Semi-supervised low-rank logistic regression for high-dimensional neuroimaging data

Danilo Bzdok, Michael Eickenberg, Olivier Grisel, Bertrand Thirion, Gaël Varoquaux

email: firstname.lastname@inria.fr

Abstract

Imaging neuroscience links human behavior to aspects of brain biology in ever-increasing datasets. Existing neuroimaging methods typically perform either discovery of unknown neural structure or testing of neural structure associated with mental tasks. However, testing hypotheses on the neural correlates of mental tasks depends on the representation of the observations. We therefore propose to optimize representation and task classification in an identical statistical learning problem. We show that this approach yields more accurate and more interpretable neural models of psychological tasks in a reference dataset.

keywords: dimensionality reduction; semi-supervised learning; bioinformatics; fMRI; systems neuroscience

1 Introduction

Methods for neuroimaging research can be grouped by discovering neurobiological structure or assessing the neural correlates associated with mental tasks. To discover, on the one hand, spatial distributions of neural activity structure across time, independent component analysis (ICA) [7] is often used. It decomposes the BOLD (blood-oxygen level-dependent) signals into the primary modes of variation. The ensuing spatial activity patterns are believed to represent brain networks of functionally interacting regions [25]. Similarly, sparse principal component analysis (SPCA) [27] has been used to separate BOLD signals into parsimonious network components. Thus extracted brain networks are probably manifestations of electrophysiological oscillation frequencies [17]. Their fundamental organizational role is further attested by continued covariation during sleep and anesthesia [10]. Network discovery by applying ICA or SPCA is typically performed on task-unrelated (i.e., *unlabeled*) "resting-state" data. These capture brain dynamics during ongoing random thought without controlled environmental stimulation. In fact, a large proportion of the BOLD signal is known not to correlate with a particular behavior, stimulus, or experimental task [10].

To test, on the other hand, the neural correlates underlying mental tasks, the general linear model (GLM) is the dominant approach [13]. The contribution of individual brain voxels is estimated according to a design matrix of experimental tasks. Alternatively, psychophysiological interactions (PPI) elucidate the influence of one brain region on another conditioned by experimental tasks [12]. As a last example, always more neuroimaging studies model experimental tasks by training classification algorithms on brain signals [22]. All these methods are applied to task-associated (i.e., *labeled*) data that capture brain dynamics during stimulus-guided behavior. Two important conclusions can be drawn. First, the mentioned supervised neuroimaging analyses typically yield results in a voxel space. This ignores the fact that the BOLD signal exhibits spatially distributed activity patterns. Second, existing supervised neuroimaging analyses do not acknowledge the abundance of easily acquired resting-state data [9]. These may allow better sampling the manifold of brain states due to the high task-rest similarities of neural activity patterns, as observed using ICA [25].

Both these biological properties can be conjointly exploited in a mixed (i.e., use task and rest data) low-rank (i.e., perform network decomposition) approach. The integration of brain-network discovery into supervised classification formally yields a semi-supervised learning framework. The most relevant neurobiological structure should hence be identified for the prediction problem at hand. Autoencoders suggest themselves because they can emulate variants of most unsupervised learning algorithms, including PCA, SPCA, and ICA [16]. Autoencoders are one-layered learning models that condense the input data to local and global representations by improving reconstruction from them. They behave like a PCA in case of one linear hidden layer and a squared error loss [4]. This architecture yields a convex optimization objective with unique global minimum. Autoencoders behave like a SPCA if shrinkage terms are added to the model weights in the optimization objective. Moreover, they have the characteristics of an ICA in case of tied weights and adding a nonlinear convex transformation [18]. These authors further demonstrated that ICA, sparse autoencoders, and sparse coding are mathematically equivalent under mild conditions. Hence, autoencoders can flexibly project the neuroimaging data onto the main axes of variation and thus reverse-engineer properties of the underlying neural processes [20].

