Brain Networks and Representational Similarity Analysis via Sparse Multitask Regression

Urvashi Oswal, Christopher Cox, Matthew A. Lambon Ralph, Timothy Rogers, and Robert Nowak, *Fellow IEEE*

***Abstract*—*Representational Similarity Analysis* (RSA) is a tool for discovering brain regions that encode representational similarities among stimuli. Existing RSA methods consider only localized networks, such as specific regions of interest or spherical volumes within cortex. In this paper we propose a new approach for *whole-brain* RSA that can discover arbitrarily structured brain networks (possibly widely distributed and non-local) that encode similarity information. We pose the RSA problem as a sparsity-regularized multi-task regression problem. This allows us to effectively search over all subsets of voxels (not just localized clusters) to detect similarity-encoding networks. A baseline approach for this regression task is the group lasso, but this approach may not select important voxels if they happen to be strongly correlated with other voxels. To address this shortcoming we present a new regularizer for multitask regression with strongly correlated covariates (voxels in fMRI applications) named *Group Ordered Weighted* ℓ_1 (GrOWL). Theoretically, we show that GrOWL automatically clusters and averages regression coefficients associated with strongly correlated variables. We apply the group lasso and GrOWL approach to whole-brain RSA, demonstrating and comparing our new approach in simulations and real-data experiments.**

I. Introduction

Network-based approaches to cognitive neuroscience typically assume that mental representations are encoded as distributed patterns of activation over large neural populations, with different populations encoding different kinds of representational structure and communicating this structure to other network components. Extensive research over the past several years has focused on testing such hypotheses using data from functional brain imaging techniques such as fMRI. The best-known approach in this vein has been *Representational Similarity Analysis* (RSA) [1], which seeks to discover brain regions whose activity encodes the known psychophysical similarities among some set of stimuli. RSA is typically applied either to a specific brain region of interest (ROI) or across the whole brain via *searchlight analysis* [2], which applied the technique to many small spherical clusters throughout the measured volume. For each such region the cosine distances between vectors of evoked responses are computed for all stimulus pairs, and the resulting neural dissimilarity matrix is correlated with the target matrix that expresses psychophysical distances of interest amongst the stimuli. If these correlations are reliably non-zero, this suggests the corresponding region may encode the similarity information.

A drawback of ROI and searchlight RSA is that these methods place strong assumptions on the anatomical structure of the regions thought to encode the similarities of interest (predefined ROIs or spherical clusters). In this paper we propose a new approach for *whole-brain* RSA that can discover arbitrarily structured brain networks (possibly widely distributed and non-local) that encode similarity information. The key insight behind our method is that RSA can be posed as a multi-task regression problem which, in conjunction with sparsity regularization methods, can automatically detect networks of voxels that appear to jointly encode similarity information.

Our new approach, called Network RSA (NRSA), is summarized as follows (see Sections IV and V for further details). Consider a set of n items and suppose we are given an $n \times n$ similarity matrix $\boldsymbol{S}$, where the ij-th element $\boldsymbol{S}_{ij}$ is the similarity [4] between item i and item j. For example, these may come from human judgments of perceptual similarity between pairs of stimuli. RSA is based on the hypothesis that there exists a set of voxels whose correlations encode the similarities in $\boldsymbol{S}$, as depicted in Figure 1.

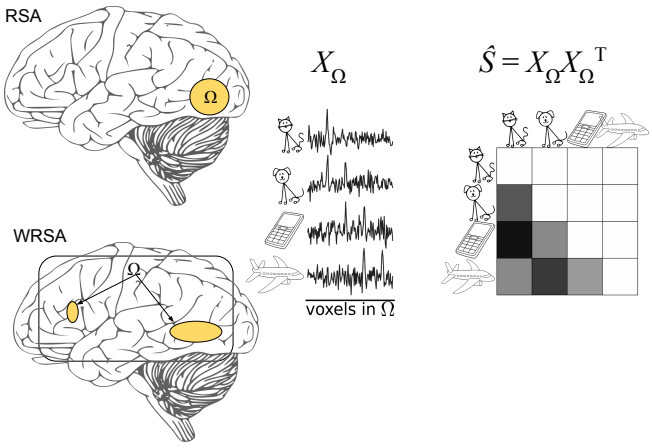


Fig. 1: Representational Similarity Analysis. Traditional RSA methods consider only localized brain networks, such as specific regions of interest or spherical clusters of the cortex (upper left) [1], [2]. We propose a new *Network* RSA (NRSA) method that can potentially identify non-local brain networks that encode similarity information (lower left). Within a set of voxels Ω (localized or non-local), the correlations between the activation patterns resulting from different stimuli approximate (perceptual) similarities between the stimuli.

U. Oswal and R. Nowak are with the Department of Electrical and Computer Engineering, University of Wisconsin-Madison, WI, 53706 USA. e-mail: uoswal@wisc.edu, nowak@ece.wisc.edu

C. Cox and T. Rogers are with the Department of Psychology, University of Wisconsin-Madison, WI, 53706 USA. e-mail: {crcox, ttrogers}@wisc.edu

M. Lambon Ralph is with the Neuroscience and Aphasia Research Unit (NARU), School of Psychological Sciences, University of Manchester, Manchester M13 9PL, UK. email: matt.lambon-ralph@manchester.ac.uk

Let $\boldsymbol{X} \in \mathbb{R}^{n\times p}$ denote a matrix of voxel activations. Each row corresponds activations in all p voxels in response to a specific stimulus, and each column corresponds to the activations in specific voxel to the n different stimuli. Our generalized notion of RSA, which encompasses conventional ROI [1] and searchlight [2] approaches, involves finding a sparse symmetric positive semi-definite matrix $\boldsymbol{W} \in \mathbb{R}^{p\times p}$ such that

$$\boldsymbol{S} \approx \boldsymbol{X}\boldsymbol{W}\boldsymbol{X}^T \ .$$

By sparse we mean that at most $k < p$ rows/columns of $\boldsymbol{W}$ are nonzero. The locations of the nonzero elements indicate which voxels are included in the similarity-encoding brain network, and the weights in $\boldsymbol{W}$ indicate the strength of the edges in the network. For instance, consider the $n \times 1$ activation vectors of two voxels $\boldsymbol{x}_k$ and $\boldsymbol{x}_\ell$ (i.e., the kth and ℓth columns of $\boldsymbol{X}$). It is easy to show that the contribution of these two voxels to the similarity representation is given by $W_{k,\ell}\,\boldsymbol{x}_k\boldsymbol{x}_\ell^T + W_{\ell,k}\,\boldsymbol{x}_\ell\boldsymbol{x}_k^T$. If $W_{k,\ell} = W_{\ell,k} \neq 0$, then the correlations between the two voxels contribute to the approximation of the similarity matrix $\boldsymbol{S}$. The complete similarity representation can be expressed as

$$\boldsymbol{S} \approx \boldsymbol{X}\boldsymbol{W}\boldsymbol{X}^T = \sum_{k,\ell=1}^{p} W_{k,\ell}\,\boldsymbol{x}_k\boldsymbol{x}_\ell^T + W_{\ell,k}\,\boldsymbol{x}_\ell\boldsymbol{x}_k^T \ .$$

The approximation problem can be posed as the least squares optimization

$$\min_{\boldsymbol{W}} \|\boldsymbol{S} - \boldsymbol{X}\boldsymbol{W}\boldsymbol{X}^T\|_F^2 \ ,$$

where the objective is the Frobenius norm of the difference between the similarity matrix $\boldsymbol{S}$ and its approximation in terms of voxel activations. If the number of voxels p exceeds the number of items n (which is usually the case in whole-brain RSA), then the system of equations is underdetermined and there will be many solutions to the optimization above. Moreover, the hypothesis underlying RSA is that a small subset of the brain encodes the similarity representations. To account for this, the optimizations can be modified by including regularizer terms that encourage sparse solutions, as discussed in the next section.

II. Learning Similarity Encodings via Group Lasso

Throughout the paper, we will assume that the similarity matrix $\boldsymbol{S}$ is symmetric and positive-semidefinite (PSD) and is exactly or approximately low-rank. Since $\boldsymbol{S}$ is known, its low-rank structure can be determined from its singular value decomposition. Similarity matrices are often low-rank because of clustering or other relationships between the items under consideration. For example, in our experimental application described later in the paper, we find that a rank $r = 3$ approximation is quite accurate. Since we suppose that the given similarity matrix $\boldsymbol{S}$ may be (approximately) low-rank, this suggests the convex optimization

$$\min_{\boldsymbol{W}\in\mathcal{S}_+^p} \|\boldsymbol{S} - \boldsymbol{X}\boldsymbol{W}\boldsymbol{X}^T\|_F^2 \ + \ \lambda_1\|\boldsymbol{W}\|_* + \lambda_2\|\boldsymbol{W}\|_1 \ ,$$

where $\mathcal{S}_+^p$ denotes the set of symmetric PSD $p \times p$ matrices, $\|\boldsymbol{W}\|_*$ is the nuclear norm of $\boldsymbol{W}$, $\|\boldsymbol{W}\|_1$ denotes the ℓ_1 norm of the elements in $\boldsymbol{W}$, and $\lambda_1, \lambda_2 > 0$ are regularization parameters. The nuclear norm encourages low-rank solutions and the ℓ_1 term promotes solutions that include a small subset of the brain voxels. This type of optimization has been studied extensively [19].

