Sparse, Low-Dimensional and Multimodal Representations of Time Series for Mind-Reading

Kiran Vodrahalli, Lydia T. Liu, Niranjani Prasad {KNV, LTLIU, NP6}@PRINCETON.EDU

Abstract

We introduce a new lens through which to analyze time-series brain data, emphasizing the importance of **sparse**, **low-dimensional** joint representations of MEG and EEG data which retain predictive power and generative modeling capabilities. Our main contribution is empirical validation suggesting that multimodal sparse CCA is able to achieve a low-dimensional (e.g. 20, 40-dim) representation of time series brain data which retains predictive and generative power.

We create featurizations of MEG data with sparse CCA and test the retained informativity of the low-dimensional space by learning a linear model between convolutional neural network image features and frequency features for MEG. The predictive power of this brain decoding task indicates that the low-dimensional space retains a surprising amount of information about the content of thoughts and generalizes across subjects. We also validate our methods using EEG representations of fMRI data. By using spatial information encoded by paired fMRI data with sparse CCA (sCCA), we verify that sCCA joint representations have predictive power by training SVMs to distinguish between states of attention and standard states, achieving a highest F1 score of 0.533 with the sCCA representation.

1. Introduction

1.1. Motivation

Three primary non-invasive forms of data are collected to study the human brain: functional magnetic resonance imaging (fMRI), electroencephalography (EEG), and magnetoencephalography (MEG). fMRI is perhaps the most popular approach, due to its high spatial resolution, though it is limited in temporal resolution. This helps explain why most experiments which use fMRI do not model its time series properties and instead focus on analyzing snapshots (Mitchell et al., 2008). On the other hand, both EEG and MEG have high temporal resolution and are thus sensitive to changes over time.

Two interesting questions arise if we want to better understand how temporal patterns in brain activity correlate with stimuli from the outside world. First, we can ask to what extent is it possible to use additional information derived from the spatially resolved fMRI data to build a predictive model of a target stimulus, across different forms of external audio or visual stimuli. Second, we would like to explore the possibility of building a generative time series model of temporal data given an external stimulus: For instance, an image of an object. That is, given the image, can we generate a time series signal of brain activity or at least, a representation of the time-series signal? Furthermore, can we reverse our model so that we can decode the time-series input to reproduce the image seen?

This paper tackles these two questions with the additional requirement of solving these problems in a low-dimensional space derived using sparse methods, and demonstrates that both may be solved reasonably convincingly with sparse CCA, a technique for combining multimodal time series data.

Few people have investigated the task of building a generative model for non-fMRI neuroimaging data. It is possible to use fMRI as a time series; however, this is more a sequence of a small number of values over a long period of time (temporal resolution is quite low). Also, the only approaches which attack this problem attempt to match a time series of images with a matching time series of fMRI snapshots. Thus, their perspective is still that of matching a single fMRI snapshot to a stationary image. Our approach is rather to match a stationary image to a time-series of brain activity, encoding the assumption that there are generator patterns in the image which induce some periodicity in the brain signal.

Preliminary work - Under review.

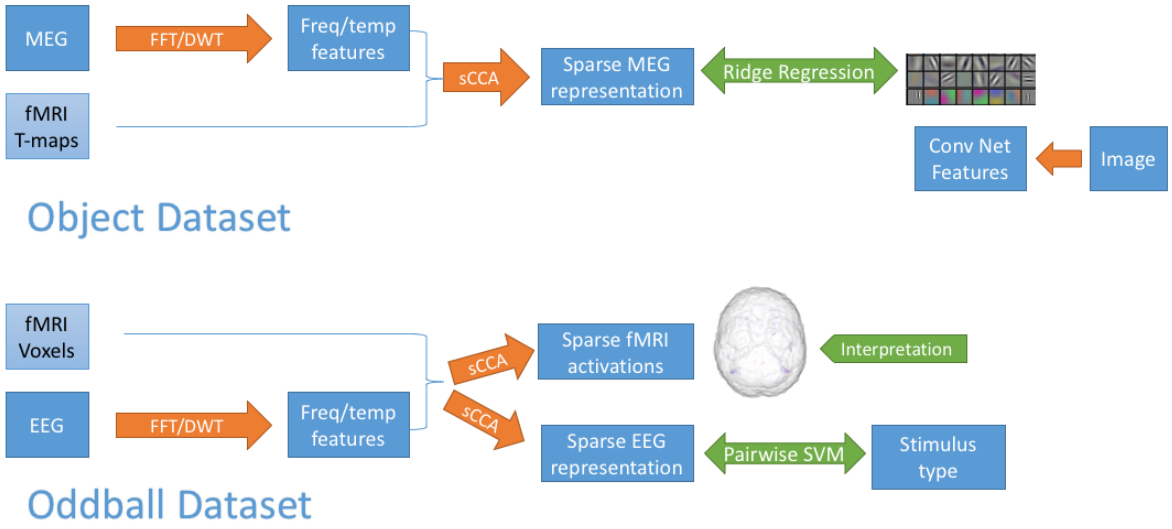


Figure 1. Summary of data analysis and generative modelling procedures employed in paper

1.2. Datasets

We investigate two datasets in this work, namely the MEG-fMRI "Object" dataset (Cichy et al., 2014) and the EEG-fMRI "Oddball" dataset (Walz et al., 2013). Both are paired with the presentation of some external stimuli.

1.2.1. MEG-fMRI

In the Object dataset, both fMRI and MEG data are collected while subjects look at 92 different images, each of an object with various classifications (human vs. non-human, natural vs. man-made, and so on). MEG recordings are taken at 306 different points on the scalp for 1300 ms (from 100ms before to 1200ms after the image is presented) for 20 different subjects.

1.2.2. EEG-fMRI

Our primary dataset is the Auditory and Visual Oddball EEG-fMRI dataset (Walz et al., 2013), available for download at https://openfmri.org/dataset/ds000116. The experiment is set up as follows: 17 healthy subjects performed separate but analogous auditory and visual oddball tasks (interleaved) while simultaneous EEG-fMRI was recorded. There were 3 runs each of separate auditory and visual tasks. Each run consisted of 125 total stimuli (each of duration 200 ms): 20% were target stimuli (requiring a button response) and 80% were standard stimuli (to be ignored). The first two stimuli in the time course are constrained to be standard stimuli, and the inter-trial interval is assumed to be uniformly distributed over $2 - 3$ seconds.

The fMRI data is an EPI sequence with 170 TRs per run, with 2 sec TR (time between scans) and 25 ms TE (echo time). There are 32 slices, and no slice gap. The spatial resolution is $3mm \times 3mm \times 4mm$. For more details on the preprocessing steps performed for fMRI data, refer to (Walz et al., 2013).

The EEG data was collected at a 1000 Hz sampling rate across 49 channels. The start of the scanning was triggered by the fMRI scan start. The EEG clock was synced with the scanner clock on each TR. We use the gradient-free EEG data provided.

1.3. Previous Work

1.3.1. Fusing modes of brain data

In the field of exploiting multimodal neuroimaging data, data fusion is defined as the use of supervised or unsupervised machine learning algorithms to combine multimodal datasets.