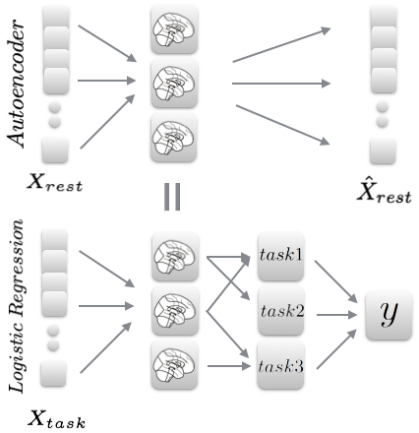


Figure 1: **Architecture.** Linear autoencoders find an optimal compression of 79,941 brain voxels into n unknown activity patterns by improving reconstruction from them. The decomposition matrix equates with the bottleneck of a factored logistic regression. Supervised multi-class learning on task data (X_{task}) is thus guided by unsupervised decomposition of rest data (X_{rest}).

In the present investigation, a linear autoencoder will be fed by (unlabeled) rest data and integrated as bottleneck into a low-rank logistic regression fed by (labeled) task data. We can then solve the unsupervised data representation and the supervised classification in an identical statistical learning problem. From the perspective of dictionary learning, the first layer represents the discovered set of basis functions whose linear combinations are learned by the second layer [20]. Neurobiologically, this allows delineating a low-dimensional manifold of brain network patterns and then distinguishing mental tasks by their most discriminative linear combinations. Theoretically, a reduction in model variance should be achieved by resting-state autoencoders that put probability mass on the most neurobiologically valid models in the hypothesis set. Taken together, the important modes of variation in brain dynamics and the neural correlates subserving mental operations have mostly been studied in isolation. We provide a principled computational framework to link these previously unconnected domains of systems neuroscience.

2 Methods

Data. As the currently biggest openly-accessible reference dataset, we chose the Human Connectome Project (HCP) resources [5]. Neuroimaging task data with labels of ongoing cognitive processes were drawn from 500 unrelated, healthy HCP participants. 18 HCP tasks were selected that are known to elicit reliable neural activity across participants. The task paradigms include 1) working memory/cognitive control processing, 2) incentive processing, 3) visual and somatosensory-motor processing, 4) language processing (semantic and phonological processing), 5) social cognition, 6) relational processing, and 7) emotional processing. All data were acquired on the same Siemens Skyra 3T scanner. Whole-brain EPI acquisitions were acquired with a 32 channel head coil (TR=720ms, TE=33.1ms, flip angle=52, BW=2290Hz/Px, in-plane FOV=280 $\times$ 180mm, 72 slices, 2.0mm isotropic voxels). The "minimally preprocessed" pipeline includes gradient unwarping, motion correction, fieldmap-based EPI distortion correction, brain-boundary-based registration of EPI to structural T1-weighted scans, nonlinear (FNIRT) registration into MNI space, and grand-mean intensity normalization. Activity maps were spatially smoothed with a Gaussian kernel of 4mm (FWHM). A GLM was implemented by FILM from the FSL suite with model regressors from convolution with a canonical hemodynamic response function and from temporal derivatives. HCP tasks were conceived to modulate activity in a maximum of different brain regions and neural sys-

tems. Indeed, at least 70% of the participants showed consistent brain activity in contrasts from the task battery, which certifies excellent coverage [5]. 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 60x72x60 space of 3mm isotropic voxels and gray-matter masked (at least 10% tissue probability). The supservised analyses were based on labeled HCP task maps with 79,941 voxels of interest representing z-values in gray matter.

These labeled data were complemented by unlabeled activity maps from HCP acquisitions of unconstrained resting-state activity. These reflect brain activity in the absence of controlled thought. In line with the goal of the present study, acquisition of these data was specifically aimed at the study of task-rest correspondence. From each participant, we included two time-series for left and right phase encoding with 1,200 maps of multiband, gradient-echo planar imaging acquired during a period of 15min (TR=720 ms, TE=33.1 ms, flip angle=52, FOV=280×180mm, and 2.0mm isotropic voxels). Besides run duration, the task acquisitions were identical to the resting-state fMRI acquisitions for maximal compatibility between task and rest data. We here drew on "minimally preprocessed" rest data from 200 randomly selected healthy, unrelated participants. PCA was applied to each set of 1,200 rest maps for denoising by keeping only the 20 main modes of variation. In sum, the HCP rest data concatenated 8000 unlabeled, noise-cleaned rest maps with 40 brain images from each of 200 randomly selected participants.

