Representational Similarity Learning with Application to Brain Networks

Urvashi Oswal UOSWAL@WISC.EDU
Christopher Cox CRCOX@WISC.EDU
Matthew A. Lambon Ralph[†] MATT.LAMBON-RALPH@MANCHESTER.AC.UK
Timothy Rogers TTROGERS@WISC.EDU
Robert Nowak RDNOWAK@WISC.EDU
University of Wisconsin-Madison, Madison, WI 53706, USA

[†]University of Manchester, Manchester M13 9PL, UK

Abstract

Representational Similarity Learning (RSL) aims to discover features that are important in representing (human-judged) similarities among objects. RSL can be posed as a sparsity-regularized multi-task regression problem. Standard methods, like group lasso, may not select important features if they are strongly correlated with others. To address this shortcoming we present a new regularizer for multitask regression called *Group Ordered Weighted* ℓ_1 (GrOWL). Another key contribution of our paper is a novel application to fMRI brain imaging. *Representational Similarity Analysis* (RSA) is a tool for testing whether localized brain regions encode perceptual similarities. Using GrOWL, we propose a new approach called *Network RSA* that can discover arbitrarily structured brain networks (possibly widely distributed and non-local) that encode similarity information. We show, in theory and fMRI experiments, how GrOWL deals with strongly correlated covariates.

1. Introduction

This paper considers the following learning task. Suppose we have a set of items along with human-judged pairwise similarities among them. For instance, the items could be visual stimuli such as advertisements, pictures, or diagrams. Assume that we also have a high-dimensional feature associated with each item. These could be numerical features quantifying the characteristics of each item or, in the case of fMRI, the features are voxel responses to stimuli. The learning task is to determine the subset of features that is most predictive of the human-judged similarities. This can be posed mathematically as follows. Let $\boldsymbol{S}$ be an $n \times n$ matrix of pairwise similarities between n items. Let $\boldsymbol{X}$ be an $n \times p$ matrix where the ith row is the $1 \times p$ vector of the features for item i. We then wish to find a weight matrix $\boldsymbol{W}$ such $\boldsymbol{X}\boldsymbol{W}\boldsymbol{X}^T \approx \boldsymbol{S}$. The weight matrix reveals which features are most important and how they are combined to represent the human-judged similarities. We call this *Representational Similarity Learning* (RSL). This problem can be viewed as special case of regression metric learning, see (Kulis, 2012), but the focus of RSL is to identify the subset of features that are most strongly influencing human-judged similarities. The theory and methods developed in this paper may also be applicable to other metric learning problems (Atzmon et al., 2015; Ying et al., 2009).

Let us illustrate this learning problem with two applications. First, suppose the items are diagrams of chemical molecules, and each diagram is also described by a vector comprised of many visual features (e.g., counts of different atom types, bonds, bond angles, etc). Novice chemistry students may miss critical similarities and differences when comparing different diagrams. After gathering pairwise similarity judgments from the students, RSL could be used to identify which features they are attending to and, thus, which important features they may be overlooking (Rau et al., 2016). Second, consider *Representational Similarity Analysis* (RSA) in fMRI brain imaging (Kriegeskorte et al., 2008). In RSA a person is scanned while viewing n different visual stimuli. Pairwise similarities are obtained through other experiments, such as asking people to look at pairs of stimuli and rate the similarity. In this case, the features are the stimuli responses of p voxels in the brain, and the goal is to determine which voxels (and hence brain regions) are encoding the similarities. RSA is depicted in Figure 1, and it is the focus of our application in Section 4.

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

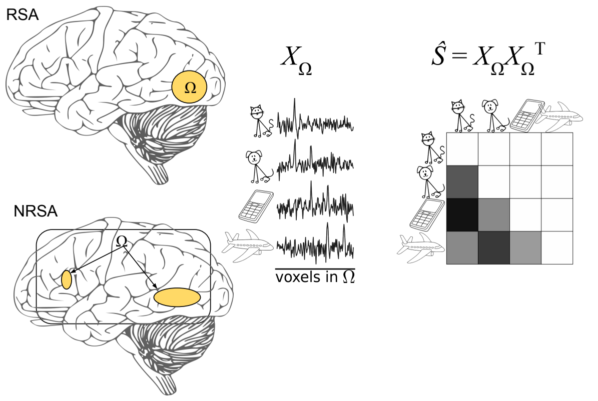

Figure 1. Representational Similarity Analysis. Traditional RSA methods consider only localized brain regions of interest or spherical clusters in the cortex (upper left) (Kriegeskorte et al., 2006; 2008). In Section 4, we propose a new *Network* RSA (NRSA) method that can potentially identify non-local brain networks that encode similarity information (lower left).

