Case 2

02582 Computational Data Analysis

s226765, s233576, s243636, s243933

May 5, 2025

Introduction

The focus of this project is to investigate the relationship between physiological responses and self-reported emotions during cognitively demanding tasks. Specifically, our research question is: **Do self-reported emotional states reflect physiological signals during a puzzle-solving task?**

To uncover underlying patterns, we apply a clustering-based unsupervised learning approach with the goal of revealing latent groups within the data. Our clustering analysis may reflect underlying differences in emotional awareness, stress sensitivity, or regulation strategies, offering insights into how people experience and report emotions under cognitive load.

Data Description

The EmoPairCompete dataset was collected from 28 participants (26 after filtering) using the Empatica E4 wristband during a structured, emotionally and cognitively demanding task. Each participant completed four rounds consisting of three phases: resting, puzzle-solving (stress), and recovery. Emotion self-reports were collected after each phase using questionnaires.

The stress-inducing task was a Tangram-based competition performed in pairs, designed to elicit cognitive load and emotional responses through challenge, communication, and time pressure. The experiment was conducted in three separate sessions across the year to introduce natural variation.

Physiological signals - like heart rate, skin temperature, and electrodermal activity (EDA-P and EDA-T) - were pre-processed into statistical features such as mean, standard deviation, skewness, and others. Self-reported emotions were assessed using the I-PANAS-SF and visual analogue scales for frustration and task difficulty.

Overview of the Data

For this project, we exclusively used the data contained in *HR_data_2.csv*, as it provided all the necessary information for our analysis. This file includes preprocessed physiological features collected from the Empatica E4 wristband, such as heart rate (HR), skin temperature (TEMP), and electrodermal activity - both phasic (EDA-P) and tonic (EDA-T) components - as well as the self-reported measures obtained through questionnaires. This data is organized per participant, per round, and per experimental phase.

While additional files like *response.csv* were available, they were not included in our analysis because all relevant self-report information from those files is already integrated into *HR_data_2.csv*. All of our relevant code and figures can be found in this GitHub repository.

Data Preparation and Exploration

Prior to applying any clustering methods, we went through data cleaning and segmentation as well as some exploratory data analysis (EDA) in order to have a better understanding of the data and its structure.

Data Cleaning

The first thing we looked at was the structure of the data at hand. We started by removing the irrelevant columns, such as "*Unnamed: 0*". Subsequently, missing values were identified. They were present in a couple of variables, like "*EDA_TD_P_RT*" and "*inspired*", for example. However, we verified that no missing values remained in the final dataset used for our main analysis, so all observations were retained.

Data Segmentation

To facilitate detailed analysis, the dataset was first divided into numerical data - where variables related to the main features present for further analysis (EDA, temperature, and heart rate) are present - and questionnaire data - where only data retrieved from the questionnaires conducted with the participants are included - capturing emotional states such as nervousness, alertness, inspiration, and attentiveness were separated as a distinct set.

Furthermore, the dataset was further segmented into subsets based on the type of physiological measure. This resulted in the separation of heart rate time-domain measures ("*HR_TD_**"), temperature-related features ("*TEMP_TD_**"), and electrodermal activity (EDA), which was further divided into phasic ("*EDA_TD_P_**") and tonic ("*EDA_TD_T_**") components. Identifier columns, such as "*Individual*", "*Phase*", "*Team_ID*", and "*Puzzler*" were also put together in the "*id_columns*" variable for grouping and filtering purposes.

Exploratory Data Analysis (EDA)

To ensure our problem formulation was grounded in the actual structure and characteristics of the dataset, we began with a focused preliminary analysis. This allowed us to understand the types of features available, data variability, and the relationship between physiological signals and task phases.

We visually explored the main physiological features - heart rate, skin temperature, EDA phasic, and EDA tonic - by generating histograms to examine their distributions and identify potential outliers among different phases. We then created feature subsets and used triangular correlation matrices and heatmaps to explore relationships between variables. In Figure 1, the histogram of the EDA phasic subset is visible.