We were further interested in the utility of the optimal 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 classifcation problem of another large sample. The ARCHI dataset [21] provides activity maps from diverse experimental tasks, including auditory and visual perception, motor action, reading, language comprehension and mental calculation. 81 right-handed healthy participants (3 not included in present analyses due to incomplete data) without psychiatric or neurological history participated in four fMRI sessions acquired under different experimental paradigms. The functional maps were warped into the MNI space and resampled to isotropic 3mm resolution. Whole-brain EPI data were acquired with the same Siemens Trio with a 32 channel head coil (TR=2400ms, TE=30ms, flip angle=60, in-plane FOV=19.2×19.2cm, 40 slices, 3.0mm isotropic voxels). Standard preprocessing was performed with Nipype [14], including slice timing, motion correction, alignment, and spatial normalization. activity maps were spatially smoothed by a Gaussian kernel of 5mm (FWHM). Analogous to HCP data, the second task dataset incorporated 1404 labeled, grey-matter masked, and z-scored activity maps from 18 diverse tasks acquired in 78 participants.

The labeled and unlabeled data were fed into a linear statistical model composed of autoencoder and low-rank logistic regression.

Linear autoencoder. The affine autoencoder takes the input $\mathbf{x}$ and projects it into coordinates of a latent representation $\mathbf{z}$ by

$$\begin{aligned} \mathbf{z} &= \mathbf{W_0}\mathbf{x} + \mathbf{b_{W_0}} \\ \mathbf{x}' &= \mathbf{W_1}\mathbf{z} + \mathbf{b_{W_1}} \end{aligned} \tag{1}$$

where $\mathbf{x} \in \mathbb{R}^{\mathbf{d}}$ denotes the vector of $d = 79,941$ voxel values from each rest map, $\mathbf{z} \in \mathbb{R}^{\mathbf{n}}$ is the n hidden dimensions (i.e., distributed neural activity patterns), and $\mathbf{x}' \in \mathbb{R}^{\mathbf{d}}$ is the reconstruction vector of the original activity map from the hidden variables. Further, $\mathbf{W_0}$ denotes the weight matrix that transforms from input space into the hidden space (encoder), $\mathbf{W_1}$ is the weight matrix for back-projection from the hidden variables to the output space (decoder). $\mathbf{b_{W_0}}$ and $\mathbf{b_{W_1}}$ are bias vectors. Note that $\mathbf{W_0}$ and $\mathbf{W_1}$ are tied, $\mathbf{W_0} = \mathbf{W_1^T}$. The optimal model parameters $\mathbf{W_0}, \mathbf{b_{W_0}}, \mathbf{b_{W_1}}$ are found by minimizing the imperfection of reconstruction according to squared error

$$\mathcal{L}_{\mathcal{AE}}(\mathbf{x}, \mathbf{x}') = ||\mathbf{x} - \mathbf{x}'||^2 \tag{2}$$

This reconstruction error criterion equates with maximizing a lower bound on the mutual information between input and the learned representation. Nonlinearities were not applied on the activations in the first layer.

Reduced-rank logistic regression. Lossy compression by a low-dimensional bottleneck is also imposed by the first layer of the low-rank multinomial logistic regression. It gives the probability of an input $\mathbf{x}$ to belong to a class $i \in \mathbb{N}$

$$\begin{aligned} P(Y = i|\mathbf{x}, \mathbf{V_0}, \mathbf{V_1}, \mathbf{b_{V_0}}, \mathbf{b_{V_1}}) &= softmax_i(\mathbf{V_1}(\mathbf{V_0}\mathbf{x} + \mathbf{b_{V_0}}) + \mathbf{b_{V_1}}) \\ &= \frac{e^{\mathbf{V_{1i}}(\mathbf{V_{0i}}\mathbf{x}+\mathbf{b_{V_{0i}}})+\mathbf{b_{V_{1i}}}}}{\sum_j e^{\mathbf{V_{1j}}(\mathbf{V_{0j}}\mathbf{x}+\mathbf{b_{V_{0j}}})+\mathbf{b_{V_{1j}}}}} \end{aligned} \tag{3}$$

where the matrix $\mathbf{V_0} \in \mathbb{R}^{\mathbf{dxn}}$ transforms the input $\mathbf{x} \in \mathbb{R}^{\mathbf{d}}$ into n latent components and the matrix $\mathbf{V_1} \in \mathbb{R}^{\mathbf{nxc}}$ projects the latent components onto hyperplanes that reflect c class probabilities. $\mathbf{b_{V_0}}$ and $\mathbf{b_{V_1}}$ are corresponding bias vectors. The loss function is given by

