Characterizing Similarity of Visual Stimulus from Associated Neuronal Response

Vikram Ravindra and **Ananth Grama**

Department of Computer Science, Purdue University, West Lafayette, IN
ravindrv@purdue.edu

Abstract

The problem of characterizing brain functions such as memory, perception, and processing of stimuli has received significant attention in neuroscience literature. These experiments rely on carefully calibrated, albeit complex inputs, to record brain response to signals. A major problem in analyzing brain response to common stimuli such as audio-visual input from videos (e.g., movies) or story narration through audio books, is that observed neuronal responses are due to combinations of "pure" factors, many of which may be latent. In this paper, we present a novel methodological framework for deconvolving the brain's response to mixed stimuli into its constituent responses to underlying pure factors. This framework, based on archetypal analysis, is applied to the analysis of imaging data from an adult cohort watching the BBC show, Sherlock. By focusing on visual stimulus, we show strong correlation between our observed deconvolved response and third party textual video annotations – demonstrating the significant power of our analyses techniques. Building on these results, we show that our techniques can be used to predict neuronal responses in new subjects (how other individuals react to Sherlock), as well as to new visual content (how individuals react to other videos with known annotations). This paper reports on the first study that relates video features with neuronal responses in a rigorous algorithmic and statistical framework based on deconvolution of observed mixed imaging signals using archetypal analysis.

1 Introduction

Understanding cognitive processes that underlie perception of sensory inputs is an active area of research in behavioural neuroscience. At the heart of these investigations is the design of suitable sensory stimuli, and analyses of observed response to these signals. In [Chen *et al.*, 2017], [Hasson *et al.*, 2004] and [Lahnakoski *et al.*, 2014], subjects are exposed to audio-visual inputs, such as episodes from TV shows or movies, also known as *naturalistic* stimulus. In [Simony *et al.*, 2016] and [Wilson *et al.*, 2008], natural audio/ speech is provided as input. A common goal of these studies is to detect patterns in neuronal responses that are persistent across cohorts. These responses are indicative of neurological mechanisms by which the brain processes inputs.

From a methodological standpoint, these studies rely on correlation measures defined on neuronal responses across subjects to quantify *similarity* in responses to stimuli [Wilson *et al.*, 2008; Hasson *et al.*, 2004]. These approaches have demonstrated success in discovering regions of the brain that manifest responses to the stimulus – regions with excited neurons, that are common across individuals. These results have motivated theories about mechanisms, representations, and processes by which the brain perceives sensory inputs. However, natural inputs are typically complex; they are mixtures of "pure" inputs, and often involve latent excitations. For instance, a scene from a thriller movie can simultaneously evoke feelings of curiosity and fear. Deconvolving the two emotions and computing neuronal responses to *pure* excitations enables accurate cataloging of regions of the brain that are responsible for processing various inputs. Conversely, we can use neuronal responses themselves to inform us about the stimulus; i.e., we can predict latent stimuli based on observed response. Finally, we can also use a catalog of learned pure stimulus-response pairs to predict neuronal responses of new (mixed) stimuli for which we have relevant features. These are profoundly important questions in behavioral neuroscience.

Motivated by these high-level challenges, our goal as part of this work, is to develop powerful new techniques for deconvolving observed neuronal responses from imaging modalities into a combination of basic, constituent representations (for instance, pure emotional responses). We then use these pure responses to reason about the stimulus. Our proposed method relies on the concept of archetypal analyses, which provides a framework for computing pure responses, or *archetypes*. These archetypes are modeled as corners of the smallest convex polytope that envelopes a suitably preprocessed connectomic dataset. In this representation, each individual's connectomic response can be expressed as a convex combination of the archetypes. These convex coefficients are used to deconvolve a mixed signal into its constituent parts based on their similarity to identified archetypes.

Archetypes are reliable representations of dynamic states of the brain, which in turn are indicative of brain activity in response to the stimulus. Hence, the archetypes can be used to make accurate inferences about the stimulus itself. Similarities in archetypes are reflected in corresponding similarities in the input. We use the functional MRI (fMRI) images of the *Sherlock* dataset by [Chen *et al.*, 2017], and show strong correlations between our computed archetypes and accompanying video annotations. Furthermore, we show that the archetypes are robust across subjects, which allows us to predict neuronal response of new subjects.

We make the following specific contributions in this paper:

- We formulate our problem of deconvolving mixed responses from brain fMRI images in the framework of archetypal analyses (AA).
- We use AA on a naturalistic viewing dataset from [Chen *et al.*, 2017] to find responses that are characteristic/ unique to scenes and similar across individuals.
- We show that connectomic features in our new framework are strongly correlated with visual/ textual features derived from our audio-visual excitation.
- We demonstrate the use of our AA framework to predict neuronal responses in new subjects (individuals who have not watches Sherlock).
- Finally, we use our AA framework to predict neuronal responses to repeated stimulus in the same subject (how do we expect the brain to respond when the individual watches the episode a second time).

