Towards EEG-based Emotion Recognition During Video Viewing: Neural-Congruency Explains User's Emotion experienced in Music Videos

Christoforos Christoforou, Maria Theodorou
Department of Computer Science, St. John's University, New York, NY, USA,
christoc@stjonhs.edu, mtheodorou@rkileaders.com

Abstract

Emotions affect our decisions, experiences, preferences, and perceptions. Understanding the neural underpinning of human emotions is a fundamental goal of neuroscience research. Moreover, EEG-based emotion recognition is a key component towards the development of affective-aware intelligent systems. However, characterizing the neural basis of emotions elicited during video viewing has been proven a challenging task. In this paper, we propose a novel machine-learning approach to isolate neural components in EEG signals that are informative of the affective content of emotionally-loaded videos. Based on these components, we define a set of neural metrics and evaluate them as potential indicators of the overall emotional content of each video. We demonstrate the predictive power of the proposed metrics, on the DEAP benchmark dataset for EEG-based emotion recognition. Our results provide novel empirical evidence that the neural components extracted by our method can serve as an informative metric in EEG-based emotion recognition during video viewing and achieving a 4-fold increase in predictive power compared to traditional frequency-based metrics. Moreover, each extracted component is associated with a spatial and a temporal profile, that allows researchers to inspect and interpret the spatiotemporal origins of the underlying neural activity. Thus, our method a framework that facilitate the study of neural correlates of emotion during video viewing.

Introduction

Electroencephalography (EEG)-based emotion recognition during video viewing has attracted the interest of many disciplines and affects domains such as artificial intelligence (AI), affective computing, cognitive neuroscience, and human-computer interaction (HCI). From the neuroscience perspective, the objective of EEG-based emotion recognition is to identify neural components that characterize a human's emotional state and to gain insights into the underlying cognitive processes involved in emotion generation. Such insights could lead to a better understanding of underlying cognitive processes and disorders affected by emotions, such as decision making and desensitization to media violence. In the context of affective computing, AI and HCI, EEG-based emotion recognition could facilitate the design of emotion-aware intelligent systems that interpret human emotions and generate responses adaptable to the user's emotional state. Moreover, implicit tagging of video using affective information can improve the performance of affective-aware video and music recommendation systems. Yet, characterizing the neural-underpinnings of emotions during video viewing remains a challenging task. There is a need for novel computational methods to identify neural components from EEG measurement during video viewing that characterize the emotional state of humans consuming video content.

Current approaches to EEG-based emotion recognition have focused on either using the power in EEG signals in specific frequency bands (i.e., frontal-asymmetry) or relied on event-related potential (ERP) paradigms. Traditionally, EEG signals are analyzed in specific frequency-bands such as theta band (3-7Hz), alpha-band (8-13 Hz), beta-bad (14-29Hz), and gamma-band (30Hz-47Hz). In particular, changes in the overall power -relative to the baseline within selected bands during a stimulus presentation- have been explored as indicators of cognitive process modulation. In the context of EEG-emotion recognition, frontal-asymmetry – which quantifies the asymmetry in the alpha-band power recorded over the frontal lobe - has been extensively used as an indicator of emotional arousal (Davidson,1992, Petrantonakis and Hadjileontiadies 2011). However, these metrics are ad-hoc in nature and exhibits large variability across individuals. Machine learning feature extraction approaches have been proposed to extract neuronal activity that maximally differentiates among cognitive conditions in selected frequency bands (Christoforou et. al. 2018). Alternatively, ERP-related methods rely on the design of experimental paradigms that present brief emotional stimuli to participants to elicit ERPs - a stereotypical neural-response followed a stimuli presentation-

and extract components in the ERP that are modulated by stimulus emotion strength. Machine learning approaches have been proposed to extract informative components from ERP either on average or on a single-trial basis (Christoforou et. al. 2013, Philiastides and Sajda 2005, Christoforou et. al. 2010). However, due to methodological constraints, the application of ERP methods is limited to simplistic image stimuli and does not apply to dynamic stimuli such as video or music content.

Recently, few studies tried to exploit the synchrony in EEG signals while participants watch video stimuli to extract informative neural components. Such methods have were applied in predicting population-wide user preferences to video advertisement (Dmochowski et. al. 2014), predicting box-office sales performance of movies (Christoforou et. al. 2017), and assessing the effectiveness of educational videos (Cohen et. al. 2018). However, to the best of our knowledge, such approaches have not been used to characterize the emotional state of individuals during video viewing.