$$\mathcal{L}_{\mathcal{LR}}(\mathbf{x}, \mathbf{y}) = \frac{1}{N_{X_{task}}} \sum_{i=0}^{N_{X_{task}}} \log(P(Y = y^{(i)}|\mathbf{x^{(i)}}; \mathbf{V_1}(\mathbf{V_0}\mathbf{x} + \mathbf{b_{V_0}}) + \mathbf{b_{V_1}}) \tag{4}$$

Layer combination. Importantly, the optimization problem of the linear autoencoder and the low-rank logistic regression are linked one two levels. First, their transformation matrices mapping from input to the latent space are identical

$$\mathbf{V_0} = \mathbf{W_0} \tag{5}$$

We thus search for a compression of the 79,941 voxel values into n latent components that represent an optimal latent code for both rest and task activity data. Second, the objectives of the autoencoder and the low-rank logistic regression are interpolated in the common loss function

$$arg\,min_\theta \;\; \mathcal{L}(\theta, \lambda) = (\lambda)\mathcal{L}_{\mathcal{LR}} + (1 - \lambda)\frac{1}{N_{X_{rest}}}\mathcal{L}_{\mathcal{AE}} + \ell_1(\theta) + \ell_2(\theta) \tag{6}$$

In so doing, we search for the combined model parameters $\theta = \{\mathbf{V_0}, \mathbf{V_1}, \mathbf{b_{V_0}}, \mathbf{b_{V_1}}, \mathbf{b_{W_0}}, \mathbf{b_{W_1}}\}$ with respect to the (unsupervised) reconstruction error and the (supervised) task classification. $\mathcal{L}_{\mathcal{AE}}$ is devided by $N_{X_{rest}}$ to equilibrate both loss terms to the same order of magnitude. For parameter shrinkage, we impose a combination of ℓ_1 and ℓ_2 penalties $\forall p \in \theta$ (i.e., elasticnet).

Optimization. The common objective was approximated in the neuroimaging data by updating parameters by derivates of the semi-supervised low-rank logistic regression. The required gradients are easily obtained by using the chain rule to backpropagate error derivatives through the linear network. As solver, we chose *rmsprop* [26], a mini-batch version of rprop. This procedure dictates an adaptive learning rate for each model parameter by scaled gradients from a running average. Gradient normalization by rmsprop is known to effectively exploit curvature information. We opted for a small batch size of 100, given the high degree of redundancy in X_{rest} and X_{task}. The matrix parameters were initalized by Gaussian random values multiplied by a gain of 0.004. The bias parameters were initalized to zero. With a slight abuse of notation, let θ denote a component of θ. The normalization factor and the update rule for θ are given by

$$\begin{aligned} \mathbf{v^{(t+1)}} &= \rho\mathbf{v^{(t+1)}} + (1 - \rho)\left(\frac{\partial f}{\partial \theta}\right)^2 \\ \theta^{(t+1)} &= \theta^{(t)} + \alpha\frac{\nabla f(\theta^{(t)})}{\sqrt{\mathbf{v^{(t+1)}} + \epsilon}}, \end{aligned} \tag{7}$$

where $0 < \rho < 1$ constitutes the decay rate. ρ was set to 0.9 to deemphasize the magnitude of the gradient. Further, α is the learning rate and ϵ a global damping factor. The hyper-parameter α was set to 0.00001 by prior studies and ϵ was set to 10^{-6}. Note that we have also experimented with other solvers (stochastic gradient descent, adadelta, and adagrad) but found that rmsprop converged faster and with higher generalization performance.