In each of these cases, we use rigorous methematical and statistical models to validate all of our conclusions.

2 Method

We present a brief overview of archetypal analysis and put it in context of other common factor analysis methods. We then model our problem as one of finding relevant archetypes that are descriptive of shared neuronal responses across individuals.

2.1 Archetypal Analysis: Preliminaries

Archetypal analysis (AA) of a set of data points in a high-dimensional feature space, represents each data-point as a convex mixture of “pure” archetypes – feature vectors that correspond to extremal representatives in the input [Cutler and Breiman, 1994]. These extremal representatives can be thought of as entities that have specialized themselves to pure functions manifest in data.

AA can be viewed as an instance of factor analysis; in the same general class as Principal Component Analysis (PCA), Singular Value Decomposition (SVD), and k-means clustering. Given a matrix $\mathbf{A} \in \mathbb{R}^{m \times p}$ and a mixing matrix $M \in \mathbb{R}^{p \times n}$, these methods minimize the following objective function:

$$\min_{\mathbf{A},\mathbf{M}} \quad ||\mathbf{X} - \mathbf{AM}||_F \tag{1}$$

The specific constraints on matrices $\mathbf{A}$ and M distinguish these methods. In PCA, $\mathbf{A}$ is constrained to be orthonormal; in k-means, each row of M is constrained to have all zeros, except a single entry, which is 1 (the cluster the corresponding entity is assigned to). Archetypal analyses constrains columns of $\mathbf{A}$ to be corners of the minimal polytope that envelops the input dataset, and rows of matrix M to be positive and sum to 1. These constraints on $\mathbf{A}$ and $\mathbf{M}$ lend themselves to intuitive interpretations. Matrix $\mathbf{A}$ can be viewed as the matrix of maximally specialized entities in the input dataset and M as a convex combination of these entities. Formally, we can write the objective function of AA as:

$$\begin{aligned} \min_{\mathbf{A},\mathbf{M}} \quad & ||\mathbf{X} - \mathbf{AM}||_F \\ \text{s.t.} \quad & m_{ij} \geq 0, i = \{1, \ldots m\}, j = \{1, \ldots, n\}, \\ & \sum_k m_{\star,k} = 1, k = \{1, \ldots, n\} \end{aligned} \tag{2}$$

Archetypes themselves can be viewed as mixtures of data-points. We can model them as convex combination of data-points, i.e., $\mathbf{A} = \mathbf{XC}, c_{ij} > 0, \sum_k c_{k,\star} = 1, \forall i, j, k$. Our problem then reduces to one of finding matrices $\mathbf{C}$ and $\mathbf{S}$ that minimizes the error:

$$\begin{aligned} \arg\min_{\mathbf{C},\mathbf{M}} \quad & ||\mathbf{X} - \mathbf{XCM}||_F \\ \text{s.t.} \quad & c_{ij} \geq 0, i = \{1, \ldots m\}, j = \{1, \ldots, n\}, \\ & \sum_k c_{k,\star} = 1, k = 1, \ldots, n, \\ & m_{ij} \geq 0, i = \{1, \ldots n\}, j = \{1, \ldots, p\}, \\ & \sum_k m_{\star,k} = 1, k = \{1, \ldots, p\} \end{aligned} \tag{3}$$

In this formulation, we need to estimate both $\mathbf{C}$ and $\mathbf{M}$. An Alternating Least Squares method can be used to iteratively improve the estimates [Cutler and Breiman, 1994]. Under weak constraints on data, AA has been shown to be unique, and can be efficiently computed using Principal Convex Hull Analysis (PCHA) [Mørup and Hansen, 2012], which we use in this paper. We summarize other AA algorithms in Section 4.

2.2 Technical Approach

The observed neuronal response to a complex stimulus, as observed in a functional MRI, results from an overlay of basic responses to individual components of such stimuli. For instance, inputs in naturalistic viewing experiments are typically audio-visual in nature. Hence, it is reasonable to expect both auditory cortex and visual cortex to be active during the course of the experiment. Since the two areas of the brain are physically distinct, it is straightforward to decouple the two responses and reason about each of them separately. The more non-obvious inference involves the decoupling, or *deconvolution* of neuronal activity when the constituent components are not easy to identify. The goal of our work is to find meaningful constituent signals (or factors) for a naturalistic viewing stimulus. To do this, we show that our analysis pipeline yields factors that are strongly correlated with mental states of viewers, as encoded in the frame-by-frame manual video annotation.