Several deep learning approaches have been explored in the analysis of EEG signals for emotion recognition (Lue et al. 2020; Zhong et al. 2020; Alhagry et. al 2017). These models formulate emotion recognition as a binary classification problem to classify low-vs-high emotional state (typically measured by valance, arousal, and dominance scores). However, such models do not capture the granularity in human emotions. The labeling of high-vs-low emotion states is arbitrary and subjective. Importantly, due to the black-box nature of deep neural networks, these models do not generate interpretable neural components that could provide insights for understanding the temporal and spatial dynamics of emotions at a neural level.

In this paper, we propose a novel framework for EEG-based emotion recognition during the viewing of emotional musical videos. The framework provides an approach to extract interpretable spatial and temporal components that can be associated with arousal, valance, and dominance. By design, the component provides full granularity over the spectrum of emotion measurements. Unlike traditional ERP it is applied directly to video stimuli and does not rely on heuristic metrics, but rather targeted features extracted from the data. The impact of the method is demonstrated on the benchmark dataset for real EEG-based emotion recognition.

Methods

Benchmark Dataset for Emotion Recognition

The growing interest in emotion recognition has motivated the creation of several benchmark datasets of electrophysiological (i.e. EEG) measurement and other modalities (such as ECG, GSR, EMG, respiration patterns, facial expression, among others. In this study, we apply our proposed framework to one of the most prominent datasets on emotion recognition, namely the DEAP dataset (Database for Emotional Analysis using Physiological Signals). In this section, we briefly describe this dataset.

The DEAP dataset contains EEG recordings from 32 participants (50% female; aged between 19 and 37; average age 26.9) while watching one-minute long segments. In total, participants watch 40 music video clips, selected to elicit different emotions. The recording used 32 EEG channels and eight peripheral channels for physiological signals (such as galvanic skin conductance [GSR], Electromyogram [EMG], and Electroculogram [EOG]). All signals were recorded at a sampling rate of 512Hz and were synchronized with an event trigger channel. For each music video clip, a one-minute segment with the maximum emotional content was extracted and presented to the participants. Each trial (i.e. a video viewing) consisted of a 2-second screen displaying the current trial number; a 5-second baseline recording (i.e. fixation cross); a one-minute display of the music video segment, followed by a self-assessment screen for arousal, valance, dominance, and liking (using the Self-Assessment manikins to visualize the scale). The dataset also provides real-value scores indicating the level of arousal, valance, dominance, and liking for each video segment, as experienced by the participants. A full description of the experimental paradigm and data collection procedure can be found in (Koelstra et al. 2011).

EEG data pre-processing

For our analysis, we considered the pre-processed version of the EEG data provided by the DEAP dataset. In this version, EEG data were down-sampled to 128Hz; eye artifacts were removed using a blind-source separation technique (Koelstra et al. 2011). Then the data were band-passed filtered between 4Hz to 45H. Continued EEG was then segmented in 60-second trials and a 3-second pre-trial baseline was removed. All channels were re-references to the average channel. Moreover, we normalize each segment by dividing each channel by the standard deviation across time. After pre-processing, the resulting dataset is defined by EEG segments of each participant and each video segment, as follows:

$$Data = \{X_v^s \in \mathbb{R}^{D\times T}, \forall s \in S, v \in V\}$$

$$Labels = \{a_v, v_v, d_v \in [0,10], \forall v\}$$

where each $X_v^s \in \mathbb{R}^{D\times T}$ correspond to the EEG segment obtained from participant s while viewing video segment *v*, *D* corresponds to the number of EEG channels (i.e. D=32 in this study) and T corresponds to the number of time samples within each segment, S is the set of all partici-

pants, V is the set of all videos. For each video *v*, the label values $a_v, v_v, d_v \in [0,10]$ correspond to the *arousal*, *valance* and *dominance* scores for video *v* calculated by aggregating the responses to the self-assessment scores across all participants.

