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

In this paper we introduce a new lens through which to analyze time-series brain data, emphasizing the importance of **sparse**, **low-dimensional** representations of EEG and MEG data which retain predictive power and generative modeling capabilities. Our EEG representations make use of spatial information encoded by paired fMRI data with sparse CCA. Sparse CCA is applied to both frequency-space and time-space expressions of the time-series data. We verify the representations have predictive power by training SVMs to distinguish between states of attention ("target states") and normal states.

We examine generative time series models from sparse, low-dimensional inputs with MEG data by learning a linear model between convolutional net image features and frequency features for MEG. Finally, we look at Gaussian processes as a method of generating both EEG and MEG time series data.

1. Introduction

1.1. Motivation

Three primary non-invasive forms of data are collected to study the human brain: functional magnetic resonance imaging (fMRI), electroencephalogram (EEG), and magnetoencephalogram (MEG). fMRI is perhaps the most popular approach due to its high spatial resolution, though it only has temporal resolution of about two seconds, perhaps explaining why most of the experiments which use fMRI do not tend to consider modeling its time series properties and instead analyze 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, 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 multiple modalities of external stimuli (for instance, audio or visual)?

Second, we would also like to ask if it is possible to build a generative time series model of temporal data given an external stimulus — for instance, a picture of an object. That is, given a picture of an object, can we generate a time series signal of brain activity, and moreover, can we reverse our model so that we can decode time-series input to get a view of what object is being looked at?

These questions have been tackled at various levels in previous work. The novelties of our approach are as follows:

- Typically, approaches to data fusion do not take advantage of sparsity. Our approach takes advantage of sparsity and data fusion and validates it with a complete predictive model as opposed to merely examining p-values. Examining sparsity is interesting because ideally, we would only use coordinates which indicate meaningful information. Sparsity can encourage this kind of behavior by only selecting a few components to be non-zero.

- Other approaches tend to use vanilla approaches to dimension reduction.

- No one has investigated Task 2 with non-fMRI 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. 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

Preliminary work - Under review.

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 only produce a time series given a time series of images (i.e. a video). We produce a time series given a *single* image, the assumption being that if a person looks at an image continuously, a fixed thought pattern will repeat continuously in the brain.

1.2. Datasets

We investigate two datasets in this work, namely the EEG-fMRI "Oddball" dataset (Walz et al., 2013) and the MEG-fMRI "Object" dataset (Cichy et al., 2014). Both are paired with external stimuli. In the case of the Oddball dataset, there are two stimuli: audio and visual, both with standard signals and target signals.

1.2.1. 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.

1.2.2. 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.

2. Previous Work

2.1. Fusing modes of brain data

In general, past approaches to exploiting multimodal neuroimaging data can be divided into three main categories:

- *fMRI-informed EEG*: aims to localize the source of the EEG data by using fMRI to construct brain model EEG-informed fMRI: we extract a specific EEG feature, assuming its fluctuations over time covaries.

- *Neurogenerative modeling*: similar to EEG-source modeling; inverse modeling based on the simulation of biophysical processes (known from neuroscience) used to generate EEG and fMRI signals.

- *Data fusion*: Using supervised or unsupervised machine learning algorithms to combine multimodal datasets. It is this approach we will focus on.

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.

2.2. 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.

2.3. Generative Models for Time Series EEG/MEG/fMRI

Most of the generative modeling work in neuroscience is focused on spatial localization, not on learning generative models for time series. The dearth of work in this area is likely due to the low temporal resolution of fMRI data: More researchers find the problem of using fMRI to spatially localize EEG/MEG easier than focusing on modeling the way the data changes over time.

2.3.1. Gaussian Processes

Gaussian processes (Williams & Rasmussen, 1996) lend themselves well for use in time series modelling, and have been used in the past as a reperesentation of EEG/MEG data. (Fox & Dunson, 2012) look at the use of multiresolution Gaussian processes to both capture long-range dependencies in noisy MEG recordings as well as allow for abrupt changes, for example at stimulus onset. In (Faul et al., 2007), Gaussian processes are trained on EEG time series data and used to classify recordings according to whether a neonatal seizure event is seen.

2.4. 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.

2.5. 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 apply 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.

3. Methods and Results

3.1. EEG-fMRI fusion

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.

3.1.1. 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 removing the stimuli where the subject response was incorrect, we select segments of the EEG time series and the fMRI data that correspond to each stimuli and call each of these an example.

More specifically, each example is -100 to 900 ms of EEG time locked to one stimulus, and 3 consecutive TRs of fMRI, the first TR of which covers 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.