We preprocess functional MRI data so that it is de-noised and corrected to account for various factors such as physiological differences, head motion, and magnetic field inhomogeneity of MRIs. The images are then co-registered and normalized to a standard space. We describe the data cleaning process, as well as the dataset in Section 3.1.

Let $|v|$ denote the number of voxels (3D pixels) in each denoised functional image in the dataset. Functional MRIs can be abstracted as groups of time-series signals – one signal per voxel. Let t denote the number of time-steps. Then, the matrix $\mathbf{X} \in \mathbb{R}^{|v| \times t}$ represents the functional MRI. As the user is being exposed to naturalistic data such as videos or audio playbacks, we break the session into epochs. These epochs correspond to scenes of a video, or chapters of audio books. For epoch e, we represent the corresponding fMRI matrix as matrix $X_e \in \mathbb{R}^{|v| \times t_e}$, where t_e is the number of time-frames in the e-th epoch. In this notation, our problem now becomes $\min_{\mathbf{C}_s, \mathbf{M}_s} ||\mathbf{X}_e - \mathbf{X}_e \mathbf{C}_s \mathbf{M}_s||_F$. However, our goal is to find archetypes that are persistent across frames of a scene across all subjects. To achieve this, for a given epoch e, we stack the frames from all (n) subjects into one population matrix $\mathbf{X}_e \in \mathbb{R}^{|v| \times (t_e \times n)}$ and solve $\sum_{e=1}^{E} \min_{\mathbf{X}_e, \mathbf{M}_e} ||\mathbf{X}_e - \mathbf{X}_e \mathbf{C}_e \mathbf{M}_e||_F$. A schematic representation of the method is shown in Figure 1

We find archetypes for each of the $\mathbf{X}_e$ matrices. For convenience, we denote the archetypes by $\mathbf{A}_e$ (i.e, $\mathbf{A}_e = \mathbf{X}_e \mathbf{C}_e$). For each column of $\mathbf{Z}_e$, we use the closest column in $\mathbf{A}_e$ as its proxy.

$$\mathbf{P}_{e_i} = \text{argmin}_{\mathbf{A}_{e_j}} ||\mathbf{X}_{e_i} - \mathbf{A}_{e_j}||_2 \qquad (4)$$

Here, $\mathbf{P}_e \in \mathbb{R}^{|v| \times (t_e \times n)}$. We define "dominant archetypes" for each time-point as the archetype that was closest to most subjects at that time-point. The matrix of these dominant archetypes is the archetypal response to a given stimulus, as shown in Figure 2.

We note that in this formulation, we consider each scene independent of all other scenes. This simplification works well in practice, since sustained neuronal response over the course of a few continuous time frames (but within one scene) are captured. Stated otherwise, images with thousands of time-points are not well approximated using a small number of archetypes because of their inherent diversity. At the same time, increasing the number of archetypes does not necessarily mean that the archetypes are interpretable. Therefore, dividing the session into smaller epochs works well in real datasets.

3 Results

We describe the dataset – the expermental design, image acquisition protocol, pre-processing and choice of Regions of Interest (ROIs). We then present three sets of results. First, we show that the factors corresponding to our archetypes are representative of neuronal responses across subjects. We then show that these archetypes encode information that is strongly correlated with video annotations. Finally, we show that shared archetypal response can predict neuronal response of new subjects.

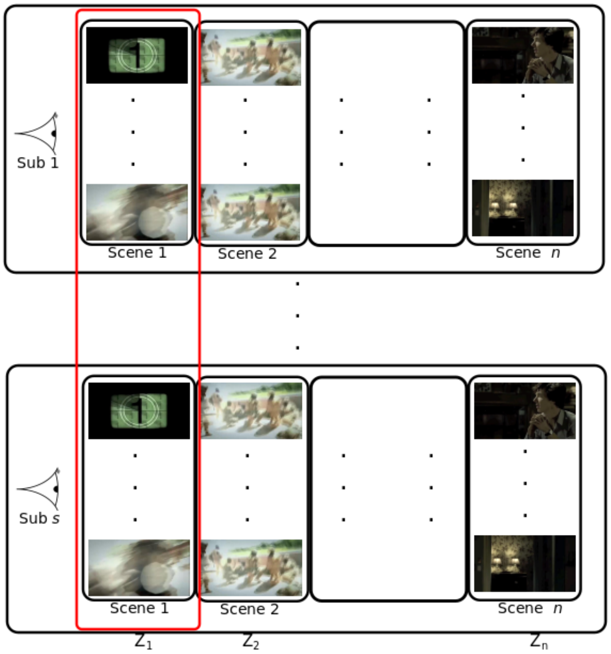


Figure 1: We stack scene-wise functional fMRIs from a population of subjects into $Z_1, ...Z_E$ and to find scene-wise archetypal responses

