UNIVERSITY OF CALIFORNIA SAN DIEGO

Decoding Affect from Intracranial Neural Response to Acoustic Stimuli

A thesis submitted in partial satisfaction of the
requirements for the degree Master of Science

in

Electrical Engineering
(Machine Learning and Data Science)

by

Jingya Huang

Committee in charge:

Professor Vikash Gilja, Chair
Professor Nicholas Antipa
Professor Gal Mishne

2021

The Thesis of Jingya Huang is approved, and it is acceptable in quality and form for publication on microfilm and electronically.

University of California San Diego

2021

EPIGRAPH

Different representation make different computation easier or harder,
by making certain information explicit
and others pushed into background.

David Marr

TABLE OF CONTENTS

LIST OF FIGURES

ACKNOWLEDGEMENTS

I would like to acknowledge Professor Vikash Gilja for his support as the chair of my committee. Through multiple drafts and many long discussions, his guidance has proved to be invaluable.

I would also like to acknowledge the Limbic Team in the TNEL Lab, without whom my research would have no doubt taken fives times as long. It is their support that helped me in an immeasureable way.

I have been even more privileged to have loving and supportive friends and family. Mom and Dad, thank you for believing in me and teaching me to always keep my goals in sight. Heye, your support got me through tough times.

This paper is coauthored with Aashish N.Patel; Jerry Shih; Vikash Gilja. The thesis author was the primary author of this thesis paper.

VITA

2015 - 2019 B. S. in Mathematics and Computer Science, University of California, San Diego

2015 - 2019 B. S. in Cognitive Science, University of California, San Diego

2019 - 2021 M. S in Electrical and Computer Engineering, University of California, San Diego

ABSTRACT OF THE THESIS

Decoding Affect from Intracranial Neural Response to Acoustic Stimuli

by

Jingya Huang

Master of Science in Electrical Engineering
(Machine Learning and Data Science)

University of California San Diego, 2021

Professor Vikash Gilja, Chair

Brain-computer Interface(BCI) provides a direct communication pathway between targeted brain region by recording its evoked neural signals and control systems that interact with neural interfaces. Our work is aiming to find an alternative brain-computer interface (BCI) control scheme to decode affective states, for assessing and interpreting changes in the user state while evaluating goal-oriented control schemes, thus improving the efficiency, usability and accuracy of the BCI. To build such a BCI system, successfully decoding the affective state is an essential step to take. There are several neural studies with non-invasive recording techniques that shed insight on the complex and subtle relationships between affective state and neural

response. However, it is challenging to capture the high resolution spatio-temporal patterns in the neural response with non-invasive recording techniques. The precise spatio-temporal pattern in cortical depth structures is important to characterize complex affective state at different time scales.

Addressing the complexity of neural representation, we bring an acoustic perception experiment into the Epilepsy Monitoring Unit (EMU) and obtain intracranial neural recordings from subjects listening to the natural acoustic stimuli that spans different dimensions of the affective states. In this work, we present several decoder models with different levels of regularization (ordered by descending model complexity: quadratic discriminative analysis (QDA), regularized QDA with class-specific Gaussian process factor analysis (GPFA), and regularized QDA with class-invariant GPFA). Through an iterative process of feature selection and model simplification we identify few models that emphasize the most informative feature for decoding effect.

Chapter 1

Introduction and Background

Brain-computer Interface(BCI) provides a direct communication pathway between targeted brain region by recording its evoked neural signals and control systems that interact with neural interfaces. In contrast to BCI which are used for directed control, an alternative approach allows us to assess and interpret changes in the user state while evaluating goal-oriented control schemes. These approaches enable affective neural responses to provide more reward-analogous complementary signals, like Pacman game studies which use input such as noisy input to influence subject's valence, arousal (e.g. Pacman [6]). Such discovery sheds some insight on the complex but subtle relationship between affective response and BCI system which related to subject's neural state control. To further explore this topic, this thesis work centers around decoding affective states using naturalistic stimuli, which is part of the ongoing neural critic project to provide an alternative brain-computer interface (BCI) control scheme to improve the efficiency, usability and accuracy of the BCI.

There are three sections we want to clarify in this chapter for BCI application, we are interested in extracting robust and informative neural features to decode affective state, thus helping the design of high performance decoder. First, in Section 1.1, we talk about some basic concepts of affective response. Recent literature reviews about existing neural features and related decoding models are discussed here. In Section 1.2, one experiment is selected to demonstrate the relationship between affective state and intracranial-EEG (iEEG) intracranial

neuron recording. Based on the selected experiment, section 1.3 functions as a guide to our development of feature extraction and neural decoding models. Note that the current model design is specific to time varying acoustic stimuli, but variations of the same model may be generalize to reflect affective response in more dynamic and naturalistic environments.

1.1 Affective Responses

Among many aspects of the neural responses due to affective state change, we are particularly interested in 1) extracting robust neural features for BCI systems (introduced above) 2) decoding affective state accurately and continuously. Several studies were able to decode induced affective (dimensional or categorical) from images, sounds, and video stimuli using selected fMRI voxels [5][6][13][14] or band-limited power feature of EEG signals [20][22][25]. However, the limited spatio-temporal resolution of those non-invasive recording techniques (e.g. fMRI, EEG) constrains the performance of the decoder. The fMRI recording provides accessibility to relevant physiological regions, notably limbic structures, but most trained decoders are only applicable to static images or videos over hours [5]. On the other hand, the EEG studies were able to record with richer temporal and spectral information, but can only target near cortical surface where complex signals are aggregated across brain regions [20]. For our work, we are privileged to record in the EMU with the intracranial-EEG (iEEG) which has more direct access to the local field potential in the depth cortical regions with higher spatio-temporal resolution.

The above paragraph summarizes existing work in affective state decoding and our motivation from the practical perspective. Moreover, our work may also provide some theoretical insight to the neural representation of affective states. There is a persistent and controversial debate between the two affective models: the categorical model (the basic emotion model proposed by [9]) and the dimensional model (the circumplex model of affect proposed by Russell [19], etc). One study shows that the categorical and dimensional representations develop along different timescales in different cerebral regions, and urges studies with precise spatio-temporal

dynamics of the cerebral networks to gain a comprehensive understanding [10]. Our modeling and analysis work on high resolution spatio-temporal dynamics of those targeted regions shed light on the neural representation of the dimensional model. Similar modeling framework and analysis techniques may be adopted to establish the neural representation of the categorical model in the future.

1.2 Acoustic Perception Experiment

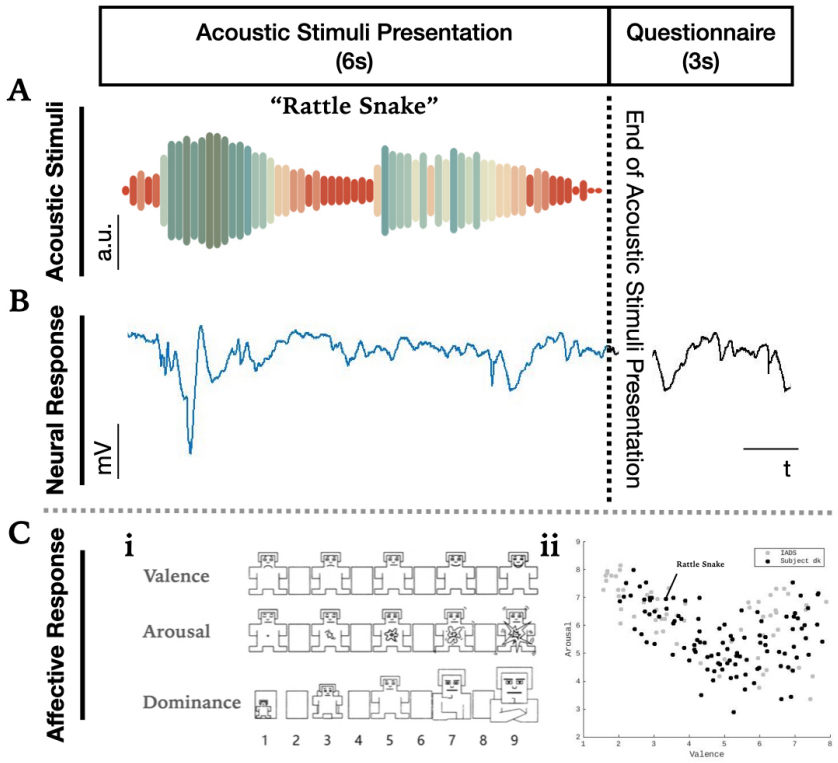

Figure 1.1. Overview of the Acoustic Perception Experiment. Each session contain approximately 20 trials, where each trial begins with the acoustic stimuli presentation for 6 seconds and then followed by the Sham Questionnaire for 3 seconds. (A) The waveform for an exemplar acoustic stimuli named "rattle snake". (B) The neural response recorded from one iEEG channel of a corresponding trial during the 6 seconds stimuli presentation (blue) and during the 3 seconds questionnaire and inter-trial interval (black). (C) The affective response for acoustic stimuli in the IADS (2017) [21]. (i) The standard self-assessment manikin (SAM) questionnaire used to collect the affective response. (ii) The distribution of average valence and arousal response for each acoustic stimuli presented (black) and not-presented (grey).