A review of the most widely used methods for data fusion is given in (Dahne et al., 2015). These are either late fusion methods, where information from one modality is not used to extract components from another, and early fusion, in which data from both modalities are decomposed together. Late fusion methods include both supervised approaches, (either using an external target signal such as stimulus type or response time) or asymmetric fusion where features from one modality are used as labels/regressors to extract factors from another modality), as well as unsupervised techniques relying on data stats,

such as PCA or ICA.

Two common forms of early fusion include joint ICA (in which features from multiple modalities are simply concatenated) and CCA. In CCA, we find the transformations for each modality that maximise the correlation between the time courses of the extracted components This method relaxes independent component assumption of joint ICA, and does not constrain component activation patterns to be the same for both modalities.

1.3.2. Sparsity and Low-dimensional Representation of EEG and fMRI

The literature itself is rather sparse on the application of sparse methods to multimodal time series brain data. (Deligianni et al., 2014) apply sparse-CCA with randomized Lasso to fMRI-connectome and EEG-connectome for resting-state data (i.e., with no supervised task) to identify the connections which provide most signal. They analyze the distance between precision matrices of the Hilbert envelopes for fMRI and EEG. Assuming brain activity patterns are described by a Gaussian multidimensional stationary process, the covariance matrix fully characterizes the statistical dependencies among the underlying signals.

1.3.3. fMRI-Image Decoding

In the past, the Gallant lab at Berkeley has produced a fixed fMRI representation (Naselaris et al., 2009), and more recently, has been able to generate a video time series given an fMRI input and also has a voxel prediction model for fMRI based on a movie input (Nishimoto et al., 2011). In the 2009 paper, they only produce a fixed time prediction for an image. In the 2011 paper, they produce a time series given a time series of images (i.e. a video).

1.3.4. Object dataset

(Cichy et al., 2014) use MEG and fMRI data to analyze the hierarchy of the visual pathway in the brain applied to object recognition. They use MEG to localize image processing in the brain through time, and fMRI to spatially localize the voxels which are involved in the processing. They validate performance with plots of predictive power based on MEG signal over time, and noting by eye that peaks correspond to neuroscientifically-known time points in the visual process. In more recent unpublished work, Cichy uses convolutional neural networks to featurize object images and then applies Representational Similarity Analysis (RSA) to conclude that the stages of the visual recognition pathway in the brain somewhat correspond to layers of the convolutional network.

1.3.5. Oddball dataset

(Walz et al., 2013) use the Oddball dataset to train a linear classifier to maximally discriminate standard and target stimuli. They create an EEG regressor out of the mean classifier output (convolved with hemodynamic response function) and use the EEG regressor, combined with other stimulus and response related regressors, to fit a linear model to fMRI data, and comment on the correlation based on the coefficients. Also, they manually looked at fMRI images at TRs that show a high degree of correlation with the regressors, to form qualitive conclusions on how well the data agrees with known neuroscientific models. with previous work.

2. Material and Methods

2.0.1. CCA and sCCA

In traditional canonical correlation analysis, we have two sets of measurements $X \in \mathbb{R}^{n\times p}, Y \in \mathbb{R}^{n\times q}$ collected on the same set of underlying phenomena, where n is the number of observations and p, q are the number of measurement channels (features) for X and Y respectively.

We posit that the features that contain pertinent information about the underlying phenomena in the two datasets are strongly correlated, since they are measurements on the same underlying phenomenon. Thus we want to find u, v that maximizes $\text{cor}(Xu, Yv)$. If X, Y are mean-centered and scaled, we have the following problem:

$$\max_{u\in\mathbb{R}^p, v\in\mathbb{R}^q} u^T X^T Y v \tag{1}$$

subject to $u^T X^T X u \leq 1$ and $v^T Y^T Y v \leq 1$. u, v are called canonical vectors. Subsequent pairs of canonical vectors maximize the same objective function with the added constraint that they are uncorrelated to the previous pairs.

We can induce sparsity in the canonical vectors directly by using sCCA, a penalized version of traditional CCA (Witten et al., 2009).

2.1. MEG Decoding and Generation

We now investigate the quality of low-dimensional representations sparse methods create for paired MEG-fMRI time-series data (from (Cichy et al., 2014)). Recall that the experiment the data was collected from involved subjects looking at 92 different images for short periods of time. Separate trials were used to collect fMRI versions of the data and MEG versions of the data.

Our goal is to demonstrate that a low-dimensional representation utilizing both the fMRI and MEG data can achieve comparable or better predictive performance than

other representations. Therefore, we use two representations of the MEG data:

1. Fourier Transform Coefficients (dimension $306 * 64$): In this case, we do not use the fMRI data. Since our featurization is built of 64 Fourier coefficients for each of the 306 MEG recording locations on the scalp, predicting this FFT featurization from an image is equivalent to generating MEG time-series data.
2. Sparse CCA Shared Feature Space with Wavelet MEG and fMRI T-maps: We construct a 20-dimensional shared feature space using the sCCA algorithm over MEG and fMRI paired over time.

We use ridge regression to learn a linear map from the image data to the low-dimensional representation. We then evaluate the quality of this map by predicting which MEG featurization corresponds to a given image from a set of 92 unlabeled MEG featurizations.

2.1.1. Featurizing Images

We pre-processed the image representations (X) by scaling each feature vector by $\frac{1}{\|\cdot\|_2}$. We also scaled the vectors representing the 92 responses for one dimension of the MEG feature vector.

In order to represent the images, we try a few different featurizations. First, the image itself (which is a $175 \times 175 \times 3$ color image) is a valid featurization and a baseline. Then, we examine a lower dimensional representation of the image derived from simply using PCA. Finally, we use a pre-trained convolutional neural network (CNN) (Jia et al., 2014) to produce activations upon sending our object images through the network. We then take a subset of these activations and join them together to give a low-dimensional featurization of the image ($\sim$ 3000-dimensional). Note that it is necessary to choose such a low-dimensionality to represent the image so that fitting our model takes place in a reasonable amount of time. Future work may involve using higher-dimensional representations of the images.

2.1.2. Featurizing MEG data

Fast Fourier Transform In order to determine the way MEG time series data encodes visual and semantic information about an object, we first introduce a hypothesis: In order to compare an static image to a time series, we must first extract parameters of the time series that describe its behavior over time. For instance, this might relate to the periodicity of the data.One approach to encoding this kind of information in a featurization is to examine the frequency values of the time series. As with the EEG data, we use FFT to get coefficients for each frequency class. We evaluate coefficients at 64 frequencies, between 1Hz and 64Hz, to obtain a low-dimensional representation of the MEG data.

Wavelets A more principled approach to frequency featurization, that also retains some temporal information, is wavelets. We run DWT decomposition on each 1300ms MEG time series, again using the Daubechies family of filters, to obtain a 195-dimensional feature vector comprising a set of time series with information in different frequency bands.

Sparse Multiple CCA Next, to 'fuse' the featurized MEG with the fMRI data, we apply sparse CCA to the wavelet featurization of the MEG data paired and fMRI T-maps to produce a 20-dimensional space for each data modality.