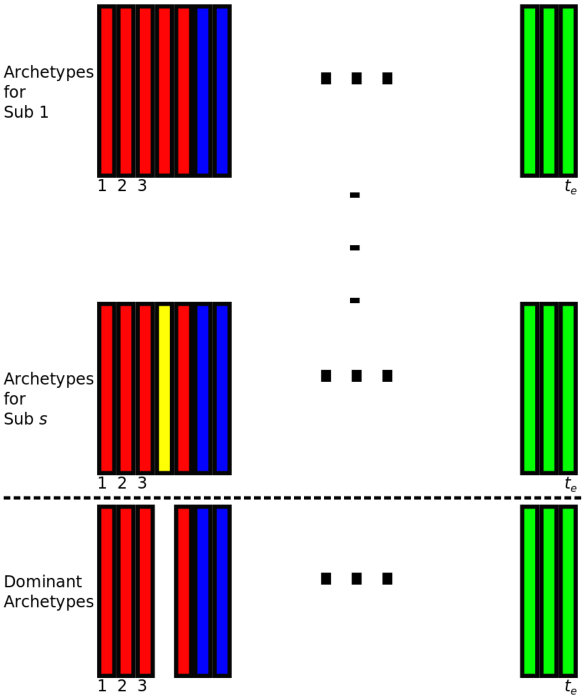


Figure 2: **Visual representation of the computation of dominant archetypes** We compute the proxy matrix P_e for epoch e and then find the closest archetype for each time-frame of each subject. The consensus across all subjects gives us the dominant archetype. This final matrix is the archetypal neuronal response. We note that there may be few gaps (e.g., at $t = 4$ in out data) due to a lack of consensus across subjects.

3.1 Dataset

We use the naturalistic movie viewing dataset described in [Chen *et al.*, 2017]. Briefly, the dataset consists of functional MRIs of 17 participants who viewed 50-minutes of the BBC show Sherlock. The entire show is divided into 50 scenes or *epochs*. The time period of each epoch was decided on the basis of various factors such as change of location, shift in context, and arrival or departure of characters.

After watching the show, the subjects were asked to recall the episode from memory to the best of their ability. The subjects were allowed to recall the scenes out of order, but were asked to describe all the details that they remembered. The 3D fMRI images were sampled every 1.5s, with isometric voxels of side 3 mm (i.e., the voxels were $3mm \times 3mm \times 3mm$). The imaging data is accompanied by movie related meta-data, such as textual transcript of the entire episode, scene location and camera angle. Further, four experts rated each visual segment on *arousal* (excitement or engagement or activity level) and *valence* (positive or negative mood).

We define two regions of interest (ROIs) in the brain. First, an antomical ROI covering the hippocampal region was selected using the Harvard-Oxford Subcortical Atlas [Desikan *et al.*, 2006]. Additionally, we also use a Default Mode Network (DMN) ROI that was constructed by identifying voxels that are strongly correlated with the Posterior Medial Cortex (PMC) across all subjects. The images were pre-processed using fMRI Standard Library (http://fsl.fmrib.ox.ac.uk/fsl). The pipeline included slice-time correction, motion correction, high-pass filter (with 140s cutoff), registration, and alignment to the standard MNI 152 standard. The images were then re-sampled to 3mm isotropic.

3.2 Archetypal Analysis Reveals Factors that are Stable Across Populations

First, we establish that the archetypes are representative of brain states in a cohort of subjects. For each epoch e of the movie, we collect all corresponding fMRI time-frames from all subjects into a matrix $\mathbf{X}_e \in \mathbb{R}^{|v| \times (n \times t_e)}$. Here $|v|$ denotes the number of voxels in each frame, n denotes the number of subjects, and t_e denotes the number of time points in epoch e. Each row of this matrix corresponds to the time series of a voxel and each column-block corresponds to a subject. We identify a small number of archetypes that are representative of the population epoch matrices $\mathbf{X}_e$ using the Principal Convex Hull Analysis algorithm [Mørup and Hansen, 2012].

In Figure 3, we visualize the results for scene 50 of the dataset by projecting to the two dominant singular vectors. In this example, we have three archetypes at the three corners of the triangle and colors represent frame numbers. The cluster of blue points indicates that the blue archetype is a good representation of the first 25 time-points, whereas the red archetype is a good representation of the last 5 data points. In the figure, we have not included frame 17 as there was no consensus. We define a consensus threshold of 70%, which is to say that the archetypal representation of a frame is said to be consistent across a population when at least 70% of the subjects are closest to the archetype. In all, we found that 93.32 % of all frames were stable across the cohort. This result is important, as it lets us use the stable or *dominant* archetypes as proxies for the actual neuronal response.