Several studies using non-invasive neural recording techniques have established the neural correlates with the acoustic stimuli sampled from the IADS dataset [11][3][15]. Most of those studies were only able to identify the importance of certain brain regions (e.g. cingulate cortex) or neural correlates over seconds of recording. To gain an understanding of neural signature under the affective response with higher resolution, we perform the acoustic perception experiment in the EMU, where we are privileged to record intracranially within the limbic system (e.g. insua, amygdala, cingulate cortex) and temporal lobe with higher spatiotemporal resolution.

The primary subject analyzed subsequently participates in 80 trials of acoustic perception. The experiment consists of four sessions (20 trials in each session) and a resting period in between. (Figure 1.1) shows an exemplar trial structure, the trial begins with acoustic stimuli presentation for 6 seconds, followed by a 3 seconds sham questionnaire.

During the first 6 seconds of each trial, an acoustic stimuli sampled from the IADS dataset is played. The IADS dataset provides a rich dataset for the sound segments of the natural scenes, objects, and human voices, each spins 6 seconds. (Figure 1.1.A) shows an example acoustic stimuli, the 'rattle snake', which consist of two consecutive rattle snake hissing. The corresponding neural response recorded from an example iEEG channel is shown in (Figure 1.1.B) as blue.

At the end of the stimuli presentation, the subject is prompted to answer a sham questionnaire. The questionnaire is designed to motivate the subject to attend to the task with simple questions like "whether a human could be identified as making the noise". The corresponding neural response recorded from the same iEEG channel (as above) is shown in (Figure 1.1.B) as black. Such neural responses recorded during these 3 seconds are treated as the baseline response.

On the behavioral side, we leverage the population affective response average (over 100 subjects) from the original IADS study[21]. The IADS experiment shares a similar trial structure as our experiment described above. A notable difference, however, is that at the end of the stimuli presentation, their subjects are asked to respond to the standard SAM questionnaire [4]

(Figure 1.1.C.i) to rate the played stimuli on three principal dimensions: valence, arousal, and dominance. (Figure 1.1.C.ii) shows the distribution of the average valence and arousal response for each acoustic stimuli. The dominance dimension is omitted here because the valence and dominance dimensions are correlated (p-value < 0.05).

We also understand that the affective response maybe highly personalized and influenced by the subject's personal experience. In the ongoing data collection, we are utilizing the SAM questionnaire to collect the subject's response at the end of stimuli presentation to reduce the variance in affective response.

1.3 Conceptual Model for Emotion Modulation

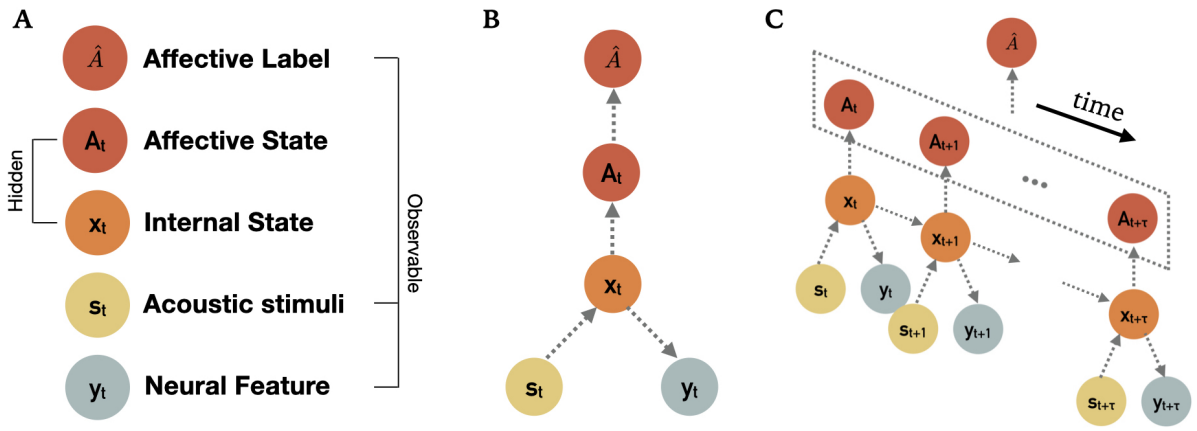

Figure 1.2. Conceptual Model for Affects Modulation. (A) The identified random variable involved in this study. Three observable variable, the affective label $\hat{A}$, the acoustic stimuli s_t, the neural feature y_t, collected during the experiment. Two hidden variable, the affective state (A_t) and the internal state (x_t). (B) A graphical representation of the proposed relationship among the affective state, acoustic stimuli and neural response at a single time point. (C) An expansion of the graphical representation in (A) across time to capture the evolution of the acoustic stimuli, affective state and neural response over time.

In this section, we will define the random variables and provide a conceptual framework for the modeling work in later chapters. Although several previous works help us gain some understanding of the affective response [11][3][15], due to its complex nature, there are limited

works that clearly explain the representation of the affective response and in particular, the relationship among the affective response, acoustic stimuli and neural response. In addition to those three observable variables, we introduce two hidden variables, the instantaneous internal neural response and the affective response. Previous works indicate that there is a lack of a one-to-one mapping between the affective response, acoustic stimuli and neural response [12]. Through the introduction of those hidden variables, we open up the opportunity to explicitly model the summarization and transformation of the information from the acoustic stimuli to the neural response that gives rise to the affective response. Here we define this particular structure (Figure 1.2) for hypothesis testing to reveal the relationship between each of those components.

There are three observable variables involved in the acoustic perception experiment. The independent variable acoustic stimuli s_t, the dependent variable neural feature y_t and affective label $\hat{A}$. Here we use the subscript t to denote the variable that is changing over time. The previous section provides more details of the acoustic stimuli s_t, the neural feature y_t and the affective label $\hat{A}$ within the context of the experiment.

To further specify the relationship among the observable variables, we define two hidden variables, the internal state x_t and the affective state A_t (Figure 1.2.A). At each time point, the acoustic stimuli s_t modulates the internal state x_t, which generates the noisy observed neural feature y_t. For the interest of this study, we follow a narrow definition of the internal state x_t, as the neural response that is informative of the affective state A_t. By specifying the time-varying affective state A_t, we could take into account the integration of the affective response at different timescales to arrive at the final affective label $\hat{A}$ collected at the end of the stimuli presentation. (Figure 1.2.B) illustrates the relationship among those random variables at each time point, which is a simplification of the full dynamic model shown in (Figure 1.2.C).

Over time, the acoustic stimuli $\{s_t\}$ influences the internal states $\{x_t\}$, which generate the neural features $\{y_t\}$ and the affective states $\{A_t\}$. While all these time-varying variables carry their own dynamics, this study focuses on the dynamics of the internal states $\{x_t\}$ which explains the variance in the neural features $\{y_t\}$ that is relevant for the affective states $\{A_t\}$.

In summary, we define five random variables important for this study (Figure 1.2.A) and identify several important relationships for hypothesis testing. In particular, the influence of the acoustic stimuli s_t on the internal state x_t, the mapping between the internal state x_t and the neural feature y_t, and between the internal state x_t and the affective state A_t. We should also note the different assumptions on temporal structure in the conceptual model (Figure 1.2.B and 1.2.C). In the later chapters when (Figure 1.2.B) is shown, we are assuming that each time point is statistically independent, while when (Figure 1.2.C) is shown, we are making assumptions on the dynamics of the internal state x_t.

Chapter 2

Conditioning and Pre-processing

2.1 Affective Label

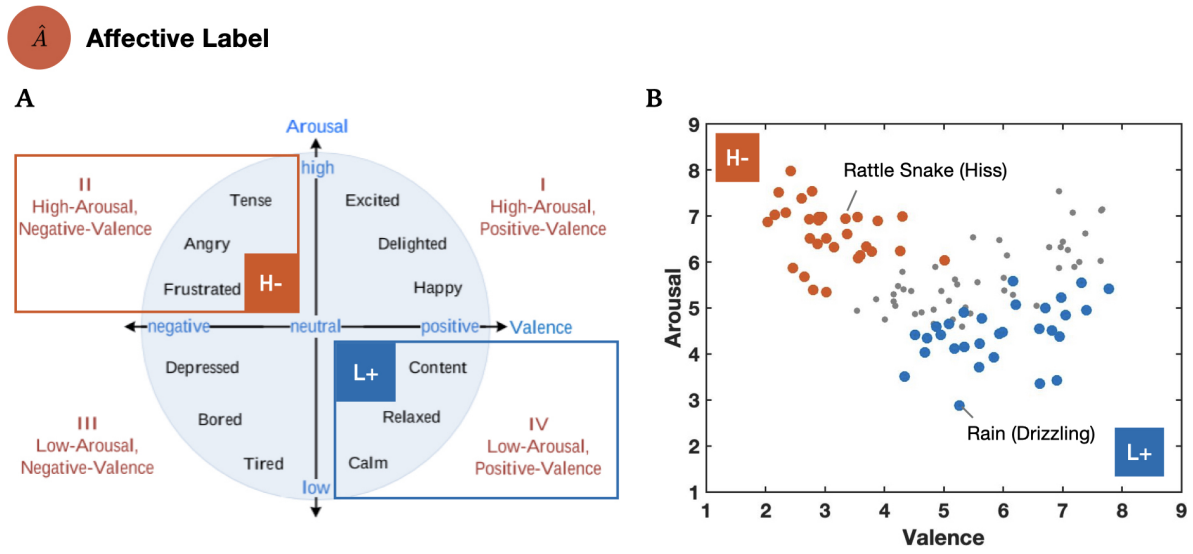

Figure 2.1. Affective Label. (A) Valence and Arousal parameterization of the behavioral affective response. H- labels the high-arousal negative valence states (e.g. tense, angry, frustrated). L+ labels the low-arousal positive valence states (e.g. content, relaxed, calm) (B) Valence and Arousal rating of the acoustic stimuli on the scale from 1 to 9, red dots for stimuli with affective label H-, blue dots for stimuli with affective label L+. Two example stimuli (rattle snake, rain) are annotated with text.