Sparse CCA, as aforementioned in 3.0.1, can be used to combine two sets of measurements, namely MEG and fMRI, on the same set of underlying phenomena. However, we also want to use data from multiple human subjects. One way to extend CCA to handle data from multiple trials and subjects is to treat the data from a different patient as yet another set of measurements. For this purpose, we can use *sparse multiple CCA*, an extension of sparse CCA to the case of $L > 2$ data sets $X_1, ..., X_L$ with features on a single set of samples (Witten & Tibshirani, 2009).

It seems sound to perform sparse multiple CCA on bimodal data (such as EEG-fMRI or MEG-fMRI) from different subjects, as long as the subjects are observing the same underlying phenomena. While this does not apply well to the Oddball dataset, which we discuss later in section 3.3.3, in the Object dataset, every subject observe the same 92 images under the same experimental conditions. Thus it is reasonable to think that CCA will be able to recover the projection that maximizes the correlation between the MEG and fMRI data modalities, as well as account for minor inter-subject differences.

Therefore, we performed sparse *multiple* CCA on the paired MEG and fMRI T-maps from $K = 3$ different subjects.

That is, given the set of data matrices $\Omega = X^{(1)}, \cdots, X^{(K)}, Y^{(1)}, \cdots, Y^{(K)}$, where $X^{(i)}$ is the featurized MEG from subject i and $Y^{(i)}$ is the fMRI T-maps from subject i all in response to the same 92 image stimuli, we find sparse $w_1, ..., w_{2K}$ ($2K = |\Omega| = 6$) such that $\sum_{M_i, M_j \in \Omega;\ i<j} w_i^T M_i^T M_j w_j$ is large. As before, we refer to these $w_i's$ as canonical components or activations.

Note that in the above formulation of the problem $M_i \in \mathbb{R}^{n \times p_i}$ ($n = 92$) and $w_i \in \mathbb{R}^{p_i}$, so this gives us one canonical component per data matrix. However, by an iterative algorithm detailed in (Witten & Tibshirani, 2009), we can

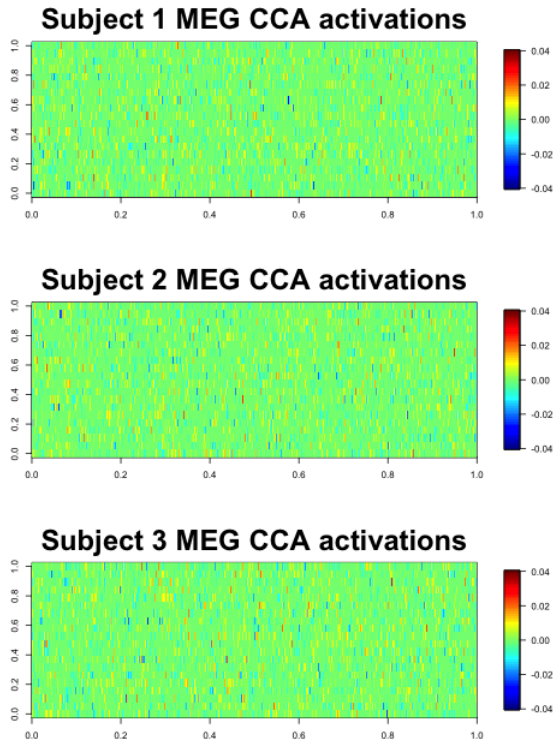


Figure 2. CCA components ("activations") for the featurized MEG of subjects 1 to 3. The Y axis is the 20 canonical components and the X axis is the Wavelet transform features.

successively obtain multiple—20 in our case—canonical components for each data matrix.

The MEG canonical components for Subjects 1 - 3 are visualized in Figure 2. First, we can see that the MEG activations are truly sparse. They do not appear to be very similar across subjects, from visual inspection and by examining the matrix norm of the pairwise differences. While we had expected canonical components to be more similar than different because the subjects are experiencing the same image stimuli after all, inter-subject differences in response (such as response time and intensity) may account for the seemingly large difference in canonical components. In order to more deeply evaluate how 'different' these MEG activations are across subjects, we will need to develop more rigorous measures of subspace similarity, but for now we will just use these canonical components to obtain features. Specifically, we project the MEG features from Wavelet transform unto their canonical components and use the resulting projections in ridge regression.

2.1.3. Ridge Regression

There are several ways we can try to fit the relationship between the image featurization and the MEG data featurization. We choose ridge regression for its low complexity. Letting Y be the featurized MEG data of dimension $m = 64 * 306$ in the FFT case and $m = 20$ in the sCCA case, and x be the featurized image input data of dimension $p = 2923$, we desire to learn C such that $Y \approx Cx$ for each of the $n = 92$ possible x. Thus, $Y = \mathbb{R}^m$ be the MEG dimension, $C \in \mathbb{R}^{m \times p}$, and $x \in \mathbb{R}^p$. This problem becomes ridge regression m times over, where each ridge regression is to learn row $C_i \in \mathbb{R}^p$ of the matrix C for $i \in [m]$. We let $\hat{y}_i \in \mathbb{R}^n$ be the values of Y_i for each of the n objects, and $X \in \mathbb{R}^{n \times p}$ be a concatenation of the p-dimensional featurizations of each of the n object images. Then, the ridge regression problem is given by

$$\text{argmin}_{C_i} \|\hat{y}_i - XC_i\|_2^2 + \lambda \|C_i\|_2^2 \tag{2}$$

for some hyperparameter λ, which has closed form solution $C_i = \left(X^T X + \lambda I\right)^{-1} X^T \hat{y}_i$. We solve for each row and concatenate them as

$$C = \begin{bmatrix} - - - C_1 - - - \\ - - - C_2 - - - \\ \vdots \\ - - - C_m - - - \end{bmatrix}$$

Therefore, we will have a linear map C from X to Y, and given a new image input z, we can featurize to get $f(z)$ and then apply $\tilde{Y} = Cf(z)$ to approximate the feature vector for MEG activity. Depending on the featurization of MEG activity (FFT or sCCA), we can use the featurization to recreate the time-series MEG data itself, given the original image z.

Incidentally, this approach is also reversible via convex optimization. Suppose we are given a new MEG sample w. We can featurize it using $g(w)$, and then solve the following convex optimization problem:

$$\text{argmin}_\theta \|g(w) - C\theta\|_2^2 \tag{3}$$

Assuming that f has an inverse f^{-1} or an approximation to an inverse, we can recover the image the subject was looking at via $f^{-1}(\theta)$, in essence performing brain decoding of time-series MEG data. The idea for this approach is due to (Mitchell et al., 2008), where the authors applied similar approaches to text data and fMRI data (not time-series). (Naselaris et al., 2009) also applied a similar approach to fixed images and fixed fMRI data. The novelty here is extending the approach to time-series MEG data.