1.1. Representational Similarity Learning

Let $\boldsymbol{X} \in \mathbb{R}^{n\times p}$ denote a feature matrix. Each row corresponds to p features associated with a specific item, and each column corresponds to the values of a specific feature for the n items. The goal of RSL is to find a sparse and symmetric 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 features are included in the similarity representation. For instance, consider the $n \times 1$ vectors corresponding features $\boldsymbol{x}_k$ 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 features 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 features contribute to the approximation of the similarity matrix $\boldsymbol{S}$. The complete similarity representation can be expressed as

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

The approximation problem can be posed as the least squares optimization

$$\min_{\boldsymbol{W}} \|\boldsymbol{S} - \boldsymbol{X}\boldsymbol{W}\boldsymbol{X}^T\|_F^2 \tag{1}$$

where the objective is the Frobenius norm of the difference between the similarity matrix $\boldsymbol{S}$ and its approximation. Classic studies of human-produced similarity judgments in many domains of interest yield low rank matrices (McRae et al., 2005; Shaver et al., 1987; Shepard, 1980) due to clustering or other representational structure amongst the items under consideration. Therefore, we suppose $\boldsymbol{S}$ is a real, symmetric and approximately rank r matrix, then there exists a matrix $\boldsymbol{Y} \in \mathbb{R}^{n\times r}$ and diagonal matrix $\boldsymbol{D} \in \mathbb{R}^{r\times r}$ which satisfies $\boldsymbol{S} \approx \boldsymbol{Y}\boldsymbol{D}\boldsymbol{Y}^T$ (e.g., obtained via eigendecomposition or Cholesky decomposition) where the diagonal entries of $\boldsymbol{D}$ correspond to the sign of the r largest eigenvalues and the columns of $\boldsymbol{Y}$ are the corresponding eigenvectors of $\boldsymbol{S}$. We will assume that $\boldsymbol{S}$ is rank r in the following discussion (if not, then we will use its best rank r approximation instead). Thus, we may instead consider the optimization

$$\min_{\boldsymbol{B}} \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{B}\|_F^2 \tag{2}$$

For any coefficient matrix $\boldsymbol{B}$ the corresponding weight matrix is given by $\boldsymbol{W} = \boldsymbol{B}\boldsymbol{D}\boldsymbol{B}^T$. Both optimizations are convex, but we will work with the latter since it automatically enforces the low-rank assumption and can be easily modified to include additional constraints or regularizers. It is easily verified that every stationary point of (2) leads to a stationary point of (1). Therefore, since both optimizations are convex, $\widehat{\boldsymbol{B}} = \arg\min_{\boldsymbol{B}} \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{B}\|_F^2$ yields $\widehat{\boldsymbol{W}} = \widehat{\boldsymbol{B}}\boldsymbol{D}\widehat{\boldsymbol{B}}^T$, which is a minimizer of (1).

In many applications the weight matrix $\boldsymbol{W}$ and the coefficient matrix $\boldsymbol{B}$ are expected to exhibit sparsity. Indeed, our hypothesis is that a small subset of the features encodes the similarity representations, hence the sparsity. Thus, the optimization above can be modified to obtain sparse and low-rank solutions, as described next.

2. RSL via Group Lasso

Consider the group lasso optimization

$$\min_{\boldsymbol{B}\in\mathbb{R}^{p\times r}} \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{B}\|_F^2 + \lambda \|\boldsymbol{B}\|_{1,2} . \tag{3}$$

Note that the optimization variable $\boldsymbol{B}$ is a $p \times r$ matrix, which guarantees a rank r (or less) solution, and thus similarity representation $\boldsymbol{X}\boldsymbol{B}\boldsymbol{D}\boldsymbol{B}^T\boldsymbol{X}^T$ will be rank r at most, which is a simple way to enforce the low-rank constraint. The parameter $\lambda > 0$ is an adjustable weight on the sparsity-promoting regularizer $\|\boldsymbol{B}\|_{1,2}$, which is defined as follows. The rows of $\boldsymbol{B}$ are denoted by $\boldsymbol{\beta}_{i\cdot}$, $i = 1, \ldots, p$, and the norm $\|\boldsymbol{B}\|_{1,2} = \sum_{i=1}^{p} \|\boldsymbol{\beta}_{i\cdot}\|_2$. This encourages solutions with only a few nonzero rows in $\boldsymbol{B}$ (Lounici et al., 2009; 2011; Obozinski et al., 2011). We note that the optimization in (1) can also be modified directly to obtain sparse and low-rank solutions. For instance, the nuclear norm of $\boldsymbol{W}$ could be penalized to obtain a low-rank solution. However, the nuclear norm optimization tends to be computationally expensive in practice.