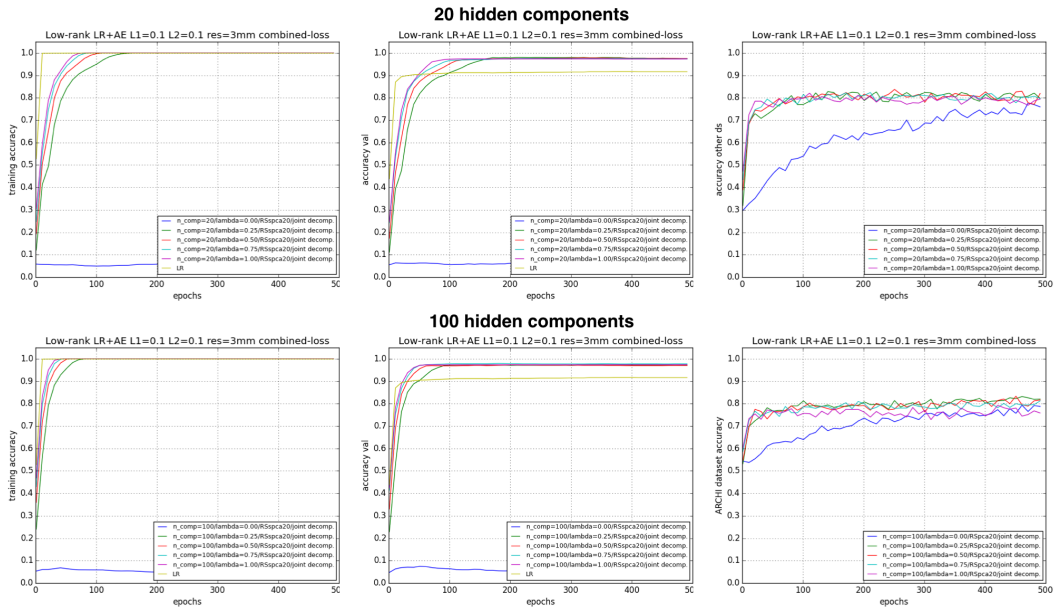


Figure 2: **Generalization performance** Depicts in-sample (apply current model on training set, *left column*), out-of-sample (current model on test set, *middle column*), and out-of-dataset (compress other dataset using the learned decomposition for classification, *right column*) accuracy. Curves for different choices of λ as well as for ordinary logistic regression (*yellowish*) across iterations over the HCP task data (epochs). XXX: Message ??

Hints. In fact, the constraint by a rest-data autoencoder qualifies as a *hint* rather than regularization in a strict sense [2]. Its purpose is not to prevent overfitting but to introduce prior knowledge on known properties of the unknown target function f. Rather than only relying on input-output pairs in the learning process, we thus narrow our hypothesis set to the biologically most plausible solutions. That is, we reduce the search space in a way that is compatible with the expected representation of BOLD activity signals.

Implementation. The analyses were performed in Python. We used *nilearn* to handle the high-dimensional neuroimaging data [1] and *theano* for automatic, numerically stable differentiation of symbolic computation graphs [6, 8]. All Python scripts that generated the results are accessible online for reproducibility and reuse `http://github.com/anonymous/anonymous`.

3 Experimental Results

General

- 20 components: high bias/low variance 100 compoentens: low bias/high variance

the features of the classes show that we captured activaiton properties, not particpant or paradigm properties

out of sample performance: extract features of general relevance beyond the supervised problem at hand

Classification performance

- This was evaluated by the prediction accuracy on a validation set (20% of the training data) at each iteration over the ARCHI task data

