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 an increasing number of mental tasks makes it crucial to represent the observations adequately. We therefore propose to optimize representation and task classification in a unified statistical learning problem. A multinomial logistic regression is introduced that is constrained by joint rank with a representation-enforcing autoencoder. We show that this approach yields more accurate and more interpretable neural models of psychological tasks in a reference dataset, and better generalization to other datasets.

keywords: dimensionality reduction; semi-supervised learning; 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) [**?**] 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 [**?**]. Similarly, sparse principal component analysis (SPCA) [**?**] has been used to separate BOLD signals into parsimonious network components. The extracted brain networks are probably manifestations of electrophysiological oscillation frequencies [**?**]. Their fundamental organizational role is further attested by continued covariation during sleep and anesthesia [**?**]. 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 variation is known not to correlate with a particular behavior, stimulus, or experimental task [**?**].

To test, on the other hand, the neural correlates underlying mental tasks, the general linear model (GLM) is the dominant approach [**?**]. 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 [**?**]. As a last example, an increasing number of neuroimaging studies model experimental tasks by training classification algorithms on brain signals [**?**]. 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 patterns of coherent neural activity. Second, existing supervised neuroimaging analyses cannot exploit the abundance of easily acquired resting-state data [**?**]. These may allow better discovery of the

manifold of brain states due to the high task-rest similarities of neural activity patterns, as observed using ICA [**?**].

Both these biological properties can be conjointly exploited in an approach that is mixed (i.e., using task and rest data) and low-rank (i.e., performing network decomposition). The integration of brain-network discovery into supervised classification can yield 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 [**?**]. Autoencoders are often one-layered learning models that condense the input data to local and global representations by optimizing reconstruction from them. They behave like a (truncated, non-centered) PCA in case of one linear hidden layer and a squared error loss [**?**]. This architecture yields an 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 function at the first layer [**?**]. 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 data-generating processes in the brain [**?**]. **XXX Bertrand: neither do I. I think you exxagerate.**

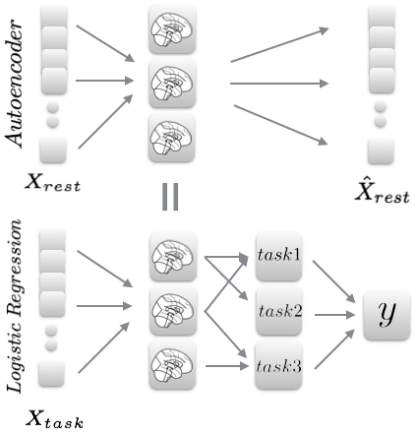

Figure 1: **Architecture.** Linear autoencoders find an optimal compression of 79,941 brain voxels into n unknown activity patterns by improving reconstruction from them. The decomposition matrix equates with the bottleneck of a factored logistic regression. Supervised multi-class learning on task data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

In the present investigation, a linear autoencoder will be fit to (unlabeled) rest data and integrated as a rank-reducing bottleneck into a multinomial logistic regression fit to (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 projectors to the discovered set of basis functions whose linear combinations are learned by the second layer [**?**]. 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 favor the most neurobiologically valid models in the hypothesis set. In summary, 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.

XXX Bertrand: I think that you could emphasize the problem that we are always data-poor, so that the set of representations that can be extracted from SPMs taken from a few tens of subjects is limited. What you propose it to extend it by i) co-analysing many problems simultaneously (multi-task) and using unlabled data that spana space of meaningful configurations XXX Bertrand: We need a sentence like: Our contribution is 1. XXX 2.YYY etc.

2 Methods

Data. As the currently biggest openly-accessible reference dataset, we chose the Human Connectome Project (HCP) resources [**?**]. 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 cog-

nition, 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=280mm $\times$ 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 systems. Indeed, at least 70% of the participants showed consistent brain activity in contrasts from the task battery, which certifies excellent activity patterns covering extended parts of the brain [**?**]. 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=280mm $\times$ 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 [**?**] 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=192mm $\times$ 192mm, 40 slices, 3.0mm isotropic voxels). Standard preprocessing was performed with Nipype [**?**], 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 an autoencoder and low-rank logistic regression.