We mention here that recently a similar approach to sparse distance metric learning has been proposed in (Atzmon et al., 2015). This method also solves a form of sparsity-regularized optimization to obtain a sparse $\boldsymbol{W}$ matrix but with weak supervision in form of rankings over triplets of items. Also, (Ying et al., 2009) have proposed and studied another optimization for sparse metric learning where they impose group sparsity directly on $\boldsymbol{W}$.

The main technical 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. This problem is illustrated in Figure 2, where we simulate a situation in which columns 5 and 7 of the data matrix $\boldsymbol{X}$ are highly correlated. Group lasso selects one of the corresponding rows in $\boldsymbol{B}$ (row 5), whereas GrOWL correctly selects both rows 5 and 7. Note that the group lasso can select at most n features (n being the number of items), since the number of nonzero rows in the solution cannot exceed the number of measurements. This can be severe limitation of the group lasso in applications where the number of features far exceeds the number of items.

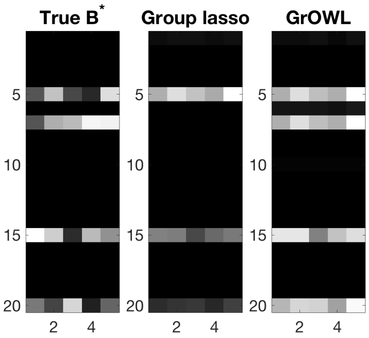

Figure 2. A comparison of group lasso (middle) and grOWL (right) optimization solutions with correlated columns in $\boldsymbol{X}$ showing that GrOWL selects relevant features (row 5 and 7) even if they happen to be strongly correlated and automatically cluster them by setting the corresponding coefficient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

In the standard (single-task) regression problem, this issue has been tackled using many techniques, including elastic net (Zou & Hastie, 2005), OSCAR (Bondell & Reich, 2008), OWL (Figueiredo & Nowak, 2016), and others. We propose a generalization of the 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 features in $\boldsymbol{X}$, unlike standard group lasso which may not select relevant features if they happen to be strongly correlated with others. Second, GrOWL encourages the coefficients associated with strongly correlated features to be near or exactly equal. In effect, this averages strongly correlated columns which can help to denoise features and improve predictions. This property of GrOWL could also be useful in other applications of multi-task regression with correlated features.

3. 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 (Figueiredo & Nowak, 2016), a special case of which is the so-called OSCAR (Bondell & Reich, 2008). 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 define the GrOWL (Group OWL) norm, and show that it automatically groups and averages highly correlated columns in $\boldsymbol{X}$ in the multiple regression setting.

In this section, we consider the general optimization

$$\min_{\boldsymbol{B}\in\mathbb{R}^{p\times r}} L(\boldsymbol{B}) + G(\boldsymbol{B}) \tag{4}$$

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. We give proof sketches for the main theorems and leave the proof details to the supplementary material.

3.1. GrOWL penalty

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

$$G(\boldsymbol{B}) = \sum_{i=1}^{p} w_i \|\boldsymbol{\beta}_{[i]\cdot}\|_2, \tag{5}$$

where $\boldsymbol{\beta}_{[i]\cdot}$ 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 (Figueiredo & Nowak, 2016) which will be useful later in the section.

Lemma 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 $-\beta_k \leq \varepsilon' \leq \varepsilon$, 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, $w_1 \geq w_2 \geq \cdots \geq w_p \geq 0$, 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

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

Proof sketch. The proof is similar to that of Lemma 2.1 in (Figueiredo & Nowak, 2016) with different sizes $\varepsilon, \varepsilon'$. The result follows since we assume that increase in k-th component is less than decrease in j-th component *i.e.,* $\varepsilon' \leq \varepsilon$.

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 1 (Identical columns)**.** *Let $\widehat{\boldsymbol{B}}$ denote the solution to the optimization in (4) 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 $\boldsymbol{x}_{\cdot j}$ and $\boldsymbol{x}_{\cdot k}$ satisfy $\boldsymbol{x}_{\cdot j} = \boldsymbol{x}_{\cdot k}$ and the minimum gap, $\Delta > 0$, then $\widehat{\boldsymbol{\beta}}_{j\cdot} = \widehat{\boldsymbol{\beta}}_{k\cdot}$.*

Proof sketch. The proof is divided into two steps. First, we show $\|\widehat{\boldsymbol{\beta}}_{j\cdot}\| = \|\widehat{\boldsymbol{\beta}}_{k\cdot}\|$ and then we further show that the rows are equal. We proceed by contradiction. Assume $\|\widehat{\boldsymbol{\beta}}_{j\cdot}\| \neq \|\widehat{\boldsymbol{\beta}}_{k\cdot}\|$ and, without loss of generality, suppose $\|\widehat{\boldsymbol{\beta}}_{j\cdot}\| > \|\widehat{\boldsymbol{\beta}}_{k\cdot}\|$. We see that there exists a modification of the solution with a smaller GrOWL norm using Lemma 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 theorems state that nearly identical variables lead to equal norm coefficient rows, and further highly correlated coefficient rows, corresponding to those variables in the solution given by the optimization using GrOWL.