Neural-Congruency Components

Our objective is to identify neural components (i.e. spatial projection of the EEG signals) that are modulated by emotions elicited while participants watch each video and thus are informative of the emotional content of each video as captured by the arousal, valance, and dominance scores. Our approach is motivated by the hypothesis that synchrony of neural responses between individuals while watching an emotional video can carry information about the underlying cognitive processes involved in emotional experiences. Here we provide details of the proposed approach to isolate such components.

Consider the set of all trials (i.e. viewings) for a video *v* and all participants $\{X_v^1, X_v^2, \dots X_v^{|S|}\}$, where S denotes the set of subject that watch the particular video. Then for a given projection vector $\boldsymbol{w} \in \mathbb{R}^D$ and a pair of participants $(i,j) \in S \times S$, the between-subject Pearson Correlation Coefficient of the projected components is given by

$$\rho(\boldsymbol{w}; i, j, v) = \frac{\boldsymbol{w}^{\mathrm{T}} \boldsymbol{R}_{ij}^{v} \boldsymbol{w}}{(\boldsymbol{w}^{\mathrm{T}} \boldsymbol{R}_{ii}^{v} \boldsymbol{w})^{\frac{1}{2}} \left(\boldsymbol{w}^{\mathrm{T}} \boldsymbol{R}_{jj}^{v} \boldsymbol{w}\right)^{\frac{1}{2}}}$$

where $R_{ij} \in \mathbb{R}^{D \times D}$ is the cross-covariance matrix between trials defined as follows:

$$\boldsymbol{R}_{ij}^{v} = \frac{1}{T} \boldsymbol{X}_v^i X_v^{j\mathrm{T}}$$

We note that the italics *T* denotes the number of temporal samples, while the non-italic superscript T denotes the transpose operation). We can now consider the inter-subject Pearson Correlation Coefficient of a projected component across all videos and all participant pairs as follows:

$$\rho(\boldsymbol{w}) = \frac{1}{M} \sum_{v \in V} \sum_{\substack{i,j \in S \times S \\ i \neq j.}} \frac{\boldsymbol{w}^{\mathrm{T}} \boldsymbol{R}_{ij}^{v} \boldsymbol{w}}{(\boldsymbol{w}^{\mathrm{T}} \boldsymbol{R}_{ii} \boldsymbol{w})^{\frac{1}{2}} \left(\boldsymbol{w}^{\mathrm{T}} \boldsymbol{R}_{jj}^{v} \boldsymbol{w}\right)^{\frac{1}{2}}} \qquad (1)$$

where M is a normalization factor $M = |S|(|S| - 1)$. With that, we seek to find a set of optimal spatial projector vector $\boldsymbol{w}$ that maximizes the average Person Product Moment Correlation Coefficient across all subject pairs. Formally, the optimization problem seeks to find $\widehat{\boldsymbol{w}}$ as :

$$\widehat{\boldsymbol{w}} = \arg_w max\ \rho(\boldsymbol{w})$$

Taking the derivative of equation (1) with respect to **w** and setting it to zero, and further assuming that the dataset have similar power levels (i.e. approximating $\boldsymbol{w}^T R_{ii} \boldsymbol{w} \approx \boldsymbol{w}^T R_{jj} \boldsymbol{w}\ \forall (i,j)$), we obtain that an optimal $\widehat{\boldsymbol{w}}$ is a solution to the following generalized eigenvalue problem (see appendix for derivation):

$$\boldsymbol{R}^{(b)}\, \boldsymbol{w} = \lambda\, \boldsymbol{R}^{(w)} \boldsymbol{w} \qquad (2)$$

where:

$$R^{(b)} = \frac{1}{M} \sum_{v \in V} \sum_{\substack{i,j \in S \times S \\ i \neq j}} R_{ij}^{v}$$

$$R^{(w)} = \frac{1}{|S|.|V|} \sum_{v \in V} \sum_{i \in S} R_{ii}^{v}$$

where $R^{(b)}, R^{(w)}$ are the between-subject and within-subject covariance matrix across all video viewings. Solutions to the generalized eigenvalue problem comprise the K eigen-vectors $\{\boldsymbol{w}_k\}_{k=1}^{K}$ of the matrix $\left(R^{(w)}\right)^{-1} R^{(b)}$, and their corresponding eigenvalues given by $\lambda_k = \frac{\boldsymbol{w_k}^T R^{(b)} \boldsymbol{w_k}}{\boldsymbol{w_k}^T R^{(w)} \boldsymbol{w_k}}$. Each eigenvector represents a component projection that captures neural activity with the largest correlation across participants while consuming an emotional video, while the corresponding eigenvalue represents the strength of that correlations. We hypothesize that the components that exhibit high correlation during the viewing of emotional videos capture neural activity that is modulated by the underlying emotions and thus can be predictive of the emotional experience of viewers. Therefore, we consider the eigenvalues associated with each component as a potential neural indicator of emotional state. We term these indicators as neural-congruency scores and their corresponding eigenvectors as neural-congruency components.

