Case 2

02582 Computational Data Analysis

Rita Saraiva (205717) & Christian Valentin Kjær (s211469)
& Eleonora Failla (s243151)& Francesca Mees (s242783)

May 8, 2025

1 Introduction and data description

Wearable biosensors such as the Empatica E4 provide continuous, multi-modal recordings that encode both tonic and phasic aspects of the wearer's physiology. The E4 biosensor captures 26 participants' heart-rate (HR), electro-dermal activity (EDA) and skin-temperature (TEMP) signals while they complete a four-round, stress-inducing tangram task. After each *rest* or *puzzle* phase, a questionnaire is collected where the emotional evaluation is noted.

Research question. Our goal is to assess how undergoing the stress condition (i.e., the puzzle phase) affects participants' emotional and physiological states. To investigate this, we fit a clustering model separately within each phase *rest*, *stress*, and *recovery* and track how participants move between clusters across phases. This approach is based on the assumption that each cluster captures a distinct, cohesive psychophysiological state. Observing consistent transitions (e.g., from a low-arousal cluster at rest to a high-arousal cluster during stress) would indicate that the task induces a measurable shift in internal state. Our central question is therefore: *does exposure to the stress condition lead to systematic and interpretable changes in cluster membership across phases?*

1.1 Data Preprocessing

The dataset preprocessing pipeline consists of several steps to prepare the data for analysis:

1. **Data Loading**: The raw dataset was loaded from a CSV file.
2. **Categorical Data Restructuring**: Categorical variables were converted to more practical numeric types:
   - `Cohort`: Prefix `D1_` was removed and converted to integer
   - `Round`: Prefix `round_` was removed and converted to integer
   - `Phase`: Prefix `phase` was removed and converted to integer
3. **Missing Value Handling**: Analysis of the dataset (shape: 312 rows × 68 columns) revealed a minimal amount of missing data, with only 5 observations having NA values across 7 parameters. Given this sparse distribution of missing values, we opted to replace them with the median of their respective columns rather than removing the observations entirely. The median was chosen over the mean to preserve robustness against potential outliers while maintaining the statistical properties of the dataset.
4. **Data Segregation**: The dataset was divided into three distinct subsets:
   - Biometric signals (columns 1-51)
   - Statistical metadata (Cohort, Individual, Round, Phase, Puzzler)

- Emotional indicators (12 emotion-related columns)

5. **Normalization**: Biometric signals were centered and normalized using `StandardScaler`[1] to ensure all features contribute equally to the analysis.

6. **Feature Rescaling**: The `Frustrated` emotional indicator was scaled by a factor of 0.5 to align its range with other emotional features.

7. **Data Integration**: The processed subsets were concatenated into a final dataset that preserves all necessary information in a standardized format.

The preprocessing steps ensure that the data is clean, normalized, and properly formatted for subsequent analysis while maintaining memory efficiency.

Model and method

2 Clustering methodology and model selection

To identify latent psychophysiological–affective states across different experimental phases, we explored several unsupervised clustering approaches introduced in the course: **K-means**, **Gaussian Mixture Models (GMM)**, and **Hierarchical Clustering**. Each model was implemented via a modular clustering class, allowing unified evaluation and parameter tuning.

2.1 Models considered

K-means. We began with K-means, which partitions the data into K spherical clusters by minimizing within-cluster variance. K-means is computationally efficient and simple to interpret, but assumes that clusters are of roughly equal size and shape. Since it relies on Euclidean distance, we standardized all features beforehand.

Gaussian Mixture Models (GMM). GMM generalizes K-means by modeling the data as a mixture of multivariate Gaussian distributions, each with its own mean and covariance. This allows for soft assignments and accommodates elliptical clusters with varying size and orientation. Fitting was done via the Expectation-Maximization (EM) algorithm, and the number of components K was selected using the Bayesian Information Criterion (BIC), as suggested in the course material[Hastie et al.(2001)].

Hierarchical Clustering. We also tested agglomerative hierarchical clustering with Ward linkage, which iteratively merges clusters that minimally increase the total within-cluster variance. This model does not require specifying K upfront and produces a dendrogram from which clusters can be extracted post hoc. However, in high-dimensional spaces, dendrograms can become hard to interpret and sensitive to noise.

[1] Subtract the mean and divide by standard deviation to obtain features with zero mean and unit variance.