XXX (Bertrand): You can simplify data description if needed for the sake of place. It's OK as long as you can refer to a more complete publication describing the data.

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

$$\begin{aligned} \mathbf{z} &= \mathbf{W_0 x} + \mathbf{b_0} \\ \mathbf{x}' &= \mathbf{W_1 z} + \mathbf{b_1} \end{aligned} \tag{1}$$

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-dimensional hidden state (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). $\mathbf{b_0}$ and $\mathbf{b_1}$ are bias vectors. The optimal model parameters $\mathbf{W_0}, \mathbf{b_0}, \mathbf{b_1}$ are found by minimizing the expected squared reconstruction error

$$\mathbb{E}\left[\mathcal{L}_{\mathcal{AE}}(\mathbf{x})\right] = \mathbb{E}\left[\|\mathbf{x} - \mathbf{W_1}(\mathbf{W_0}\mathbf{x} + \mathbf{b_0}) + \mathbf{b_1})\|^2\right], \tag{2}$$

This reconstruction error criterion equates with maximizing a lower bound on the mutual information between input and the learned representation. Here we choose $\mathbf{W_0}$ and $\mathbf{W_1}$ to be tied, i.e. $\mathbf{W_0} = \mathbf{W_1^T}$. In particular, this means that the weights learned are forced to take a two-fold function: That of signal *analysis* and that of signal *synthesis*. The first layer *analyzes* the data in order to obtain the cleanest latent representation, while the second layer represents building blocks from which to *synthesize* the data using the latent activations. Tying these processes together makes the analysis layer interpretable and pulls all non-zero singular values towards 1. Nonlinearities were not applied to 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 $i \in \{1, \ldots, l\}$

$$P(Y = i|\mathbf{x}; \mathbf{V_0}, \mathbf{V_1}, \mathbf{c_0}, \mathbf{c_1}) = \mathrm{softmax}_i(f_{RL}(\mathbf{x})) \tag{3}$$

where $f_{LR}(\mathbf{x}) = \mathbf{V_1}(\mathbf{V_0}\mathbf{x} + \mathbf{c_0}) + \mathbf{c_1}$ computes multinomial logits and $\mathrm{softmax}_i(x) = \exp(x_i)/\sum_j \exp(x_j)$. 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 $\mathbf{V_1} \in \mathbb{R}^{\mathbf{nxl}}$ projects the latent components onto hyperplanes that reflect l label probabilities. $\mathbf{c_0}$ and $\mathbf{c_1}$ are corresponding bias vectors. The loss function is given by

$$\mathbb{E}\left[\mathcal{L}_{\mathcal{LR}}(\mathbf{x}, \mathbf{y})\right] \approx \frac{1}{N_{X_{task}}} \sum_{k=0}^{N_{X_{task}}} \log(P(Y = y^{(k)}|\mathbf{x^{(k)}}; \mathbf{V_0}, \mathbf{V_1}, \mathbf{c_0}, \mathbf{c_1}) \tag{4}$$

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

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

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

$$\mathcal{L}(\theta, \lambda) = \lambda\mathcal{L}_{\mathcal{LR}} + (1 - \lambda)\frac{1}{N_{X_{rest}}}\mathcal{L}_{\mathcal{AE}} + \Omega(\theta). \tag{6}$$

In so doing, we search for the combined model parameters $\theta = \{\mathbf{V_0}, \mathbf{V_1}, \mathbf{c_0}, \mathbf{c_1}, \mathbf{b_0}, \mathbf{b_1}\}$ 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. $\Omega(\theta)$ represents a regularization, for which we choose the ElasticNet-type, i.e., a combination of ℓ_1 and ℓ_2 penalties $\forall p \in \theta$.

Optimization. The common objective was optimized in the neuroimaging data by gradient descent in the parameters of the semi-supervised low-rank logistic regression (SS-LR LogReg). The required gradients are easily obtained by using the chain rule to backpropagate error derivatives through the linear network. As solver, we chose *rmsprop* [**?**]. 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.