Spatial and temporal profiles of neural-congruency components.

Given the solutions to the generalized eigenvalue problem, the temporal profile of each components is calculated as the product of each component $\boldsymbol{w}_k$ with the EEG recordings of each video viewing. Moreover, the topographical profile (i.e. the forward model) of each component is calculated as

$$\boldsymbol{a}_k = \frac{R^{(w)}\, \boldsymbol{w}_k}{\boldsymbol{w}_k^{\mathrm{T}} R^{(w)} \boldsymbol{w}_k}$$

The temporal profile provides a visual representation of the modulation of neural activity relating to a particular emotion during the viewing of the video. Visual inspection of the temporal components can provide insights as to which sections of the video elicit the strongest emotional responses. Similarly, the spatial profile provides a topographical map that captures the covariance between each component's activity as measured by each electrode and can be used to estimate the areas of the brain the elicit the neural activity (i.e. using source estimation algorithms)

Relation of neural-congruency to emotion scores.

To evaluate the ability of the proposed neural-congruency metrics to predict the emotional content of each video, we used linear regression to model the relation between each of the extracted components to valance, arousal, and dominance scores. We report the explained variance of the model and regression statistics corrected for multiple comparisons using the false discovery rate.

Results

We applied the proposed method for extracting informative neural components on the EEG measurements of 32 participants watching the 40 music videos to identify neural-congruency components. The solution to the generalized eigenvector problem in equation 2, resulted in a total of 32 eigenvectors, of which we consider the eight with the highest associated eigenvalues. The value for the number of selected components (K=8) roughly corresponds to the knee point of the eigenvalue-spectrum of the auto-covariance matrix $R^{(w)}$. The forward models of the eight selected components and their associated eigenvalues are shown in Figure 1. The topography of each forward model informs of the approximate location of the underlying neuronal activity eliciting the components; while the associated eigen value shows the degree, this neural activity is "synchronously" observed across participants.

A correlation analysis was performed between each of the resulting neural components and each of the three emotion measurements. Correlation analysis showed a strong positive correlation between the seventh neural-synchrony component (i.e. the component with the 7th highest eigenvalue) and the population-wide valance scores (r=0.49, p<0.001); A strong positive correlation exists between neural-synchrony component and the dominance metric (r=0.39, p < 0.01). No correlation was shown between the 7th component and the arousal matric (r=0.09, p> 0.57). None of the other components showed a strong or moderate correlation between the three emotion-metrics after correction for multiple comparison using false discovery rate.

Moreover, we sought to investigate whether the neural congruency extracted with the proposed framework is predictive of population-wide emotional experiences associated with each video. In particular, we considered three univariate regression models, each respectively uses one of the population-wide emotion scores as its dependent variable (i.e. valance, arousal, dominance) and the neural-congruency scores of component 7 as an independent variable. Regression analysis shows that neural congruency score significantly predicted population-wide valance scores, b=0.4836, t(38)=3.473, p<0.001; the neural-

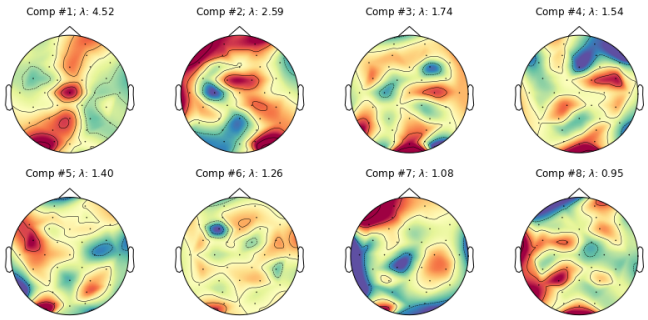

Figure 1: Forwards model of the eight components with highest inter-subject correlation extracted by the proposed method. Components are ordered from highest to lowest correlation; lambda corresponds to the eigenvalue associated with each component.