2.2 Evaluation of cluster number (K)

To determine the optimal number of clusters, we employed three established techniques:

- **Silhouette score**: Measures how similar each observation is to its own cluster vs. the next-closest one. We applied this to both K-means and hierarchical clustering, scanning $K \in [2, 20]$.
- **Gap statistic**: Compares the within-cluster dispersion of real data to that of uniform reference samples. This was used for K-means to support or refine the Silhouette-based selection.
- **Bayesian Information Criterion (BIC)**: Used for GMMs to penalize overfitting as the number of components increases. We selected the K that minimized BIC, following the methodology from the lectures.

Each evaluation method was implemented in code (see `Clustering.get_optimal_k()`) and applied with K ranging from 2 to 20.

2.3 Final model selection

After comparing all methods, we selected **K-means** as our final clustering approach. All methods when applying the Silhouette evaluation method consistently identified K=2 as the optimal number of clusters, with K-Means achieving a Silhouette score of 0.34 compared to GMM's 0.35 and Hierarchical's 0.28. While GMM showed slight better separation, K-Means was preferred due to its superior interpretability and computational efficiency. Importantly, K-Means makes no assumptions about data distribution, unlike GMM which requires Gaussian-distributed features - an assumption violated by our PANAS questionnaire data. This makes K-Means more robust for our mixed data types. Hierarchical clustering performed significantly worse in both separation quality (Silhouette = 0.28) and computational efficiency.

K-means is an unsupervised classification algorithm that groups objects into k groups based on their characteristics. To select the optimal number of groups (k), we took into consideration the elbow method on the Within-Cluster-Sum of Squared Errors and the Silhouette Coefficient. The analysis of these parameters suggested that the optimal number of groups was three (see Appendix 4 for more details). Then, we performed feature importance analysis to discern which were the most important features for the model, using a cluster-based feature weighting technique based on which features minimize the within-cluster sum of squares the most. Finally, the psychophysiological features of each group were examined to define the groups in terms of their emotional states and physiological responses.

Summary of AI-pipeline

To summarize our AI workflow, we performed three main steps: data pre-processing, model selection, and clustering analysis. First, we cleaned, normalized, and structured the biometric and emotional data to prepare it for modeling. Next, we compared several unsupervised clustering algorithms and selected K-means based on the Silhouette score. Finally, we applied K-means clustering across experimental phases to uncover stable psychophysiological-emotional profiles and observe how participants transitioned between these states in response to stress. [AI-Pipeline(C2)]

Results

3 Association between participants perception of their induced emotions and data from E4 biosensors

The correlation matrix in Figure 1 shows the association between participants perception of the induced emotions and electrodermal activity features. We can appreciate the relationships between the electrodermal activity features (on the x-axis) and participants perception of their induced emptions features (on the y-axis).

The colors represent the strength and direction of the correlation, following the scale shown on the right (from +1 to -1). Positive correlations (red) indicate that as one variable increases, the other tends to increase. Negative correlations (blue) indicate that as one variable increases, the other tends to decrease. The asterisks indicate significant correlations, meaning that one asterisk (*) means a significant correlation at $p < 0.01$, and two asterisks (**) indicate a highly significant correlation at $p < 0.001$.

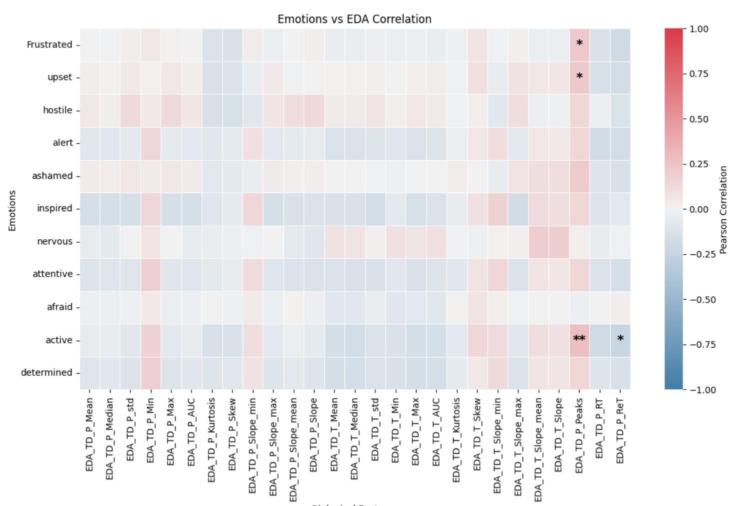


Figure 1: Association between participants perception of the induced emotions and electrodermal activity features, measured using Pearson's correlation. The significance testing was corrected for multiple comparisons using the Benjamini-Hochberg false discovery rate control method ($^{*}p < 0.01$, $^{**}p < 0.001$).