3.1.2. Frequency Transform

We performed Discrete Fourier Transform (using FFT) on each 1000 ms section of EEG, time-locked to the onset of each stimuli. Then we extracted average spectral power of EEG components divided in the frequency band, including delta (1-3 Hz), theta (4-7 Hz), alpha (8-13 Hz), beta (14-30 Hz) and gamma (31-50 Hz), for each of the 34 channels.

3.1.3. 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 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 a penalized version of traditional CCA (Witten et al., 2009), which we will refer to as Sparse CCA (sCCA) in this paper. 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 2 and 3, after minimal smoothing with a Gaussian kernel (radial with σ=0.65).

The green 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 2, 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 4, 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 3 and 5).

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.

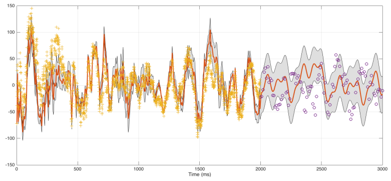

Figure 1. GP Regression for EEG time series.

3.1.4. 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.

	Accuracy	F1
EEG in original space (34000-dim)	0.708	0.343
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 + fMRI in sCCA space (40-dim)	0.583	0.358
EEG + fMRI in sCCA space (80-dim)	0.625	0.388
FFT EEG (70-dim)	0.675	0.295
EEG in PCA space (40-dim)	0.641	0.460
0-mean Gaussian random vector (20-dim)	0.675	0.160

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

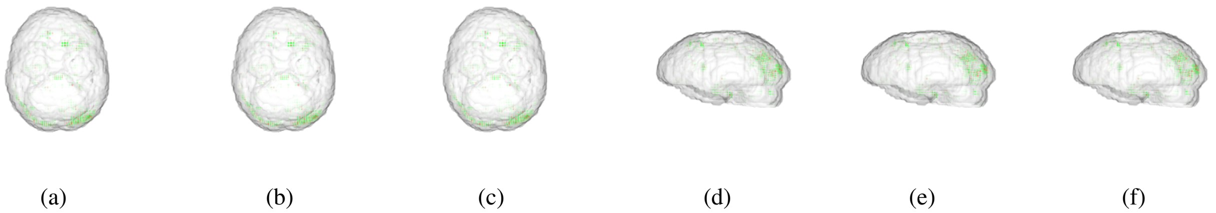

Figure 2. fMRI activations corresponding to the highest correlation canonical vector (correlation= 0.982) [Audio stimuli]

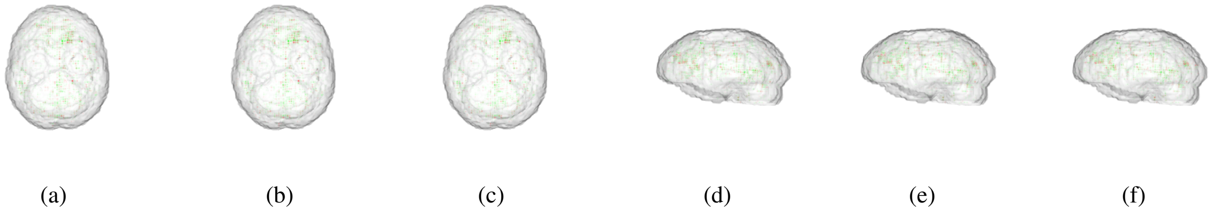

Figure 3. fMRI activations corresponding to the 2^{nd} highest correlation canonical vector (correlation= 0.980) [Audio stimuli]

more effective that these more popular methods of dimensionality reduction.

	Accuracy	F1
EEG in original space (34000-dim)	0.725	0.369
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 + fMRI in sCCA space (40-dim)	0.735	0.161
EEG + fMRI in sCCA space (80-dim)	0.608	0.210
FFT EEG (70-dim)	0.741	0.443
EEG in PCA space (40-dim)	0.700	0.476
0-mean Gaussian random vector (20-dim)	0.658	0.253

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

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.

3.1.5. Gaussian Process applied to EEG

We briefly investigate the extent to which a Gaussian Process (GP) is able to model the EEG time series. Figure 1 plots the GP trained on the first 2000 points of channel 1 of 37 in the EEG recordings for a given subject in a audio trial. The GPML[1] Matlab toolkit was used to train the GPs. The covariance function used here is the sum of a Matèrn kernel, which allows to model sharp changes in the time series, and a periodic kernel for long term variations. A mean squared reconstruction error of $2.3538e - 06$ is achieved for the 2000 training points.