We simplify the circumplex model of affect [17], which is the most popular two-dimensional model of affective experience, introduced in Chapter 1, to focus on the two diagonal quadrants. There are two reasons why this particular simplification is adopted. From the data

perspective, the acoustic stimuli in the IADS dataset do not disperse the full quadrants. Acoustic stimuli that elicit low-arousal & negative-valence affective state (e.g. depressed, bored, tired) or high-arousal & positive-valence affective state (excited, delighted, happy) are not salient enough in the IADS dataset, a similar result also exist in the follow-up expanded IADS dataset [23]. From the machine learning perspective, the imbalanced data counts in the multi-class model will introduce the variance too difficult to be captured. As an initial investigation of affective decoding using iEEG recording, we define the binary affective class focusing on the diagonal quadrants.

In our experiment, we mainly focused on label H- as the negative valence & high arousal state, characterizing emotions such as tense, angry, frustrated; L+ as the positive valence & low arousal state, characterizing emotions such as content, relaxed and calm (Figure 2.1.A). From each affective class, we sampled 50 acoustic stimuli (Figure 2.1B). Take the sound of rattlesnake (hissing) as an example of H- stimuli and rain (drizzling) as an example of L+ stimuli. Rattlesnake hissing commonly drives people to high alert or even to be frustrated, while drizzling sound sedates their temper and minds.

2.2 Neural Features

To extract the neural feature from the iEEG recording channels, we follow a standard pre-processing pipeline (Figure 2.2.A). To de-noise, we use the mean as reference for each probe (16 channels per probe). With the de-noised signal, we first bandpass-filter the time-series in 6 well-established frequency band: delta [2Hz, 3.5Hz], theta [4Hz, 7Hz], alpha [9Hz, 11Hz], beta [15Hz, 30Hz], gamma [30Hz, 50Hz], high gamma [90Hz, 110Hz] (Figure 2.2.B). We then extracted the instantaneous power for the band-limited neural recording using the Hilbert transform. We use the 1/f fall of as the exclusion criteria and result in 91 channels in recording sites: Medial Insula, Anterior Insula, Posterior Cingulate, Anterior Cingulate, Orbital Frontal, and Hippocampal Tail. The electrodes that have significant broadband deviation from 1/f fall off

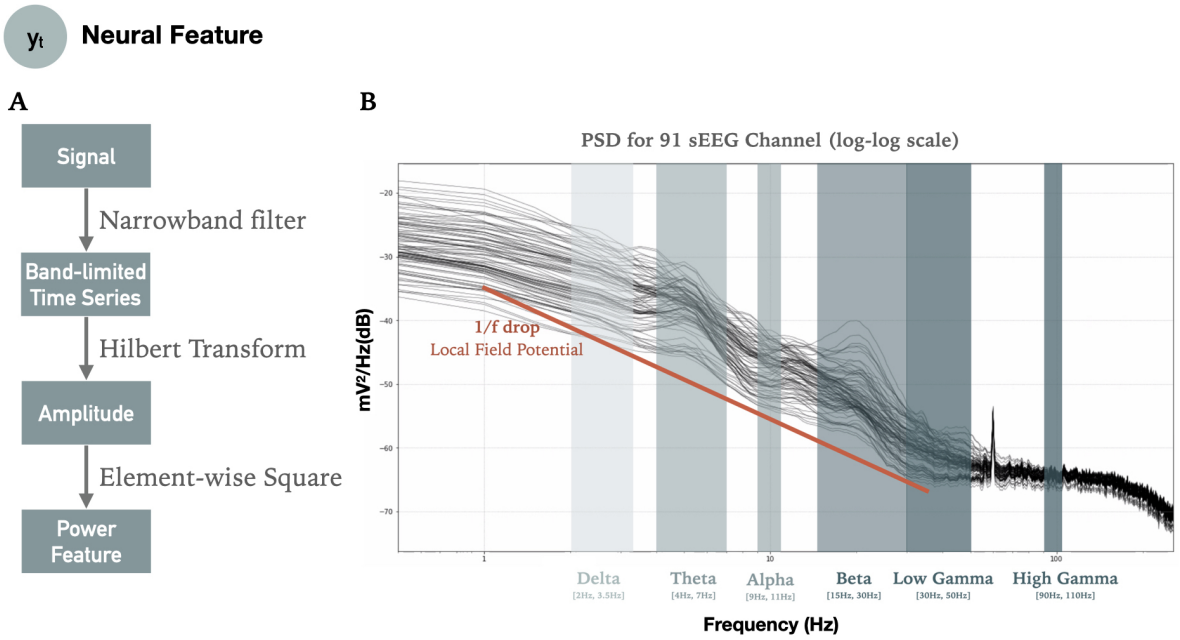

Figure 2.2. Neural Feature. (A) Neural feature extraction pipeline. This study uses the narrow-banded instantaneous power as the neural feature. (B) The PSD for 91 iEEG channels on the log-log scale. Each shaded region denotes an identified frequency band: delta [2Hz, 3.5Hz], theta [4Hz, 7Hz], Alpha [9Hz, 11Hz], Beta [15Hz, 30Hz], Low Gamma [30Hz, 50Hz], High Gamma [90Hz, 110Hz]

are considered noisy or artifact-prone and removed. We are particularly interested in the 1/f fall off (Figure 2.2.B); several neuro-physiology studies [2][1][16][18][8] have shown that the 1/f fall off in the power spectral density represents the local field potential, which is a mixture of the local spiking activity and postsynaptic potential. As such, the extracted band-limited power could be interpreted as the fluctuations in the local field potential at different timescales.

Chapter 3

Decoding Affective State

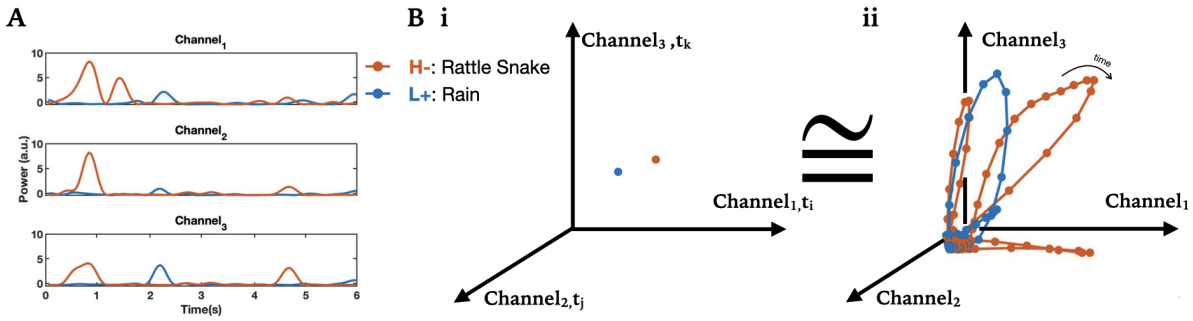

Figure 3.1. Schematic figure of the Population Analysis. (A) The dynamic activity recorded from a channel (alpha-band power feature) during the 6 seconds stimuli presentation. Neural feature for two example trials are shown, red for H-: Rattle snake, blue for L+: Rain. (B) Two equivalent representations of the trajectory of a population vector for two example trials in (A), with recording in 3 iEEG channels and sampled at 122-time points. (i) In a 366D space (3 dimensions are shown), each dimension represents a iEEG channel at a single time point. A dot represents an example trial, rattlesnake (red), rain (blue). (ii) In a 3D space, each dimension represents a iEEG channel. The dots represent the response at each time point in the example trial, rattlesnake (red), rain (blue), and the line indicate the elapse of time

In this chapter, we will discuss a sequence of modeling and hypothesis testing in the improvement of the decoding performance under the population analyses framework. Three attractive characteristics of the population analyses are taken into consideration. [7]

Firstly, compared to the widely used trial average technique, the population analysis allows us to make a trade-off between the multiple trial single channel and the single-trial multiple channels for statistical power. Noting two factors such that 1) the time course of the neural

response across different trials are very different from each other because the acoustic stimuli each have different dynamics; 2) the affective response, as a type of high-cognitive response, may arrive at different time course. Considering both factors, trial-averaging is not proper in our case. Further evaluation of real experimental scenarios also shows that the experiment time is particularly precious with recording in the EMU, which makes it less meaningful to repetitively present the same stimuli. Not even considering that the human subjects may be fatigued after several times of repetition and introduces non-controllable variance for trial-averaging.

Secondly, the single-channel responses may bear no obvious moment-by-moment relationship with the affective response that can be externally measured. Heterogeneity in the neural response do exist. In addition, the iEEG channel response is an aggregated response of the local spiking activity from several nearby neurons, which adds another layer of complexity.

Lastly, using the population analysis as the exploratory analysis provides a comprehensive initial assessment of the salient features of the data to guide subsequent analyses.