congruency score explained 24% of variance in popula-

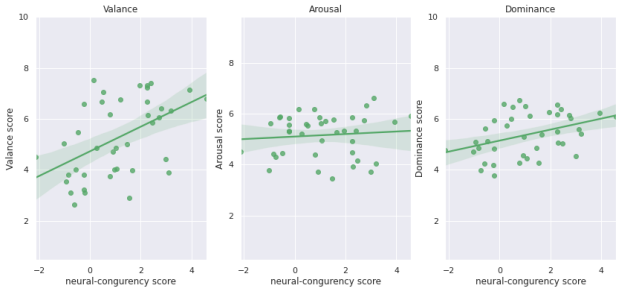

Figure 2: Scatter plots showing the regression line of the three models; each using the neural-congruency scores of component 7 as an independent variable, and respectively the valance, arousal and dominance scores as the dependent variable.

tion-wide valance (R^2 = 0.24; F(1,38)=12.06; p < 0.001). Neural-congruency also significantly predicted population-wide dominance score b=0.2159, t(38)=2.669, p<0.01; and it explained16% of the variance in dominance (R^2 = 0.16; F(1,38)=7.126; p < 0.01). The neural congruency scores did not predicted the population-wide arousal scores b=0.0495, t(38)=0.565, p>0.57); F(1,38)=0.3196;). The scatter plots in figure 2 show the regression line for the three models.

Finally, we inspected the correlation between the three observed emotion matrices to each other. Correlation analysis shows a strong positive correlation between valance and dominance metrics ($r=0.8$, $p<0.0001e-7$); the analysis also showed a moderate negative correlation between arousal and dominance metrics ($r=0.48$, $p< 0.002$). A weak correlation between arousal and valance we observed; however, it failed to reach statistical significance ($r=0.08$, $p>0.06$). We further checked the correlation between the three emotion-metrics and population-wide, self-reported likeability scores for each video. The analysis shows a positive correlation between the valance and likeability, $r=0.78$, $p< 0.001$; and dominance and likeability, $r=0.46$, $p<0.002$).

Discussion and Conclusion

In this paper, we propose a novel approach to isolate neuronal components elicited during the viewing of emotionally loaded musical video clips and are informative of the emotional content of those videos. Specifically, we formulated an optimization problem to extract spatial components from EEG measurements that maximize the correlation across viewings and subjects. Based on the resulting optimal components, we defined a set of neural-congruency metrics which we then evaluated as potential indicators of the overall emotional content of each video, as experience by viewers in terms of valance, arousal, and dominance. Moreover, each of the extracted components is associated with a corresponding temporal and spatial profile. These profiles enable researchers to inspect and interpret the spatiotemporal origins of the underlying neural activity and thus study the neural-correlates on emotions. The neural-congruency metric is validated, on a benchmark dataset, to carry predictive information as to the level of valance and dominance each video elicits to viewers.

Our results demonstrate the neural-congruency components extracted using our approach carry predictive information as to the underlying emotional metrics associated with the music video in the DEAP dataset. In particular, the neural-congruency component (component 7) explains 24% of the variance ($R=0.49$) in valance scores and 16% of the variance in dominance scores ($R=0.39$); both results were statistically significant. These results constitute novel evidence that the synchrony in neuronal activity in EEG measures, extracted by our method, can be an informative metric that can be used in EEG-based emotion recognition and to evaluate the emotional content of videos. In comparison, results reported in (Koelstra et al. 2011) on the same DEAP benchmark dataset showed that traditional power-based features on few selected channels exhibit only a small, correlation with the valance scores ($R<0.08$, the average across participant; indicatively, valance x theta-power on channel PO4: $R=0.05$; valance x alpha power in channel PO4: $R=0.05$; valance x beta power in channel CP1, $R=-0.07$; valance x gamma power in channels CP6, CP2, and C4, $R=0.08$), albeit statistically significant. The proposed neural-congruency components demonstrate a 4-fold increase in correlation compared to traditional power-based features. Thus, our results suggest that the synchrony in EEG signals can be predictive of emotional state and can serve as an informative metric for EEG-based recognition systems.