In order to actually perform the ridge regression, we ran Matlab code in parallel on Fujitsu RX200 S8 servers with dual, eight-core 2.8GHz Intel Xeon E52680 v2 processors with 256GB RAM running the Springdale distribution of Linux. We implemented ridge regression so that it was easily parallelizable by re-writing the solution $\left(X^T X + \lambda I\right)^{-1} X^T \hat{y}_i$ as $V \left(\Sigma^2 + \lambda I\right)^{-1} V^T \hat{y}_i$, utilizing

the singular value decomposition $X = U\Sigma V^T$ where U and V have orthonormal columns and Σ is diagonal, which only needs to be calculated once. Notably, we avoid having to recalcuate difficult inverses. This implementation is also more stable for smaller values of λ. The speedup from this algorithmic change is dramatic: Finding C took 2 hours with the standard Matlab implementation `ridge`, and just over one second with our implementation for an approximate $5500\times$ speedup. This speedup was essential to the calculation of many of our results in a reasonable amount of time.

We also had to choose a value for the λ-parameter involved in ridge regression: We tested 100 different λ values logarithmically evenly spaced across $[10^{-8}, 10^{8}]$. We report the performance for all λ-values.

After learning the C matrix, we are given a featurized MEG vector y^* with unknown label. Then for each x_i, $i \in [n]$, we can approximate y^* with Cx_i. Ideally, the correct class i^* has $d(Cx_{i^*}, y^*)$ is the smallest over all $\{x_i\}_{i=1}^n$ for some distance measure d. We use the negative cosine distance for d. Formally, the predicted class is

$$\text{argmax}_i \frac{\langle Cx_i, y^* \rangle}{\|Cx_i\|_2 \|y^*\|_2} \tag{4}$$

Ranking the x_i (each i is a class) by cosine distance allows us to assign a rank $\in [n]$ to the the correct class i^* as a measure of quality of the learned matrix C. Therefore, the best rank is 0 (no other image-predictions are closer to the true MEG featurization) and the worst rank is 91 (all other images are closer). Note that random chance would give i^* a rank of $n/2 = 45$.

For both the FFT and sCCA featurizations, we examine two forms of generalization: subject generalization and image generalization. For the former, we test how well the C matrix learned on one set of individuals generalizes to a different individual. For the latter, we test how well C learned on a subset of 92 images generalizes to an out-of-set image. To test subject generalization for the FFT featurization, we learned a matrix C_{12} where the MEG frequency responses for subject 1 and subject 2 was averaged. Then, we predicted the MEG response for subject 3 and evaluated performance. In the case of sCCA, we averaged over the first three subjects and tested on the fourth. To test image generalization for the FFT representation of the MEG data, we randomly sampled 5 distinct object classes in $[n]$. For each left-out image l, we learned a matrix C_{-l} where no training samples from image l were seen. Due to the small sample size ($n = 92$) and the low-dimensionality of our image feature embedding, we did not expect great generalization over images. Since the sCCA representation was much more low-dimensional (only 20 dimensions), we were able to learn C_{-l} for all 92 possible left-out images.

2.2. EEG-fMRI fusion

In this section, we show how sparse CCA can be used for fusion of EEG and fMRI data modalities and produce potentially interpretable brain activation components.

In order to investigate the efficacy of combining fMRI and EEG data, we define a prediction task as follows: Detect whether a signal at a given time point is a target signal or a non-target signal. Our goal is to demonstrate that with a sparse, low-dimensional representation of both the fMRI and EEG data, we can achieve comparable predictive performance as in the setting where we do not use the sparse low-dimensional representation. Since we have paired EEG-fMRI data over time, we use the Canonical Correlations Analysis (CCA) algorithm to map the dual-input into a low dimensional embedding space. In order to use sparsity, we use a variant of CCA known as sparse Canonical Correlations Analysis (sCCA). We use a popular discriminative classifier, the Support Vector Machine (SVM) to train using the low-dimensional representation to detect whether or not there was a target stimulus at a given point in time. We perform experiments for both the audio stimuli and the visual stimuli.

2.2.1. Data Cleaning

We examine the data from a single experiment for one subject which consisted of 125 audio stimuli over a 340 second duration. After excluding stimuli for which the subject response was incorrect, we select segments of the EEG time series and the fMRI data that correspond to each stimuli and treat each of these as an example. More specifically, each example is -100 to 900 ms of EEG time locked to one stimulus, and 3 consecutive TRs of fMRI, where the first TR is from the time slice coinciding with the onset of the stimulus.

We want to know if we can extract enough information from these snippets of EEG and fMRI to determine if the stimulus that occurred was standard or target.

2.2.2. Frequency Transform

To explore the effectiveness of frequency-space representations of the EEG time series, we also looked at first featurizing EEG using Fourier Transform and Wavelets transform before applying CCA.

We performed the Fast Fourier Transform (FFT) to extract coefficients from 1000ms segments of the EEG recordings, time-locked to the onset of each stimulus. We retained the first 64 coefficients, from 1Hz to 64Hz. This comprises the majority of brain activity (most relevant are gamma waves generated during conscious perception, typically at 40Hz) while filtering out high frequency artifacts.

We also looked at featurization using the Discrete Wavelet Transform (DWT) which allows us to keep both frequency and temporal information in the EEG. Multilevel wavelet decomposition was run using Daubechies 4-tap wavelet filters, to produce of set of time series with information from the delta (0-4Hz), theta(4-8Hz), alpha (8-16Hz), beta (16-32Hz) and gamma (32-64Hz) frequency bands, along with a low-frequency approximation of the EEG signal. This featurization has been widely used in literature for representing EEG, for example (Yu et al., 2010) and (Tumari et al., 2013). This set of multiresolution time series is concatenated to produce a 157-dimensional feature vector.

3. Results

3.1. MEG-fMRI Ridge Regression Performance

3.1.1. Regression results and discussion for FFT representation

First we present the performance of the FFT featurization of in Figure 3. We present training and subject generalization results for image 5, and average over four randomly chosen images to produce the image generalization graph. For small λ, the linear regression is perfect on the training set, and even for very large lambda, the rank of the correct image drops to no worse than second place. In other words, we attained perfect accuracy in predicting which image was connected with a given MEG frequency feature vector on the training data. This result could potentially be due to overfitting. Even though we use the cosine distance, $\|x-y\|_2^2 = \|x\|_2^2 + \|y\|_2^2 - 2\langle x, y\rangle$. Since $\|y\|_2^2 = 1$ for all y and ridge regression makes $\|x\|_2^2 = c$, some constant, we have that minimizing l_2 distance is equivalent to maximizing cosine distance, and thus we are effectively training for the same objective, explaining why overfitting despite using a different training metric can happen.

However, regarding the portion of Figure 3 dealing with subject generalization, we see that for large lambda the rank was at lowest (and best) 31. This means that 31 of 91 other images were closer to the true MEG featurization, and thus the correct answer ranked in the top 35%, which is somewhat better than average. Thus we can claim some generalization across different brains.

The case for image generalization is considerably less convincing. The third image of Figure 3 demonstrates that the rank was below 60 for all λ; in other words, doing worse than average.