3.1 Quadratic Discriminate Classifier on neural response

In this section, we explore the performance of the quadratic discriminate classifier (QDA) on the neural feature. As shown in the graphical model (Figure 3.2.A), in order to decode the affective state A_t from the neural feature y_t, there are three assumptions made.

1. The class conditional distribution of the neural feature is Gaussian

$$p(y|\hat{A} = k) \sim \textit{Gaussian} \text{ for } k = \{H-, L+\} \tag{3.1}$$

2. Assume the affective state A_t at each time-point within the trial share the same affective label $\hat{A}$ as the stimuli presented in that trial.

$$A_1 = A_2 = \cdots = A_t = \hat{A} \tag{3.2}$$

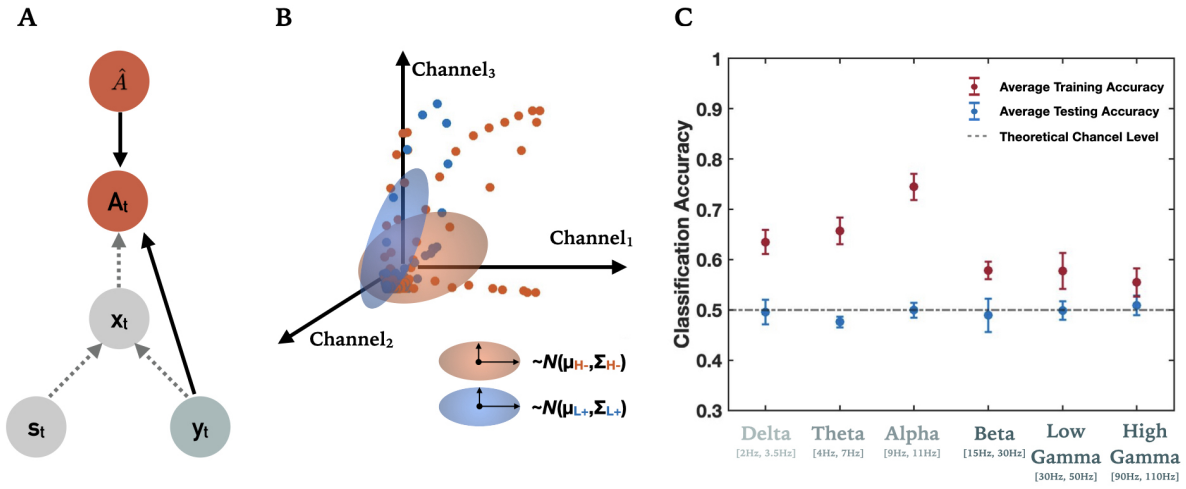

Figure 3.2. Quadratic Discriminate Classifier (QDA) on Neural Feature. (A) Graphical representation of the QDA decoder. The neural feature (y_t) propagates through time, at each point in time generating the affective state (A_t). (B) Schematic figure for the state view of the class conditioned distribution, red for affective state H-, blue for L+. Assume the neural response is Gaussian, fitting the mean vector (μ_{H-}, μ_{L+}) and the full covariance matrix (Σ_{H-}, Σ_{L+}). (C) Classification accuracy with four fold cross validation, assuming every time-point (At) within the trial share the same label as the trial label ($\hat{A}$).

3. Assume the neural feature at each time point is statistically independent.

$$p(y_{t_1}, y_{t_2}) = p(y_{t_1})p(y_{t_2}) \text{ for } t_1, t_2 = 1, \cdots, t \tag{3.3}$$

With those three assumptions above (Equation 3.1, 3.2, 3.3), we only treat the 91 iEEG channels as feature dimension and consider the neural feature at each time point as separate data points (Figure 3.2.B) that are equally informative of the affective label $\hat{A}$.

For each class $\hat{A} = \{H-, L+\}$, we specify the distribution of the 91-dimensional feature across all class-specific trials and all-time points as a Gaussian distribution.

$$p(y|\hat{A} = k) \sim N(\mu_k, \Sigma_k) \text{ where } k = \{H-, L+\} \tag{3.4}$$

During the training time, we approximate the mean vector μ_{H-}, μ_{L+} and the covariance matrix Σ_{H-}, Σ_{L+} using all time points within the training trials. Combining the prior class distribution $p(\hat{A} = H-) = p(\hat{A} = L+) = 0.5$, we test the classifier on the hold-out trials.

$$\hat{A} = argmax_{k=H-,L+}[logp(\hat{A} = k|\theta) + logN(y; \mu_k, \Sigma_k)] \text{ where } k = \{H-, L+\} \tag{3.5}$$

With four-fold cross-validation (Figure 3.2.C), the model achieves decent training accuracy across all frequency bands, but the testing accuracy sets around the theoretical chance level. The gap between the training and testing accuracy seems to imply overfitting due to variance in the model. In particular, estimating 8,372 parameters (91D Gaussian distribution) with 17,446 data points(143 trials each with 122 timepoints), we are data limited. Thus, proper regularization is needed. In addition, the neural response at the time when acoustic stimuli are silent may introduce unenchanted variance under the assumption that each time point within the trial is equally informative of the affective label. In conclusion, there is a decodable difference in neural feature y_t across the affective states. To further improve the performance of the decoder we need to properly regularize the decoder and explicitly model the temporal structure.

3.2 Gaussian Classifier on neural response with class-specific GPFA Model

At the end of the last section, we proposed two directions for improving the performance of the QDA decoder: to enforce the regularization and to model the temporal structure.

Explicit modeling of the temporal structure ensures the robustness of the data analysis and thus improves subsequent decoding performance. We can imagine a scenario when all the timepoint in the acoustic stimuli is randomly shuffled, listening to this constructed acoustic stimuli must end up with a different affective response compared to the original one. This simple mental experiment elucidates the importance of the evolution of the affective state and the internal state across time. To model such temporal structure, we need to expand the feature dimension

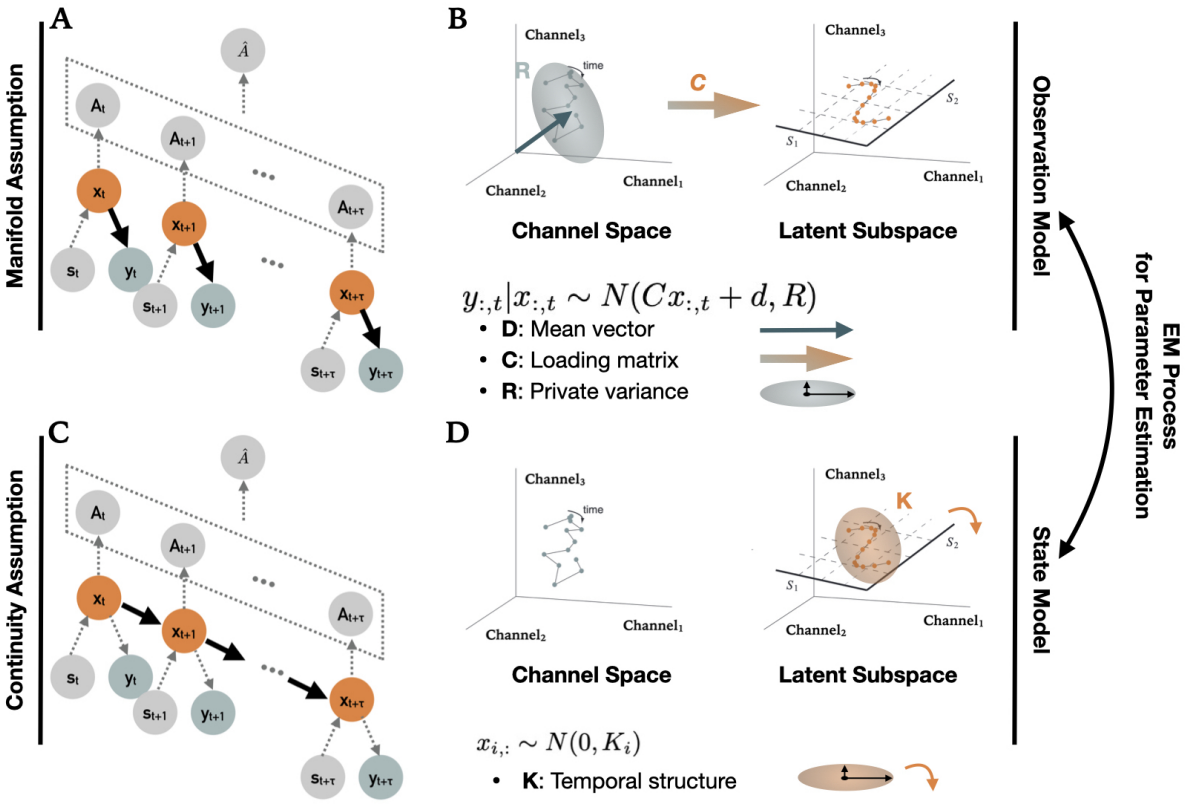

Figure 3.3. Schematic of the GPFA Model Assumption. (A) A graphical representation of the manifold assumption: neural response (y_t) lie on a low dimensional Population space. (B) The factor analysis based observation model in GPFA leverages the assumption in (A) and uses the mean vector d to capture the mean response across all trials, the loading matrix C for the transformation from the high-dimensional channel space to the low-dimensional latent subspace, the private variance R for the variance specific to each channel. (C) A graphical representation of the continuity assumption: internal state (x_t) close to each other in time are close by in the latent subspace. (D) The Gaussian process based state model in GPFA leverages the assumption in (C) and uses the GP covariance matrix K to capture the temporal structure of the inferred latent state (x_t).

by including each time point as a separate dimension, which results in 11,102 dimensions (91 sEEG channel, 122 time points). In this high-dimensional neural feature space, we are even more scarce in the data, so enforcing proper regularization in the decoder is even more necessary.