We observed that participants with better overall well-being (measured via PANAS) have more stable biosignals. We also observe that participants that are feeling more upset, frustrated and active have more phasic peaks often related to a greater level of responsiveness or emotional reactivity to the stimuli. While no significant correlations were identified between the participants emotions and the measured heart-rate or temperature features after correcting for multiple comparisons (see Appendix Fig. 5 and 6 for more details), the analysis revealed a slight positive correlation trend between perceived emotions and heart-rate features, and a slight negative correlation trend between perceived emotions and temperature features. The absence of significant correlations indicates that, within the scope of this study and the sensitivity of our measurements, there was no statistically reliable linear relationship observed between how participants subjectively experienced the induced emotions and these specific physiological indicators.

4 Clustering analysis

The results of the clustering analysis led to the identification of three different groups (Fig. 2 and see Appendix Fig. 7 for more details):

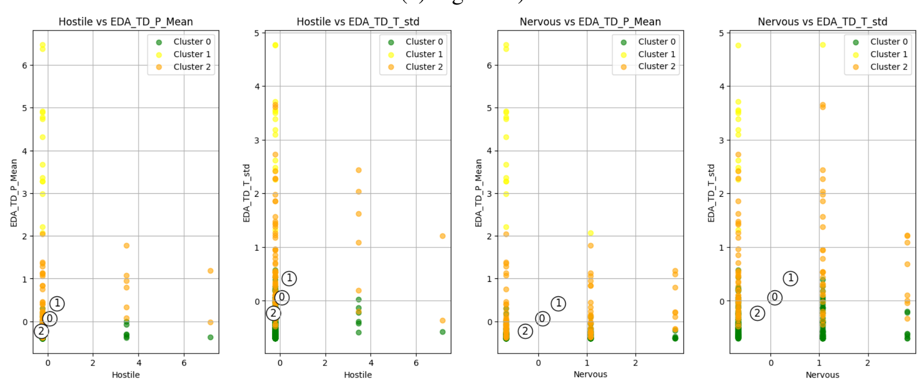
- Group 0: Moderate Emotional State with Low Physiological Responses (electrodermal activity)
- Group 1: Low Emotional Arousal with High Physiological Responses (electrodermal activity)
- Group 2: Moderate to Slightly Hostile/Frustrated State with Moderate Physiological Responses (electrodermal activity)

The graphical representation of these groups as a function of mean phasic EDA and standard deviation of the tonic EDA vs participant hostile allows this grouping to be clearly visualized (Figure 2b, left panels). The grouping is also visible when plotting mean phasic EDA and standard deviation of the tonic EDA vs participant nervous (Figure 2b, right panels).

The values of the representative points of each group for the top 12 most important variables (descending sorting) for the algorithm are presented in Figure 2a, and allow us to see common and differential characteristics of each group. In this case, several features show that clusters are primarily distinguished by phasic EDA components (e.g., rapid, stimulus-driven responses captured by the Response Slope Phasic or teh AUC of Phasic). Tonic features like Standard Deviation of Tonic suggest secondary slower baseline arousal differences between clusters. For example, the AUC of the Phasic is markedly lower in group 0 (-0.332) compared to group 1 (4.105).

		0	1	2
Most discriminative centroid feature values	**Emotional State**	Moderate	Low	Slight High
	Electrodermal Activity	Low	High	Moderate
	Median Phasic EDA	-0.279	4.145	0.015
	Area Under the Curve of Phasic EDA	-0.332	4.105	0.155
	Mean Phasic EDA	-0.330	4.096	0.153
	Mean Slope of Phasic EDA	-0.273	3.735	0.069
	Response Slope Phasic EDA	-0.240	3.371	0.046
	Standard Deviation of Phasic EDA	-0.445	3.318	0.570
	Maximum Phasic EDA	-0.469	3.281	0.636
	Maximum Slope of Phasic EDA	-0.451	3.250	0.595
	Standard Deviation of Tonic EDA	-0.445	3.186	0.590
	Minimum Slope of Phasic EDA	0.449	-3.138	-0.608
	Maximum Slope of Tonic EDA	-0.457	3.024	0.648
	Minimum Slope of Tonic EDA	0.437	-2.937	-0.611

(a) Figure A)



(b) Figure B)

Figure 2: **A)** Clinical labeling of the k-means groups health in terms of Emotional State and Electrodermal Activity (gray shading). Centroid feature values of the most discriminative features of the model. **B)** Visualization of clustering in terms of Mean Phasic EDA and Std of Tonic EDA, vs. participant Hostile and Nervous.

