.J., Frith, C.D.: Dysconnection in schizophrenia: from abnormal synaptic plasticity to failures of self-monitoring. Schizophr Bull 35(3), 509–27 (2009)

[50] Tononi, G., Edelman, G.M., Sporns, O.: Complexity and coherency: integrating information in the brain. Trends in cognitive sciences 2(12), 474–484 (1998)

[51] Vinci, G., Freeman, P., Newman, J., Wasserman, L., Genovese, C.: Estimating the distribution of galaxy morphologies on a continuous space. arXiv preprint arXiv:1406.7536 (2014)

[52] Vogt, C., Vogt, O.: Allgemeine Ergebnisse unserer Hirnforschung, vol. 21. JA Barth (1919)

[53] Witten, D.M., Tibshirani, R.: A framework for feature selection in clustering. Journal of the American Statistical Association 105(490) (2010)

[54] Young, M.P.: The organization of neural systems in the primate cerebral cortex. Proc Biol Sci 252(1333), 13–8 (1993)

[55] Yuste, R.: From the neuron doctrine to neural networks. Nat Rev Neurosci 16(8), 487–497 (2015)

[56] Zeki, S.M.: Functional specialisation in the visual cortex of the rhesus monkey. Nature 274(5670), 423–428 (1978)

[57] Zhao, P., Rocha, G., Yu, B.: The composite absolute penalties family for grouped and hierarchical variable selection. The Annals of Statistics pp. 3468–3497 (2009)

[58] Zilles, K., Amunts, K.: Receptor mapping: architecture of the human cerebral cortex. Current opinion in neurology 22(4), 331–339 (2009)