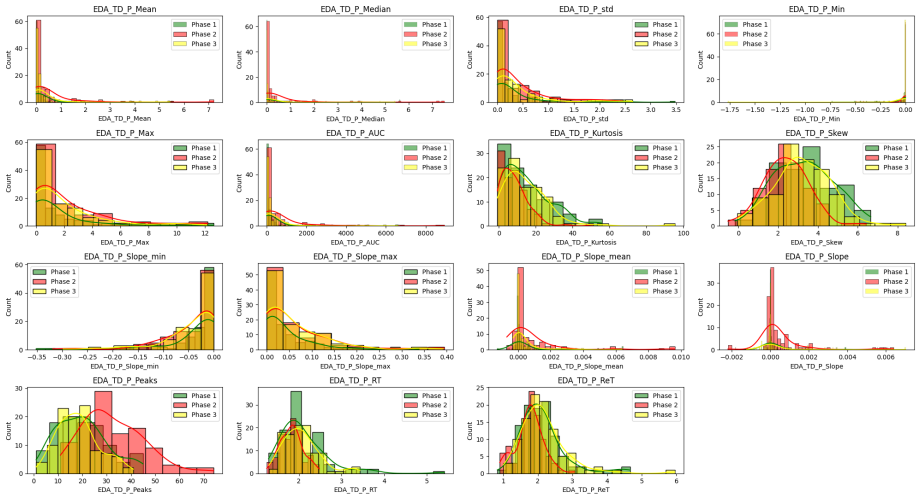

Figure 1: Histogram of EDA phasic parameters

In parameters, like "*EDA_TD_P_Peaks*" (bottom left), it is evident that there are visible differences between the phases of the experiment. As one might expect, during Phase 2 - where participants are solving the puzzles - the distribution deviates a bit, perhaps showing a more stressed or concentrated period of time. This will be further investigated in later sections.

We applied a similar approach to the self-reported questionnaire data, plotting distributions and computing correlations between evaluative self-reported responses.

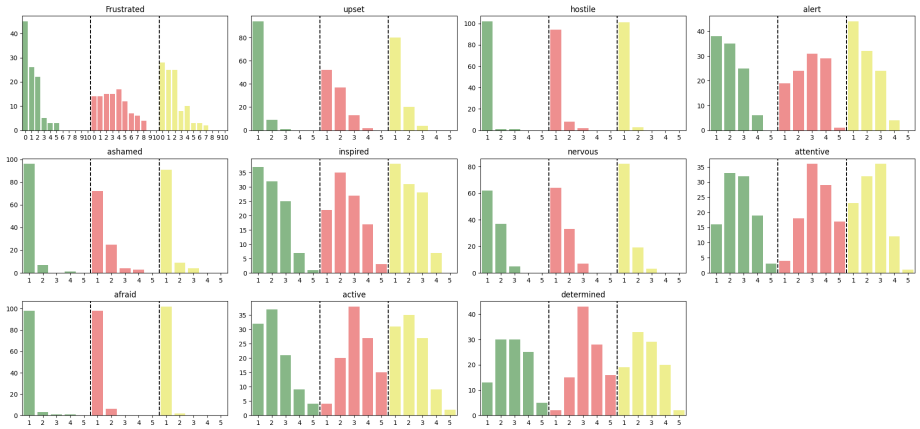

Figure 2: Histogram of questionnaire variables throughout phases

As can be seen in Figure 2, the histogram plots reveal phase-specific emotional responses (Phase 1 in green, Phase 2 in red, and Phase 3 in yellow). Most negative affect (e.g., frustration, upsetness, fear) peaked slightly during the active puzzle-solving phase (Phase 2), while positive engagement indicators such as determination also increased, suggesting cognitive and emotional arousal. Recovery patterns in Phase 3 indicate a partial return to baseline for most emotions. Therefore, this exploratory analysis helped us identify the most relevant features and interactions, which ultimately formed a clearer direction for our project.

Feature Selection

Given our EDA and our general aim, we decided to focus solely on the phase 2 data, as it represents the cognitively demanding task phase where participants were being stimulated. We then selected a subset of physiological features that were most relevant for capturing variations in stress responses during this phase: the EDA phasic variables. Multiple sources (including [1] and [2]) state that the phasic component of EDA represents the changes that occur during event-related activity, which in this experiment is the Tangram-based task. The statistics of this feature, such as mean, minimum, maximum, reaction time, peaks, etc., are all useful in determining the stress level of an individual.