Theorem 2 (Correlated columns 1)**.** *Let $\widehat{\boldsymbol{B}}$ denote the solution to the optimization in (4) with $L(\boldsymbol{B}) = \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{B}\|_1$ or $L(\boldsymbol{B}) = \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{B}\|_F^2$. If $\boldsymbol{x}_{\cdot j}$ and $\boldsymbol{x}_{\cdot k}$ satisfy $\|\boldsymbol{x}_{\cdot j} - \boldsymbol{x}_{\cdot k}\|_1 \leq \frac{\Delta}{\sqrt{r}}$ or $\|\boldsymbol{x}_{\cdot j} - \boldsymbol{x}_{\cdot k}\|_2 \leq \frac{\Delta}{\|\boldsymbol{Y}\|_F}$ respectively, then $\|\widehat{\boldsymbol{\beta}}_{j\cdot}\| = \|\widehat{\boldsymbol{\beta}}_{k\cdot}\|$.*

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

Theorem 3 (Correlated columns 2)**.** *Let $\widehat{\boldsymbol{B}}$ denote the solution to the optimization in (4) with $L(\boldsymbol{B}) = \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{B}\|_1$ or $L(\boldsymbol{B}) = \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{B}\|_F^2$. If $\boldsymbol{x}_{\cdot j}$ and $\boldsymbol{x}_{\cdot k}$ satisfy $\|\boldsymbol{x}_{\cdot j} - \boldsymbol{x}_{\cdot k}\|_1 \leq \frac{\Delta}{\phi\sqrt{r}}$ or $\|\boldsymbol{x}_{\cdot j} - \boldsymbol{x}_{\cdot k}\|_2 \leq \frac{\Delta}{\phi\|\boldsymbol{Y}\|_F}$ respectively, then $\|\widehat{\boldsymbol{\beta}}_{j\cdot} - \widehat{\boldsymbol{\beta}}_{k\cdot}\| \leq \frac{8\phi\|\widehat{\boldsymbol{\beta}}_{k\cdot}\|}{4\phi^2+1}$ which further implies that*

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

where $\phi \geq 1$.

Proof sketch. By contradiction, suppose $\|\widehat{\boldsymbol{\beta}}_{j\cdot} - \widehat{\boldsymbol{\beta}}_{k\cdot}\| \geq \frac{8\phi\|\widehat{\boldsymbol{\beta}}_{k\cdot}\|}{4\phi^2+1} \geq \frac{2\|\widehat{\boldsymbol{\beta}}_{k\cdot}\|}{\phi}$. 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 $\boldsymbol{w}$. We study the GrOWL-Lin weights with linear decay (equivalent to the OSCAR in single-task regression), and the GrOWL-Spike weight sequence which puts a big weight on the maximum magnitude while the rest of the coefficients are weighted equally.

GrOWL-Lin:	$w_i = \lambda + \lambda_1(p-i)/p$ for $i = 1, \ldots, p$
GrOWL-Spike:	$w_1 = \lambda + \lambda_1, w_i = \lambda_1$ for $i = 2, \ldots, p$

3.2. Proximal algorithms

We derive the proximal operator for the optimization using the GrOWL norm here. The computational algorithms to solve the GrOWL optimization based on the proximity operators can be found in (Parikh & Boyd, 2013). The proximal operator of the GrOWL norm is given by

$$\mathrm{prox}_G(\boldsymbol{V}) = \arg\min_{\boldsymbol{B}} \frac{1}{2}\|\boldsymbol{B} - \boldsymbol{V}\|_F^2 + G(\boldsymbol{B}) \tag{6}$$

In the following theorem, we solve for the proximity operator of GrOWL in terms of the proximity of OWL. For the exact formulation of $\mathrm{prox}_{\Omega_{\boldsymbol{w}}}$, see (Bogdan et al., 2013), (Zeng et al., 2014).

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

$$\widehat{\boldsymbol{v}}_{i\cdot} = (\mathrm{prox}_{\Omega_{\boldsymbol{w}}}(\tilde{\boldsymbol{v}}))_i \frac{\boldsymbol{v}_{i\cdot}}{\|\boldsymbol{v}_{i\cdot}\|} \tag{7}$$

Proof Sketch: The proof proceeds by finding a lower bound for the objective function in (6) and then we show that the proposed solution achieves this lower bound.

Efficient O($p\log p$) algorithms to compute $\mathrm{prox}_{\Omega_{\boldsymbol{w}}}$ have been proposed by Bodgan *et al* (Bogdan et al., 2013; 2014).