Since $\boldsymbol{S} \in \mathcal{S}_+^n$, we can work with the "square-root" of $\boldsymbol{S}$ instead (e.g., its Cholesky decomposition). Suppose that $\boldsymbol{S}$ has rank $r < n$ and let $\boldsymbol{Y} \in \mathbb{R}^{n\times r}$ satisfy $\boldsymbol{Y}\boldsymbol{Y}^T = \boldsymbol{S}$ (or take $\boldsymbol{Y}$ corresponding to the best rank-r approximation to $\boldsymbol{S}$, given by the SVD). Consider the optimization

$$\min_{\boldsymbol{B}\in\mathbb{R}^{p\times r}} \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{B}\|_F^2 \ . \tag{1}$$

This optimization has the form of a multi-task regression [6], [10], [11], and a solution $\widehat{\boldsymbol{B}}$ yields a weight matrix $\widehat{\boldsymbol{W}} = \widehat{\boldsymbol{B}}\widehat{\boldsymbol{B}}^T$. Note that this optimization automatically enforces a rank r solution, eliminating the need for the nuclear norm term above. Although $\widehat{\boldsymbol{W}}$ is not exactly a solution to the optimization $\min_{\boldsymbol{W}} \|\boldsymbol{S} - \boldsymbol{X}\boldsymbol{W}\boldsymbol{X}^T\|_F^2$, it is easy to relate the two optimizations as follows. Let $\widetilde{\boldsymbol{W}}$ be a solution to $\min_{\boldsymbol{W}} \|\boldsymbol{S} - \boldsymbol{X}\boldsymbol{W}\boldsymbol{X}^T\|_F^2$ and let $\widetilde{\boldsymbol{B}}\widetilde{\boldsymbol{B}}^T$ be its Cholesky factorization. Assuming the normalization $\|\widehat{\boldsymbol{B}}\|_F, \|\widetilde{\boldsymbol{B}}\|_F \leq 1$, we have

$$\begin{aligned}
\|\widetilde{\boldsymbol{W}} - \widehat{\boldsymbol{W}}\|_F &= \|\widetilde{\boldsymbol{B}}\widetilde{\boldsymbol{B}}^T - \widehat{\boldsymbol{B}}\widehat{\boldsymbol{B}}^T\|_F \\
&\leq \|\widetilde{\boldsymbol{B}} + \widehat{\boldsymbol{B}}\|_F\|\widetilde{\boldsymbol{B}} - \widehat{\boldsymbol{B}}\|_F \\
&\leq (\|\widetilde{\boldsymbol{B}}\|_F + \|\widehat{\boldsymbol{B}}\|_F)\|\widetilde{\boldsymbol{B}} - \widehat{\boldsymbol{B}}\|_F \\
&\leq 2\|\widetilde{\boldsymbol{B}} - \widehat{\boldsymbol{B}}\|_F \ .
\end{aligned}$$

The optimization $\min_{\boldsymbol{B}\in\mathbb{R}^{p\times r}} \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{B}\|_F^2$ can be modified to encourage sparse solutions by including a regularization term. The most common approach to promote sparsity in multi-task regression is the group lasso:

$$\min_{\boldsymbol{B}\in\mathbb{R}^{p\times r}} \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{B}\|_F^2 \ + \ \lambda\|\boldsymbol{B}\|_{1,2} \tag{2}$$

Here $\lambda > 0$ is a regularization parameter and $\|\boldsymbol{B}\|_{1,2}$ is defined as follows. The rows of $\boldsymbol{B}$ are denoted by $\boldsymbol{\beta}_{i*}$, $i = 1, \ldots, n$, and the norm $\|\boldsymbol{B}\|_{1,2} = \sum_{i=1}^n \|\boldsymbol{\beta}_{i*}\|_2$. This encourages solutions with only a few nonzero rows in $\boldsymbol{B}$ [6], [10], [11].

The main signal processing innovation in this paper is a new approach to the group lasso that is designed to cope with strongly correlated covariates (i.e., cases in which certain columns of $\boldsymbol{X}$ may be close to, or even exactly, collinear). This is a concern in fMRI, since certain voxels may have very correlated activation patterns. In the standard (single-task) regression problem, this issue has been tackled using many techniques, including the elastic net [9], OSCAR [7] and OWL [15], and others. We propose a generalization of the recently proposed Ordered Weighted ℓ_1 (OWL) approach to the multi-task setting, and thus call our new approach Group OWL (GrOWL).

We show that GrOWL shares many of the desirable features of the OWL method, namely it automatically clusters and averages regression coefficients associated with strongly correlated columns of $\boldsymbol{X}$. This has two desirable effects, in terms of both model selection and prediction. First, GrOWL can select all of the relevant voxels in $\boldsymbol{X}$, unlike standard group lasso

which may not select relevant voxels if they happen to be strongly correlated with others. Second, GrOWL encourages the coefficients associated with strongly correlated voxels to be near or exactly equal. In effect, this averages strongly correlated voxelss which can help to denoise activation patterns and improve predictions.

III. GrOWL

Here we discuss modifications of the group lasso in order to deal with strongly correlated columns in $\boldsymbol{X}$. Our approach is motivated by the recently proposed OWL [15] norm, a special case of which is the so-called OSCAR [7]. These methods are designed to automatically cluster and effectively average highly correlated columns in the data matrix, and have been shown to outperform conventional lasso in many applications, particularly in cases of strong correlations. Both OWL and OSCAR deal only with the single regression setting. The main innovation here is the development of new norms, in the spirit of OWL, that allow us to deal with correlated columns in the multiple regression / multitask setting. We present two variants of the GrOWL (group OWL), and show that they automatically group and average highly correlated columns in $\boldsymbol{X}$ in the multiple regression setting.

In this section, we consider the general optimization

$$\min_{\boldsymbol{B}\in\mathbb{R}^{p\times r}} \; L(\boldsymbol{B}) \; + \; G(\boldsymbol{B}) \tag{3}$$

where typical loss functions considered here are absolute error, $L(\boldsymbol{B}) = \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{B}\|_1$, or squared Frobenius error, $L(\boldsymbol{B}) = \|\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{B}\|_F^2$, and $G(\boldsymbol{B})$ is the GrOWL norm defined later in the section. The following results can be extended to the solution of the optimization with squared Frobenius norm loss but, for the sake of simplicity, we consider the absolute error loss in this section (details for extending the theory to Frobenius norm loss are presented in the Appendix). We give proof sketches for the main theorems and leave proofs of the theorems to the Appendix.

A. GrOWL penalty

Let $\boldsymbol{B} \in \mathbb{R}^{p\times r}$ and let $\boldsymbol{\beta}_{i*}$ and $\boldsymbol{\beta}_{*j}$ denote the ith row and jth column of $\boldsymbol{B}$. Define the GrOWL penalty

$$G(\boldsymbol{B}) \; = \sum_{i=1}^{p} w_i \|\boldsymbol{\beta}_{[i]*}\|_2, \tag{4}$$

where $\boldsymbol{\beta}_{[i]*}$ is the row of $\boldsymbol{B}$ with the i-th largest 2-norm and $\boldsymbol{w}$ is a vector of non-negative and non-increasing weights. Before we analyze the GrOWL regularization, we state a generalization of Lemma 2.1 in [15] which will be useful later in the section.

Lemma III.1. *Consider a vector $\boldsymbol{\beta} \in \mathbb{R}^p_+$ and any two of its components β_j and β_k, such that $\beta_j > \beta_k$. Let $\boldsymbol{v} \in \mathbb{R}^p_+$ be obtained by applying a transfer of size $\varepsilon, \varepsilon'$ to $\boldsymbol{\beta}$ such that $\varepsilon \in (0, (\beta_j - \beta_k)/2]$ and $-\beta_k \le \varepsilon' \le \varepsilon$, that is: $v_j = \beta_j - \varepsilon, v_k = \beta_k + \varepsilon'$, and $v_i = \beta_i$, for $i \neq j, k$. Let $\boldsymbol{w}$ be a vector of non-increasing non-negative real values, $w_1 \ge w_2 \ge \cdots \ge w_p \ge 0$, and Δ be the minimum gap between two consecutive components of vector $\boldsymbol{w}$, that is, $\Delta = \min\{w_i - w_{i+1}, i = 1, \cdots, p-1\}$. $\Omega_{\boldsymbol{w}}(\cdot)$ is the OWL norm with weight vector $\boldsymbol{w}$, then*

$$\Omega_{\boldsymbol{w}}(\boldsymbol{\beta}) - \Omega_{\boldsymbol{w}}(\boldsymbol{v}) \; \ge \; \Delta\varepsilon$$

Proof. The proof is similar to that of Lemma 2.1 in [15] with different sizes $\varepsilon, \varepsilon'$ and the result follows because we assume that the increase in k-th component is less than the decrease in j-th component *i.e.,* $\varepsilon' \le \varepsilon$.
More intuitively, if β_k doesn't go up by ε in magnitude, then increase its magnitude so that it does and call this $\boldsymbol{v}'$ with $v'_j = \beta_j - \varepsilon$ and $v'_k = \beta_k + \varepsilon$. Then we apply Lemma 2.1 in [15] to $\boldsymbol{v}'$ and $\Omega_{\boldsymbol{w}}(\boldsymbol{v}) \le \Omega_{\boldsymbol{w}}(\boldsymbol{v}')$. □

The following theorem states that identical variables lead to equal coefficient rows corresponding to those variables in the solution given by the optimization using GrOWL.