The self-reported measure equivalent to stress is the "*Frustrated*" variable, which captures the frustration/stress of each participant on a scale of 0 to 10 after performing the task during the second phase. This subjective report serves as a representation of how stressed the individual thinks they are, whereas the physiological features represent how stressed the individual actually is. Using this, we can now answer the question of how well individuals are able to identify their stress levels by using the chosen subsets of the physiological and self-reported data.

Clustering Methods

To uncover patterns in the physiological data, we applied three different clustering techniques: K-means clustering, Hierarchical clustering and Gaussian Mixture Models (GMM). These methods were chosen for their ability to identify different types of structures in the data and their flexibility in handling various types of clusters.

The goal was to group individuals with similar phasic EDA responses during the task phase (Phase 2), which we specifically focused on. This specific focus allowed us to afterwards examine how these groups' physiological responses related to their self-reported levels of frustration in the questionnaire.

Before applying any clustering techniques, we standardized the relevant data to ensure consistency. This step was crucial, as both the clustering methods and Principal Component Analysis (PCA) are sensitive to the scale of the data. Additionally, given the relatively high number of variables compared to the number of samples, PCA was performed to reduce dimensionality and as it can help improving model performance and interpretation. Analyzing the explained variance, we determined that retaining three components accounted for over 90% of the total variance, making it an optimal choice for further analysis.

K-Means Clustering

Being a partition-based method, K-means clustering assigns data points to a predefined number of clusters K. It minimizes the within-cluster variance, iterating through steps of assigning data points to the closest cluster centroid and updating the centroids based on the assignments.

To choose an optimal number of clusters we relied on the elbow method: calculated and plotted the within-cluster sum of squares (WCSS) for different values of K and identified the point where adding more clusters no longer significantly reduces the inertia. This plot (Figure 3 in the Appendix) showed an inflection point at $K = 5$, suggesting that it was an appropriate choice for our data.

To visualize the results of how individuals were grouped, the clustering results for the first two principal components were plotted, which can be found in the appendix (Figure 4 in the Appendix). The clusters were clearly separated, indicating meaningful distinctions in the physiological data.

Hierarchical Clustering

Hierarchical clustering is an agglomerative method that builds a hierarchy of clusters by progressively merging data points based on their dissimilarity, which can be visualized through a dendrogram. We used Ward's linkage method, which minimizes the variance within each cluster and typically results in more balanced groupings.

By analyzing the generated dendrogram (Figure 5 in the Appendix), three seems to be the most reasonable number of clusters as it had the longest verti-

cal distance between merges, indicating a natural separation in the data structure. The cluster labels were then applied to the PCA-reduced data and visualized over the first two principal components (Figure 6 in the Appendix). This approach also showed distinct clusters based on individuals' physiological responses.

Gaussian Mixture Models

Being a more flexible clustering method, GMM assumes the data is generated from a mixture of several Gaussian distributions, each representing a different cluster, and estimates the parameters using the Expectation-Maximization algorithm. Each data point is assigned a probability of belonging to each cluster rather than being assigned to only one.

Given the high-dimensionality and low sample size in our dataset, we defined that all components share the same covariance matrix, which helps avoid overfitting and simplifies the model. Additionally, a regularization term of $1e-3$ was added to the covariance matrix to further stabilize the model estimation.

To determine the optimal number of components (K), we evaluated models with varying values of K using the Bayesian Information Criterion (BIC) and Akaike Information Criterion (AIC) scores, as these balance model complexity and goodness of fit. These were plotted in Figure 7 in the Appendix. We selected $K = 5$ as it provided a good trade-off, yielding one of the lowest BIC scores and relatively low AIC values compared to other values of K. After fitting the GMM to the PCA-reduced data, we visualized the resulting clusters. Similarly, it also produced distinct groups based on individuals' physiological responses, as seen in Figure 8 in the Appendix.

Method Selection

In order to evaluate clustering performance among the various models, we performed model validation by using the Davies-Bouldin metric. This measure is defined as the average similarity between the clusters, where similarity is the ratio of the distances within the cluster to the distances between clusters [3]. The lower the score, the better, as it implies that the clusters are far apart and dense.