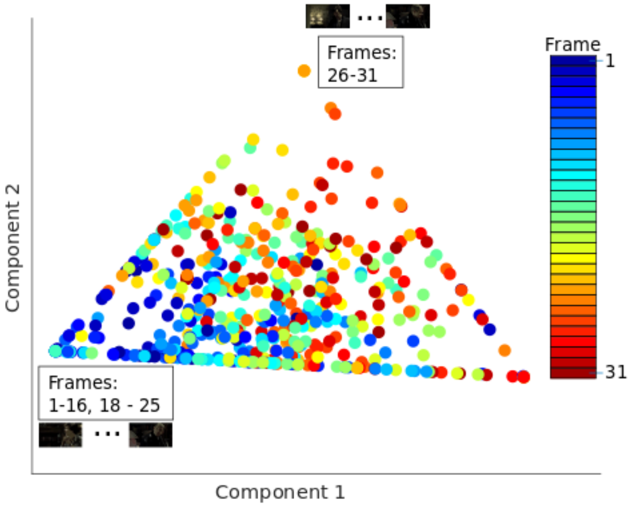
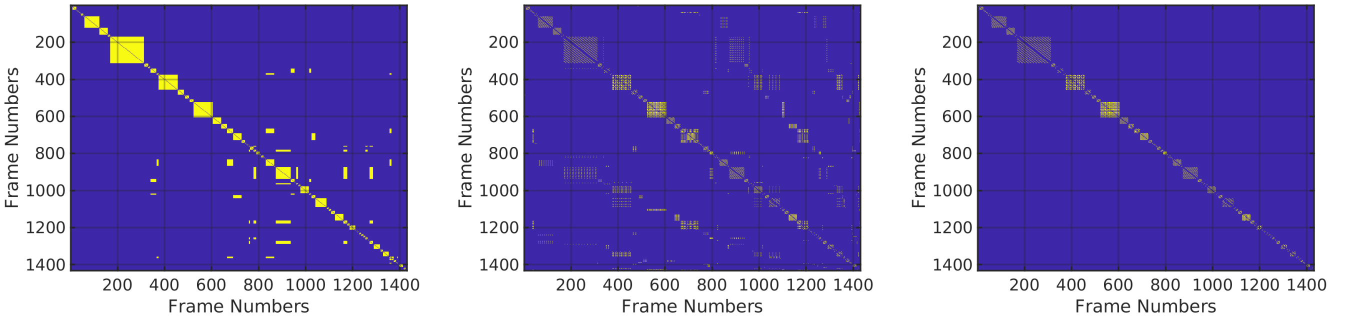


Figure 3: An example of scene-wise archetypes for scene 50. The clusters around the blue archetype show consensus across subjects for frames 1-16 and 18-25. The red archetype has fewer frames associated with it. This scene did not use the third archetype. For the purpose of visualization, the archetypes were projected down to two dimensions (singular vectors). The epochs corresponding to scene 50 were given by [Chen *et al.*, 2017]

3.3 Archetypal Analysis Identifies Correlated Responses from Visual Stimulus

We now present our key result of this paper – *video frames with similar expert scores can be identified from functional MRI data*. As before, we construct the matrix $\mathbf{X}_e$ for each epoch. We find a small number of archetypes using PCHA. Each column vector of $\mathbf{X}_e$ is then assigned to its closest archetype . We define the dominant archetype for time t_{e_i} to be the archetype closest to at least 70% subjects at t_{e_i}. We restrict our analysis to continuous time-frames (at least 3 for visual persistence), that are assigned the same dominant archetype and stack them into a matrix $\mathbf{D}_e$. This matrix represents segments of the video that are identical in terms of their archetype assignment. We repeat the procedure for all scenes to compile the aggregate matrix $\mathbf{D}$ formed by concatenating D_es ($\mathbf{D} = [\mathbf{D}_{e_1}|\mathbf{D}_{e_2}|...|\mathbf{D}_{e_E}]$).