To regularize the decoder, there are two important and popular assumptions, in the population analysis framework, applicable to our decoding problem. The manifold assumption, which specifies that the neural response y_t lies on a low dimensional latent subspace (Figure

3.3.A); and the continuity assumption, which specifies that the internal state x_t close to each other in time are close by in the latent subspace (Figure 3.3.C). Those two assumptions lead to the method Gaussian process factor analysis (GPFA), which is a generative model that simultaneously models the correlation among channels and temporal structure across time [24].

The factor analysis-based observation model and the Gaussian process-based state model are two important components of the GPFA that leverage the two assumptions above to constraint the covariance of the neural feature. The factor analysis based observation model in GPFA leverages the manifold assumption (Figure 3.3.B) and uses the mean vector d to capture the mean response across all trials, the loading matrix C for the transformation from the high-dimensional channel space to the low-dimensional latent subspace, the private variance R for the variance specific to each channel. Another important component of the GPFA is the Gaussian process-based state model in GPFA, which leverages the continuity assumption (Figure 3.3.D) and uses the GP covariance matrix K to capture the temporal structure of the inferred latent state (x_t). Taken together, the GPFA model specifies the following covariance structure for the neural feature y.

$$\Sigma_y = CKC^T + R \tag{3.6}$$

With a class-specific GPFA model, the neural internal (x_t) state propagates through time by obeying the smooth property, at each point in time generating the observed neural feature (y_t) through the FA mapping. The combination of the observation model and the state model specifies the structure of the spatiotemporal pattern of the neural feature (Figure 8.B.ii) with the class-specific parameters C_k, R_k, d_k, K_k where $k = H-, L+$. Those GPFA parameter parameterizes the class conditional distribution in Equation (Figure 3.4.B.iii).

$$p(y|\hat{A} = k) \sim N(\mu_k, \Sigma_k) \text{ where } k = H-, L+, \mu_k = d_k, \Sigma_k = C_k K_k C_k^T + R_k \tag{3.7}$$

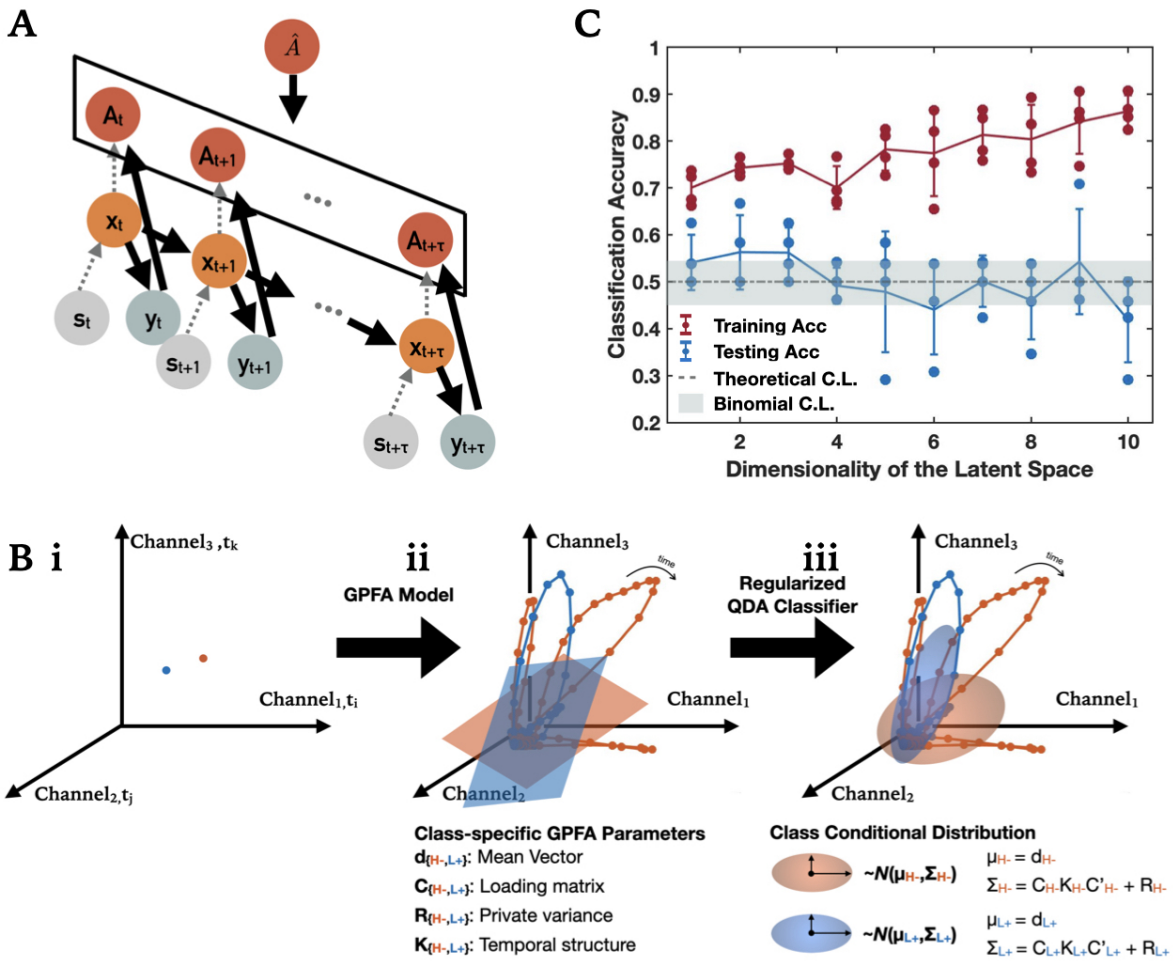

Figure 3.4. Gaussian Classifier on Neural Feature with Class-specific GPFA Model. (a) Graphical representation of the regularized QDA decoder. The neural internal (x_t) state propagates through time by obeying modeled dynamics, at each point in time generating the observed neural feature (y_t). The observed neural feature generates the affective response (A_t). (B) Schematic figure for the state view of the class conditional distribution, red for affective state H-, blue for L+. In (i) each dimension represent a channel at a time point, the neural feature scarcely populates this high-dimensional space. (ii) The GPFA model leverages the manifold and continuity assumption and specifies the neural response with few parameters. (iii) Parameterize the class conditional distribution with the GPFA parameters d, C, K, R. (C) Classification accuracy with four fold cross validation, assuming every time-point (At) within the trial shares the same label as the trial label ($\hat{A}$).

During the training time, we approximate the GPFA parameters C_k, R_k, d_k, K_k. Combining the prior class distribution $p(\hat{A} = H-) = p(\hat{A} = L+) = 0.5$, we test the classifier (Equation

3.5) on the hold-out trials.With four-fold cross-validation (Figure 3.4.C), we see marginal improvements in the testing decoding accuracy especially with GPFA model at the dimension 1, 2 and 3. In order to further improve the performance, we look into the components of the neural response that is most informative. In particular, the current formulation of the class conditional distribution with GPFA (Equation 3.7) fits a joint distribution of two important components of the neural response, the shared response and the private response, that are distinct in their characteristics and each could be informative of the affective state. To further regularize the decoder, it is important to simplify structure by de-emphasize less informative components while emphasize more informative components of the neural response. As the first step, we test the hypothesis: whether the latent state x_{H-} and x_{L+} occupy the same latent subspace, i.e. $C_{H-} \cong C_{L+}$. To compare the latent subspace, we compute the angle between the latent subspace as the measurement for similarity [26]. With $C_1^T C_1 = C_2^T C_2 = 1$

$$\cos\theta(C_1, C_2) = |C_1^T C_2| \text{ where } 0 \leq \theta(C_1, C_2) \leq \frac{\pi}{2} \tag{3.8}$$

As the dimensionality of the space increase, any two random subspace are increasingly likely to be orthogonal of each other [7]. To establish the baseline for the latent subspace alignment, we compute the within-class subspace alignment (Figure 3.5.A). If we can robustly estimate the latent subspace, the latent subspace learned with two sets of trials sharing the same affective label should be well aligned. The latent subspace captures the shared variance in the neural response. To avoid distortion of alignment due to overlapping trials, we randomly split the trials sharing the same affective label and learn the latent subspace for each corresponding set of trials. When the latent dimension is 1, the learned subspace is well aligned (Figure 3.5.A.II). As the latent dimensionality increases, the learned subspace soon becomes orthogonal of each other. It is possible that there is not enough data to robustly estimate the latent subspace, but with the angle between the principal subspace of the 5D latent subspace, the first two-dimensional principal subspace is still well aligned. It provides us confidence that the estimation of the latent