We chose this over other measures as it does not assume the shape of the cluster to be circular, and it also serves as a measure to determine whether the clusters are equidistant, which is helpful when mapping the "*Frustrated*" scores from the questionnaire that are ordinal and equally spaced [4], [5]. The scores along with the model and the number of clusters used within each model are shown below in Table 1.

Model	No. of clusters	Davies-Bouldin score
K-means clustering	5	0.838
Hierarchical clustering	3	0.690
Gaussian Mixture Model	5	0.832

Table 1: The Davies-Bouldin scores for each clustering model.

According to these results, the hierarchical clustering model is our best model as it has the lowest score. The 3 clusters resulted by this hierarchical model are then chosen for the analysis of the aforementioned research question.

Results

As mentioned within the introduction, we aim to compare the physiological and self-reported measures to determine whether the individuals actually felt the stress that they reported they experienced during the task. In order to accomplish this, one measure is needed to report the "Felt stress" and another to report the "Perceived stress". The former is a result of the analysis of the physiological features with clustering models, and the latter is a categorical mapping of the "*Frustrated*" variable taken from the questionnaire features.

Using the clusters from the hierarchical clustering model, the mean of the features is calculated with respect to each cluster individually. To determine which feature should be used to assign stress level for "Felt stress", the most important feature of the first principal component is used. This ended up being the "*EDA_TD_P_AUC*" feature that represents the area under the curve of the phasic component of the EDA signal. Each cluster is then mapped to a "Felt stress" level based on this feature. The results are shown in Table 2.

Cluster	EDA_TD_P_AUC_Task	Felt stress
1	1408.764	Medium
2	131.291	Low
3	7328.021	High

Table 2: The felt stress categorical mappings of each cluster based on the AUC of the EDA phasic component.

The "Perceived stress" mappings are made by mapping certain ranges of the mean of the "*Frustrated*" variable obtained during Phase 2 over multiple rounds for each test subject individually. As the variable values varied between 0.5 to 5.5 with an increment of 0.25, the following categorical mappings shown in Table 3 are made to match the categories made for the "Felt stress" categories obtained in Table 2.

"Frustrated" values from questionnaire	Perceived stress
[0, 0.25, 0.5, 0.75, 1, 1.25, 1.5, 1.75, 2]	Low
[2.25, 2.5, 2.75, 3, 3.25, 3.5, 3.75, 4, 4.25]	Medium
[4.5, 4.75, 5, 5.25, 5.5, 5.75, 6]	High

Table 3: The perceived stress categorical mappings of category based on specific ranges of the "*Frustrated*" variable from the questionnaire.

Now that there are measures for both "Felt stress" and "Perceived stress" for each individual, we simply explore how many were able to rightly identify

their felt stress/frustration by simply equating the perceived and felt stress labels. It turns out that only 9 individuals were able to match their reported stress levels with what they actually felt! Among the 17 individuals who were not able to match their reported stress level to the actual felt stress level, 15 individuals reported more stress than what they actually felt, and consequently, 2 individuals reported less stress than what they actually felt.

Conclusions

Taking everything into account, our analysis highlights a clear mismatch between self-reported stress and physiological indicators. By comparing physiological responses with questionnaire-based frustration scores, we categorized each participant's felt and perceived stress levels.

This comparison revealed that only a small number of individuals accurately assessed their own stress (9 out of 26 in the study conducted). A majority (15 participants) overreported their stress, perceiving themselves as more stressed than their physiological signals suggested. This discrepancy points to a gap in stress self-awareness, likely influenced by cognitive bias, emotional state, or context rather than actual physiological arousal. Therefore, we conclude that relying solely on self-report may not provide a reliable assessment of an individual's true physiological experience.

References

[1] Cem Ersoy Yekta Said Can Bert Arnrich. "Stress detection in daily life scenarios using smart phones and wearable sensors: A survey". In: (2019). URL: https://www.sciencedirect.com/science/article/pii/S1532046419300577#s0055.

[2] A Dorsey et al. "Measurement of Human Stress: A Multidimensional Approach". In: (2022). URL: https://www.ncbi.nlm.nih.gov/books/NBK589926/#ch5.

[3] Yanchi Liu; Zhongmou Li; Hui Xiong; Xuedong Gao; Junjie Wu. "Understanding of Internal Clustering Validation Measures". In: (2010). URL: https://ieeexplore.ieee.org/document/5694060.