3.1.2. Regression results and discussion for sCCA multimodal representation

We present the performance for the sCCA of the wavelet-MEG and T-map-fMRI data in Figures 4 and 5. We present

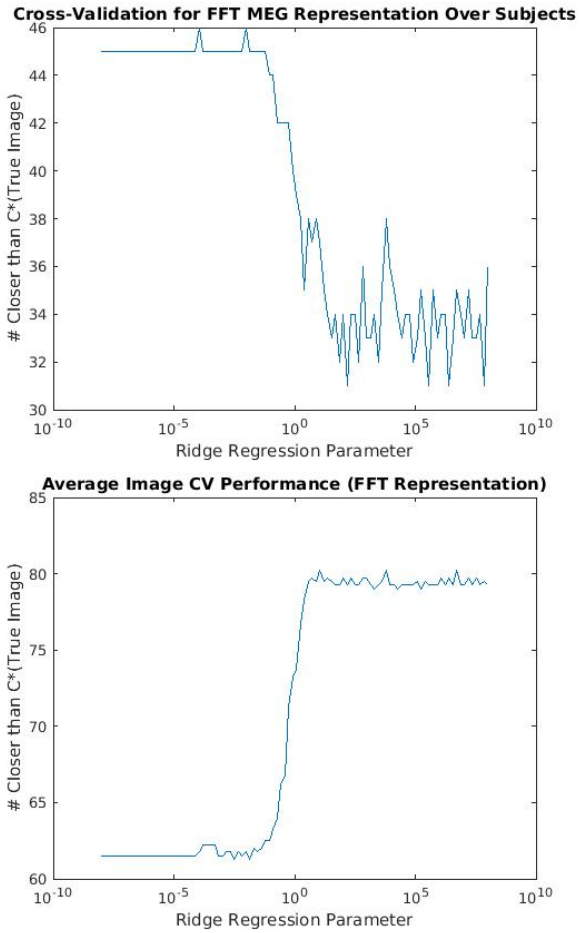


Figure 3. FFT Featurization Performance

training and subject generalization results for image 5. For the image generalization results, we exhibit an average over all left-out images as well as the performance for image 5, 13, 78, and 77, which were selected as representative of very-low rank, good-rank, and bad-rank respectively.

For training, performance was perfect, as with the FFT represenation. The same argument for potential overfitting applies. In Figure 4, the first image demonstrates that subject generalization has much better performance than the FFT representation: In fact, performance is perfect. The second image of Figure 4 demonstrates that on average, image generalization is better than for the FFT representation, nearly attaining random performance with low λ. The improvement upon the FFT representation is still considerable, with the sCCA representation gaining nearly 13% in rank.

However, Figure 5 demonstrates that there is more to the story for image generalization. Images 5 and 13 have a top rank of 9 and 1 respectively, meaning that the linear maps C_{-5} and C_{-13} learned without seeing images 5 and 13 do an excellent job of generating a 20-dimensional MEG rep-

resentation for these images. Image 78 achieves a rank less than 30, which is also quite good. Image 77 is an example of one of the images which does not generalize under the learned C_{-77}. Roughly half of the images have a rank less than 50, and 10 images have rank ≤ 10. There are also some images which perform with very low rank scores. We summarize the distribution of ranks in Figure 6.

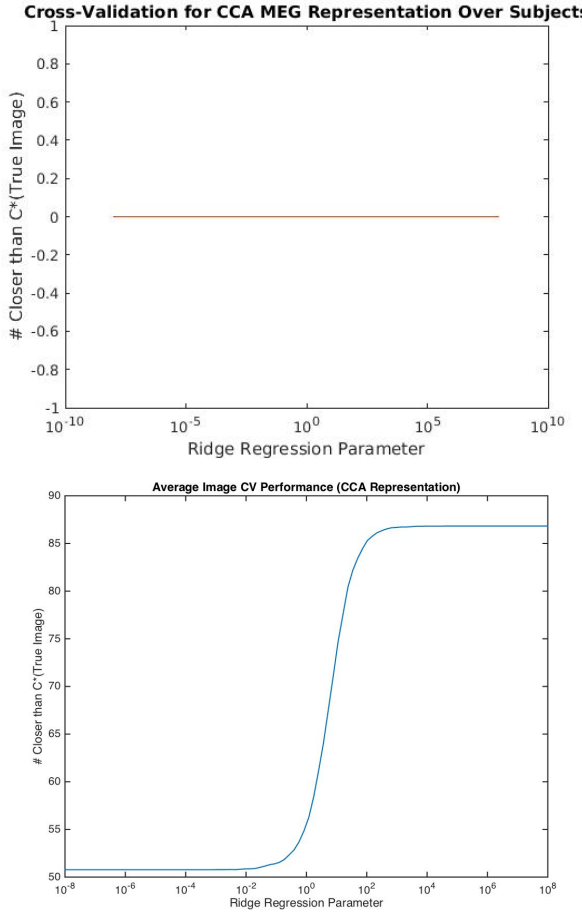


Figure 4. sCCA Featurization Performance

Now we attempt to find a pattern in the images which achieved generalization. We consider the ten images which got ranks ≤ 10. In Figure 7 we exhibit each of the images and its classification. Regarding classifications, we notice that the majority of the top images were inanimate and non-human. Looking at the images themselves, a large number of them tend to be circular in shape.

The better subject and image generalization of the 20-dimensional sCCA representation of combined MEG and fMRI data suggests that multimodal, low-dimensional representations derived from sparse methods retain predictive power in the MEG setting, in addition to the EEG setting.

3.2. Interpretability of EEG-fMRI canonical vectors

To investigate the effect of the number of canonical vectors used to project the data, we compute the 20-dimensional as well as the 40-dimensional CCA space for the paired EEG-fMRI.

The canonical vector for fMRI from the pair of canonical vectors with the highest correlation (40-dimensional CCA) is visualized in Figures 8 and 9, after minimal smoothing with a Gaussian kernel (radial with σ=0.65).

The blue points indicate negative coefficients in the canonical vector while the red points indicate positive coefficients. The transparency of the points are scaled according to the magnitude of the coefficients. Thus we can view the more intensely colored regions as highly 'activated' regions that were found to be most correlated with EEG activity. Notably, even though sCCA does not enforce any form of spatial regularization, the canonical vector activations clearly exhibit some spatial clustering. This could suggest that the canonical vectors are indeed picking out voxels in a way that is consistent with the regions of brain function (rather than in completely random locations); thus we can hope that the canonical vectors lend themselves reasonably to neuroscientific interpretation.

In Figure 8, the intensely colored region at the back of the brain corresponds to the some of the strong correlates in fMRI that were found in (Walz et al., 2013). In Figure 10, the highlighted regions, which are symmetric, appear to correspond to the visual cortex. We also note that the activations for the 2nd highest correlation canonical vector looks similar for visual and audio stimuli (Figures 9 and 11).

Another observation of interest is that the locations of the activations appear similar across the TRs, while the color or transparency of the activated voxels do differ slightly. This is again good sign that suggests that the canonical vectors are picking out functionally meaningful brain regions and tracking their development over time. While this is not our objective, the particular activated voxels that change over the TRs should be of interest to neuroscientists who might find a connection with brain function.