4. RSL from fMRI Data

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 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 RSA (Kriegeskorte et al., 2008). RSA is typically applied either to a specific brain region of interest (ROI) or to many localized regions throughout the brain in a process called *searchlight analysis* (Kriegeskorte et al., 2006). For a given region, RSA computes the cosine distances between the evoked responses for all stimulus pairs. The resulting dissimilarity matrix is correlated with a target matrix of known psychophysical distances amongst 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). We propose a new approach called *Network RSA* (NRSA) 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.

Network RSA is summarized as follows. 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 known psychophysical similarity (Tversky & Gati, 1982) 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 across stimuli encode the similarities in $\boldsymbol{S}$, as depicted in Figure 1. In RSA, the features are $\boldsymbol{X} \in \mathbb{R}^{n \times p}$, a matrix of voxel activations. Each row corresponds to activations in all p voxels in response to a stimulus, and each column corresponds to the activations in specific voxel to the n stimuli. Our generalized notion of RSA, which encompasses conventional ROI (Kriegeskorte et al., 2008) and searchlight (Kriegeskorte et al., 2006) approaches, involves finding a sparse and symmetric matrix $\boldsymbol{W} \in \mathbb{R}^{p \times p}$ such that $\boldsymbol{S} \approx \boldsymbol{X}\boldsymbol{W}\boldsymbol{X}^T$. 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.

4.1. Network RSA application: Simulated Data

Before applying our framework to real fMRI data, we consider a simulation study that allows us to compare results against a known ground-truth. We compare group lasso and GrOWL by analyzing synthetic data generated from a deep neural network model trained to generate distributed representations of a word's sound (phonology) and meaning (semantics) from its spelling (orthography; Figure 3 top left). The network structure is motivated by the influential "triangle" model of the human reading system (Plaut et al., 1996). Specifically, phonological outputs receive contributions from 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 outputs. This architecture is interesting because different kinds of similarity 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. In contrast, orthographic and semantic similarity structures are unsystematic: similarity of word spelling does not necessarily predict similarity of meaning and vice versa. In learning to map from orthography to semantics and on to phonology, the indirect path thus comes to encode quite different similarity relations amongst the words than does the direct path (Harm & Seidenberg, 2004; Plaut et al., 1996).

To capture these properties we generated model "orthographic" representations as patterns sampled from 6 overlapping clusters of binary input features, roughly corresponding to different orthographic neighborhoods. For every word a "phonological" pattern was generated by flipping each orthographic feature with probability 0.1. Thus phonological patterns were distorted variants of orthographic patterns, creating high systematicity between these. We also created a "semantic" pattern for each word from a set of binary 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 left bottom panel in Figure 3 shows the cosine distances encoded amongst the 30 "words" in each layer of one trained model. Layers in the direct path each encode roughly the same distances amongst items, while the semantic layer encodes a quite different set of distances that is weakly reflected in two of the three hidden layers in the indirect path. Thus the different components of this simple word-reading network contribute differentially to the encoding of semantic versus ortho-phonological similarity structure.

We trained 100 models with different initial weights, corresponding to 100 model subjects, and presented each with 30 orthographic inputs. Each input generated a vector of activations over the 100 model units. To ensure high redundancy amongst units this vector was concatenated 5 times and perturbed with independent noise, yielding 500 measurements per model subjects. These were treated as analogs of the estimated BOLD response at each of 500 model voxels in a brain imaging study. We then applied group lasso and GrOWL to find the voxel subsets that encode either the semantic or phonological distances (derived from target values for the output layers of the network). We fit models by searching a grid of parameters (λ, λ_1), including $\lambda_1 = 0$ as the special case of GrOWL that is group lasso. For each grid point we counted a voxel as "selected" if it received a non-zero weight, and assessed how accurately the model selected the voxels encoding phonological structure (all those along the direct pathway) or semantic structure (the semantic layer hidden layers 2 and 3 in the indirect path) by computing hit rates and false alarm rates). All three models showed low and equivalent cross-validation error; however GrOWL achieved this error rate while selecting considerably more voxels. The ROC plots in Figure 4 further show that GrOWL did not select additional voxels at random: it outperformed group lasso considerably in discriminating signal-carrying from non-signal carrying voxels. The right panel of Figure 3 shows the frequency with which each model unit is selected for the best-performing solution of each method and structure type. The strong sparsity enforced by group lasso is clearly apparent: target units are selected less consistently than with GrOWL, which consistently discovers 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 rightmost plots of Figure 3 show 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.