Theorem III.1 (Identical columns)**.** *Let $\widehat{\boldsymbol{B}}$ denote the solution to the optimization in (3) with $L(\boldsymbol{B}) = \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{B}\|_1$ or $L(\boldsymbol{B}) = \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{B}\|_F^2$. If columns $\boldsymbol{x}_{*j}$ and $\boldsymbol{x}_{*k}$ satisfy $\boldsymbol{x}_{*j} = \boldsymbol{x}_{*k}$ and the minimum gap, $\Delta > 0$, then $\widehat{\boldsymbol{\beta}}_{j*} = \widehat{\boldsymbol{\beta}}_{k*}$.*

Proof sketch. The proof is divided into two steps. First, we show $\|\widehat{\boldsymbol{\beta}}_{j*}\| = \|\widehat{\boldsymbol{\beta}}_{k*}\|$ and then we further show that the rows are equal. We proceed by contradiction. Assume $\|\widehat{\boldsymbol{\beta}}_{j*}\| \neq \|\widehat{\boldsymbol{\beta}}_{k*}\|$ and, without loss of generality, suppose $\|\widehat{\boldsymbol{\beta}}_{j*}\| > \|\widehat{\boldsymbol{\beta}}_{k*}\|$. We see that there exists a modification of the solution with a smaller GrOWL norm using Lemma III.1 and same data-fitting term, and thus smaller overall objective value which contradicts our assumption that $\widehat{\boldsymbol{B}}$ is the minimizer of $L(\boldsymbol{B}) + G(\boldsymbol{B})$.

The following theorem states that nearly identical variables lead to equal norm coefficient rows corresponding to those variables in the solution given by the optimization using GrOWL.

Theorem III.2 (Correlated columns 1)**.** *Let $\widehat{\boldsymbol{B}}$ denote the solution to the optimization in (3) with $L(\boldsymbol{B}) = \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{B}\|_1$. If $\boldsymbol{x}_{*j}$ and $\boldsymbol{x}_{*k}$ satisfy $\|\boldsymbol{x}_{*j} - \boldsymbol{x}_{*k}\|_1 \le \frac{\Delta}{\sqrt{r}}$, then $\|\widehat{\boldsymbol{\beta}}_{j*}\| = \|\widehat{\boldsymbol{\beta}}_{k*}\|$.*

Proof sketch. The proof is similar to the identical columns theorem. By contradiction and without loss of generality, suppose $\|\widehat{\boldsymbol{\beta}}_{j*}\| > \|\widehat{\boldsymbol{\beta}}_{k*}\|$. We show that there exists a transformation of $\widehat{\boldsymbol{B}}$ such that the increase in the data fitting term is smaller than the decrease in the GrOWL norm.

The following theorem states that nearly identical variables lead to highly correlated coefficient rows corresponding to those variables in the solution given by the optimization using GrOWL.

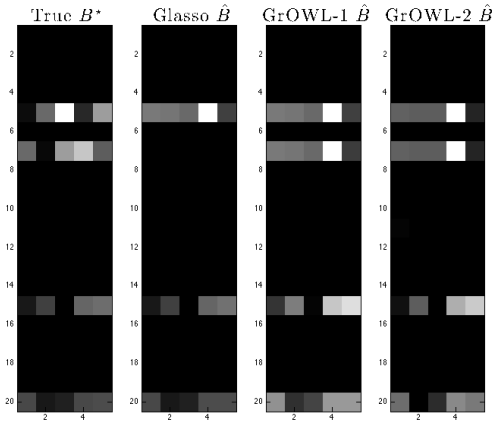
Theorem III.3 (Correlated columns 2)**.** *Let $\widehat{\boldsymbol{B}}$ denote the solution to the optimization in (3) with $L(\boldsymbol{B}) = \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{B}\|_1$. If $\boldsymbol{x}_{*j}$ and $\boldsymbol{x}_{*k}$ satisfy $\|\boldsymbol{x}_{*j} - \boldsymbol{x}_{*k}\|_1 \le \frac{\Delta}{\phi\sqrt{r}}$, then $\|\widehat{\boldsymbol{\beta}}_{j*} - \widehat{\boldsymbol{\beta}}_{k*}\| \le \frac{8\phi\|\widehat{\boldsymbol{\beta}}_{k*}\|}{4\phi^2+1}$*
which further implies that

$$1 \ge \frac{\widehat{\boldsymbol{\beta}}_{j*}^T \widehat{\boldsymbol{\beta}}_{k*}}{\|\widehat{\boldsymbol{\beta}}_{j*}\|\|\widehat{\boldsymbol{\beta}}_{k*}\|} \ge 1 - \frac{1}{2}\left(\frac{8\phi}{4\phi^2+1}\right)^2 \left(\ge 1 - \frac{2}{\phi^2}\right)$$

where $\phi \geq 1$.

Proof sketch. The proof is similar to the identical columns theorem. By contradiction, suppose $\|\widehat{\boldsymbol{\beta}}_{j*} - \widehat{\boldsymbol{\beta}}_{k*}\| \geq \frac{8\phi\|\widehat{\boldsymbol{\beta}}_{k*}\|}{4\phi^2+1} \geq \frac{2\|\widehat{\boldsymbol{\beta}}_{k*}\|}{\phi}$. We show that there exists a transformation of $\widehat{\boldsymbol{B}}$ such that the increase in the data fitting term is smaller than the decrease in the GrOWL norm. This contradicts our assumption that $\widehat{\boldsymbol{B}}$ is the minimizer of $L(\boldsymbol{B}) + G(\boldsymbol{B})$ and completes the proof.

So far, we have seen that the GrOWL penalty has desirable clustering properties that lead to nearly identical coefficient rows. We study two variants of GrOWL with different weight sequences $\boldsymbol{w}$. First, we study the weights with linear decay (equivalent to the OSCAR in single-task regression) and call it GrOWL-I. Next, we study the L1 + Linf weight sequence and call it GrOWL-II (see Figure 2).



GrOWL-I:	$w_i = \lambda(p-i)$ for $i = 1, \ldots, p$
GrOWL-II:	$w_1 = \lambda_1 + \lambda_2, w_i = \lambda_1$ for $i = 2, \ldots, p$

Fig. 2: A comparison of group lasso and grOWL optimization solutions with correlated columns in $\boldsymbol{X}$ showing that GrOWL-I and GrOWL-II select relevant features (row 5 and 7) even if they happen to be strongly correlated and automatically cluster them by setting the corresponding coefficient rows to be equal (or nearly equal).

B. Proximal algorithms

We present computational algorithms for the two variants of the GrOWL norm here. The algorithms rely on the computation of the proximity operator [16] of the grOWL norm given by

$$\text{prox}_G(\boldsymbol{V}) = \arg\min_{\boldsymbol{B}} \frac{1}{2}\|\boldsymbol{B} - \boldsymbol{V}\|_F^2 + G(\boldsymbol{B}) \tag{5}$$

In the following theorem, we solve for the proximity operator of grOWL in terms of the proximity of OWL. For the exact formulation of $\text{prox}_{\Omega_{\boldsymbol{w}}}$, see [12], [21].

Theorem III.4. *Let* $\tilde{v}_i = \|\boldsymbol{v}_{i*}\|$ *for* $i = 1, \cdots, p$. *Then* $\text{prox}_G(\boldsymbol{V}) = \widehat{\boldsymbol{V}}$, *where* i*-th row of* $\widehat{\boldsymbol{V}}$ *is*

$$\widehat{\boldsymbol{v}}_{i*} = (\text{prox}_{\Omega_{\boldsymbol{w}}}(\tilde{\boldsymbol{v}}))_i \frac{\boldsymbol{v}_{i*}}{\|\boldsymbol{v}_{i*}\|} \tag{6}$$

Proof Sketch: The proof proceeds by finding a lower bound for the objective function in (5) and then we show that the proposed solution achieves this lower bound.

IV. Network RSA: Simulated Data

In this section we illustrate comparative properties of group lasso, GrOWL-I and GrOWL-II by applying these to the analysis of synthetic data generated from a simple neural network motivated by the triangle model of word-reading (Figure 3 top left). This feed-forward network takes a model analog of word spelling as input (orthographic layer) and is trained to generate distributed representations of its meaning (semantic layer) and pronunciation (phonological layer). The network is deep in that mappings from orthography to phonology, from orthography to semantics, and from semantics to phonology, are all mediated by one or more hidden layers of units. The model's ability to generate phonological output is mediated by two separate pathways: a *direct* route mediated by a single hidden layer, and an "indirect" route composed of three hidden layers, which must first compute mappings from orthography to semantics, then project onward to contribute to the phonological output units.

On each learning trial the orthographic pattern corresponding to a particular word is clamped over input units. Each unit's net input is then computed as the dot product of the activations of connected sending units and the values of interconnecting weights. Activation values are then taken as a sigmoid function of the unit's net input. Output activations generated across semantic and phonological units are compared to target values for the corresponding words, and the model is trained with gradient descent to reduce the squared error. Training proceeds until all semantic and phonological output units are within 0.1 of their target values for all patterns.