[4] Unknown. "Why Davies-Bould chose a number ob cluster higher than Silhouette or Calinsky Harabasz?" In: (2021). URL: https://datascience.stackexchange.com/questions/78345/why-davies-bould-chose-a-number-ob-cluster-higher-than-silhouette-or-calinsky-ha.

[5] Haitian Wei. "How to measure clustering performances when there are no ground truth?" In: (2020). URL: https://medium.com/@haataa/how-to-measure-clustering-performances-when-there-are-no-ground-truth-db027e9a871c.

Appendix

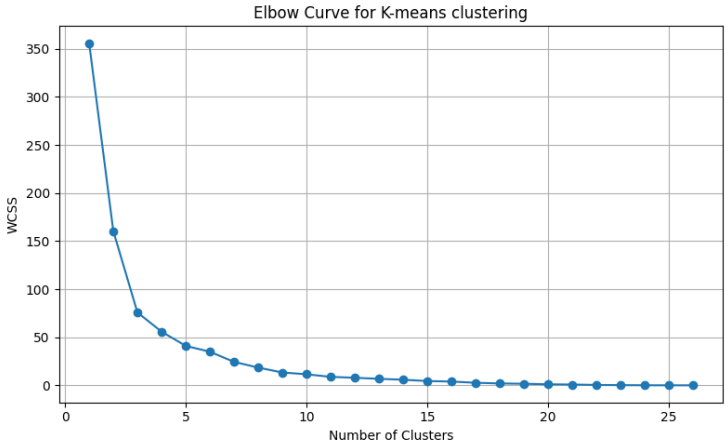

Figure 3: Elbow plot for K-means clustering