XXX: Sorry, I don't like magic numbers. Please explain where this comes from ? Maybe $\sqrt{1/p}$?

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

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

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.

Hints. In fact, the constraint by a rest-data autoencoder qualifies as a *hint* rather than regularization in a strict sense [**?**]. 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 [**?**] and *Theano* for automatic, numerically stable differentiation of symbolic computation graphs [**?**, **?**]. All Python scripts that generated the results are accessible online for reproducibility and reuse `http://github.com/anonymous/anonymous`.

3 Experimental Results

Serial versus parallel structure discovery and classification. We first tested for an advantage of combined unsupervised decomposition and supervised classification learning. We benchmarked against performing data reduction on the (unlabeled) first half of the HCP data by ICA and SPCA ($n = 5, 20, 50, 100$ components) and learning classification models in the (labeled) second half by ordinary logistic regression. ICA performed iterative blind source separation by a parallel FASTICA implementation (200 maximum iterations, per-iteration tolerance of 0.0001, initialized by random mixing matrix, whitening). SPCA separated the BOLD signals into network components with few regions by a regression-type optimization problem constrained by ℓ_1-penalty (no orthogonality assumptions, 1000 maximum iterations, per-iteration tolerance of 1 * 10^{-8}, sparsity alpha=1). The second half of the data was projected onto the latent components discovered in first data half. The ensuing component loadings were submitted to ordinary logistic regression (one hidden layer, $\ell_1 = 0.1$, $\ell_2 = 0.1$, 500 maximum iterations). This serail two-step approach was compared against SS-LR LogReg (two hidden layers, $\ell_1 = 0.1$, $\ell_2 = 0.1$, 500 maximum iterations, $\lambda = 0.75$). Importantly, all trained classification models were tested on a large, unseen test set (20% of data) in the presented analyses. Across choices for n, SS-LR LogReg achieved more than 96% out-of-sample accuracy, whereas supervised learning based on ICA and SPCA loadings ranged from 38% to 87% and 32% to 84%, respectively (Table 1). These explorations attest to the advantage of directly searching for classification-relevant structure in the fMRI data, rather than solving supervised and unsupervised problems independently. This effect was particularly pronounced when assuming few hidden dimensions.

Model performance. SS-LR LogReg was subsequently trained (500 epochs) across parameter choices for the hidden components ($n = 5, 20, 100$) and the balance between autoencoder and

n	ICA + LogReg	SPCA + LogReg	SS-LR LogReg
5	37,53 %	32,19 %	96,50 %
20	80,98 %	78,15 %	97,33 %
50	84,19 %	83,97 %	97,69 %
100	87,28 %	82,19 %	97,80 %

Table 1: Serial versus parallel dimensionality reduction and classification XXX: only one decimal is meaningful

logistic regression ($\lambda = 0.25, 0.5, 0.75, 1$). Assuming 5 latent directions of variation should yield models with higher bias and smaller variance then SS-LR LogReg with 100 latent directions. Given the 18-class problem of HCP, setting λ to 0 consistently yields generalization performance at chance-level (5,6%) because only the unsupervised layer of the estimator is optimized. At each epoch (i.e., iteration over the data), the out-of-sample performance of the trained classifier was assessed on 20% of unseen HCP data. Additionally, the "out-of-dataset performance" of the learned decomposition was assessed by using it as dimensionality reduction of an independent labeled dataset (i.e., ARCHI) and conducting ordinary logistic regression on the ensuing component loadings.