We also examined how the grouping changed pre-puzzle and post-puzzle rest, and puzzling. Figure 3 revealed consistent cluster distributions across experimental phases, suggesting minimal state transitions during the puzzle task. In the pre-puzzle rest (Phase 1), participants predominantly exhibited Moderate Emotional States with Low Physiological Responses (Cluster 0: 71.2%), with smaller proportions in Moderate-Hostile States and Moderate Physiological Responses (Cluster 2: 25%) and Low Arousal with High Physiological Responses (Cluster 1: 3.8%). During puzzling (Phase 2), these proportions remained stable (Cluster 0: 67.3%, Cluster 2: 26.9%, Cluster 1: 5.8%), indicating the task provoked minor shifts in psychophysiological states. Post-puzzle (Phase 3) distributions closely mirrored initial baselines (Cluster 0: 66.3%, Cluster 2: 29.8%, Cluster 1: 3.8%), demonstrating resilience to sustained state changes. This stability implies that: (1) individual baseline emotional-physiological profiles strongly predict their task-state responses, and (2) the puzzle activity did not substantially alter group-level response patterns. The slight increase in Cluster 2 (hostile/frustrated states) during active puzzling and post-puzzling (+1.8-4.8%) may suggests the task has elevated frustration/Hostile responses in a subset of participants.

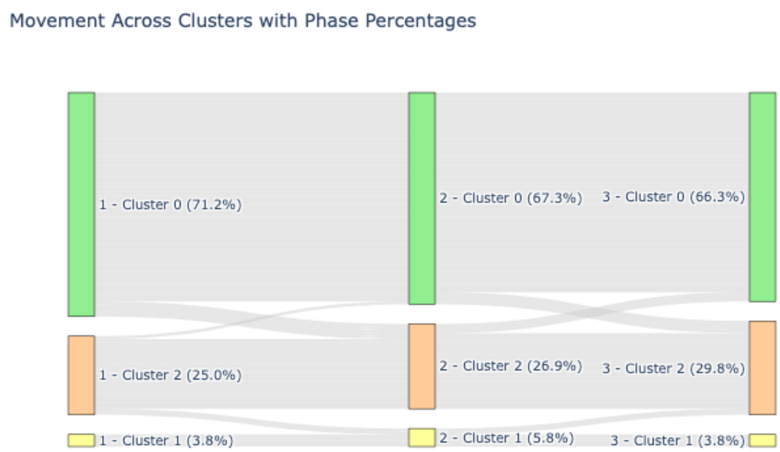


Figure 3: Sankey diagram representing which group each participant belonged Pre-puzzle (Phase 1), Puzzle (Phase 2) and Post-Puzzle (Phase 3).

Most Cluster 1 (high arousal) and Cluster 2 (frustrated/hostile) participants came from fall cohorts (D12/D13) (see Appendix Fig. 8 for more details), unlike winter Cohort 1 (D11). The consistency of cluster membership despite phase transitions implies that individual stress-response profiles were more strongly influenced by cohort-specific conditions than by the puzzle task itself.

Discussion

While this study provides valuable insights into transient frustration and individual physiological patterns, several limitations should be acknowledged. First, the controlled experimental setting may not fully reflect the complexity of real-world frustration experiences, reducing the applicability of the results to everyday contexts. Second, the subjective nature of self-reported emotional states introduces variability and potential bias. Lastly, the relatively small sample size limits the statistical power and generalizability of the findings. Future studies with larger and more diverse samples in real-life settings are necessary to confirm and expand on these results.

Conclusion

In this work, we explored whether unsupervised clustering models could reveal systematic and interpretable changes in cluster membership across different phases of stress exposure. We selected K-means clustering with three clusters, based on the Silhouette score. The three clusters identified were characterized by (1) a moderate emotional state with low physiological responses, (2) low emotional arousal with high physiological responses, and (3) a more hostile/frustrated emotional state with moderate physiological responses. The cluster memberships remained relatively stable across the different phases, suggesting that individual stress-response profiles were more influenced by cohort-specific conditions than by the task itself. Overall, we have demonstrated how K-means clustering can be used to identify and understand distinct emotional-physiological response patterns across different stress conditions.

References

[AI-Pipeline(C2)] AI-Pipeline. C2. Case 2 Pipeline. `https://github.com/Rita-Saraiva/CDA_Case_2`. Git repository.

[Hastie et al.(2001)] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. 2001. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer New York, New York, Chapter 14.3.7. Gaussian Mixtures as Soft K-means Clustering.