3.2.1. Other benchmarks

Our hypothesis is that CCA gives the most 'informative' low-dimensional projection of the EEG time series data. Therefore, we benchmark against other projection schemes, such as PCA and Random Projection.

Our random projection matrix was generated using a random Gaussian matrix with zero mean and variance $1/\sqrt{40}$, based on suggestion in (Indyk & Motwani, 1998), where 40 is the chosen dimension of the projection space.

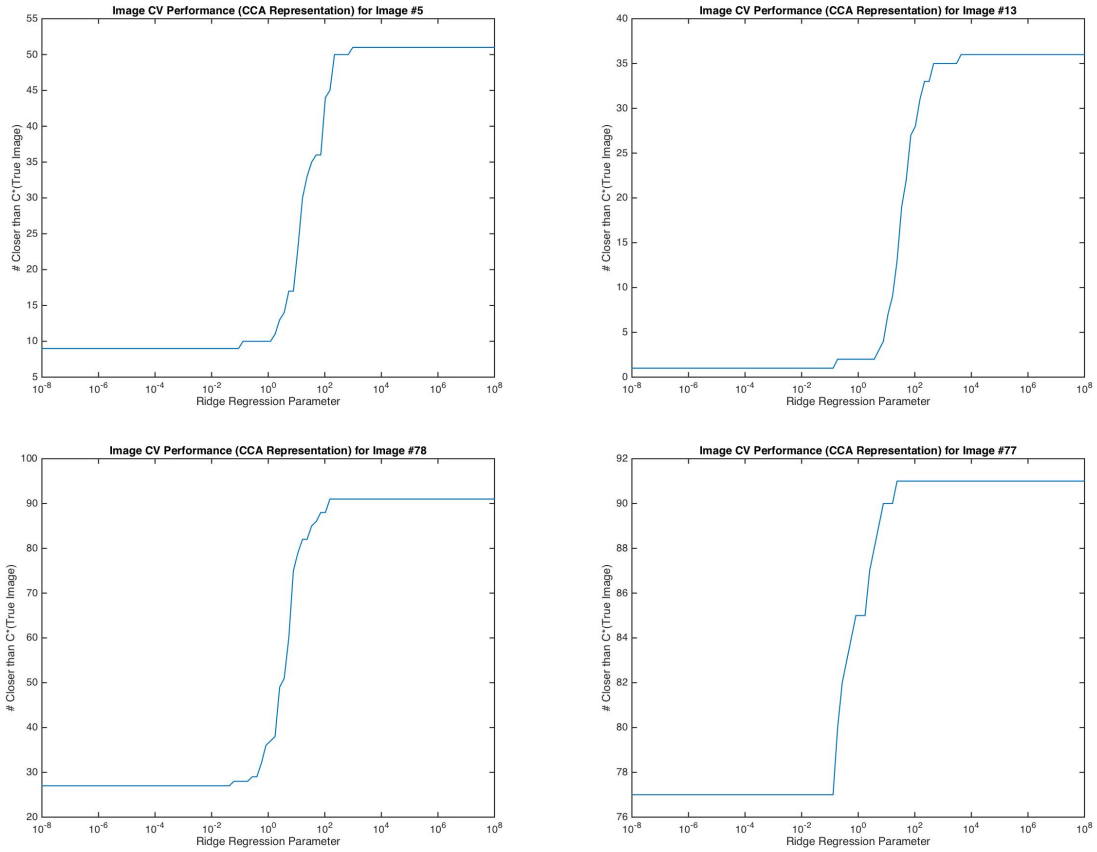


Figure 5. sCCA Featurization: Image Generalization

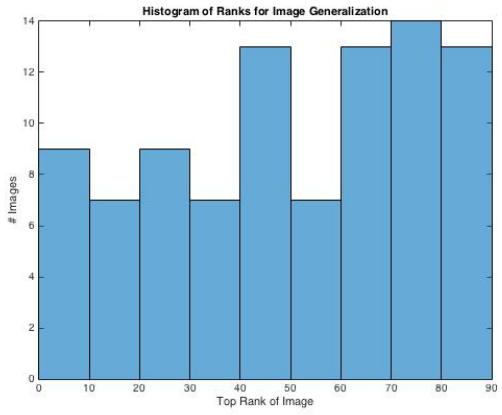


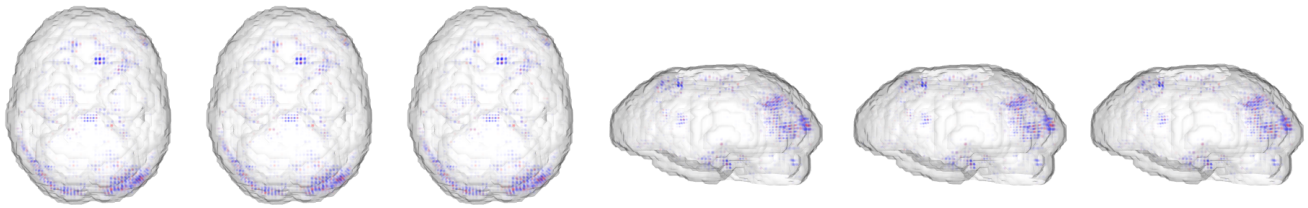
Figure 6. Histogram of Rank Values for Image Generalization

3.2.2. EEG-fMRI Classification results and discussion

We evaluate the quality of our sparse representations by assessing their performance in classification of target and standard stimuli. SVM is our method of choice for performing these binary classifications, as it is widely used in the literature for classifying EEG time series (Zhong et al.; Lin et al., 2008)

We performed SVM binary classification of the examples and report the 10-fold cross validation accuracy (out of 1) and F1 scores in Table 1 and Table 2, which contain the results for an Audio stimuli run and a Visual stimuli run respectively.

Figure 7. sCCA Featurization: Top Rank Images. (Left to right) 5: human bodypart, 13: human face, 33: nonhuman bodypart, 47: nonhuman face, 55, 62: natural inanimate, 74, 80, 89, 90: artificial inanimate



(a) TR1 (b) TR2 (c) TR3 (d) TR1 (e) TR2 (f) TR3

Figure 8. fMRI activations corresponding to the highest correlation canonical vector (correlation $= 0.982$) [Audio stimuli]

	Accuracy	F1
EEG in original space (34000-dim)	0.708	0.343
CCA projection of EEG		
EEG in sCCA space (20-dim)	0.683	0.387
EEG in sCCA space (40-dim)	0.758	0.533
EEG in CCA space, no sparsity constraint (40-dim)	0.625	0.162
EEG concatenated with fMRI		
EEG + fMRI in sCCA space (40-dim)	0.583	0.358
EEG + fMRI in sCCA space (80-dim)	0.625	0.388
Frequency space transformations of EEG		
EEG Fast Fourier Transform (64-dim)	0.691	0.235
EEG Wavelets smooth approximation (2346-dim)	0.675	0.456
EEG Wavelets hierarchical approximation (5338-dim)	0.733	0.446
EEG FFT in sCCA space (40-dim)	0.650	0.268
EEG Wavelets smooth approx. in sCCA space (40-dim)	0.567	0.416
EEG Wavelets hier. approx. in sCCA space (40-dim)	0.692	0.345
Other benchmarks		
EEG in PCA space (40-dim)	0.641	0.460
EEG in random projection space (40-dim)	0.650	0.289

Table 1. Classification accuracies for SVM on various projections of the data [Audio Stimuli]

1. Original space vs. CCA space:

   Best Accuracy and F1 was for the projection of EEG onto 40 CCA vectors.