	n = 5				*n = 20*				*n = 100*			
	$\lambda = 0.25$	$\lambda = 0.5$	$\lambda = 0.75$	$\lambda = 1$	$\lambda = 0.25$	$\lambda = 0.5$	$\lambda = 0.75$	$\lambda = 1$	$\lambda = 0.25$	$\lambda = 0.5$	$\lambda = 0.75$	$\lambda = 1$
Out-of-sample accuracy	*88,90%*	95,13%	**96,49%**	95,72%	97,44%	**97,80%**	97,33%	*97,32%*	97,21%	*97,03%*	**97,80%**	97,38%
Precision (mean)	*86,98%*	94,86%	**96,28%**	95,36%	**97,38%**	97,06%	*96,98%*	97,02%	96,90%	*96,47%*	**97,50%**	96,91%
Recall (mean)	*88,27%*	95,19%	**96,57%**	95,70%	**97,51%**	97,50%	*97,35%*	97,36%	97,20%	*97,15%*	**97,88%**	97,37%
F1 score (mean)	*86,64%*	94,89%	**96,38%**	95,43%	**97,42%**	97,22%	*97,11%*	97,13%	97,00%	*96,71%*	**97,65%**	97,07%
Out-of-dataset accuracy	60,83%	*54,30%*	60,69%	**62,92%**	79,72%	**81,94%**	79,72%	*79,44%*	**82,08%**	81,66%	81,25%	*75,83%*

Table 2: Performance of SS-LR LogReg across model parameter choices. XXX: recall what the chance-level performance is.XXX: please add lambda=0

We made several noteworthy observations (Table 2). First, the purely supervised estimator ($\lambda = 1$) achieved in no instance the best accuracy, precision, recall or f1 scores on HCP data. Classification by low-rank logistic regression is therefore facilitated by imposing structure from the unlabeled rest data. Second, the higher the number of latent components n, the higher the out-of-dataset performance with small choices for λ. This suggests that the presence of more rest-data-inspired hidden components results in more effective feature representations in unrelated task data. Third, for $n = 20$ and 100 (but not 5) the purely rest-data-trained decomposition matrix ($\lambda = 0$) resulted in noninferior out-of-dataset performance of X and Y, respectively (not shown in Table 2 XXX: please show it). This confirms that guiding model learning by task-unrelated structure extract features of general relevance beyond the supervised problem at hand.

Individual effects of dimensionality reduction and rest data. We first quantified the gain of introducing a bottleneck layer disregarding the autoencoder. To this end, ordinary logstic regression was juxtaposed with LR LogReg at $\lambda = 1$ (no autoencoder). For this experiment, we increased the difficulty of the classification problem by including data from all 38 HCP tasks. Indeed, increased class separability in component space, as compared to voxel space, entails differences in generalization perforamnce of $\approx 17\%$ (Figure 2).

We then quantified the impact of structure from rest data keeping all model parameters constant ($n = 20$, $\lambda = 0.5$, $\ell_1 = 0.1$, $\ell_2 = 0.1$). At the beginning of every epoch, 2000 task and 2000 rest maps were drawn with replacement from same amounts of task maps but varying amounts of rest maps. In frequently encountered data-scarce settings (≈ 100 samples), but not in data-rich settings (≈ 1000 samples), the repertoire of rest structure modulated model performance (Figure 3).

Feature identification. We finally examined whether the models were fit for purpose (Figure 4). To this end, we computed Pearson's correlation between the classifier weights and the averaged neural activity map for each of the 18 tasks. Ordinary logistic regression thus yielded a mean correlation of of $\rho = 0.28$. For SS-LR LogReg ($\lambda = 0.25, 0.5, 0.75, 1$), a per-class-weight map was computed by matrix multiplication of the two inner layers. Feature identification thus ranged between $\rho = 0.35$ and $\rho = 0.55$ for $n = 5$, between $\rho = 0.59$ and $\rho = 0.69$ for $n = 20$, and between $\rho = 0.58$ and $\rho = 0.69$ for $n = 100$. Consequently, SS-LR LogReg puts higher absolute weights on relevant structure. This reflect a higher signal-to-noise ratio, in part explained by the more BOLD-typical local contiguity. Conversely, SS-LR LogReg puts lower probability mass on irrelevant structure.