The GP is also used to forecast the next 1000 points. The periodic kernel allows the prediction makes some attempt at modeling the fluctuations in the EEG, though change-points in the predicted trajectory are not as sharp. With further tuning, this may provide a reasonable approach to generative modelling of EEG signals.

3.2. MEG Time-series Generation and Decoding

3.2.1. Preprocessing

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.

3.2.2. Featurizing MEG data

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. We use a Fast Fourier Transform (FFT) to get coefficients for each frequency class. We evaluate coefficients

[1] http://www.gaussianprocess.org/gpml/code/matlab/doc/

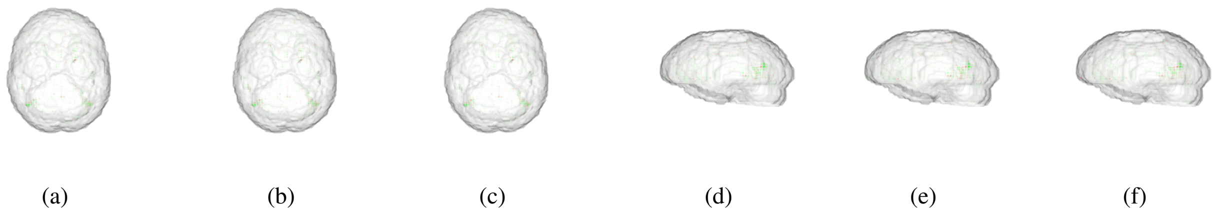
(a) (b) (c) (d) (e) (f)

Figure 4. fMRI activations corresponding to the highest correlation canonical vector (correlation= 0.966) [Visual stimuli]

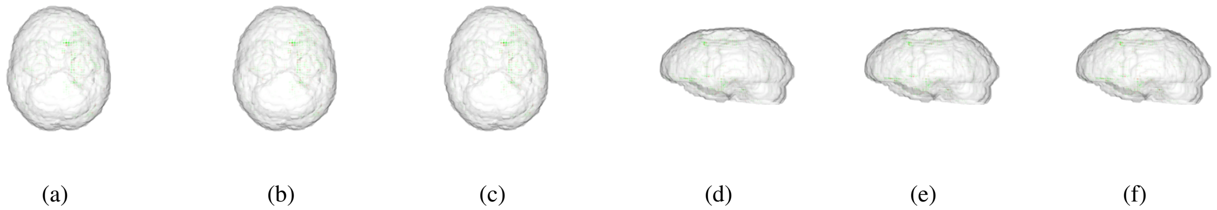
(a) (b) (c) (d) (e) (f)

Figure 5. fMRI activations corresponding to the 2^{nd} highest correlation canonical vector (correlation= 0.950) [Visual stimuli]

at just 64 frequencies, spaced between 0 and the Nyquist frequency (500Hz), to obtain a low-dimensional representation of the MEG data.

As an alternative approach to featurization of the time series, we consider modelling the MEG time series using a Gaussian process, and using the covariance matrix trained as features of the MEG data. The solid red line in Figure 6 plots the GP for a 1301-point time series from a channel of MEG recording from a single trial. The covariance function here is based on a pure Matèrn kernel, and the mean squared reconstruction error is $4.0656e - 09$. However, as we do not model periodicity here, when we attempt to predict future points, we see the signal immediately decays to 0.

Though there is some smoothing, we find that key features of the MEG signal are modelled well by GPs trained on much fewer than 1301 points - the dashed line in Figure 6 uses just 128 evenly spaced points. This still generates a 128×128 dimensional covariance matrix, so does not satisfy our requirements for a sparse, low-dimensional representation of the MEG data from which to map to the image features. It may however be worth exploring the possibility of separating images into different classes based on information in the covariance matrix.

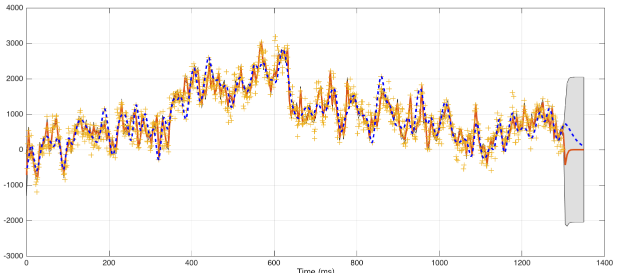

Figure 6. GP Regression for MEG time series

3.2.3. Featurizing Images

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.