The triangle model provides an interesting test case for discovery of representational similarity structure, because different kinds of structure emerge through learning in different network components. The central idea is that orthographic and phonological similarities are highly systematic: items that are similar in spelling are likely (though not guaranteed) to be similar in pronunciation. These regularities are easily learned within the direct pathway mapping from orthography to phonology, allowing the system to generate appropriate pronunciations for previously unseen word-forms. In contrast, the relationship between orthographic and semantic similarity structure is unsystematic: similarity of word spelling does not necessarily predict similarity of meaning. Thus other pathways within the network, in learning to map from orthography to semantics and from semantics to phonology, come to encode different similarity relations amongst the words [5], [14].

To capture these properties of the triangle framework, we generated model "orthographic" word representations in which individual patterns were sampled from 6 overlapping clusters of binary input features, roughly corresponding to different orthographic neighborhoods. For each such "word," a corresponding "phonological" pattern was generated by flipping each binary feature from the orthographic pattern with probability 0.1. Thus phonological patterns were distorted variants of the orthographic patterns, creating high system-

aticity between these. Finally, for each word we also created a semantic pattern by generating a set of binary semantic features also organized into clusters. Across items, these vectors expressed a hierarchical similarity structure with two broad superordinate clusters each composed of three tighter clusters. Importantly, the similarity structure expressed by the semantic vectors was independent of the structure expressed in the orthographic/phonological patterns.

The top right panel in Figure 3 shows, for each layer of one trained model, the cosine distances encoded amongst the 30 model "words." Just from inspection it is clear that units in the direct pathway from orthography to phonology all encode roughly the same distances amongst items, reflecting the high systematicity between orthographic and phonological similariy structure. The semantic layer encodes a very different set of distances amongst the items while, again from inspection, two of the three hidden layers in the indirect pathway encode weaker versions of this structure. Finally the first hidden layer between orthography and semantics appears to encode a blend of the orthographic and semantic distances. Thus the different components of this simple word-reading network contribute differentially to the encoding of semantic versus ortho-phonological similarity structure.

To create synthetic "brain imaging data" we trained 5 models with different initial weights, corresponding to 5 model subjects. We then presented each trained model with all 30 orthographic inputs and generated a vector of unit activations for each input over the 100 model units. To ensure a high degree of redundancy within our synthetic dataset, this vector was next concatenated 5 times and then perturbed with independent noise, yielding measurements from 500 model "voxels" in each of 5 different "subjects." We then applied group lasso and GrOWL to find the voxel subsets that encode the targeted semantic or phonological distances (derived from the target values for the semantic and phonological output layers of the network).

We fit statistical models by searching a two-dimensional grid of parameters (λ , λ_1), including $\lambda_1 = 0$ as the special case of GrOWL that is group lasso. At each grid point we computed the number of "voxels" selected (*i.e.,* having non-zero weights). We assessed how well each fitted model identified the "voxels" that encode phonological structure (all those along the direct pathway) and those that encode semantic structure (the semantic layer itself and the middle layer and third hidden layer in the indirect pathway) by computing hit rates and false alarm rates. Figure 4 shows these data for group lasso, GrOWL-I and GrOWL-II. All three models show relatively low and equivalent cross-validation error; however GrOWL-II achieves this error rate while selecting considerably more voxels. The ROC plots in panel 3 further show that GrOWL-II is not just selecting additional voxels at random: its ability to discriminate signal-carrying from non-signal carrying voxels outstrips the group lasso considerably.

The bottom right and middle panels of Figure 3 show the frequency with which each model unit is selected for the best-performing solution in group lasso and GrOWL, in decoding phonological and semantic similarity. While each method hones in on approximately the correct subset of network units, the strong sparsity enforced by group lasso is clearly apparent: target units are much less consistently selected. GrOWL, in contrast, consistently discovers much more of the signal.

Finally, we considered the ability of GrOWL to reveal the network structure encoding each kind of similarity, treating the weights in the matrix $\boldsymbol{W}$ as direct estimates of the joint participation of pairs of units in expressing the target similarity. The bottom-right panel of Figure 3 shows the estimated connectivity, thresholded to show the 25% of the non-zero weights with the largest magnitudes. The detected edges clearly express the network representational substructure: units in the direct pathway are shown as highly interconnected with one another and weakly or disconnected from those in the indirect pathway, and vice versa. Thus the search for different kinds of similarity reveals different functional subnetworks in the model.

V. Network RSA: Real Data

We next consider the application of group lasso and GrOWL to the discovery of similarity structure in neural responses measured by fMRI across the whole brain while participants perform a cognitive task. As with the well-known searchlight RSA ([1], [4]), we begin with a measurement of the $n \times n$ similarities existing amongst a set of n items in some cognitive domain. Using fMRI, we measure the neural responses evoked by each item at the scale of single voxels (3mm cubes), and treat these d voxels as features of the n items. We then compute a rank-r approximation of the target similarity matrix $\boldsymbol{S} = \boldsymbol{Y}\boldsymbol{Y}^T$, and use this as the target $\boldsymbol{Y} \in \mathbb{R}^{n \times r}$ matrix for a sparse-regression analysis of the $n \times d$ matrix of fMRI responses , $\boldsymbol{X}$, evoked by each item across the whole cortex. The model is then fit to optimize the object functions specified in (2) for group lasso and (3) for GrOWL. The best regularization parameter is selected through cross-validation, and a final model is fit with that parameter and used to predict the similarities existing amongst a set of items in an independent hold-out set. Model predictions are compared to expected results expected from a null hypothesis that no features encode the target similarity structure. If predictions are more accurate than expected from random data, this provides evidence that the model has discovered voxel subsets that jointly encode some of the target similarity structure. Moreover, because the model is constrained to be sparse, most voxels will receive coefficients of zero, and the presence of non-zero coefficients can be taken as evidence that the corresponding voxel encodes information important to representing the target similarity structure.

The current experiment serves as a proof-of-concept with the aim of answering three questions. First, does either approach work, in the sense of generating above-chance predictions of the similarities existing amongst a set of stimuli given the neural responses they generate across the whole brain? Second, does one approach work better than the other in generating such predictions? Third, does the approach generate solutions that are consistent with what is known about representation of information in the brain?

To answer these questions, we applied the approach to discover voxels that work to encode the visual similarities

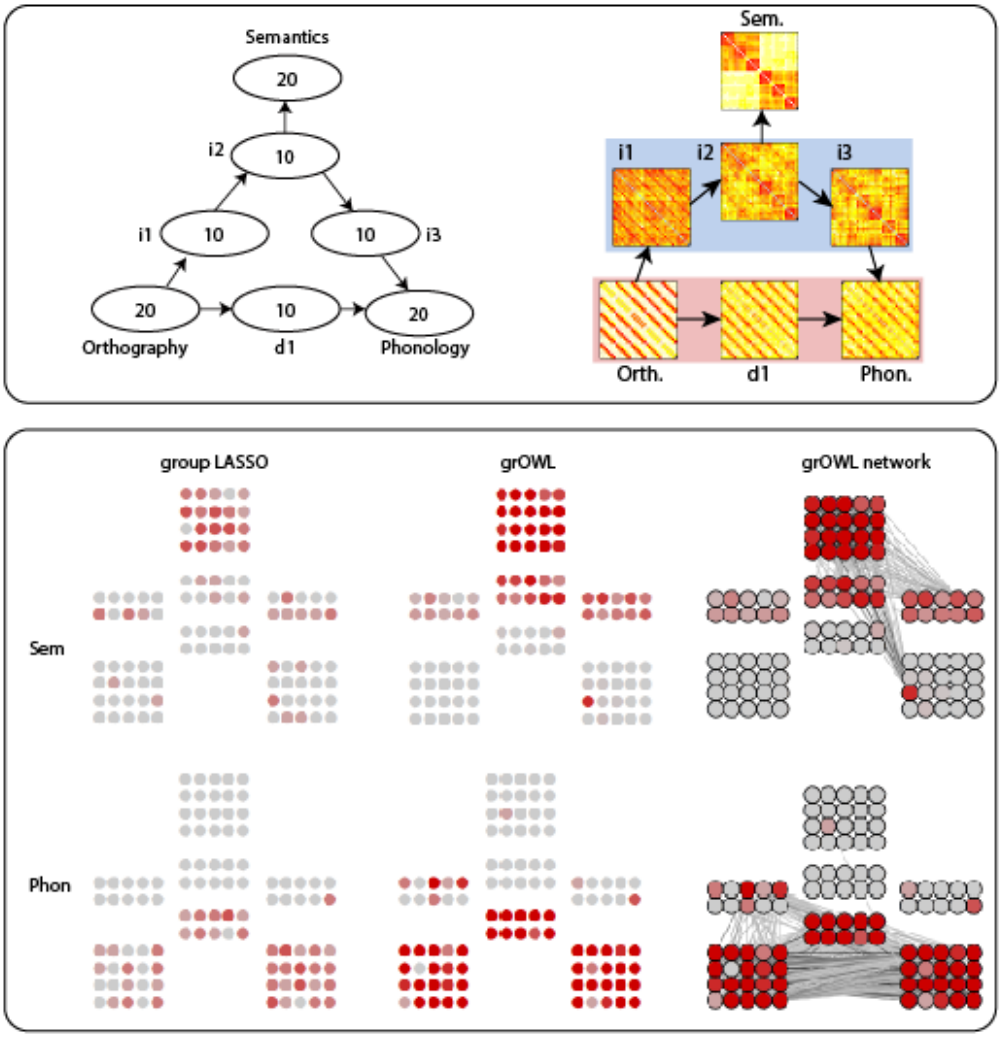


Fig. 3: Top panel: Network architecture (left) and the similarity structure expressed in each layer (right). Red background shows the direct pathway and blue the indirect pathway from orthography to phonology. Layers in the two pathways encode different similarity structures. The target similarity matrices for the analysis express either the semantic structure (top layer) or the phonological structure (bottom right layer). Arrows indicate feed-forward connectivity. Bottom panel: Units selected by group LASSO (right) and GrOWL (middle) when decoding semantic (top) or phonological (bottom) structure. Colors show the proportion of times across subjects and unit concatenations that the unit received a non-zero weight, with red indicating 1 and gray 0. The rightmost plots show the largest weights in the associated matrix W for each GrOWL model, which pick out two subnetworks in the model.

existing amongst a set of line drawings of common objects. We chose this task and dataset because (a) there exist well-understood methods for objectively measuring the degree of visual similarity amongst such items [18] and (b) it is well known that visual similarity is encoded by neural responses in occipital and posterior temporal cortices.