Our proposed method learns the neural components for the data and does not rely on heuristic, hand-selected features such in the case with frontal-asymmetry and frequency-power approaches (Petrantonakis and Hadjileontiadies 2011). Moreover, unlike discriminant deep-neural networks approaches (Lue et al. 2020; Zhong et al. 2020; Alhagry et. al 2017) that focus on differentiating between categorical variables of high vs low valance, arousal, and dominant, the proposed neural-cognitive congruency metric provide regression scores; hence it provides granular, real-value ratings for the EEG-based emotion recognition for each video. Importantly, our approach not only serves as a neural indicator of emotions during emotional video viewing but also isolates spatial and temporal profiles of each informative neuronal activity. These profiles can be used to study the temporal dynamics of emotion during video and estimate which brain areas elicit the neural activity. As noted in the introduction, paradigms for studying emotions have been limited to simplistic image stimuli. Hence, the interpretability of the spatial-temporal neural profiles makes them a valuable tool in neuroscience research and in the study of emotions during video viewing, opening new frontiers in emotion understanding.

In conclusion, we proposed a novel approach for EEG-based recognition that extracts interpretable and informative neural components that predict the emotional experience of viewers watching the video. Our approach can find applications in the design of emotion-aware AI systems, HCI, emotional video tagging, and the study of human emotions during video content. In future research, we plan to further explore the per-subject modulation of the resulting components, as well as extending our proposed method of EEG signals filtered in specific frequency bands.

Reference

Alhagry, S.; Fahmy, A.; and El-Khoribi, A. 2017. Emotion recognition based on eeg using LSTM recurrent neural Networks, *Emotion* 8 (10) pp. 255-258.

Christoforou, C.; Constantinidou, F.; Shoshilou, P.; and Simos, P. 2013.Single-trial linear correlation analysis: application to characterization of stimulus modality effects. *Frontiers in Computational Neuroscience* 7, 15

Christoforou, C.; Haralick, R.M.; Sajda, P.; and Parra L.C. 2010. The bilinear brain: towards subject-invariant analysis, *In 2010 4th International Symposium on Communications, Control and Signal Processing (ISCCSP)*, pp. 1–6. IEEE.

Christoforou, C.; Hatzipanayioti. A.; and Avraamides M. 2018. Perspective Taking vs Mental Rotation: CSP-based single-trial analysis for cognitive process disambiguation. *In Wang, S., Yamamoto, V., Jianzhong S., Yang Y., Jones, E., Iasemidis, L., Mitchell, T., (Eds.) Proceedings of International Conference, Brain Informatics* (pp. 109-199). Arlington, TX, USA.

Christoforou, C.; Papadopoulos, T.C.; Constantinidou, F.; Shoshilou, P.; and Theodorou, M. 2017. Your Brain on the Movies: A Computational Approach for Predicting Box-office Performance from Viewer's Brain Responses to Movie Trailers. *Frontiers in Neuroinformatics* 7, 15

Cohen, S.S.; Madsen, J.; Touchan, G.; Robles, D.; Lima, S.F.A; Henin, S.; and Parra, L.C. 2018. Neural engagement with online educational videos predicts learning performance for individual students, *Neurobiology of Learning and Memory*.

Davidson, R.J. 1992. Anterior cerebral asymmetry and the nature of emotion. *Brain Cognition*, pp. 75-80.

Dmochowski, J.P.; Bezdek, M.A.; Abelson, B.P; Johnson, J.S.; Schumacher, E.H.; and Parra, L.C. 2014. Audience preferences are predicted by temporal reliability of neural processing. *Nature Communications*.

Koelstra, S.; Muhl, C.; Soleymani, M; Lee, J.; Yazdani, A.; Ebrahimi, T.; Pun, T.; Nijholt, A; and Patras, I. 2011. DEAP: A database for emotion analysis using physiological signals, *IEEE Transactions on Affective Computing* 3 (1) pp. 18-31

Liu, Y.; Ding, Y.; Li, C.; Cheng, J.; Song, R.; Wan, F.; and Chen, X. 2020. Multi-channel EEG-based Emotion Recognition via a Multi-level Features Guided Capsule Network, *Computers in Biology and Medicine*

Petrantonakis, P.C.; and Hadjileontiadies, L.J. 2011. A novel emotion elicitation index using frontal brain asymmetry for enhanced EEG-based emotion recognition. *IEEE Trans. Info. Technolo. Biomed.*, 15 (5), pp. 737-746.