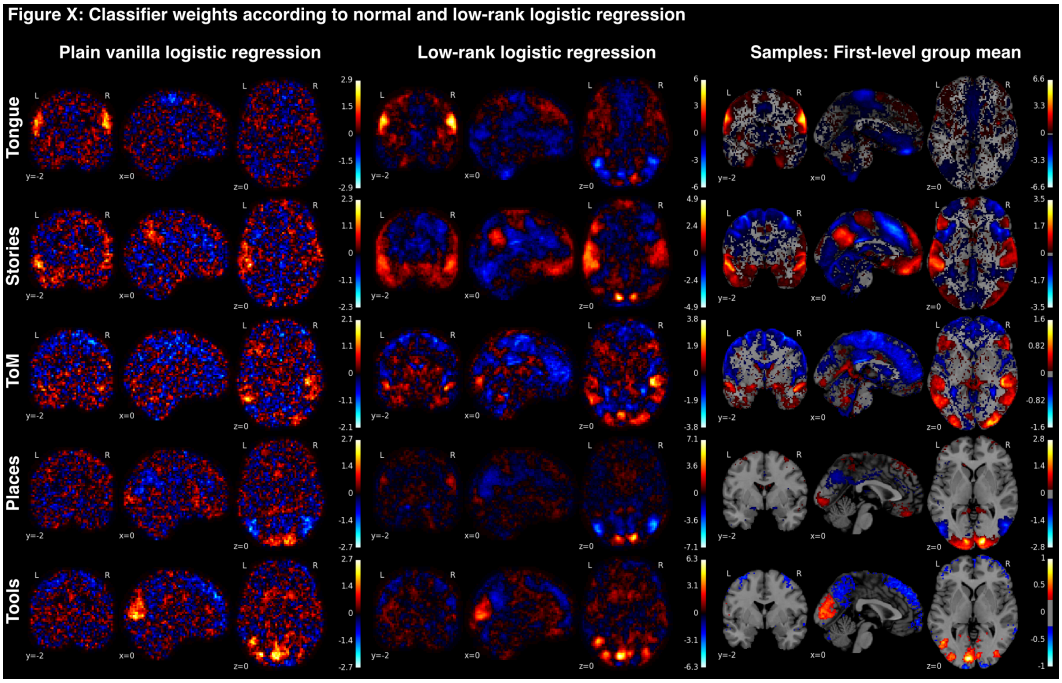
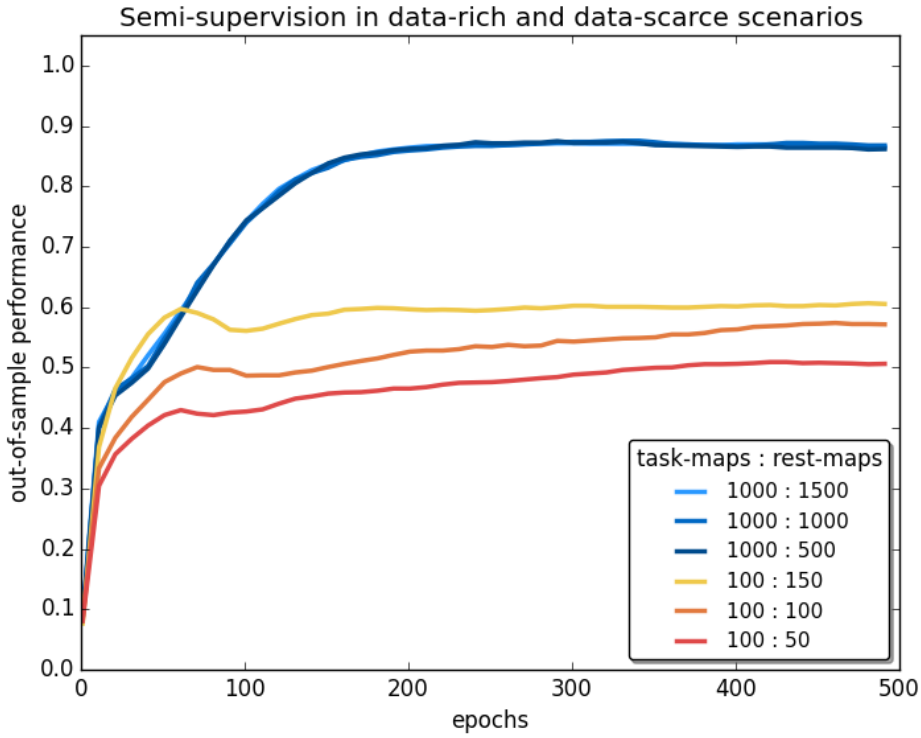


Figure 3: **Comparing the interpretability of classification weights** The voxel predictors corresponding to 5 exemplary (of 18 total) psychological tasks (*rows*) from the HCP dataset [5]. *Left column:* ordinary logistic regression (same implementation but without bottleneck or autoencoder), *middle column:* semi-supervised low-rank logistic regression (latent components $n = 20$, $\lambda = 0.5$, $\ell_1 = 0.1$, $\ell_2 = 0.1$), *right column:* voxel-wise average of whole-brain activity maps from each task (thresholded at 0.25). Low-rank logistic regression puts a) higher absolute weights on relevant structure b) lower ones on irrelevant structure, and c) yields BOLD-typical local contiguity. All values are z-scored and smoothed by a Gaussian (fwhm=4).