A. fMRI dataset

The data were collected as part of a larger study from 23 participants at the University of Manchester who were compensated for their time. Each participant viewed a series of line drawings depicting common objects while their brains were scanned with fMRI. The line drawings included 37 items, each repeated 4 times for a total of 148 unique stimulus events. At each trial participants pressed a button to indicate whether the item could fit in a "wheely bin" (a form of trash can common in the UK). Scans were collected in a sparse event-related design and underwent standard pre-processing to align functional images to the anatomy and to remove movement and scanner artifact and temporal drift. Responses to each stimulus event were estimated at each voxel using a deconvolution procedure with a standard HRF kernel. For each participant a cortical surface mask was generated based on T1-weighted anatomical images, and functional data were filtered to exclude non-cortical voxels. Voxels with estimated responses more than 5 standard deviations from the mean response across voxels were excluded from the analysis. 10k-15k voxels were selected for each participant, and neural responses across all voxels for each of 148 stimulus events were entered into the analysis. The mean response across the 4 repeated observations of each item were taken to give 37 item responses for each participant. Each column corresponding to

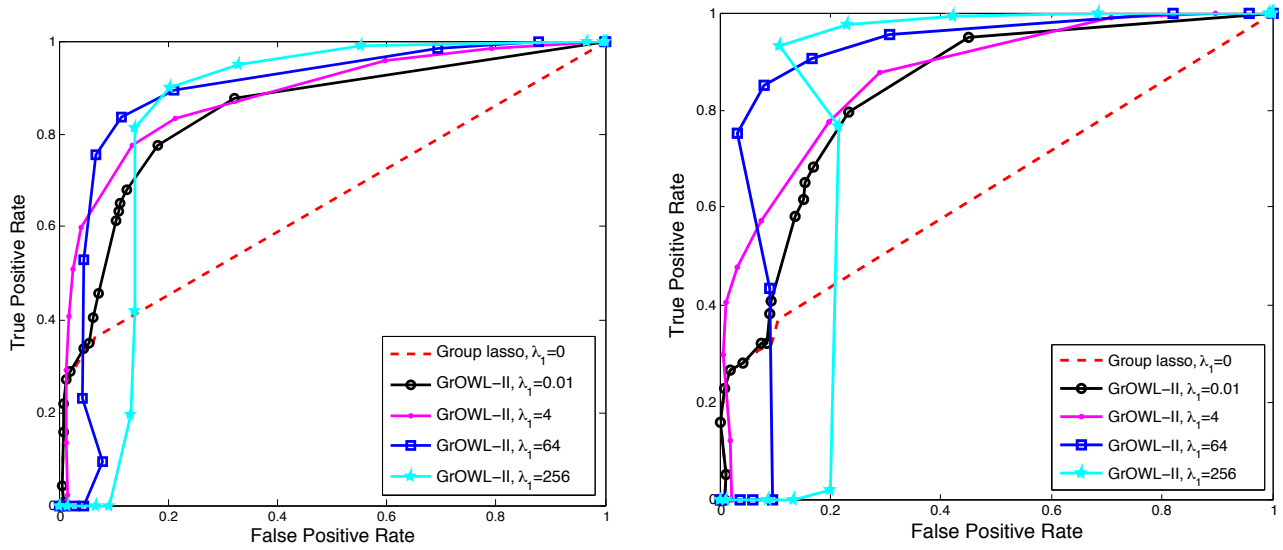


Fig. 4: ROC curves generated by sweeping through λ values (for $\lambda = 0$, all units are selected and as λ is increased fewer units are given non-zero weight). Each curve represents a fixed value of λ_1, where the curve $\lambda_1 = 0$ corresponds to group lasso. ROC curves are averaged across participants for each method, considering both similarity structures, Semantics (left panel) and Phonology (right panel).

a voxel was normalized to be of standard deviation equal to one and a column of ones was added.

Target similarities: Each stimulus was a bitmap of a black-and-white line drawing. We took pairwise Chamfer distance as a proxy for inter-item visual dissimilarities. r = 3 is the smallest value to attain $\|\boldsymbol{S} - \boldsymbol{Y}\boldsymbol{Y}^T\|_F \leq 0.2$. This 37×3 matrix $\boldsymbol{Y}$ was used as the target matrix for the analysis.

Model fitting: For each participant, training data were divided into 9 subsets containing 4-5 stimulus events each. One subset was selected as a final hold-out set. Models were then fit at each of 10 increasing values of λ using 8-fold cross validation. At each fold we assessed the model using the Frobenius norm of the difference between the target $\boldsymbol{Y}$ entries and the predicted $\widehat{\boldsymbol{Y}} = \boldsymbol{X}\widehat{\boldsymbol{B}}$ entries for hold-out items (henceforth the model error). We selected the λ with the lowest mean error for each subject subjects, then fit a full model for each subject at this value and assessed it against the final hold-out set, considering the model error on hold-out items. We repeat this process for 9 different final hold-out sets.

B. Results

Figure 5(a) shows performance on the final hold-out sets for each participant and each method, considering error between predicted ($\boldsymbol{Y}_z$) and actual dissimilarities ($\boldsymbol{Y}$). Both approaches show significantly non-random prediction. As in our simulations, the GrOWL-II shows somewhat better performance (lower error, higher correlation) though all methods show comparable prediction error. We also note that, as in the simulations, GrOWL-II selected almost double the number of voxels in each participant.

Figure 5(b) shows the locations of selected voxels (i.e., those with non-zero coefficients) across all 23 participants, mapped into a common anatomical space with 4mm full-width-half-max spatial smoothing and projected onto a model of the cortical surface. The left column shows the voxels selected for *at least five* (out of nine) cross-validation runs while the right column shows the voxels selected for *all* the nine cross-validation runs. As seen in the maps, both methods pick voxels prominently in the occipital and posterior temporal cortices and GrOWL-II picks consistently more voxels than group lasso.

Finally, Figure 6 shows the largest magnitude edges in the $\boldsymbol{W}$ matrix for the best-performing parameterization of group LASSO (top) and GrOWL-II (bottom) in one subject. Two observations are of note. First, both methods uncover a similar network structure, with many interconnections in visual cortical regions and some edges connecting to anterior regions in frontal and temporal cortex. Second, as in the simulations, GrOWL-II reveals a much denser network. The results suggest the possibility that subregions of frontal and temporal cortex may, together with occipeto-temporal cortex, participate in networks that serve to encode visual similarity structure.

VI. Conclusion

We have developed and demonstrated a new approach for whole-brain Representational Similarity Analysis called Network RSA (NRSA). Unlike traditional RSA methods that consider only specific regions of interest or spherical clusters of the cortex, NRSA can discover arbitrarily structured brain networks (possibly widely distributed and non-local) that encode similarity information. NRSA is posed as a sparsity-regularized multi-task regression problem. This allows us to effectively search over all subsets of voxels (not just localized clusters) to detect similarity-encoding networks. We further proposed a new sparsity regularizer for multi-task regression, the GrOWL, that is able to cope with strongly correlated covariates, a serious challenge for sparsity-based approaches to fMRI analysis.