We then find the Pearson Correlations (similarity) between all pairs of vectors in $\mathbf{D}$. The resulting matrix is shown in Figure 4a, after thresholding ($>$ 0.9) to retain only highly correlated frames. We can see a strong block-diagonal structure in the matrix. Indeed this is expected because we retain continuous blocks of time-points that were assigned the same dominant archetype. To validate the hypothesis that highly correlated archetypes are indicative of similar stimuli, we use the expert ranked meta-data for the video. We z-score normalize the frame-wise scores and compute the correlation between them. The resulting similarity matrix after threshold-



(a) Correlation of neuronal response between pairs of frames

(b) Correlation of pairs of frame annotations

(c) Product Matrix of (a) and (b) showing excellent agreement between archetype correlation and corresponding annotation correlation

Figure 4: Heatmaps showing (a) similarity of neuronal responses while viewing scenes and (b) similarity of annotations. (c) Overlap of subfigures (a) and (b). The correlations were thresholded to highlight the similarity of block diagonal structure in the two matrices

ing is shown in Figure 4b. Note the similarity in the block diagonal structure between the two distance matrices. The Hadamard Product (or entry-wise product) of the two matrices is shown in Figure 4c.

To quantify these results, we performed a statistical significance test. We restrict ourselves to the common block diagonal parts of the two matrices (i.e, the non-zeros in Figure 4c). In over 1000 trials, we observe that the similarity of frame annotations along the block-diagonal is higher than off-diagonal entries. In fact, on average, the archetypal similarity and average annotation similarity along the block-diagonal are both > 0.9. In contrast, for the remaining elements, the average archetypal similarity is -0.12 and for average annotation similarity is 0.02.

3.4 Dominant Archetype Predicts Neuronal Response

We show that the archetypes we choose as "dominant archetypes" can generalize to new subjects. To do this, we use archetypes as proxies for neuronal responses, and show that they approximate the real signals well. This exploits the fact that archetypes are representatives of function, therefore they capture the general cognitive state of subjects.

We split the set of subjects into a training set of 12 subjects and a test set of 5 subjects. As before, for each scene e, we create the matrix $\mathbf{X}_e$ and find representative archetypes. We compute the frame-wise dominant archetypes as before (i.e., most commonly occurring archetype for each frame). We then create our predicted response matrix $\mathbf{D}_e \in \mathbb{R}^{|v| \times t_e}$. For each of the test subjects, we compute the l-2 norm between their (actual) neuronal response and our archetypal response. As a control (background distribution), we compute the distance between the neuronal response of all other archetypes. We find that we predict the correct response in $79.31 \pm 4.4\%$ frames, across different runs of test subjects. We fit the two (predicted response and the background) distributions to Standard Gaussians, as shown in Figure 5. The dominant archetypes (blue) fit to a Gaussian with $\mu = -0.79$ and $\sigma = 0.45$, whereas the background archetypes (red) fit to the Gaussian parameterized by $\mu = 0.93$ and $\sigma = 0.66$.

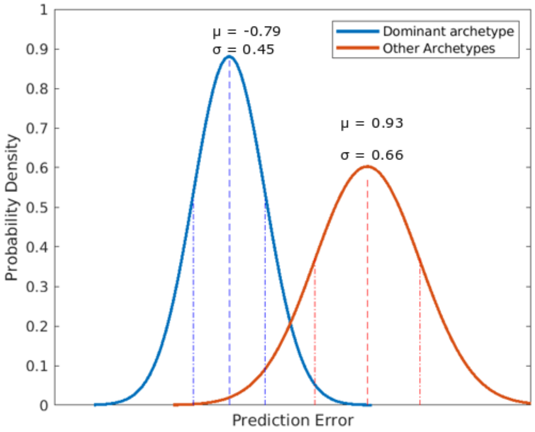


Figure 5: The error (2-norm distance) while using archetypes as proxies to neuronal responses of subjects in test set. The errors are fitted to Gaussian distributions. The blue distribution represents the distance between fMRI response of test subjects and dominant archetypes (across subjects in the training set) for the same set of movie frames (i.e., the same visual stimulus). The red distribution represents the distance between fMRI response of test subjects and all non-dominant archetypes. The separation between the two distributions demonstrates the power of our method.

3.5 Archetypes as Predictors for Repeated Stimuli

Finally, we show that archetypes that describe an epoch of visual stimulus can be used as predicted responses for the same visual stimulus captured in another session. Recall that the functional MRIs collected by [Chen *et al.*, 2017] were collected over the course of two sessions, with a short break between them. The first scene in both the sessions is an animated musical *Let's all go to the lobby*. We use this limited data in a test-retest framework: we perform archetypal analysis on the first (test) scene to find four archetypes that are closest to each time-frame for each subject separately. Then, we use the closest archetype as the predicted response and measure distance to actual response (retest) of the same subject. We find that on average, 70.19 ± 8.21 frames across subjects in the retest scene are closer the corresponding closest archetype of the test scene, than any other archetype. Note that a random assignment would have a one-in-four chance of success. A caveat for this final result is the limited amount of available data (functional MRI responses to same scene), compounded by the fact that the first few frames are usually dropped. However, this experiment suggests that over longer test-retest sessions archetypal responses are reproducible with strong statistical significance.