whenever I tried to have 3+ times more rest than task images results dropped dramatically...naive interpretation: if the rest information is too rich it overrules task patterns and hinders the supervised learning

whenever I tried to have 3+ times more rest than task images results dropped dramatically...naive interpretation: if the rest information is too rich it overrules task patterns and hinders the supervised learning

Figure 4: **Relation between support recovery and out-of-dataset performance** *Upper row:* the fit-for-purpose of the classification models is quantified by linear correlation between the z-scored average task image and the z-scored weights from ordinary (*red*) and present (*blue colors*) logistic regression ($\lambda = 0.25, 0.50, 0.75, 1.00$). *Lower row:* ordinary logistic regression is applied to the ARCHI dataset after projection into the latent dimensions learned from the HCP task data.
whenever I tried to have 3+ times more rest than task images results dropped dramatically...naive interpretation: if the rest information is too rich it overrules task patterns and hinders the supervised learning

- the less lambda, the faster the convergence, with small differences though; smaller differences with larger number of components
- both 20 and 100 hidden components achieve 97% out-of-sample accuracy
- ordinary logistic regression converges faster but with stagnation at 92% out-of-sample accuracy
- classification at chance 6% in case of lambda==0 because only the unsupervised layer of the estimator is optimized
- important: the thus only resting-state-trained decomposition achieves the same usefulness in the ARCHI classification than the decomposition derived in semi-supervised fashion

Support recovery

- fit for purpose; the right model should have high probability in the right places
- comparing to ordinary logistic regression, low-rank logreg yielded class weights that were much more similar to feature of the respective training examples, for any choice of λ
- the higher λ, the better support recovery
- conversely, the higher lambda (i.e., bigger emphasis on the resting-state autoencoder), the more useful the low-dimensional transformations for classification in an independent dataset (ARCHI) as measured by f1 score for higher (i.e., 50 and 100) but not lower (i.e., 20) numbers of latent components

Anecdotal observations

- Modifications of the modek that did not improve the generalization performance: drop-out, input corruption (p=0.1/0.3/0.5), RELUs
- no penalty or only L1 or L2 penalty generally yielded worse generalization performance
- low L1/L2 penalties (0.1) yielded better results than higher values (0.3, 0.5)
- "pretraining" of the bottleneck by settings weights to preliminary PCA, SPCA or ICA weights did not give better accuracies, neither for autoencoder pretraining
- present approach better than first PCA, SPCA or ICA and subsequently Logreg on loadings, even if each on the whole data
- introducing a nonlinearity (sigmoid, tanh, softplus) into the system did not improve predictive accuracy but elastic did -¿ most useful decomposition has characteristics of a SPCA and not PCA or ICA
- introduced an addition layer with many units (50/100) did not improve generalization

4 Discussion and Conclusion

Using the flexibility of autoencoders, we learn the optimal decomposition from high-dimensional voxel brain space into the most important activity patterns for a supervised learning question. The higher generalization accuracy and support recovery, comparing to ordinary logistic regression, hold potential for adoption in various neuroimaging analyses. Besides increased performance, these models are more interpretable by automatically learning a mapping to and from a brain-network space. This domain-specifc classification algorithms encourages departure from the artificial and statistically less attractive voxel space. Neurobiologically, neural activity underlying defined mental operations can be explained by linear combinations of the main activity patterns. That is, fMRI data probably concentrate near a low-dimensional manifold of brain networks. Extracting fundamental building blocks of brain organization might facilitate the quest for the cognitive primitives of human thought. We hope that these first steps stimulate development towards powerful semi-supervised classification methods in systems neuroscience.

In the future, automatic reduction of brain images to their neurobiological essence may leverage data-intense neuroimaging studies. Initiatives for data collection are rapidly increasing in neuroscience [23]. These promise structured integration of accumulating neuroscientific knowledge from

neuroimaging databases. Tractability by condensed feature representations can avoid the ill-posed problem of learning the full distribution of activity patterns. This is not only relevant to the multi-class challenges spanning the human cognitive space [24] but also the multi-modal combination with high-resolution 3D models of brain anatomy [3] and high-throughput genome analyses [19]. The biggest socioeconomic potential may lie in across-hospital clinical studies that predict disease trajectories and drug responses in psychiatric/neurological populations [11, 15].