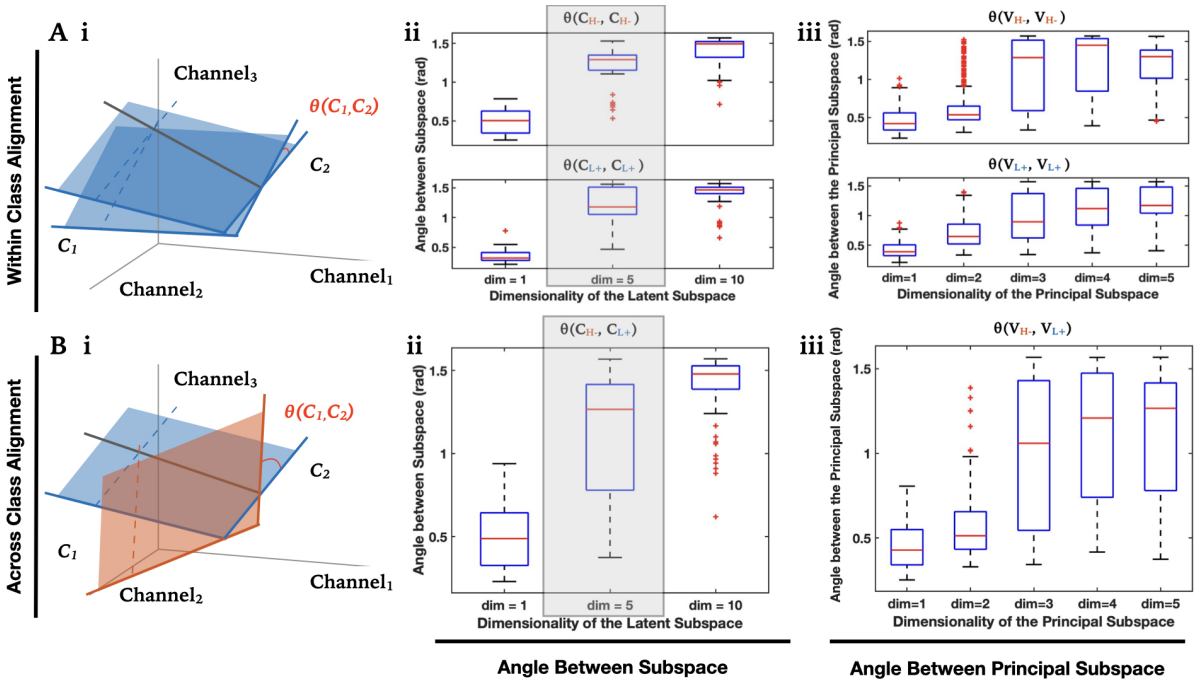

Figure 3.5. Latent Subspace Alignment with Bootstrap Analysis. Hypothesis testing for the subspace alignment with the bootstrap analysis, using the cosine angle between the subspace as the test statistics. (A) Establishing the chancel level using the within class alignment. (i) Randomly split the trials sharing the same affective label (H-, L+) and learn the latent subspace C1,C2 for each corresponding set of trials. (ii) The distribution of the angle between subspace in radius after 1000 bootstrap iterations. (iii) For each dimension of the latent subspace, align the subspace using the principal axis. Increase the dimensionality of the principal subspace by gradually include principal axis in the order of fraction of the variance explained from high to low. (B) Test the across class alignment. Randomly pair the splits of trials obtained above for each affective label (H-, L+). (ii) and (iii) are the result after 1000 bootstrap iterations

subspace is robust, but we also need to be aware that the measurement of alignment with cosine angle is very sensitive to the dimensionality of the space. High dimensional alignment may be missed due to this sensitivity.

After establishing the baseline for the latent subspace alignment, we test the across-class alignment by randomly pair the splits of trials obtained above for each affective label (H-, L+). Similar distribution in the angle between subspace (Figure 3.5.B.ii) and principal subspace (Figure 9.B.ii) are observed compared to that for within-class alignment. Though more stringent hypothesis testing is still needed to test this hypothesis, it opens up the opportunities to consider

fitting a class-invariant GPFA and leverage the data across all affective classes to constrain the learned GPFA model.

3.3 Gaussian Classifier on shared / private response with class-invariant GPFA Model

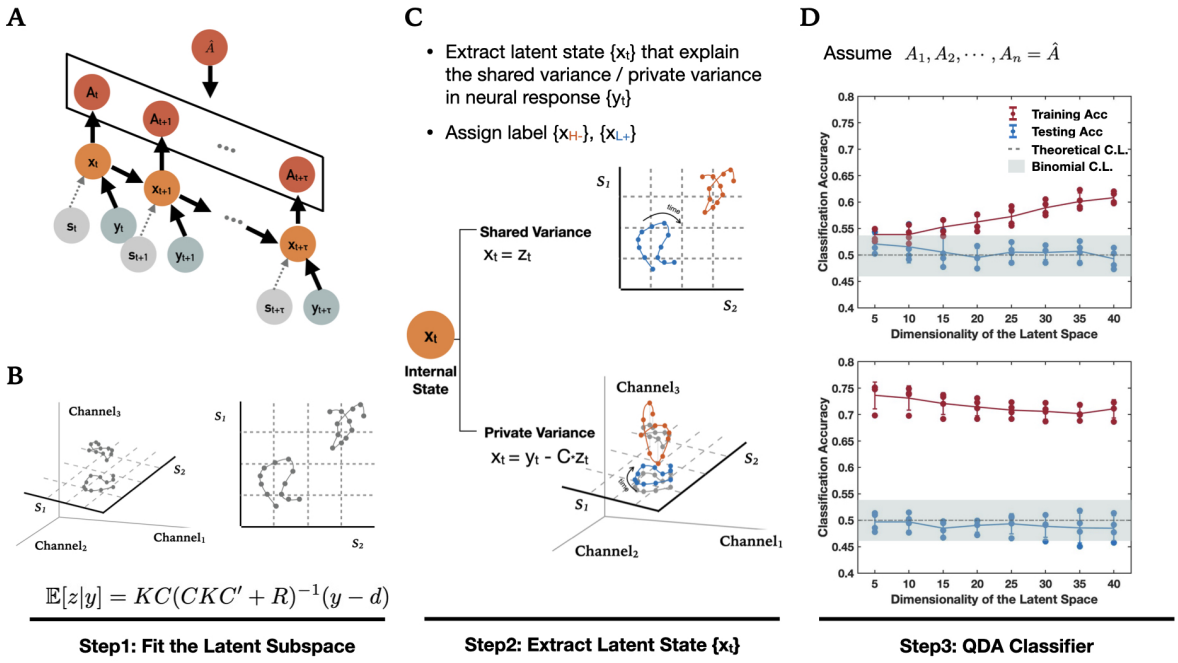

Figure 3.6. Gaussian Classifier on Shared / Private Response with Class-invariant GPFA Model. (A) A graphical representation of a proposed neural dynamical filter, modeling the dynamics of the neural state (x_t). The neural internal (x_t) state propagates through time obeying modeled dynamics, at each point in time generating both the affective response (A_t) and the observed neural feature (y_t). (B) A schematic example for fitting the latent subspace using all trials across affective label. (C) Define two types of internal state. The internal state (x_t) as the inferred trajectory (z_t) from the GPFA, that explains the shared variance in neural feature (y_t). The internal state (x_t) as the residual response ($x_t = y_t - Cz_t$) from the GPFA, that explains the private variance in neural feature (y_t) specific for each channel. (D) The performance of the QDA classifier on the two internal state (x_t) defined in (C), corresponding to the component of neural feature (y_t) that explains the shared variance / private variance.

The shared subspace alignment analysis towards the end of last section seems to hinting a shared latent subspace between the two affective states. In this section, we learn a class-invariant

GPFA model using all trials across the affective state, and rethink the definition of the internal state in consideration of the components of the neural response that explain the shared variance or the private variance. To distinguish the internal state x_t from the inferred neural trajectory in GPFA from last section, we denote the inferred GPFA neural trajectory as z in this section.

As the first step, we fit a class-invariant GPFA model with all trials and learn the associated GPFA parameters C, R, d, K (Figure 3.6.B). Then define the internal state x either by the inferred trajectory z, which explains the shared variance, or the residual response $y - Cz$, which explains the private variance (Figure 3.6.C). We train the QDA classifier with full covariance on each of the defined internal state. The training performance for the shared response is low especially in the lower dimension subspace may suggest that there is an affective-state independent neural response that explains most of the variance across channels. In contrast, we did see a higher decoding accuracy in training for the private response, but the gap between the training and testing again seems to point to regularization for further improvements (Figure 3.6D.).

Considering the two affective states share the same latent subspace, we may further refine our hypothesis for the structure in the shared / private response to two categories: the hypothesis associate with the distribution and that associate with the dynamics (Figure 3.7). In particularly, test whether the inferred internal state x_{H-} differs from x_{L+} in its distribution in the latent subspace (Figure 3.7.A) or the residual response x_{H-} differs from x_{L+} in its probability and intensity diverging the latent subspace (Figure 3.7.C). To decide on the salient feature for the decoder. In addition, test whether the inferred internal state (or the residual response) x_{H-} differs from x_{L+} in its dynamics (e.g. smooth, oscillatory) to specify the corresponding forms the GP covariance for the dynamic pattern (Figure 3.7B,D).