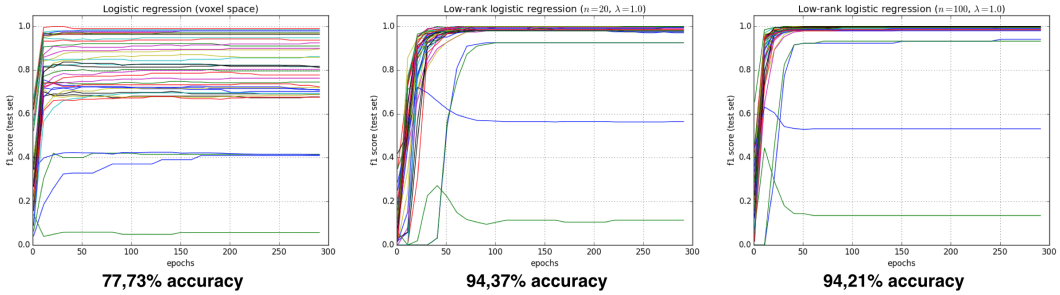

Figure 2: **Effect of bottleneck in a 38-task classificaton problem** Depicts f1 scores for prediction of each of 38 classes across training epochs. Ordinary logistic regression operating in voxel space (*left plot*) converged faster but performed worse than low-rank logistic regression for few (*middle plot*) and many (*right plot*) latent modes. Autoencoder or rest data were not used for these analyses ($\lambda = 1$). Hence, projecting the input data into a reduced space for classification yields higher class separability. XXX: ugly figure: use larger font for labels and titles. XXX: includes pdf figures, not a png file.

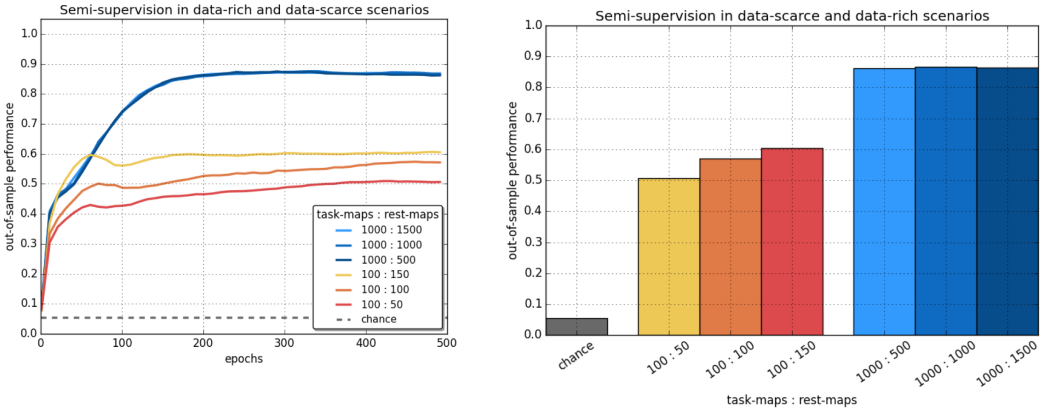

Figure 3: **Effect of richness in rest data** Gradient descent was performed on 2000 task and 2000 rest images. These were drawn with replacement from identical versus varying quantities of task versus rest maps, at the begining of each epoch.