Acknowledgment Data were provided by the Human Connectome Project. The study was supported by the German National Academic Foundation (D.B.).

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] Abu-Mostafa, Y.S.: Learning from hints. Journal of Complexity 10(1), 165–178 (1994)

[3] Amunts, K., Lepage, C., Borgeat, L., Mohlberg, H., Dickscheid, T., Rousseau, M.É., 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)

[4] Baldi, P., Hornik, K.: Neural networks and principal component analysis: Learning from examples without local minima. Neural networks 2(1), 53–58 (1989)

[5] 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)

[6] 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)

[7] 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)

[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] Biswal, B.B., Mennes, M., Zuo, X.N., Gohel, S., Kelly, C.e.a.: Toward discovery science of human brain function. Proc Natl Acad Sci U S A 107(10), 4734–9 (2010)

[10] Fox, D.F., Raichle, M.E.: Spontaneous fluctuations in brain activity observed with functional magnetic resonance imaging. Nat Rev Neurosci 8, 700–711 (2007)

[11] Frackowiak, R., Markram, H.: The future of human cerebral cartography: a novel approach. Philosophical Transactions of the Royal Society of London B: Biological Sciences 370(1668), 20140171 (2015)

[12] Friston, K.J., Buechel, C., Fink, G.R., Morris, J., Rolls, E., Dolan, R.J.: Psychophysiological and modulatory interactions in neuroimaging. Neuroimage 6(3), 218–29 (1997)

[13] Friston, K.J., Holmes, A.P., Worsley, K.J., Poline, J.P., Frith, C.D., Frackowiak, R.S.: Statistical parametric maps in functional imaging: a general linear approach. Hum Brain Mapp 2(4), 189–210 (1994)

[14] Gorgolewski, K., Burns, C.D., Madison, C., Clark, D., Halchenko, Y.O., Waskom, M.L., Ghosh, S.S.: Nipype: a flexible, lightweight and extensible neuroimaging data processing framework in python. Front Neuroinform 5, 13 (2011)

[15] Gustavsson, A., Svensson, M., Jacobi, F., Allgulander, C., Alonso, J., Beghi, E., Dodel, R., Ekman, M., Faravelli, C., Fratiglioni, L., et al.: Cost of disorders of the brain in europe 2010. European Neuropsychopharmacology 21(10), 718–779 (2011)

[16] Hinton, G.E., Salakhutdinov, R.R.: Reducing the dimensionality of data with neural networks. Science 313(5786), 504–507 (2006)

[17] Hipp, J.F., Siegel, M.: Bold fmri correlation reflects frequency-specific neuronal correlation. Curr Biol (2015)

[18] Le, Q.V., Karpenko, A., Ngiam, J., Ng, A.Y.: Ica with reconstruction cost for efficient overcomplete feature learning. In: Advances in Neural Information Processing Systems. pp. 1017–1025 (2011)

[19] Need, A.C., Goldstein, D.B.: Whole genome association studies in complex diseases: where do we stand? Dialogues in clinical neuroscience 12(1), 37 (2010)

[20] Olshausen, B.A., et al.: Emergence of simple-cell receptive field properties by learning a sparse code for natural images. Nature 381(6583), 607–609 (1996)

[21] 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)

[22] Poldrack, R.A., Halchenko, Y.O., Hanson, S.J.: Decoding the large-scale structure of brain function by classifying mental states across individuals. Psychol Sci 20(11), 1364–72 (2009)

[23] Poldrack, R.A., Gorgolewski, K.J.: Making big data open: data sharing in neuroimaging. Nature neuroscience 17(11), 1510–1517 (2014)

[24] Schwartz, Y., Thirion, B., Varoquaux, G.: Mapping cognitive ontologies to and from the brain. arXiv preprint arXiv:1311.3859 (2013)

[25] 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)

[26] Tieleman, T., Hinton, G.: Lecture 6.5rmsprop: Divide the gradient by a running average of its recent magnitude. COURSERA: Neural Networks for Machine Learning (2012)

[27] Varoquaux, G., Gramfort, A., Pedregosa, F., Michel, V., Thirion, B.: Multi-subject dictionary learning to segment an atlas of brain spontaneous activity. Information Processing in Medical Imaging pp. 562–573 (2011)