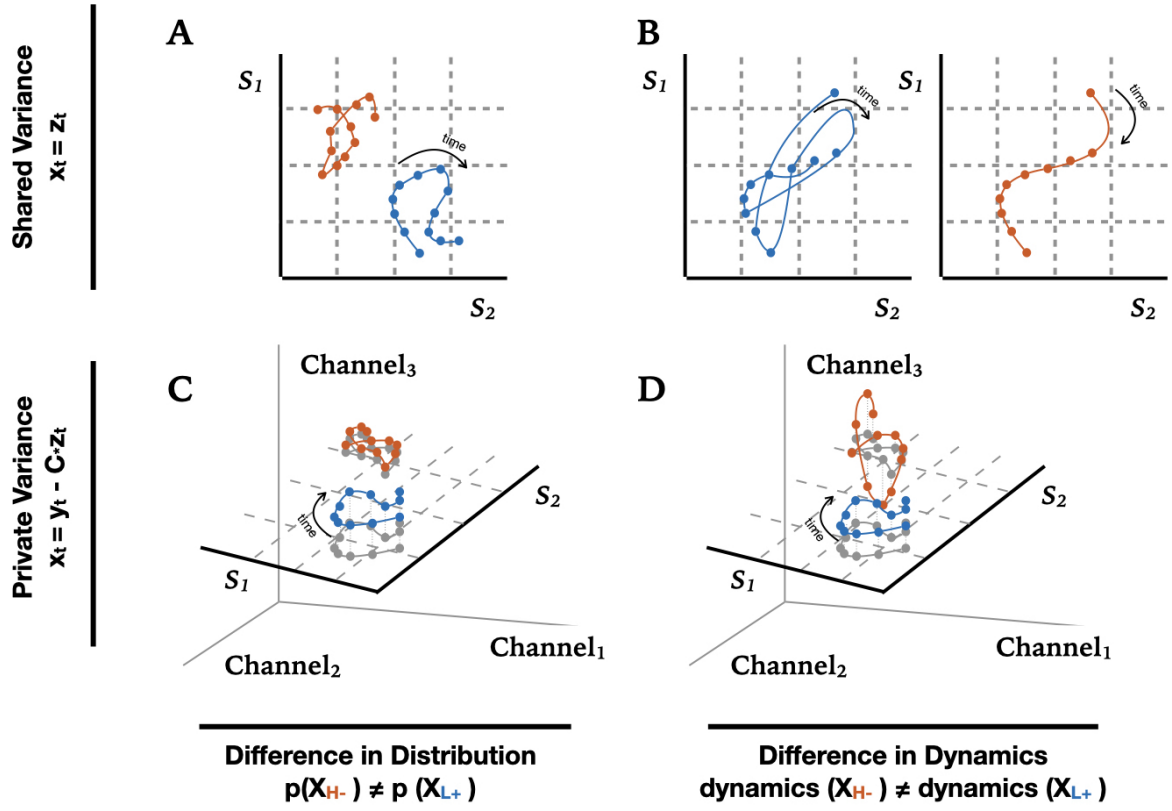

Figure 3.7. Hypothesis Testing for Statistical Difference on Shared / Private Response. Assume the the two affective class H- and L+ share the same GPFA latent subspace. Define the internal state x_t to explain the shared variance $x_t = z_t$. (A) Hypothesis: the inferred internal state X_{H-} differs from X_{L+} in its distribution in the latent subspace. (B) Hypothesis: the inferred internal state X_{H-} differs from X_{L+} in its dynamics (e.g. smooth, oscillatory). Define the internal state $x_t = y_t - Cz_t$ (C) Hypothesis: the residual response X_{H-} differs from X_{L+} in its probability and intensity diverging the latent subspace. (D) Hypothesis: the residual response X_{H-} differs from X_{L+} in its dynamics.

Chapter 4

Conclusion and Future Work

4.1 Summary of the Results

In conclusion, we conceptualize the neural representation of the affective response with three observable variables (neural feature y_t, acoustic stimuli s_t, affective label $\hat{A}$) and two hidden variables (internal state x_t, affective state A_t) through the graphic model specified in (Figure 1.2). We sequentially improve the decoder performance through feature selection driven by the result of the statistical analysis.

We started with the QDA analysis on neural response, and found decodable differences of neural features across all frequency bands. To further improve the performance of the decoder, we introduce the generative model of GPFA to regularize the form of covariance matrix in QDA. In addition, the GPFA model characterizes the temporal structure of the internal state previously assumed to be discrete and independent. Marginal improvements in the decoding performance are observed especially in low dimension latent subspaces. The additional shared subspace alignment analysis suggests a simplified model to overcome the challenge of data limit.

Due to the similiarity in the latent subspace learned with class-specific GPFA, it is more optimal to use data from all trials to constrain one class-invariant GPFA model and examine the characteristic of the neural response through the lens of the class-invariant GPFA model. With the shared latent subspace, it makes sense to provide two complementary definitions of the internal state: the neural response of the shared variance (projection on the latent subspace) and

that of the private variance (divergence away from the latent subspace). Here we generate four working hypothesis for feature selection and decoder model improvement in the future.

4.2 Future work

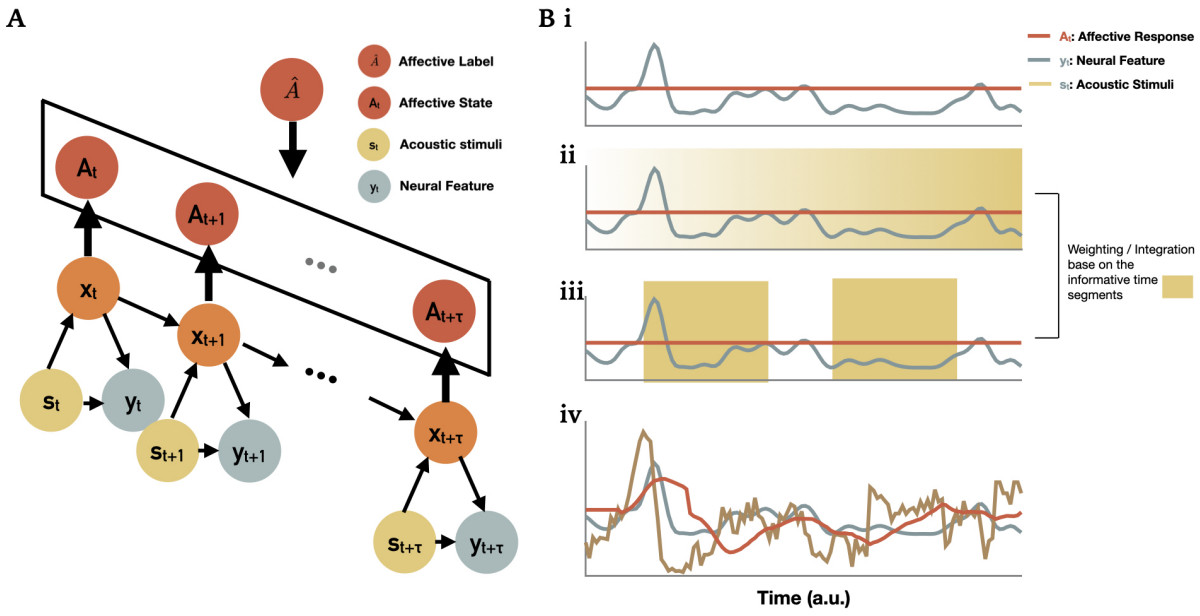

Figure 4.1. Integrate the Acoustic Feature. (A) Graphical representation of the regularized QDA decoder incorporate the acoustic feature. The acoustic stimuli (s_t) at each time influences the neural internal state (x_t) and neural feature (y_t). The neural internal state (x_t) propagates through time by obeying modeled dynamics, at each point in time generating both the affective response (A_t) and the observed neural feature (y_t). (B) Schematic figure for the proposed analysis. The existing analysis assume the model (i), where each time points is treated as equally informative of the the affective label. Model (ii) hypothesizes a linear integration of the affective response as a function of time. When approaching the end of the trial, the neural response become more informative of the affective label. Model (iii) provides an alternative structure in contrast to model (ii) by assuming that the affective response is transient, only informative in certain segments of the trial where acoustic events happened. Model (iv) is a more expressive model which explicitly model the instantaneous affective response (A_t) with the temporal structure of the acoustic stimuli (s_t).

The discussion within this section centers around the temporal influence of the acoustic stimuli on the neural representation of the affective response, motivated from two separate perspectives.

From the modeling and analysis standpoint, we see marginal improvements in the decoder when incorporating the temporal structure compared to the decoder regarding each time point as a discrete and independent data point. This marginal but noticeable improvement can be explained by the oversimplification in using the time-invariant affective response as the class label for decoder. We are only able to decode such nosiy affective state label from the 2D neural subspaces. Considering that the estimated intrinsic dimensionality of the neural response (using the GPFA method) is 10D, we see overfitting when including the remaining 8D neural response, whose variability could be explaining the time characteristics of the affective responses.

From the theory standpoint, we are interested in designing a universal BCI control system that dynamically interacts with a neural interface. Using the affective response as conditional control, we need to decode the behavior and affective response from the neural response changing over time. Such a system requires temporal characteristics to carry its own meaning, a chronological logic, because it implies a natural evolution of the affective state which could be generalizable to other tasks. By extracting the time characteristic from the task, which interplay with the affective state, we no longer limit ourselves to the features specific to the task, but rather focus on a more generalizable feature that can be incorporated into the model. For example, the time characteristics we observe in our experiment while listening to the acoustic stimuli may also apply to more interactive tasks such as gameplay.

Revisit the mental experiment introduce in Chapter 3. Imagining, if you will, by introducing the new sound ”I love you” in the experiment for the subjects. How could such sound lead to his/her strong emotional response? Here we can assume “love” sound is a high valence, positive arousal emotion state for most human beings. The acoustic stimuli “I love you” can be depicted and analyzed from different perspectives. We mainly want to analyze the time configuration of acoustic stimuli’s influence on the affective state. To be precise, we want to illustrate the above perspectives using a high cognitive analogy while redefined the model through an engineering angle of view.