4.2. Network RSA application: 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 (Kriegeskorte et al., 2008), 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 p 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 $n \times p$ matrix of fMRI responses, $\boldsymbol{X}$, evoked by each item across the whole cortex. The model is then fit to optimize the objective functions specified in (3) for group lasso and (4) with squared Frobenius loss 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 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 aims to answer three questions. (1) Does either approach learn a model from whole-brain fMRI that can accurately predict the pairwise similarities among stimuli? (2) Does group lasso or GrOWL learn a more accurate model? (3) Do the fitted models identify voxels in areas that are consistent with known neural representations? To answer these questions, we applied the approach to discover voxels that work to encode the visual similarities 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 (Antani et al., 2002) and (b) it is well known that visual similarity is encoded by neural responses in occipital and posterior temporal cortices (Kriegeskorte & Kievit, 2013).

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 trash can. 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

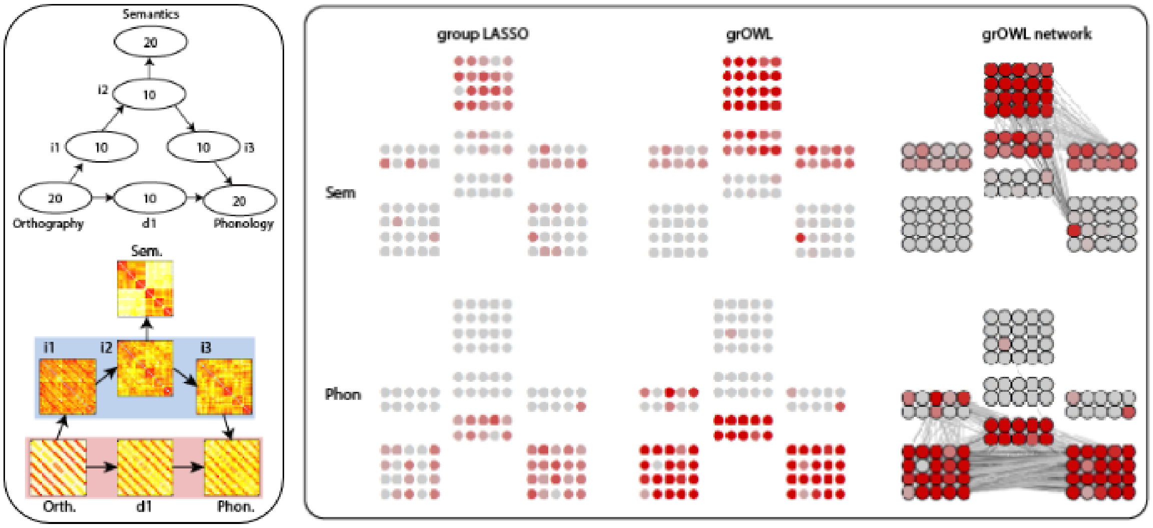

Figure 3. Left panel: Network architecture (top) and the similarity structure expressed in each layer (bottom). Red background shows the direct pathway and blue the indirect pathway from orthography to phonology. Layers in the two pathways encode different similarity structures. The target similarity matrices for the analysis express either the semantic structure (top layer) or the phonological structure (bottom right layer). Arrows indicate feed-forward connectivity. Right 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.

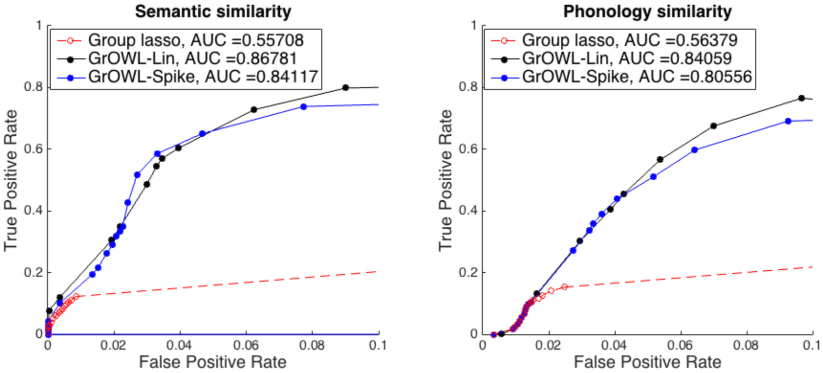

Figure 4. Trade-off curves for FPR ≤ 0.1 generated by sweeping through (λ, λ_1) values (for $\lambda = 0$, all units are selected and as λ is increased fewer units are given non-zero weight). Each point corresponds to a combination of λ and λ_1 that gives the best trade-off (where setting $\lambda_1 = 0$ results in the group lasso). The pareto-frontier for group lasso (red), GrOWL-Lin (black), GrOWL-Spike (blue) is averaged across 100 participants for each method, considering both similarity structures, Semantics (left panel) and Phonology (right panel). Note for any $\lambda > 0$, the group lasso solution will include *at most* $n = 30$ voxels, since the number of selected voxels will not exceed the number of measurements. If $\lambda = 0$, then the group lasso will select all voxels. Thus, group lasso curve beyond $n = 30$ selections (around 0.01 FPR) is shown as a dashed line, which extends linearly to the point $(FPR, TPR) = (1, 1)$.