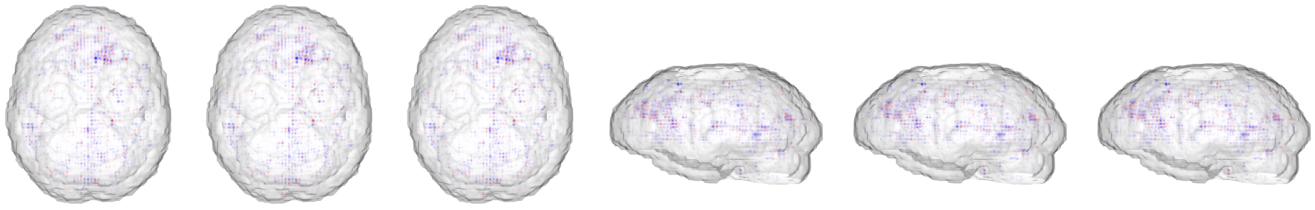
2. EEG in CCA space vs. EEG + fMRI in CCA space:

   Including the fMRI projections did not improve the results. This suggests that just using the projection of the EEG into the CCA is sufficient to encode the most relevant information about target vs. standard stimuli from the fMRI, so the projection of the fMRI data into the CCA space does not contain additional information that is useful for classification.

3. Dimensionality reduction by FFT, PCA, non-sparse CCA:

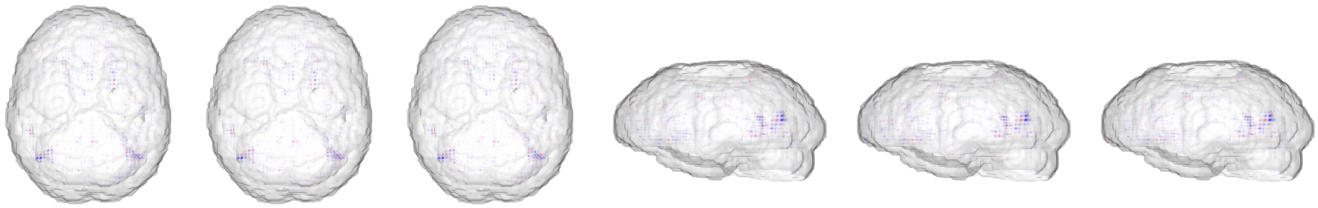
   All performed worse than dimensionality reduction by CCA, suggesting that sparse CCA may indeed be more effective that these more popular methods of dimensionality reduction.

Results from the Visual Stimuli trial are similar to that from the Audio Stimuli trial though slightly more ambiguous. Here, it is not as clear that sCCA is the best performer, though it still had the highest accuracy and the 3rd highest F1 score. The FFT and PCA representations did better here.



(a) TR1 (b) TR2 (c) TR3 (d) TR1 (e) TR2 (f) TR3

Figure 9. fMRI activations corresponding to the 2^{nd} highest correlation canonical vector (correlation= 0.980) [Audio stimuli]



(a) TR1 (b) TR2 (c) TR3 (d) TR1 (e) TR2 (f) TR3

Figure 10. fMRI activations corresponding to the highest correlation canonical vector (correlation= 0.966) [Visual stimuli]

	Accuracy	F1
EEG in original space (34000-dim)	0.725	0.369
CCA projection of EEG		
EEG in sCCA space (20-dim)	0.801	0.166
EEG in sCCA space (40-dim)	0.742	0.432
EEG in CCA space, no sparsity constraint (40-dim)	0.658	0.337
EEG concatenated with fMRI		
EEG + fMRI in sCCA space (40-dim)	0.735	0.161
EEG + fMRI in sCCA space (80-dim)	0.608	0.210
Other benchmarks		
EEG in PCA space (40-dim)	0.700	0.476
EEG in random projection space (40-dim)	0.600	0.223

Table 2. Classification accuracies for SVM on various projections of the data [Visual Stimuli]

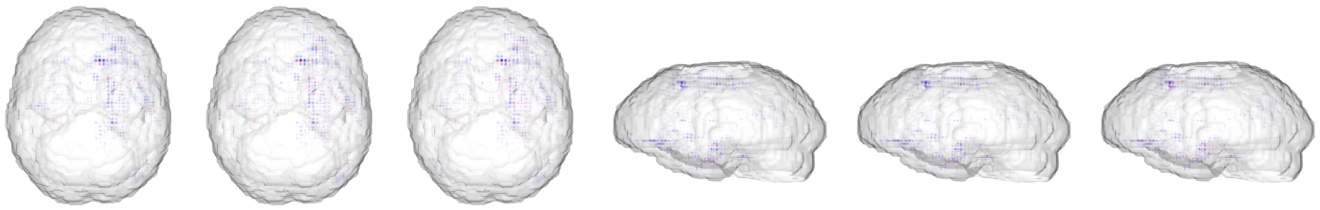
3.2.3. Analyzing multiple trials and subjects for EEG-fMRI

In the above experiments, we performed CCA on the data for a single trial at a time, because we found there to be significant differences in the distribution of the EEG time series between trials. This could be due to measurement artifacts or other trial-specific noise factors. In addition, for the Oddball dataset, each trial has its own unique sequence of stimuli, making it harder to justify combining the data from different trials.

We have used sparse CCA to combine two sets of measurements, namely EEG and fMRI, on the same set of underlying phenomena. While it seemed sound to perform sparse multiple CCA on bi-modal data from different subjects, as long as the subjects are observing the same underlying phenomena, this does not apply well to the Oddball dataset, where every trial from every subject has its own sequence of stimuli as well as noise characteristics. Our attempt to directly apply sparse multiple CCA to the EEG-fMRI data resulted in canonical components with very low correlations. More work is needed to understand how to apply sparse CCA on experimental data where the experimental stimuli differs from trial to trial.

4. Conclusions

The goal of this paper was to establish the possibility and utility of sparse, low-dimensional, multimodal representations and generative models of time-series brain data which retain information, as verified by the predictive power of the models over various tests. We were able to achieve sparsification of the neuroimaging data, a significant improvement over many existing methods for multimodal data analysis which do not take sufficient advantage of sparse methods. Therefore our work is part of a larger effort to better utilize principled machine learning methods in neuroscience. We have also tested our procedure on two real datasets with classification tasks that are tailored to each dataset, rather than simulated datasets. In doing so, we have tackled the difficulties in neuroimaging data



(a) TR1 (b) TR2 (c) TR3 (d) TR1 (e) TR2 (f) TR3

Figure 11. fMRI activations corresponding to the 2^{nd} highest correlation canonical vector (correlation= 0.950) [Visual stimuli]

analysis that arise due to noise, as well as inter-subject and inter-trial differences.