In an ideal scenario, with the perfect decoder, we can estimate the ground-truth trajectory

of the affective response, and leverage it to learn the one-on-one mapping with the neural response. In reality, such ground-truth trajectory of the affective response doesn't exist, so we are trying to design a system that simultaneously approximates the affective response trajectory and its mapping to the neural response, as an integration process. Such a Hypothesis-driven model design with inspiration from the cognitive process assumes a certain time structure for the integration process. Four different models will be examined here, each associate with the different time integration configuration.

Listening to an acoustic stimulus and a new construction of this stimulus by shuffling its time points could elicit a very different affective response. As an illustrative example, will you be in the same affective state when hearing 'I love you ' or 'love I you' or even 'I you love'? The affective experience is different. Such differences imply the importance of time in the meaningful affective response. For the model introduces in Chapter 3, the QDA model treats the response at all time points as the same (Figure 4.1.B.i) and disregards the meaningful evolution of the affective and neural state we highlight here. Considering the importance of the time factor on the configuration of the integration process, we are inspired to think through the subsequent model variation.

The neural response accumulates for the affective state as time elapses. Specifically, the neural representation of the affective response at the beginning of the trial may contain more uncertainty than that at the end of the trial. For example, from 'I .....' to 'I lo.....', then to 'I love you', as we hearing more words, we have a more certain affective response to this sound. Its implication to the decoder model is that we may need to consider weighting the importance of the time point for decoding base on its distance away from the end of the trial (Figure 4.1.B.ii). To test whether such a relationship exists in our neural feature, we will use the KL divergence to examine the informativeness of the neural response as a function of time away from the end of the trial.

Our affective state may only be driven by certain key elements of the sound. Those key elements alone can trigger a strong affective response that dominates the affective response of

the entire sound. For example, the affective response (high valence positive arousal) to the sound 'i love you' may be directly driven by the 'love' sound. To incorporate such observation in our model we may only need to consider a brief integration window for the neural representation. Specifically, the neural response at time points around the identified salient elements of the sound and treat the rest of the neural response as non-informative spontaneous response (Figure 4.1.B.iii). The limitation of this approach is that the annotation of the salient time points is based on the human experience. There are two parts that might introduce the bias: the identification of the salient point; the beginning and end of the integration window. In addition, there might be a complex integration structure within the time window. One example is that the time points closer to the salient point could be more informative of the affective response than those further away, but there're many more possible integrations to be explored.

Instead of testing all variations of the proposed design shown above (Figure 4.1.B.ii and 4.1.B.iii), we could consider adopting a more expressive model (Figure 4.1.B.iv) to estimate such integration function through a data-driven approach. To explore the integration structure and overcome the bias in the human labeling, we may use the acoustic features (tone, etc.) as the noisy estimation of the ground-truth affective trajectory and learn mapping between this estimated affective trajectory and the neural response. Such decoder takes time configuration into account while maintaining enough flexibility in the structure of the integration process. In addition to the acoustic stimuli, other physiology measurements may also be considered as potential source of information to estimate this ground-truth affective trajectory (e.g. the heart-rate).

Last but not least, we want to reanalyze a question: does the sound 'i love you' itself alone incite a strong cognitive response? Or 'I love you 'sound reminds subjects of the scene like 'confess sb' love', 'proposal', or even 'wedding' that actually trigger a strong affective response. Such questions seem too hard to answer for now. but it shed some insight for our future research interest: Is there any principal element, beyond sound, image, or sense, that can be viewed as the leading factor for trigger neuron response that driven the affective response?

Acknowledgment

This paper is coauthored with Aashish N.Patel; Jerry Shih; Vikash Gilja. The thesis author was the primary author of this thesis paper.

Bibliography

[1] Claude Bedard, Helmut Kroeger, and Alain Destexhe. Does the 1/f frequency scaling of brain signals reflect self-organized critical states? *Physical review letters*, 97(11):118102, 2006.

[2] Joydeep Bhattacharya and Hellmuth Petsche. Universality in the brain while listening to music. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 268(1484):2423–2433, 2001.

[3] Danny Oude Bos et al. Eeg-based emotion recognition. *The influence of visual and auditory stimuli*, 56(3):1–17, 2006.

[4] Margaret M Bradley and Peter J Lang. Measuring emotion: the self-assessment manikin and the semantic differential. *Journal of behavior therapy and experimental psychiatry*, 25(1):49–59, 1994.

[5] Hang-Yee Chan, Ale Smidts, Vincent C Schoots, Alan G Sanfey, and Maarten AS Boksem. Decoding dynamic affective responses to naturalistic videos with shared neural patterns. *Neuroimage*, 216:116618, 2020.

[6] Junichi Chikazoe, Daniel H Lee, Nikolaus Kriegeskorte, and Adam K Anderson. Population coding of affect across stimuli, modalities and individuals. *Nature neuroscience*, 17(8):1114–1122, 2014.

[7] John P Cunningham and M Yu Byron. Dimensionality reduction for large-scale neural recordings. *Nature neuroscience*, 17(11):1500–1509, 2014.

[8] Nima Dehghani, Claude Bédard, Sydney S Cash, Eric Halgren, and Alain Destexhe. Comparative power spectral analysis of simultaneous elecroencephalographic and magnetoencephalographic recordings in humans suggests non-resistive extracellular media. *Journal of computational neuroscience*, 29(3):405–421, 2010.

[9] Paul Ekman. An argument for basic emotions. *Cognition & emotion*, 6(3-4):169–200, 1992.

[10] Bruno L Giordano, Caroline Whiting, Nikolaus Kriegeskorte, Sonja A Kotz, Joachim Gross, and Pascal Belin. The representational dynamics of perceived voice emotions evolve from categories to dimensions. *Nature Human Behaviour*, pages 1–11, 2021.

[11] Dirk T Hettich, Elaina Bolinger, Tamara Matuz, Niels Birbaumer, Wolfgang Rosenstiel, and Martin Spüler. Eeg responses to auditory stimuli for automatic affect recognition. *Frontiers in neuroscience*, 10:244, 2016.

[12] Xin Hu, Jingjing Chen, Fei Wang, and Dan Zhang. Ten challenges for eeg-based affective computing. *Brain Science Advances*, 5(1):1–20, 2019.

[13] Karim S Kassam, Amanda R Markey, Vladimir L Cherkassky, George Loewenstein, and Marcel Adam Just. Identifying emotions on the basis of neural activation. *PloS one*, 8(6):e66032, 2013.

[14] Brian Knutson, Kiefer Katovich, and Gaurav Suri. Inferring affect from fmri data. *Trends in cognitive sciences*, 18(8):422–428, 2014.

[15] Sukhbinder Kumar, Katharina von Kriegstein, Karl Friston, and Timothy D Griffiths. Features versus feelings: dissociable representations of the acoustic features and valence of aversive sounds. *Journal of Neuroscience*, 32(41):14184–14192, 2012.

[16] E Novikov, A Novikov, D Shannahoff-Khalsa, B Schwartz, and J Wright. Scale-similar activity in the brain. *Physical Review E*, 56(3):R2387, 1997.

[17] Jonathan Posner, James A Russell, and Bradley S Peterson. The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology. *Development and psychopathology*, 17(3):715, 2005.

[18] Walter S Pritchard. The brain in fractal time: 1/f-like power spectrum scaling of the human electroencephalogram. *International Journal of Neuroscience*, 66(1-2):119–129, 1992.

[19] James A Russell. A circumplex model of affect. *Journal of personality and social psychology*, 39(6):1161, 1980.

[20] Xinke Shen, Xin Hu, Shizhao Liu, Sen Song, and Dan Zhang. Exploring eeg microstates for affective computing: decoding valence and arousal experiences during video watching. In *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 841–846. IEEE, 2020.

[21] Ryan A Stevenson and Thomas W James. Affective auditory stimuli: Characterization of the international affective digitized sounds (iads) by discrete emotional categories. *Behavior research methods*, 40(1):315–321, 2008.

[22] Ante Topic and Mladen Russo. Emotion recognition based on eeg feature maps through deep learning network. *Engineering Science and Technology, an International Journal*, 2021.

[23] Wanlu Yang, Kai Makita, Takashi Nakao, Noriaki Kanayama, Maro G Machizawa, Takafumi Sasaoka, Ayako Sugata, Ryota Kobayashi, Ryosuke Hiramoto, Shigeto Yamawaki, et al. Affective auditory stimulus database: An expanded version of the international affective digitized sounds (iads-e). *Behavior research methods*, 50(4):1415–1429, 2018.

[24] Byron M Yu, John P Cunningham, Gopal Santhanam, Stephen I Ryu, Krishna V Shenoy, and Maneesh Sahani. Gaussian-process factor analysis for low-dimensional single-trial analysis of neural population activity. *Journal of neurophysiology*, 102(1):614–635, 2009.

[25] Guanhua ZHANG, Minjing YU, Guo CHEN, Yiheng HAN, Dan ZHANG, Guozhen ZHAO, and Yong-Jin LIU. A review of eeg features for emotion recognition. *SCIENTIA SINICA Informationis*, 49(9):1097–1118, 2019.

[26] Peizhen Zhu and Andrew V Knyazev. Angles between subspaces and their tangents. *Journal of Numerical Mathematics*, 21(4):325–340, 2013.