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

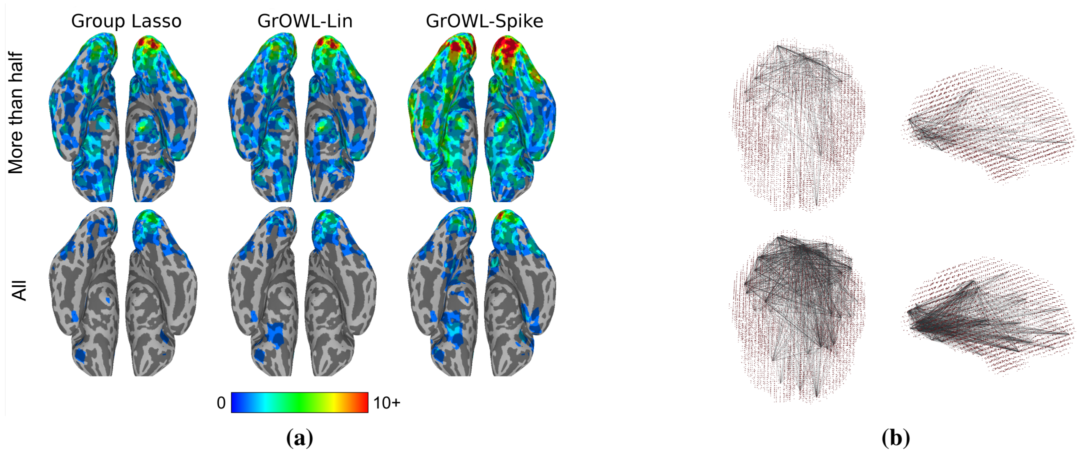

Figure 5. Panel (a) shows surface maps corresponding to group lasso (left), GrOWL-Lin (middle) and GrOWL-Spike (right) showing the voxels selected for the tuning parameters with smallest prediction error on the hold-out data for *at least five* and *all nine* cross-validations in the top and bottom rows 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). Panel (b) is a network plot showing the top edges from the $\boldsymbol{W}$ matrix for the best-performing parameterization of group LASSO (top) and GrOWL-Spike (bottom) in one subject. The thickness of the edges is proportional to the edge weights.

Target similarities. Each stimulus was a bitmap of a black-and-white line drawing. We took pairwise Chamfer distance (Borgefors, 1988) as a proxy for inter-item visual dissimilarities. r = 3 is the smallest value to attain $\|\boldsymbol{S} - \boldsymbol{Y}\boldsymbol{Y}^T\|_F / \|\boldsymbol{S}\|_F \leq 0.15$. 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 each λ and λ_1 parameter (grid points) 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, 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 with 9 different final hold-out sets.

Results. The table below shows performance on the final hold-out sets (H) for each participant and each method, considering error between predicted ($\widehat{\boldsymbol{Y}}$) and actual dissimilarities ($\boldsymbol{Y}$) where MSE = $\|\boldsymbol{Y}_H - \widehat{\boldsymbol{Y}}_H\|_F / \|\boldsymbol{Y}_H\|_F$. Both approaches show significantly non-random prediction. As in our simulations, all methods show comparable prediction error on hold-out sets. We also note that, as in the simulations, GrOWL selected almost double the number of voxels in each participant. Our assertion that both approaches show significantly non-random predictions is based on a permutation-based paired t-test, where chance would have been zero difference. By this measure, for example, GrOWL-Spike's performance is significantly better than chance (t-value=8.59, p<0.0001).

Method	MSE (p-value)
Group Lasso	0.5266 (1.045e-07)
GrOWL-Lin	0.5271 (6.456e-08)
GrOWL-Spike	0.5213 (1.774e-08)

Figure 5(a) shows the locations of selected voxels (i.e., those with non-zero coefficients) across all 23 participants for the tuning parameters with smallest mean prediction error on hold-out data, mapped into a common anatomical space with 4mm full-width-half-max spatial smoothing and projected onto a model of the cortical surface. To look into the stability of selection across different training sets, the top row shows the voxels selected for *at least five* (out of nine) cross-validation runs while the bottom row 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 picks consistently more voxels than group lasso.

Figure 5(b) shows the largest magnitude edges in the $\boldsymbol{W}$ matrix for the best-performing parameterization of group LASSO (top) and GrOWL-Spike (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 reveals a much denser network. The results suggest the possibility that subregions of frontal and temporal cortex may, together with occipito-temporal cortex, participate in networks that serve to encode visual similarity structure.

Acknowledgements

This work was partially supported by NSF grant CCF-1218189.