Philiastides, M.G.; and Sajda, P. 2005. Temporal characterization of the neural correlates of perceptual decision making in the human brain. *Cerebral Cortex* 16, 509-518

Zhong, P.; Wang, D.; and Miao, C. 2020. EEG-Based Emotion Recognition Using Regularized Graph Neural Networks, *IEEE Transactions on Affective Computing*.

Appendix

In this section we provide the derivation the solution to the optimization problem defined in the methods section 2. Recall we are trying to maximize the expression in equation (1) with respect to the vector w. Under the assumption that $\boldsymbol{w^T} R_{ii} \boldsymbol{w} \approx \boldsymbol{w}^T R_{ii} \boldsymbol{w}\ \forall (i,j)$, we define the covariance matrix in equation (1), with respect to the average covariance matrix $R^{(w)} = \frac{1}{P} \sum_{i \in P} R_{ii}$ and we can re-write equation (1) as follows:

$$\rho(\boldsymbol{w}) = \frac{1}{M} \sum_{v \in V} \sum_{\substack{(i,j) \in S \times S \\ i \neq j.}} \frac{\boldsymbol{w}^{\mathrm{T}} \boldsymbol{R}_{ij}^{v} \boldsymbol{w}}{(\boldsymbol{w}^{\mathrm{T}} \boldsymbol{R}^{(w)} \boldsymbol{w})}$$

Taking the derivative of $\rho(\boldsymbol{w})$ with respect to **w** and setting it to zero we get the following

$$\frac{\partial \rho(\boldsymbol{w})}{\partial \boldsymbol{w}^T} = \frac{1}{M} \sum_{v \in V} \sum_{\substack{(i,j) \in S \times S \\ i \neq j}} \frac{R_{ij}\, \boldsymbol{w}(\boldsymbol{w}^T R^{(w)} \boldsymbol{w}) - R^{(w)} \boldsymbol{w}(\boldsymbol{w}^T R_{ij} \boldsymbol{w})}{(\boldsymbol{w}^T R^{(w)} \boldsymbol{w})^2}$$

$$= \frac{1}{(\boldsymbol{w}^T R^{(w)} \boldsymbol{w})^2} \left(\frac{1}{M} \sum_{v \in V} \sum_{\substack{(i,j) \in S \times S \\ i \neq j}} R_{ij} \right) \boldsymbol{w}(\boldsymbol{w}^T R^{(w)} \boldsymbol{w})$$

$$- R^{(w)} \boldsymbol{w}(\boldsymbol{w}^T \left(\frac{1}{M} \sum_{v \in V} \sum_{\substack{(i,j) \in S \times S \\ i \neq j}} R_{ij} \right) \boldsymbol{w})$$

Setting the derivative to zero, we get:

$$R^{(b)} \boldsymbol{w}(\boldsymbol{w}^T R^{(w)} \boldsymbol{w}) - R^{(w)} \boldsymbol{w}\left(\boldsymbol{w}^T R^{(b)} \boldsymbol{w}\right) = 0$$

$$=> R^{(b)} \boldsymbol{w}(\boldsymbol{w}^T R^{(w)} \boldsymbol{w}) = R^{(w)} \boldsymbol{w}\left(\boldsymbol{w}^T R^{(b)} \boldsymbol{w}\right)$$

$$=> R^{(b)} \boldsymbol{w} = R^{(w)} \boldsymbol{w} \frac{\left(\boldsymbol{w}^T R^{(b)} \boldsymbol{w}\right)}{\left(\boldsymbol{w}^T R^{(w)} \boldsymbol{w}\right)}$$

$$=> R^{(b)} \boldsymbol{w} = \lambda R^{(w)} \boldsymbol{w}$$

where we set $\lambda = \frac{(w^T R^{(b)} w)}{(w^T R^{(w)} w)}$ and $R^{(b)} = \left(\frac{1}{M} \sum_{v \in V} \sum_{\substack{(i,j) \in S \times S \\ i \neq j}} R_{ij} \right)$. Thus, the optimal **w** is a solution to the generalized eigenvalue problem of equation (2)