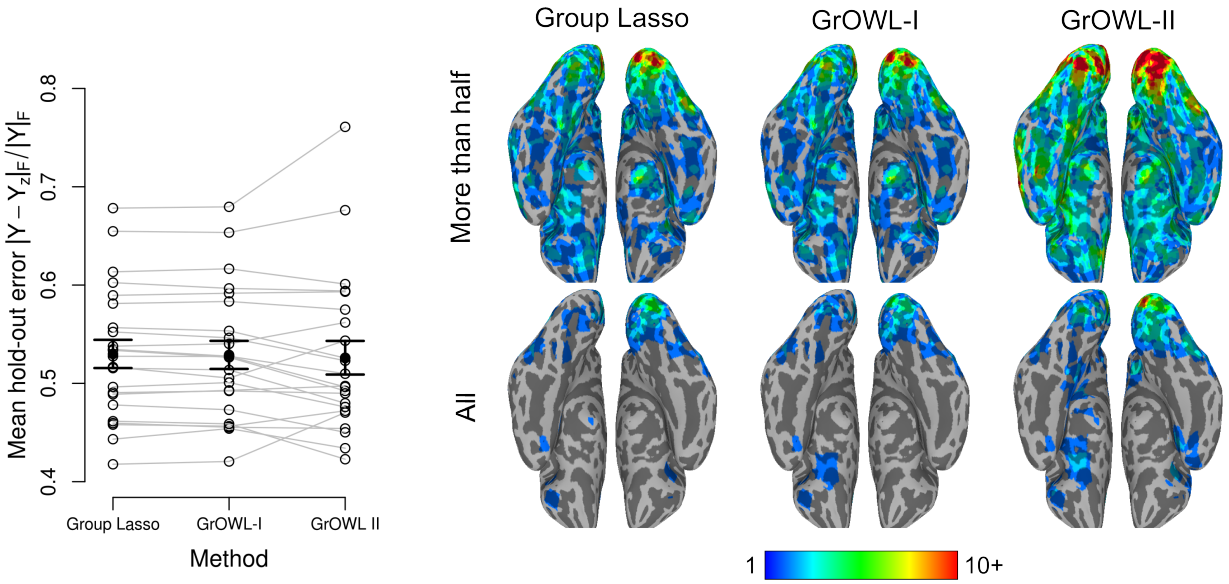


Fig. 5: Panel (a) shows mean hold-out prediction error for group lasso and GrOWLs for 23 subjects. Panel (b) shows surface maps corresponding to group lasso (left), GrOWL-I (middle) and GrOWL-II (right) showing the voxels selected for *at least five* and *all nine* cross-validations in the top and bottom rows respectively. The heat map shows the number of subjects for which those voxels were picked. Blue is the least (1 subject) and red is the most (10 or more subjects).

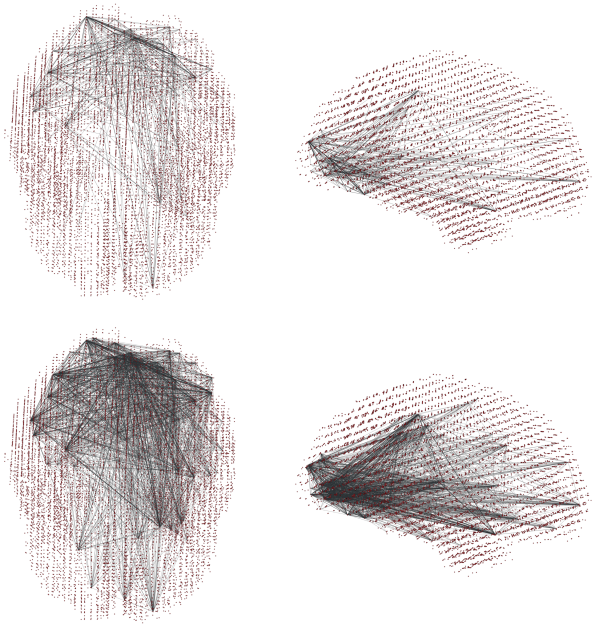


Fig. 6: Network plot showing the top edges from the $\boldsymbol{W}$ matrix for the best-performing parameterization of group LASSO (top) and GrOWL-II (bottom) in one subject. The thickness of the edges is proportional to the edge weights.

Our analysis of synthetic data generated from a neural network model of word-reading showed that the GrOWL identifies signal-carrying features more consistently than group lasso when signal is redundant; that the approach can discover different feature subsets encoding different kinds of similarity structure; and that the weight matrix $\boldsymbol{W}$ can be used to uncover subnetworks of features that jointly work to encode such structure. Analysis of visual similarity structure in fMRI data from a picture-viewing task further established that the approach can be used to find cortical regions and subnetworks that likewise express a target similarity structure. In future work these methods may be useful for discovering such structure for cases where the interesting cortical regions and networks have proven elusive.

Appendix A
Clustering properties of GrOWL with absolute error loss function

Proof of Theorem III.1

Proof. The proof is divided into two steps. First, we show $\|\widehat{\boldsymbol{\beta}}_{j*}\| = \|\widehat{\boldsymbol{\beta}}_{k*}\|$ and then we further show that the rows are equal. We proceed by contradiction. Assume $\|\widehat{\boldsymbol{\beta}}_{j*}\| \neq \|\widehat{\boldsymbol{\beta}}_{k*}\|$ and, without loss of generality, suppose $\|\widehat{\boldsymbol{\beta}}_{j*}\| > \|\widehat{\boldsymbol{\beta}}_{k*}\|$. We see that there exists a modification of the solution with a smaller GrOWL norm and same data-fitting term, and thus smaller overall objective value which contradicts our assumption that $\widehat{\boldsymbol{B}}$ is the minimizer of $L(\boldsymbol{B}) + G(\boldsymbol{B})$.

Consider the modification, $\boldsymbol{V} = \widehat{\boldsymbol{B}}$ except $\widehat{\boldsymbol{v}}_{j*} = \widehat{\boldsymbol{\beta}}_{j*} - \boldsymbol{\varepsilon}$ and $\widehat{\boldsymbol{v}}_{k*} = \widehat{\boldsymbol{\beta}}_{k*} + \boldsymbol{\varepsilon}$ where $\boldsymbol{\varepsilon} = \delta\widehat{\boldsymbol{\beta}}_{j*}$ and δ is chosen such that $\|\boldsymbol{\varepsilon}\| \in \left(0, \frac{\|\widehat{\boldsymbol{\beta}}_{j*}\| - \|\widehat{\boldsymbol{\beta}}_{k*}\|}{2}\right]$

Let $L(\boldsymbol{B}) = \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{B}\|_1 = \|\boldsymbol{Y}' - \boldsymbol{x}_{*j}\widehat{\boldsymbol{\beta}}_{j*} - \boldsymbol{x}_{*k}\widehat{\boldsymbol{\beta}}_{k*}\|_1$ where $\boldsymbol{Y}'$ is the residual term given by $\boldsymbol{Y}' = \boldsymbol{Y} - \sum_{i\neq j,k} \boldsymbol{x}_{*i}\widehat{\boldsymbol{\beta}}_{i*}$. Since $\boldsymbol{x}_{*j} = \boldsymbol{x}_{*k}$, L is invariant under this transformation, *i.e.*, $L(\boldsymbol{V}) = L(\widehat{\boldsymbol{B}})$. Same is true for $L(\boldsymbol{B}) = \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{B}\|_F^2$.

Observe that the GrOWL norm of $\boldsymbol{B}$ is equal to the OWL norm of the vector of euclidean norms of rows of $\boldsymbol{B}$. Since

$\|\boldsymbol{v}_{k*}\| = \|\boldsymbol{\beta}_{k*} + \boldsymbol{\varepsilon}\| \leq \|\boldsymbol{\beta}_{k*}\| + \|\boldsymbol{\varepsilon}\|$, this transformation is equivalent to that defined in Lemma III.1 and we have

$$G(\widehat{\boldsymbol{B}}) - G(\boldsymbol{V}) \quad \geq \quad \Delta\|\boldsymbol{\varepsilon}\|$$

This leads to a contradiction to our assumption that $\widehat{\boldsymbol{B}}$ is the minimizer of $L(\boldsymbol{B}) + G(\boldsymbol{B})$ and completes the proof that $\|\widehat{\boldsymbol{\beta}}_{j*}\| = \|\widehat{\boldsymbol{\beta}}_{k*}\|$. Now, let $\widehat{\boldsymbol{\beta}}_{j*} + \widehat{\boldsymbol{\beta}}_{k*} = \boldsymbol{z}$, then the minimizer satisfies

$$\min_{\widehat{\boldsymbol{\beta}}_{j*}, \widehat{\boldsymbol{\beta}}_{k*}} w_j\|\widehat{\boldsymbol{\beta}}_{j*}\| + w_k\|\widehat{\boldsymbol{\beta}}_{k*}\|$$

$$\text{such that } \widehat{\boldsymbol{\beta}}_{j*} + \widehat{\boldsymbol{\beta}}_{k*} = \boldsymbol{z} \text{ and } \|\widehat{\boldsymbol{\beta}}_{j*}\| = \|\widehat{\boldsymbol{\beta}}_{k*}\|$$

It is easy to see that the solution to this optimization is $\widehat{\boldsymbol{\beta}}_{j*} = \widehat{\boldsymbol{\beta}}_{k*} = \boldsymbol{z}/2$ □

Proof of Theorem III.2

Proof. The proof is similar to the identical columns theorem. By contradiction and without loss of generality, suppose $\|\widehat{\boldsymbol{\beta}}_{j*}\| > \|\widehat{\boldsymbol{\beta}}_{k*}\|$. We show that there exists a transformation of $\widehat{\boldsymbol{B}}$ such that the increase in the data fitting term is smaller than the decrease in the GrOWL norm.