4 Related Research

We use archetypal analysis (AA) over other formulations of factor analysis. We briefly discuss the other alternatives and present our justification for AA. We also summarize other AA methods, which find application in fMRI studies.

4.1 Rationale for Archetypal Analysis

To reiterate from Section 2.1, we do not use PCA/ SVD due to the unnecessary orthogonality constraint and the absence of a non-negativity constraint. Independent Component Analysis (ICA) is another popular method used in fMRIs studies. However, ICA also does not impose a non-negativity constraint, and is therefore unsuitable for our purpose. Non-negative matrix factorization (NMF) methods constrain $(\mathbf{XC})_{ij} > 0$. There are indeed connections between NMF and AA, which have been explored by [Damle and Sun, 2017] and [Javadi and Montanari, 2019]

Clustering algorithms such as k-means provide alternative approaches. However, there are two arguments against such methods: (i) hard cluster assignments can hide subtle information. For instance, if a subject strongly experiences fear, while also feeling somewhat curious clustering algorithms can ignore the latter; (ii) Cluster centers are chosen on the basis of (some notion of) distance from other points, but are not indicative of cognitive states.

Finally, Generalized Linear Models (GLMs) present another possible solution. Indeed, many fMRI studies successfully use GLMs in their applications. Briefly, the observed fMRI response $\mathbf{Y}$ is expressed in terms of design matrix $\mathbf{X}$ and coefficients β using a linear model., i.e., $\mathbf{Y} = \mathbf{X}\beta + \epsilon$, where ϵ is the estimated error. The design matrix encodes expected haemodynamic response to a given stimulus. These methods can be used to model block experiments and other simpler stimuli. Block experiments are carefully crafted experiments with well defined on-periods and off-periods, which makes the design matrix reasonably simple to estimate. However, this is extremely hard to estimate in our application of naturalistic viewing. While workarounds to this have been implemented in [Hasson *et al.*, 2004], the strength of our approach is that it does not assume any knowledge of expected response. To the best of our knowledge, the only other work that uses archetypal analysis on fMRI data is [Hinrich *et al.*, 2016]. Their method finds a common matrix $\mathbf{C}$ across subjects, which creates a shared convex combination of data points to create convex hulls for each of the subjects. In our application, we find strong correlation with video annotation (which is our goal) when we find archetypes that are common across subjects (i.e., a shared $\mathbf{A}_e = \mathbf{X}_e\mathbf{C}_e$).

4.2 Other Archetypal Analysis Methods

In this study, we have used Principal Convex Hull Analysis (PCHA), as described by [Mørup and Hansen, 2012]. However, other algorithms include Minimum Volume Simplex Analysis (MVSA) [Li and Bioucas-Dias, 2008], where the linear mixing model is solved by a sequence of quadratically constrained subproblems to fit a minimumum volume simplex, minimum-volume enclosing simplex (MVES) [Chan *et al.*, 2009], where the authors do not assume that pure datapoints are realizable, and alternating decoupled volume max-min (ADVMM) and successive decoupled volume max-min (SDVMM) [Chan *et al.*, 2012], which constructs a simplex by minimizing the Winter Criterion. A probabilistic framework for AA was developed by [Seth and Eugster, 2016]. This framework accommodates binary and integer vectors. Indeed, many of these methods also perform well for our application. Our objective is not to compare and contrast these methods, rather, our focus has been to model the problem so as to make it amenable to archetypal analysis, and to demonstrate its power in accurately characterizing neuronal response.

5 Conclusions and Discussion

Naturalistic stimuli are realistic inputs to the brain that model many day-to-day situations, where complex multi-modal inputs are received and processed by the brain. Since the inputs are complex, the corresponding brain responses are also mixed. In this paper, we present a novel method to deconvolve brain-states using archetypal analysis. We show that the computed archetypes are stable across populations, correlate well with manually annotated video labels which encode higher-order cognitive states, and that they can be used to predict neuronal responses of new individuals. An interesting follow-up to our work is to find shared response between the video viewing and recall sessions, which would allow us to develop theories of common mechanisms while creating memories and when recalling from memory.

While our experiments were focused on the dataset due to [Chen *et al.*, 2017], other datasets may yield interesting insights. For instance, [Honey *et al.*, 2012] release a dataset in which subjects listen to the same audio-books in two different languages. AA on this dataset may reveal language-independent archetypal responses to the storyline.