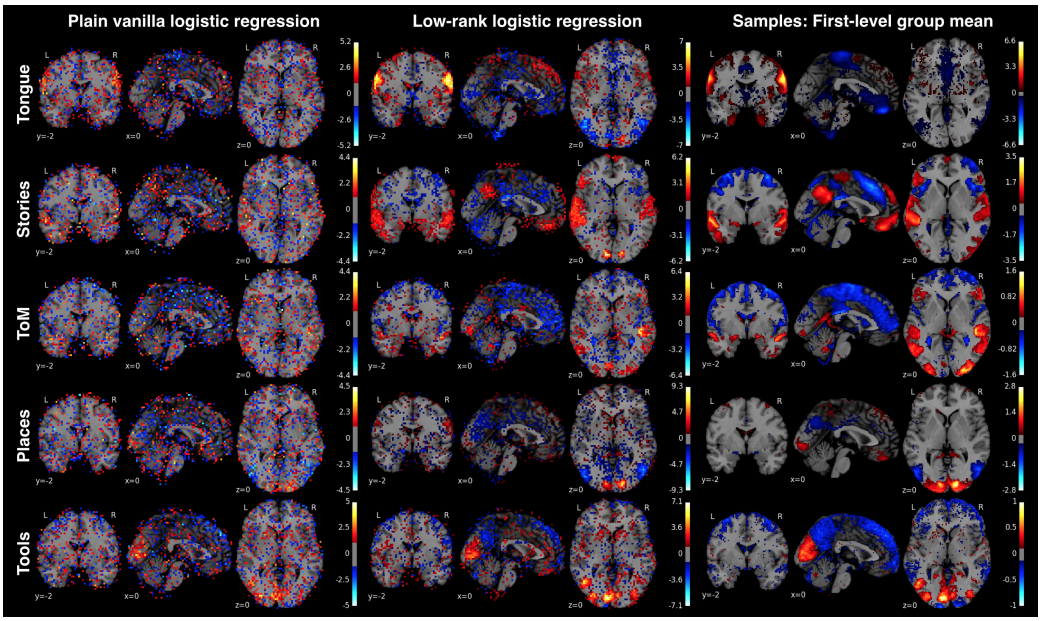

Figure 4: **Classification weight maps** The voxel predictors corresponding to 5 exemplary (of 18 total) psychological tasks (*rows*) from the HCP dataset [**?**]. *Left column:* ordinary logistic regression (same implementation but without bottleneck or autoencoder), XXX: sorry, but is this a multinomial or plain logistic ? Please clarify *middle column:* semi-supervised low-rank logistic regression ($n = 20$ latent components, $\lambda = 0.5$, $\ell_1 = 0.1$, $\ell_2 = 0.1$), *right column:* voxel-wise average of whole-brain activity maps from each task. Low-rank logistic regression a) puts 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 thresholded at the 75th percentile.

Despite lower interpretability, the salt-and-pepper-like weight maps obtained from ordinary logistic regression are sufficient for good classification performance. Hence, SS-LR LogReg yielded class weights that were much more similar to features of the respective training samples for any choice of λ. SS-LR LogReg therefore captures genuine properties of neural activity patterns, rather than participant- or study-specific artefacts. XXX: But what you show here is mostly a denoising or regularization effect.

Miscellaneous observations. For the sake of completeness, we informally report modifications of the model that did not improve the generalization performance. Introducing stochasticity into model learning by a) input corruption of $\mathbf{X_{task}}$ or b) drop-out at the bottleneck $\mathbf{W_0}$ using binomial distributions ($p = 0.1, 0.3, 0.5$) deteriorated model performance in all instances. Adding c) rectified linear units (ReLU) to $\mathbf{W_0}$ or other commonly used nonlinearities (d) sigmoid, e) softplus, f) hyperbolic tangent) all led to smaller classification accuracies. **XXX: Please give explanations why this is the case. My feeling is that we're still not rich enough in the sample dimension to make these tricks worthwhile.** Further, g) "pretraining" of the bottleneck $\mathbf{W_0}$ (i.e., non-random initialization) by either corresponding SPCA or ICA loadings did not exhibit improved accuracies, neither did h) autoencoder pretraining. Moreover, introducing an additional i) overcomplete layer (100 units) after the bottleneck was not advantageous. Finally, imposing imposing either j) only ℓ_1-penalty or k) only ℓ_2-penalty was disadvantageous in all tested cases, which favored ElasticNet regularization chosen in the above analyses.

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 feature 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 that correspond to peculiar brain networks combinations. 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 [**?**]. 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 [**?**] but also the multi-modal combination with high-resolution 3D models of brain anatomy [**?**] and high-throughput genome analyses [**?**]. The biggest socioeconomic potential may lie in across-hospital clinical studies that predict disease trajectories and drug responses in psychiatric/neurological populations [**?**, **?**].

Acknowledgment Data were provided by the Human Connectome Project. The study was supported by the German National Academic Foundation (D.B.). XXX: Note that this will be removed at submission for the sake of anonymization.