References

Antani, Sameer, Kasturi, Rangachar, and Jain, Ramesh. A survey on the use of pattern recognition methods for abstraction, indexing and retrieval of images and video. *Pattern recognition*, 35(4):945–965, 2002.

Atzmon, Yuval, Shalit, Uri, and Chechik, Gal. Learning sparse metrics, one feature at a time. *Journal of Machine Learning Research*, 1:1–48, 2015.

Bogdan, Malgorzata, Berg, Ewout van den, Su, Weijie, and Candes, Emmanuel. Statistical estimation and testing via the sorted l1 norm. *arXiv preprint arXiv:1310.1969*, 2013.

Bogdan, Malgorzata, Berg, Ewout van den, Sabatti, Chiara, Su, Weijie, and Candes, Emmanuel J. Slope–adaptive variable selection via convex optimization. *arXiv preprint arXiv:1407.3824*, 2014.

Bondell, Howard D and Reich, Brian J. Simultaneous regression shrinkage, variable selection, and supervised clustering of predictors with oscar. *Biometrics*, 64(1):115–123, 2008.

Borgefors, Gunilla. Hierarchical chamfer matching: A parametric edge matching algorithm. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 10(6):849–865, 1988.

Figueiredo, M.A.T. and Nowak, R. D. Ordered weighted l1 regularized regression with strongly correlated covariates: Theoretical aspects. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, pp. 930–938, 2016.

Harm, Michael W and Seidenberg, Mark S. Computing the meanings of words in reading: cooperative division of labor between visual and phonological processes. *Psychological review*, 111 (3):662, 2004.

Kriegeskorte, Nikolaus and Kievit, Rogier A. Representational geometry: integrating cognition, computation, and the brain. *Trends in cognitive sciences*, 17(8):401–412, 2013.

Kriegeskorte, Nikolaus, Goebel, Rainer, and Bandettini, Peter. Information-based functional brain mapping. *Proceedings of the National Academy of Sciences of the United States of America*, 103(10):3863–3868, 2006.

Kriegeskorte, Nikolaus, Mur, Marieke, and Bandettini, Peter. Representational similarity analysis–connecting the branches of systems neuroscience. *Frontiers in systems neuroscience*, 2, 2008.

Kulis, Brian. Metric learning: A survey. *Foundations and Trends in Machine Learning*, 5(4):287–364, 2012.

Lounici, Karim, Pontil, Massimiliano, Tsybakov, Alexandre B, and Van De Geer, Sara. Taking advantage of sparsity in multitask learning. *arXiv preprint arXiv:0903.1468*, 2009.

Lounici, Karim, Pontil, Massimiliano, Van De Geer, Sara, and Tsybakov, Alexandre B. Oracle inequalities and optimal inference under group sparsity. *The Annals of Statistics*, pp. 2164–2204, 2011.

McRae, Ken, Cree, George S, Seidenberg, Mark S, and McNorgan, Chris. Semantic feature production norms for a large set of living and nonliving things. *Behavior research methods*, 37 (4):547–559, 2005.

Obozinski, Guillaume, Wainwright, Martin J, and Jordan, Michael I. Support union recovery in high-dimensional multivariate regression. *The Annals of Statistics*, pp. 1–47, 2011.

Parikh, Neal and Boyd, Stephen. Proximal algorithms. *Foundations and Trends in optimization*, 1(3):123–231, 2013.

Plaut, David C, McClelland, James L, Seidenberg, Mark S, and Patterson, Karalyn. Understanding normal and impaired word reading: computational principles in quasi-regular domains. *Psychological review*, 103(1):56, 1996.

Rau, M., Mason, B., and Nowak, R. D. How to model implicit knowledge? Similarity learning methods to assess perceptions of visual representations. In *Proceedings of the 9th International Conference on Educational Data Mining*, 2016.

Shaver, Phillip, Schwartz, Judith, Kirson, Donald, and O'connor, Cary. Emotion knowledge: further exploration of a prototype approach. *Journal of personality and social psychology*, 52(6): 1061, 1987.

Shepard, Roger N. Multidimensional scaling, tree-fitting, and clustering. *Science*, 210(4468):390–398, 1980.

Tversky, Amos and Gati, Itamar. Similarity, separability, and the triangle inequality. *Psychological review*, 89(2):123, 1982.

Ying, Yiming, Huang, Kaizhu, and Campbell, Colin. Sparse metric learning via smooth optimization. In *Advances in neural information processing systems*, pp. 2214–2222, 2009.

Zeng, Xiangrong, Figueiredo, Mario, et al. The atomic norm formulation of oscar regularization with application to the frankwolfe algorithm. In *Proceedings of the European Signal Processing Conference, Lisbon, Portugal*, 2014.

Zou, Hui and Hastie, Trevor. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, 2005.