More generally, AA can be used as an alternative to Independent Component Analysis. However, this may require specialized algorithms that handle instrumentation-related complexities that arise due to differences in MRI hardware across sites, and the variance in neuronal responses for large populations. Furthermore, it may be interesting to develop and test multi-stage AA pipelines that find archetypes within subjects, and to combine them into a group analysis, analogous to the two-stage Generalized Linear Models.

Acknowledgements

This work is supported by NSF awards IIS-1546488, OAC-1908691, and the NSF Center for Science of Information STC, CCF-0939370.

References

[Chan *et al.*, 2009] T. Chan, C. Chi, Y. Huang, and W. Ma. A convex analysis-based minimum-volume enclosing simplex algorithm for hyperspectral unmixing. *IEEE Transactions on Signal Processing*, 57(11):4418–4432, Nov 2009.

[Chan *et al.*, 2012] T. Chan, J. Liou, A. Ambikapathi, W. Ma, and C. Chi. Fast algorithms for robust hyperspectral endmember extraction based on worst-case simplex volume maximization. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1237–1240, March 2012.

[Chen *et al.*, 2017] Janice Chen, Yuan Chang Leong, Christopher J Honey, Chung H Yong, Kenneth A Norman, and Uri Hasson. Shared memories reveal shared structure in neural activity across individuals. *Nature Neuroscience*, 20(1):115, 2017.

[Cutler and Breiman, 1994] Adele Cutler and Leo Breiman. Archetypal analysis. *Technometrics*, 36(4):338–347, 1994.

[Damle and Sun, 2017] Anil Damle and Yuekai Sun. A geometric approach to archetypal analysis and nonnegative matrix factorization. *Technometrics*, 59(3):361–370, 2017.

[Desikan *et al.*, 2006] Rahul S Desikan, Florent Ségonne, Bruce Fischl, Brian T Quinn, Bradford C Dickerson, Deborah Blacker, Randy L Buckner, Anders M Dale, R. Paul Maguire, Bradley T Hyman, Marilyn S Albert, and Ronald J Killiany. An automated labeling system for subdividing the human cerebral cortex on mri scans into gyral based regions of interest. 31(3):968–980, 2006.

[Hasson *et al.*, 2004] Uri Hasson, Yuval Nir, Ifat Levy, Galit Fuhrmann, and Rafael Malach. Intersubject synchronization of cortical activity during natural vision.(research article). *Science*, 303(5664):1634, 2004.

[Hinrich *et al.*, 2016] Jesper Love Hinrich, Sophia Elizabeth Bardenfleth, Rasmus Erbou Roge, Nathan William Churchill, Kristoffer Hougaard Madsen, and Morten Morup. Archetypal analysis for modeling multisubject fmri data. 10(7):1160–1171, 2016.

[Honey *et al.*, 2012] Christopher J Honey, Christopher R Thompson, Yulia Lerner, and Uri Hasson. Not lost in translation: neural responses shared across languages. 32(44):15277, 2012.

[Javadi and Montanari, 2019] H. Javadi and A. Montanari. Nonnegative matrix factorization via archetypal analysis. *Journal of the American Statistical Association*, 2019.

[Lahnakoski *et al.*, 2014] Juha M. Lahnakoski, Enrico Glerean, Iiro P. Jääskeläinen, Jukka Hyönä, Riitta Hari, Mikko Sams, and Lauri Nummenmaa. Synchronous brain activity across individuals underlies shared psychological perspectives. *NeuroImage*, 100:316 – 324, 2014.

[Li and Bioucas-Dias, 2008] Jun Li and J.M Bioucas-Dias. Minimum volume simplex analysis: A fast algorithm to unmix hyperspectral data. In *IGARSS 2008 - 2008 IEEE International Geoscience and Remote Sensing Symposium*, volume 3, pages III – 250–III – 253. IEEE, 2008.

[Mørup and Hansen, 2012] Morten Mørup and Lars Kai Hansen. Archetypal analysis for machine learning and data mining. *Neurocomputing*, 80:54, 2012.

[Seth and Eugster, 2016] Sohan Seth and Manuel J. A. Eugster. Probabilistic archetypal analysis. *Machine Learning*, 102(1):85–113, Jan 2016.

[Simony *et al.*, 2016] Erez Simony, Christopher J Honey, Janice Chen, Olga Lositsky, Yaara Yeshurun, Ami Wiesel, and Uri Hasson. Dynamic reconfiguration of the default mode network during narrative comprehension. *Nature Communications*, 7(1), 2016.

[Wilson *et al.*, 2008] Stephen M. Wilson, Istvan Molnar-Szakacs, and Marco Iacoboni. Beyond superior temporal cortex: Intersubject correlations in narrative speech comprehension. *Cerebral Cortex*, 18(1):230–242, January 2008.