Consider the modification, $\boldsymbol{V}$, as defined in the proof of Theorem III.1. By triangle inequality, the difference in loss function L that results from this modification satisfies

$$L(\boldsymbol{V}) - L(\widehat{\boldsymbol{B}}) \quad \leq \quad \left\|\boldsymbol{x}_{*j} - \boldsymbol{x}_{*k}\right\|_1\|\boldsymbol{\varepsilon}\|_1$$

Invoking Lemma III.1 as in the previous theorem and $\|\boldsymbol{\varepsilon}\|_1 \leq \sqrt{r}\|\boldsymbol{\varepsilon}\|$, we get

$$L(\boldsymbol{V}) + G(\boldsymbol{V}) - (L(\widehat{\boldsymbol{B}}) + G(\widehat{\boldsymbol{B}})) \leq (\left\|\boldsymbol{x}_{*j} - \boldsymbol{x}_{*k}\right\|_1 - \frac{\Delta}{\sqrt{r}})\|\boldsymbol{\varepsilon}\| < 0$$

This contradicts our assumption that $\widehat{\boldsymbol{B}}$ is the minimizer of $L(\boldsymbol{B}) + G(\boldsymbol{B})$ and completes the proof. □

Proof of Theorem III.3

Proof. The proof is similar to the identical columns theorem. By contradiction, suppose $\|\widehat{\boldsymbol{\beta}}_{j*} - \widehat{\boldsymbol{\beta}}_{k*}\| \geq \frac{8\phi\|\widehat{\boldsymbol{\beta}}_{k*}\|}{4\phi^2+1} \geq \frac{2\|\widehat{\boldsymbol{\beta}}_{k*}\|}{\phi}$. We show that there exists a transformation of $\widehat{\boldsymbol{B}}$ such that the increase in the data fitting term is smaller than the decrease in the GrOWL norm.

Consider the modification, $\boldsymbol{V}$, as defined in the proof of Theorem III.1 with $\boldsymbol{\varepsilon} = \frac{\widehat{\boldsymbol{\beta}}_{j*} - \widehat{\boldsymbol{\beta}}_{k*}}{2}$. By triangle inequality, the difference in loss function L that results from this modification satisfies

$$L(\boldsymbol{V}) - L(\widehat{\boldsymbol{B}}) \quad \leq \quad \left\|\boldsymbol{x}_{*j} - \boldsymbol{x}_{*k}\right\|_1\|\boldsymbol{\varepsilon}\|_1$$

We now bound the decrease in the GrOWL norm. Note by parallelogram law,

$$\begin{aligned}
&\|\widehat{\boldsymbol{\beta}}_{j*} + \widehat{\boldsymbol{\beta}}_{k*}\|^2 \\
&= 2\|\widehat{\boldsymbol{\beta}}_{j*}\|^2 + 2\|\widehat{\boldsymbol{\beta}}_{k*}\|^2 - \|\widehat{\boldsymbol{\beta}}_{j*} - \widehat{\boldsymbol{\beta}}_{k*}\|^2 \\
&\leq 2\|\widehat{\boldsymbol{\beta}}_{j*}\|^2 + 2\|\widehat{\boldsymbol{\beta}}_{k*}\|^2 + \left(\frac{1}{4\phi^2} - \frac{1}{4\phi^2} - 1\right)\|\widehat{\boldsymbol{\beta}}_{j*} - \widehat{\boldsymbol{\beta}}_{k*}\|^2 \\
&\leq 4\|\widehat{\boldsymbol{\beta}}_{j*}\|^2 + \left(\frac{\|\widehat{\boldsymbol{\beta}}_{j*} - \widehat{\boldsymbol{\beta}}_{k*}\|}{2\phi}\right)^2 - \frac{1+4\phi^2}{4\phi^2}\|\widehat{\boldsymbol{\beta}}_{j*} - \widehat{\boldsymbol{\beta}}_{k*}\|^2 \\
&\leq 4\|\widehat{\boldsymbol{\beta}}_{j*}\|^2 + \left(\frac{\|\widehat{\boldsymbol{\beta}}_{j*} - \widehat{\boldsymbol{\beta}}_{k*}\|}{2\phi}\right)^2 - 2\frac{\|\widehat{\boldsymbol{\beta}}_{j*}\|\|\widehat{\boldsymbol{\beta}}_{j*} - \widehat{\boldsymbol{\beta}}_{k*}\|}{\phi} \\
&\leq \left(\|\widehat{\boldsymbol{\beta}}_{j*}\| + \|\widehat{\boldsymbol{\beta}}_{k*}\| - \frac{\|\widehat{\boldsymbol{\beta}}_{j*} - \widehat{\boldsymbol{\beta}}_{k*}\|}{2\phi}\right)^2
\end{aligned}$$

Thus, we have

$$\begin{aligned}
G(\widehat{\boldsymbol{B}}) - G(\boldsymbol{V}) &\geq \Delta\left(\|\widehat{\boldsymbol{\beta}}_{j*}\| + \|\widehat{\boldsymbol{\beta}}_{k*}\| - \|\widehat{\boldsymbol{\beta}}_{j*} + \widehat{\boldsymbol{\beta}}_{k*}\|\right) \\
&\geq \frac{\Delta\|\widehat{\boldsymbol{\beta}}_{j*} - \widehat{\boldsymbol{\beta}}_{k*}\|}{2\phi} = \frac{\Delta\|\boldsymbol{\varepsilon}\|}{\phi}
\end{aligned}$$

Combining this with $\|\boldsymbol{\varepsilon}\|_1 \leq \sqrt{r}\|\boldsymbol{\varepsilon}\|$, we get

$$\begin{aligned}
&L(\boldsymbol{V}) + G(\boldsymbol{V}) - (L(\widehat{\boldsymbol{B}}) + G(\widehat{\boldsymbol{B}})) \\
&\qquad \leq \left(\sqrt{r}\left\|\boldsymbol{x}_{*j} - \boldsymbol{x}_{*k}\right\|_1 - \frac{\Delta}{\phi}\right)\|\boldsymbol{\varepsilon}\| < 0
\end{aligned}$$

This contradicts our assumption that $\widehat{\boldsymbol{B}}$ is the minimizer of $L(\boldsymbol{B}) + G(\boldsymbol{B})$ and completes the proof. □

APPENDIX B
CLUSTERING PROPERTIES OF GROWL WITH SQUARED FROBENIUS LOSS FUNCTION

In this section, we consider the optimization

$$\min_{\boldsymbol{X}} \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{B}\|_F^2 \; + \; G(\boldsymbol{B}) \tag{7}$$

Here we derive an upper bound on the increase in the squared loss term after applying the transformation, $\boldsymbol{V}$. We assume that the columns of the matrix, $\boldsymbol{X}$, are normalized to a common norm, *i.e.,* ($\|\boldsymbol{x}_{*i}\| = c$ for $i = 1, \cdots, p$). Define $L(\boldsymbol{X}) = \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{B}\|_F^2 = \|\boldsymbol{Y}' - \boldsymbol{x}_{*j}\boldsymbol{\beta}_{j*} - \boldsymbol{x}_{*k}\boldsymbol{\beta}_{k*}\|_F^2$ where $\boldsymbol{Y}'$ is again the residual term.

Lemma B.1. *Let $\widehat{\boldsymbol{B}} \in \mathbb{R}^{p\times r}$ and if $\boldsymbol{V}$ is as defined in (?), then we have*

$$L(\boldsymbol{V}) - L(\widehat{\boldsymbol{B}}) \leq \|\boldsymbol{\varepsilon}\|\|\boldsymbol{Y}'\|_F\|\boldsymbol{x}_{*j} - \boldsymbol{x}_{*k}\|$$

Proof.

$$\begin{aligned}
L(\boldsymbol{V}) - L(\widehat{\boldsymbol{B}}) = &\frac{1}{2}\|\boldsymbol{Y}' - \boldsymbol{x}_{*j}(\widehat{\boldsymbol{\beta}}_{j*} - \boldsymbol{\varepsilon}) - \boldsymbol{x}_{*k}(\widehat{\boldsymbol{\beta}}_{k*} + \boldsymbol{\varepsilon})\|_F^2 \\
&- \frac{1}{2}\|\boldsymbol{Y}' - \boldsymbol{x}_{*j}\widehat{\boldsymbol{\beta}}_{j*} - \boldsymbol{x}_{*k}\widehat{\boldsymbol{\beta}}_{k*}\|_F^2
\end{aligned}$$

Expanding the Frobenius norm terms, canceling the common $\frac{1}{2}\|\boldsymbol{Y}'\|_F^2$ terms and using the common norm of columns

($\|\boldsymbol{x}_{*i}\| = c$ for $i = 1, \cdots, p$) we get

$$\begin{aligned}
&L(\boldsymbol{V}) - L(\widehat{\boldsymbol{B}}) \\
&= \frac{c^2}{2}\text{tr}((\widehat{\boldsymbol{\beta}}_{j*} - \boldsymbol{\varepsilon})(\widehat{\boldsymbol{\beta}}_{j*} - \boldsymbol{\varepsilon})^T + (\widehat{\boldsymbol{\beta}}_{k*} + \boldsymbol{\varepsilon})(\widehat{\boldsymbol{\beta}}_{k*} + \boldsymbol{\varepsilon})^T \\
&- \widehat{\boldsymbol{\beta}}_{j*}\widehat{\boldsymbol{\beta}}_{j*}^T - \widehat{\boldsymbol{\beta}}_{k*}\widehat{\boldsymbol{\beta}}_{k*}^T) + \text{tr}(\boldsymbol{Y}'^T(\boldsymbol{x}_{*j} - \boldsymbol{x}_{*k})\boldsymbol{\varepsilon}) \\
&+ \text{tr}((\widehat{\boldsymbol{\beta}}_{j*} - \boldsymbol{\varepsilon})\boldsymbol{x}_{*j}^T\boldsymbol{x}_{*k}(\widehat{\boldsymbol{\beta}}_{k*} + \boldsymbol{\varepsilon})^T - \widehat{\boldsymbol{\beta}}_{j*}\boldsymbol{x}_{*j}^T\boldsymbol{x}_{*k}\widehat{\boldsymbol{\beta}}_{k*}^T)
\end{aligned}$$