Appendix

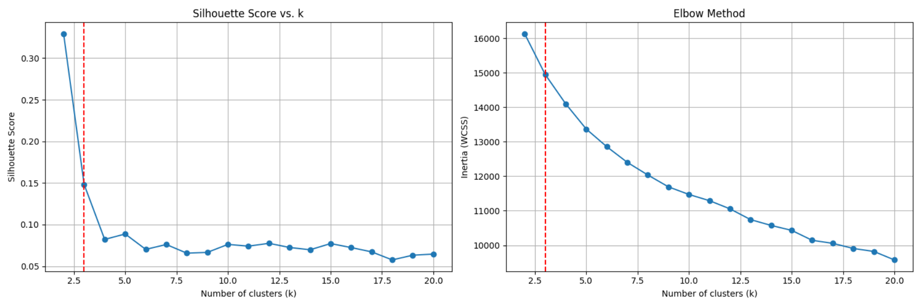


Figure 4: On the left, we can see the Silhouette coefficient as a function of the number of cluster. On the right, we can see the sum of squared distances as a function of the number of clusters. Considering we want the minimum sum of squared distances and the maximum Silhouette coefficient, the optimal number of clusters seems to be three.

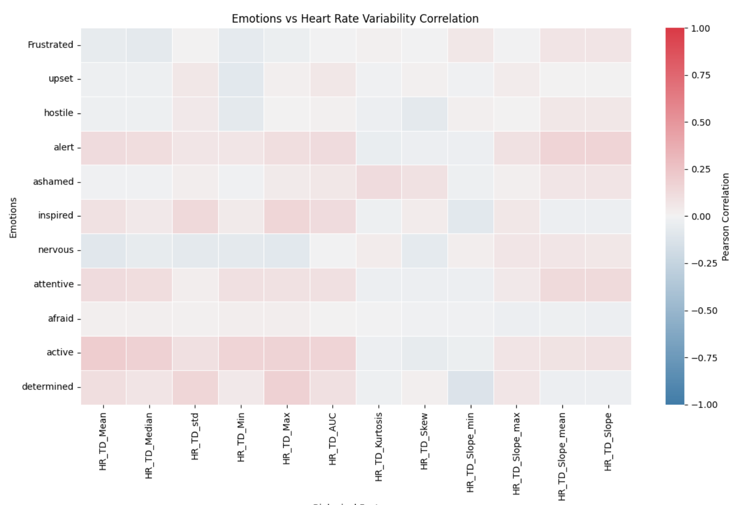


Figure 5: Association between participants perception of the induced emotions and heart-rate features, measured using Pearson's correlation. The significance testing was corrected for multiple comparisons using the Benjamini-Hochberg false discovery rate control method ($^{*}p < 0.01$, $^{**}p < 0.001$).

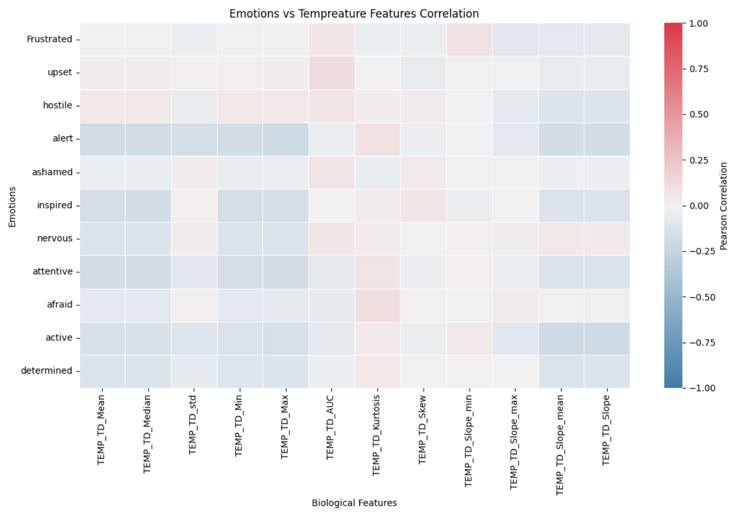


Figure 6: Association between participants perception of the induced emotions and temperature features, measured using Pearson's correlation. The significance testing was corrected for multiple comparisons using the Benjamini-Hochberg false discovery rate control method ($^{*}p < 0.01$, $^{**}p < 0.001$).

Figure 7: Centroid Emotional Features Mean Values of the K-means Groups.

Figure 8: Sankey diagram representing which group each participant belonged Pre-puzzle (Phase 1), Puzzle (Phase 2) and Post-Puzzle (Phase 3) for each Cohort. Each Cohort number representing when the experiment was run (three separate occasions). Cohort 1 (D11) were completed in the winter and Cohort 2 (D12) and Cohort (3,4,5,6) (D13) were completed in the fall. D13 were conducted in four separate sessions (Cohort 3 and 5 in the morning and Cohort 4 and 6 in the evening.