The next task at hand is to verify the soundness of these approaches on other sets of MEG or EEG data gathered under similar conditions.

For the Object task, we would like to explore further featurizations of the images: Using different convolutional network architectures and different layers could prove interesting. Another interesting line of inquiry might be to investigate mapping the convolutional network features themselves into the same shared space as the MEG and fMRI T-maps. It might then be possible to see directly which low-dimensional features of the MEG representation are most correlated with different layers of the convolutional network. We could also examine other regression models, particularly nonlinear ones. For instance, we could use Gaussian Process regression instead of ridge regression, or try to fit a deep neural network.

Further work remains to be done at the algorithmic level as well. Sparse CCA was relatively slow to run. Coming up with a parallel implementation or a faster implementation of sCCA is desirable in order to speed up experiments so that more data can be utilized more quickly.

We would also like to find ways to better interpret the fMRI activations from sCCA, based on previously established functional regions of the brain in the neuroscience literature.

5. Acknowledgements

We would like to thank Radoslaw Cichy for kindly organizing and sharing the Object dataset from his lab with us.

We would like to thank Barbara Engelhardt for her invaluable guidance and support throughout the writing of this paper. We would also like to thank Greg Darnell and the members of the class of COS 513 (Princeton, Fall 2015) for their helpful suggestions and comments.

References

Cichy, Radoslaw Martin, Pantazis, Dimitrios, and Oliva, Aude. Resolving human object recognition in space and time. *Nature Neuroscience*, 17(3):1–10, 2014. ISSN 1546-1726. doi: 10.1038/nn.3635. URL http://dx.doi.org/10.1038/nn.3635$\delimiter"026E30F$npapers3://publication/doi/10.1038/nn.3635.

Dahne, Sven, Bieszmann, Felix, Samek, Wojciech, Haufe, Stefan, Goltz, Dominique, Gundlach, Christopher, Villringer, Arno, Fazli, Siamac, and Muller, Klaus-Robert. Multivariate Machine Learning Methods for Fusing Multimodal Functional Neuroimaging Data. *Proceedings of the IEEE*, 103(9):1507–1530, 2015. ISSN 0018-9219. doi: 10.1109/JPROC.2015.2425807. URL http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=7182735.

Deligianni, Fani, Centeno, Maria, Carmichael, David W., and Clayden, Jonathan D. Relating resting-state fMRI and EEG whole-brain connectomes across frequency bands. *Frontiers in Neuroscience*, 8(August):1–16, 2014. ISSN 1662-453X. doi: 10.3389/fnins.2014.00258. URL http://journal.frontiersin.org/article/10.3389/fnins.2014.00258/abstract.

Indyk, Piotr and Motwani, Rajeev. Approximate nearest neighbors: Towards removing the curse of dimensionality. In *Proceedings of the Thirtieth Annual ACM Symposium on Theory of Computing*, STOC '98, pp. 604–613, New York, NY, USA, 1998. ACM. ISBN 0-89791-962-9. doi: 10.1145/276698.276876. URL http://doi.acm.org/10.1145/276698.276876.

Jia, Yangqing, Shelhamer, Evan, Donahue, Jeff, Karayev, Sergey, Long, Jonathan, Girshick, Ross, Guadarrama, Sergio, and Darrell, Trevor. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*, 2014.

Lin, Yuan P., Wang, Chi Hong, Wu, Tien L., Jeng, Shyh Kang, and Chen, Jyh Horng. Support vector machine for EEG signal classification during listening to emotional music. *Proceedings of the 2008 IEEE 10th Workshop on Multimedia Signal Processing, MMSP 2008*, pp. 127–130, 2008. doi: 10.1109/MMSP.2008.4665061.

Mitchell, T., Shinkareva, S. V., Carlson, A., Chang, K., Malave, V. L., Mason, R. A., and Just, M. A. Predicting human brain activity associated with the meanings of nouns. *Science*, 320:1191–1195, 2008. doi: 10.1126/science.1152876.

Naselaris, Thomas, Prenger, Ryan J, Kay, Kendrick N, Oliver, Michael, and Gallant, Jack L. Article Bayesian Reconstruction of Natural Images from Human Brain Activity. *Neuron*, 63(6):902–915, 2009. ISSN 0896-6273. doi: 10.1016/j.neuron.2009.09.006. URL http://dx.doi.org/10.1016/j.neuron.2009.09.006.

Nishimoto, Shinji, Vu, AnT., Naselaris, Thomas, Benjamini, Yuval, Yu, Bin, and Gallant, JackL. Reconstructing Visual Experiences from Brain Activity Evoked by Natural Movies. *Current Biology*, 21(19):1641–1646, 2011. ISSN 09609822. doi: 10.1016/j.cub.2011.08.031. URL http://linkinghub.elsevier.com/retrieve/pii/S0960982211009377.

Tumari, S. Z. Mohd, Sudirman, R., and Ahmad, a. H. Selection of a Suitable Wavelet for Cognitive Memory Using Electroencephalograph Signal, 2013. ISSN 1947-3931. URL http://www.scirp.org/journal/PaperDownload.aspx?DOI=10.4236/eng.2013.55B004.

Walz, J. M., Goldman, R. I., Carapezza, M., Muraskin, J., Brown, T. R., and Sajda, P. Simultaneous EEG-fMRI Reveals Temporal Evolution of Coupling between Supramodal Cortical Attention Networks and the Brainstem. *Journal of Neuroscience*, 33(49):19212–19222, 2013. ISSN 0270-6474. doi: 10.1523/JNEUROSCI.2649-13.2013. URL http://www.jneurosci.org/cgi/doi/10.1523/JNEUROSCI.2649-13.2013.

Witten, D. M., Tibshirani, R., and Hastie, T. A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics*, 10(3):515–534, 2009. ISSN 1465-4644. doi: 10.1093/biostatistics/kxp008. URL http://biostatistics.oxfordjournals.org/cgi/doi/10.1093/biostatistics/kxp008.

Witten, Daniela M and Tibshirani, Robert J. Statistical Applications in Genetics and Molecular Biology Extensions of Sparse Canonical Correlation Analysis with Applications to Genomic Data Extensions of Sparse Canonical Correlation Analysis with Applications to Genomic Data . 8(1), 2009.

Yu, Hongbin, Lu, Hongtao, Ouyang, Tian, Liu, Hongjun, and Lu, Bao-Liang. Vigilance detection based on sparse representation of EEG. pp. 2439–2442, 2010. doi: 10.1109/IEMBS.2010.5626084. URL http://ieeexplore.ieee.org/ielx5/5608545/5625939/05626084.pdf?tp={&}arnumber=5626084{&}isnumber=5625939$\delimiter"026E30F$nhttp://ieeexplore.ieee.org/xpl/articleDetails.jsp?arnumber=5626084.

Zhong, Mingjun, Lotte, Fabien, Girolami, Mark, and L, Anatole. Classifying EEG for Brain Computer Interfaces Using Gaussian Process Gaussian Process for Binary Classification. *Computing*, 33(0):1–8.