Expanding terms and making further cancellations gives

$$\begin{aligned}
&L(\boldsymbol{V}) - L(\widehat{\boldsymbol{B}}) \\
&= \text{tr}(\boldsymbol{Y}'^T(\boldsymbol{x}_{*j} - \boldsymbol{x}_{*k})\boldsymbol{\varepsilon}) - (c^2 - \boldsymbol{x}_{*j}^T\boldsymbol{x}_{*k})\ \text{tr}((\widehat{\boldsymbol{\beta}}_{j*} - \widehat{\boldsymbol{\beta}}_{k*} - \boldsymbol{\varepsilon})\boldsymbol{\varepsilon}^T) \\
&\leq \text{tr}(\boldsymbol{Y}'^T(\boldsymbol{x}_{*j} - \boldsymbol{x}_{*k})\boldsymbol{\varepsilon}) \\
&- (c^2 - \boldsymbol{x}_{*j}^T\boldsymbol{x}_{*k})\|\boldsymbol{\varepsilon}\|(\|\widehat{\boldsymbol{\beta}}_{j*}\| - \|\widehat{\boldsymbol{\beta}}_{k*}\| - \|\boldsymbol{\varepsilon}\|) \\
&\leq \text{tr}(\boldsymbol{Y}'^T(\boldsymbol{x}_{*j} - \boldsymbol{x}_{*k})\boldsymbol{\varepsilon}^T) \\
&\leq \|\boldsymbol{Y}'\|_F\|(\boldsymbol{x}_{*j} - \boldsymbol{x}_{*k})\boldsymbol{\varepsilon}\|_F \\
&= \|\boldsymbol{\varepsilon}\|\|\boldsymbol{Y}'\|_F\|\boldsymbol{x}_{*j} - \boldsymbol{x}_{*k}\|
\end{aligned}$$

where the first inequality follows from simplification and Cauchy-Schwarz inequality. The second inequality follows from $c^2 > \boldsymbol{x}_{*j}^T\boldsymbol{x}_{*k}$ and $\|\widehat{\boldsymbol{\beta}}_{j*}\|_2 - \|\widehat{\boldsymbol{\beta}}_{k*}\|_2 - \|\boldsymbol{\varepsilon}\| > 0$ (by assumption). The third inequality follows, again, by Cauchy-Schwarz inequality.

□

Using this Lemma one can easily extend the clustering properties of GrOWL to the optimization in (7).

APPENDIX C
PROXIMAL ALGORITHMS FOR GROWL

Proof. Outline: the proof proceeds by finding a lower bound for the objective function in (5) and then we show that the proposed solution achieves this lower bound.

First, note that the following is true for any $\boldsymbol{B}$ and $\boldsymbol{V}$,

$$\begin{aligned}
\|\boldsymbol{B} - \boldsymbol{V}\|_F^2 &= \sum_{i=1}^{p}\|\boldsymbol{\beta}_{i*} - \boldsymbol{v}_{i*}\|^2 \\
&\geq \sum_{i=1}^{p}(\|\boldsymbol{\beta}_{i*}\| - \|\boldsymbol{v}_{i*}\|)^2 = \|\tilde{\boldsymbol{\beta}} - \tilde{\boldsymbol{v}}\|^2
\end{aligned}$$

where the inequality follows from reverse triangle inequality.

Combining this with $G(\boldsymbol{B}) = \Omega_{\boldsymbol{w}}(\tilde{\boldsymbol{\beta}})$, we have a lower bound on the objective function in (5). For all $\boldsymbol{B} \in \mathbb{R}^{p\times r}$

$$\frac{1}{2}\|\boldsymbol{B} - \boldsymbol{V}\|_F^2 + G(\boldsymbol{B}) \geq \frac{1}{2}\|\text{prox}_{\Omega_{\boldsymbol{w}}}(\tilde{\boldsymbol{v}}) - \tilde{\boldsymbol{v}}\|^2 + \Omega_{\boldsymbol{w}}(\text{prox}_{\Omega_{\boldsymbol{w}}}(\tilde{\boldsymbol{v}}))$$

Finally, we show that $\boldsymbol{B} = \widehat{\boldsymbol{V}}$ achieves this lower bound,

$$\begin{aligned}
&\frac{1}{2}\|\widehat{\boldsymbol{V}} - \boldsymbol{V}\|_F^2 + G(\widehat{\boldsymbol{V}}) \\
&= \frac{1}{2}\sum_{i=1}^{p}\|(\text{prox}_{\Omega_{\boldsymbol{w}}}(\tilde{\boldsymbol{v}}))_i\frac{\boldsymbol{v}_{i*}}{\|\boldsymbol{v}_{i*}\|} - \boldsymbol{v}_{i*}\|_2^2 + \Omega_{\boldsymbol{w}}(\text{prox}_{\Omega_{\boldsymbol{w}}}(\tilde{\boldsymbol{v}})) \\
&= \frac{1}{2}\|\text{prox}_{\Omega_{\boldsymbol{w}}}(\tilde{\boldsymbol{v}}) - \tilde{\boldsymbol{v}}\|_2^2 + \Omega_{\boldsymbol{w}}(\text{prox}_{\Omega_{\boldsymbol{w}}}(\tilde{\boldsymbol{v}}))
\end{aligned}$$

□

REFERENCES

[1] N. Kriegeskorte, M. Mur, and P. Bandettini, "Representational similarity analysis–connecting the branches of systems neuroscience", *Frontiers in systems neuroscience*, 2, 2008.

[2] N. Kriegeskorte, R. Goebel, and P. Bandettini, "Information-based functional brain mapping", *Proceedings of the National Academy of Sciences of the United States of America, 103* vol. 10, pp. 3863–3868, 2006.

[3] A. Argyriou, T. Evgeniou, and M. Pontil. "Convex multi-task feature learning", *Machine Learning, 3*, vol. 73,, pp. 243–272, 2008.

[4] A. Tversky, and I. Gati, "Similarity, separability, and the triangle inequality", *Psychological review, 2* vol. 89, pp. 123, 1982.

[5] D. Plaut, J. McClelland, M. Seidenberg, and K. Patterson, "Understanding normal and impaired word reading: computational principles in quasi-regular domains." *Psychological review 1* vol. 103, pp. 56, 1996.

[6] G. Obozinski, M. Wainwright and M. Jordan, "Support union recovery in high-dimensional multivariate regression," *Ann. Stat.*, pp. 1–47, 2011.

[7] H. Bondell and B. Reich, "Simultaneous regression shrinkage, variable selection, and supervised clustering of predictors with OSCAR," *Biometrics*, vol. 64, pp. 115–123, 2007.

[8] H. Dalton, "The measurement of the inequality of incomes," *The Economic Journal*, vol. 30, pp. 348–361, 1920.

[9] H. Zou, and T. Hastie, "Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 67* vol. 2, pp. 301–320, 2005

[10] K. Lounici, M. Pontil, A. Tsybakov, and S. van de Geer. "Taking advantage of sparsity in multi-task learning." http://arxiv.org/abs/0903.1468, 2009.

[11] K. Lounici, M. Pontil, S. van de Geer, and A. Tsybakov, "Oracle inequalities and optimal inference under group sparsity", *The Annals of Statistics, 39* vol. 4, pp. 2164–2204, 2012.

[12] M. Bogdan, E. van den Berg, W. Su, and E. Candès, "Statistical estimation and testing via the sorted ℓ_1 norm," available at http://arxiv.org/abs/1310.1969, 2013.

[13] M. Bogdan, E. van den Berg, C. Sabatti, W. Su, and E. Candès, "SLOPE – adaptive variable selection via convex optimization", available at http://arxiv.org/abs/1407.3824, 2014.

[14] M. Harm, and M. Seidenberg, "Computing the meanings of words in reading: cooperative division of labor between visual and phonological processes." *Psychological review 3*, vol. 111, pp. 662, 2004.

[15] M. Figueiredo, and R. Nowak, "Sparse Estimation with Strongly Correlated Variables using Ordered Weighted l1 Regularization", http://arxiv.org/abs/1409.4005, 2014.

[16] N. Parikh, and S. Boyd, "Proximal algorithms.", *Foundations and Trends in optimization 3* vol. 1, pp. 123–231, 2013.

[17] P. Bühlmann, P. Rüttiman, S. van de Geer, and C.-H. Zhang, "Correlated variables in regression: Clustering and sparse estimation," *Journal of Statistical Planning and Inference*, pp. 1835–1858, 2013.

[18] S. Antani, R. Kasturi, and R. Jain, "A survey on the use of pattern recognition methods for abstraction, indexing and retrieval of images and video," *Pattern recognition 4* vol. 35, pp. 945–965, 2002.

[19] S. Oymak, A. Jalali, M. Fazel, Y. Eldar, and B. Hassibi, "Simultaneously structured models with application to sparse and low-rank matrices", http://arxiv.org/abs/1212.3753, 2012.

[20] X. Zeng, M. Figueiredo, "Decreasing Weighted Sorted ℓ_1 Regularization", *IEEE Signal Processing Letters,* vol. 21, pp. 1240–1244, 2014.

[21] X. Zeng, M. Figueiredo, "The atomic norm formulation of OSCAR regularization with application to the Frank-Wolfe algorithm", *Proceedings of the European Signal Processing Conference*, Lisbon, Portugal, 2014.