3.2.4. Ridge Regression

There are several ways we can try to fit the relationship between the image featurization and the MEG data featurization. The first approach we take is ridge regression. Letting Y be the featurized MEG data of dimension $m = 64 * 306$ 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 (frequency or Gaussian Process covariance matrix), we can use the featurization to re-create the time-series MEG data itself, given the original image z. 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 code in parallel on a 4-core system. We also had to choose a value for the λ-parameter involved in ridge regression: Since running a single experiment was very expensive time-wise, we experimented with a few λs. We settled on using $\lambda = 0.00005$ after getting very bad results with other options.

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. Note that random chance would give i^* a rank of $n/2$. We also calculate the average cosine distance from y^* for all Cx_i as well as the maximum cosine distance, and compare these values to Cx_{i^*}.

3.2.5. Regression results and discussion

We learned C_1, C_2, C_3 for three different subjects separately and found that the average rank was the maximum possible - in other words, we attained perfect accuracy in predicting which object was connected with a given MEG frequency feature vector.

However, these results are probably 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.

To verify that this result was due to overfitting, we ran two basic tests: One to test generalization over subjects and another to test generalization over images. To test subject generalization, we learned matrices C_{ij} where the MEG frequency responses for subject i and subject j were averaged. Then, we predicted the MEG response for subject k and evaluated performance. To test image generalization, we randomly sampled 5 distinct object classes in $[n]$. For each left-out class l, we learned a matrix C_{-l} where no training samples from class 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.

In both cases, generalization was not great. The correct Cx_{i^*} was very close to the average $\sum_{i \in [n]} Cx_i$, and i^* had roughly mean average rank, indicating that in cases where added images are seen, the model is unable to generalize. This result most likely has to do with the low-dimensional representation of the images, and perhaps also the choice of convolutional network features.

4. Future Work

4.1. Next Steps

For the Oddball task, we would like to try including spatial regularization in our sparse CCA implementation to encourage interpretability of the results. We can include spatial regularization by using sparse kernel CCA and using the Graph Laplacian matrix to penalize voxels which are not close to each other. Another next step for the Oddball task is to explore using the highest correlated EEG timepoints to fit a Gaussian Process model. We would effectively be generating a time series model of EEG using a low-dimensional, sparse representation.

The main next steps in the Object task is fixing the problems with ridge regression and overfitting. One approach is to change the featurizations of the MEG (for instance, using Gaussian Process features, or sparse CCA with fMRI-MEG paired data) and images, and another approach is to change the method of fit: i.e., use Gaussian Process Regression instead of ridge regression. After we fix the overfitting problem, we would like to generate MEG signals from the MEG featurization predicted from the image featurization, living up to the goal of generating time-series brain data from sparse, low-dimensional, multimodal inputs.

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.

4.2. Future Directions

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. The first task at hand is to verify the soundness of these approaches on other sets of EEG and MEG data gathered under similar conditions. Another interesting topic to investigate is sparse representations of connectivity matrices as they evolve over time in the same vein as in (Deligianni et al., 2014). The goal would be to give a generative model for a connectivity matrix as it evolves through time, using EEG and fMRI or MEG and fMRI together to give a multimodal representation of connectivity through time.

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`.

Faul, Stephen, Gregorčič, Gregor, Boylan, Geraldine, Marnane, William, Lightbody, Gordon, and Connolly, Sean. Gaussian process modeling of EEG for the detection of neonatal seizures. *IEEE Transactions on Biomedical Engineering*, 54(12):2151–2162, 2007. ISSN 00189294. doi: 10.1109/TBME.2007.895745.

Fox, Eb and Dunson, Db. Multiresolution Gaussian Processes. *Advances in Neural Information Processing Systems (NIPS)*, 25, 2012. ISSN 10495258. URL `https://papers.nips.cc/paper/4682-multiresolution-gaussian-processes.pdf`.

Jia, Yangqing, Shelhamer, Evan, Donahue, Jeff, Karayev, Sergey, Long, Jonathan, Girshick, Ross, Guadarrama, Sergio, Darrell, Trevor, and Eecs, U C Berkeley. Caffe: Convolutional Architecture for Fast Feature Embedding. 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. M., Shinkareva, S. V., Carlson, A., Chang, K.-M., Malave, V. L., Mason, R. A., and Just, M. A. Predicting Human Brain Activity Associated with the Meanings of Nouns. *Science*, 320(5880):1191–1195, 2008. ISSN 0036-8075. doi: 10.1126/science.1152876. URL `http://www.sciencemag.org/cgi/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.

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.

Williams, Christopher K.I. and Rasmussen, Carl Edward. Gaussian processes for regression. *Advances in Neural Information Processing Systems (NIPS)*, 8(August), 1996. URL http://eprints.aston.ac.uk/651/.

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.

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.