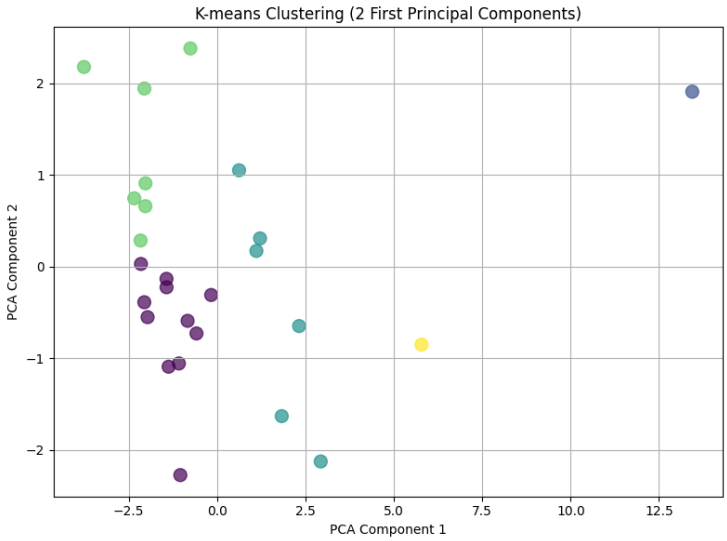

Figure 4: K-means clustering of the first two principal components

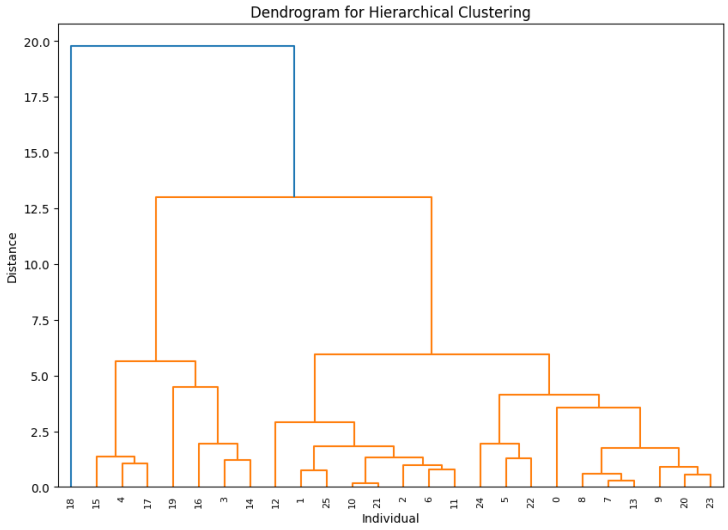

Figure 5: Dendrogram showing the hierarchical clustering process

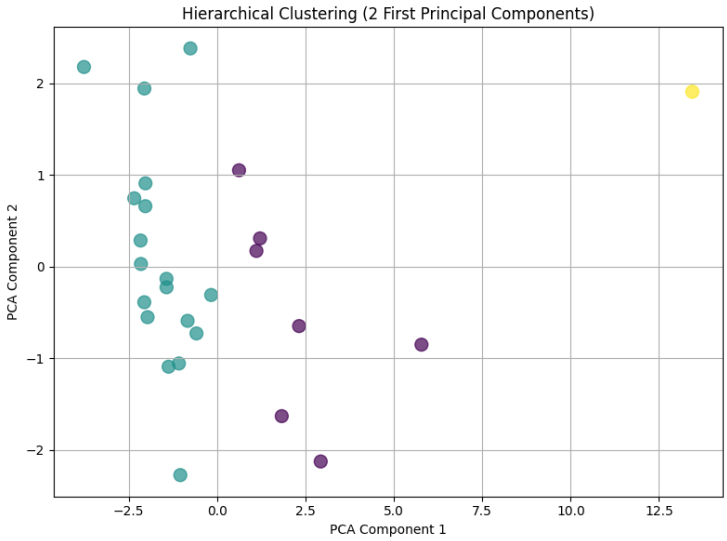

Figure 6: Hierarchical clustering of the first two principal components

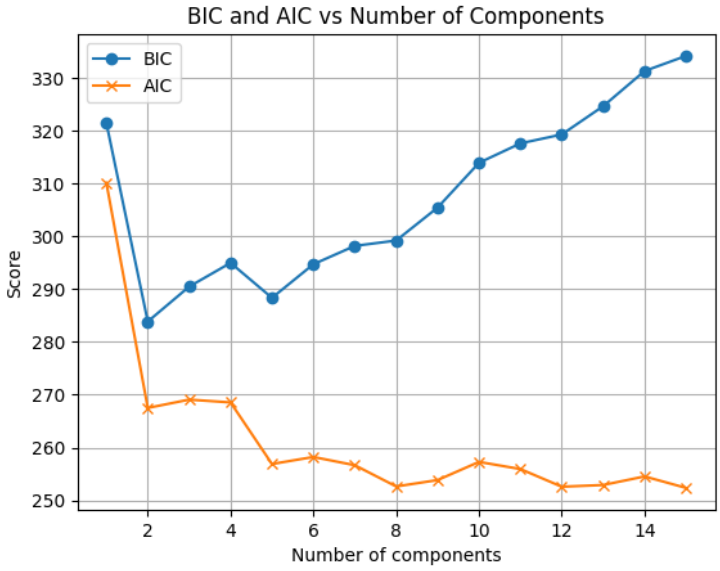

Figure 7: BIC and AIC values for different numbers of components in the GMM

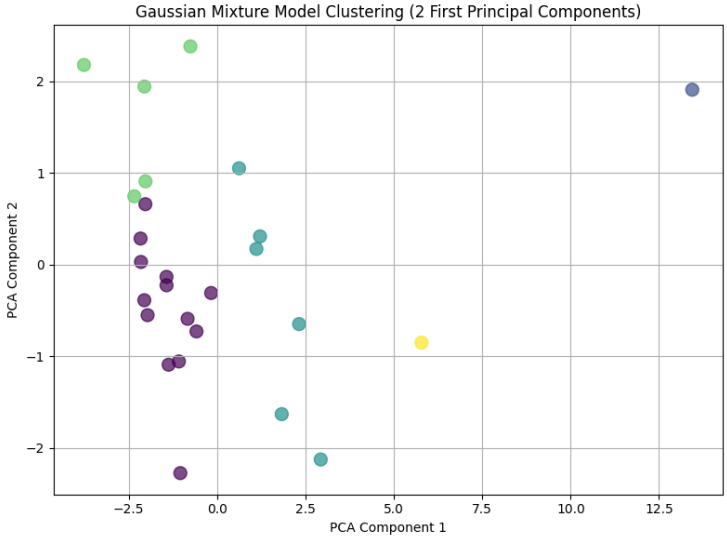

Figure 8: GMM of the first two principal components